

Statistical Decision Making for Bivariate and Multivariate Analysis with Visirule

Ismail O. MURAINA^{1*}, Mukaila A. RAHMAN², Imran A. ADELEKE³

¹School of Science, Adeniran Ogunsanya College of Education Otto/Ijanikin, Nigeria

²Faculty of Science, Lagos State University, Nigeria

³ School of Science, Adeniran Ogunsanya College of Education Otto/Ijanikin, Nigeria

* corresponding author: niyi2all@yahoo.com

Abstract

Without any doubt, research is sometimes designed to solve a particular problem. In social sciences and education, data collected during research studies are often subjected to statistical analysis for deducing desirable results. Reported research that involved data collection without proper analysis to back these results often makes the work non-robust enough for wider acceptance in the research community. One of the greatest challenges often faced by amateur researchers and students is how to select appropriate statistical tools which are suitable and relevant to the data at hand. Having a firm knowledge of suitable statistical tool goes a long way in motivating and enhancing researcher's attitudes towards the work. This paper presents procedures necessary for researchers to select appropriate statistical instrument for a desired purpose. The Visirule software (a free software) was used as a decision supporting tool, in which the rules are basically and precisely presented graphically based on logic programming model. The bivariate and multivariate analysis discussed in this work can be of great use to researchers and to both undergraduate and post-graduate students in making a productive and realizable decision in utilizing suitable statistical data analysis in researches.

Keywords. Statistics; Visirule; Bivariate; Multivariate; Decision support tool

1. Introduction and Background

Research studying and writing is a sine-qua-non to all students in higher institutions to embark upon at last stage/level of their schooling. There are courses and topics that equip students to understand all basic requirements in research and statistical analysis yet many students' still find it confusing while other see it as a threat to their graduation. Reports have shown that students writing project move at a fast rate from chapter one to chapter three before many of them boycott or abandon the project totally or continue sporadically. To some students, taking clues after old projects is an alternative. These and many more reasons interest researchers in this study. Reading meaning to already collected data has remained a serious drawback to students writing projects. In this study, the researchers take a further step to look into proper way of selecting most relevant statistical tool to analyze a particular data at hand: focus is set towards bivariate and multivariate analysis/statistics using visirule as decision tree software.

Research simply means a search for knowledge. It can also be seen as a scientific and systematic search for pertinent information on a specific topic. [1,2] define research as a systematized effort to gain new knowledge while some people consider research as a movement, a movement from the known to the unknown. In fact, it is a voyage of discovery. It is a careful, scholarly, purposeful and systematic investigation or study that is designed to add or delete from a body of knowledge ([3]). To [4,5], research is the orderly investigation of a subject matter for the purpose of adding to knowledge. Research can mean "re-search" implying that the subject matter is already known but, for one reason or another, need to be studied again. Alternatively, the expression can be used without a hyphen and in this case it typically means investigation a new problem or phenomenon. Research in an academic activity and as such the term should be used in a technical sense. According to [6] research comprises defining and redefining problems, formulating hypothesis or suggested solution; collecting, organizing and evaluating data; making deductions and reaching conclusions; and at last carefully testing the conclusions to determine whether they fit the formulating hypothesis. In short, the search for knowledge through objective and systematic method of finding solution to a problem is research.

Statistics is a part of the everyday language of education and the other social sciences ([7,8]). These research based disciplines routinely use statistics to express knowledge and to discuss theory and research. Students as well as educationists should be literate in the vocabulary of research, data analysis and scientific thinking. Fluency in statistics will go a long way to help understanding the research reports one encounters in every day life. This will also help one to be able to conduct quantitative research, contribute to the growing body of education and reach full potential as an educational scientist. Although, essential learning statistics can be challenging. Students in writing projects should be equipped with knowledge necessary to conduct a successful statistical analysis in research.

In research, the kind of variable available determine the type of approach to the analysis. A variable according to [9] is any characteristics, number, or quantity that can be measured or counted. A variable may also be called a data item. Age, sex, business income and expenses, country of birth, capital expenditure, class grades, eye colour and vehicle type. All these are called variables because the value may vary between data units in a population and may change in value over time. Variables have types: Numeric variables – Which can be a continuous variable (Observations can take any value between a certain set of real numbers) or a discrete variable (Observations can take a value based on a count from a set of distinct whole values), categorical variables. Numeric variables have values that describe a measurable quantity as a number, like 'how many' or 'how much'. Therefore, numeric variables are quantitative variables while categorical variables which can be nominal (Observations can take a value that is not able to be organized in a logical sequence) or ordinal (Observations can take a value that can be logically ordered or ranked) have values that describe a 'quality' or 'characteristic' of a data unit, like 'what type' or 'which category'. Categorical variables fall into mutually exclusive (in one category or in another) and exhaustive (include all possible options) categories. Therefore, categorical variables are qualitative variables and tend to be represented by a non-numeric value.

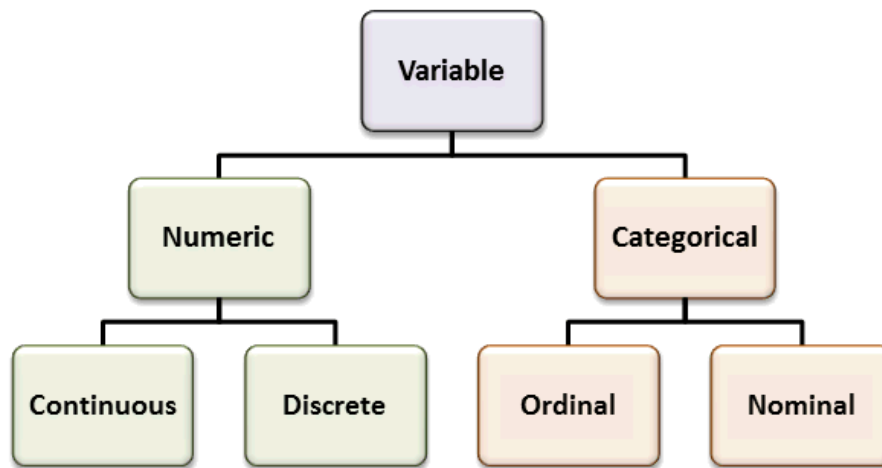


Fig. 1 A Summary of variable types under numeric and categorical

Similarly, variables can be classified into dependent and independent variables. When conducting research, according to [10] experimenters often manipulate variables. For example, an experimenter might compare the effectiveness of four types of antidepressants. In this case, the variable is "type of antidepressant." When a variable is manipulated by an experimenter, it is called an independent variable. The experiment seeks to determine the effect of the independent variable on relief from depression. In this example, relief from depression is called a dependent variable. In general, the independent variable is manipulated by the experimenter and its effects on the dependent variable are measured. Another way of looking at types of variables is qualitative variables and quantitative variables. Qualitative variables are those that express a qualitative attribute such as hair color, eye color, religion, favorite movie, gender, and so on. The values of a qualitative variable do not imply a numerical ordering while quantitative variables are those variables that are measured in terms of numbers. Some examples of quantitative variables are height, weight, and shoe size. After researchers look into the type of variables on ground it is pertinent to consider the type of statistics available for the type of variables. In research, the two major common classifications of statistics are parametric

statistics and non-parametric statistics. In [11] it was clearly stated that in the literal meaning of the terms, a parametric statistical test is one that makes assumptions about the parameters (defining properties) of the population distribution(s) from which one's data are drawn. Examples of parametric tests include:

- ✓ All forms of T-tests
- ✓ Pearson product Moment Correlation Coefficient Test
- ✓ All forms of ANOVA Test
- ✓ All forms of ANCOVA Test
- ✓ MANOVA Test

A non-parametric test is one that makes no such assumptions. In this strict sense, "non-parametric" is essentially a null category, since virtually all statistical tests assume one thing or another about the properties of the source population(s). Examples of non-parametric tests include:

- ✓ the various forms of chi-square tests,
- ✓ the Fisher Exact Probability test,
- ✓ the Mann-Whitney Test,
- ✓ the Wilcoxon Signed-Rank Test,
- ✓ the Kruskal-Wallis Test,
- ✓ and the Friedman Test.

[12] observed that one should use nonparametric tests when data don't meet the assumptions of the parametric test, especially the assumption about normally distributed data. This assumption may not be enough to use any test; it is also necessary to note that parametric analysis is used to test group means while non-parametric analysis is to test group medians. In research, A statistical hypothesis is an assumption about a population parameter. This assumption may or may not be true. Hypothesis testing refers to the formal procedures used by statisticians to accept or reject statistical hypotheses ([13]). A hypothesis test is a statistical test that is used to determine whether there is enough evidence in a sample of data to infer that a certain condition is true for the entire population.

A hypothesis test examines two opposing hypotheses about a population: the null hypothesis and the alternative hypothesis. The null hypothesis is the statement being tested. Usually the null hypothesis is a statement of "no effect" or "no difference". The alternative hypothesis is the statement one wants to be able to conclude is true.

Based on the sample data, the test determines whether to reject the null hypothesis. One should use a p-value, to make the determination. If the p-value is less than or equal to the level of significance, which is a cut-off point, then one can reject the null hypothesis.

A common misconception is that statistical hypothesis tests are designed to select the more likely of two hypotheses. Instead, a test will remain with the null hypothesis until there is enough evidence (data) to support the alternative hypothesis.

Condition for the use Parametric and Non-Parametric Statistics

Parametric Statistics

The parametric statistics is used when

- The dependent variable is measured at the interval or ratio level, that is, using a continuous scale rather than discrete categories.
- The samples should be selected randomly from the population
- The observations that make up data must be independent of one another. That is, each observation or measurement must not be influenced by any other observation or measurement.

Examples of parametric statistics are: T-TEST, ANOVA, ANCOVA, MANOVA, etc

Non-parametric Statistics

The non-parametric statistics is used:

- When measured on nominal (categorical) and ordinal (ranked) scale

- When they have very small samples
- When data do not meet assumptions of parametric techniques

Examples of non-parametric statistics are: Chi-square, Mann-Whitney V test, Spearman Rank-order Correlation, etc

The table1 below shows differences between parametric and non-parametric statistics according to [14].

Table1. Lists of Parametric and Non-parametric statistics

NON-PARAMETRIC	PARAMETRIC
Chi-square	No equivalent test
Mann-Whitney V test	Independent-sample t-test
Wilcoxon Signed Rank Test	Paired Sample t-test
Kruskal-Wallis Test	One way between group ANOVA
Friedman Test	One-way repeated measures ANOVA
Spearman Rank order Correlation	Pearson's product moment correlation

Correlation Techniques- Non-Experimental Design

Correlation techniques are used by researcher engaged in non-experimental designs WHERE predictor should correlate highly with dependent variable not other independent variables. Different kinds of correlation are as follow:

- (1) Explore the association between pairs of variables – **Correlation**
- (2) Predict scores on one variable from scores on another variable – **Bivariate regression**
- (3) Predict scores on a dependent variable from one score of a number of independent variables –

Multiple regression

- (4) Identify the structure underlying a group of related variables – **Factor analysis**

All these are used to test models and theories, predict outcomes and access reliability and validity of scales

Correlation Techniques- Experimental Design

Correlation

- To describe the strength and direction of the relationship between two variables
- Two continuous variables or one continuous and dichotomous variables (e.g two values, sex: male and female)
- Pearson Product-Moment co-efficient
- Non-parametric – spear-mans rank other correlation

Partial Correlation

- To explore the relationship between two variables while statistically controlling for a third variable. This is useful when one is suspecting that the relationship between two variables of interest may be influenced or confounded by the impact of the third variable
- Three variables: all continuous, two variables to explore the relationship and one variable to control

Multiple Regression

- To predict a single dependent continuous variable from a group of independent variables. It can be used to test the predictive power of set of variables and to access the relative contribution of each individual variable.
- One continuous dependent variable and two or more independent variables or one dichotomous independent variables e.g (sex: male & female)

Logistic Regression

- To allow one to test models to predict categorical outcome with two or more categories. Your predictor (independent) variables can either be categorical or continuous or a mix of both in one model
- One categorical (dichotomous) dependent variable (problem with sleep:) and two or more continuous or categorical predictor (independent) variables (note: we must code dichotomous variables as 0 and 1) (e.g sex, trouble getting to sleep, trouble staying asleep, coded as 0 and 1)

Factor Analysis

- This is useful in reducing a large number of related variables (items that make up a scale) to a smaller, more manageable, number of dimensions or components

Comparing Group – Experimental Design

Independent-Sample Test

- It is used when you want to compare the mean score of two different groups of people or conditions
- One categorical independent variable (males / females or experimental /control) and one continuous dependent variable (achievement score)

Paired Sample T-Test (Repeated Measure)

- It is used when you have only one group of people (students, companies etc) and one collects data from them on two different occasions or under two different conditions (pretest / posttest)
- One categorical independent variable (time 1 and time 2) and one continuous dependent variable (performance scores measured on two different occasions or under different conditions)

One-Way ANOVA(Between-group)

- When we are interested in comparing the mean scores of more than two groups. That is, when you have different subjects or cases in each of your groups (independent group)
- One categorical independent variable with three or more distinct categories (this can also be a continuous variables that has been recoded to give three equal groups e.g age group: 20-24 age, 25-29 age and 30-44 age.) and one continuous dependent variables (performance test score)

One-Way Repeated Measures ANOVA

- It is designed to measure ANOVA design that each subject is exposed to two or more different conditions, or measure on the same continuous scale more than one occasions.
- One independent variable (categorical: e.g time1/ time2/ time3) and one dependent variable (continuous: performance test score)

Two Ways Between Groups ANOVA

- It is used when there are two independent variables and between-groups indicates that different people are in each of the groups. The techniques allow us to look at the individual and joint effect of two independent variables on one dependent variable
- Two categorical independent variables (sex: Males/Females, Age: young, middle, old) and one continuous dependent variable (achievement scores)

Mixed Between-Within Subject ANOVA

- It is used when you have two independent variables: one is a between-subject variable (gender: males/females), the other is a within-subjects variable(time). In this case you would expose a group of both males and females to the treatment at time1(pre-test) and time2(post-test)
- One categorical independent between-subjects variable with two or more levels (group1/group2)
- One categorical independent within-subjects variable with two or more levels (time1/ time2/ time3) and one continuous dependent variable (scores of students in mathematics test measure at each time period)

Multivariate ANOVA

- MANOVA is an extension of ANOVA for use when you have more than one dependent variable. The dependent variables must be related in some way or there should be some conceptual reason for considering them together

- One categorical independent variable (e.g. sex) and two or more continuous dependent variables. MANOVA can also be extended to two-way and higher-order designs involving two or more categorical independent variables

One Way ANCOVA

- It is used when you have a two-group pretest/post-test design (e.g. comparing the impact of two different interventions, taking before and after measures for each group). The scores on the pretest, are treated as a covariate to `control` for pre-existing differences between the groups
- One categorical independent variable with two or more levels (group1/group2); One continuous dependent variable (score on mathematics test at time 2) and One or more continuous covariates (scores on mathematics test at time 1)

Two Way ANCOVA

- ANCOVA will control for scores on your covariate and then perform a normal two-way ANOVA. This will tell you if there is a significant main effect on your first independent variable (e.g. group). A main effect for your second independent variable (sex). A significant interaction between the two.
- Two categorical independent variables with two or more levels (sex Males/Females), group: Mathematics / Computer). One continuous dependent variables, one or more continuous covariates.

Comparing Group – Non Experimental Design

Chi-Square: The test is used when you wish to explore the relationship between two categorical variables.

- Two categorical variables, with two or more categories in each
- For 2 x 2 table use Fisher's Exact Probability Test

Mann Whitney V Test

- This test is used to test for differences between two independent groups on a continuous measure. Mann Whitney V Test actually compares medians
- One categorical variable with two groups e.g. sex. One continuous variable

Wilcoxon Signed Rank Test

- Also called Wilcoxon Matched pairs signed rank test is designed for use with repeated measures; that is when your subjects are measured on two occasions or under two different conditions. It converts scores to ranks and compares them at time1 and time2
- One group of subjects measured on the same continuous scale or criterion on two different occasions

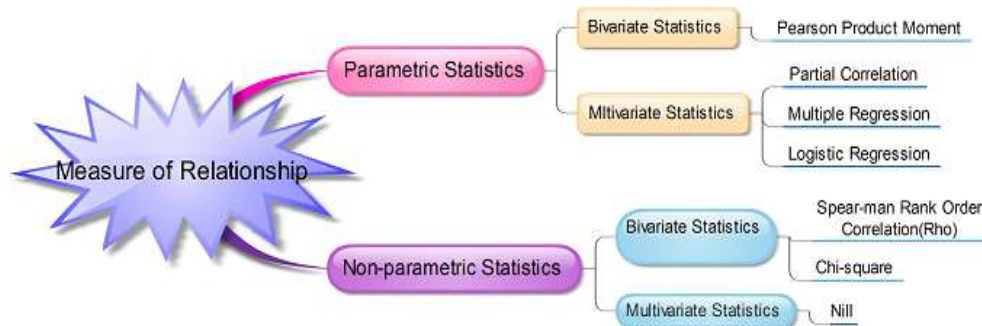
Kruskal-Wallis test

- This test (also known as Kruskal Wallis test) allows you to compare the scores on some continuous variables for three or more groups. It converts the scores to ranks and the mean rank for each group is computed
- One categorical independent variable with three or more categories (age 18-29, 30-44, 45+). One continuous dependent variable

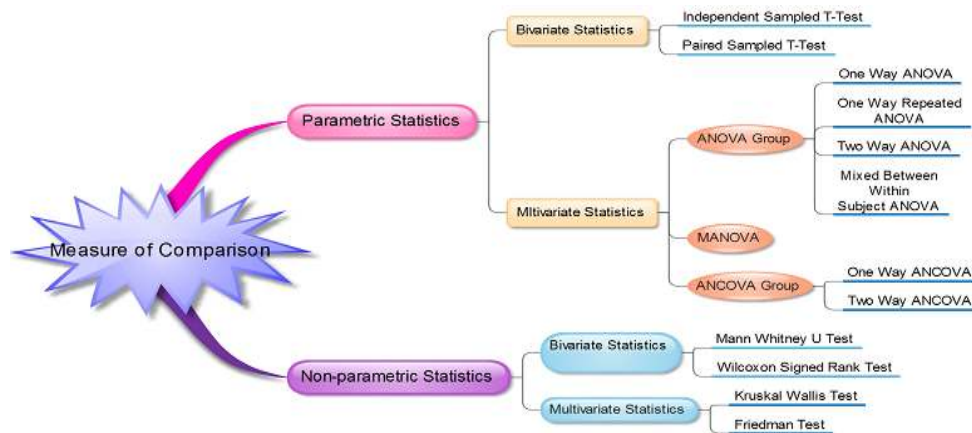
Friedman Test

- This test is used when you take the same sample of subjects or cases and you measure them at three or more points in time or under three different conditions
- One sample of subjects measured on the same scale or measured at their different time or periods or under three different conditions

Spearman's Rank Order Correlation (ρ) is used to calculate the strength of the relationship between two continuous variables



Summary Graph for Bivariate Statistics: **Figure 1B**



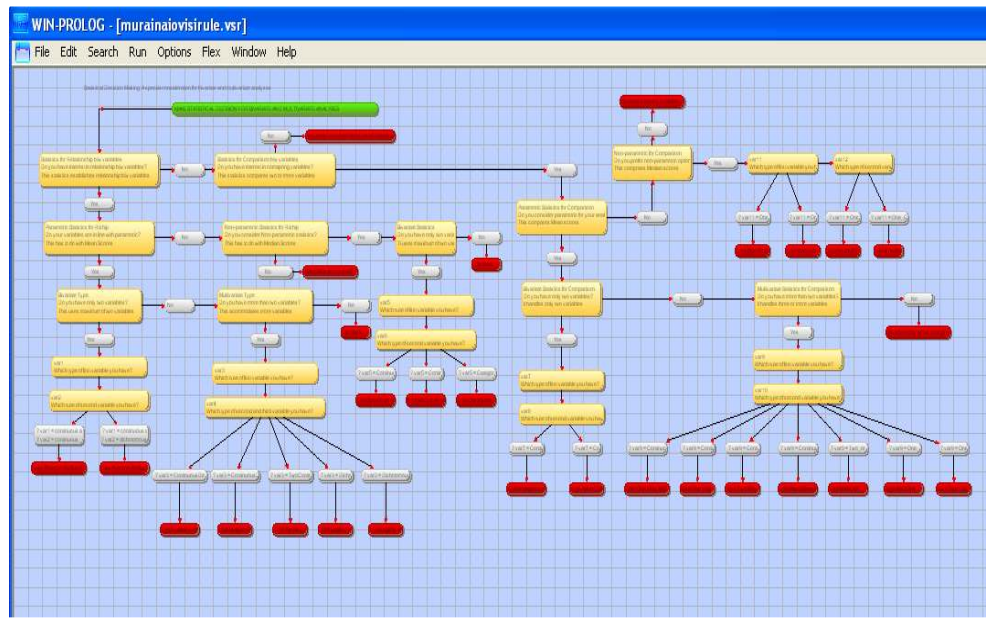
Summary Graph for Multivariate Statistics: **Figure 2**

2. Methods

This paper presents procedures necessary for researchers to select appropriate statistical instrument for a desired purpose. The Visirule software was used as a decision supporting tool, in which the rules are basically and precisely presented graphically based on logic programming model. Graphical illustrations were also given which clearly depicted the relationships between the statistics discussed. Subsequently, Visirule software was used to implement the easy selection of those discussed tools.

In this structure, more than twenty-three (23) questions were raised relating to particular structural analysis relevant to data collected, which are depicted by yellow single choice boxes. Followed by forty-five (45) white boxes indicating options to be taken by the proceeding questions. Finally, there are twenty-nine (29) red boxes implying the result or outcome (conclusion) which usually terminate or end series of inferences ([15,16]). Visirule was used by authors only covered bivariate and multivariate analysis.

The application of visirule will greatly assist the researchers, students and other interested people to just follow the questions based on the kind of data or variable they have. The software (Visirule) will generate the appropriate statistical tool that is best fit the kind of data or variable the researcher has. This software is chosen because it helps in making good decision without necessarily memorize any commands.

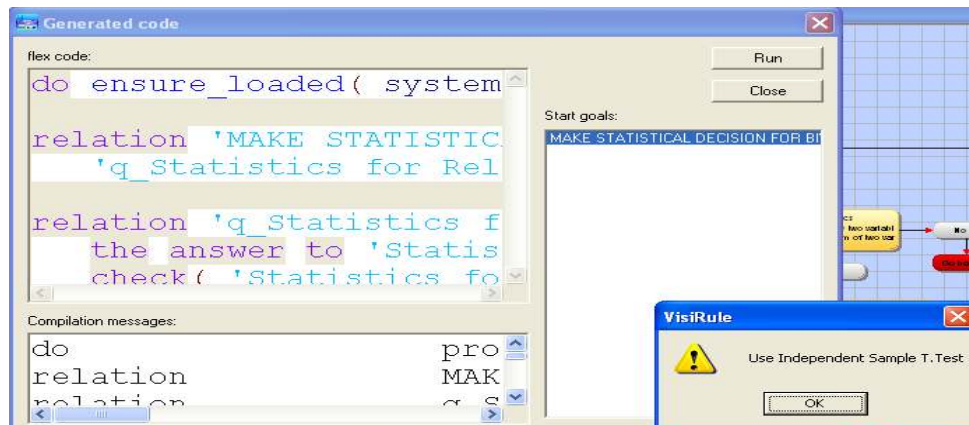


Visirule Application Screenshots: **Figure 3**

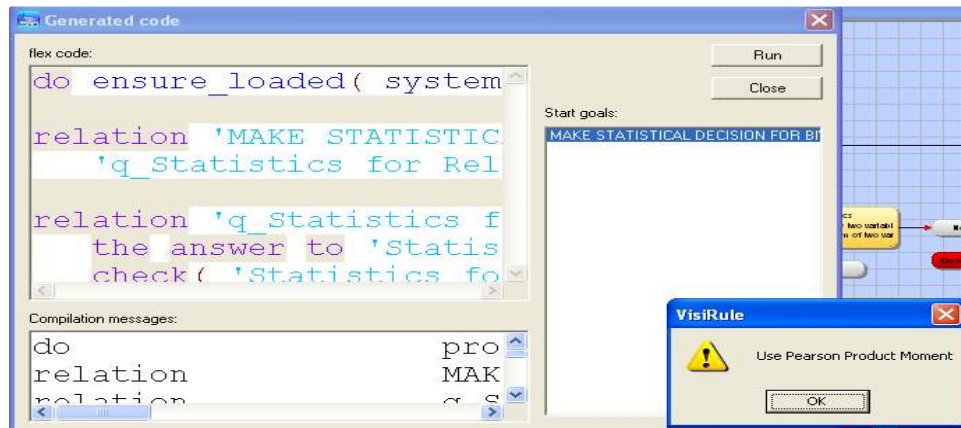
3. Results and Discussion

The bivariate and multivariate analysis discussed in this work can be of great use to researchers and to both undergraduate and post-graduate students in making a productive and realizable decision in utilizing suitable statistical data analysis in researches.

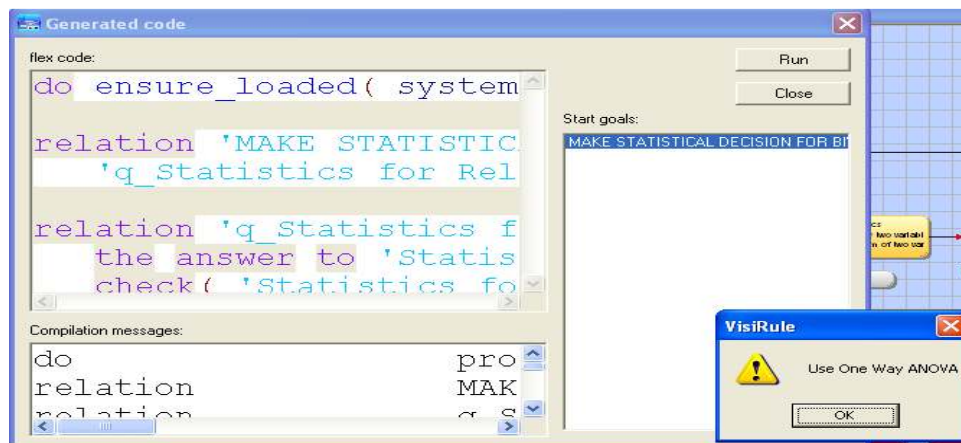
This study will be of an advantage to both lecturers and research students who at the end of their course would write project on their fields of study. Also, students do not need any special knowledge of programming before selecting appropriate statistical tool for analysis. It can also serve as a supportive material for teacher teaching research related courses.



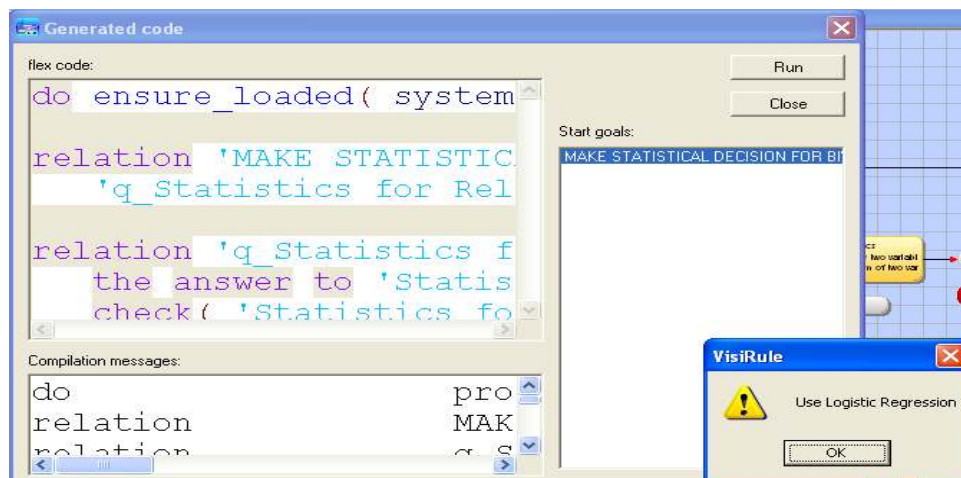
Visirule Output Screenshots: **Figure 4**



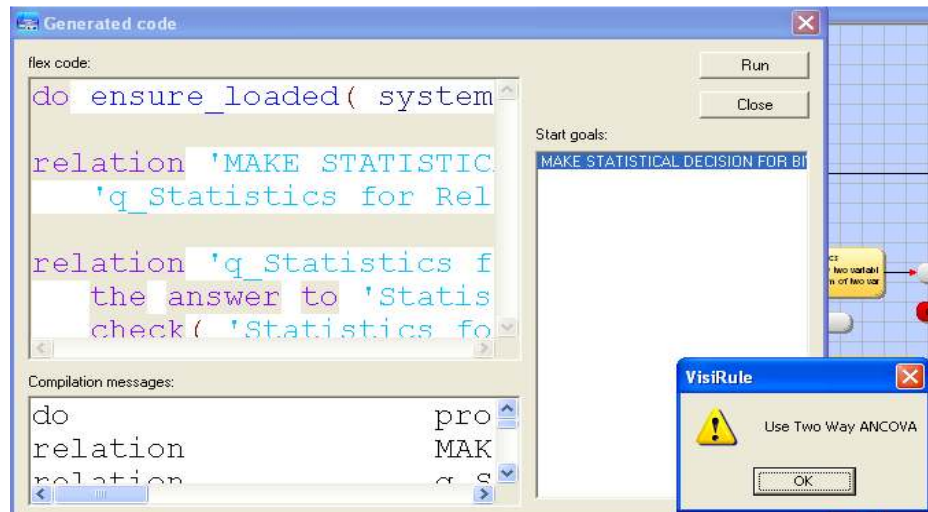
VisiRule Output Screenshots (cont.): **Figure 5**



VisiRule Output Screenshots (cont.): **Figure 6**



VisiRule Output Screenshots (cont.): **Figure 7**



Visirule Output Screenshots (cont.): **Figure 8**

4. Conclusion and Beneficiaries

This study is a step further of earlier research that was primarily on univariate and bivariate statistics [17], here this particular study emphasizes on how problem of selecting an appropriate statistical tool for data analysis would be resolved. At the same time, authors gave details of conditions required of using one statistical tool. Explanations were given to parametric and non-parametric, the statistics that are under parametric and non-parametric and finally the statistics were grouped into bivariate and multivariate.

Graphical illustrations were also given which clearly depicted the relationships between the statistics discussed. Subsequently, Visirule software was used to implement the easy selection of those discussed tools

Hence, below are the beneficiaries:

1. This study will be of an advantage to both lecturers and research students who at the end of their course would write project on their fields of study.
2. Also, students with little or no knowledge of programming and research will find the study more useful.
3. Teacher at various part of institutions would find it more helpful to teach students statistics.

References

- [1] Redman, L. V. and Mory A. V. H. (1923). Research Methodology: An Introduction. Retrieved on the 12-10-2014 from www.google/research/methodology
- [2] Muijs, D. (2006). Doing qualitative research in Education with spss. New Delhi, SAGE Publications Ltd.
- [3] Akinade, E.A and Owolabi, T. (2009). Research Methods. A pragmatic approach for social sciences, Behavioral Science Education. Lagos, Connel Publications.
- [4] Ross R.H (2014). Motivating learners using graphical Visirule based system
i. International Journal of Information Studies Volume (1): issue (4).
- [5] Ejifugba, A.U (1998). Fundamentals of Research in Health Education Benin, Barloz publishers Inc.
- [6] Woody, C. (2014). Meaning of Research. Retrieved on the 12-10-2014 from www.google/research

- [7] Healey, J. F. (2014). The essential of Statistics: A tool for Social Research. Retrieved on the 12-10-2014 from www.askme.com
- [8] Giarratano, J.C (2007). Principles and programming Fourth Indian Edition Thomson.
- [9] Australian Bureau of Statistics (2015). What is a variable? Retrieved on the 4 December, 2015 from <http://www.abs.gov.au/websitedbs/a3121120.nsf/home/statistical+language+-+what+are+variables>
- [10] Heidi Ziemer (2015). Online Statistics Education: An Interactive Multimedia Course of Study. Developed by Rice University (Lead Developer), University of Houston Clear Lake, and Tufts University
- [11] Jim Frost (2015). Choosing between a Non-Parametric Test and a parametric Test. Retrieved on the 5 December, 2015 from <http://blog.minitab.com/blog/adventures-in-statistics/choosing-between-a-nonparametric-test-and-a-parametric-test>
- [12] Vassarstats_Online_Textbook (2015). Parametric and Non-parametric. Retrieved on the 5, December 2015 from <http://vassarstats.net/textbook/parametric.html>
- [13] Stat Trek (2015). What is Hypothesis Testing? Retrieved on the 6 December, 2015 from <http://stattrek.com/hypothesis-test/hypothesis-testing.aspx>
- [14] Donna R (2012). Examine the differences between univariate and bivariate data. Retrieved on the 5th May, 2012 from <http://regentsprep.org>
- [15] Alwan, R.H and Benrachi, B. (2009). Motivating learners using graphical Visirule based system International Journal of Information Studies Volume (1): issue (4).
- [16] Bilgi, N.B, kulkarni, R.V and Spenser .C (2010). An expert system using of Decision Logic charting Approach For Indian legal Domain with Specific reference to transfer of property Act. International Journal of Artificial Intelligence and Expert system (IJAE), Volume (1): issue (2).
- [17] Muraina I.O; Rahman M.A; Adeleke I.A and Aiyegbusi E.A. (2013). Use of a Rule Tool in Data Analysis Decision Making. *International Institute of Science, Technology and Education (IISTE) (Information and Knowledge Management)*. Vol.3, No.3, ISSN 2224-5758 (Paper) ISSN 2224-896X (Online), Pp 131-142.