SELECTION OF STATISTICAL TEST IN EDUCATIONAL RESEARCH AND ASSESSMENT

Pawan Mijar

Mr. Mijar is a Statistic Officer working at the Education Review Office and data analyst expertise on applied statistics. Correspondence regarding this article can be addressed to him on his address. Email: pawan.ero.statistics@gmail.com

Abstract

Statistical analysis has vital and significant roles in educational research and assessment. In today's era of science and technology research should be rational with logical reasoning so, with flow of time educational assessment has gradually shifted from traditional assessment towards alternative assessment where more attention has been paid to the core research methodology, finding and its presentation. This article will try to update the reader with the basic research tools that are utilized while conducting various educational research and assessment in Nepal. For choosing right statistical tests on the basis of study design (univariate, bivariate, multivariate), level of measurement, and distribution of the data in the population. First researcher should deal with data types and its analysis. As per objectives descriptive or inferential or both analyses can be done. Likewise, it will be helpful in selection of parametric and nonparametric statistical tests. The relationships between two variables (bivariate) by means of parametric tests (*t*-test, ANOVA, Pearson´s correlation coefficient, simple linear regression etc.) will be more suitable. For interval or continuous, normally distributed data the non-parametric tests will be considered. Then, multivariate analyses (e.g. multi-way ANOVA and multiple linear regression) for determining the independent contribution of different factors to a single outcome, with tests chosen on the basis of the nature of the outcome variables and on the hypothesized relationship between variables.

Keywords: statistical reasoning, descriptive statistics, inferential statistics, parametric test, non-parametric test

Introduction

Over the past decade, the use of data and evidence based educational practice has increased dramatically and become the standard for educational decision making and policy formulation. Today, more than ever, educators are required to gather and analyze various forms of data for purpose of educational assessment and data driven decision making. Educational assessment is a tool and a way of managing the educational practice, besides serving as a response and information about correct or incorrect learning methods. Assessment is an important part in the teaching and

learning process (Jamil, 2012) which can provide a clearer picture on what the students have learnt and problems they encountered and help to maintain educational quality (Akkya, and Durmus, 2005). There are two types of educational assessments, i.e. formative assessment and summative assessment. Formative assessment is a planned process that regularly determines students' understanding in the instructional activities. Meanwhile, summative assessment is a cumulative assessment that may generate an ultimate grade at the end of the course.

The wrong selection of the statistical method carries serious problems in the implementation of the finding so, researcher should have good statistical knowledge to select the appropriate statistical method. Mishra (2009) stated that various statistical methods are available for a specific situation and condition to analyze the data. The assumptions and conditions of different statistical methods are different. So, in a selection of statistical methods for data analysis, good knowledge of the assumptions and conditions is essential and the proper statistical method can be selected in data analysis.

Statistics has become an integral part of our daily lives. It is widely considered as a mathematical science pertaining to the collection, analysis, interpretation or explanation and presentation of data. Statistics primarily concerned with decision making and policy formulation so it is widely used in educational research. Statistics are collected in systematic manner to achieve predetermined purpose or objectives so in educational research and assessment statistics should align with research from questionnaire development to analysis to policy formulation and implementation (Singh et al., 2020).

A researcher should have good knowledge to select the appropriate statistical method because the result of the wrong selection of the statistical method carries serious problems in the implementation of the finding. Various statistical methods are available for a specific situation and condition to analyze the data. The assumptions and conditions of different statistical methods are different. So, in a selection of statistical methods for data analysis, good knowledge of the assumptions and conditions is essential and the proper statistical method can be selected in data analysis. Likewise, the type and nature of the data and objective of the research also play a very important role in the data analysis procedure. Hence, a particular statistical method is used for a particular objective. Nowadays, various statistical software such as SPSS, SAS, Stata, R, etc. is available in data processing and analysis. Two main statistical methods are used in data analysis called descriptive statistics and inferential statistics. A descriptive statistic is a summary statistic that quantitatively describes or summarizes features from a collection of information with main indexes mean and variance.

The purpose of this article is to review basic statistical concepts, its procedure and the use of selected common Statistical tests in educational research. Particularly important is the ability to examine research for the appropriate statistical test use and interpretation. Knowledge of statistical concepts and common statistical tests assist in the appraisal of educational research for evidence-based practice. In this paper I explain how to select the correct statistical test depending on the type of data and purpose of the analysis. When choosing the appropriate statistical test for educational research, the first step is to decide what scale of measurement of your data and how will this will affect your decision. The next stage is to consider the various statistical test and to analyze and to interpret them.

Research Questions

1) What is the nature of data?
2) What scales of measurements have been used?
3) Which test is used to carry inferential statistics or descriptive statistics?
4) Are the data suitable for parametric or non-parametric test?
5) What kind of test is used for data analysis?

Objectives of the Study

The main objective of the study is to select right statistical test for right decision making process in educational research and assessment. The specific objectives of this study are as follows.

- To identify the nature of data
- To evaluate the scales of measurement
- To explore the patterns and relationships in the data
- To be aware of inferential statistics and descriptive statistics
- To select parametric or non-parametric test

Procedure to Select Right Statistical Test

Choosing the right statistical test involves a systematic process that ensures the test selected is appropriate for the data and research objectives. Here are the steps to guide a researcher through this process:

Step 1: Define the Research Question and Hypotheses

Identify the main research question(s) and formulate the null and alternative hypotheses.

Step 2: Identify the Types of Variables

Determine the dependent variable & independent variable(s) with model specification then classify variables on the basis of their nature. They can be nominal, Ordinal, interval or ratio.

Step 3: Determine the Number of Groups or Conditions
Total number of group of data is vital in data analysis so identify the groups whether they are categorized into Single group, two groups or multiple groups

Step 4: Assess the Study Design
Group may be independent or paired, independent groups are separate and unrelated. Paired or matched groups are related or the same participants are measured more than once.

Step 5: Check Assumptions of the Data
Assess normality if the data follows a normal distribution using tests like the Shapiro-Wilk test. To consider normality sample size should have large enough and variance of each group should be equal.

Step 6: Select the Appropriate Statistical Test
Based on the above factors, choose a test either it is parametric or non-parametric test. To analyze relationship researcher can test correlation and regression. To test association between variables it is better to go for Chi – square test.

Step 7: Conduct the Test and Interpret the Results
After conducting a statistics test, researcher need to interpret the result and draw conclusion based on the finding of study. Proper data analysis and interpretation may helful for policy maker and other academic & non-academic user.

Step 8: Validate the Results
Validation acts as an approval stamp attesting the quality of data and its finding so after finding output researcher need to validate result and should report accordingly.

Data Types and its Graphical visualization

According to Stevens (1946), there are four types of measurements (nominal, ordinal, interval, and ratio) and the types of measurements are determined by their basic empirical operations. Nominal measurement consists of category labels (e.g., numbers or symbols) that can be assigned to observations (or individuals) so that those with different labels are not equivalent. With an ordinal measurement, category labels are assigned to observations to rank and order them with respect to one another. The ordinal scale arises from the operation of rank-ordering. Categorical data with a meaningful order but no consistent are ordinal data. In educational research rankings (first, second, third) or satisfaction ratings (satisfied, neutral and dissatisfied) can be done in ordinal data (Stevens, 1946). Using an interval measurement, numbers are assigned to observations. The numbers have the property of order, and equal differences between any two adjacent numbers reflect equal magnitude. All the properties of an interval measurement apply to a ratio measurement, and in addition,

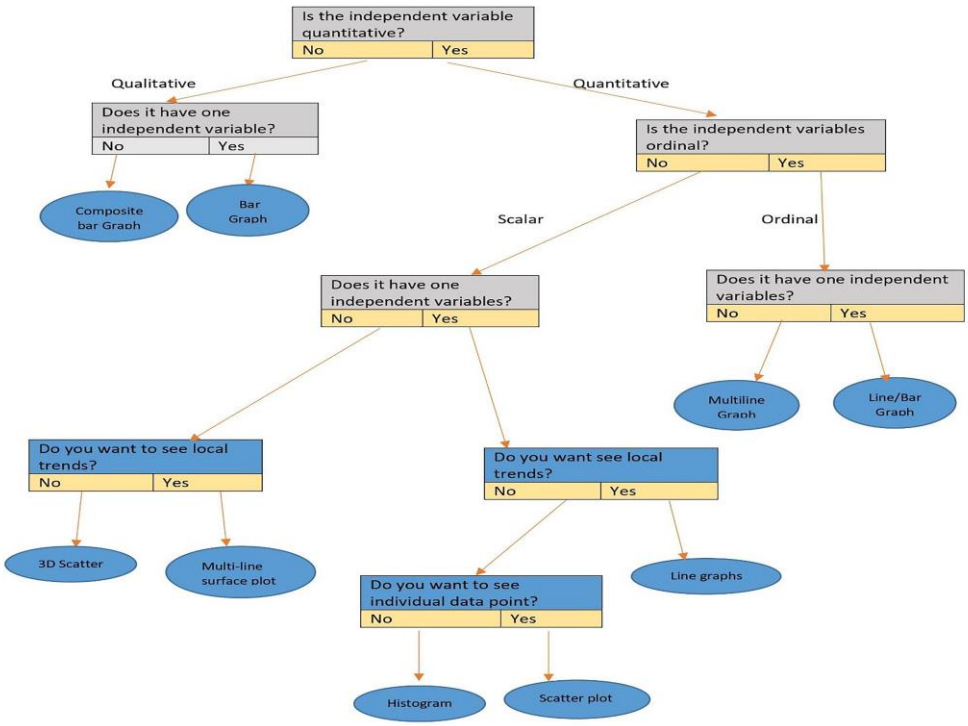
there is a true zero point for a ratio measurement to reflect the absence of the measured characteristic. Ratio data is a form of quantitative (numeric) data. It measures variables on a continuous scale, with an equal distance between adjacent values (Swierstra, 2008).

The statistical graphs are used to represent a set of data to make it easier to understand and interpret statistical information. Recognizing and interpreting variability in data lies at the heart of statistical reasoning. Since graphical displays should facilitate communication about data, statistical literacy should include an understanding of how variability in data can be gleaned from a graph. Before constructing the graph, the graph constructor should have a clear purpose in mind, along with an adequate understanding of variables and graph types (Berg and Smith, 1994). For a graph to be an effective communication piece for both the creator and the observer, four main components should be considered: 1) data form, 2) graph choice, 3) graph mechanics, and 4) aesthetics and visuospatial aspects. While these are four distinct components, they are all interrelated and influence the type and quality of the message communicated by the graph.

The graphical chart you pick depends on the type of data you have and the objective of plotting the data – what is the question you are trying to answer? There are separate charts for qualitative & quantitative data and separate charts for discrete & continuous data. The following flow charts can help researcher to pick the right graphical chart.

Flow Chart 1

Selection of data base graphical chart



Descriptive and Inferential Statistics

Fisher (2009) states that descriptive statistics involves describing and summarizing the data from selected sample. It focused on visible characteristics of dataset. Generally categorical and numerical variables are suitable for descriptive analysis. Measure of central tendency, mean, median, mode, range, standard deviation, variation etc. By 'describe' we generally mean either the use of some pictorial or graphical representation of the data (e.g. a histogram, box plot, radar plot, stem-and-leaf display, icon plot or line graph) or the computation of an index or number designed to summarize a specific characteristic of a variable or measurement (e.g., frequency counts, measures of central tendency, variability, standard scores). Along the way, we explore the fundamental concepts of probability and the normal distribution.

Inferential statistics is the analysis of random sample of data taken from population to describe and make inference about population. It compares, test and predicts the data and provides conclusion about population. Estimation, hypothesis testing and regression are common statistical test in inferential statistics. Inferential statistics measures the significance i.e. whether any difference e.g. between two samples is due to chance or a real effect, of a test result. This is represented using p values. The type of test applied to a data set relies on the sort of data analysed, i.e. binary, nominal, ordinal, interval or ratio data; the distribution of the data set (normal or not); and whether a potential difference between samples or a link between variables can be studied.

Selecting the appropriate graph based on the data type enhances the clarity and interpretability of the data, aiding in better communication of research findings. In statistics, different types of data are best represented by specific types of graphs. Following table shows detailed breakdown of data types and the appropriate graphs for each:

P-value and Effect Size

A p-value or significance level indicates the probability that a result is obtained by chance. In education research, the most common significance levels are 0.05 or 0.01, which indicate a 5% or 1% chance, respectively of rejecting the null hypothesis when it is true. A smaller p-value of .01 as compared to a p-value of .05 will decrease the chances of rejecting the null hypothesis when it is true. When a p-value is less than or equal to the significance level designated by the researcher should have rejected the null hypothesis and reported a difference in the groups or a relationship among the variables (Gravetter and Wallnau, 2012). While a significant p-value indicates statistical significance, effect size denotes the relative magnitude of the differences or the relationship (Wallnau, 2012). There are many different measures of effect size, which correspond to the statistical test utilized

(Cumming,2012). Effect size calculators are available online and the reader may calculate effect sizes if the researcher did not calculate the value.

Parametric and Non-Parametric Test

Parametric and nonparametric tests are broad classifications of statistical testing procedures. They are perhaps more easily grasped by illustration than by definition. In Statistics, a parametric test is based on assumptions related to population or data sources. According to (Mohanty and Misra, 2016) parametric statistical tests is a test whose model specifies certain conditions about the parameters of the population from which the research sample was drawn. It is a kind of hypothesis test which gives generalizations for generating records regarding the mean of the primary/original population. Parametric statistics consist of parameter like mean, standard deviation, variance etc. Parametric test make assumption about population parameters. Z – Test and t-test are often carried out in this test analysis. The t-statistic test holds on the underlying hypothesis, which includes the normal distribution of a variable. In this case, the mean is known, or it is considered to be known. For finding the sample from the population, population variance is identified. It is hypothesized that the variables of concern in the population are estimated on an interval scale.

In Non-Parametric tests are usually referred to as distribution free or assumption free test (Sheskin, 2011). Parametric test can be used on nominal and ordinal data (Wash, 1992). Non parametric test also applied to interval and ratio data which do not follow normal distribution. Non parametric statistical analysis differs from statistical analysis in that it only uses + or – sign or the rank of the data sizes instead of original values (Nahm, 2016). When the sample size is small and researcher is unsure about normality of the data is used. Similarly, if the data is better represented by median rather than mean then non-parametric test will be perfect. With reference to non-parametric statistics, statistical techniques like Spearman‟s rank order correlation, chi- square, Mann Whitney U test can be termed as non-parametric statistical techniques.

A chi-square test ($\chi 2$) is a statistical test that examines the relationship in variables measured at the categorical level. The $\chi 2$ test compares the frequency of data observed with the expected frequencies of the data expected if there is no relationship between the variables resulting in a Pearson's Chi-Square (Gravette and Wallnau, 2012). A chi-square test indicated no significant relationship between gender and dropping out of cardiovascular rehabilitation χ 2 (1, n = 438 = .37, p = .56, phi = -.03). In a significant finding, the phi coefficient can indicate effect size; however, in the hypothetical example the findings were not significant.

Table 1

Selection of parametric and non-parametric test

S.N	Category	Parametric statistical test	Non-Parametric statistical test	Remarks
1	Correlation	Pearson Correlation	Spearman Rank coefficient (Rho), Kendall' Tau	
2	Two groups Independences measure	Independent t-test	Mann-Whitney test	
3	More than Two groups Independences measure	One way ANOVA	Kruskal - Wallis one way ANOVA	
4	Two groups repeated measure	Paired t-test	Wilcoxon matched pair signed rank test	
5	More than, Two groups repeated measure	One way ANOVA	Friedman's Two way Analysis of Variance	

Reliability, Stability and Validity

Reliability is essentially concerned with 'error in measurement (Bannigan & Watson, 2009) i.e. how consistently or dependably does a measurement scale measure what it is supposed to be measuring (Jeffers, 2002). The reliability of a scale indicates how free it is from random error. Psychologists considered three types of consistency: over time (test-retest reliability), across items (internal consistency), and across different researchers (inter-rater reliability). The most commonly used statistic is Cronbach's coefficient alpha (available using IBM SPSS, see Chapter 9). This statistic provides an indication of the average correlation among all of the items that make up the scale. Values range from 0 to 1, with higher values indicating greater reliability. While different levels of reliability are required, depending on the nature and purpose of the scale, Nunnally (1978) recommends a minimum level of .7. Cronbach alpha values are dependent on the number of items in the scale. When there are a small number of items in the scale (fewer than 10), Cronbach alpha values can be quite small. In this situation it may be better to calculate and report the mean inter-item correlation for the items. Optimal mean inter-item correlation values range from .2 to .4 (as recommended by Briggs and Cheek 1986). Reliability can be assessed in different ways; test-retest reliability for stability, inter-item reliability for internal consistency and interrater reliability or parallel scale for equivalence.

Stability

Stability refers to the ability of a system to remain operational and responsive even in the face of unexpected events or changes. A measurement scale's stability is the extent to which the same results are obtained on repeated administrations of the instrument. The estimation of reliability here focuses on the instrument's susceptibility to extraneous factors from one administration to the next' (Polit and Hungler, 1995). This is assessed through 'test-retest reliability', a commonly used indicator of the reliability of a measurement scale (Watson, 1995).

Validity

Once a measurement scale has been shown to be reliable over time it should be assessed to establish whether or not it is reliably measuring what you want it to measure (Watson, 1995). Validity is concerned with the meaning and interpretation of a scale. There are many ways of testing validity and it has been suggested that 'A variety of approaches should be used in testing any index, rather than relying on a single validation procedure' (McDowell and Newell 1996, p. 37). This is because validity is not absolute. It is a matter of degree rather than an 'all or nothing' concept' (Carmines and Zellar 1979). 'In reality…it is not possible to take one form of measurement validity in isolation, as several forms may be applicable' (Gould, 1994).

Impact of Wrong Selection of the Statistical Methods

A wrong selection of the statistical method not only creates some serious problem during the interpretation of the findings but also affect the conclusion of the study. The selection of wrong statistical method and test gives wrong result and may fail to fulfill objectives of research assessment. There are specific statistical methods for every situation. Failing to select an appropriate statistical method, our significance level as well as their conclusion is affected. Due to incorrect practice, we detected a statistically significant difference between the groups although actually difference did not exist.

Conclusion

In education, statistics are used for educational planning, policymaking, quality assurance, and evaluating different aspects of the education system. Statistics help provide the quantitative foundation needed for projecting future development and play a key role in strengthening the educational planning and evaluation process. Some specific ways statistics are used in education include assessing economic factors related to education, measuring results in natural and social sciences experimentation, and determining the reliability and validity of educational tests. It is important for drawing conclusions and inferences from facts. In education, statistics helps with constructing and standardizing tests, understanding individual student differences, comparing evaluation methods, and making predictions about student progress.

To bring out right conclusion and to meet objectives of research the selection of right statistical test, tools and technique has significant role. The selection and undertaking of the appropriate statistical test with graphical presentation has pivotal role in modern educational assessment. In this articles, different types of statistical tests were explained for the purpose of educational research. From different review of literature researchers was concluded that skill of selecting appropriate statistical test is very essential for making good and specific conclusion in educational research & assessment.

References

Angra, A., & Gardner, S. M. (2017). Reflecting on graphs: Attributes of graph choice and construction practices in biology. *CBE—Life Sciences Education*, *16* (3).

Akkaya, R. (2016). Research on the development of middle school mathematics pre-service teachers' perceptions regarding the use of technology in teaching mathematics. *Eurasia Journal of Mathematics, Science and Technology Education*, *12*(4), 861-879.

Bannigan, K., and Watson, R. (2009). Reliability and validity in a nutshell. *Journal of clinical nursing*, *18*(23), 3237-3243.

Blanchard, J., and Carey, J. (1987). Scales of measurement and appropriate statistical tests. Literacy Research and Instruction, 26(4), 302-308.

Carmines, E. G., and Zeller, R. A. (1979). *Reliability and validity assessment*. Sage publications.

Hippel, Paul T. von (2005). "Mean, Median, and Skew: Correcting a Textbook Rule". Journal of Statistics Education. 13 (2).

Jamil, M., and Muhammad, Y. (2019). Teaching Science Students to Think Critically: Understanding Secondary School Teachers' Practices. *Journal of Research and Reflections in Education (JRRE)*, *13*(2).

Jeffers, B. R. (2002). Continuing education in research ethics for the clinical nurse. *The Journal of Continuing Education in Nursing*, *33*(6), 265-269.

Main, M. E., and Ogaz, V. L. (2016). Common statistical tests and interpretation in nursing research. *International Journal of Faith Community Nursing*, *2*(3), 5.

Nilsson, J., Parker, M. G., and Kabir, Z. N. (2004). Assessing health-related quality of life among older people in rural Bangladesh. *Journal of Transcultural Nursing*, *15*(4), 298-307.

Nunnally, J. C. (1978). An overview of psychological measurement. *Clinical diagnosis of mental disorders: A handbook*, 97-146.

Piedmont, R. L., and Hyland, M. E. (1993). Inter-item correlation frequency distribution analysis: A method for evaluating scale dimensionality. *Educational and psychological measurement*, *53*(2), 369-378.

Singh, S., Dhir, S., Das, V. M., & Sharma, A. (2020). Bibliometric overview of the Technological Forecasting and Social Change journal: Analysis from 1970 to 2018. *Technological Forecasting and Social Change*, *154*, 119963.

Swierstra, W. (2008). Data types à la carte. *Journal of functional programming*, *18*(4), 423-436.