Propellers Journal of Education
Vol. 2 No. 1, 1-8 (June 2023)
ISSN: 2971-639X (Online)
*ijvocter.com*

# Choosing the Appropriate Statistical Methods for Data Analysis in a Research: A Solution to the Puzzle Faced by Beginners

**Gideon, S.N. & Nwogu, V.U.**
Department of Statistics, Abia State Polytechnic, Aba
Correspondence: dearngos2012@gmail.com



**Received:** May 06, 2023|**Accepted:** June 25, 2023|**Published:** June 30, 2023

**Abstract:** *Choosing the appropriate statistical method to be employed in the analysis of data collected for a given research has remained a puzzle to many researchers most especially beginners. This paper provided, in a simple and clear form, a guide in the selection of statistical methods for analysis. The paper highlighted the factors to be considered in the selection of a statistical method for research data analysis and also gave the statistical methods that are appropriate for some research hypothesized problem situations. Schedules that can aid beginners in the selection of statistical methods to be employed in research for each analysis situation were also provided.*
**Keywords:** Research objectives, research questions, data collection, and selection statistical methods

## 1.0 Introduction

Data analysis is fundamental in most research work and the reliability of the findings of research depends to a large extent on how well the data analysis was carried out. Sometimes the appropriate data are collected and the wrong method of analysis is employed leading to spurious results and deceitful conclusions. Khusainova *et al,* (2016) and Mishra *et al,* (2019) emphasized that incorrect choice of method of analysis for experimental data leads to an erroneous conclusion, and incorrect interpretation of research results and thereby distorts or even leads to the loss of scientific value of such research. Therefore choosing the right statistical method, in research, is a decision-making situation and should be done carefully.

Choosing the right statistical method for data analysis may sometimes be a confusing and challenging task for many researchers, most especially beginners. The reason for this confusion is given by Pallant (2001)

*"Although most statistics courses teach you how to calculate a correlation coefficient or perform a t-test, they typically do not spend much time helping students learn how to choose which approach is appropriate to address a particular research question".*

A researcher is faced with several statistical methods to choose from, depending on the problem at hand, when carrying out research. Some of these statistical methods look similar in practice but address different problems. Therefore researchers should carefully choose the method to be use in order to adequately address the problems their various types of research seek to address.

Efforts have been made towards guiding researchers in choosing the right statistical method to be adopted in their studies. Marusteri and Bacarea (2010) gave a step-by-step guide on how to

select a suitable test for the comparison of two or more groups for statistical differences. Their work is limited to only the test of hypotheses which is just an aspect of the areas of application of statistical methods. Khusainova *et al,* (2016) developed an algorithm that can aid researchers in the selection of an appropriate statistical method for the analysis of research data. Mishra *et al,* (2019) approached the selection of appropriate statistical methods from the parametric and non-parametric statistical points of view only. They discussed some statistical methods and the assumptions of these methods. It may not be easy to write an article that will cover all problem situations in research and the appropriate statistical methods for the analysis of data accrue from research. Khusainova *et al*, (2016) were more elaborate in their work than the other cited authors. However, these works did not discuss the selection of a method of analysis based on the possible forms of research objectives.

This paper aims to throw more light on the factors to be considered in selecting a statistical method. The paper explains how to choose statistical methods that are appropriate for some research problem situations. Also, schedules that can aid beginners in the selection of statistical methods to be employed based on specific research problem situations were provided.

The rest of the paper is planned as follows; Section 2.0 covers some research questions and the appropriate statistical methods to address them. Sub-section 2.1 discussed research situations and statistical methods of data analysis under the exploration of relationships and/or degree of relationships among variables, Sub-section 2.2 dealt with research situations and methods of analysis for testing the significance of group differences and Sub-section 2.3 covers methods of data analysis when prediction of group membership is of interest. The paper was concluded in Section 3.0.

## 2.0 Some Research Questions and Appropriate Statistical methods

The choice of appropriate statistical method to be used in analysing a set of data collected in the course of a given research depends on:

i. The objective of the research.
ii. The nature of the variables in the research.
iii. The nature of the data collected.

The concepts above cannot be explained in isolation from one another since they are linked together, that is, in the course of explaining one the other is made mention of. The first thing to do when one wants to embark on research is to identify and clearly state the problem of the research. The statement of a problem is further translated in specific action points, called objectives of the study, which are thereafter framed in question forms referred to as research questions. Achieving research objectives simply means adequately addressing the research questions. When the objectives are clearly stated the variables of the study are identified from it.

A variable is a quantity that changes from one hypothetical unit of a population to another. A variable can be random when its values are determined by chance or non-random when its values are fixed or determined by the investigator. A variable can also be dependent when its values are influenced by the values of another variable or independent when its values are not influenced by the values of another variable. Another form a variable can take is the discrete and continuous forms. A discrete variable assumes only integer values (whole numbers) which always result from counts while a continuous variable assumes all conceivable values between two integers and always results from measurements.

Once the research variables are identified, data on the variables are collected. Depending on the type of variable of interest, the data can be categorical or continuous (counts or measures). Another useful consideration in data analysis that is directly linked to the nature of the data collected is the assumptions required by the statistical method to be adopted. Every statistical method has underlying assumptions which may be the normality of observations, independence of observations, serial correlation of observations, etc. Most statistical methods assume the normality of the data for analysis and when this assumption is met, in the case of hypothesis testing, the parametric tests are adopted otherwise the non-parametric tests are used. Parametric tests are statistical tests in which an assumption is made on a given population parameter or distribution while in non-parametric tests no assumption is made on any parameter or distribution (Marusteri and Bacarea, 2010).

Statistical methods are employed in most research to answer research questions. A typical research question may fall under any of the following common headings.

### 2.1 Relationship and/or Degree of Relationship among Variables

When the interest of a researcher is on investigating the relationship among variables, questions like (1a*) "Is there a relationship between variable A and variable B?"* (2a) *"What is the degree of relationship between variable A and variable B?"* etc. arise. Such questions that try to explore the relation among variables and the predictive power of variables are answered using regression analysis, correlation analysis, discriminant analysis, and factor analysis.

To answer question (1a) *regression analysis* is appropriate. *Regression analysis* studies the relationship, which could be linear or otherwise, among variables. The most commonly used is the linear regression analysis which could be performed for univariate and multivariate cases. Univariate linear regression analysis studies the relationship between a dependent random variable and one or more independent fixed variables. When one independent variable is involved, we have *simple linear regression* and when there is more than one independent variable, we have a *multiple linear regression*. *Multivariate linear regression analysis* involves more than one dependent random variable against one or more independent fixed variables. We also have *hierarchical multiple regression analysis* which is used when we have numerous independent variables and wished to statistically determine which independent variable has the most predictive power (Ross and Wilson, 2017).

Question (2a) can be answered using *correlation analysis*. *Correlation analysis* studies the degree of relationship among independent random variables. For a bivariate case, *Simple correlation analysis* is performed for two independent random variables to ascertain the degree of linear relationship between the variables. Therefore *correlation analysis* is not suitable for cause and effect study which *regression analysis* handles since the variables involved in correlation analysis are random variables. Another form of *correlation analysis* is the *partial correlation analysis* which involves more than two independent random variables and for a case of three independent random variables, studies the degree of relationship between two random variables while keeping the effect of the other random variables constant. In multivariate cases, *canonical correlation analysis* is used to determine the relationship between two sets of variables, and *Factor analysis* allows a researcher to summarize a large set of continuous variables into a smaller number of factors that effectively captures the underlying structure of correlation in a set of variables (Tabachnick & Fidel, 1996). *Discriminant analysis* is also employed to determine the predictive ability of a set of independent continuous variables on a dependent categorical variable

A schedule that can assist a researcher in the selection of the appropriate statistical method for data analysis when the objective or the research question bothers on exploring the relationship of variables is given in Table 1.

**Table 1. Schedule for Selection of Statistical Methods for Exploration of the Relationship among Variables**

| Type of Research Question | Number and Type of Dependent Variable | Number and Type of Independent Variable | Parametric Method | Non-Parametric Method | Remark |
|---|---|---|---|---|---|
| What is the degree of relationship between two variables? | One categorical variable | One categorical variable | None | Chi-Square | The categories of the variable are not considered as scores and may not be equal |
| | None | Two Continuous Variables | Pearson Product Moment Correlation | Spearman Rank Correlation | Non parametric method applicable when normality is not met |
| Is there any relationship between two variables? | One continuous variable | One continuous variable | Simple Regression Analysis | None | The dependent variable must be random and the independent variable fixed. Linear and non - linear relationships can be explored |
| Is there relationship between one variable and two or more other variables? | One continuous variable | Two or more continuous variables | Multiple Regression Analysis | None | The dependent variable must be random and the independent variables fixed. Linear and non -linear relationships can be explored |
| Is there any relationship between two variables after controlling the effect of another variable? | None | Three continuous variables of which one is to be controlled | Partial Correlation Analysis | None | When the variables are categorical, the Chi-Square can be used here to explore the relationships |
| Is there any relationship between two set of variables? | None | Two sets of more than one continuous variables each | Canonical Correlation Analysis | None | When the variables are categorical the Chi-Square can be used here to explore the relationships |
| How many factors adequately represents the underlying correlation structure of some variables? | None | Several continuous variables | Factor Analysis | None | When the variables are categorical the Chi-Square can be used here to explore the relationships |
| Is there any relationship between two set of variables? | Two or more continuous variables | Two or more continuous variable | Multivariate Regression Analysis | None | The dependent variables must be random and the independent variables fixed. Linear relationships is often explored |
| Which variables have the most predictive power for the dependent variable? | One continuous variable | A set of continuous variables | Hierarchical Multiple Regression Analysis | None | The dependent variables must be random and the independent variables fixed. Linear relationships is often explored |
| Which variables best predict group membership of the dependent variable? | One categorical variable | A set of continuous or a set of categorical variables | Logistic Regression Analysis | None | One may find other statistical tools that can do same work as logistic regression |
| Which variables best predict group membership of the dependent variable? | One categorical variable | A set of continuous variable | Discriminant Analysis | None | One may find other statistical tools that can do same work as discriminant analysis |

**Significance of Group Parameters and Group Difference**

Questions like (1b) *"Is the measure obtained from the sample (mean, variance, proportion) significantly different from what the unknown population parameter is?"* (2b) "*Is group A significantly different from group B based on given characteristic?"* (3b) *"Is there significant differences among several groups?"* etc.

Question (1b) can be addressed using one sample *student's t-test* and *z-test*. The question involves only one population from which a sample is selected to test a hypothesis made on an unknown population parameter. *Student's t-test* is used when small samples (samples of less than thirty observations) are drawn from a normally distributed population whose variance is not known while the *z-test* is employed when large samples (samples of thirty observations or more) are drawn from a normally distributed population whose variance is known (Amadi and Gideon, 2020). To answer Question (2b), an *independent samples student's t-test* whose non-parametric equivalent is *Mann Whitney test* or *paired samples student's t-test* whose non-parametric equivalent is *Wilcoxon signed test*, *z-test*, and *one-way ANOVA* can be used. The question involves the determination of the difference between the means of a given characteristic in the two populations on the bases of two independent samples drawn from the two populations under study. Though *One-way ANOVA* can also do the same job, researchers often find it easier to apply *student's t-test* and *z-test* in addressing Question (2b) rather than *one-way ANOVA* which is often used in cases where there are more than two groups to be compared. Question (3b) involves several groups and the statistical method that can be applied to answer the question is *one-way ANOVA*, a parametric test approach whose non-parametric equivalents are the *Kruskal Wallis test* used when measures of a dependent continuous variable are obtained from different units at each level of a categorical variable and *Friedman's test* employed when measures of a dependent continuous variable are obtained from different periods that constitute different levels of a categorical variable (Pallant, 2001). *One-way ANOVA* performs a test of the significance of means of several populations based on one-factor classification of observation but when the observations are classified based on two factors and one wishes to compare the means for the two factors at the same time, the *two-way ANOVA* is employed.

In multivariate analysis group differences are tested using *Analysis of Covariance (ANCOVA), Hoteling's $t^2$-tes*t, and *Multivariate Analysis of Variance (MANOVA)*. *ANCOVA* is a statistical tool that combines regression analysis and analysis of variance in the determination of differences in several groups in the presence of a concomitant variable (Michael *et al.*, 2004; Huitema, 2011). *Hoteling's $t^2$-test* is a multivariate tool for testing difference between two groups in consideration of the means of several variables per group (Aslam and Arif, 2020). MANOVA is a procedure of testing group differences based on two or more dependent continuous variables (Tabachnick and Fidel, 1996).

Table 2 offers a useful guide for the selection of statistical methods for testing the significance of group parameters and parameter differences.

**Table 2. Schedule for Selection of Statistical Methods for Testing the Significance of Group Parameters**

| Type of Research Question | Number and Type of Dependent Variable | Number and Type of Independent Variable | Parametric Method | Non-Parametric Method | Remark |
|---|---|---|---|---|---|
| Is the mean of a particular variable differ from a specified value? | None | One continuous variable | Student's t-test | None | Applicable when sample of size is less than thirty is drawn from a population and the population variance is not known. |
| Is the mean of a particular variable differ from a specified value? | None | One continuous variable | Z-test | None | Applicable when sample of size thirty or more is drawn from a population and the population variance is not known. |
| Does one particular group (males) poses a given characteristics (school dropout) more than another (females)? | One categorical variable (school dropout) | One categorical variable (gender) | None | Chi-square | Both variables gender and number of school dropouts are categorical. School dropouts are classified based on gender. Variables may have unequal categories that may be unequal. |
| Does one particular group (males) poses a given characteristics (heart beat) more than another (females)? | One continuous variable (heart beat) | One categorical variable (gender) | Independent sample students t-test | Mann Whitney U-test | Measures of heart beat are classified based on gender. Applicable for samples less than thirty. The categorical variable must have only two categories |
| Does one particular group (males) poses a given characteristics (heart beat) more than another (females)? | One continuous variable (heart beat) | One categorical variable (gender) | z-test for difference in two population means | Mann Whitney U-test | Measures of heart beat are classified based on gender. Applicable for samples are equal to thirty or more. The categorical variable must have only |
| Is there a change in a given variable (blood pressure) before and after treatment? | One continuous variables (blood pressure) | One categorical variable (periods: before and after treatment) | Paired sample students t-test | Wilcoxn Signed Rank test. | Measure of blood pressure are classified based on gender. Applicable for samples less than thirty. The categorical variable must have only two categories |
| Is there any difference in a variable among the levels of another variable? | One continuous variable | One categorical variable with three or more levels. | One-way ANOVA | Kruskal Wallis test | Measures of the dependent continuous variable are obtain from different units of each levels of the independent categorical variable. |
| Is there any difference in a variable among the levels of another variable? | One continuous variable | One categorical variable with three or more levels. | One-way ANOVA | Freidman test | Measures of the dependent continuous variable are obtain from same units in three or more different times (levels of the independent categorical variable). |
| Is there any difference in a variable among the levels of another two different variables? | One continuous variable | Two categorical variable with three or more levels. | Two-way ANOVA | None | Measures of the dependent continuous variable are obtain from different units of each levels of the two independent categorical variable. |

**Table 2 Contd. Schedule for Selection of Statistical Methods for Testing the Significance of Group Parameters**

| Type of Research Question | Number and Type of Dependent Variable | Number and Type of Independent Variable | Parametric Method | Non-Parametric Method | Remark |
|---|---|---|---|---|---|
| Is there any difference in a variable among the levels of another variable after controlling a concomitant variable? | Two continuous variable | One categorical variable with three or more levels. | ANCOVA | None | One of the continuous variables is a concomitant variable. Measures of the dependent continuous variables are obtain from different units of each levels of the independent categorical variable. |
| Is there any difference between two groups based on the means of several variables? | Several continuous variables | None | Hotelling $t^2$-test | None | The number of continuous variables observed in the two groups must be equal while the sample sizes may differ. Two-way ANOVA can be a substitute for Hotelling $t^2$-test. |
| Is there any difference in two variables among the levels of another variable? | Two continuous variables | One categorical variable with three or more levels. | MANOVA | None | Measures of the dependent continuous variables are obtain from different units of each levels of the independent categorical variable. |

## 2.2 Prediction of Group Membership

In certain research, interest may be in the identification of specific independent variables that best predict the membership of units to certain groups as defined by the dependent variable. The statistical methods that can be employed in such situations are *discriminant analysis* and *logistic regression analysis*. *Discriminant analysis* is employed to determine the predictive ability of a set of independent continuous variables on a dependent categorical variable (Tabachnick and Fidel, 1996) while *Logistic regression analysis* studies the association between a categorical dependent variable and a set of independent variables (Sperandei, 2014).

**Table 3. Schedule for Selection of Statistical Methods for Prediction of Group Membership**

| Type of Research Question | Number and Type of Dependent Variable | Number and Type of Independent Variable | Parametric Method | Non-Parametric Method | Remark |
|---|---|---|---|---|---|
| Which variables best predict group membership of the dependent variable? | One categorical variable | A set of continuous variable | Discriminant Analysis | None | One may find other statistical tools that can do same work as discriminant analysis |
| Which variables best predict group membership of the dependent variable? | One categorical variable | A set of continuous or a set of categorical variables | Logistic Regression Analysis | None | One may find other statistical tools that can do same work as logistic regression |

## 3. Conclusion

This paper has highlighted several research situations such as exploration of relationships or degree of relationships amongst variables, testing the significance of group parameters and group differences, and prediction of group memberships. The study provided schedules under the mentioned research situations aimed towards helping researchers most especially beginners

in solving the puzzle associated with the selection of the appropriate statistical methods for analysis of data collected in a research.

**References**

Amadi, E. P. and Gideon, S. N. (2020).Essential Statistics for Higher Education: A Guide to Estimation and Hypothesis Testing, Eagles Press and Publishers, Aba Nigeria.

Aslam, M. and Arif, O.H (2020) Multivariate Analysis under Indeterminacy: An Application to Chemical Content Data. *Journal of Analytical Methods in Chemistry*, https.//doi.org/10.1155/2020/1406029

Huitema, B.F. (2011) The Analysis of Covariance and Alternative Statistical Methods for Experiments, Quasi-Experiments and Single-Case Studies,2nd Edition, John Willey and Sons, Inc. Publication, Hoboken, New Jersey.

Khusainov, R.M., Shilova, Z.V. and Curteva, O.V. (2016) Selection of Appropriate methods for Research Results Processing, *Mathematics Education*, Vol. 11(1),303-315.

Mishra, P., Pandey, C.M., Singh, U., Gupta, A, Sahu, C. and Keshri, A. (2019) Descriptive Statistics and Normality Tests for Statistical Data. *Ann Card Anaesth*, Vol. 22:67-72

Marusteri, M. and Bacarea, V. (2010) Comparing Groups for Statistical Difference: How to Choose the Right Statistical Test? *Biochemia Medica*, Vol. 20(1):15-32.

Michael, H.K, Christopher, J.N., John, N and William, L. (2004) Applied Linear Statistical Models 5th Edition, McGraw-Hill Companies Inc., New York.

Pallant, J. (2001) SPSS Survival Manual: A Step by Step Guide to Data Analysis using SPSS, Open University Press, Berkshire SL6 2QL, UK.

Ross, A. and Wilson, V.L. (2017) Hierarchical Multiple Regression Analysis Using at Least Two Sets of Variables (In Two Blocks).In: Basic and Advanced Statistical tests. SensePublishers, Rotterdam.

Sperandei, S. (2014) Understanding Logistics Regression Analysis. *Biochemia Medica*, Vol. 24(1):12-18.

Tabachnick, B. and Fidel, L. (1996). Using Multivariate Statistics, 3rd Edition, New York HarperCollins.