

SELECTING AND USING STATISTICAL TOOLS FOR DATA ANALYSIS

Baba Wachiko, and Ahmad, Manko Umar PhD

Department of Primary Education Studies,

Niger State College of Education, Minna

Email: wachikobaba68@gmail.com

Abstract

Preparing this piece of article was borne out to respond to the call from young researchers on how to select appropriate statistical tools and apply it in data analyses. Because the goal of any researcher is to avoid errors which may result into incorrect findings and making wrong conclusion. To avoid such errors is to select appropriate statistical tools and apply it. Statistics is concerned with scientific methods of collecting, organizing, presenting and analyzing data as well as making valid decisions. The selection of appropriate statistical test for data analyzing depends on the nature of data collected and the type and research design. Some conceptual clarifications were discussed to serve as background of the article. These terms includes: data, classification of data, population, sample, variable and its types, and others. Preparing an ungrouped frequency distribution table, the researcher needs three basic steps: List the scores either in ascending or descending, tallying the scores and determining the number of times each scores occurred. The number of strokes at each score is counted and written down as frequency of that particular score, that in column 3 of the table. Preparing a grouped frequency distribution table, whenever the researcher has a large number of scores, it may be laborious to list all the scores. A convenient and faster method is to group the scores before tallying them. The following procedure should be adopted, when preparing a grouped frequency distribution table: group the scores into class intervals. To do this one need to understand the following concepts: class interval, class size, class limit and class boundaries, tally the scores in column 2 using the class intervals and write down the frequency of occurrence of each class interval in column 3. The descriptive and inferential statistics were also discussed and the steps involved in selecting appropriate statistics tests. Therefore, it is against this background, that the writers give details step-by-step of statistical tests that are generally employed in analyzing research data.

Keywords: Statistics Terms, Presentation of Data, Measures of Central Tendency, Parametric Statistics, Non-parametric Statistics.

Background

The goal of any researcher in the process of collecting, organizing, analyzing and presenting data is to avoid errors which may result into incorrect findings and making wrong conclusion. To avoid such errors is to

select appropriate statistical tools and apply it. The nature of data collected and research design gives the researcher basic ingredients on which appropriate statistical measures to be used, so that the researcher derives intelligible decisions. According to

Muhammad (2020) that the basic ingredients for any research work are qualitative and quantitative data. Quantitative data are data which numerical values can be assigned. Examples are examination scores; ages of students, height, attitude etc. Qualitative data are data that has non numerical in nature. These data are in alphabetical form, e.g., marital status, sex, state of origin, etc.

Therefore, statistical tests assist the researcher to establish strong evidence for arriving at any possible conclusion from limited amount of data. For a given data to be adequately analyzed, appropriate methods for data collection must be taken into consideration as submitted by (Medinat, 2018 & Muhammad, 2020). Some of the researchers find it difficult and challenging in selecting appropriate statistical tools for their data analysis. Hence there is need to use these approaches to have a basic knowledge in selecting appropriate statistical tools to enhance young researchers' performance in data collection and analyses in their research work. Therefore, this paper is concerned with the statistical methods of data analysis, which assist in providing valid predictions and conclusions in research study. Criteria for choosing appropriate statistical tool, conceptual clarifications, sources of data, procedure for data collection and analysis are discussed.

Conceptual Clarifications

This section deals with definitions based on the works of Awotunde and Ugodulunwa (2002) and Maruf and Sambo (2013).

Statistics: This is an aspect of mathematics which is concerned with scientific methods for collecting, organizing, summarizing, presenting and analyzing data as well as drawing a valid conclusion and making

reasonable decisions on the basis of such analysis. Statistics is applied to solve educational problems by various stakeholders. Statistics is therefore of immense importance in education. Because of contemporary usage of statistics, it is of two types; descriptive and inferential statistics. Descriptive statistics, the event or outcomes of events are described without drawing conclusions. These includes: mean, mode, median, standard deviation, range, percentile, kurtosis, correlation coefficient and proportions etc. It is used only for collection, organization, summarizing, analysis and presentation of data. Inferential statistics builds further on descriptive According to Maruf and Sambo (2013) and Muhammad (2020), data is classified into two, these are: statistics to make interpretation, generalizing and conclusions. These include: t-test, f-test, z-test, chi-square, etc.

Data: simply means collection of information, qualitative or quantitative. For example, information about score, age, height, distance and number.

Classification of Data

According to Maruf and Sambo (2013) and Muhammad (2020), data is classified into two, these are:

- I. **Qualitative data:** this is data in which numerical values can be assigned to it. Examples; examination scores, ages, weight, height, etc.
- II. **Quantitative data:** this is the form data that numerical values cannot be assigned. Such as marital status, sex, religion. State of origin, etc.

Types of Data

According to Maruf and Sambo (2013) and Muhammad (2020), that data is of into three types, these are:

- I. **Time series data:** data that is generated based on timely interval like every second, minutes, hourly, daily, weekly, annually etc. Example weather report, rainfall, inflation rate, clock stock etc. All these data collected based on these sequential observations is refer to time series data.
- II. **Cross section data:** data collected at the same time; example of Population census conducted by National Population Commission every 10 years.
- III. **Pooled data:** data that are collected which combined elements or features of time series and cross section data.

Data are generated through either primary or secondary source. Primary source are data collected directly by the researcher while the secondary source are data generated through second person, then to the researcher. These are examples of references, documents, recorded items and bibliographies etc.

Population: this refers to the group of measurements that are of interest in research i.e the aggregate of units to the covered. This may be people, objects, materials, measurement or things. Population could be finite or infinite.

Sample: this is part or subset of a population. It is any subgroup or aggregate drawn by some appropriate method from a population. The sample is usually the portion of the population appropriately selected for observation.

Parameter: this is a descriptive measurement or characteristics, true value of population. When such features as mean, standard deviation or variable are computed is known as parameter.

Distribution: this is the arrangement of a set of numbers classified according to some property.

Skewness of a Distribution: this is a distribution having a longer tail at one end than at the order. It is an asymmetrical distribution.

Kurtosis: this is the extent of peakness in a distribution

Normal Distribution: this is a symmetrical distribution having its mean, mode and median equal. Also, the frequencies of the variable extend equally both to the left and to the right of the mode.

Variables: are those that vary in quality or those that vary in quantity.

Quantitative Variables: are those that take values in terms of magnitude. These may be scores obtained in a test, weight, height, age, distance, number etc.

Qualitative Variables: are those that differ in kind such as marital status, gender, nationality, and social economic status, educational qualifications etc. quantitative variables maybe discrete or continuous.

Discrete Variable: is one which can take only a finite set of values, implying that fractional values are usually not allowed. These variables are generated by counting process usually in whole numbers. Example number of teachers in a school, number of male and female in a class, etc.

Continuous Variable: is that which can take on any value over a range of feasible values. Measured data can be whole numbers or fractions, such as height, weight, distance values etc. variable could also be independent variable or dependent variable.

Independent Variable: is one that is manipulated or treated. The effect of this manipulation is manifested on the other variable. Effect of teaching method on academic performance. The teaching method is the independent variable while dependent variable depends on that of the independent variable. Here performance is dependent variable. Note that in grouping the dependent variable is placed on the vertical, Y – axis while the independent variable is placed on the horizontal, x – axis. There are two types of independent variables – treatment or active variables and organism or attribute variables.

Treatment or Active variable: this is defined as one that can be directly manipulated by the researchers. This group includes method of teaching, method of grouping and reinforcement procedures.

Organismic or Attribute Variables: these are those variables that cannot be actively manipulated by the researchers. They are characteristics of individuals that cannot be manipulated at will. Such as age, aptitude, social class, race and intelligence level had already been determined but the researchers can decide to include or remove them as variables to be studied.

Organization and Presentation of Data

The field of statistics can be divided into two areas, namely descriptive and inferential statistics. Descriptive statistics deals with collection; presentation and description of numeral data while inferential statistics deals with set of methods used for drawing inferences about a population from data available on only a representative sample of that population.

Organization of Data

In any research investigation, after the researcher has collected relevant data, you

need to organize or prepare your data by coding the data if the computer will handle the analysis or by summarizing and present the data in tables or graphs.

Coding Data

If researcher wants to use computer in analyzing data, the computer is easier and more economical. For the data to make sense to the computer, you need to represent each category with numerical values. You may use “1” to represent male and “2” to represent female values to data that are in categories (i.e data that are not numerical).

Frequency Distribution Table

Another way of organizing and presenting data is by using frequency distribution table. There are basically two kinds of tables used in organizing and presenting data. They are ungrouped and grouped frequency tables.

Frequency distribution table is a table which shows how many times each score in a distribution occurred. It consists basically of three columns – score column, tally column and frequency column. More columns could be add depending on what other information is required.

Preparing an ungrouped frequency distribution table, the researcher needs three basic steps.

Step I: List the scores either in ascending or descending. That is in order of magnitude in first column of the table.

Step II: Tallying the scores. Tallying of scores involves placing a stroke against a score each time the score occurs in the distribution in the column 2 of the table.

Step III: Determining the number of times each scores occurred. The number of strokes at each score is counted and written

down as frequency of that particular score, that in column 3 of the table.

Preparing a grouped frequency distribution table, whenever the researcher has a large number of scores, it may be laborious to list all the scores. A convenient and faster method is to group the scores before tallying them. The following procedure should be adopted, when preparing a grouped frequency distribution table.

Step I: Group the scores into class intervals. To do this one need to understand the following concepts: class interval, class size, class limit and class boundaries.

Step II: Tally the scores in column 2 using the class intervals.

Step III: Write down the frequency of occurrence of each class interval in column 3.

Using Graphs to Represent Data

Graphs are often used in presenting frequency distributions. Graphical methods include: Pie chart, bar chart, histogram and frequency polygon. But the most commonly used graphs in representing data in educational research are the bar chart and the histogram.

Bar Chart

To prepare the bar chart, draw the vertical and horizontal axes. Label this horizontal axis a "X" and the vertical axis a frequency, "F". The bars are of the same width and that is evenly spread over the score.

Histogram

In drawing the histogram, you exactly repeat the steps involved in drawing a bar chart. The only differences are that in drawing the bars, you allow the points to make contact. The first step towards analyzing data is to prepare and organise

them in a way that will make analysis possible and easier. If the researcher has done that, then the simple statistics to be used for analysis which are descriptive in nature are: Percentage, Mean, Median, Mode and Range.

Percentage

This is a simple yet useful statistic for describing information about people, events or situations. It is particularly useful in describing the characteristics of two or more groups of people or objects where the numbers of people or objects in the groups are different.

The Mode

The mode is a measure of central tendency that indicates the most frequently occurring score in a distribution. It is important to note that a distribution of scores may have more than one mode. A distribution with one mode is said to be unimodal while that with two modes is said to be bimodal. If these are more than two modes in a distribution such distribution is said to be multi-modal in nature.

The Median

The median is the score of value at the middle or centre of any set of scores or values. It divides the set of scores or values into two, each containing some number of scores. To find the media, first arrange the scores or values in order of their size (magnitude) (either ascending or descending order). When the number of scores is odd. The score find in middle is the median but if the number of scores is even, then two scores will appear at the centre of the distribution. To find the median when the number of scores is even, you add the two scores at the center and divide the sum by 2. The result is median.

The Mean

The mean is the most common measure of central tendency. It is the arithmetic average of a set of scores. To find the mean, you add all the values in a distribution and divide by the number of values in the distribution.

The Range

The range is a simple statistic for indicating how far apart the scores in distribution are from one another. In other words, it shows how scattered the scores in a distribution are. There are other statistics which also perform this function better than the range. These include the mean deviation and standard deviation.

Standard Deviation

It shows how much variation or dispersion from the average exists. A low standard deviation indicates that the data points tend to be very close to the mean, a high standard deviation indicates that the data points are spread out over a large range of values.

Hypothesis Testing

The word hypothesis can be referred to as an assumption used as a basis of action. It is therefore an assumption or verification and not established facts. Hypotheses are used as indicators of the realistic answers which researchers have to their stated problems or questions in research.

Generally, hypotheses are stated in two forms, namely: null hypothesis symbolized by H_0 and alternative hypothesis, symbolized by H_1 and H_a .

A null hypothesis is a hypothesis of no difference or relationship. it is a statement to the effect that there is no significant difference or relationship between two populations between two parameters of two population or more parameters. An alternative hypothesis is in form: there is a significant difference or relationship

between two populations between two parameters of two population or more parameters.

It should be noted that for every hypothesis is in line with specific research question. For example.

Is there any significant difference in the mean score of students taught statistics using scaffolding method and those taught using lecture method? The appropriate null hypothesis will be.

H₀: There is no significant difference in the mean score of students taught statistics using scaffolding method and those taught using lecture method.

H₁: There is significant difference in the mean scores of students taught statistics using scaffolding method and those taught using lecture method.

Hypothesis testing can be regarded as a special type of decision-making process. Testing a hypothesis depends on several steps and the steps are presented below:

Step 1: Statement of Hypothesis

Step 2: Selection of level of significance

Step 3: Selection of relevant test statistics and computation of its value from sample data.

Step 4: Determination of rejection and acceptance regions (i.e critical region).

Step 5: Statistical decision

Step 6: Conclusion

Statement of Hypothesis

Discussion was made on how to state hypotheses. Hypotheses are suggested

answers to the research problem under investigation. All hypotheses stated must go through the process of hypothesis testing.

Selection of Significance Level

Any time, a researcher is testing a null hypothesis, it is essential to give a chance of rejection. On that basis when a researcher wants to analyze data collected, risk level is selected for rejecting or accepting the null hypothesis. This risk level is normally stated in terms of probability level of significance. It is therefore the probability value that would enable a researcher to either reject or not to reject a null hypothesis. It can also be regarded as the amount of error involved in a given statistical decision about the null hypothesis.

This therefore means that whenever a researcher selects a particular level of significance, then you know precisely the amount of error risk involved in decision. Level of significance is symbolized by " α " (alpha) which is the probability value that forms the boundary between rejection or non-rejection of null hypothesis. In educational and psychological research, it is good and conventional to use 0.05 and 0.01 levels of significance.

Statistical Analysis

To select appropriate statistical test for any data analysis should be based on one or two of the following:

- i. The research design of study either experimental or non-experimental
- ii. Type of statistics: descriptive and inferential
- iii. Features of the population and sample data that is parametric and non parametric
- iv. Level of measurement, that is nominal, ordinal, interval & ratio

Research design by the researchers' specific objective, which is link with research questions and hypotheses and data collected for the study. If the data is quantitative then the design is either experimental or non experimental. In experimental design, two groups are compared based upon some conditions and provide room for randomization, control and manipulation. If otherwise, it is quasi-experimental, which offer fewer well controls and intact class is used. Most educational experiment falls into this category. Expost-facto or correlation design is used when data generated is from secondary source and the required data is from a large sample of population. Again, it is not possible to manipulate variables, to assign subjects or conditions at random. Thus, descriptive/survey design is appropriate here, while data collected will be used to describe and compare. If the data generated are purely words (alphabetic), then the appropriate design is qualitative resign. The design can be historical, documentary, case study, etc.

On using the statistics, the general rule is that if you only have research questions, you will use simple statistics like mean, median and percentage. If, however you have hypotheses which are more specific then you need to use powerful statistics for hypotheses testing, like T-test, Z-test, Chi-square etc.

Generally, means, mode, range and percentage are calculated in order to get an idea about the data. These can assist to answer research questions. If you intend to analyse the data manually i.e by calculating things with pen and paper by yourself then you should be careful not to make mistake. Doing the analysis manually can also be hectic and time consuming, but these days, there are people employed in the computer

rooms of schools and other educational institutions. They can assist you to analyze your data for you, if you explain to them what your research hypotheses are. Now let going into selection of appropriate statistical tool for hypotheses testing.

Selection of Relevant Statistical Tool and Computation

Statistical tests are categorized into two groups, namely parametric and non-parametric test, any of these tests are selected depend on the data collected.

Parametric tests: These are testing whose efficacy tests whether the variable being studies is least appropriately normally distributed.

It can be referred to as statistical procedure which makes inferences about the population parameters such as mean. In using the test, three important assumptions must be made to test a hypothesis.

- i. The normal distribution is normally distributed in the population.
- ii. The normal distributions have the same standard deviation.
- iii. The data is taken from an interval or ratio scale.

If above assumptions are made, then the parametric test such as z-test, t-test, PPMC test and F-test could be applied.

The Z-test

The Z-statistic is used in testing hypothesis involving one sample. It is recommended for testing hypothesis where the sample size is large i.e It is equal to or greater than 30 (≥ 30).

The T-Test

The t-test statistics was developed by William Gosset an employee of Guinness

Brewery Dublin. The t-test is used to determine whether two means are significant when the sample size is small (i.e $n > 30$). The original data in comparison of two groups, there must always, exist a difference no matter how identical the groups are. There are two different types of t-tests. The T-test for independent samples and the T-test for non-independent samples.

Independent samples are samples which are randomly formed. The members of one group are not related to members of the other group in any systematic way other than that they are selected from the same population.

In a nutshell, the t-test of independent samples is used to determine whether there is probably a significant difference between the means of two independent samples. For instance, a researcher maybe interested in finding out whether a significant difference exists between mean scores of two groups of students who were taught mathematics with Instructional scaffolding and traditional method after the administration of the two treatments.

Non-independent samples formed by some type of matching. When samples are not independents, the members of one group are systematically related to the members of a second group. Therefore, the t-test for the non-independent samples is used to determine whether there is probably a significance difference between the means for one sample at two different times. For example, researchers want to investigate whether there is significant difference on scores in both mathematics and statistics of students. t-test for non-independent can be used here.

Analysis of Variance (F-Test)

Analysis of variance abbreviated ANOVA was developed by R.A Fisher and sometimes

been referred to F-test. ANOVA is an effective way of determining whether the means of two or more independent samples drawn from populations having the same mean suppose the researcher want to find out whether there is a significant difference in the mean scores of a sample of students in four subjects. The methods of F-test will enable you to test the hypotheses.

PPMR-Test

This is an acronym for Pearson Product Moment Correlation Coefficient. It is type of correlation coefficient, named after the man who developed it, it is used when the two sets of data are continuous or interval data. The PPMR – test is used to test hypothesis of no significant relationship between variables, that are 30 or more pairs of variables. For instance, the researcher wants to test the null hypothesis that whether there is significant relationship between students attitudes to school and academic performance.

Non-Parametric Test

There are so many non-parametric tests than parametric test. These tests are developed without reference to the distribution of variables. The non-parametric statistics certain assumptions cannot be made about the populations involved, but the researchers can use the tests, because they are easier to learn, the mathematical derivation are more readily understood, they can be applied to data of a classificatory nature that are common in behavioural research (Normal or Ordinal Scale), they handle ranked data better and the computations are relatively simple. Such tests are:

CHI-SQUARE (X^2) TEST

The Chi-square test is a measure of relationship, association or independence, introduced by Karl Pearson in 1900. The

Chi-square test is best known and most important of all non-parametric test. There are five basic conditions that must be met for Chi-square analysis to be validly applied.

These are:

1. The sample observations are independent of each other
2. Sample data are drawn at random from the population
3. Sample data are expressed in original units
4. The sample should contain at least 50 observations
5. There should be no less than five observations in any one cell
6. Not more than 20% of the expected frequency should be less than 5 and data are classified into nominal, non-ordered categories.

Sign Test

The sign test is based on the difference between two pairs of observations without regards to magnitudes. It gets its name from the fact that plus and minus signs are its raw data. The only necessary assumption for this test is that the variable being studies has a continuous distribution.

Wilcoxon Matched-Pairs Signed Ranks Test

It is s test used to passes the significance of difference between two samples consisting of matched pairs of subjects. Such matched pairs of subjects would include two measures taken on the same subject as in a pre and post comparison of individuals. Wilcoxon is the non-parametric counterpart of the t-test for correlated data. The null hypothesis tested is that there is no difference between the two populations

from which the matched comparisons are drawn.

Statistical Decision

In testing null hypotheses, the researcher has to compare the computed value of test statistics with the critical values presented in tables to enable you to retain or reject the hypotheses. Before, this comparison, the researcher needs to set level of significance and also determine degree of freedom. Level of significance has been discussed.

Now let briefly explain the concept of degree of freedom (df) Degree of freedom is the number of observations which are free to vary when certain restrictions have been placed on the data being considered. Take for instance, in your mathematics class, you ask your students to provide any four numbers which would be added to 5 to add up to 25. In this case, it is fixed, other numbers can vary. But no matter how they vary 5 must be added to sum up to 25. Therefore, the degree of freedom here is $N - 1$ where N is the total number of choices while 1 is the fixed variable. Different test statistics have different ways of determining their degrees of freedom. With the knowledge of the degree of freedom and level of significance, the researcher can now find the critical value of a particular test that is required for either rejecting or retaining the null hypothesis. Statistical decisions are made in terms of the test statistics.

Drawing Conclusion

Thus, after making the critical decision about whether rejecting or retaining a null hypothesis. Then the next step is to draw conclusion. The conclusion is normally stated in terms of the parameters of population which the test is concerned. If the null hypothesis is rejected then we conclude that there are significance differences or significance relationship but

if the null hypothesis is not rejected, we conclude that the difference between the samples is therefore not significant.

Researchers should therefore adopt the foregoing procedure (step 1 -6) in testing hypotheses formulated to guide the studies.

References

- Agwagah, U. (2013). Improving the teaching of mathematics language on errors commitment by senior secondary school students in bearing problems in Omuo- Ekiti, Nigeria.
- Abacus the Journal of the Mathematical Association of Nigeria*. 38(1) 50 -56.
- Ahmad, M. U., Etsu, S.M & Abubakar, S. N (2019). Students' perception of factors affecting academic performance in mathematics.
- Proceeding of September 2019 Annual Conference, MAN*. Pp508 – 514.
- Awotunde, P.O & Ugoduluwa (2002). An Introduction to statistical methods in education. Printed and published in Nigeria Fab Anieh Nig. Ltd. Jos.
- Chianson, M.M. Kurumeh, M.S & Obida, J.A (2011). Effect of cooperative learning strategies on Students Achievement in geometry in secondary schools in Benue State. *Journal of Mathematical Science* 1(1) 115 – 121.
- Joy, N.A(2014). Students' perception of factors affecting academic performance in chemistry. *Journal of Educational Research and Development*. 8(2),241-256.
- Kajuru, Y.K & Isah, A (2014). Influence of cooperative learning strategy on performance in geometry among junior secondary school students in

Sokoto metropolis for employment generation and sustainability. *Proceeding of September 2014 Annual Conference, MAN. Pp33 – 53.*

mobilizations, management for access and quality education. *Proceeding of Multicultural African Conference 2013. A.B.U. Zaria.*

Kajuru, Y.K & Popoola, F.R (2010). Pedagogical strategies for improving the teaching and learning of mathematics at College of Agriculture in Nigeria. *A Journal of Studies in Science and Mathematics Education. 1(1): 33 – 41.*

Odetola, C.A & Salom, M.F. (2014). Effect of mathematical language on errors committed by senior secondary schools in bearing problem in Omuo-Ekiti, Nigeria. *Abacus, the Journal of Mathematical Association of Nigeria 37(1) 101 – 109.*

Maruf, O.I & Sambo, S. (2013). Teachers and students' perceptions on statistical analysis as a tool for resource

West African Examination Council (WAEC) (2016, 2017, 2018). Chief Examiner reports.