ISSN: 2315-5701

OSCOTECH JOURNAL OF ARTS AND SOCIAL SCIENCES (OJASS)

A BI-ANNUAL ACADEMIC JOURNAL OF THE FACULTY OF MANAGEMENT SCIENCES,

OSUN STATE COLLEGE OF TECHNOLOGY, ESA OKE

SEPTEMBER, 2016 EDITION

http://ojass.oscotechesaoke.edu.ng/en/

Vol. 3 No. 1

Page 203 - 216

Statistical Analysis and Tests; A Vital Tool For Research Study

S. O. Jabaru, & T. A. Salami

Department of Mathematics and Statistics, Osun State College of Technology, Esa-Oke, Osun State

&

K. Jimoh

Department of Statistics, Al-Hikmah University, Ilorin

Tel: 08036209117

E-Mail : jimkaminsha@yahoo.com

Abstract

Many research works are characterized by abuse and misuse of statistical tools/concepts due to lack of proper understanding of how and why such tools are used. Misuse of statistical tools always yields misleading results, the consequences of which may be calamitous. This research work therefore examines common statistical tests and concepts, including their assumptions as regards their applications in research work.

Keywords: statistics, tests, assumptions, t-test, z-test, chi-square, correlation

1.0 INTRODUCTION

Statistics is the art of learning from data. It is concerned with the collection, organization, presentation, description and analysis of data which often leads to drawing reasonable conclusions (Salami et al, 2008). It is a known fact today that in order to learn about something (Research), one must deal with data (numeric or alphabetic), thus the six highlighted points above is a must in order to carry out a successful and meaningful research.

In any research involving statistical analysis, you must have a clear idea of the statistical procedure you want to perform. Using statistical package to do statistical analysis is easier when a data analysis plan is available. Therefore, you must have a data analysis plan. A data analysis plan includes; coding procedure, how some variables will be transformed, attributes of the variable that will be presented in the final report (descriptive statistics), required statistical tests, and other statistical analysis that will be done.

Based on their major cluster of functions, statistics could be categorized into two types – Descriptive and Inferential. It could be further dichotomized into parametric and non – parametric types (Isyaku, 1994). Descriptive statistics attempt to organize, summarize, tabulate and describe collection of data such as test scores, frequency counts, ranks etc which may be represented in various forms, e.g. in numerical forms, tables and graphs. Under descriptive statistics are studies such as frequency distribution, measures of central tendency, measures of association and measures of variability.

Inferential statistics on the other hand, deals with conclusions, assumptions, postulations, summations and extrapolations arrived at

scientifically based on available data. For instance, an inference on the population could be made based on analysis of data from a sample of that population. The various tests of significance in educational research come under inferential statistics, e.g. z – test, t – test, analysis of variance, chi square test, Wilcoxon Rank Sum test, etc.

2.0 TYPES OF VARIABLE

Variables are events or qualities that can vary and so assume more than one value. Variables also refer to observable characteristics of an object or person that belongs to a group of objects or persons respectively (Olaitan and Ndomi, 2000). In another perspective, Anikweze (2010:140) explains variable as "a property by which the members of a group differ from one another". Example include differences in sex, age, complexion, ethnic group, location and religion – all the six examples are classification variables used for identification of a group of individuals.

Ferguson (1989) refers to this as labelling. Labels can be used to classify variables into levels, e.g. male and female for sex, rural and urban for location, and varied number of levels for multi – valued variables such as occupation into farmers, teachers, shoemakers, pilots, nurses, accountants, etc.

In research, two distinct types of variables are reckoned with. These are *independent and dependent variables*. An independent variable is a factor that is selected and manipulated by the researcher in order to study its effects on the dependent variables. Such factors are often described as *active, stimulus, input, treatment or predictor variables*. Some independent variables

are sometimes controlled in order to eliminate their effect on the dependent variable. There could also be intervening variables which are factors that are intangible but capable of affecting the dependent variables through their effects on the independent variables whether treatment or moderator types.

A dependent variable is also called *target variable*. It refers to a measurable behaviour exhibited by the participant or subject of experiment. The behaviours vary as the independent variable is manipulated. Thus, any change in the independent variable results into a corresponding change in the dependent variables. Note that there could be *extraneous variables* that are not connected with an experiment but which affect the outcome which is the dependent variable. In a good research design, such variables are tactically controlled or excluded. To permit comparability, dependent variables are measured with the same type of scale as their independent variables.

3.0 LEVELS OF MEASUREMENT

The type of model or statistical test we choose to analyze our data will depend upon the level at which the data is measured. Data measurement is traditionally characterized as being divided into four possible levels, although for practical purposes we can usually speak of just three (the difference between interval and ratio measures being of little real importance to the analyst):

Nominal data

This type of data has no order, and the assignment of numbers to categories is purely arbitrary (e.g. alive = 1, dead = 2). Because of lack of order, it is not possible to perform arithmetic or logical operations on nominal

data. Nominal data is also commonly referred to as categorical, as it assigns observations to qualitative categories.

Ordinal data

This type of data has qualitative order but the intervals between scale points are unequal. For instance, although we can order all the football teams in the Nigerian premier league in terms of their level of ability, the interval distance from the top team to the second highest team may be great, but the interval from the second ranked team to the third ranked team may be very close. Because of lack of equal distances, arithmetic operations are inappropriate with ordinal data which are restricted to logical operations (more than, less than, equal to). For example, given a team finishing 12th and a team finishing 6th in a league of 24 where 1 is top of the table, we cannot divide 12 into 6 to conclude that the second team has twice the achievement of the first (although statistically unsophisticated, sports fans often attempt such operations). However, one can say that the first team represents more achievement than the second one.

Interval data

This type of data has quantitative order and equal intervals. Counts are interval, such as counts of income, years of education or number of votes.

Ratio data

This is the type of data that are interval data which also have a true zero point. The Celsius temperature scale is not ratio because “zero degrees is not no temperature” but income is ratio because zero pounds is truly “no income”. For most statistical procedures, the distinction between interval and ratio does not matter and it is common to use the term interval to refer to ratio

data as well. Occasionally however, the distinction between interval and ratio becomes important. With interval data, one can perform logical operations situation, add, and subtract, but one cannot multiply or divide. e. g. if a liquid is 40 degrees and we add 10 degrees, it will be 50 degrees. However, a liquid at 40 degrees does not have twice the temperature of a liquid at 20 degrees because 0 degrees does not represent "no temperature" – to multiply or divide in this way we would have to be using the Kelvin Temperature scale, with a true zero point (0 degree Kelvin = -273.15 degrees Celsius). Researchers should be aware of the statistical issues involved.

4.0 MEASURES AND APPROPRIATE SCALES

Two ordinal/interval variables: Correlation

One nominal and one ordinal/interval variable: Comparisons of means with t – test

Two nominal variables: Crosstabulation and Pearson Chi Square.

For common multivariate procedures, the following applies:-

Multiple linear regression: ordinal/interval dependent variable; ordinal/interval or nominal independent variables, although nominal variables need to be recoded to dummy variables.

Factor analysis: ordinal/interval variables although for some more advanced procedures, ordinal dichotomous variables can be used in some cases. Note there are no dependent or independent variables in factor analysis.

Logistic Regression: nominal/ordinal; (dichotomous) dependent variable, ordinal/interval or nominal independent variables, although nominal variables need to be recoded to dummy variables.

5.0 STATISTICAL ANALYSIS AND ASSOCIATED TESTS

In inferential statistics, the Z and student t – test prove useful when carrying out test of hypothesis on differences in means of one or two samples depending on whether the population parameter is known or unknown. For three or more sample test, differences in means can be tested using Analysis of Variance (ANOVA). In experimental research involving pre – test and post – test designs, one can use the pre – test as co – variance in order to obtain more robust result using Analysis of Co - variance (ANCOVA). Multivariate data analysis is best done using the tool of statistical software. According to Adedayo (2006), outputs of advanced statistical techniques like factor analysis, discriminant analysis, path analysis, Multivariate Analysis of Variance (MANOVA) enables one to get deeper insights to the data and obtain high level of sophistication of the results.

1. Data types that can be analyzed with Z – tests

i. Data points that is independent of one another.

ii. Data with sample size (n) greater than 30.

iii. Data having normal distribution most especially when the sample size (n) is small (less than 30). If however n > 30 the distribution of the data does not need to be normal.

iv. Sample data with equal variances

v. Data with individuals selected at random from the population.

vi. Data with equal sample sizes, though some differences are allowed.

2. Data types that can be analyzed with t – tests

i. Data sets that are independent of one another except in the case of the paired sample t – test.

ii. Data sets with n < 30.

iii. Data sets with normal distribution for the equal and unequal variance.

iv. Data sets having the same sample variance.

v. Data set whereby individuals are selected at random from the population.

vi. Data sets with equal sample sizes, however some differences are allowed.

NOTE: There are one sample t-test, two sample t-test and independent sample t-test.

3. Analysis of Variance (ANOVA)

1. Observations are independent. (The value of one observation is not related to any other observations).
2. Homogeneity of variances: Variances of the dependent variables are equal across groups.
3. Normality: The dependent variable in each group is normally distributed.

4. Data types that can be analyzed with Chi-square test

1. Data with independent samples.
2. Data that are treated as nominal variable.

5. Data types that can be analyzed with multiple linear regression model

1. Independent samples.
2. Data with normally distributed errors.
3. Data with constant variance of the residuals (homoscedastic error variances).
4. Data with uncorrelated predictor variables (no collinearity).

6. Nature of data and different types of correlation coefficient

Recall that a correlation coefficient is a single number that tells to what extent two things are related and the extent of which variations in one go with variations in other. The simplest methods of computing correlation coefficient are the Pearson's Product Moment (r) and the Spearman's Rho (ρ). Occasions arise when nature of data demands different approaches to determining the correlation coefficient.

7. Different types of correlation coefficient

1. ***Pearson's Product Moment***: This is the method used if the first variable is normally distributed and the second variable is also normally distributed.
2. ***Spearman's Rho***: This is used if both variables are normally distributed but expressed in ranks.
3. ***Biserial correlation***: This is used if one variable is normally distributed and the other is normally distributed but artificially dichotomized. For instance, we may wish to find out if any relationship exists between salary earnings and support or opposition to the ruling party. In a test that determines whether a student passes or fails, we assume a continuum along which individual differs with respect to achievement

that is required for passing. Those scoring above the certain critical point pass and those scoring below it fail. What we have done is to bundle the heterogeneous variables in achievement to only two artificial groups.

4. ***Point – Biserial Correlation:*** This is computed when one variable is normally distributed and the other variable is genuinely dichotomized, e.g. relationship between performance in Mathematics and Sex, or the correlation between scores in R.K. and being alcoholic/non – alcoholic. This is expressed as
5. ***Phi Coefficient***: This is a measure of the degree of association between two binary variables (both variables are genuinely dichotomized) e.g. relationship between gender (male or female) of students and their academic performance (pass or fail). Two binary variables are considered positively associated if most of the data falls along the diagonal cells while they are considered negatively associated if most of the data falls off the diagonal. (Simon, 2005). The Phi coefficient is given by

$$\phi = \frac{bc - ad}{\sqrt{(a+b)(c+d)(b+d)(a+c)}}$$

Examples

1. Suppose we need to relate 20 students' performance (scores) in religious knowledge to their interest in religious activities. We may show further interest in relating the interests of the students in religious activities to sex as a factor. The data in the table below may be collected for analysis.

Table 1: Students' Scores in Religious Knowledge by Interest in Religious Activities and Sex

S/N	R.K. SCORES	INTEREST	SEX
1	40	Active	M
2	30	Dull	M
3	15	Dull	F
4	15	Active	F
5	70	Active	F
6	80	Dull	F
7	65	Dull	M
8	30	Active	M
9	15	Active	F
10	50	Dull	M
11	45	Dull	F
12	60	Active	F
13	70	Dull	M
14	60	Dull	F
15	40	Dull	M
16	25	Active	F
17	30	Dull	M
18	70	Dull	M
19	50	Dull	F
20	40	Active	F

2. Suppose it is required to test if the average score of students in a particular course is equal to 40% given the following scores (in %) of 10 students in that course.

45, 56, 66, 42, 45, 53, 36, 61, 35, 39.

Solution: **Use One sample t-test**

3. To test if the average scores of Arts and Science students in a general (GNS) course differ given the following scores (in %) of 10 and 12 students in that course from the two groups.

Art: 45, 56, 66, 42, 45, 53, 36, 61, 35, 39.

Science: 59, 63, 66, 53, 63, 62, 52, 66, 59, 67, 57, 61.

***Solution: Use t*wo sample t-test**

4. The efficacies of four newly developed chemotherapies (A, B, C, D) for curing Hepatitis B are to be compared. The table below shows the number of cured and uncured patients (outcome) after administering the four therapies on 397 patients. Base on this information can we conclude that cure from the ailment depends on the type of chemotherapy treatment received?

Drugs	Cured	Not Cured	Total
A	78	2	80
B	90	7	97
C	88	18	106
D	130	15	45

Solution: Use Chi-Square test

6.0 COMMON MISUSE OF STATISTICAL TOOLS

1. Data collection before setting the research objectives: Statisticians are not magicians that can mold or manipulate the data to suit the researcher's expectations. Data collected before research objectives are set might not possess the needed information to achieve the stated objectives.

2. Not caring about the types of variables (scales of measurement) being investigated.

3. Assuming all the results must always be statistically significant.

7.0 CONCLUSION

Researchers use a wide range of statistical methods to analyze data. However, many forms of data analysis can be done with statistical packages such as Excel, SPSS etc (Hall, 2011). The simplest type of educational research is survey. Data from survey could be easily analyzed using descriptive statistics. The relevance of the validity and reliability of instruments for collecting research data cannot be overemphasized. For one thing, if a research instrument is invalid automatically, the outcome of the research endeavour would also be invalid. The research effort would have been a waste. Furthermore, it is only valid and reliable instruments that can produce data useful for inferences that have predictive meaning. Only researches based on reliable and valid instruments can lead to solutions that can stand the test of time. In addition, conclusion based on data from invalid instruments would be baseless if not laughable. On the other hand, judgements based on the results from valid instruments cannot be contradicted because they are based on reliable empirical evidence.

8.0 REFERENCES

Anikweze, C. M. (2010). Measurement and Evaluation for Teacher Education, 2nd Edition, Enugu, SNAAP Press (Nig.) Ltd.

Ferguson, G. A. (1989). Statistical Analysis in Psychology and Education (5th Ed.), Auckland: McGraw-Hill International Book Company.

Hall, S. (2011).What Statistical Tools are Used in Survey Research? Downloaded from www.ehow.com/about statistical analysis used in survey research.

Isyaku K. (1994). Basic Statistics for Education and Social Sciences, Kaduna, National Teachers Institute Press.

Olaitan S. O. & Ndomi B. M. (2000). Vocametrics – A High-Tech problem solving Quantitative Text with Computer Skills, Onitsha: Cape Publishers Int. Ltd.

Salami T. A., Jimoh K and Jabar S. O. (2008). Basic concept of economic and social statistics, published by Sunyus ventures, Esa-Oke, Osun State.

Yahya W.B. (2013). Statistical Packages and data analysis in postgraduate research in Nigeria, a paper presented at the National Stakeholders' Summit on postgraduate Research in Nigeria