BASIC STATISTICAL TECHNIQUES IN RESEARCH

Chapter · April 2018

CITATIONS
0

READS
54,331

2 authors, including:

Musibau Adetunji Babatunde
University of Ibadan
90 PUBLICATIONS **769** CITATIONS

SEE PROFILE

Data Collection, Management and Analysis in Academic Research

Proceedings of a Workshop

Edited by
Labode Popoola, Olajide Olorunnisola and
Olusegun Ademowo

2009

CONTENTS

CHAPTER 1

BASIC STATISTICAL TECHNIQUES IN RESEARCH

F.A. Adesoji and M.A. Babatunde

Introduction

The essence of research is to solve problem(s). In order to do this, we have to select a plan for this study which is being referred to as "design of the experiment". In this case, research can be regarded as the application of the scientific method to the study of a problem. Question/hypotheses are thought of in an attempt to find solution to the problem at hand. In order to gather information or data, instruments must be in place. The data collected are not useful because they are referred to as raw data or untreated data until they are analyzed by using appropriate statistical tools.

The design of experiment is inseparable from the statistical treatment of the results. If the design of an experiment is faulty, no amount of statistical manipulation can lead to the drawing of valid inference. Experimental design and statistical procedures are two sides of the same coin. Any research that deals with the manipulation of variables which are basically of two types. These are, numerical and categorical. Numerical variables are recorded as numbers such as height, age, scores, weight, etc. Categorical variables could be dichotomy (for example, male or female), trichotomy (for example, high, medium and low economic status) or polychotomy (for example, birth places). Statistical techniques have to do with data generation, manipulation and interpretation. In order to generate data, measurement is necessary. Measurement is the ascribing of symbols or figures to entities and it is thus basic

to data analysis, interpretation and research in general. This is done with the aid of well validated instruments.

The study of data is called STATISTICS and in a more refined way, it deals with scientific and analytical methods of collecting, organizing, analyzing and presenting data in such a way that some meanings and conclusions could be made out of something that appears to be jungle of data. Statistics can be very broadly classified into two categories, viz, descriptive and inferential statistics. Descriptive statistics refers to the type of statistics, which deal with collection, organizing, summarizing and describing quantitative data. For example, the average score describe the performance of the class but does not make any generalization about other classes. Examples of descriptive statistics are graphs, charts (pie charts, columinal charts, bar charts, histogram, etc), Pictograms, tables and any form whereby data are displayed for easier understanding. Other examples are measures of central tendency (mode, mean, median), correlation coefficient (degree of relationship), kurtosis, skewedness etc.

Inferential statistics deals with the methods by which inferences are made on the population on the basis of the observations made on the smaller sample. For example, a researcher may want to estimate the average score of two or more classes of an English course by making use of the average score of one class. Any procedure of making generalization that goes beyond the original data is called inferential statistics. The statistics provide a way of testing the significance of results obtained when data are collected. It thus uses probability, that is, the chance of an event occurring. Examples of inferential statistical tools are student t-test, Analysis of Variance, Analysis of Covariance, Correlation Analysis, Multiple regression, Multivariate Analysis of Variance etc.

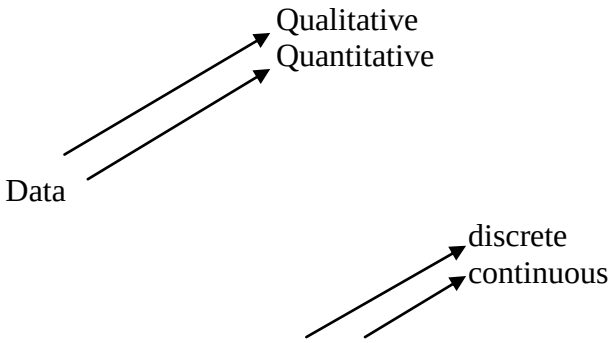
An area of inferential statistics called hypothesis testing is a decision making process for evaluating claims about a population based on information obtained from samples. Relationship among variables can also be determined. Also, by studying past and

present, data and conditions; it is also possible to make predictions based on this information. It would be observed that descriptive statistics consists of the collection, organization, summarization, and presentation of data while inferential statistics on the other hand, consists of generalizing from samples to populations, performing estimation and hypothesis testing, determining relationships among variables, and making predictions.

Nature of Variables and Types of Data

Variables can be classified as qualitative or quantitative. Qualitative variables are variables that can be placed into distinct categories, according to some characteristics or attributes. For example, categorization according to gender (male or female) then, variable gender is qualitative and it takes categorical data, let us say 1 or 2. Other examples are religious preference, ability level and geographical locations. Quantitative variables are numerical and can be ordered or ranked. For example, the variable age is numerical and people can be ranked in order according to the value of their ages. Other examples are heights, weights and body temperatures.

Quantitative can be further classified into two groups: discrete and continuous. Discrete variables can be assigned values such as 0,1,2,3 and are said to be countable. Examples of discrete variables are the number of children in a family, the number of student in the classroom etc. Thus, discrete variables assume values that can be counted. Continuous variables, by comparison can assume all value in an interval between any two specific values. Temperature is a continuous variable since it can assume all values between any two given temperatures. Data could also be categorized as purely numerical and not purely numerical. Mean cannot be determined in a not purely numerical data. For example, data involving the number of people and their mode of collecting information.



The mean can not be determined in this case compare with the scores of group of students in a course of study. Data can also assume nominal level (assigning A, B, C), or ordinal (ordered or ranked), interval (precise difference exist) or ratio.

For example, scores of 50% and 51%- a meaningful one point difference exists, but there is no true zero point. For example, 0^{o}F does not mean no heat at all, and the ratio- here, in addition to the difference between units, there is a time zero point and true ratio between values. Note that there is no complete agreement among statisticians about the classification of data into one of the four categories. Also, data can be altered so that they fit into a different category. For example, if you categorize income of workers into low, average and high, then a ratio variable becomes an ordinal variable.

Parametric and Non-Parametric Tests

The distribution of many test statistics can be said to be normal or follows some form that can be derived from the normal distribution. A characteristic property of the normal distribution is that 68% of all of its observations fall within a range of ±1 standard deviation from the mean, and a range of ±2 standard deviations includes 95% of the scores. In other words, in a normal distribution, observations that have a standardized value of less

than -2 or more than +2 have a relative frequency of 5% or less.[1] The exact shape of the normal distribution (the characteristic "bell curve") is defined by a function which has only two parameters: mean and standard deviation.

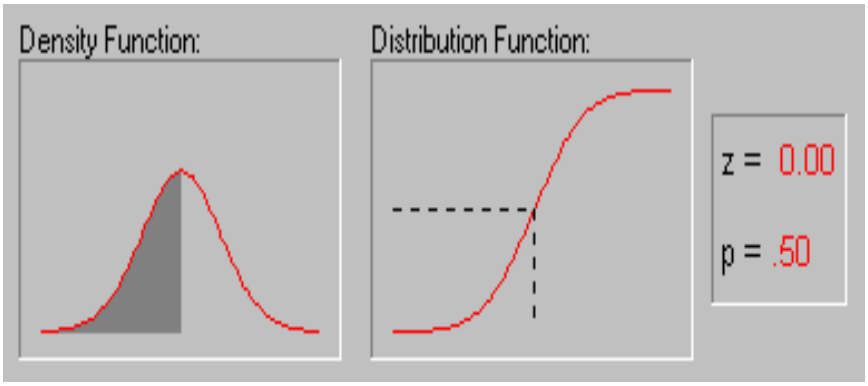


Figure 1: Probability Density Function. The left hand side of the graph represents a Standard Normal Distribution Function

The attempt to choose the right test to compare measurements may however be a bit difficult, since we must choose between two families of tests: parametric and nonparametric. Many statistical tests are based upon the assumption that the data are sampled from a normal distribution. These tests are referred to as parametric tests. Parametric statistics are statistics where the population is assumed to fit any parametrized distributions (that is, most typically the normal distribution).[2] Commonly used parametric

[1] Standardized value means that a value is expressed in terms of its difference from the mean, divided by the standard deviation.

[2] The normal distribution, also called the Gaussian distribution, is an important family of continuous probability distributions, applicable in many fields. Each member of the family may be defined by two parameters, the mean ("average", μ) and variance (standard deviation squared) σ^2, respectively. The standard normal distribution is the normal distribution with a mean of zero and a variance of one.

tests include the Mean, Standard deviation, the t-test (One-sample *t* test, Unpaired *t* test, Paired *t* test), One-way ANOVA, Pearson product moment correlation, Simple linear regression, Multiple linear regression. For example, Analysis of Variance (ANOVA) assumes that the underlying distributions are normally distributed and that the variances of the distributions being compared are similar. The Pearson product-moment correlation coefficient also assumes normality.

Although parametric techniques are robust, that is, they often retain considerable power to detect differences or similarities even when these assumptions are violated, some distributions violate the assumptions so markedly that a non-parametric alternative is more likely to detect a difference or similarity.[3] Hence, tests that do not make assumptions about the population distribution are referred to as nonparametric-tests**.**

Specifically, nonparametric methods were developed to be used in cases when the researcher knows nothing about the parameters of the variable of interest in the population (hence the name nonparametric). In more technical terms, nonparametric methods do not rely on the estimation of parameters (such as the mean or the standard deviation) describing the distribution of the variable of interest in the population. Therefore, these methods are also sometimes (and more appropriately) called parameter-free methods or distribution-free methods. All commonly used nonparametric tests rank the outcome variable from low to high and then analyze the ranks.

For many variables of interest, we simply do not know for sure that they are normally distributed. For example, we can not state categorically that incomes of Universities staff in Nigeria are distributed normally in the population. In addition, the incidence rates of rare diseases are not normally distributed in the population, the number of car accidents is also not normally distributed, and neither are very many other variables in which a researcher might

[3] The power of a statistical test is the probability that the test will reject a false null hypothesis.

be interested. Examples of nonparametric tests include Median, Interquartile range, Wilcoxon test, Mann-Whitney test, Kruskal-Wallis test, Friedman test for dependent samples, Chi square test, Spearman correlation, Coefficient of concordance.

Many observed variables actually are normally distributed, which is why the normal distribution represents an empirical discussion in the literature. The problem may occur when one tries to use a normal distribution-based test to analyze data from variables that are themselves not normally distributed. Another factor that often limits the applicability of tests based on the assumption that the sampling distribution is normal is the size of the sample of data available for the analysis (sample size; n). We can assume that the sampling distribution is normal even if we are not sure that the distribution of the variable in the population is normal, as long as our sample is large enough (for example, 100 or more observations). However, if our sample is very small, then those tests can be used only if we are sure that the variable is normally distributed, and there is no way to test this assumption if the sample is small.

Applications of tests that are based on the normality assumptions are further limited by a lack of precise measurement. For example, the assumption in most common statistical techniques such as Analysis of Variance (and *t*- tests), regression is that the underlying measurements are at least of interval, meaning that equally spaced intervals on the scale can be compared in a meaningful manner. However, this assumption is very often not tenable because the data may represent a rank ordering of observations (ordinal) rather than precise measurements. Choosing between parametric and nonparametric tests is sometimes easy. You should definitely choose a parametric test if you are sure that your data are sampled from a population that follows a normal distribution. Non-parametric tests could be selected under three conditions:

- The outcome is a rank or a score and the population is clearly not Gaussian. Examples include class ranking of students.
- Some values are off the scale, that is, too high or too low to measure. Even if the population is Gaussian, it is impossible to analyze such data with a parametric test since you do not know all of the values. Using a nonparametric test with these data is simple. Assign values too low to measure an arbitrary very low value and assign values too high to measure an arbitrary very high value. Then perform a nonparametric test. Since the nonparametric test only knows about the relative ranks of the values, it does not matter that you did not know all the values exactly.
- If the data are not sampled from normal distribution, we can consider transforming the values to make the distribution become normal. For example, you might take the logarithm or reciprocal of all values.

Despite the three cases stated above, it is not always easy to decide whether a sample comes from a normal distributed population. For example:

- If we collect many data points (maybe over a hundred or so), we can look at the distribution of the data and it will be fairly obvious whether the distribution is approximately bell shaped. With few data points, it is difficult to tell whether the data are Gaussian by inspection, and the formal test has little power to discriminate between Gaussian and non-Gaussian distributions.
- We can look at previous data as well. What is important is the distribution of the overall population, and not the distribution of our sample. In deciding whether a population is Gaussian, we should look at all available data, not just data in the current experiment.

When in doubt, some people choose a parametric test (because they are not sure the Gaussian assumption is violated), and others choose a nonparametric test (because they are also not sure

whether the Gaussian assumption is met). There are four cases to consider in the answer whether one chooses between a parametric or nonparametric test sample sizes:

1. The central limit theorem ensures that parametric tests work well with large samples even if the population is non-Gaussian. In other words, parametric tests are robust to deviations from Gaussian distributions, so long as the samples are large. The problem is that it is impossible to say how large is large enough, as it depends on the nature of the particular non-Gaussian distribution. Unless the population distribution is quite distorted, you are probably safe choosing a parametric test when there are at least two dozen data points in each group.
2. Nonparametric tests work well with large samples from Gaussian populations. The probability (P) values tend to be a bit too large, but the discrepancy is small. In other words, nonparametric tests are only slightly less powerful than parametric tests with large samples.
3. In a small sample situation, we can not rely on the central limit theorem if we use a parametric test with data from non-Gaussian populations because the P value may be inaccurate.
4. In addition, when we use a nonparametric test with data from a Gaussian population, the P values also tend to be quite high. The nonparametric tests lack statistical power with small samples.

Thus, large data sets present no problems. It is usually easy to tell if the data come from a Gaussian population, but it does not really matter because the nonparametric tests are so powerful and the parametric tests are so robust. However, small data sets present a dilemma. It is difficult to tell if the data come from a Gaussian population, but it matters a lot. The nonparametric tests are not powerful and the parametric tests are not robust.

Basic Statistical Techniques

A multitude of different statistical tools is available, some of them simple, some complicated, and often very specific for certain purposes. In analytical work, the most important common operation is the comparison of data, or sets of data, to quantify accuracy (bias) and precision. The value of statistics lies with organizing and simplifying data, to permit some objective estimate showing that an analysis is under control or that a change has occurred. Statistical techniques can be used to describe data, compare two or more data sets, determine if a relationship exists between variables, test hypotheses and make estimates about population measures. Some well known statistical tests and procedures for research observations are:

The t-test

The *t*-test is the most commonly used method to evaluate the differences in means between two groups. For example, the *t*-test can be used to test for a difference in test scores between a group of patients who were given a drug and a control group who received an injection. Theoretically, the t-test can be used even if the sample sizes are very small (e.g., as small as 10) as long as the variables are normally distributed within each group and the variation of scores in the two groups is not reliably different.

The *p*-level reported with a *t*-test represents the probability of error involved in accepting the research hypothesis about the existence of a difference. It is the probability of error associated with rejecting the hypothesis of no difference between the two categories of observations (corresponding to the groups) in the population when, in fact, the hypothesis is true. If the calculated p-value is below the threshold chosen for statistical significance (usually the 0.05 level), then the null hypothesis which usually states that the two groups do not differ is rejected in favor of an alternative hypothesis, which typically states that the groups do differ.

There are however specific assumptions underlying the use of the t-test:

1. The sample data should be normally distributed.
2. The sample must be representative of the population so that we can made generalizations at the end of the analysis.
3. Equality of variances. Equal variances are assumed when two independent samples are used to test a hypothesis.
4. The dependent measurements involved in the calculation of the means must come from either interval or ratio scales.

Since all calculations are carried out subject to the null hypothesis, it may be very difficult to come up with a reasonable null hypothesis that accounts for equal means in the presence of unequal variances. Consequently, the null hypothesis is that the different treatments have no effect which therefore makes unequal variances untenable. A fundamental issue in the use of the t-test is often whether the samples are independent or dependent. Independent samples typically consist of two groups with no relationship while dependent samples typically consist of a matched sample (alternatively a paired sample) or one group that has been tested twice (repeated measures).

Dependent *t*-tests are also used for matched-paired samples, where two groups are matched on a particular variable. For example, if we examined the heights of men and women in a relationship, the two groups are matched on relationship status. This would call for a dependent *t*-test because it is a paired sample (one man paired with one woman). Alternatively, we might recruit 100 men and 100 women, with no relationship between any particular man and any particular woman; in this case we would use an independent samples test. Another example of a matched sample would be to take two groups of students, match each student in one group with a student in the other group based on a continous assessment result, then examine how much each student reads. An example pair might be two students that score 70 and 71 or two students that scored 55 and 50 on the same continous assesment. The hypothesis would be that students that did well on

the test may or may not read more. Alternatively, we might recruit students with low scores and students with high scores in two groups and assess their reading amounts independently.

An example of a repeated measures *t*-test would be if one group were pre- and post-tested. For example, if a teacher wanted to examine the effect of a new set of textbooks on student achievement, he/she could test the class at the beginning of the year (pre-test) and at the end of the year (post-test). A dependent *t*-test would be used, treating the pre-test and post-test as matched variables (matched by student).

Table 1: T-Test for Comparison of Mean Score of LPT and CT groups

Treatment Groups	N	- x	SD	t-value	Sig.(t)
LPT	120	11.69	3.19	7.92	0.000
CT	120	10.27	3.37		

In all the testing of this nature, the probability value is very important when interpreting the print-out. If the probability is less than 0.05 (the probaibility for committing type 1 error), the result is significant and we reject the null hypothesis. The reverse is the case if the probability is greater than 0.05.

ANOVA/ANCOVA /MCA

The purpose of analysis of variance (ANOVA) is to test differences in means (for groups or variables) for statistical significance. This is accomplished by analyzing the variance, that is, by partitioning the total variance into the component that is due to true random error and the components that are due to differences between means. These latter variance components are then tested for statistical significance, and, if significant, we reject the null

hypothesis of no differences between means, and accept the alternative hypothesis that the means (in the population) are different from each other.

The variables that are measured (for example, a test score) are called dependent variables. The variables that are manipulated or controlled (for example, a teaching method or some other criterion used to divide observations into groups that are compared) are called factors or independent variables. Analysis of covariance (ANCOVA) is a general linear model with one continuous explanatory variable and one or more factors. ANCOVA is a merger of ANOVA and regression for continuous variables. ANCOVA tests whether certain factors have an effect after removing the variance for which quantitative predictors (covariates) account. If there is a significant difference among the groups, pairwise comparison is carried out. The pairwise tests that could be used in order to detect the pairs which are significantly different include Scheffe, Duncan or Turkey.

Multiple Classification Analysis (MCA), which is part of ANOVA or ANCOVA, is a technique for examining the interrelationship between several predictor variables and one dependent variable in the context of an additive model. The MCA can handle predictors with no better than nominal measurements and interrelationships of any form among the predictor variables or between a predictor and dependent variable. It is however essential that the dependent variable should be either an interval-scale variable without extreme skewness or a dichotomous variable with frequencies which are not extremely unequal. In addition, it gives the percentage contributions of treatment and other categorical variables to the variance of the dependent measure.

However, the overall percentage contribution of all the variables (composite) to the variance of the dependent measure could also be obtained from the Multiple R square, which is 0.207*100=20.7%.

Table 2: Summary of Analysis of Covariance of Subjects Post-test Achievement Scores in Narrative Composition

Source of Variation	Sum of Squares	Df	Mean square	F	Sig of F
Covariates PRE ECAT	591.274	1	591.274	33.019	000
Main Effects TRT	370.979	3	123.660	6.905	000*
Explained	962.273	4	240.568	13.434	000
Residual	3689.026	206	17.708		
Total	4651.229	210	22.143		

*Significant at $p<0.05$

Table 2 shows that there is significant main effects of treatment since the probability is less than 0.05. In Table 3, if Adjusted Deviation is added or subtracted to the Grand Mean, the result is the mean score for each of the treatment groups. For example, the mean score of WLBIS group is 9.42-0.09=9.33. Table 4 shows that significant differences existed between groups 1, 2 and 3 separately, and group 4.

Table 3: Multiple Classification Analysis (MCA) of Posttest Achievement Scores in Narrative Composition

Grand Mean=9.42

Variable Category	N	Unadjusted Deviation	ETA	Adjusted for Independents +Covariates Deviation	BETA
TRT					
1.WLBIS	55	-0.18		-.09	
2.PLEBIS	52	2.26		1.65	
3.WLPLEBIS	53	.45		.52	
4.CS	50	-2.68		-2.20	
			.37		.29
Multiple R Square				.207	
Multiple R				.445	

Table 4: Scheffe Post-Hoc Analysis of Posttest Means of Achievement in Narrative According to Treatment Groups

Achievement		GRP 4	GRP 3	GRP 2	GRP 1
Mean	Treatment				
7.22	GRP 4				
9.33	GRP 1	*			
9.94	GRP 3	*			
11.07	GRP 2	*	*		

(*) Indicates significant differences that are shown in the lower triangle

Multivariate Analysis of Variance (MANOVA)

The Multrivariate Analysis of Variance (MANOVA) is designed to test the significance of group differences. The only substantial difference between the two procedures is that MANOVA can include several dependent variable, wheras ANOVA can handle only one dependent variable. MANOVA is based on the following assumptions:

- The observations within each sample must be randomly sampled and must be independent of each other.
- The observations on all dependent variables follow a multivariate normal distribution in each group.
- The population covariance matrices for the dependent variables in each group must be equal (homogeneity of covariance matrices).
- The relationships among all pairs of dependent variables for each cell in the data matrix must be linear.

Example

A research is interested in examining whether age of some set of workers could affect their income and hours worked per week. The two dependent variable being income and hours worked per week.The three tables below show the Multivariate Analysis of Variance.

Table 5: Box's Test of Equality of Covariance Matrices

Box's M	6.936
F	0.766
δf_1	9
δf_2	2886561
sig	0.648

Table 6: Multivariate Tests for Income and Hours Worked by Age Category

Effect	value	F	Hypothesis δf	Error δf	Sig.	Eta-Squared
Intercept Pillai's Trace	0.957	7507.272	2.00	680.0	0.000	0.957
Wilk's Lamdba	0.043	7507.272	2.00	680.0	0.000	0.957
Hotelling's Trace	22.080	7507.272	2.00	680.0	0.000	0.957
Roy's Largest root	22.880	7507.272	2.00	680.0	0.000	0.957
Age Pillai's Trace	0.091	10.791	6.00	1362.00	0.000	0.045
Wilk's Lamdba	0.909	11.035	6.00	1360.00	0.000	0.046
Hotelling's Trace	0.100	11.279	6.00	1358.00	0.000	0.047
Roy's Largest root	0.099	22.457	3.00	681.00	0.000	0.090

Table 5 reveals that there is no significant diefference in the observed variance ($F_{(9,289)}$=0.766). Therefore Wilk's Lambda f ratio will be used (If there is a significant difference, Pillai's Trace F ratio will be used). Table 6 reveals that there is a significant difference among the age groups with respect to income income and hours worked per week. (F(6,1360) = P<0.05; Z^2 =0.046).

Table 7: Univariate ANOVA Summary Table

Source	Dependent Variable	Sum of Squares	df	Mean Square	F	Sig	Eta-Squared
Corrected Model	Income	1029.016	3	343.005	20.995	0.000	0.085
	HRS	64.281	3	21.427	0.167	0.919	0.001
Intercept	Income	128493.5	1	128493.5	7864.97	0.000	0.920
	HRS	1410954	1	1410954	10972.71	0.000	0.942
Age	Income	1029.016	3	343.005	20.995	0.000	0.085
	HRS	64.281	3	21.427	0.167	0.919	0.001
Error	Income	11125.807	681	16.337			
	HRS	87568.119	681	128.588			
Total	Income	149966.0	685				
	HRS	1575151	685				
Corrected Total	Income	12154.82	684				
	HRS	87632.4	684				

R^2 = 0.085 (Adjusted R^2 =0.081) ; R^2 = 0.001 (Adjusted R^2 =0.004)

Table 7 shows that there is significant difference among the age groups in income ($F_{(3, 681)}$ =20.995; P< 0.05; Z^2 =0.001). Post hoc analysis could be used to determine the source of significant difference. This is done in the same way it is done in ANOVA or ANCOVA.

Regression Analysis

Regression analysis is a statistical technique used for the modeling and analysis of numerical data consisting of values of a dependent variable (response variable) and of one or more independent variables (explanatory variables). The dependent variable in the regression equation is modeled as a function of the independent variables, corresponding parameters (constants), and an error term. The error term is treated as a random variable. It represents unexplained variation in the dependent variable. The parameters are estimated so as to give a best fit of the data. Most commonly the best fit is evaluated by using the least squares method, but other criteria have also been used. For some kinds of research questions, regression can be used to examine how much a particular set of predictors explains differences in some outcome. In other cases, regression is used to examine the effect of some specific factor while accounting for other factors that influence the outcome.

Regression analysis requires assumptions to be made regarding probability distribution of the errors. Statistical tests are made on the basis of these assumptions. In regression analysis the term model embraces both the function used to model the data and the assumptions concerning probability distributions. Regression can be used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships. These uses of regression rely heavily on the underlying assumptions being satisfied. Regression analysis has been criticized as being misused for these purposes in many cases where the appropriate assumptions cannot be verified to hold.

The set of underlying assumptions in regression analysis is that:

- The sample must be representative of the population for the inference prediction.
- The dependent variable is subject to error. This error is assumed to be a random variable, with a mean of zero.
- The independent variable is error-free.

- The predictors must be linearly independent, that is, it must not be possible to express any predictor as a linear combination of the others.
- The errors are uncorrelated, that is, the variance-covariance matrix of the errors is diagonal and each non-zero element is the variance of the error.
- The variance of the error is constant (homoscedasticity).
- The errors follow a normal distribution. If not, the generalized linear model should be used.

The violation of any of these implied conditions could have potentially negative effects for your research. In a simple linear regression, we examine a causal relationship between a dependent variable , Y, and one independent variable, X. Linear regression is synonymous with equation of a straight line. This is because it is important to make certain that a linear relationship exists between the factors before running a regression model. For example, we can express the relationship between the dependent variable *y* and the independent variable *x* through the following formula:

$$y = m_0 + m_1 x_1 \ldots \quad (1)$$

That is, equation (1) says that *y* can be expressed as a linear function of *x*. The constant for the model is represented by the parameter m_0. Regression is a tool that allows us to take data from some sample and use these data to estimate m_0 and m_1. These values are then used to create predicted values of the outcome, with the observed or true value from the data designated as y and the predicted value as $\hat{y}$. Furthermore, in equation (1), the value m_1 measures the causal effect of a one unit increase of X on the value of Y. The parameter m_1 is also referred to as the regression coefficient for X, and is the average amount the dependent variable increases when the independent variable increases one unit and other independents are held constant. Thus, when independent measure increases by 1, the dependent variable increases by m_1 units.

In multiple linear regression, there are more than one independent variable in the model. Multiple regression allows researchers to examine the effect of many different factors on some outcome at the same time. The general purpose of multiple regression is to learn more about the relationship between several independent or predictor variables and a dependent variable. This is because the simple linear regression is a simplification of reality. In real life, there are more than one variable that affects the behaviour of the dependent variable. It can be specified as follows:

$$Y = m_0 + m_1x_1 + m_2x_2 \ldots + m_nx_n \quad (2)$$

Where Y is the dependent variable; x_1 and x_2 are the independent variables; and c_0, c_1, c_2, are the parameters. The intercept of the regression is c_0 while c_1 and c_2 are referred to as the partial regression coefficients. The error term is represented by u_i. Hypothesis testing are conducted to show whether the parameters that have been estimated are statistically significant or, whether the independent variables contribute to the explanation of variation in the dependent variable. If we are able to reject the null hypothesis at an acceptable significance level, then we conclude that the parameter is not statistically significant. The quality of the fitness of the model is determined by the R^2. The R^2 has a value that is between 0 and 1. High values of R^2 will indicate that the model fits the data well. A limitation in the use of R^2 is that its value increases with the number of explanatory variables. It does not usually penalize for the consequence loss of degrees of freedom as the number of explanatory variables increases. The power of the test is therefore affected. Thus, the adjusted R-square was developed to take care of the inadequacies. In addition, the F-statistic test for the joint significance of all the parameters in the model.

The composite and relative contributions of independent variables to the dependent variable are usually determined through multiple regression. These are explained in Tables 8, 9 and 10 respectively.

Table 8: Composite Effect of Independent Variables (School Environment and Teacher Competency) on Dependent Variable (achievement in Integrated Science).

Multiple Correlation (R)	R Square	Adjusted R-Square	Standard Error of The Estimate	F	Sign F
0.962	0.926	0.923	4.2834	307.025	0.000*

Table 8 shows Multiple Correlation R, the square of this correlation (R^2), Standard Error and F value and the probability value (Sig. F). The variables under consideration correlated significantly with the dependent variable. R value is 0.926. If R is squared and the result is multiplied by 100, the percentage contribution of all the independent variables taken together to the variance of the dependent variable is obtained. In this case, R^2=0.923, R^2*100=92.3%.

Table 9: Analysis of Variance

Source of Variance	Sum of Squares	DF	Mean	F	Significance
Regression	39432.31	7	5633.188	307.025	0.000*
Residual	3155.80	172	18.348		
Total	42588.11	179			

*Significant of $p<0.05$.

Table 9 gives the Analysis of Variance. The F ratio here is different from those of ANOVA and ANCOVA. The value here is significant because the probability is less than 0.05. What this means is that the R value earlier obtained is not due to chance.

Table 10: Relative Contributions (Beta Weights) of the Seven Independent Variables to Students' Achievement in Integrated Science

Independent Variables (Predictors)	Unstandarized Coefficients		Beta Weight (β)	Rank	T	Significant
	B	Std. Error				
Constant	12.187	3.118			3.908	0.000
Special Learning Environment	0.687	0.061	0.548	1st	11.273	0.000*
Special Equipment	0.0097	0.122	0.020	5th	0.790	0.431
Speech Theraphy	2.742	0.301	0.435	2nd	9.104	0.000*
Knowledge of Science and Education	0.335	0.159	0.054	3rd	2.102	0.037
Knowledge of learners	0.0063	0.119	0.012	7th	0.539	0.591
Teachers' strategies	0.0027	0.041	0.015	6th	0.670	0.504
Counselling of parents	0.0033	0.029	0.025	4th	1.131	0.260

*Significant at $p<0.05$

Table 10 gives the relative effects of the independent variables on the dependent variable. The B is referred to as partial correlation, the Beta weight (β) is the weight contribution of each variable. We also have t values for all the variables and the probability values (sig.t). Note the ranking of the variables according to their weight contributions. You will notice that the constant is under B and not under Beta weight (β).

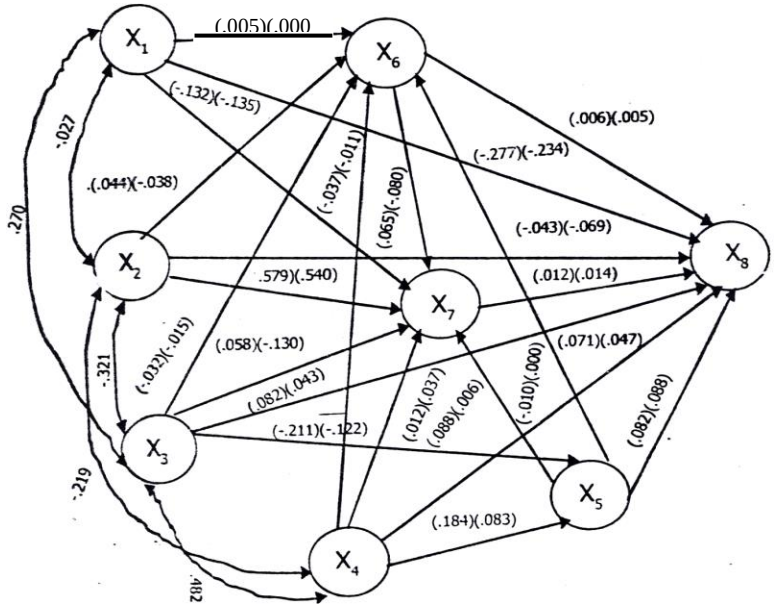


Fig.2: Hypothesized Recursive Model Showing Path Coefficient and Zero Order Correlations

Figure 2 is a follow-up to Figure 1 after trimming of the paths in Figure 1. The detail of the trimming could not be given but it involves removing any path whose path-coefficient is less than 0.05. The model in Figure 2 is usually referred to as parsimonious model. It shows the paths which have direct and indirect influence on the dependent variable.

Path Analysis

This is an extension of multiple regression analysis. The use of path analysis enables the researcher to calculate the direct and indirect influence of independent variables on a dependent variable. These influences are reflected on the path coefficients, which are actually standardized regression coefficients (Beta weights). Path analysis is one of the techniques for the study and analysis of causal relations in ex-post facto research. It usually

starts with hypothesized model and end up with parsimonous model. This is after carrying out trimming of the paths by using structural equations. In the hypothesized model belo, variables X_1, X_2, X_3, X_4, X_5, X_6, and X_7 re called independent variables or exogenous variables while variable X_8 is referred to as dependent variable or endogenous variable. The model is hypothetical.

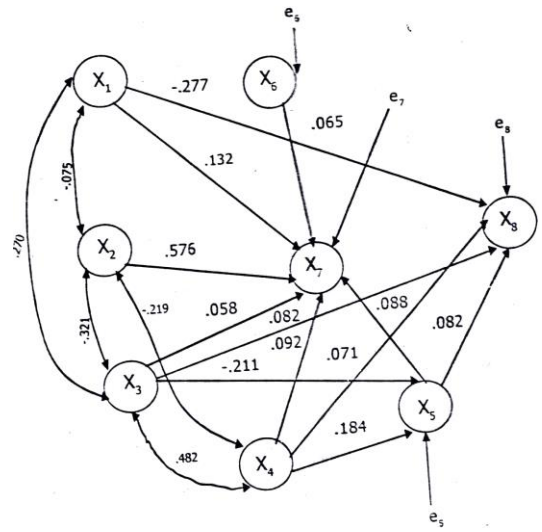


Fig. 3: The Parsimonious Model

Factor Analysis

Factor analysis is useful in reducing a mass of information to a manageable and economical description. For example, data on fifty characteristics for 300 states are unwiedly to handle, descriptively or analytically. Reducing them to their common factor patterns facilitates the management, analysis and understanding of such data. These factors concentrate and index characteristics without much loss of information. States can be more easily discussed and compared on economic, development, size and public dimensions other than on the hundreds of characteristics each dimension involves.

It should be noted that standardizing one's variables before applying factor analysis is not necessary because result of factor analysis are not affected by the standardization, which is built into the procedure.

A Hypothetical Situation

Suppose a 25 item questionnaire on students' attitude towards Physics was administered to 40 students. Factor analysis could be carried out to find out the commonalities of the test items such that the 25 items would be reduced to a fewer number of items and the instrument would still be able to measure validly and reliably the construct attitudes towards Physics. Also, the scale items could be sorted into their various components so that the items, which correlate highly with themselves are group together.

Table 11 shows the initial eigen values which provide information on the % variance explained by each of the variables. It could be observed that out of the 25 items, the first 9 items account for 76.75 of the total variance. The 25 items have been reduced to 9 and the 9 items could be assumed to have measured the construct, which the 25 items were designed to measure. This shows that since the 9 items were found to account for 76.75 of the total variance, if items 10-25 are explained, no serious harm would be done to the scale of measurement.

The analysis was carried out to establish the number of meaningful factors. Nine factors have thus been found to be meaningful or nontrivial. These are the factors considered as peculiar factors perceived by the students as their attitudes toward Physics.

Table 11: Total Variance Explained Students Attitudes Towards Physics

Items	Initial Eigen Values			Extraction Sum of Squared Loadings		
	Total	% of Variance	Cumulative	Total	% of Variance	Cumulative
1	4.179	16.718	16.718	4.179	16.718	16.718
2	3.745	14.991	31.748	3.748	14.991	31.709
3	2.662	10.647	42.256	2.662	10.647	42.356
4	1.927	7.709	50.065	1.927	7.709	50.065
5	1.703	6.813	56.878	1.703	6.813	56.878
6	1.458	5.834	62.711	1.458	5.834	62.711
7	1.388	5.552	68.264	1.388	5.552	68.264
8	1.107	4.427	72.690	1.107	4.427	72.690
9	1.017	4.069	76.759	1.017	4.069	76.759
10	0.931	2.724	80.483			
11	0.826	2.306	83.789			
12	0.686	2.746	86.535			
13	0.653	2.613	89.148			
14	0.536	2.144	91.292			
15	0.443	1.771	93.062			
16	0.381	1.525	94.588			
17	0.364	1.456	96.043			
18	0.329	1.317	97.360			
19	0.208	0.830	98.190			
20	0.151	0.604	98.794			
21	0.105	0.422	99.216			
22	0.0830	0.333	99.549			
23	0.05917	0.237	99.786			
24	0.03782	0.151	99.937			
25	0.01579	0.06316	100.000			

Extraction Method: principal Component Analysis

Names are usually given to the isolated factors and different ietms are loaded on each factors.

Correlations

Correlation is a measure of the relation between two or more variables. It indicates the strength and direction of a linear relationship between two random variables. The correlation is 1 in

the case of an increasing linear relationship, −1 in the case of a decreasing linear relationship, and some value in between in all other cases, indicating the degree of linear dependence between the variables. The closer the coefficient is to either −1 or 1, the stronger the correlation between the variables. If the variables are independent then the correlation is 0, but the converse is not true because the correlation coefficient detects only linear dependencies between two variables.

Pearson Product-Moment Correlation Coefficient

The most widely-used type of correlation coefficient is Pearson product-moment correlation coefficient. Pearson's correlation coefficient is a parametric statistic. It is a common measure of of the correlation between two variables *X* and *Y*. When measured in a population the Pearson Product Moment correlation is designated by rho (ρ). When computed in a sample, it is designated by the letter *r* and is sometimes called *Pearson's r*. Pearson's correlation reflects the degree of linear relationship between two variables. Pearson correlation, assumes that the two variables are measured on at least interval scales, and it determines the extent to which values of the two variables are proportional to each other. However, the value of correlation (that is, correlation coefficient) does not depend on the specific measurement units used. For example, the correlation between height and weight will be identical regardless of whether inches and pounds, or centimeters and kilograms are used as measurement units.

Spearman's rank correlation coefficient

The Spearman's rank correlation coefficient is a non-parametric measure of correlation – that is, it assesses how well an arbitrary monotonic function could describe the relationship between two variables, without making any assumptions about the frequency distribution of the variables.

Spearman's rank correlation coefficient does not require the assumption that the relationship between the variables is linear, nor does it require the variables to be measured on interval scales; it

can be used for variables measured at the ordinal level unlike the Pearson product-moment correlation coefficient. However, Spearman's correlation coefficient does assume that subsequent ranks indicate equidistant positions on the variable measured.

Choosing Appropriate Statistical Technique in a Research Enterprise

Statistical techniques can be used to describe data, compare two or more data sets, determine if a relationship exists between variables, test hypothesis and make estimates about population measures. Not only it is important to have a sample size that is large enough, but also it is necessary to see how the subjects in the sample were selected. Volunteers generally do not represent the population at large.

However, student should realize that computer merely give numerical answer and save time and effort of doing calculations by hand. It is the duty of students to understand and interprets computer print-out correctly. Note that data can be subjected to parametric and nonparametric statistics depending on the nature of the data. Purely numerical data like student score in a chemistry test could be subjected to a parametric test. Note that this score has a zero origin like a score of 60% for example. On the other hand, the number of people who read newspapers in the morning is not a purely numerical data and it can only be subjected to a nonparametric test. In this case, you can not perform all the mathematical operations with the data. You can add, subtract but you can not determine mean score neither can you carry out division.

The variable type determines to some extent the type of statistical (descriptive or inferential) method that it will support. For example, while trichotomous and polychotomous variables allow for the use of ANOVA/ANCOVA statistics, the dichotomous and monochotomous allow for t-test analysis. Variable types also influences the language of the hypotheses and hence, inferences that can be made from such hypotheses.

In any study, you describe, compare data or determine if relationship exists between variables. If the percentage score of group A is higher than that of group B, you are describing the performance of the groups. When you go a step further by finding out whether the performance are significantly different you are at the realm of making inference and this informed the usage of the parametric statistics; in this case, t-test. When we compare more than two mean scores, determination of F ratio is involved; in this case the statistical tool could be ANOVA or ANCOVA depending on the design employed for the study.

If a study involves determination of relationship, we can use Spearman Rank order correlation, Pearson Product moment correlation, Chi-square statistics or even multiple regression analysis. All depends on the nature of the research. Chi-square shows the degree of association between two different bases of classification. It should be noted that the z-test is used only when the population parameters are known and the variable of interest is normally distributed in the parent population. If the two conditions are met, the z-test is used as an exact test, even for small samples ($n<30$). However, if the variable is not normally distributed, a large sample permits the use of a z-test. In this case, the use of z-test is regarded as an approximate test. In most research situations, the z-test for single mean is rarely encountered because the conditions of normality and known parameters (σ) are rarely met. However z-test could be used for two means from independent samples. Normally, z-test is used to test for the mean of a large sample, and the t-test used for the mean of small sample.

Examples of Topics, Research Questions, Hypotheses and Selection of Appropriate Statistical Tools

As pointed out earlier, the selection of statistical tool in a study is a function of the design of the experiment. Also, the language of the hypotheses is determined by the design. However, a researcher need not shy away from the fact that the topic of a research is an offshoot of the statement of the problem. How then shall we determine the statistics to apply in the analysis of data based on the

language of the topics, design, and the hypotheses? The following examples would throw more light on the answer to the question.

1. **Topic:** Student, teacher and school environment factors as determinants of achievement in Senior Secondary School Chemistry in Oyo State, Nigeria.

Questions

i. What is the most meaningful causal model for students' achievement in secondary school Chemistry?

ii. To what extent will the seven independent variables when taken together, predict students' achievement?

Interpretation

The key word in the topic is 'determinant'. This is a situation where a variable (independent) could determine another variable (dependent), a kind of causal relationship. The independent variables here are: Student, teacher and school environment variables while the dependent variable is achievement in Senior Secondary school Chemistry. It would be observed that hypothesis is not necessary here. This is an ex-post facto study of the survey type in which the researcher need not manipulate the independent variables, they have already manifested.

The first question demands that the experimenter would construct hypothesized causal model, which he has to trim in order to get the parsimonoius model. The second question has to do with the determination of the composite effect of the independent variables on the dependent variable. The answer to this question could be obtained through multiple regression analysis. Path analysis is an extension of multiple regression. Path coefficients are the beta-weights in multiple regression.

2. **Topic:** Effects of guided discovery and self-learning strategies on secondary school students' learning outcome in Chemistry.

Hypotheses

1. There is no significant main effect of treatment on
   i. Learning outcome in Chemistry
   ii. Attitude to Chemistry
2. There is no significant main effect of ability on:
   i. Learning outcome in Chemistry
   ii. Attitude to Chemistry
3. There is no significant main effect of gender on
   i. Learning outcome in Chemistry
   ii. Attitude to Chemistry
4. There is no significant interaction effect of treatment and gender on
   i. Learning outcome in Chemistry
   ii. Attitude to Chemistry

Interpretation

The topic shows effects of two methods of instruction (independent) variables on two dependent variables (learning outcome in Chemistry and attitude to Chemistry). There should be a control group. The design then becomes: Pretest, Posttest, control group experimental design. Therefore, the appropriate statistical tool is Analysis of Covariance (ANCOVA) with pretest scores as covarite. Two categorical varaibles are being investigated. These are gender (dichotomous) and academic ability level (trichotomous). And because of the interaction hypotheses, there must be factorial design. Here it is 3 x 2 x`3, which is interpreted as follows: treatment at 3 levels, gender at 2 levels nd academic ability at 3 levels.

3. **Topic**: A comparative analysis of leadership styles of male and female managers in the banking industry in southwestern Nigeria.

Hypothesis

There is no significant relationship in the leadership styles of male and female managers in the banking industry.

Interpretation

Other hypotheses could be based on other vaiables investigated in the study. Here, the nature of the data generated cannot be purely numerical. Therefore, non-parametric statistics is the best bet. The candiadte could use Chi-square statistic for analyzing te data collected.

4. Topic: An evaluation of extra-mural studies programmes of the University of Ibadan 1989/90 and 1998/99 sessions.

Hypotheses:

1. The extra-mural studies significantly influenced achievement of self-actualization and self-esteemvalues of candidates.
2. There is no significant difference in the academic achievement of candidates in the 1989/90 and 1998/99 sessions.
3. Ther is no significant difference in the cademic achievement of candidates from three different locations in the 1989/90 session.

Interpretation

It could be observed that there are different languages in the three hypotheses. Hypothesis 1 could be tested using Chi-Square statistic because of the word “influence”. Hypothesis 2 and 3 has to do with purely numerical data (academic achievements). While hypothesis 2 is a test of significnce between two sets of candidates from two sessions, hypothesis 3 is a test of significance among candidates from three different locations. In this case, hypothesis 2 will be tested using t-test while hypothesis 3 will be tested uning Analysis of Varianace (ANOVA). Note that the word “Evaluation” in the topic can accommodate virtually all statistical procedures; it all depends on the nature of the hypothesis.

5. **Topic**: Impact of government labour integration policy on Nigerian dockworkers' productivity in Nigerian Ports Authority, Nigeria.

Hypotheses:

1. Ther is no significant impact of government labour integration policy on Nigerian workers' productivity in public and private sectors of the economy.
2. There is no significant impact of government labour integration policy on Nigerian male and female workers' productivity.
3. There is no significant impact of government labour integration policy on three different categories of workers' productivity.

Interpretation

It could be observed that this could also be a cause and effect study. The word "impact" in the topic is synonymous with "effect". Thus, the study demands the usage of t-test and Analysis of Variance (ANOVA) depending on the nature of the hypotheses. The study is ex-post facto but of the experimental type. Hypothesis 1refers to public and private sectors therefore, t-test will be appropriate. Hypothesis 2 will also be tested with t-test beause it is between male and female workers while hypothesis 3 can only be tested with Analsysi of Variance (ANOVA) because three categories of workers are involved.

6. Topic: Relationship between some school factors and Secondary School system efficiency in Ogun State, Nigeria.

Hypotheses

1. There is no significant relationship between the school related factors and school system efficiency.
2. There is no significant relationship between school location and school system efficiency.

Other hypotheses could be generated based on other school-related factors.

Interpretation

This is a descriptive survey research and based on the languages of the topic and hypotheses, it is a relational study. The hypotheses could be tested by Pearson Product Moment Correlation or Chi-square statistic. If ranking is involved the candidate could still make use of Spearman rank order correlation.

Data Analysis and Interpretation

Unprocessed data are called raw data. Data have to be processed by making use of computers to for analysis. This is because manual procedures for estimating and computing relevant statistics have become increasingly tedious or entirely impossible. Computers are now applied in all aspects of statistical analyses, from the calculation of simple sums to the estimation of large scale stochastic models.

There are numerous statistical programs for analyzing data. These are known as package or canned programs. The popular programs in this class include SPSS, SAS, GAUSS, E-Views, RATS, LIMDEP, and STATA. However, the most popular series for the educational researcher is the Statistical Package for Social Sciences (SPSS). SPSS contains many of the most common statistical procedures needed by the students.

References

Adesoji, F.A. (2006). Statistical Methods for Data Analysis and Data Interpretation In Alegbeleye, G.O, Mabawonku, I and Fabunmi, M (eds.) *Research Methods in Education*, University of Ibadan, Ibadan.

Bluman, A.G. (1990) *Elementary Statistics.* McGraw Hill, Higher Education, New York.

Gbadegesin, A., R. Olopoenia, and A. Jerome (2005) "Statistics for the Social Sciences". Ibadan University Press.

Gujarati, D. N (1995) Basic Econometrics. 3rd Edition. McGraw Hill, New York.