INTERNATIONAL SCHOLARS JOURNAL OF ARTS AND SOCIAL SCIENCE RESEARCH
(ISJASSR)
Website: www.theinterscholar.org/journals/index.php/isjassr
E-mail: editorialoffice@theinterscholar.org

Volume 3, Issue 4, May, 2021 ISSN: 2507 -1528

Application of Statistical Tools for Data Analysis: Relationship Techniques

By

Ugwu Paschal Anayochukwu, PhD Candidate
Department of Marketing,
Nnamdi Azikiwe University, Awka ,
Anambra State, Nigeria.

Ambakederemo, Owulupu Samuel
Department of Maritime Transport and Business Studies,
DeltaState School of Marine Technology, Burutu

Opuware, Kenny Alubeze
Department of Maritime Transport and Business Studies,
Delta State School of Marine Technology, Burutu

Abstract: *Statistical analysis is divided into descriptive and inferential statistical analyses. Whereas descriptive statistics is interested in data description, the interest of inferential statistics is making inferences and extrapolations from data. In application of statistical techniques, one needs to determine the type of relationship that exists. If the data variables can be divided into dependent and independent classification, indicating whether dependence or interdependence techniques should be utilized. Dependent relationship is when variables are divided into dependent and independent variables, where dependent variables depend on independent variables. In interdependence relationship, variables are not divided into dependent and independent, rather they are analyzed simultaneously to establish their underlying relationships. Canonical correlation, Multivariate analysis of variance, Analysis of variance, Multiple discriminant analysis, Multiple regression, Conjoint analysis are all dependent techniques. Factor analysis, Cluster analysis, Correspondence analysis are interdependent techniques. SEM can be thought of being both dependent and interdependent techniques, mostly because its foundation lies in two familiar multivariate techniques: factor analysis and multiple regression analysis, and its ability to analyze multiple relationships simultaneously.*

Key words: Descriptive Statistics, Inferential Statistics, Dependent Relationship, Interdependent Relationship, Correlation, Regression, Structural Equation Modeling(SEM).

Introduction

Statistical analysis is broadly divided into two. Descriptive and Inferential statistical analysis**.** The major concern of descriptive statistics is to present information in a convenient, usable, and understandable form. For example, once data have been collected, the first things that a researcher would want to do is to calculate their frequency, to graph them, to calculate the

measures of central tendency (means, median, and mode), to calculate the dispersion of the scores (variances and standard deviations), and to identify outliers in the distribution of the scores. These procedures are called descriptive statistics because they are aimed primarily at describing the data. Inferential statistics, on the other hand, is not concerned with just describing the obtained data. Rather, it addresses the problem of making broader generalizations or inferences from sample data to population. Descriptive statistics is used to describe a set of data in terms of its frequency of occurrence, its central tendency, and its dispersion.

Although the description of data is important and fundamental to any analysis, it is not sufficient to answer many of the most interesting problems that researchers encounter. Consider an experiment in which a researcher is interested in finding whether a particular drug can improve people's memory.

The researcher offers the drug to one group but not to the control group, and then compares the means of the two groups on a memory test. Descriptive statistics will not tell the researcher, for example, whether the difference between a sample mean and a hypothetical population mean, or the difference between two obtained sample means, is small enough to be explained by chance alone or whether it represents a true difference that might be attributable to the effect of the experimental treatment, i.e., the drug. To address these issues, the researcher must move beyond descriptive statistics and into the realm of inferential statistics, and, particularly, on to the statistical procedures that can be employed to arrive at conclusions extending beyond the sample statistics themselves. A basic aim of inferential statistics then is to use the sample scores for hypothesis testing.

According to Taylor and Francis (2006), in choosing an appropriate statistical test, the first issue that the researcher must consider is the nature of the hypothesis. Is the intention of the hypothesis to test for differences in mean scores between groups, or is it testing for relationships between pairs of variables? Testing for differences means that the researcher is interested in determining whether differences in mean scores between groups are due to chance factors or to real differences between the groups as a result of the study's experimental treatment.

Testing of relationships among two or more variables involves asking the question, "Are variations in variable X associated with variations in variable Y?" For example, do students who do well in high school also perform well in university? Do parents with high intelligence tend to have children of high intelligence? Is there a relationship between cost of advertising and sales volume? Is there a relationship between competition and product life cycle? All these questions concern the relationships among variables, and to answer these questions, researchers must choose statistical tests that will appropriately test for the relationships among these variables.

According to Joseph Hair, William Black, Barry Babin and Rolph Anderson (2010), when considering the application of statistical techniques, one needs to determine the type of relationship that exists. Can the data variables be divided into dependent and independent classification, indicating whether dependence or interdependence techniques should be utilized?

Dependence Techniques:

The different dependence techniques can be categorized by two features:

a. The number of dependent variables
b. The type of measurement scale employed by the variables.

Regarding the number of dependent variables, dependence techniques can be classified as those having a single dependent variable, several dependent variables or even several dependent/independent relationships.
Dependence techniques can be further classified as those with metric (quantitative/numerical) or nonmetric (qualitative/categorical) dependence variables.

If the analysis involves a single dependent variable that is metric, multiple regression technique is appropriate, conjoint technique is also appropriate. Conjoint analysis is a dependence procedure that may treat the dependence variable as either metric or nonmetric, depending on the type of data collected. In contrast, if the single dependent variable is nonmetric, multiple discriminant analysis and linear probability model are the appropriate tools.

Multiple discriminant analysis is appropriate when the nonmetric dependent variable is dichotomous(male-female) or multichomous (high-medium-low). It is applicable in situation in which the total sample can be divided into group based on a nonmetric dependent variable characterizing several known class. The primary objectives of multiple discriminant analysis is to understand group differences and to predict the likelihood that any entity (object or individual) will belong to a particular class or group based on several metric independent variables.

A **linear probability model** (LPM) is a regression **model** where the outcome variable is a binary variable, and one or more explanatory variables are used to predict the outcome. Explanatory variables can themselves be binary, or be continuous.

When the research problem involves several dependent variables:

If the several dependent variables are metric, we must look to the independent variables. If the independent variables are nonmetric, Multivariate analysis of variance (MANOVA) is appropriate. MANOVA is used to simultaneously explore the relationship between several categorical independent variables (treatments) and two or more metric dependent variables. ANOVA is when one dependent metric variable is involved, with two or more independent variables.

Multivariate analysis of covariance (MANCOVA) can be used in conjunction with MANOVA to remove (after treatment) the effect of any uncontrolled metric independent variables (covariates) on the dependent variables. It is similar to bivariate partial correlation in which the effect of the third variable is removed from the correlation.
MANOVA is useful when the researcher designs experimental situation (manipulation of several nonmetric treatment variables) to test hypotheses concerning the variance in group responses on two or more metric variables.

If the independent variables are metric or if the several dependent variables are nonmetric and can be transformed through dummy variable coding (0-1), canonical correlation analysis can be used.

Canonical correlation is a logical extension of multiple regression analysis; the objective is to simultaneously correlate several dependent and several independent variables. Multiple regression involves a single metric dependent variable but canonical correlation involves many dependent variable. The underlying principle is to develop a linear combination of each set of variables (both dependent and independent variables) in a manner that maximizes the correlation between the two sets, that is obtaining a set of weights for dependent and independent variables that provides the maximum simple correlation between the set of dependent variable and set of independent variable.

Interdependence Techniques/Relationships:

Here, variables are not divided into dependent and independent, rather they are analyzed simultaneously to establish their underlying relationships. The question here is "is the structure of relationships among variable?" If the structure of the relationship is among variables, **factor analysis** is the appropriate technique. If the structure of the relationship is among case/respondents, **cluster analysis** can go. But if the structure of relationship is among objects, the question is "how are the attributes measures?" **Multi dimensional scaling** and **correspondence analysis** are the appropriate techniques to use if the attributes are measured **metrically** and **nonmetrically** respectively.

Structural Equation modeling (SEM) is appropriate when the research problem involves multiple relationships of dependent and independent variables. SEM provides opportunity to measure multiple relationships at the same time. It has the ability to assess the contribution of each indicator or observed variable in representing its associated construct and measure how well the combined set of indicator variables represent the construct, while accounting for measurement errors.

Table 1.1: The Relationship Between Multivariate Dependent Methods

Canonical Correlation

$Y_1 + Y_2 + Y_3 + ... + Y_n$	=	$X_1 + X_2 + X_3 + \ldots + X_n$
(metric, nonmetric)		(metric, nonmetric)

Multivariate Analysis of Variance (MANOVA)

$Y_1 + Y_2 + Y_3 + \ldots + Y_n$	=	$X_1 + X_2 + X_3 + \ldots + X_n$
(metric)		(nonmetric)

Analysis of variance (ANOVA)

Y_1	=	$X_1 + X_2 + X_3 + \ldots + X_n$
(metric)		(metric, nonmetric)

Multiple Discriminant Analysis

Y_1	=	$X_1 + X_2 + X_3 + \ldots + X_n$
(nonmetric)		(metric)

Multiple Regression Analysis

Y_1	=	$X_1 + X_2 + X_3 + \ldots + X_n$
(metric)		(metric, nonmetric)

Conjoint Analysis

Y_1	=	$X_1 + X_2 + X_3 + \ldots + X_n$
(metric, nonmetric)		(nonmetric)

Structural Equation Modeling (SEM)

Y_1	=	$X_{11} + X_{12} + X_{13} + \ldots + X_{1n}$
Y_2	=	$X_{21} + X_{22} + X_{23} + \ldots + X_{2n}$
Y_m	=	$Xm_1 + X_{m2} + X_{m3} + \ldots + X_{mn}$
(metric)		(metric, nonmetric)

Source : Hair et al (2010)

Example of simple correlation and Simple regression:

Correlation is primarily concerned with finding out whether a relationship exists and determining its magnitude and direction. When two variables vary together, such as loneliness and depression, they are said to be **correlated** (Taylor and Francis, 2010). Accordingly, correlational studies are attempts to find the extent to which two or more variables are related. Typically, in a correlational study, no variables are manipulated as in an experiment — the researcher measures naturally occurring events, behaviors, or personality characteristics and then determines if the measured scores covary. The simplest correlational study involves obtaining a pair of observations or measures on two different variables from a number of individuals. The paired measures are then statistically analyzed to determine if any relationship exists between them. For example, marketing researchers have explored the relationship between variables such as advertising cost and sales volume, competition and product life cycle, market research and customers' satisfaction. To quantitatively express the extent to which two variables are related, it is necessary to calculate a **correlation coefficient**. There are many types of correlation coefficients, and the decision of which one to employ with a specific set of data depends on the following factors:

- The level of measurement on which each variable is measured
- The nature of the underlying distribution (continuous or discrete)
- The characteristics of the distribution of the scores (linear or nonlinear)

The two most used are **Pearson product moment correlation coefficient**(**r**), employed with interval or ratio scaled variables, and the **Spearman rank order correlation coefficient**(**r**rho),

employed with ordered or ranked data. It is important to note that, regardless of which correlational technique the researcher uses, they all have the following characteristics
in common:
1. Two sets of measurements are obtained on the same individuals or on pairs of individuals who are matched on some basis.
2. The values of the correlation coefficients vary between +1.00 and –1.00. Both of these extremes represent perfect relationships between the variables, and 0.00 represents the absence of a relationship.
3. A **positive relationship** means that individuals obtaining high scores on one variable tend to obtain high scores on a second variable. The converse is also true, i.e., individuals scoring low on one variable tend to score low on a second variable.
4. A **negative relationship** means that individuals scoring low on one variable tend to score high on a second variable. Conversely, individuals scoring high on one variable tend to score low on a second variable.

Assumptions

• For each subject in the study, there must be related pairs of scores,i.e., if a subject has a score on variable X, then the same subject must also have a score on variable Y.
• The relationship between the two variables must be linear, i.e., the relationship can be most accurately represented by a straight line.
• The variables should be measured at least at the ordinal level.
• The variability of scores on the Y variable should remain constant at all values of the X variable. This assumption is called **homoscedasticity or Homogeneity of variance.**

Table 1.1: Correlationbetween ice-cream consumption and drowning.

		Ice-cream con.	Drowning
Ice-cream con.	Pearson Correlation	1	.652**
	Sig. (2-tailed)		.000
	N	256	256
Drowning	Pearson Correlation	.652**	1
	Sig. (2-tailed)	.000	
	N	256	256

**. Correlation is significant at the 0.01 level (2-tailed).

Source: SPSS output, version 20

Based on the table 4.9 above, there is strong positive relationship between ice-cream consumption and drowning. The correlation coefficient is .652. According to Evans (1996), the strength of correlation between variable can be determined with the following guides:
- .00-.19 “very weak”
- .20-.39 “weak”
- .40-.59 “moderate”
- .60-.79 “strong”
- .80-1.0 “very strong”

Based on information on table 1.1, the strong positive relationship between ice-cream consumption and drowning is significance because its probability value (.000) is less than the alpha value (0.01) therefore we accept the alternative hypothesis that there is significant relationship between ice-cream consumption and drowning.

It should be noted that the relationship between ice-cream consumption and drowning does not in any way suggest that icecream consumption causes drowning or vice versa. It only suggest that they covary. It can be said that increase in heat leads to increase in icecream consumption, and also increase in swimming, which may cause drowning.

For causality to be established, the following conditions must be present:

- **Covariation:** Because causality means that a change in a cause brings about a corresponding change in an effect, systematic covariance (correlation) between the cause and effect is necessary, but not sufficient to establish causality.
- **Sequence:** A second requirement for causation is the temporal sequence of events. It simply means that the cause must occur before the effect.
- **Nonspurious covariance:** Nonspurious association must exist between the cause and effect. A spurious association is one that is false or misleading.
- **Theoretical support:** The final condition for causality is theoretical support, or a compelling rationale to support a cause and effect relationship. A model can demonstrate relationships between any constructs that correlate between one another. But unless theory can be used to establish a causal ordering and rationale the observed covariance, the relationship remains simple association and should not be attributed with any further causal power (Hair et al, 2010).

These conditions are mostly demonstrated in regression analysis.

Simple Regression

Regression and correlation are closely related. Both techniques involve the relationship between two variables, and they both utilize the same set of paired scores taken from the same subjects. However, whereas correlation is concerned with the magnitude and direction of the relationship, regression focuses on using the relationship for prediction. In terms of prediction, if two variables were correlated perfectly, then knowing the value of one score permits a perfect prediction of the score on the second variable. Generally, whenever two variables are significantly correlated, the researcher may use the score on one variable to predict the score on the second. There are many reasons why researchers want to predict one variable from another. For example, knowing a person's I.Q., what can we say about this person's prospects of successfully completing a university course? Knowing a person's prior voting record, can we make any informed guesses concerning his vote in the coming election? Knowing his mathematics aptitude score, can we estimate the quality of his performance in a course in statistics? Or will the knowledge of advertising cost aid the estimate of sales volume? These questions involve predictions from one variable to another, and psychologists, educators, biologists, sociologists, and economists are constantly being called upon to perform this function.

Assumptions

• For each subject in the study, there must be related pairs of scores. That is, if a subject has a score on variable X, then the same subject must also have a score on variable Y.
• The relationship between the two variables must be linear, i.e., the relationship can be most accurately represented by a straight line.
• The variables should be measured at least at the ordinal level.
• The variability of scores on the Y variable should remain constant at all values of the X variable. This assumption is called **Homoscedasticity or Homogeneity of variance.**

Table 1.2: Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.748[a]	.560	.55	1.152

a. Predictors: (Constant), Advertising

Source: SPSS output, version 20.

Table 1.3: ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2.306	1	9.632	11.103	.000[b]
	Residual	495.412	373	.868		
	Total	497.717	374			

a. Dependent Variable: Sales volume

b. Predictors: (Constant), Advertising

Source: SPSS output, version 20.

Table 1.4: Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.316	.158		14.612	.000
	Advertising	.560	.057	.560	1.318	.000

a. Dependent Variable: Sales volume

Source: SPSS output, version 20.

Table 1.2 above shows the model summary results which sought to establish the explanatory power of the independent variable (Advertising) for explaining and predicting the dependent variable (Sales volume).

"R" is the multiple correlation coefficient, (i.e the linear correlation between the observed and model predicted values of the dependent variable), a value of .748 indicates a strong positive correlation.

R- square: is the coefficient of determination (i.e the squared value of the multiple correlation coefficient). It means that 0.560 (56.0%) variation in the dependent variable (Sales volume) is accounted for by the independent variable (Advertising) guiding the study. While the remaining 44% change in sales volume is explained by other factors other than advertising.

Adjusted R-square: is the improvement in R- square. It is an adjustment of the R-squared that controls the addition of extraneous predictors to regression model. The value of the Adjusted R is 0.55. It means that precisely 55% of the variations in sales volume is accounted for by advertising, after the coefficient of determination(R-square) has been adjusted to be sensitive to the number of included variables (predicting variables or predictors) and insensitive to extraneous or exogenous variables.

The Anova table 1.3 above tests the overall validity of the model. F-statistic and p-value were associated. The f- statistic is mean square (Regression) divided by the mean square (Residual): 9.632/0.868 =11.103. The p-value (F- Significance) is compared to some alpha level in testing the null hypothesis that the model coefficient is zero. The p- value (.000) is smaller than 0.05 (alpha value). This means that the explanatory variable is significant, and therefore the validity of the model.

The coefficient of f-statistic (11.103) is significantly different from zero (0) because its p-value is 0.000, which is lesser than 0.05. This can be interpreted thus: $F=11.103$, $P = .000 < .05$, we accept the alternative hypothesis that the influence of the tested independent variables is significant, and therefore cannot be ignored in explaining variations in sales volume.

In table 1.4 above, the "**B**" coefficient is the value for the regression equation for predicting the dependent variable from the independent variable.

The regression equation is stated thus: Estimated Sales volume $SV= \alpha + b1ADV$. Where SV= Sales volume, α = constant, ADV = Advertising. The b1 is the regression coefficients, which indicate the amount of change in Sales volume given a unit change advertising (Predictor).

Advertising - The coefficient for Advertising is .560. So based on the findings of this study, for every unit change in Advertising, a .56% change in sales volume is expected.

Multiple regression is the extension of simple regression that is when there are more than one independent variables guiding the study. It is a statistical technique through which one can analyze the relationship between a dependent or criterion variable and a set of independent or predictor variables.

Conclusion

Statistical analysis is broadly divided into two. **Descriptive and Inferential statistical analyses.** Whereas the concern of descriptive statistics is description of data, inferential statistics is more concern in making inferences and extrapolations from data. In application of statistical

techniques, one needs to determine the type of relationship that exists. If the data variables can be divided into dependent and independent classification, indicating whether dependence or interdependence techniques should be utilized. Dependent relationship is when variables are divided into dependent and independent variables, where dependent variables depend on independent variables. In interdependence relationship, variables are not divided into dependent and independent, rather they are analyzed simultaneously to establish their underlying relationships. SEM can be thought of being both dependent and interdependent techniques, because SEM's foundation lies in two familiar multivariate techniques: **factor analysis and multiple regression analysis**, and has ability to analyze multiple relationships simultaneously.

References

Joseph F. Hair Jr, William C. Black, Barry J. Babin and Rolph E. Anderson (2010). Multivariate Data Analysis: Prince Hall, Upper Saddle River, NJ07458.

Richard Taylor and William Francis (2006). Hand Book of Univariate and Multivariate Data Analysis and Interpretation with SPSS. Chapman & Hall/CRC Taylor & Francis Group, 600

Broken Sound Parkway NW, Suite 300 Boca Raton, FL 33487-2742.

Berger R D, Mishoe J W. 1976. GSMP simulation of several growth functions to describe epidemic progress. Proc Am Phytopathol Soc, 3: 217.

Bharadwaj C L, Singh B M. 1983. The stability of resistance to Pyricularia oryzae Cav. in rice. Ind Phytopathol, 36: 422–426.

Bradbury P J, Zhang Z, Kroon D E, Casstevens T M, Ramdoss Y, Buckler E S. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics, 23: 2633–2635.

Campbell C L, Madden L V, Pennypacker S P. 1980. Structural characterization of bean root rot epidemics. Phytopathology, 70(2): 152–155.

Campbell C L, Madden L V. 1990. Introduction to Plant Disease Epidemiology. New York: John Wiley & Sons. Carlson G A. 1969. Bayesian analysis of pesticide use. In: Proceedings of the American Statistic Association. Business and Economic Statistics Section: 411–416.

Carlson G A. 1970. A decision theoretic approach to crop disease prediction and control. Am J Agric Econ, 52(2): 216–223.

Chiang K S, Huang Y T. 2005. Analysis of the spatial pattern of rice leaf blast. Plant Prot Bull Taipei, 47(2): 129–142.

Choi W J. 1987. A computer simulation model for rice leaf blast. Suweon: Korea Seoul National University.

Collett D. 1991. Modelling Binary Data. London: Chapman & Hall. Crossa J, Gauch H G J, Zobel R W. 1990. Additive main effects and multiplicative interaction analysis of two international maize cultivar trials. Crop Sci, 30(3): 493–500.

Eberhart S A, Russell W A. 1966. Stability parameters for comparing varieties. Crop Sci, 6(1): 36–40.

Eskridge K M. 1995. Statistical analysis of disease reaction data using nonparametric methods. Hort Sci, 30(3): 478–480.

Ezuka A, Horino O. 1974. Classification of rice varieties and Xanthomonas oryzae strains on the basis of their differential interaction. Bull Tokai Kinki Natl Agric Exp Stat,27: 1–19. (in Japanese) Faris M A, de A Lara M, de S Leao Veiga A F. 1979. Stability of sorghum midge resistance. Crop Sci, 19(5): 577–580.