THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

Scientific Levels of Field Data Analysis in Computing Research

Bostley Muyembe Asenahabi
Lecturer, Department of Physics and Computer Science, Alupe University College, Kenya
Peters Anselemo Ikoha
Lecturer, Department of Information Technology, Kibabii University, Kenya

Abstract:
Scientific research is a study that is systematically planned with a view of solving real life problems through creating new knowledge. A researcher requires to draft questions in line with the constructs and concept being handled. Data analysis begins with proper analysis of variables and literature to be reviewed. However, field data analysis involves descriptive, exploratory, inferential, predictive, causal and mechanistic. A desktop literature review of papers published in accredited research journals, books and popular articles was conducted. This review paper will be of great help to both occasional and seasonal researchers as they analyze data to solve problems scientifically and generate new knowledge.

Keywords: *Scientific research, descriptive analysis, exploratory analysis, inferential analysis, predictive analysis, causal analysis*

1. Introduction

Scientific research is a way of creating new knowledge in a planned manner through systematic collection, interpretation and evaluation of data. Before beginning the scientific research, the researcher should determine the objectives to be achieved, do planning and specify the methodology (Çaparlar & Dönmez, 2016). Scientific research majorly requires quantitative data which can be analyzed using different data analysis tools.

Data analysis brings out order, structure and meaning to the raw data that has been collected. A researcher should select the appropriate data analysis tools to analyze the data for each data set before the study commences (Thompson, 2009). The researcher should first plan to describe the sample, a process that is done through identifying the important demographic characteristics of the sample for instance gender, age, religion, race and academic background.

The researcher should plan the analysis of each data set by indicating the research question/hypothesis, all relevant variables within the question and the analysis tools to be used. This process helps researchers to ensure that they have collected all the relevant data. This is in tandem with Thompson and Panacek, (2006) who suggest that a researcher has to develop a list of questions to be in the data collection tool which should all reflect the research question. The researcher should also be able to defend the choice of analysis tools applied in the research.

A data analysis plan speeds up the analysis process as the researcher is not left stranded with how to manipulate the collected data. Burns and Grove, (2005) postulate that a data analysis plan increases scientific integrity as it decreases the chance of making a type I statistical error which basically happens when a researcher repeats analysis with an aim of forcing to bring out something that is statistically significant. The researcher is also expected to identify the computer system to be used for data analysis, the method to be used for data entry (Thompson & Panacek, 2008) and the data analysis application that will be used for the study. The data analysis process will run smoothly if a researcher knows what he wants from the collected data and the process required to achieve it.

It is essential for a researcher to perform the various levels of data analysis after collecting primary data for scientific computing research. By performing the various levels of data analysis, a researcher is able to build up a framework or model which forms the end product of the study. A researcher is also able to come up with a solution to the research problem and generate new knowledge from the study. Besides, a researcher is able to identify new areas of research which can be ventured into. This paper describes the different levels of data analysis essential for scientific research ranging from the low level to the high level data analysis.

2. Levels of Data Analysis

There are basically six levels of data analysis ranging from low level (descriptive and explorative) to high level data analysis (inferential, predictive, causal and mechanistic). Figure 1 summarizes the different levels of data analysis.

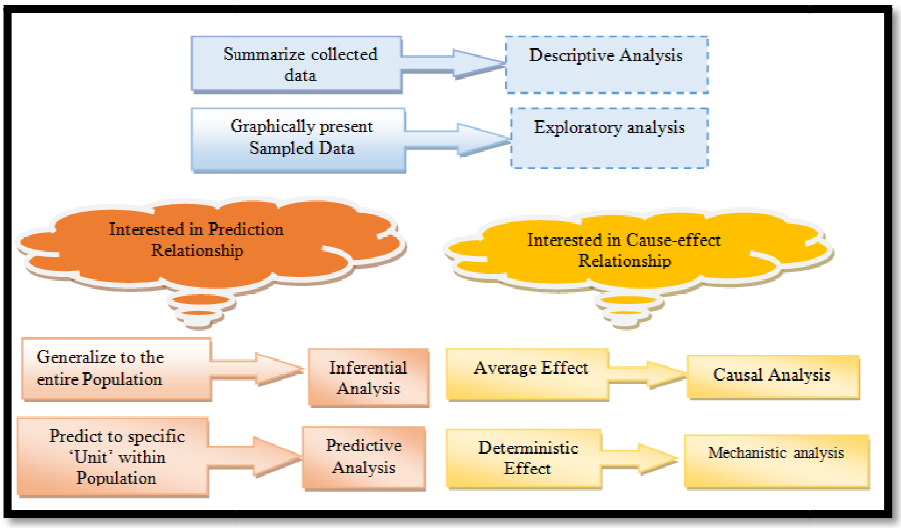

Figure 1: Summarized Levels of Scientific Computing Research Data Analysis

2.1. Descriptive Data Analysis

Descriptive data analysis process summarizes data elements with an aim of describing what happened in the sample. It enables the researcher to detect sample characteristics which may in the long run have an impact on the conclusion. It forms the initial data analysis process performed on demographic variables like age, gender, academic qualifications and race (Omair, 2014) on the available data set by describing the characteristics of the sample and checking the variables for any violation of assumptions underlying the analysis techniques used to address the research questions.

For instance, if a researcher is out to analyze the usability of a certain virtual learning environment in a learning institution, he is required to collect and summarize demographic data about the respondents who interact with the virtual learning environment. Data elements can be analyzed descriptively using different data analysis tools as illustrated in Table 1: Types of descriptive data analysis tools.

Data Analysis Tools	Function
Frequency distribution	Gives a summary of the variables in number or percentage form. Describes categorical data for instance ratio of users who use laptops to users who use smartphones to attend to web conferences.
Measure of central tendency - mean, median and mode	Gives the average, most common or center of a given set of data. Mean is applied on continuous data types to describe the average number or score.
Measure of variability – range, standard deviation, Variance	Provides a sense of how much the responses are spread out. The rate of vary displays how much spread there is in the most extreme response scores. The standard deviation is applied on continuous data and informs how far each score lies from the mean. Variance displays how much spread there is in the data.
Skewness and Kurtosis	Skewness value displays the symmetry of the distribution while Kurtosis indicates its 'peakedness'.

Table 1: Descriptive Data Analysis Tools

2.2. Exploratory Data Analysis

Exploratory data analysis (EDA) is used for visualization and studying the data set so that statistical regularities that might not be apparent can be uncovered. It establishes relationships that were previously unknown, presents the context required to come up with the appropriate model for the problem being handled and to efficiently interpret its results. Exploratory data analysis relies on graphical techniques to bring out data patterns (Myatt & Johnson, 2014) with respect to the sampled data only, which implies that it can not be generalized to the entire population.The graphical techniques that are commonly used for exploratory data analysis are illustrated in Table 2: Types of Exploratory data analysis tools

Data Analysis Tool	Presentation Tool	Function
Univariate analysis	Box plot	Graphically depicts a five-number summary of numerical data: minimum, first quartile, median, third quartile, and maximum. Outliers are shown by dots above or below the whisker.
	Histogram	It approximately represents the frequency distribution of continuous data with height proportional to the number of cases in each bin. It allows for inspection of data for its underlying distribution – normal distribution, outliers or skewness
Multivariate analysis	Bar chart	It displays the distribution of categorical data
	Line graph	It is used to display changes over time
	Map chart	It is used to visually present location data
	Pareto chart	It contains both bars and a line graph, where individual values are represented in descending order by bars, and the cumulative total is represented by the line.
	Pie charts	It is used to display the relationship between a part to a whole
	Scatter plot	It is used to visually present relationship between two quantitative variables using dots. It can suggest positive correlation (rising), negative correlation (falling), or null (uncorrelated) relationship.

Table 2: Types of Exploratory Data Analysis Tools

2.3. Inferential Data Analysis

Inferential data analysis allows a researcher to generate a conclusion based on evidence and reasoning rather than explicit statements. It enables a researcher to explore the strength of relationship between variables; whether there are differences between two or more samples and whether these differences are likely to be present in the population of interest. Inferential analysis is used to generalize results obtained from a random sample to the population which it was drawn from. A researcher can use inferential data analysis to compare results between different subgroups of the sampled data to determine if there is any statistically significant association or difference between the independent and dependent variables with respect to the study concepts. Bhasin, (2019) points out that this type of data analysis can have excellent prediction results if a proper sampling technique is applied and good data analysis tools are used to analyze the data sets. Inferential data analysis is basically used to: make estimates about the populations and test hypothesis to draw conclusions about populations.

Data analysis tests used for inferential data analysis are: tests of comparison and as describe in Tables 3 and 4.

- Comparison tests are used to verify if differences exist in the means, medians or rankings of scores of two or more groups. The test to be used should be selected considering the parameters of the data to be analyzed. Means are mainly calculated for interval or ratio data, whereas medians and rankings are measures associated with ordinal data.

Type of Test	Function
Independent t-test (Parametric)	Used to compare the means between two independent/unrelated groups when the dependent variable is continuous
Mann-Whiney U (Non-Parametric)	
Dependent t-test (Parametric)	Used to compare the means of two related samples on the same continuous dependent variable
Wilcoxon signed-rank (Non-Parametric)	
ANOVA (Parametric)	Used to determine if there are statistically significant differences between the means of two or more groups of an independent (unrelated) variable on a continuous or ordinal dependent variable
Kruskal-Wallis H (Non-parametric)	
Mood's Median (Non-parametric)	Used to determine whether the medians of two independent samples are equal

Table 3: Comparison tests

- Correlation Tests are Used to Describe the relationship between two variables. Despite Pearson's *r* being a more statistically significant test, Spearman's *r* is recommended for interval and ratio variables when the analyzed data is not normally distributed.

Tye of Test	Nature of Data	Variables
Pearson's r	Parametric	Interval/Ratio
Spearman's r	Non-Parametric	Ordina/Interval/Ratio
Chi square test of independence	Non-Parametric	Nominal/Ordinal

Table 4: Correlation tests

2.4. Predictive Data Analysis

Predictive data analysis enables a researcher to make predictions for specific individuals or units within a population. This kind of data analysis is applied by a researcher in analyzing the current trends along with historic facts to come up with conclusions which predict about trends of future events (Moghimi, 2016). The prediction outcome and success of the model depends upon choosing and measuring the right variables (Bhasin, 2019). This kind of analysis applies known results to come up with a model that can be used to either predict values for different data or new data (Banumathi & Aloysius, 2017).

Regression analysis is the main technique used to conduct predictive data analysis. The focus is mainly to design a mathematical equation as a model to represent the interaction among different variables in consideration. Table 5 indicates the different analysis tools used for predictive data analysis while Table 6 indicates the different regression tests.

Tool Used	Function
Linear regression	Predicts the response variable as a linear function of the parameters with unknown coefficients.
Multiple regression	Used when you want to explore the predictive ability of a set of independent variables on one continuous dependent measure
Logistic regression	Used to assign outcome probabilities to observations. It transforms information about the binary dependent variable into an unbounded continuous variable and estimates a regular multivariate model.

Table 5: Predictive Analysis Tools

Regression tests are used to illustrate if a change in the independent variables leads to a change in a dependent variable.

Type of Test	Independent Variable	Dependent Variable
Simple linear regression	1 interval/ratio	1 interval/ratio
Multiple linear regression	2+ interval/ratio	1 interval/ratio
Logistic Regression	1+ any variable type	1 binary variable type
Nominal regression	1+ any variable type	1 nominal
Ordinal regression	1+ any variable type	1 ordinal variable

Table 6: Regression Tests

2.5. Causal Data Analysis

Causal data analysis enables a researcher to determine the cause-and-effect relationship between variables. The researcher is able to visualize the effect of eliminating a variable on another variable. Application of causal studies usually requires randomized studies. This type of analysis is considered to be a high-level data analysis.

Explicit causal data analysis can be performed by a researcher after carrying out an experiment so as to compare the control and treatment group using variance analysis. Researchers are always out to identify if a particular independent variable affects the dependent variable. There is no particular data analysis tool used for determining causal relationship. However, a criterion of association, time ordering (temporal precedence) and non-spuriousness can be used to establish causality. An association (correlation) between the independent and dependent variable is an indicator of causality, however, it should not be mistaken to imply causality.

After establishing association, the researcher has to determine the temporal precedence of the variables of interest. Logically, for the independent variable to have an effect on the dependent variable, it has to occur first. This can easily be achieved in experimental research if a researcher controls the exposure to the treatment (independent variable) and measure the outcome (dependent variable). In survey design, time ordering can turn out to be much difficult to monitor. Determining temporal precedence may involve application of logic, existing research and common sense.

Non-spuriousness is another issue to be considered for a researcher to state that causality exists. A spurious or false relationship exists if an extraneous variable affects the association between the independent and dependent variable under study. Researchers in the scientific field are always challenged in ruling out spurious relationships since other factors exist which affect the association of two variables.

For a researcher to effectively determine causal relationship, he has to use appropriate research design, collect data carefully besides using statistical controls and triangulation of different data sources. Table 7 depicts the different analysis tools used for causal data analysis.

Tool Used	Function
Independent t-test (Parametric)	Used to compare the means between two independent/unrelated groups when the dependent variable is continuous
Mann-Whitney U Test (Non-parametric)	
Dependent t-test (Parametric)	Used to compare the means of two related samples on the same continuous dependent variable
Wilcoxon Signed Rank test (non-parametric)	
One-way ANOVA (Parametric)	Used to determine if there are statistically significant differences between the means of two or more groups of an independent (unrelated) variable on a continuous or ordinal dependent variable
Kruskal-Wallis test	
Pearson r correlation (Parametric)	Used to measure the degree of linear dependence between two variables
Spearman's correlation/Chi-square test (Non-parametric)	

Table 5: Causal Data Analysis Tools

2.6. Mechanistic Data Analysis

Mechanistic data analysis is a high-level data analysis process and requires maximum amount of effort for a researcher to develop a model from the collected data. According to Bhasin, (2019) the main idea behind this kind of data analysis is understanding the nature of exact changes in variables which affect other variables. Mechanistic model involves writing down scientifically motivated equations which describe the collection of dynamic systems giving rise to the observations on each unit (Breto, Ionides, and King, 2020). Factor analysis is an umbrella term used for both Standard Factor Analysis and Principal Component Analysis. Table 8 highlights the different mechanistic data analysis tools.

Tool Used	Function
Factor analysis	Used to develop a linear combination with fewer variables with respect to the original variables so that the reduced variables account for or capture and provide an explanation of most the variance/variability in correlation matrix pattern.
Standard Factor Analysis	The factors are estimated using mathematical model to develop a theory or come up with a theoretical solution for an ideal/perfect world situation without massive variance. IT generates factors (a combination of variables) that go into development of the model
Principal component analysis	Used to describe the data patterns and direct the data by bringing out their relationship. It also transforms the original variables into a set of components having very strong linear correlations that go into the measurement tool/technique. Used to come up with a practical application/ empirical solution or real-world summary of data set
Exploratory Factor Analysis	Used in the early stages of research to develop a theory or gather information about relationship among variables
Confirmatory Factor Analysis	A multivariate statistical procedure that is used to test or confirm specific hypothesis or theories developed concerning the structure for the constructs underlying a set of variables
Structural Equation Model	SEM is a multivariate statistical analysis technique combining factor analysis and multiple regression analysis and is used to analyze structural relationships, show causal relationships between measured variables and latent constructs.
Empirical Orthogonal function	Used to decompose a data set in terms of orthogonal basis functions which are determined from the data. It is typically found by computing the eigenvectors of the covariance matrix of the data set.

Table 6: Mechanistic Data Analysis Tools

3. Conclusion

Data analysis is essential to a researcher when using scientific research to come up with a solution to a problem. This paper has analyzed the different levels of data analysis and the required data analysis tools for each level.

4. References

i. Banumathi, S., and Aloysius, A. (2017). Predictive analytics concepts in big data - A survey. *International Journal of Advanced Research in Computer Science , 8*(8), 27 - 30.

ii. Bhasin, H. (2019). *8 Types of Analysis in Research.* Retrieved from Marketing management articles.

iii. Breto, C., Ionides, E. L., and King, A. A. (2020). Panel Data Analysis via Mechanistic Models. *Journal of the American Statistical Association, 115*(531), 1178 - 1188.

iv. Burns, N., and Grove, S. K. (2005). *The practice of nursing research: Conduct, critique, and utilization.* (5th ed.). St. Louis: Elsevier Saunders.
v. Çaparlar, C. Ö., and Dönmez, A. (2016). What is Scientific Research and How Can it be Done? *Turkish Journal of Anaesthesiology and Reanimation, 44*, 212 - 218.
vi. Hoda Moghimi, S. V. (2016). How Do Business Analytics and Business Intelligence Contribute to Improving Care Efficiency? *49th Hawaii International Conference on System Sciences.*
vii. Myatt, G., and W. Johnson. (2014). *Making sense of data I: A practical guide to exploratory data analysis and data mining.* Hoboken, NJ: Wiley.
viii. Omair, A. (2014). Understanding the process of statistical methods for effective data analysis. *Journal of Health Specialties, 2*(3), 100 - 104.
ix. Panacek, E. A., and Thompson, C. B. (2007). Basics of research part 5: Sampling methods: Selecting your subjects. *Air Medical Journal, 26*, 75–78.
x. Shuttleworth, M. (2008). *Correlation and Causality. Experiment Resources.*
xi. Thompson, C. B. (2009). Descriptive Data Analysis. *Air Medical Journal, 28*(2), 56 - 59.
xii. Thompson, C. B., and Panacek, E. A. (2008). Data management. *Air Medical Journal*, 156 - 158.
xiii. Thompson, C., and Panacek, E. (2006). Clinical research and critical care transport: How to get started. . *Air Medical Journal, 25*, 107 - 111.