Different Types of Data Analysis; Data Analysis Methods and Techniques in Research Projects Authors

Hamed Taherdoost

International Journal of Academic Research in Management (IJARM)
Vol. 9, No. 1, 2020, Page: 1-9, ISSN: 2296-1747

www.elvedit.com

Different Types of Data Analysis; Data Analysis Methods and Techniques in Research Projects

Authors

Hamed Taherdoost
Research Club, Research and Development Department | Hamta Group, Hamta Business Corporation

hamed.taherdoost@gmail.com
hamed@hamta.org
Vancouver, Canada

Abstract

This article is concentrated to define data analysis and the concept of data preparation. Then, the data analysis methods will be discussed. For doing so, the first six main categories are described briefly. Then, the statistical tools of the most commonly used methods including descriptive, explanatory, and inferential analyses are investigated in detail. Finally, we focus more on qualitative data analysis to get familiar with the data preparation and strategies in this concept.

Key Words

Data Analysis, Data Preparation, Data Analysis Methods, Data Analysis Types, Descriptive Analysis, Explanatory Analysis, Inferential Analysis, Predictive Analysis, Explanatory Analysis, Causal Analysis and Mechanistic Analysis, Statistical Analysis.

I. Data Analysis and Data Preparation

Data analysis is simply the process of converting the gathered data to meaningful information. Different techniques such as modeling to reach trends, relationships, and therefore conclusions to address the decision-making process are employed in this process (Start, 2006). However, the data needs to be prepared before being used in the data analysis process.

Data preparation is the process in which data is converted to the numerical format which is machine-

readable to be used in specific analyzing programs such as SAS or SPSS.

The steps to follow for the data preparation process are data coding, data entry, missing values, and data transformation. These steps are described briefly here:

Data Coding: Converting data to numerical values happens during the data coding process. It uses a codebook which is a document including different information such as an explanation of the variables, measures, and format of variables, the response, and finally codding them. In this process response means determining the types of scales for instance, whether the scale is chosen as nominal, ratio, ordinal, or interval; whether the scale is five-point, seven-point, etc. For example, to code the industry type, we can use a numerical form, and the coding scheme can be considered as 1 for healthcare, 2 for manufacturing, 3 for retailing, and 4 for financial.

Data entry: In this process, the coded data from the previous step is entered into text files or spreadsheets. It also can be directly added to the statistical program.

Missing data: As some respondents may not answer all the questions because of different reasons, a method should be used to face these missed values. For example, you need to add the value -1 or 999 in some programs, some of them automatically address the missed values, and others use a listwise deletion technique facing the missing values which drop all the answers even with a single missed value.

Data transformation: Transforming data is needed before interpreting them in some cases. Reverse coded items can be considered as an example that should be transformed before comparing or combining with not reversed ones. This concept is used where the meaning of the item is opposite to their underlying construct (Bhattacherjee, 2012).

II. Types of Data Analysis

In this section, we discuss the main types of data analysis methods in brief. For this purpose, the data analysis can be categorized into the following six main methods (Taherdoost, 2021):

- Descriptive
- Exploratory
- Inferential
- Predictive
- Explanatory or Causal
- Mechanistic

A. Descriptive: Recognized as the first type of data analysis, it is known as the method with the least amount of effort. Thus, it can be used for large volumes of data. Here the data is used to perform a data set (Start, 2006).

B. Exploratory: This method is used to explore the unknown relationships and discover new connections, and define future studies or questions (Start, 2006).

C. Inferential: Inferential analyzing method uses a small sample to conclude a bigger population. It means, data from a subject sample of the world is used to test a general theory about its nature. The types of data sets that can be used in this method are observational, retrospective data set, and cross-sectional time study (Bhattacherjee, 2012).

D. Predictive: Predictive analysis utilizes historical and current facts to reach future predictions. It can also use data from a subject to predict the values of another subject. There are different predictive models; however, a simple model with more data can work better in general. Therefore, the prediction data set and also the determination of the measuring variables are important aspects to consider (MacGregor, 2013).

E. Explanatory: This analyzing method is used to determine the consequences happening to one variable when changing another one using randomized trial data sets (Bhattacherjee, 2012).

F. Mechanistic: This method needs the most effort to determine the exact changes in the variables which can lead to changes in other ones using randomized trial data sets. It can be also concluded that mechanistic analysis is hardly inferable. Thus, when you need high precision in your result and you should minimize your errors, for example in the engineering and physical sciences, it can be a choice.

Next, we will focus on the statistical analysis tools of three common data analysis types (descriptive, explanatory, and inferential).

III. Descriptive Analysis

This method summarizes the data to reach a simple presentation as a result. This method can be categorized into Univariate and Bivariate analyses (Taherdoost, 2021).

Univariate is a set of different statistical tools which look for characteristics and general properties of one variable. These statistical techniques are Frequency, Central Tendency, and Dispersion.

Frequency disruption is the most basic method to determine the disruption of variables. It determines all possible values for a specific variable and the number of times or the frequency that each of those values is in the data set.

The *central tendency* of disruption which is also known as three Ms is used to determine the number of the most represented value which can help you to use a single variable in comparison to a set of data. The mean, median, and mode are the most commonly used central tendency methods. Mean is simply the average of the values, Medium is the middle value in the data set, and Mode is the amount of the

most frequently occurring value.

Dispersion is the way of spreading the variables around the central tendency. Common tools are range, variance, and the square root of the variance which is known as standard deviation. The range is the amount of the difference between the highest and the lowest values. The variance shows the concentration of values around the average value. It is used when we have two variables in our data set and we also need a comparison between two data sets. Therefore, it can show the relationship and connections between these two variables. Bivariate correlation or simply correlation is the most common measure. This measure uses a specific formula to calculate correlation using the sample mean values and standard deviations. This method also can be used when the number of variables is more than two. Although due to the complexity of using formulas manually, these programs can be simply solved by computing with software such as SPSS (Bhattacherjee, 2012).

IV. Explanatory Analysis

Explanatory analysis as discussed before looks to find influences. It means explanatory analysis tries to reach the answer to the research questions related to the connections, relations, and patterns between variables (Taherdoost, 2014; Taherdoost & Madanchian, 2020). The main explanatory analysis techniques are Dependence and Interdependence methods. Dependence is concerned with the impact of a set of predictor variables on a single outcome variable. Interdependence techniques are multivariate analyses that aim to determine the interrelationships between the variables considering no assumption for the influence direction. The main tools for these techniques are shown in the Figure.1.

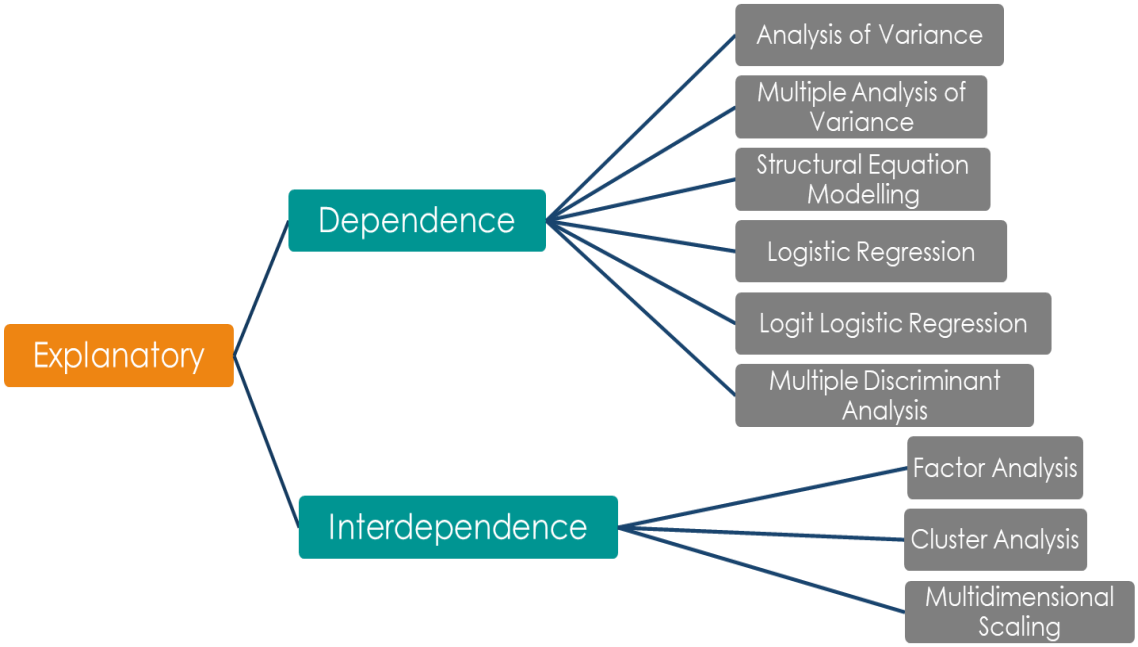

Figure 1: Explanatory Analysis Types

For dependence techniques the following common tools can be used.

Analysis of variance (ANOVA): It is utilized to make a comparison between the calculated separate outcomes of different groups which are the predictor variable and also multi-category. This tool can be used in inferential analysis, and it is discussed in more detail in that section.

Multiple analysis of variance (MANOVA): As the extension of the former tool, the comparison can be gained across two or more outcome variables. Here the groups are the equivalent of a categorical predictor variable. Both ANOVA and MANOVA are specifically used in experimental research.

Structural equation modeling: the relationships between a series of interrelated and predictor variables can be handled using this method which is also known as LISREL. The structural relationship between the variable which is measured and the latent construct can be defined using this tool. As a comprehensive method, it is beneficial for testing hypotheses as well as testing, presenting, and estimating, theoretical networks of relations that are mostly linear between variables. For measurement equivalence it can also be applied in complex analyses for first and higher-order constructs (Taherdoost, 2018).

Logistic regression: The logistic regression method contains multiple regressions with dichotomous outcome variables and categorical or metric predictor variables. Logistic regression is similar to the former tool, but in this type, the outcome variables include more than two categories.

Multiple discriminant analysis: As an alternative method multiple discriminant analysis can be used instead of the forms of logistic regression described before. Different numbers of predictor variables and a single outcome variable also can be handled in both dichotomous and multichotomous categories by using this tool.

As discussed before, the interdependence technique is employed to reach the relationships of a set of variables and it considers neither explanation nor influence. The most common ones including factor analysis, cluster analysis, and multidimensional scaling are explained following:

Factor analysis: This method is useful to analyze and discover the patterns and relationships when we face a large number of variables by reducing the variables to either a new variate or a derived smaller set of factors. It is not a method aiming to predict the outcome variables but as a series of statistical techniques, it is used to perform latent factors driving observable variables (Taherdoost, 2017).

Cluster analysis: This method is used to classify objects or individuals. For this purpose, mutual groups are constructed to maximize both the homogeneity and heterogeneity between clusters. Factor analysis groups variables as factors; however, the cluster method groups objects or people together by considering different criteria.

Multidimensional scaling: It is also called perceptual mapping and is used to identify key dimensions considering individuals' judgments and perceptions. For this purpose, using distances that are represented in multidimensional space, judgments and perceptions can be transformed. In this method, you have the outcome variable which is recognized as judgment or perception here and you must determine the independent variables which are the perceptual dimensions. This is, in fact, the main difference between this method and cluster analysis (Blaikie, 2003).

V. INFERENTIAL ANALYSIS

The inferential analysis is simply the way from sample to population. Here, we provide a basic overview of the most widely used inferential statistical procedures, including the t-test, analysis of variance (ANOVA), chi-square, regression, and time series.

T-Test: T-test also known as student's t-test is a method commonly used to test a hypothesis in comparison to means or averages between the groups. It uses single dichotomous independent and continuous dependent variables. The T-test can be categorized into the non-directional or two-tailed test and directional or one-tailed test. A non-directional or two-tailed test is used to understand the differences of the means of two groups from each other statistically. On the other hand, the directional or one-tailed test determines whether the mean of one group is statistically larger than the other one. To sum, a t-test can be used to compare the mean values of two independent or dependent samples, the difference between the sample mean value with the assumed mean, and finally for the sample mean, to determine the confidence interval. It can be added that this test is designed for hypothesis testing. These hypotheses include null hypothesis and alternative hypothesis, and they can be shown as:

$H0: \mu_1 \leq \mu_2$ (null hypothesis)

$H1: \mu_1 > \mu_2$ (alternative hypothesis)

Where μ is the mean population.

To compare two different methods based on their superiority, and based on some specific assumptions, both can be considered equally good, these assumptions are called the null hypothesis. So generally, the rejection of the null hypothesis is the aim of the statistical significance tests. In contrast, the alternative hypothesis is stated when we want to show one of the methods is more superior to the other (Bhattacherjee, 2012).

Analysis of Variance (ANOVA): The results of an ANOVA are equal to the results of using multiple t-tests. Therefore, this method can be more efficient, and decrease the added experiment-wise error as well as the chances of spurious results. It can also address the issues to the validity of the complicated statistical conclusion which happens using a series of t-tests. It is interesting to note that, ANOVA despite its variance-based name, instead of using variance, uses the differences between the mean values for comparison of the groups. However, during the process of determining it is utilized whether the

means are different or not. Using the mean value, ANOVA can compare more than a group. It makes t-test a special format of ANOVA which can just consider one group in each test.

Two main categories of ANOVA are one-way ANOVA and multifactor ANOVA. But as discussed before another variant of this analysis is MANOVA.

One-way ANOVA is applied to determine the existence of a statistical difference between the mean or average value of two or more levels of a single independent variable. On the other hand, the multifactor technique is used for two or more independent factors or variables. In the MANOVA technique, there are more than two or more independent variables that can be considered related in some way (Taherdoost, 2021).

Chi-Square ($\chi 2$): Chi-square is used to detect the relationships between the categories of two variables. These two variables which are also categorical are from the same population. In contrast with the other discussed inferential tools (applying for interval and ratio data), this technique must be only applied for nominal and ordinal data because it is simply a test of proportions. The basis for this is comparing one set of proportions with your expected value by chance. It compares the discrepancy of observed and expected frequencies. Smaller chi-squares values show smaller discrepancies between those values.

Regression: In the regression method, the values of one or more independent variables can be utilized to predict a value on some dependent variable. Thus, this statistical technique is similar to a correlation which shows the relationship between variables. However, they are different based on the primary predictive object in regression. An important application of regression is to estimate the possibility and the degree of predicting the onset of specific criteria using a set of hypothesized risk factors (Taherdoost, 2021).

Regression can be simple or multiple. The simple method uses a single independent variable, but the multiple methods may use several independent variables to determine and estimate the dependent variable (Marczyk, et al, 2010).

Time Series Analysis: The variables which are changing continuously based on time can be analyzed using time series. It is also a good choice for longitudinal research designs. In these designs using regular intervals and a large number of observations, single units or objects are calculated repeatedly. In general, the aims are summarizing time-series data, fitting models with low dimensions, and finally making predictions. Therefore, it can be added that where the time series compares different points in single series, regression is used to test the relationship between time series (Bhattacherjee, 2012).

VI. Qualitative Data Analysis

The main difference between the analysis of the qualitative data and quantitative data is the important role of the researcher in the qualitative analysis based on his/her techniques, knowledge, and integrative

skills on the analysis results. Data coding is also necessary for qualitative data analysis. Common qualitative techniques including grounded theory and content analysis are reviewed here.

Grounded Theory: It is one of the techniques used for text data analysis. Grounded theory is an inductive method, which uses the data to build theories about that phenomenon. Simply, this method categorizes the data which is gathered in a text format to codes, categories, and relationships. The main phase of the coding process in the grounded theory are as follows:

- Open coding: In this step, data is transformed into codes with a descriptive label. But for this purpose, data should be turned to the small and discrete components.
- Axial coding: In this step, the relationships and connections between the codes should be determined, then the codes must be condensed into the broader categories by considering their connections.
- Selective coding: This phase uses a central and core category that can be chosen from previous ones or can be a newly generated one. Then, the connections of data and codes with this core category must be identified. It must be considered that in this step we must eliminate the codes and categories which do not have enough supporting data.

After the central concept is generated, the theory should be written by pulling together all the categories. This can be done using different integration techniques like storylining, memoing, or concept mapping. In general, these methods are used to find the connections and integrate the categories around the central category (Taherdoost, 2021).

Content analysis: This method uses a qualitative, or quantitative manner to analyze the text. It is a systematic approach that uses different steps to analyze the contents. The first step is sampling to select a set of text from a large population. This process is not based on random selection; however, texts with more pertinent content should be selected as samples. The next step is dividing the texts into segments or "chunks" using specific rules. Codes, in the next step, are applied to the segments. You can use one or more than one code for each of them. Finally, the codes are analyzed to find the most frequent ones. This final process can be both quantitative and qualitative (Bhattacherjee, 2012).

VII. Conclusion

This article provided a summary of the most common data analysis techniques. It first describes data preparation methods which are an essential process in analyzing data. Then, common methods are reviewed, and the tools for the most important techniques are discussed. Qualitative data analysis and its strategies are also discussed more specifically in the final section.

REFERENCES

[1] Bhattacherjee, A. (2012). Social science research: Principles, methods, and practices (2nd ed.).

[2] Taherdoost, H. (2014). Exploratory Factor Analysis; Concepts and Theory. Advances in Applied and Pure Mathematics, 375-382.

[3] Taherdoost, H. (2021). Handbook on Research Skills: The Essential Step-By-Step Guide on How to Do a Research Project: Amazon Kindle.

[4] Blaikie, N. (2003). Analyzing quantitative data: From description to explanation. Sage.

[5] MacGregor, J. (2013). Predictive Analysis with SAP®. Bonn: Galileo Press.

[6] Taherdoost, H. (2018). Development of an adoption model to assess user acceptance of e-service technology: E-Service Technology Acceptance Model. Behaviour & Information Technology, 37(2), 173-197.

[7] Marczyk, G. R., DeMatteo, D., & Festinger, D. (2010). Essentials of research design and methodology (Vol. 2). John Wiley & Sons.

[8] Start, S. (2006). Introduction to Data Analysis Handbook Migrant & Seasonal Head Start Technical Assistance Center Academy for Educational Development. Journal of Academic, 2(3), 6-8.

[9] Taherdoost, H. (2017). Understanding of E-service Security Dimensions and its effect on Quality and Intention to Use. Information and Computer Security, Accepted and in Press. .

[10] Taherdoost, H., & Madanchian, M. (2020). Developing and Validating a Theoretical Model to Evaluate Customer Satisfaction of E-Services. In K. Sandhu (Ed.), Digital Innovations for Customer Engagement, Management and Organizational Improvement (pp. 46-65.): IGI Global.

AUTHORS' BIOGRAPHY

Hamed Taherdoost is an award-winning leader and R&D professional. Hamed was lecturer at IAU and PNU universities, a scientific researcher and R&D and Academic Manager at IAU, Research Club, MDTH, NAAS, Tablokar, and Hamta Academy. Hamed has authored over 120 scientific articles in peer-reviewed international journals and conference proceedings (h-index = 24; i10-index = 50), as well as eight book chapters and seven books in the field of technology and research methodology. He is the author of the Research Skills book and his current papers have been published in Behaviour & Information Technology, Information and Computer Security, Electronics, Annual Data Science, Cogent Business & Management, Procedia Manufacturing, and International Journal of Intelligent Engineering Informatics.

He is a Senior Member of the IEEE, IAEEEE, IASED & IEDRC, Working group member of IFIP TC and Member of CSIAC, ACT-IAC, and many other professional bodies. Currently, he is involved in several multidisciplinary research projects among which includes studying innovation in information technology & web development, people's behavior, and technology acceptance.