RESEARCH METHODOLOGY

UNIT – 4

Syllabus - UNIT-IV : STATISTICAL APPLICATIONS

Factor Analysis - bivariate and Multivariate Analysis. (Practical problems.)

Expected Leaning Outcome:

1. **To impart knowledge for enabling students to develop data analytics skills and meaningful interpretation to the data sets so as to solve the business/Research problem.**
2. **Have adequate knowledge on measurement & scaling techniques as well as the quantitative data analysis**
3. **Have basic knowledge on qualitative research techniques**
4. **Have basic awareness of data analysis-and hypothesis testing procedures**

ANALYSIS OF DATA

In the beginning the data is raw in nature but after it is arranged in a certain format or a meaningful order this raw data takes the form of the information. The most critical and essential supporting pillars of the research are the analysation and the interpretation of the data.Both these aspects of the research methodology are very sensitive in nature and hence it is required that both these concepts are conducted by the researcher himself or under his very careful and planned supervision. With the help of the interpretation step one is able to achieve a conclusion from the set of the gathered data.

Analysis of the data can be best explained as computing some of the measures supported by the search for relationship patterns, existing among the group of the data.Research depends a great deal on the collected data but it should be seen that this collected data is not just a collection of the data but should also provide good information to the researcher during the various research operations. Hence to make data good and meaningful in nature and working, data analysis plays a very vital and conclusive role. In this step data is made meaningful with the help of certain statistical tools which ultimately make data self explanatory in nature.

According to Willinson and Bhandarkar, **analysis of data 'involves a large number of operations that are very closely related to each other and these operations are carried out with the aim of summarizing the data that has been collected and then organizing this summarized data in a way that helps in getting the answers to the various questions or may suggest hypothesis.'**

Purpose of Analysis of data

The purpose of the scientific analysis was first explained by Leon Festinger and Daniel Katz and according to both of them; the purpose of the analysis of the data can be explained as follows

1. Should be very productive in nature, with high significance for some systematic theory.
2. Should be readily disposed to the quantitative treatment.

Procedure for the Analysis of the data

1. Data collected can be used in the best possible effective manner by performing the following activities
2. Carefully reviewing all the data collection.
3. Analyzing the data then with the help of certain suitable techniques.
4. . Results obtained from the analysation of the data should then be related to the study's hypothesis.

Analysation Steps

The various steps of the analysation of the data were given by Herbert Hyman and can be summarized as follows –

1. Tabulation of the data after conceptualization, relating to every concept of the procedure is done which ultimately provides an explanation based on the quantitative basis.
2. Tabulation in the same way is carried out for every sub group, which gives quantitative description.

3. To get statistical descriptions consolidating data for different aspects is brought into use.
4. Examination of such data is then done, which helps in improving the evaluation of the findings.
5. Different qualitative and non statistical methods are brought into the use for obtaining quantitative description but only if it is needed.

Types of Analysis

1. Descriptive Analysis

- Also referred to as the One Dimensional Analysis.
- Mainly involves the study of the distribution of one variable.
- Depicts the benchmark data.
- Helps in the measurement of the condition at a particular time.
- Acts as the prelude to the bi – variate and multivariate analysis.
- Such an analysis may be based on the one variable, two variables or more than two variables.
- Helps in getting the profiles of the various companies, persons, work groups etc.

2. Casual analysis:

- Also referred to as the Regression Analysis.
- Has their root in the study of how one or more variables affect the changes in the other variable.
- Explains the functional relationship between two or more variables.
- Helps in experimental research work.
- Explains the affect of one variable on the other.
- Involve the use of the statistical tools.

3. Co – Relative Analysis –

- Involves two or more variables.
- Helps in knowing correlation between these two or more variables.
- Offers better control and understanding of the relationships between the variables.

4. Inferential Analysis –

- Involves tests of significance for the testing of the hypothesis.
- Helps in the estimation of the population values.
- Helps in the determination of the validity data which can further lead to draw some conclusion.
- Takes an active part in the interpretation of the data.

After data collection, the researcher must prepare the data to be analyzed. Organizing the data correctly can save a lot of time and prevent mistakes. Most researchers choose to use a database or statistical analysis program (e.g. Microsoft Excel, SPSS) that they can format to fit their needs and organize their data effectively. Once the data has been entered, it is crucial that the researcher check the data for accuracy. This can be accomplished by spot-checking a random assortment of participant data groups, but this method is not as effective as re-entering the data a second time and searching for discrepancies. This method is particularly easy to do when using numerical data because the researcher can simply use the database program to sum the columns of the spreadsheet and then look for differences in the totals. One of the best methods of checking for accuracy is to use a specialized computer program that cross-checks double-entered data for discrepancies.

Descriptive Statistics

Correlation

Correlation is one of the most often used (and most often *mis*used) kinds of descriptive statistics. It is perhaps best described as "a single number that describes the degree of relationship between two variables."(3) If two variables tend to be "correlated," that means that a participant's score on one variable tends to vary with a score on the other. For example, people's height and shoe size tend to be positively correlated. This means that for the most part, if a person is tall, they are likely to have a large shoe size, and conversely, if they are short, they are likely to have a smaller shoe size. Correlation can also be negative. For example, warmer temperatures outside may be negatively correlated with the number of hot chocolates sold at a local coffee shop. This is to say that as the temperature goes up, hot chocolate sales tend to go down. Although causality may seem to be implied in this situation, it is important to note that on a statistical level, **correlation does not imply causation**. A good researcher knows that there is no way to assess from *correlation alone* that a causal relationship exists between two variables. In order to assert that “X caused Y,” a study should be experimental, with control groups and random sampling procedures. Determining causation is a difficult thing to do, and it is a common mistake to assert a cause-and-effect relationship when the study methodology does not support this assertion.

Inferential Statistics

Inferential statistics allow the researcher to begin making inferences about the hypothesis based on the data collected. This means that, while applying inferential statistics to data, the researcher is coming to conclusions about the population at large. Inferential statistics seek to generalize beyond the data in the study to find patterns that ostensibly exist in the target population. This course will not address the specific types of inferential statistics available to the researcher, but a succinct and very useful summary of them, complete with step-by-step examples and helpful descriptions, is available

Statistical Significance

Researchers cannot simply conclude that there is a difference between two groups in a well-constructed study. This difference must be due to the manipulation of the independent variable. No matter how well a researcher designs the study, there always exists a degree of error in the results. This error may be due to individual differences within and between experimental groups, or the error may be due to systematic differences within the researcher’s sample. Irrespective of its source, this error acts as “noise” in the data and affects participants’ scores on study measures even though it is not the variable of interest. Statistical significance is aimed at determining the probability that the observed result of a study was due to the influence of something other than chance. A result is “statistically significant” at a certain level. For example, a result might be significant at p<.05. "P" represents the probability that the result was due to chance, and .05 represents a 5% probability that the result was due to chance. Therefore, in a well-run study, p<.05 means that inferential statistical analysis has indicated that the observed results have over a 95% probability of being due to the influence of the independent variable. The 5% cutoff is generally thought of as the standard for most scientific research. Note that it is theoretically impossible to ever be entirely certain that one's results are not due to chance, as the nature of science is one of observing trends and testing

Bivariate and multivariate analyses

Bivariate and multivariate analyses are statistical methods to investigate relationships between data samples. Bivariate analysis looks at two paired data sets, studying whether a relationship exists between them. Multivariate analysis uses two or more variables and analyzes which, if any, are correlated with a specific outcome. The goal in the latter case is to determine which variables influence or cause the outcome.

Bivariate Analysis

Bivariate analysis investigates the relationship between two data sets, with a pair of observations taken from a single sample or individual. However, each sample is independent. You analyze the data using tools such as t-tests and chi-squared tests, to see if the two groups of data correlate with each other. If the variables are quantitative, you usually graph them on a scatterplot. Bivariate analysis also examines the strength of any correlation.

Bivariate Analysis Examples

One example of bivariate analysis is a research team recording the age of both husband and wife in a single marriage. This data is paired because both ages come from the same marriage, but independent because one person's age doesn't cause another person's age. You plot the data to showing a correlation: the older husbands have older wives. A second example is recording measurements of individuals' grip strength and arm strength. The data is paired because both measurements come from a single person, but independent because different muscles are used.

You plot data from many individuals to show a correlation: people with higher grip strength have higher arm strength.

Multivariate Analysis

Multivariate analysis examines several variables to see if one or more of them are predictive of a certain outcome. The predictive variables are independent variables and the outcome is the dependent variable. The variables can be continuous, meaning they can have a range of values, or they can be dichotomous, meaning they represent the answer to a yes or no question. Multiple regression analysis is the most common method used in multivariate analysis to find correlations between data sets. Others include logistic regression and multivariate analysis of variance.

Multivariate Analysis Example

Multivariate analysis was used in by researchers in a 2009 Journal of Pediatrics study to investigate whether negative life events, family environment, family violence, media violence and depression are predictors of youth aggression and bullying. In this case, negative life events, family environment, family violence, media violence and depression were the independent predictor variables, and aggression and bullying were the dependent outcome variables. Over 600 subjects, with an average age of 12 years old, were given questionnaires to determine the predictor variables for each child. A survey also determined the outcome variables for each child. Multiple regression equations and structural equation modeling was used to study the data set. Negative life events and depression were found to be the strongest predictors of youth aggression.

Factor Analysis:

Factor analysis is a technique that is used to reduce a large number of variables into fewer numbers of factors. This technique extracts maximum common variance from all variables and puts them into a common score. As an index of all variables, we can use this score for further analysis. Factor analysis is part of general linear model (GLM) and this method also assumes several assumptions: there is linear relationship, there is no multicollinearity, it includes relevant variables into analysis, and there is true correlation between variables and factors. Several methods are available, but principal component analysis is used most commonly.

Uses of factor analysis in market research and analysis

Factor analysis has proved to be very beneficial in market research and analysis of variables that determine consumer behavior:

1. It helps to make sense of large data with interlinked relationships
2. It may point out relationships that may not have been obvious
3. It can point out to the underlying relationships with respect to consumer tastes, preferences, etc.
4. It is easier to condense and correlate data through factor analysis and also to draw conclusions from the data gathered in market research and analysis.
5. It can be used to form empirical clusters of variables and underlying factors that affect them

Types of factoring:

There are different types of methods used to extract the factor from the data set:

1. Principal component analysis: This is the most common method used by researchers. PCA starts extracting the maximum variance and puts them into the first factor. After that, it removes that variance explained by the first factors and then starts extracting maximum variance for the second factor. This process goes to the last factor.

2. Common factor analysis: The second most preferred method by researchers, it extracts the common variance and puts them into factors. This method does not include the unique variance of all variables. This method is used in SEM.

3. Image factoring: This method is based on correlation matrix. OLS Regression method is used to predict the factor in image factoring.

4. Maximum likelihood method: This method also works on correlation metric but it uses maximum likelihood method to factor.

5. Other methods of factor analysis: Alfa factoring outweighs least squares. Weight square is another regression based method which is used for factoring.

Types of factor analysis

A factor analysis is mainly used for interpretation of data and in analyzing the underlying relationships between variable and other underlying factors that may determine consumer behavior. Instead of grouping responses and response types, factor analysis segregates the variable and groups these according to their co relevance.

There are mainly three types of factor analysis that are used for different kinds of market research and analysis.

1. Exploratory factor analysis
2. Confirmatory factor analysis
3. Structural equation modeling

Exploratory factor analysis is used to measure the underlying factors that affect the variables in a data structure without setting any predefined structure to the outcome. Confirmatory factor analysis on the other hand is used as tool in market research and analysis to reconfirm the effects and correlation of an existing set of predetermined factors and variables that affect these factors. Structural equation modeling hypothesizes a relationship between a set of variables and factors and tests these casual relationships on the linear equation model. Structural equation modeling can be used for exploratory and confirmatory modeling alike, and hence it can be used for confirming results as well as testing hypotheses.

Factor analysis will only yield accurate and useful results if done by a researcher who has adequate knowledge to select data and assign attributes. Selecting factors and variables so as to avoid too much similarity of characteristics is also important. Factor analysis, if done correctly, can allow for market research and analysis that helps in various areas of decision making like product features, product development, pricing, market segmentation, penetration and even with targeting.

Factor loading:

Factor loading is basically the correlation coefficient for the variable and factor. Factor loading shows the variance explained by the variable on that particular factor. In the SEM approach, as a rule of thumb, 0.7 or higher factor loading represents that the factor extracts sufficient variance from that variable.

Eigenvalues: Eigenvalues is also called characteristic roots. Eigenvalues shows variance explained by that particular factor out of the total variance. From the commonality column, we can know how much variance is explained by the first factor out of the total variance. For example, if our first factor explains 68% variance out of the total, this means that 32% variance will be explained by the other factor.

Factor score: The factor score is also called the component score. This score is of all row and columns, which can be used as an index of all variables and can be used for further analysis. We can standardize this score by multiplying a common term. With this factor score, whatever analysis we will do, we will assume that all variables will behave as factor scores and will move.

Criteria for determining the number of factors: According to the Kaiser Criterion, Eigenvalues is a good criteria for determining a factor. If Eigenvalues is greater than one, we should consider that a factor and if Eigenvalues is less than one, then we should not consider that a factor. According to the variance extraction rule, it should be more than 0.7. If variance is less than 0.7, then we should not consider that a factor.

Rotation method: Rotation method makes it more reliable to understand the output. Eigenvalues do not affect the rotation method, but the rotation method affects the Eigenvalues or percentage of variance extracted. There are a number of rotation methods available: (1) No rotation method, (2) Varimax rotation method, (3) Quartimax rotation method, (4) Direct oblimin rotation method, and (5) Promax rotation method. Each of these can be easily selected in SPSS, and we can compare our variance explained by those particular methods.

Assumptions:

1. No outlier: Assume that there are no outliers in data.
2. Adequate sample size: The case must be greater than the factor.
3. No perfect multicollinearity: Factor analysis is an interdependency technique. There should not be perfect multicollinearity between the variables.
4. Homoscedasticity: Since factor analysis is a linear function of measured variables, it does not require homoscedasticity between the variables.
5. Linearity: Factor analysis is also based on linearity assumption. Non-linear variables can also be used. After transfer, however, it changes into linear variable.
6. Interval Data: Interval data are assumed.

Key concepts and terms:

Exploratory factor analysis: Assumes that any indicator or variable may be associated with any factor. This is the most common factor analysis used by researchers and it is not based on any prior theory.

Confirmatory factor analysis (CFA): Used to determine the factor and factor loading of measured variables, and to confirm what is expected on the basic or pre-established theory. CFA assumes that each factor is associated with a specified subset of measured variables. It commonly uses two approaches:

The traditional method: Traditional factor method is based on principal factor analysis method rather than common factor analysis. Traditional method allows the researcher to know more about insight factor loading.

The SEM approach: CFA is an alternative approach of factor analysis which can be done in SEM. In SEM, we will remove all straight arrows from the latent variable, and add only that arrow which has to observe the variable representing the covariance between every pair of latents. We will also leave the straight arrows error free and disturbance terms to their respective variables. If standardized error term in SEM is less than the absolute value two, then it is assumed good for that factor, and if it is more than two, it means that there is still some unexplained variance which can be explained by factor. Chi-square and a number of other goodness-of-fit indexes are used to test how well the model fits.

Eigen value (or latent root): When we take the sum of squared values of factor loadings relating to a factor, then such sum is referred to as Eigen Value or latent root. Eigen value indicates the relative importance of each factor in accounting for the particular set of variables being analysed.

Factor scores: Factor score represents the degree to which each respondent gets high scores on the group of items that load high on each factor. Factor scores can help explain what the factors mean. With such scores, several other multivariate analyses can be performed.We can now take up the important methods of factor analysis.

Univariate, Bivariate and Multivariate data and its analysis

When it comes to the level of analysis in statistics, there are three different analysis techniques that exist. These are –

- **Univariate analysis**
- **Bivariate analysis**
- **Multivariate analysis**

The selection of the data analysis technique is dependent on the number of variables, types of data and focus of the statistical inquiry. The following section describes the three different levels of data analysis –

1. Univariate data –

Univariate analysis is the most basic form of statistical data analysis technique. When the data contains only one variable and doesn't deal with a causes or effect relationships then a Univariate analysis technique is used. For instance, in a survey of a class room, the researcher may be looking to count the number of boys and girls. In this instance, the data would simply reflect the number, i.e. a single variable and its quantity as per the below table. The key objective of Univariate analysis is to simply describe the data to find patterns within the data. This is be done by looking into the mean, median, mode, dispersion, variance, range, standard deviation etc.

How Univariate analysis is conducted?

Univariate analysis is conducted through several ways which are mostly descriptive in nature

- Frequency Distribution Tables
- Histograms
- Frequency Polygons
- Pie Charts
- Bar Charts

2. Bivariate data –

Bivariate analysis is slightly more analytical than Univariate analysis. When the data set contains two variables and researchers aim to undertake comparisons between the two data set then Bivariate analysis is the right type of analysis technique. For example – in a survey of a classroom, the researcher may be looking to analysis the ratio of students who scored above 85% corresponding to their genders. In this case, there are two variables – gender = X (independent variable) and result = Y (dependent variable). A Bivariate analysis is will measure the correlations between the two variables as shown the table below.

How Bivariate analysis is conducted?

1. **Correlation coefficients**

Correlations is a statistical association technique where strength of relationship between two variables are observed. It shows the strength as strong or weak correlations and are rated on a scale of –1 to 1, where 1 is a perfect direct correlation, –1 is a perfect inverse correlation, and 0 is no correlation.

2. Regression analysis

Regression analysis is used for estimating the relationships between two different variables. It includes techniques for modelling and analysing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. It helps to understand how the value of the dependent variable changes when any one of the independent variables is changed. Regression analysis is used for advanced data modelling purposes like prediction and forecasting. There are a range of different regression techniques used depending on the nature of variable and the type of analysis sought by the research. For example –

- Linear regression
- Simple regression
- Polynomial regression
- General linear model
- Discrete choice
- Binomial regression
- Binary regression
- Logistic regression

3. Multivariate data

Multivariate analysis

Multivariate analysis is a more complex form of statistical analysis technique and used when there are more than two variables in the data set. Here is an example –

A doctor has collected data on cholesterol, blood pressure, and weight. She also collected data on the eating habits of the subjects (e.g., how many ounces of red meat, fish, dairy products, and chocolate consumed per week). She wants to investigate the relationship between the three measures of health and eating habits?

In this instance, a multivariate analysis would be required to understand the relationship of each variable with each other.

How Multivariate analysis is conducted

1. Commonly used multivariate analysis technique include
2. Factor Analysis
3. Cluster Analysis
4. Variance Analysis
5. Discriminant Analysis
6. Multidimensional Scaling
7. Principal Component Analysis
8. Redundancy Analysis

Key Terms to Remember

“**Variable** is a property that takes on different values"

Dependent variable is the variable that is affected by the independent variable. The

Independent variable is the antecedent while the dependent variable is the consequent

Factor is an underlying dimension of several related variables.

Factor loadings are the values that explain how closely the variables are related to each other.

Explanatory variable is causal or independent variable and is also called extragenous variable.

Criterion variable is resultant or dependent variable and is also called endogeneous variable.

Observable variable is directly observable.

Latent variable is unobservable variable which may influencecriterion variable.

Discrete variable takes only integer value when measured.

Continuous variable can assume any real value.

Cluster analysis is a technique of measuring some measure ofsimilarity.

Discriminant analysis is used primarily to identify variables that contribute to differences in the priory defined groups with the use of discriminant functions.

ANOVA is analysis of variance, where the comparison between means of samples drawn from some population having same mean values by testing the significance of difference between more than two sample means and inferences are made.

F-Ratio is the ratio between mean square between columns and mean square of residual.

Multiple Regression describes the relationship between two or more independent variables.

Important Questions

1. What Is Factor Analysis and How Does It Simplify Research Findings?
2. What are the types of Factor Analysis?
3. What statistical analysis method should I use with multivariate categorical data?
4. What are the types of Bivariate Analysis
5. Explain n the Application of Multivariate Analysis
6. In what situations is discriminant analysis used?

MCQ – Questions

1. A numerical value used as a summary measure for a sample, such as a sample mean, is known as a

   a) Population Parameter

   b) Sample Parameter

   c) Sample Statistic

   d) Population Mean

2. Statistics branches include

   a) Applied Statistics

   b) Mathematical Statistics

   c) Industry Statistics

   d) Both A and B

3. To enhance a procedure the control charts and procedures of descriptive statistics are classified into

    a) Behavioural Tools

    b) Serial Tools

    c) Industry Statistics

    d) Statistical Tools

4. Sample statistics are also represented as

    a) Lower Case Greek Letter

    b) Roman Letters

    c) Associated Roman Alphabets

    d) Upper Case Greek Letter

5. Individual respondents, focus groups, and panels of respondents are categorised as

    a) Primary Data Sources

    b) Secondary Data Sources

    c) Itemised Data Sources

    d) Pointed Data Sources

6. The variables whose calculation is done according to the weight, height and length and weight are known as:

    a) Flowchart Variables

    b) Discrete Variables

    c) Continuous Variables

    d) Measuring Variables

7. A method used to examine inflation rate anticipation, unemployment rate and capacity utilisation to produce products is classified as

    a) Data Exporting Technique

    b) Data Importing Technique

    c) Forecasting Technique

    d) Data Supplying Technique

8. Graphical and numerical methods are specialized processes utilised in

    a) Education Statistics

    b) Descriptive Statistics

    c) Business Statistics

    d) Social Statistics

9. The scale applied in statistics which imparts a difference of magnitude and proportions is considered as

    a) Exponential Scale

    b) Goodness Scale

    c) Ratio Scale

    d) Satisfactory Scale

10. Review of performance appraisal, labour turnover rates, planning of incentives and training programs and are examples of

    a) Statistics in Production

    b) Statistics in Marketing

    c) Statistics in Finance

    d) Statistics in Personnel Management