Prof. Amalendu Bhunia
Professor, Dept. of Commerce.
Dean, Faculty of Arts & Commerce,
University of Kalyani
Nadia, West Bengal, India
www.klyuniv.ac.in
bhunia.amalendu@gmail.com

Statistical Analysis of Survey Data

Amalendu Bhunia

Abstract:

This review observes the statistical analysis of survey-based quantitative data. Numerous statistical techniques including descriptive statistics, correlation analysis, and regression analysis are applied to examine survey data. The statistical tools are selected on the basis of research questions, research problems, selected variables, research hypotheses, scales of measurement, and research objectives. Statistical analysis is the method of collecting and organizing data to draw logical conclusions

1. Background :

The topic 'Statistics' is extensively used in more or less all the subjects. While undertaking research, the researchers should have a few consciousnesses in using the statistical tools that help them in drawing precise conclusions. The statistical analysis relies on the goal of the study. The goal of a survey is to get information about the condition of the population study. The first statistical assignment is then is to carry out a descriptive analysis of variables. In the descriptive analysis, it is obligatory to present outcomes acquired for each kind of variable. For qualitative and dichotomous variables, statistical results must be offered as frequencies and percentages. For quantitative variables, the appearance is central tendencies and variability. After descriptive analysis, the relationship between variables and predictive examination based on regression models is presented using SPSS software. Survey study is one of the most generally employed research techniques, scholars, researchers, and organizations of all magnitudes exercise surveys to assess public attitude. Researchers employ an extensive variety of statistical techniques to analyze survey data.

2. Statistical Analysis on Survey Data:

The statistical analysis basically depends on (a) research questions, (b) research problems including selected variables, (c) research hypotheses, (d) scales of measurement, and (e) research objectives.

A survey is an action that gathers information in a systematized and precise manner about attributes of interest from a few or all units of a population using definite perceptions, techniques, and methods and accumulates such information into a helpful summary shape. A survey generally starts with the requirement for information where no data or inadequate data exist. Usually, the researcher desires to study the traits of a population, make a database for diagnostic purposes or check a hypothesis. A survey can be reflected to involve some interrelated steps which comprise defining the objects, choosing a survey outline, shaping the sample design, creating the questionnaire, gathering and processing the data, analyzing and spreading the data, and filing the survey. The steps of a survey are discussed below:

(i) Formulation of the Research Question:

One of the vital jobs in a survey is to formulate the research question. A good research question structures spine of good research, which in turn is very important in extrication secrecies of character and providing insight into a problem. The research question makes out the problem to be considered and directs to the methodology. It guides to the strengthening of a suitable hypothesis. Therefore, the research question aims to search for an active ambiguity in an area of anxiety and involves a need for purposeful examination. A good research question helps prop up a purposeful debatable theory and building of a rational argument. In this area, the researchers generally describe the characteristics and patterns of the social phenomena, explain the causes for the characteristics of the particular phenomena and the behaviour of the individuals involved, and bring about change and the outcomes of change. Types of research questions should be (a) general research question and (b) specific research questions.

(ii) Formulation of the Research Problem:

Before collecting data, selecting a research method, and analyzing data, a researcher needs to formulate a specific research problem, which can be investigated by scientific procedures. Identifying a research problem is indicating a specific area for answering some research questions. A research problem is a report about a vicinity of worry, a situation to be developed, a complexity to be removed or a disturbing question that survives in academic literature, in theory, or in practice that involves the need for significant understanding and purposeful examination. In social science disciplines, the research problem is normally created in the shape of a question. A research problem does not declare how to do somewhat, propose an unclear or broad proposal, or give a value question. In this section, the researchers have to answer the following questions-

- What is the purpose of the study?
- What is or how much is already known?
- Is extra information required?
- What is to be explored?
- What are the variables?
- How is it to be measured?
- The data which will be collected i.e., will be respondents give correct information?
- Is the time is appropriate for undertaking the research?
- Can a hypothesis be formulated?

(iii) Formulation of Hypothesis:

A hypothesis may be a statement of expectation or prediction that may be tested by research. Before formulating the hypothesis, the researchers have to study existing literature rigorously (articles, books, and/or cases) wherever accurately the researcher will build a fresh contribution. The researchers ought to focus on the subsequent kinds of hypotheses before finalizing.

(a) Null hypothesis: This presumes a status quo, no significant change. This is a statistical hypothesis that the researcher seeks to reject.

(b) Alternative hypothesis: This is the experimental hypothesis. It can either be directional or non-directional. It is the hypothesis that the researcher seeks to support.

Directional hypothesis shows the direction of the relationship between independent and dependent variables. The non-directional hypothesis shows the existence of a relationship between variables but no direction is specified.

(iv) Scales of Measurement:

Measurement is the numbers or other symbols that are used as characteristics of objects. Scaling is basically an addition of measurement. Scaling is the process by which respondents would be classified as having a negative, neutral or positive attitude. Nominal, ordinal, interval, and ratio are the scales of measurement. A criterion for good measurement includes reliability, validity, and sensitivity. Reliability is the degree on the way to reduce the measurement error. Validity is the ability of a scale or measuring instrument to measure what it is intended to measure Sensitivity is the ability of a measurement instrument to accurately measure variability in stimuli or responses. There are various types of scaling techniques. The researcher has to think about when to use which type of scaling method. Various scaling techniques are Graphic Rating Scale, Numerical Rating Scale, Descriptive Rating Scale, Comparative Rating Scale, Likert Scale, Semantic Differential Scale, Stapel's Scale, Bogardus Scale, Multi-Dimensional Scaling, Thurston Scales, Guttman Scales, the Q-Sort Technique, Behavior Intention Scale, and Rank Order Scale, etc.

(v) Formulation of Research Objectives:

For developing the research objectives, the researcher has to consider the research gap, research questions, and the research problem. It should be non-biased verbs and should not be biased verbs. The researcher ensures that objectives flow logically from the statement of need and address the problem, make objectives fall within the range of results, which are expected to be achieved, and the objectives should.be hierarchical. The objectives should be general objectives and specific objectives. General objective states what researchers expect to achieve by the study in general terms and specific objectives are the smaller, logically connected parts of the general objective.

3. Data Analysis:

It is the process of collecting and organizing data in order to draw helpful logical conclusions from it. There are various types of data analysis. It includes
Descriptive analysis: What is happening (positive or negative), figure out.
Diagnostic analysis: What is happening to figure out the reason?
Predictive analysis: What is likely to happen in the future?
Causal analysis: Find out the cause and effect.
Prescriptive analysis: Forms a plan of action based on the above analysis.

4. Statistical Tests:

(a) Reliability and Validity:

Reliability and validity are the two vital properties that the data can have. They are frequently stated jointly, but they provide us diverse types of information. Reliability informs us how consistently the data evaluate something. Validity informs whether the data are determining the right things for an exacting employ of the test.

Cronbach's Alpha is used for internal consistency/reliability when you have multiple Likert scale questions in a survey/questionnaire. If you have a Likert scale or other types of items, use the Spearman-Brown formula/Cronbach's Alpha. If you have dichotomous items (e.g., yes-no answers) as you would with multiple-choice exams, the KR-20 formula is the best-accepted statistic. Rule of thumb of Cronbach's Alpha: 0.70 and above is good, 0.80 and above is better, and 0.90 and above is best.

When researchers determine a construct that they supposed to be reliable across time, then the data they acquire should also be reliable across time. Test-retest reliability is the degree to which this is really the case. An assessment of the test-retest reliability needs to use the measure on a group of people. This is usually completed by graphing the data in a scatter plot and calculating Pearson's *r*.

Cohen's kappa statistic measures interrater reliability. Interrater reliability or precision happens when your data collectors give the same score to the same data item. This statistic should only be calculated when two collectors each rate one trial on each sample, or one collector rates two trials on each sample. Fleiss's Kappa statistic should only be calculated when multiple collectors each rate one trial on each sample. Rule of thumb of interrater reliability: 0.41 – 0.60 = moderate agreement, 0.61 – 0.80 = substantial agreement, 0.81 – 0.99 = near-perfect agreement and 1 = perfect agreement.

Internal validity measures the appropriateness of the research methods. Internal validity is buoyancy where the causal relationships between variables are independent. Internal validity may be checked through regression analysis or any other causality analysis between dependent and independent variables.

(b) Descriptive Statistics:

Descriptive statistics is the process of categorizing and describing the data or information in a study. It is a technique to collect, summarizes, organizes, exhibit, and analyze sample data obtained from a population. It is a collection of tools or techniques, which quantitatively describes the data in summary and graphical shapes.

Mostly descriptive statistics is applied to describe the behaviour of sample data. Because, in a study, there are several variables that are to be calculated or measured, therefore descriptive statistics are used to split this enormous quantity of data into the easiest structure. For instance, it would be attractive to observe the average sales made by a Salesman in a company. At this moment, as there are reasonably numerous activities in one day, therefore we can exercise descriptive statistics to create that simpler. Descriptive statistics can be calculated either by method of central tendency or variability or both.

Descriptive statistics are the suitable analyses when the researcher tackles the following research questions.

1) Has there been a significant change in the mean sales price of a company?
2) Has there been an increase in the number of sales units in a company?

The above research questions are distinctive, which need statistical analysis for the answers. To answer these questions, a good number of samples should be obtained from the target population. Then the researchers have to use descriptive statistics to classify and abridge the sample data. There are three most important types of descriptive statistics. These are (a) the distribution concerns the frequency of each value, (b) the central tendency concerns the averages of the values, and (c) the variability or dispersion concerns how to spread out the values are. Again, on the basis of the number of variables, descriptive statistics are two types. These are (a) univariate descriptive statistics and (b) bivariate/multivariate descriptive statistics. The attention of univariate descriptive statistics is on one variable. It’s significant to check data from each variable independently using numerous measures of distribution, central tendency, and spread. In bivariate analysis, the researcher concurrently studies the frequency and variability of two variables to observe if they fluctuate jointly. The researcher can also compare the central tendency of the two variables before performing additional statistical tests. Multivariate analysis is similar to bivariate analysis however with more than two variables.

(i) Frequency percentages are a mainly helpful technique of stating the relative frequency of survey answers and other data. A percentage frequency distribution is an exhibit of data that indicates the percentage of observations that survive for each data point or grouping of data points. Percentage frequency distributions are also displayed as bar graphs or pie charts.

(ii) The arithmetic means or the average is the sum of the observations divided by the number of observations. When the data is interval and/or ratio data, the arithmetic mean is used to describe the current status of the variable. When the data is ordinal data, the median is used to describe the current status of the variable. When the data is nominal data, the modal value is used to describe the current status of the variable.

(iii) The researcher employs a statistic of dispersion to provide a single number that explains how compact or widely spread a set of observations is. A statistic of dispersion informs how widely spread a set of measurements is. Standard deviation is the most familiar.

When the data is interval and/or ratio data, the standard deviation and coefficient of variation are used. The standard deviation has useful statistical properties that create it the basis of many statistical tests. The coefficient of variation is helpful when comparing the amount of variation for one variable among groups with unlike means, or among unlike measurement variables. When the data is ordinal data, the inter-quartile range is used for the study of variability. When the data is ordinal data, an index of qualitative variation is used for the study of variability.

(c) Correlation and Regression Analysis or Further Tests:

(i) Check Linearity:

Linearity means that two variables are related by a mathematical equation. The importance of testing for linearity lies in the fact that many statistical techniques need an assumption of the linearity of data. The linearity of the variable is checked through a scatter plot or F-test. If the probability of the F-test of deviation from linearity based on the means test is more than 0.05, then the relationship between the independent variables is linearly dependent. Checking linearity is helpful for Pearson correlation analysis and linear regression analysis.

(ii) Check Homoscedasticity:

Homoscedasticity means whether the residuals (the residuals are simply the error terms or the differences between the observed value of the dependent variable and the predicted value) are equally distributed. It is checked through Levene’s statistic or Breush-Pagan test or Non-Constant Variance Test. If the p-value is more than 0.05, therefore we cannot reject the null hypothesis that the variance of the residuals is constant and infer that heteroscedasticity is not present. Checking homoscedasticity is helpful for Pearson correlation analysis and linear regression analysis.

(iii) Check Normality:

The normal distribution is the leading probability distribution in statistics as it fits lots of usual facts. The normal distribution is a probability function that explains how the values of a variable are distributed. When the data is interval and/or ratio data, normality of the data should be checked using Skewness and Kurtosis Test, Kolmogorov-Smirnov Test, Shapiro-Wilk Test, D'Agostino Test, Anderson-darling Test, Jarque-Bera Test, Q-Q plots. If the interval or ordinal data is normal, the researcher can use a parametric test of the hypothesis. When the data is ordinal or nominal, there is no need to check the normality and the researcher can use a non-parametric test of hypothesis.

After checking the above three tests, the researcher can use correlation analysis to know the strength and direction of the relationship between two variables. If the data is ratio or interval data, Pearson correlation is used. If your data is ordinal data, Spearman’s rank correlation is used. If your data is nominal data, the chi-square test is used. Kendall's tau-b (τb) correlation coefficient is a nonparametric measure of the strength and direction of association that exists between two variables measured on at least an ordinal scale.

5. Conclusions:

The aim of this topic has been to talk about subjects engaged in the study of survey data. These topics include the exercise of both descriptive and logical approaches to survey data. This issue has proposed a few statistical methods that can provide helpful tools for survey data analysis. I propose that survey researchers critically think about these techniques, suitable to survey objects, in survey data analysis, with the object of taking out as much information as possible from obtained survey (quantitative) data.