EDUCATIONAL STATISTICS

B.ED (1.5-YEARS)

Course Code: (8614) **Units: 1–9**

Faculty of Education

Early Childhood Education and Elementary Teacher Education Department

ALLAMA IQBAL OPEN UNIVERSITY, ISLAMABAD

1st Edition... 2017

Quantity 20000

Composed by .... Zahid Javed Bajwa

Printer ....

Publisher .... Allama Iqbal Open University, Islamabad.

COURSE TEAM

Chairman:

Prof Dr Nasir Mahmood
Chairman
Early Childhood Education and Elementary Teacher
Education Department

Course Development:

Coordinator:

Prof. Dr. Nasir Mahmood
Professor
Allama Iqbal Open University, Islamabad

Memebers:

1. Prof. Dr. Nasir Mahmood
   Professor
   Allama Iqbal Open University, Islamabad
2. Aftab Ahmad Khan
   Government High School Aamgah Haripur
3. Miss Sumbal Asghar
   Beacon House Education System
4. Salman Khalil Chaudhary
   Allama Iqbal Open University, Islamabad

Reviewers:

1. Dr. Rizwan Akram Rana-University of Punjab
2. Dr. Raana Malik-University of Punjab

Editor:

1. Fazal Karim

Layout & Design:

Malik Mateen Ishfaq
Word Processor Operator, PPU
Allama Iqbal Open University

ACKNOWLEDGEMENTS

Statistics is of vital importance in educational research. It enters in the process right from the beginning of the research when whole plan for the research, selection of design, population, sample, analysis tools and techniques etc., is prepared. Planning and carrying out research project and trying to analyze data without using statistical techniques will carry away from the objectives of the study. In this book, efforts have been made to provide updated knowledge and relevant skills to prospective Teachers.

This book reflects the efforts of many people. The Department of Early Childhood Education and Elementary Teacher Education is grateful to all the members of the course team for writing units and giving comments and feedback to improve the materials.

It is worth mentioning that course coordinator has very skillfully mediated over the whole process of course development, kept the course team engaged and adhered to timeline. The composing and formatting of the scripts was very laborious work requiring extend hours of sitting. Mr. Zahid Javed Bajwa has very efficiently completed the task in short time.

Thanks to our Vice-Chancellor Prof. Dr. Shahid Siddiqui for his vision, support, guidance and motivating work environment for all of us to focus on academic excellence at university.

Prof. Dr. Nasir Mahmood
Chairman/Dean
Faculty of Education

INTRODUCTION

Statistics is of vital importance in vast variety of fields. Particularly it is invaluable for the field of research. In research and particularly in educational research following questions cannot be answered without the use of proper statistical techniques.

- What kind and how much data we need to collect?
- How should we organize and summarize the data?
- How can we analyze the data and draw conclusion from it?
- How can we assess the strength of the conclusion and evaluate their uncertainty?

Owing to the importance, this course is included for prospective B Ed. graduates. The very first unit of the course introduces, its characteristics, functions, its importance and limitations and its application in educational research. Basic overview of descriptive and inferential statistics, variables and its types, scientific method and notation used in the subject is also given in this unit. Unit 2 explains some basic concepts like variable, data, population sample. Unit 3 elaborate graphical representation or exploratory data analysis techniques. Unit 4 highlights some basic techniques of measures of dispersion like range, mean deviation, variance and standard deviation, and measures of shape like skewness and kurtosis. Measures of central tendency like mean, median and mode are described in unit 5. Unit 6 deals with inferential statistics, its logic and importance in educational research. Hypothesis testing, its logic, errors in hypothesis testing and *t*-test and its types are also discussed in this unit. Correlation along with Pearson and Spearman correlation method and regression and its types are discussed in unit 7. Unit 8 deals with ANOVA, logic behind using ANOVA, F-distribution, one-way ANOVA and multiple comparison procedures. Chi-square (χ^2) distribution, its uses and types are discussed in unit 9.

Prof. Dr. Nasir Mahmood
Course Development Coordinator/Program Coordinator

CONTENTS

OBJECTIVES

After completion of this course the students will be able to:

1. demonstrate basic understanding of statistics.

2. explain the application of statistics in educational research.

3. distinguish and between descriptive and inferential statistics.

4. distinguish between the levels of measurement.

5. explain variable and data and their types.

6. explain population sample and their types.

7. demonstrate the basic understanding of graphical representation of data.

8. tell the basic purpose of measure of central tendency, measures of dispersion, and numerical measures of shape.

9. explain and use inferential techniques like *t*-test and ANOVA.

10. explain correlation, regression and their types.

11. explain chi-square (χ^2) distribution, its uses and types.

UNIT-1

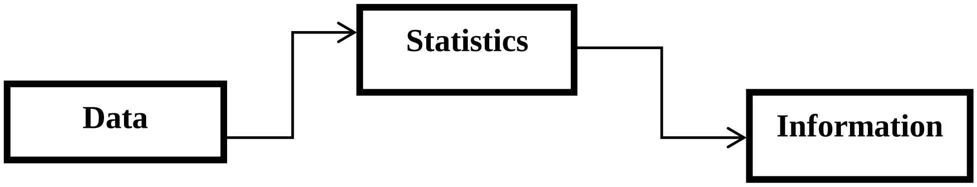
INTRODUCTION TO STATISTICS

Written By:
Aftab Ahmad

Reviewed By:
Dr. Rizwan Akram Rana

Introduction

Statistics is a broad subject with applications in vast variety of fields. The word “statistics” is derived from the Latin word “Status”, which means a political state. Statistics is a branch of knowledge that deals with facts and figures. The term statistics refers to a set of methods and rules for organizing, summarizing, and interpreting information. It is a way of getting information from data.



We can say that Statistics is a science of collecting, organizing, interpreting and reporting data. It is a group of methods which are used for collecting, displaying, analyzing, and drawing conclusions from the data.

In other words, statistics is a methodology which a researcher uses for collecting and interpreting data and drawing conclusion from collected data (Anderson & Sclove, 1974; Agresti & Finlay, 1997).

Statistical data can be used to answer the questions like:

- What kind and how much data we need to collect?
- How should we organize and summarize the data?
- How can we analyze the data and draw conclusion from it?
- How can we assess the strength of the conclusion and evaluate their uncertainty?

Above discussion lead us to the conclusion that statistics provides methods for:

i) Design: Planning and carrying out research studies.
ii) Description: Summarizing and exploring data.
iii) Inferences: Making predictions and generalization about phenomena represented by the data.

Objectives of Unit

After reading this unit the students will be able to:

12. demonstrate basic understanding of statistics.
13. know the characteristics of statistics.
14. explain the functions of statistics.
15. Enlist the characteristics of statistics.
16. tell the importance and limitations of statistics.
17. briefly explain the application of statistics in educational research.
18. distinguish between descriptive and inferential statistics.
19. describe variables and its types.
20. distinguish between the levels of measurement.
21. identify various statistical notations.

1.1 Functions of Statistics

Functions of Statistics are summarized under following headings.

i) To present facts in a definite form

Daily we encounter millions of pieces of information which are often vague, indefinite and unclear. When such pieces of information undergo certain statistical techniques and are represented in the form of tables or figures, they represent things in a perspective which is easy to comprehend. For example, when we say that some students out of 1000 who appeared for B. Ed examination were declared successful. This statement is not giving as much information. But when we say that 900 students out of 1000 who appeared for B. Ed examination were declared successful; and after using certain statistical techniques we conclude that “90% of

B. Ed. students were successful"; now the sentence becomes more clear and meaningful.

ii) To simplify unmanageable and complex data

In our daily life and in research also, we often get large amount of information. To get a clear picture, statistics helps us either by simplifying such information by taking few figures to serve as a representative sample or by taking average to give a bird's eye view of the large masses. Complex data may be simplified by presenting them in the form of a tables, graphs or diagrams, or representing it through an average etc.

iii) To use techniques for making comparisons

Often in research things become more clear and significant when they are compared with others of the same type. The comparison between two different groups is courtesy of certain statistical techniques, such as average, coefficients, rates, ratios, etc.

iv) To enlarge individual experience

As an individual our knowledge is limited to what we can observe and see; and that is a very small part of the ocean of knowledge. Statistics extends our knowledge and experiences by presenting various conclusions and results, based on numerical investigations. For example, we daily listen and also have general impression that the cost of living has increased. But to know to what extent the increase has occurred, and how far the rise in prices have affected different income groups, it would be necessary to have a comparison of the rise in prices of articles consumed.

v) To provide guidance in the formulation of policies

Statistics enable us to make correct decisions, whether they are taken by a businessman or government. In fact statistics is a great servant of business in management, government. Statistical methods are employed in industry in tackling the problem of standardization of products. Large industries maintain a separate department for statistical intelligence or statistical bureau, the work of which is to collect, compare and coordinate figures for formulating future policies of the firm regarding production and sales.

vi) To enable measurement of the magnitude of a phenomenon

Statistics enables us to measure the magnitude of a phenomenon under investigation. Estimate of the population of a country or the quantity of wheat, rice and other agricultural commodities produced in the country during any year are examples of such phenomena.

1.2 Characteristics of Statistics

Following are the characteristics of statistics.

i) Statistics consists of aggregate facts

The facts which can be studied in relation to time, place or frequency can be called statistics. A single isolated and unconnected fact or figure is not statistics because we cannot study it in relation to other facts and figures. Only aggregate of facts e.g. academic achievement of the students, I.Q. of a group of students, weight of students in a class, profit of a firm etc. are called statistics.

ii) Multiple causes affect Statistics

A phenomena may be affected by so many factors. We cannot study the effects of one factor on the phenomena only by ignoring others. To have a true picture we will have to study the effects of all factors on the phenomena separately as well as collectively, because effects of the factors can change with change of place, time or situation. For example, we can say that result of class X in board examination does not depend on any single factor but collectively on standard of teachers, teaching methods, teaching aids, practical's performance of students, standard of question papers, environment of the examination hall, exam supervisory staff and standard of evaluation of answers after the examination.

iii) Data should be numerically expressed, enumerated of estimated

Data to be called statistics should be numerically expressed so that counting or measurement of data can be made possible. It means that the data or the fact must be in quantitative form as achievement scores 60, 50, 85, 78, and 91 out of 100. If it is not in quantitative form it should be quantified.

iv) Statistics are enumerated or estimated according to reasonable standard of accuracy

For a clear picture of the phenomena under investigation, it should be researched using reasonable standard of accuracy depending upon the nature and purpose of collection of data. Data collection should be free from personal prejudices and biases. Biased and personally prejudiced data leads to inaccurate conclusion.

v) Statistics are collected in a Systematic Manner

In order to have reasonable standard of accuracy statistics/data must be collected in a very systematic manner. Any rough and haphazard method of collection will not be desirable for that may lead to improper and wrong conclusion.

vi) Statistics for a Pre-determined Purpose

Before collection of data**,** investigator/researcher must have a purpose and then should collect data accordingly. Data collected without any purpose is of no use. Suppose we want to know intelligence of a section of people, we must collect data relating to I.O. level and data relating to income, attitude and interest level of that group of people will be of no use. Without having a clear idea about the purpose we will not be in a position to distinguish between necessary data and unnecessary data or relevant data and irrelevant data.

vii) Statistics are Capable of being placed in Relation to each other

Statistics is a method for the purpose of comparison etc. It must be capable of being compared; otherwise, it will lose much of its significance. Comparison can be made only if the data are homogeneous. Data on memory test can be compared with I.Q. It is with the use of comparison only that we can illustrate changes which may relate to time, place, frequency or any other character, and statistical devices are used for this purpose.

1.3 Importance and Scope of Statistics

Statistics is important in our daily life. We live in the information world and much of this information is determined mathematically with the help of statistics. It means statistics keeps us informed about day to day happening. Importance of statistics in our daily life is discussed under following headings.

i) Every day we watch weather forecasting. It is possible due to some computer models based on statistical concepts. These models compare prior weather with the current weather and predict future weather.

ii) Statistics is frequently used by the researchers. They use statistical techniques to collect relevant data. Otherwise there may be loss of money, time and other resources.

iii) In business market statistics play a greater role. Statistical techniques are the key of how traders and businessmen invest and make money. Also, in industry, these tools are used in quality testing. Production managers are always interested to find out whether the product is confirming the specification or not. He uses statistical tools like inspection plan, control chart etc.

iv) Statistics also has a big role in the medical field. Before any drugs prescribed, pharmacists show statistically valid rate of effectiveness. Similarly statistics is behind all other medical studies. Doctors predict diseases on the bases of statistical concepts.

v) Print and electronic media use statistical tools to make predictions of winner of elections and coming government.

vi) Statistics has widely been used in psychology and education to determine the reliability and validity to a test, factor analysis etc.

vii) Apart from above statistics has a wide application in marketing, production, finance, banking, investment, purchase, accounting and management control.

1.4 Limitations of Statistics

The science of Statistics has following limitations:

i) The use of statistics is limited to numerical studies

We cannot apply statistical techniques to all type of phenomena. These techniques can only be applied to the phenomena that are capable of being quantitatively measured and numerically expressed. For example, the health, intelligence, honesty, efficacy etc. cannot be quantitatively measured, and thus are unsuitable for statistical study. In order to apply statistical techniques to these constructs, first we will have to quantify them.

ii) Statistical techniques deal with population or aggregate of individuals rather than with individuals

For example, when we say that the average height of a Pakistani is 1 meter and 80 centimeters, we mean to shows the height not of an individual but as found by the study of all individuals living in Pakistan.

iii) Statistics relies on estimation and approximations

Statistical techniques are not exact laws like mathematical or chemical laws. They are derived by taking a majority of cases and are not true for every individual. Thus the statistical inferences are uncertain.

iv) Statistical results might lead to fallacious conclusions

Statistical results are represented by figures, which are liable to be manipulated. Also the data placed in the hands of an expert may lead to fallacious results because figures may be stated without their context or may be applied to a fact other than the one to which they really relate. An interesting example is a survey made some years ago which reported that 33% of all the girl students at John Hopkins University had married University teachers. Whereas the University had only three girls student at that time and one of them married to a teacher.

1.5 Application of Statistics in Educational Research

Statistics is of vital importance in educational research. It does not include measurement of problems such as construction of indices or the scoring of items on a questionnaire. Rather, it involves a manipulation of numbers under the assumption that certain requirements have been met in the measurement procedure. Statistics practically seems to work at the analysis stage of the research process when data have been collected. It does not mean that social scientists can plan and carry out entire research projects without any knowledge of statistics. Planning and carrying out research project and trying to analyze data without using statistical techniques will carry away from the objectives of the study.

Statistics enters in the process right from the beginning of the research when whole plan for the research, selection of design, population, sample, analysis tools and techniques etc., is prepared.

1.6 Descriptive and Inferential Statistics

Researchers use a variety of statistical procedures to organize and interpret data. These procedures can be classified into two categories – *Descriptive Statistics* and *Inferential Statistics*. The starting point for dealing with a collection of data is to organize, display, and summarize it effectively. It is the major objective of descriptive statistics. Descriptive Statistics, as the name implies, describes the data. Descriptive statistics consist of methods for organizing and summarizing information. These are statistical procedures that are used to organize, summarize, and simplify data. In these techniques raw scores are taken and undergone some statistical techniques to obtain more manageable form. These techniques allow the researcher to describe large amount of information or scores in a few indices such as mean, median, standard deviation etc. When these indices are calculated for a sample, they are called statistics; and when they are calculated from entire population, they are called parameters (Fraenkel, Wallen, & Hyun, 2012). Descriptive statistics organizes scores in the form of a table or a graph. It is especially useful when the researcher finds it necessary to handle interrelationship among more than two variables.

Only summarizing and organizing data is not the whole purpose of a researcher. He often wishes to make inferences about a population based on data he has obtained from a sample. For this purpose, he uses inferential statistics. Inferential Statistics are techniques that allow a researcher to study samples and then make generalizations about the populations from which they are selected.

Population of a research study is typically too large and it is difficult for a researcher to observe each individual. Therefore a sample is selected. By analyzing the results obtained from a sample, a researcher hopes to make general conclusion about the population. One problem with using sample is that a sample provides only limited information about the population. To address this problem is the notion that the sample should be *representative* of the population. That is, the general characteristics of the sample should be consistent with the characteristics of the population.

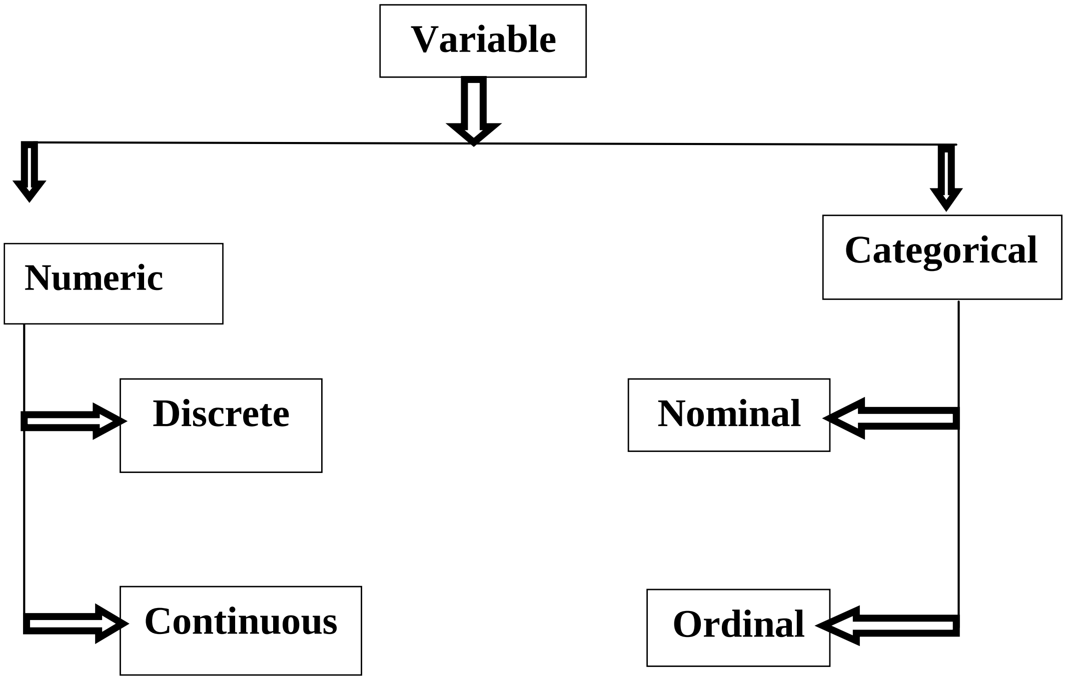
1.7 Variable

A variable is something that is likely to vary or something that is subject to variation. We can also say that a variable is a quantity that can assume any of a set of values. In other words, we can say that a variable is a characteristic that varies from one person or thing to another. It is a characteristic, number or quantity that increases or decreases over time or takes different value in different situations; or in more precise words, it is a condition or quality that can differ from one case to another. We often measure or count it. A variable may also be called a data item. Examples of variables for human are height, weight, age, number of siblings, business income and expenses, country of birth, capital expenditure, marital status, eye color, gender, class grades, and vehicle type, etc.

The variables that yield numerical information/measurement are called quantitative or numerical variable and the variable that yield non-numerical information or measurement are called qualitative or categorical variable. In the above example, first seven are the examples of quantitative variable and last five are the examples of categorical variables.

Quantitative variables can further be classified as either discrete or continuous. A discrete variable consists of separate, indivisible categories/values. No values can exist between two neighboring categories/values – for example, seven dots or eight dots – no other value can be observed in between them. These variables are commonly restricted to whole countable numbers – for example, the number of children in a family or the number of students attending the class. If anyone observes a class attending from day to day, he may find 30 students one day and 29 students the next day. A discrete variable may also consist of observations that differ qualitatively. For example, a psychologist observing patients may classify some as having panic disorders, others as having dissociative disorders, and some as having psychotic disorders. The type of disorder is a discrete variable because there are distinct and finite categories that can be observed.

On the other hand, variables such as time, height, and weight are not limited to a fixed set of separate, indivisible categories. They are divisible in an infinite number of fractional parts. Such variables are called continuous variables. For example, a researcher is measuring the amount of time required to solve a particular mental arithmetic problem. He can measure time in hours, minutes, seconds, or fractions of seconds



.

1.8 Level of Measurement

There are two basic types of variables – quantitative and categorical. Each uses different type of analysis and measurement, requiring the use of different type of measurement scale. A scale of a variable gives certain structure to the variable and also defines the meaning of the variable. There are four types of measurement scales: nominal, ordinal, interval, and ratio.

Nominal Scale

A nominal scale is the simplest form of measurement researchers can use. The word nominal means "having to do with names." Measurements made on this scale involve merely naming things. It consists of a set of categories that have different names. Such measurements label and categorize observations but do not make quantitative distinctions between them. For example, if we wish to know the sex of a person responding to the questionnaire, we would measure it on nominal scale consisting of two categories (male or female). A researcher observing the behavior of a group of infant monkeys might categorize responses as playing, grooming, feeding, acting aggressively or showing submissiveness. As the researcher merely gives names to each category so, this is a nominal scale of measurement. The nominal scale consists of qualitative distinctions.

Although, a nominal scale consists of qualitative differences, yet it does not provide any information about quantitative differences between individuals. Numerical values like 0 and 1 are merely used as code for nominal categories when entering data into computer programs.

Ordinal Scale

In ordinal scale of measurement, the categories that make up the scale not only have separate names but also are ranked in terms of magnitude. This scale consists of a set of categories that are organized in an ordered sequence. For example, a manager of a company is asked to rank employees in term of how well they perform their duties. The collected data will tell us who the manager considers the best worker, the second best, and so on. The data may reveal that the worker, who is ranked second, is viewed as doing better work than the worker who is ranked third. However, we can get no information about the amount that the workers differ in job performance, i.e. we cannot get the answer of the question "How much better?" Thus, an ordinal scale provides us information about the direction of difference between two measurements, but it does not reveal the magnitude of the difference.

Interval Scale

An interval scale possesses all the characteristics of an ordinal scale, with additional feature that the categories form a series of intervals that are exactly of the same size. This additional information makes it possible to compute distances between values on an interval scale. For example, on a ruler 1-inch interval is the same size at every location on the ruler. Similarly 4-inch distance is exactly the same size no matter where it is measured on the ruler. Similarly, the distance between the scores of 70 and 80 is considered to be the same as the distance between scores of 80 and 90. For all practical purposes these numbers can undergo arithmetic operations to be transformed into meaningful results. Interval scale answers the question "How much better?" or "How much is the difference?" But there is no intrinsic zero, or starting point. The zero point on the interval scale does not indicate a total absence of what is being measured. For example, 0^o (zero degree) on the Celsius or Fahrenheit scale does not indicate no temperature.

Ratio Scale

A ratio scale has all the characteristics of an interval scale but adds an absolute zero point. It means on a ratio scale a value of zero indicates complete absence of the variable being measured. Advantage of absolute zero is that a ratio of numbers on scale reflects ratio of magnitude for the variable being measured. We can say that one measurement is three times larger than another, or one score is only half as large as another. Thus, ratio scale not only enables us to measure the difference between two individuals, but also to describe the difference in terms of ratios.

1.9 The Scientific Method

There are many disciplines ranging from medicine and astrophysics to agriculture, zoology and social sciences, where scientists a process called scientific method is used to advance their knowledge and understanding.

Scientific method is a process for explaining the world we see. It is a process used to validate observations while minimizing observer bias. This method is a series of steps that lead to answers that accurately describe the things we observe. Its goal is to conduct research in a fair, unbiased and repeatable manner.

Scientific method is a tool for: (a) forming and framing questions, (b) collecting information to answer those questions, and (c) revising old and developing new questions.

The scientific method is not the only way, but the best-known way to discover how and why the world works. It is not a formula. It is a process with a manner of sequential steps designed to create an explainable outcome that increases our knowledge base. The process is as follows:

i) Ask a question

Asking a question is the first step of scientific method. Good questions come from careful observations. Our senses are a good source of observation. Sometime certain instruments like a microscope or a telescope are also used. These instruments extend the range of senses. During the observation many questions come in the mind. These questions derive the scientific method.

ii) Define the Problem

The question raised during the observation led to state a problem.

iii) Forming a Hypothesis

A hypothesis is a clear statement of what one expect to be the answer of the question. A hypothesis represents the best educated guess based on the one's observation and what he already knows. A good hypothesis is testable. It provides some specifics that lead to method of testing. The hypothesis can also lead to predictions.

iv) Conducting the Experiment / Testing the Hypothesis

After forming the hypothesis, it is tested. There are different methods to test a hypothesis. The most familiar method is to conduct an experiment.

v) Analyzing the Results

After the experiment (or whatever method is used to test a hypothesis), all information, that are gathered, are analyzed. Tables and graphs are used in this step to organize the data.

vi) Drawing Conclusions

On the basis of analysis, it is concluded whether or not the results support the hypothesis. If, in case, hypothesis is not supported by the data, the researcher checks for errors. Sometime he may have to reject the hypothesis and make a new one.

vii) Communicate the Results

After any scientific investigation, results should be communicated to let others know the new piece of knowledge.

1.10 Statistical Notations

Commonly used statistical notations are given in the following table.

Sr. No	**Notation/ Symbol**	**Used for**
1	P	Population proportion
2	p	Sample proportion
3	X	Set of population elements
4	x	Set of sample elements
5	N	Population size (Number of elements in the population)
6	n	Sample size (Number of elements in the sample)
7	µ (mew)	Population mean
8	x	Sample mean
9	σ (Sigma)	Standard deviation of the population
10	s	Standard deviation of the sample
11	σ^2	Variance of the population
12	s^2	Variance of the sample
13	ρ or ϱ (rho)	Population correlation coefficient based on all the elements of the population (Spearman’s rank order correlation)
14	r	Sample correlation coefficient based on all the elements of the sample
15	B_0	The intercept constant in a population regression line
16	b_0	The intercept constant in a sample regression line
17	B_1	The regression coefficient (the slope)in a population regression line
18	b_1	The regression coefficient (the slope)in a sample regression line
19	R^2	Coefficient of determination
20	s_{b1}	Standard error of the slope of a regression line
21	H_0	Null hypothesis
22	H_1 or H_a	Alternate hypothesis
23	*p*	Probability value
24	α (alpha)	Level of significance
25	β (beta)	Probability of committing a Type II error
26	Z or z	Standardized score or z-score
27	∑	Summation, used to compute sum over a range of values
28	∑X	Sum of a set of n observations. Thus $\sum X = X_1 + X_2 + X_3 + \ldots + X_n$
29	χ^2	Chi-square statistics
30	Var(X)	Variance of random variable X
31	SD(X)	Standard deviation of random variable X
32	M	Mean of the sample
33	SE	Standard error of a statistic

34	ME	Margin of error
35	DF or Df	Degree of freedom
36	Q_1	Lower/first quartile (25% of population are below this value)
37	Q_2	Median/second quartile (50% of population are below this value, also median of the sample)
38	Q_3	Upper/third quartile (75% of population are below this value)
39	IQR	Inter-quartile range ($Q_3 – Q_1$)
40	X~	Distribution of random variable X
41	$N (\mu,\sigma^2)$	Normal distribution / Gaussian distribution
42	U (a, b)	Uniform distribution (equal probability in range a, b)
43	gamma (c, λ)	Gamma distribution
44	χ^2 (k)	Chi-square distribution
45	Bin (n, p)	Binomial distribution
46	F (k_1, k_2)	F distribution
47	Poisson (λ)	Poisson distribution

1.11 Self-Assessment Questions

Q. 1 What do you understand by statistics?
Q.2 What are the characteristics of statistics?
Q. 3 Explain the functions of statistics.
Q. 4 Write down the characteristics of statistics.
Q. 5 Why is statistics important for educational research? Also state its limitations.
Q. 6 How will you apply statistics in educational research?
Q. 7 How will you distinguish descriptive statistics from inferential statistics?
Q. 8 What is a variable? Also write its types.
Q. 9 Briefly state the levels of measurement.

1.12 Activities

1. Diagrammatically show how “data” becomes “information”.
2. Make a list of the questions that can be answered using statistics.
3. Make a list of the “functions of statistics”.
4. Think and write down any two characteristics not given in the unit.
5. Make a diagram to show the types of variables.
6. Draw a hierarchy of levels of measurement.
7. Make a list of the steps of scientific method.

1.13 Bibliography

Agresti, A. & Finlay, B. (1997). *Statistical Methods for Social Sciences*, (3rd *Ed.*). Prentice Hall.

Anderson, T. W., & Sclove, S. L. (1974). *Introductory Statistical Analysis*, Finland: Houghton Mifflin Company.

Dietz, T., and Kalof, L. (2009). *Introduction to Social Statistics*. UK: Wiley-Blackwell

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to Design and Evaluate in Education*. (8th Ed.) McGraw-Hill, New York

Gravetter, F. J., & Wallnau, L. B. (2002). *Essentials of Statistics for the Behavioral Sciences (4th Ed.)*. Wadsworth, California, USA.

UNIT-2

BASIC STATISTICAL CONCEPTS

Written By:
Aftab Ahmad

Reviewed By:
Dr. Rizwan Akram Rana

Introduction

In this unit you will study some basic concepts like variable, data, population, and sample. Types of variable, types of data, types of population and types of sample are also discussed. The purpose of this unit is to give an awareness of these commonly used concepts.

Objectives

After reading this unit the students will be able to:

1. explain variable and its types.
2. explain data and its types.
3. explain population and its types.
4. explain sample and its types.

2.1 Variable and Data

Variable

A variable is something that varies or something that is subject to variation. It has no definite value but can assume any set of values. In other words we can say that a variable is a characteristic that varies from one person or thing to another. It is a characteristic, number or quantity that increases or decreases over time or takes different value in different situations; or in more precise words it is a condition or quality that can differ from one case to another. It may also be called a data item. In some other words, a variable is an image, concept or a perception that can be measured. It should be kept in mind that a concept cannot be measured. It must be converted to some measureable form; and measureable form of a concept is called a variable. Examples of variables for human are height, weight, age, number of siblings, business income and expenses, country of birth, capital expenditure, marital status, eye color, gender, class grades, and vehicle type are examples of variables.

Variable = A Concept that can be measured

2.1.1 Types of Variables

Variables can be categorized in three different ways, (a) The causal relationship (b) The design of study, and (c) The unit of measurement. Let us describe these variables in some details.

The Causal Relationship

In causal relationship studies four types of variables may operate. These may be:

i) Change variables that are responsible for bringing about change in a phenomena;
ii) Variables which affect the link between cause and effect variables;
iii) Outcome variables which result from the effects of a change variable;
iv) Connecting or linking variables, which in certain situation are important to complete relationship between cause and effect.

In research, change variables are referred to as independent variables while the outcome variables are known as dependent variables. In cause effect relationship, there are some unmeasured variables affecting the relationship. These are called extraneous variables. The variables linking cause-effect relationship are called intervening variables. A brief summary of above mentioned variables is given in the following table.

Table 2.1: *Types of Variables (causal relationship)*

Variable	**Description**
Independent Variable	It is a cause that brings changes in the situation
Dependent Variable	It is a change that occurs due to dependent variable
Extraneous Variable	It is a situation/factor in everyday life that influences changes in dependent variable. As these factors are not measured in the research study, they can increase or decrease the magnitude of relationship between the independent and dependent variables.
Intervening Variable	It is a link between independent and dependent variable. Sometimes, without the intervention of another variable, it is impossible to establish a relationship between independent and independent variables.

Design of the Study

A study that investigates causation or association may be controlled, contrived experiment, a quasi-experiment or an ex post facto or non-experimental study. Normally, there are two types this category of variables.

i) Active Variables: these variables can be changed or controlled; and

ii) Attribute Variables: these variables can be changed or controlled and refer to characteristics of the research study population. Demographic features like age, gender, education, qualification and income etc. are attributive variables.

Some common types of variables are given below.

i) Binary Variable

These variables take only two values. For example, male or female, true or false, yes of no, improved or not improved, completed task or failed to complete task etc. These variables can be divided into two types; opposite binary variables, and Conjunct binary variables. Opposite binary variables are polar opposite to each other. For example, success or failure, true or false etc. There is no third or middle value. On the other hand conjunct binary variables assume two values but also have middle value. For example, agreeing 20% with the policies of one party and 80% with others.

ii) Categorical Variable

Usually an independent variable or predictor contains values indicating membership in more than one possible categories. For example, gender (male or female), marital status (married, single, divorced, widow), or brand of a product.

iii) Confounding Variable

A variable that has hidden effect on the experiment.

iv) Continuous Variable

A variable with infinite number of values. And its values are obtained by measuring. For example, height and weight of students in a class, time it takes to get to school, distance between Lahore and Karachi etc.

v) Dependent Variable

Outcome or response of an experiment. An independent variable has direct or inverse effect upon dependent variable. In graph it is plotted on y-axis.

vi) Independent Variable

The variable that is manipulated by the researcher. In graph it is plotted on x-axis.

vii) Nominal Variable

It is another name of categorical variable.

viii) Ordinal Variable

Similar as categorical variable, but there is clear order. For example, income level of low, middle and high.

ix) Interval Variable

An interval variable is a measurement where the difference between two values is meaningful. The difference between the temperature of 100^o and 90^o is the same as 80^o and 70^o.

x) Ratio Variable

Similar to interval variable, but has meaningful zero.

xi) Qualitative Variable

A broad category of any variable that can't be counted "i.e. has no numerical value". Nominal and ordinal variable fall under this umbrella.

xii) Quantitative Variable

A broad category of any variable that can be counted "i.e. has numerical value associated with it". Variable fall in this category include discrete variable and ratio variable.

2.1.2 Less Common Types of Variables Data

Some less common types of variables are given below.

i) Attribute Variable

Another name for a categorical variable (in statistical software) or a variable that isn't manipulated (in design of experiments).

ii) **Collider Variable**
A variable represented by a node on a causal graph that has paths pointing in as well as out.

iii) **Covariate Variable**
Similar to an independent variable, it has an effect on the dependent variable but is usually not the variable of interest.

iv) **Criterion Variable**
Another name for a dependent variable, when the variable is used in non-experimental situations.

v) **Dichotomous Variable**
Another name for a binary variable.

vi) **Dummy Variables**
Used in regression analysis when you want to assign relationships to unconnected categorical variables. For example, if you had the categories "has dogs" and "owns a car" you might assign a 1 to mean "has dogs" and 0 to mean "owns a car."

vii) **Endogenous Variable**
Similar to dependent variables, they are affected by other variables in the system. Used almost exclusively in econometrics.

viii) **Exogenous Variable**
Variables that affect others in the system.

ix) **Indicator variable**
Another name for a dummy variable.

x) **Intervening variable**
A variable that is used to explain the relationship between variables.

xi) **Latent Variable**
A hidden variable that can't be measured or observed directly.

xii) **Manifest variable**
A variable that can be directly observed or measured.

xiii) **Manipulated variable**
Another name for independent variable.

xiv) **Mediating variable**
Variables that explain how the relationship between variables happens. For example, it could explain the difference between the predictor and criterion.

xv) **Moderating variable**
Changes the strength of an effect between independent and dependent variables. For example, psychotherapy may reduce stress levels for women more than men, so sex moderates the effect between psychotherapy and stress levels.

xvi) **Nuisance Variable**
An extraneous variable that increase variability overall.

xvii) **Observed Variable**
A measured variable (usually used in SEM).

xviii) **Outcome variable**
Similar in meaning to a dependent variable, but used in a non-experimental study.

xix) **Polychotomous variables**
variables that can have more than two values.

xx) **Predictor variable**
Similar in meaning to the independent variable, but used in regression and in non-experimental studies.

xxi) **Test Variable**
Another name for the Dependent Variable.

xxii) **Treatment variable**
Another name for independent variable.

Data

The term “data” refers to the kind of information a researcher obtains to achieve objectives of his research. All research processes start with collection of data, which plays a significant role in the statistical analysis. This term is used in different contexts. In general, it indicates facts or figures from which conclusions can be drawn. Or it is a raw material from which information is obtained. Data are the actual pieces of information that you collect through your study. In other words data can be defined as collection of facts and details like text, figures, observations, symbols, or simply description of things, event or entity gathered with a view of drawing inferences. It is a raw fact which should be processed to get information

2.1.3 Types of Data

In research, different methods are used to collect data, all of which fall into two categories, i.e. primary data and secondary data. It is a common classification based upon who collected the data.

Primary data

As the name suggests, is one which is collected for the first time by the researcher himself. Primary data is originated by the researcher for the first time for addressing his research problem. It is also known as first hand raw data. The data can be collected using various methods like survey, observations, physical testing, mailed questionnaire, questionnaire filled and sent by enumerators, personal interviews, telephonic interviews, focus groups discussion, case studies, etc.

Secondary data

Point towards the second hand information already collected and recorded by any other person with a purpose not relating to current research problem. It is readily available form of data and saves time and cast of the researcher. But as the data is gathered for the purpose other than the problem under investigation, so the usefulness of the data may be limited in a number of ways like relevance and accuracy. Also, the objectives and methods adopted to collect data may not be suitable to the current situation. Therefore, the researcher should be careful when using secondary data. Examples of secondary data are censuses data, publications, internal records of the organizations, reports, books, journal articles, websites etc.

2.1.4 Key Differences Between primary And Secondary Data

Some key differences between primary and secondary data are given in the following lines.

i) Primary data refers to the data originated by the researcher for the first time. Secondary data is already existing data, collected by other researchers, agencies, and organizations.

ii) Primary data is real-time data whereas secondary data is one which relates to the past.

iii) Primary data is collected to address the problem in hand while the purpose behind collection of secondary data is different from the problem in hand.

iv) Collection of primary data is a laborious process. On the other hand collection of secondary data is easy and rapid.

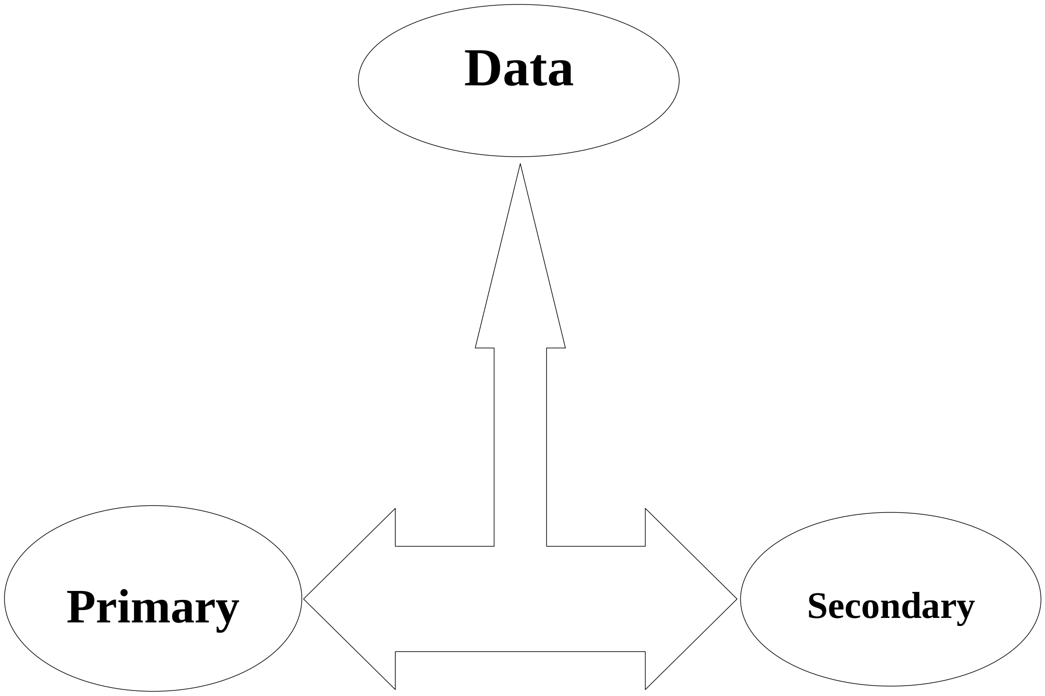
v) Sources of primary data are survey, observations, physical testing, mailed questionnaire, questionnaire filled and sent by enumerators, personal interviews, telephonic interviews, focus groups discussion, case studies, etc. On the other hand sources of secondary are censuses data, publications, internal records of the organizations, reports, books, journal articles, websites etc.

vi) Collection of primary data requires a large amount of resources like time, cost, and human resources. On the other hand collection of secondary data is expensive and easily available.

vii) Primary data is specific to the researcher's needs. He can control the quality of research. On the other hand, secondary data is neither specific to researcher needs nor has he control over the quality of data.

viii) Primary data is available in the raw form while secondary data has undergone some statistical procedures and is refined from primary data.

ix) Data collected from primary sources are more reliable and accurate than the secondary sources.



2.2 Population and Sample

Population

A research population is a large collection of individuals or objects to which the researcher wants the results of the study to apply. Population is the main focus of a research question. A research population is also known as a well-defined collection of individuals or objects known to have similar characteristics. All individuals or objects within a certain population usually have a common, binding characteristic or trait. Population can also be defined as all individual that meet a set of specification or a specific criteria. All researches are done for the benefit of population.

2.2.1 Types of Population, Sample

In educational research, we commonly come across two types of populations.

i) **The Target Population** is also known as the theoretical population and refers to the entire group of individuals or objects to which a researcher is interested to generalize the conclusions. This type of population usually has varying degree of characteristics.

ii) **The Accessible Population** is also known as the study population. It is the population to which a researcher can apply the conclusions of the study. This population is a subset of the target population.

Sample

A sample is simply a subset or subgroup of population (Frey, Carl, & Gary, 2000).The concept of sample arises from the inability of the researchers to test all the individuals in a given population. Sampling is the process of selecting some individuals from the accessible population, in a way that these individuals represent whole accessible population. The sample should be representative in a sense that each individual should represent the characteristics of the whole population (Lohr, 1999). The main function of the sample is to allow the researchers to conduct the study to individuals from the population so that the results of their study can be used to derive conclusions that will apply to the entire population.

2.2.2 Types of Sample

Generally researchers use two major sampling techniques: probability sampling and non-probability sampling.

Probability sampling

Is a process that utilizes some form of random selection. In probability sampling, each individual in chosen with a known probability. This type of sampling is also known as random sampling or representative sampling; and depends on objective judgment. Various types of probability are as under:

i) Simple Random sampling

In random sampling each member of the population has an equal chance of being selected as subject. Each member is selected independently of the other member of population. Many methods are used to proceed with random sampling. In a commonly used method each member of the population is assigned a unique number. All assigned numbers are placed in bowl and mixed thoroughly. The researcher, then blind-folds and picks numbered tags from the bowl. All the numbers picked are the subjects of the study. Another method is to use computer for random selection from the population. For smaller population first method is useful and for larger population computer-aided method is preferred.

Advantages of Simple Random Sampling

It is an easy way of selecting a sample from a given population. This method is free from personal bias. As each member of the population is given equal opportunities of being selected so it a fair way and one can get representative sample.

Disadvantages of Simple Random Sampling

One of the most obvious limitations of random sampling method is its nee of a complete list of all members of the population. For larger population, usually this list is not available. In such case, it is better to use other sampling techniques.

ii) Systematic Random Sampling

In systematic random sampling, the researcher first randomly picks the first item or the subject from the population. Then he selects each nth subject from the list. The procedure involved in this sampling is easy and can be done manually. The sample drawn using this procedure is representative unless certain characteristics of the population are repeated for every nth member, which is highly risky.

Suppose a researcher has a population of 100 individuals and he needs 12 subjects. He first picks his starting number 7. He then picks his interval 8. The members of his sample will be individual 7, 15, 23, 31, 39, 47, 55, 63, 71, 79, 87, and 95

Advantages of Systematic Random Sampling

The main advantage of using this technique is its simplicity. It allows researcher to add a degree of system or process into the random selection of subjects. Another advantage is its assurance that the population will be evenly sampled.

Disadvantages of Systematic Random Sampling

Systematic sampling assumes that the size of the population is available or can be approximated. Suppose a researcher wants to study the behavior of monkeys of a particular area. If he does not have any idea of how many monkeys there are, he cannot systematically select a starting point or interval size. If any population has a type of natural standardized pattern, the risk accidently choosing very common cases is more apparent.

iii) Stratified Random Sampling

In this type of sampling, the whole population is divided into disjoint subgroups. These subgroups are called stratum. From each stratum a sample of pre-specified size is drawn independently in different strata, using simple random sampling. The collection of these samples constitutes a stratified sample.

Advantages

This type of sampling is appropriate when the population has diversified social or ethnic subgroups.

Disadvantages

While using this type of sampling, there is greater chance of overrepresentation of subgroups in the sample.

iv) Cluster Sampling

It is a simple random sample in which each sampling unit is a collection or cluster, or elements. For example, a researcher who wants to study students may first sample groups or cluster of students such as classes, and then, select the sample of students from among the clusters.

Advantages

This type of sampling is appropriate for larger population. It saves time and resources.

Disadvantages

In this type of sampling, there is a greater chance of selecting a sample that is not representative of the whole population.

Non-Probability Sampling or Judgmental Sampling

This technique depends on subjective judgment. It is a process where probabilities cannot be assigned to the individuals objectively. It means that in this technique samples are gathered in a way does not give all individuals in the population equal chances of being selected. Choose these methods could result in biased data or a limited ability to make general inferences based on the findings. But there are also many situations in which choosing this kind of sampling techniques is the best choice for a particular research question or the stage of research.

There are four kinds of non-probability sampling techniques.

i) Convenience Sampling

In this technique a researcher relies on available subjects, such as stopping peoples in the markets or on street corners as they pass by. This method is extremely risky and does not allow the researcher to have any control over the representativeness of the sample. It is useful when the researcher wants to know the opinion of the masses on a current issue; or the characteristics of people passing by on streets at a certain point of time; or if time and resources are limited in such a way that the research would not be possible otherwise. What may be the reason for selecting convenience samples, it is not possible to use the results from a convenience sampling to generalize to a wider population.

ii) Purposive or Judgmental Sampling

In this technique a sample is selected on the bases of the knowledge of population and the purpose of the study. For example, when an educational psychologist wants to study the emotional and psychological effects of corporal punishment, he will create a sample that will include only those students who ever had received corporal punishment. In this case, the researcher used purposive sample because those being selected fit a specific purpose or description that was necessary to conduct the research.

iii) Snowball Sample

This type of sampling is appropriate when the members of the population are difficult to locate, such as homeless industry workers, undocumented immigrants etc. a snowball sample is one in which the researcher collects data on a few members of the target population he or she can locate, then asks to locate those individuals to provide information needed to locate other members of that population whom they know. For example, if a researcher wants to interview undocumented immigrants from Afghanistan, he might interview a few undocumented individuals he knows or can locate, and would then rely on those subjects to help locate more undocumented individuals. This process continues until the researcher has all the interviews he needed, until all contacts have been exhausted. This technique is useful when studying a sensitive topic that people might not openly talk about, or if talking about the issue under investigation could jeopardize their safety.

iv) Quota Sample

A quota sample is one in which units are selected into a sample on the basis of pre-specified characteristics so that the total sample has the same distribution of characteristics assumed to exist in the population. For example, if a researcher wants a national quota sample, he might need to know what proportion of the population is male and what proportion is the female, as well as what proportion of each gender fall into different age category and educational category. The researcher would then collect a sample with the same proportion as the national population.

2.3 Self-Assessment Questions

Q. 1 What is a variable?
Q. 2 What are commonly used types of variable?
Q. 3 What do you understand by the term “data”?
Q. 4 Write down the types of data.
Q. 5 What is population?
Q. 6 What do you understand by the target population?
Q. 7 What do you mean by the assessable population?
Q. 8 What do you mean by the term “sample”?
Q. 9 Write down the types of probability sampling.
Q. 10 Write down the types of non-probability sampling.

2.4 Activities

1. Suppose a scientist is conducting an experiment to test the what extant a vitamin could extend a person’s life expectancy. Identify:
   i) Independent Variable of the experiment.
   ii) Dependent Variable of the experiment.

2. Suppose a Lahore-based company is launching a new product for senior citizens of Pakistan and tests that product for senior citizens of Lahore. Identify:
   i) Target Population of the company.
   ii) Assessable Population of the company.

2.5 Bibliography

Bartz, A. E. (1981). *Basic Statistical Concepts (2nd Ed.)*. Minnesota: Burgess Publishing Company

Deitz, T., & Kalof, L. (2009). *Introduction to Social Statistics*. UK: Wiley_-Blackwell

Frey, L. R., Carl H. B., & Gary L. K. (2000). *Investigating Communication: An Introduction to Research Methods.*2nd Ed. Boston: Allyn and Bacon

Gay, L. R., Mills, G. E., & Airasian, P. W. (2010). *Educational Research: Competencies for Analysis and Application*, *10th Edition*. Pearson, New York USA.

Gravetter, F. J., & Wallnau, L. B. (2002). *Essentials of Statistics for the Behavioral Sciences (4th Ed.)*. Wadsworth, California, USA.

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Albany: Duxbury Press.

UNIT-3

STATISTICAL GRAPHICS / EXPLORATORY DATA ANALYSIS

Written By:
Miss Sumbal Asghar

Reviewed By:
Dr. Rizwan Akram Rana

Introduction

Graphical representation of data is for the purpose of easier interpretation. Facts and figures as such do not catch our attention unless they are presented in an interesting way. Graphical representation of data is the most commonly used interesting modes of presentation. The purpose of this unit is to make you familiar with this interesting mode of presentation.

Objectives

After reading this unit, you will be able to explain:

1. Bar Chart
2. Pictograms
3. Histogram
4. Frequency Polygon or Ogive
5. Scatter Plot
6. Box Plot
7. Pie Chart

3.1 Bar Chart

Bar charts are one of the most commonly used graphical representations of data used to visually display compare values to each other. They are easy to create and interpret. They are also flexible and have several variations of standard bar charts including vertical or horizontal bar charts, component or grouped charts, and stacker bar charts.

Data for a bar chart are entered in columns. Each numeric data value becomes a bar. The chart is constructed such that lengths of the different bars are proportional to the size of the category they represent. X-axis represents the different categories and has no scale; the y-axis does have a scale and indicates the units of measurement, in case of vertical bar charts, and vice versa in case of horizontal bar charts.

In the following figure result of first, second and third term of a student in the subjects of English, Urdu, Mathematics and Pak-Studies.

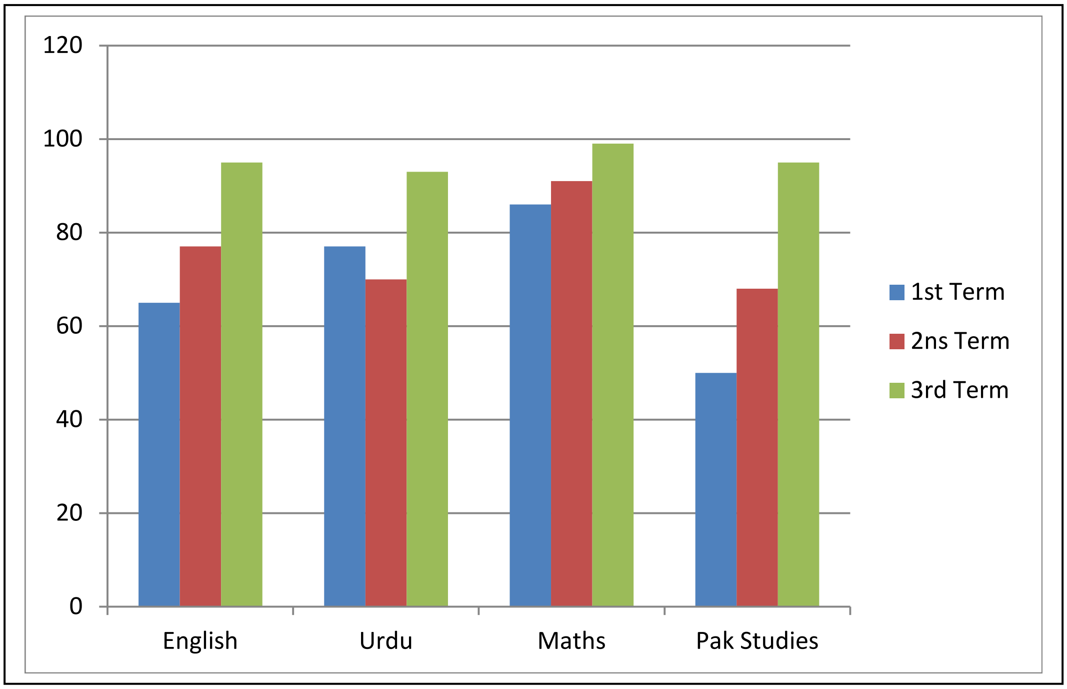


Fig 1: Vertical bar chart

Bar chart can also be represented in horizontal form.

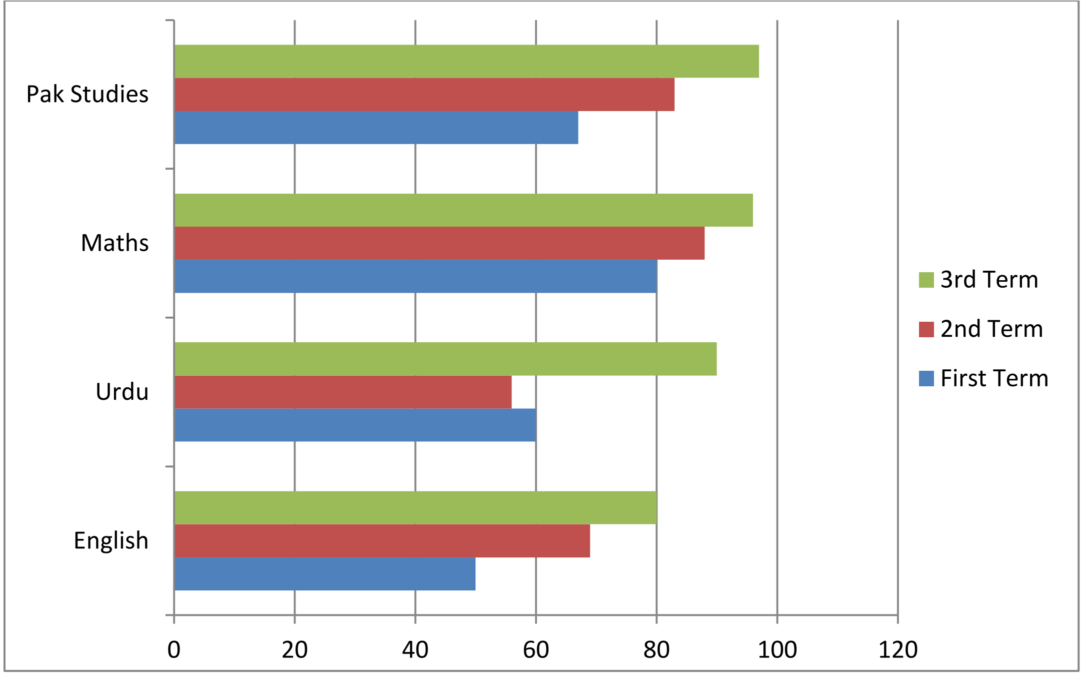


Fig 2: Horizontal bar chart

3.1.1 Advantages and Disadvantages of Bar Charts

Following are the advantages of bar charts.

i) They show data category in a frequency distribution.
ii) They display relative numbers / proportions of multiple categories.
iii) They summarize a large amount of data in an easily interpretable manner.
iv) They make trends easier to highlight than tables do.
v) By bar charts estimates can be made quickly and accurately.
vi) They are easily accessible to everyone.

Following are the disadvantages of bar charts.

i) They often require additional explanation.
ii) Thy fail to expose key assumptions, causes, impacts and patterns
iii) T hey can be manipulated to give false impressions.

3.2 Pictograms

A pictogram is a graphical symbol that conveys its meaning through its pictorial resemblance to a physical object. A pictogram may include a symbol plus graphic elements such as border, back pattern, or color that is intended to covey specific information s. we can also say that a pictogram is a kind of graph that uses pictures instead of bars to represent data under analysis. A pictogram is also called “pictograph”, or simply “picto”.

A pictogram or pictograph represents the frequency of data as pictures of symbols. Each picture or symbols may represent one or more units of data.

Pictograms form a part of our daily lives. They are used in transport, medication, education, computers etc. they indicate, in iconic form, places, directions, actions or constraints on actions in either the real world (a road, a town, etc) or in virtual world (computer, internet etc.).

To successfully convey the meaning, a pictogram:

i) Should be self-explanatory.
ii) Should be recognizable by all people.
iii) Must represent a general concept.
iv) Should be clear concise and interesting.
v) Should be identifiable as a set, through uniform treatment of scale, style and subject.
vi) Should be highly visible, easy to reproduce in any scale and in positive or negative form.
vii) Should not be dependent upon a border and should work equally well in positive or negative form.
viii) Should avoid stylistic fads or a commercial appearance and should imply to wide audience that has a sophisticated, creative culture.
ix) Should be attractive when used with their design, elements and typestyles.

3.2.1 Advantages and Drawbacks of Pictograms

Following are the advantages of pictograms:

i) Pictograms can make warnings more eye-catching.

ii) They can serve as an "instant reminder" of a hazard or an established message.

iii) They may improve warning comprehension for those with visual or literacy difficulties.

iv) They have the potential to be interpreted more accurately and more quickly than words.

v) They can be recognized and recalled far better than words.

vi) They can improve the legibility of warnings.

vii) They may be better when undertaking familiar routine tasks.

There are a number of disadvantages of relying on pictograms.

i) Very few pictograms are universally understood.

ii) Even well understood pictograms will not be interpreted equally by all groups of peoples and across all cultures, and it takes years for any pictogram to reach maximum effectiveness.

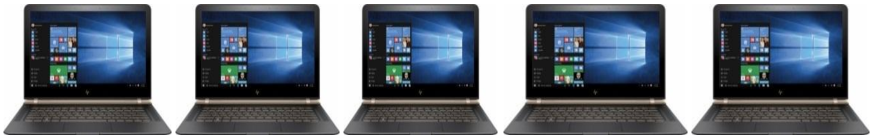
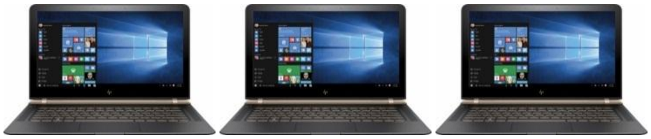
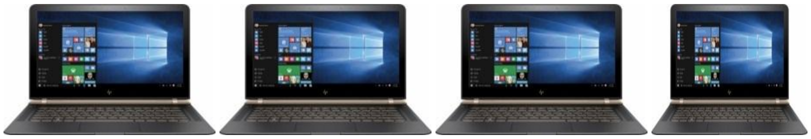
iii) They have the potential for interpreting the opposite or often undesired meaning which can create additional confusion.

Example

The following table shows the number of laptops sold by a company for the months January to March. Construct a pictograph for the table.

Month	January	February	March
Number of laptops	25	15	20

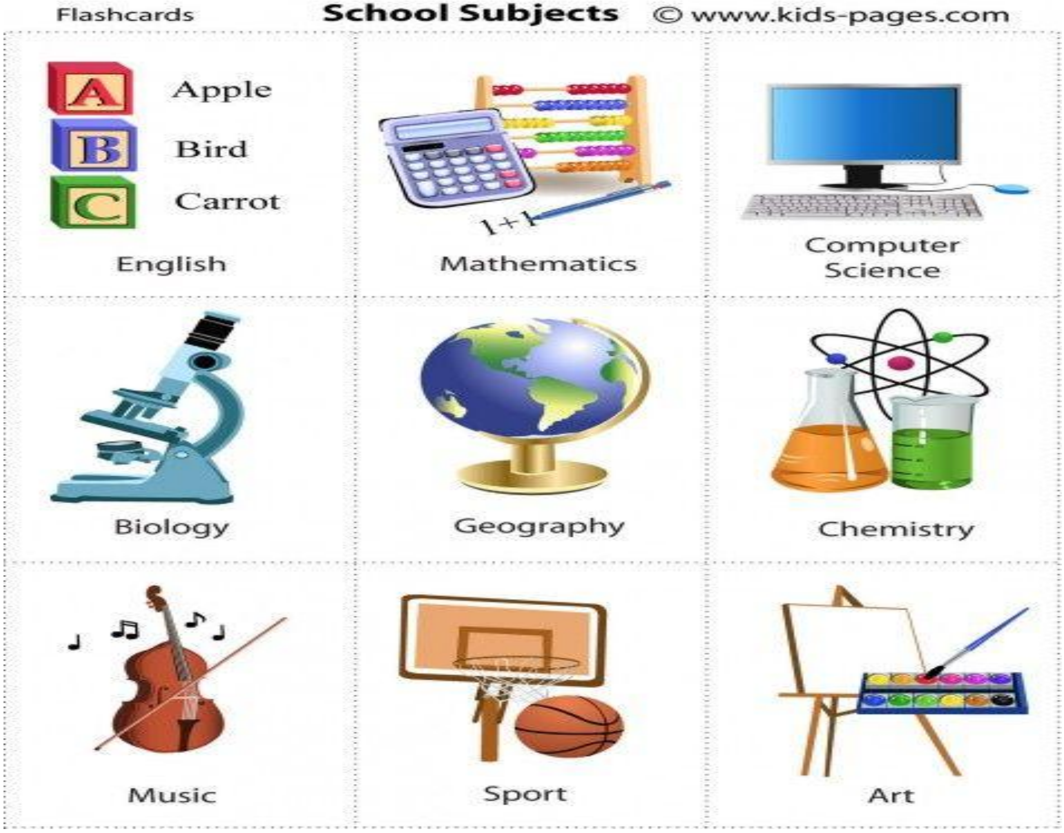
Solution:

January	
February	
March	

- represents 5 laptops

Example

School Subject pictogram



Source: www.kids-pages.com

3.3 Histogram

A histogram is a type of graph that provides a visual interpretation of numerical data by indicating the number of data points that lie within the range value. These range values are called classes or bins.

A histogram looks similar to bar charts. Both are ways to display data set. The height of the bar corresponds to the relative frequency of the amount of data in the class. The higher the bar is, the greater the frequency of the data will bean vice versa. The main difference between these graphs is the level of measurement of the data. Bar graphs are used for data at nominal level of measurement. It measures the frequency of categorical data. On the other hand histograms are used for data that is at least ordinal level of measurement. As a common practice the bars of bar graph are rearranged in order for decreasing height. However the bars of cannot be rearranged. They must be displayed in order that the classes occur.

A bar graph presents actual counts against categories. The height of the bar indicates the number of items in that category. A histogram displays the same categorical variables in bins. While creating a histogram, you are actually creating a bar graph that shows how many data points are there within the range (an interval), called a bin.

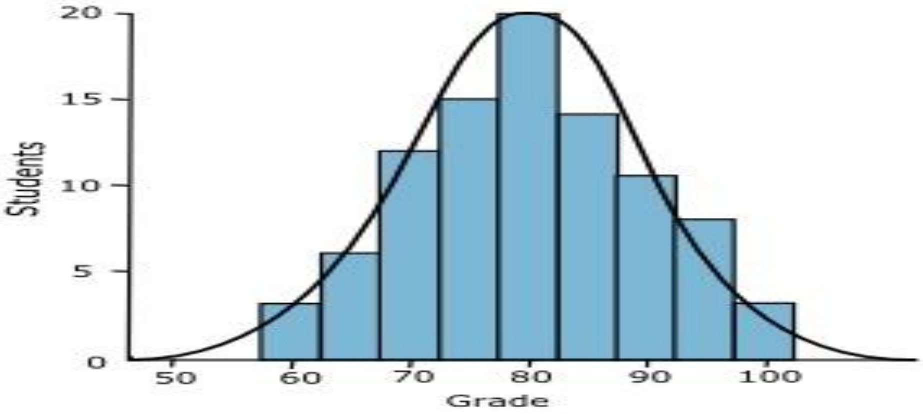
There are no hard and fast rules about how many bins there should be. But the rule of thumb is 5-20 bins. Less than 5 bins will have little meaning and more than 20 bins, will make data hard to read and interpret. Ideally 5-7 bins are enough.

3.3.1 Shapes of Histogram

Histogram may be of different shapes. Following are some of the shapes.

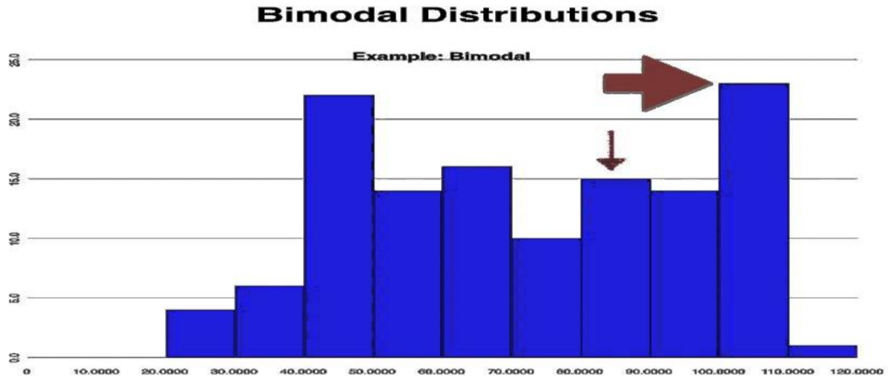
i) Bell-shaped

A bell-shaped picture, shown below, usually presents a normal distribution.



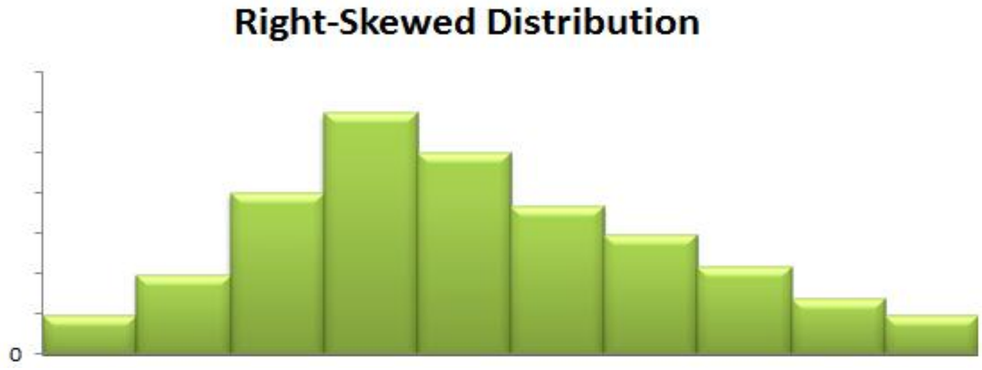
ii) Bimodal

A bimodal shape, shown below, has two peaks. This shape may show that the data has come from two different systems. Often in a single system, there may be two modes in the data set.



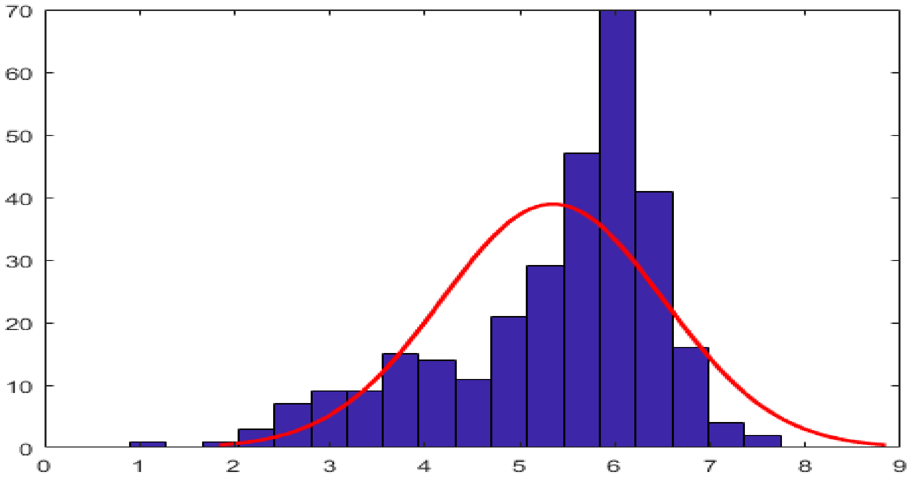
iii) Skewed right

Some histograms will show a skewed distribution to the right, as shown below. A distribution skewed to the right is said to be positively skewed. This kind of distribution has a large number of occurrences in the lower value cells (left side) and few in the upper value cells (right side). A skewed distribution can result when data is gathered from a system with has a boundary such as zero. In other words, all the collected data has values greater than zero.



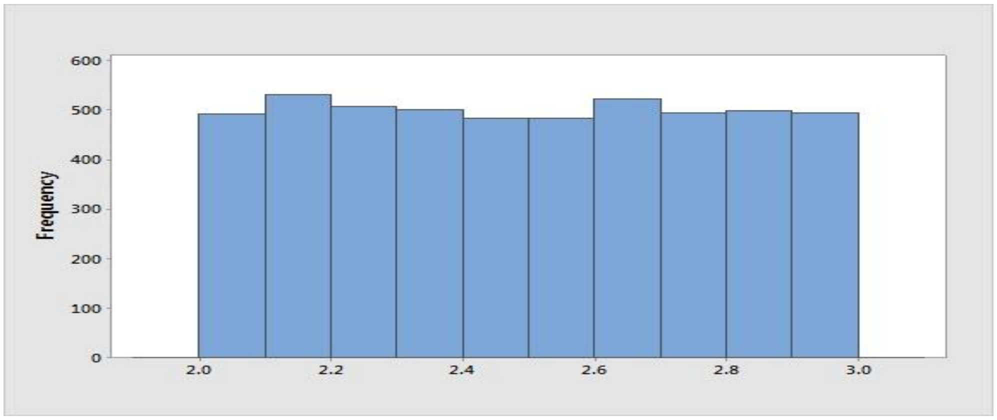
iv) Skewed left

Some histograms will show a skewed distribution to the left, as shown below. A distribution skewed to the left is said to be negatively skewed. This kind of distribution has a large number of occurrences in the upper value cells (right side) and few in the lower value cells (left side). A skewed distribution can result when data is gathered from a system with a boundary such as 100. In other words, all the collected data has values less than 100.



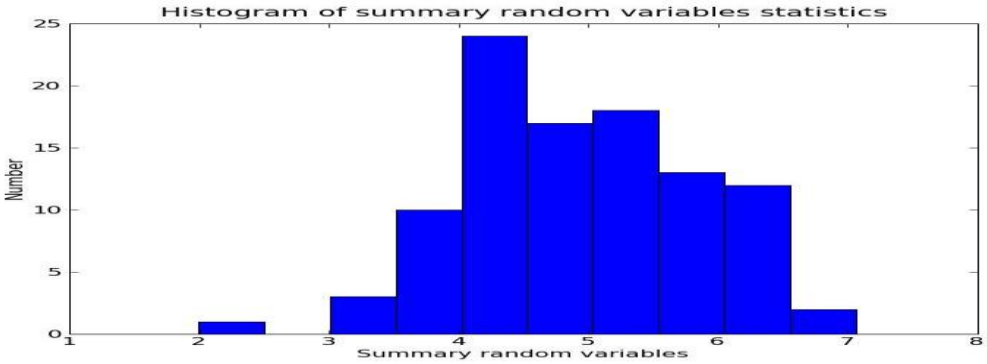
v) Uniform

A uniform distribution, as shown below, provides little information about the system. It may describe a distribution which has several modes (peaks). If your histogram has this shape, check to see if several sources of variation have been combined. If so, analyze them separately. If multiple sources of variation do not seem to be the cause of this pattern, different groupings can be tried to see if a more useful pattern results. This could be as simple as changing the starting and ending points of the cells, or changing the number of cells. A uniform distribution often means that the number of classes is too small.



vi) Random

A random distribution, as shown below, has no apparent pattern. Like the uniform distribution, it may describe a distribution that has several modes (peaks). If your histogram has this shape, check to see if several sources of variation have been combined. If so, analyze them separately. If multiple sources of variation do not seem to be the cause of this pattern, different groupings can be tried to see if a more useful pattern results. This could be as simple as changing the starting and ending points of the cells, or changing the number of cells. A random distribution often means there are too many classes.



Source: http://www.pqsystems.com/qualityadvisor/DataAnalysisTools/histogram.php

3.4 Frequency Polygon

The frequency polygon is as graph that displays data by using lines that connect points plotted for the frequencies at the midpoint of the classes. This graph is useful for understanding the shape of distribution. They are good choice for displaying cumulative frequency distribution.

A frequency polygon is similar to histogram. The difference is that histogram tends to be rectangles while a frequency polygon resembles a line graph.

3.5 Cumulative Frequency Polygon or Ogive

The cumulative frequency is the sum of the frequencies accumulated up to the upper boundary of a class in the distribution. A graph that can be used to represent the cumulative frequencies for the classes is called cumulative frequency graph or ogive.

An ogive is drawn on the basis of cumulative frequency. To construct cumulative frequency, first we have to form cumulative frequency table. The upper limits of the classes are taken on the x-axis and the cumulative frequencies on the y-axis and the points are plotted.

There are two methods for of drawing a cumulative frequency curve or ogive.

i) The less than method
In this method a frequency distribution is prepared which gives the number of items that are less than a certain size. It gives a series which is cumulatively upward.

ii) The greater than method
In this method a frequency distribution is prepared that gives the number of items that exceed a certain size and gives a series which is cumulatively downward.

Example

Marks of 30 students of a class, obtained in a test out of 75, are given below: 42, 21, 50, 37, 38, 42, 49, 52, 38, 53, 57, 47, 29, 59, 61, 33, 17, 17, 39, 44, 42, 39, 14, 7, 27, 19, 54, 51.

Classes	Frequency	Cumulative Frequency	
		Less Than	Greater Than
0-10	1	1	29 + 1 = 30
10-20	4	1 + 4 = 5	22 + 7 = 29
20-30	3	5 + 3 = 8	15 + 7 = 22
30-40	7	8 + 7 = 15	8 + 7 = 15
40-50	7	15 + 7 = 22	5 + 3 = 8
50-60	7	22 + 7 = 29	1 + 4 = 5
60-70	1	29 + 1 = 30	1
Total	30		

3.6 Scatter Plot

A scatter plot is used to plot data in XY- plane to show how much one variable or data set is affected by another. It has points that show the relationship between two variables or two sets of data. These points are sometimes called markers and position of these points depends on the values in the columns sets on the XY axis. Scatter plot gives good visual picture of the relationship or association between two variables or data sets, and aids to interpretation of the correlation coefficient or regression model.

The relationship between two data sets or variables is called correlation. If the markers are close together and make a straight line in the scatter plot, the two variables of data sets have high correlation. If the markers are equally distributed in the scatter plot, the correlation is low, or zero.

Correlation may be positive or negative. Correlation is positive when the values increase together, i.e. if one value increases the other will also increase or if once value decreases the other will also decrease. On the other hand, correlation is negative when one value increases the other decreases, and vice versa.

Scatter plot provides answers of the following questions.

i) Are variables X and Y or two data sets related?
ii) Are variables X and Y or two data sets linearly related?
iii) Are variables X and Y or two data sets non-linearly related?
iv) Does the variation Y or one data set change depending on X or other data set?
v) Are there outliers?

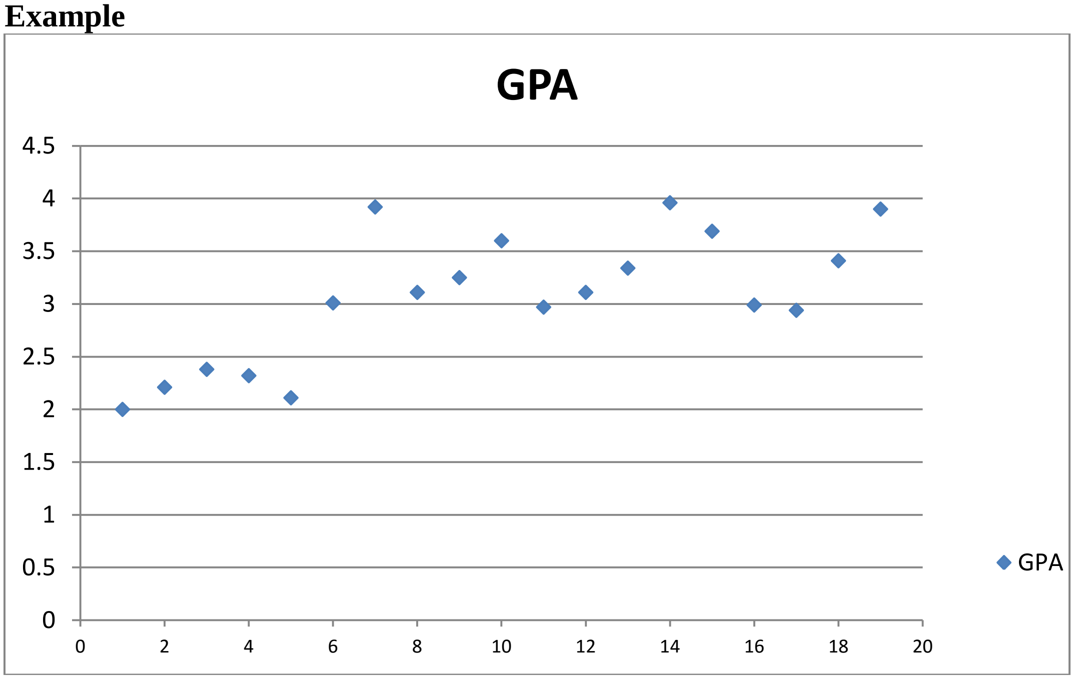
3.6.1 When to Use Scatter Plot?

Following situations provide a rationale to use a scatter plot.

i) When there is paired numerical data.
ii) When the dependent variable have multiple values for each value of independent variable.
iii) When the researcher tries to determine whether the two variables are related, such as:
    a) When trying to identify potential root causes of the problems.
    b) To determine objectively whether a particular cause and effect are related.
    c) When determining whether two effects those appear to be related both occur with the same cause.
    d) When testing for autocorrelation before constructing a control.

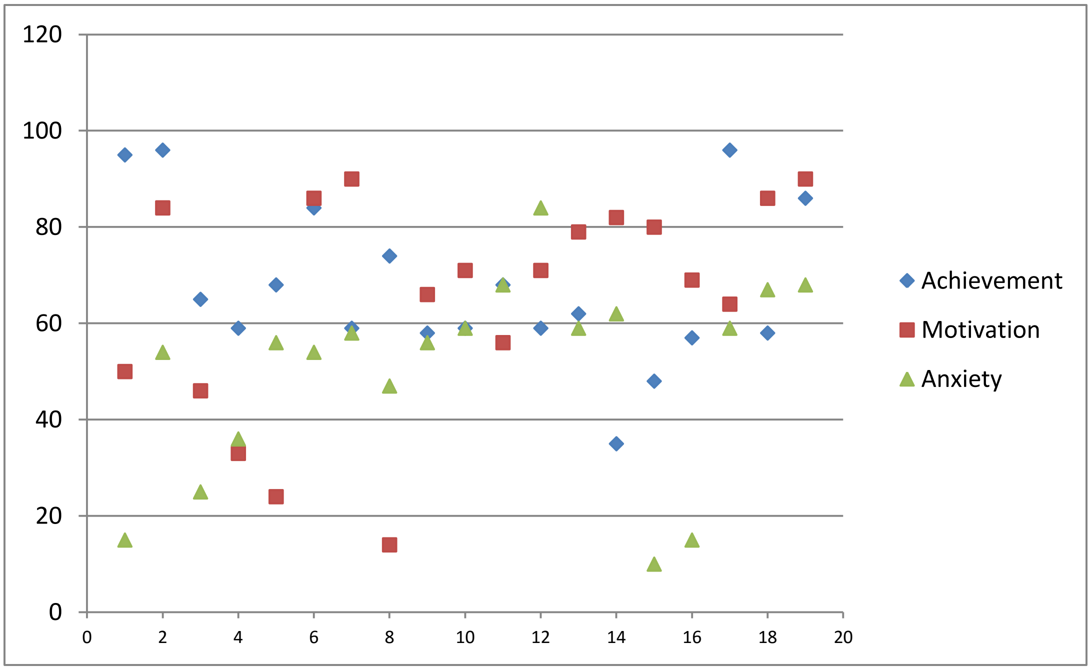
Name of Student	GPA
A	2.0
B	2.21
C	2.38
D	2.32
E	2.11
F	3.01
G	3.92
H	3.11
I	3.25
J	3.60
K	2.97
L	3.11
M	3.34
N	3.96
O	3.69
P	2.99
Q	2.94
R	3.41
S	3.90

Example



Example

Name of Student	Achievement	Motivation	Anxiety
A	95	50	15
B	96	84	54
C	65	46	25
D	59	33	36
E	68	24	56
F	84	86	54
G	59	90	58
H	74	14	47
I	58	66	56
J	59	71	59
K	68	56	68
L	59	71	84
M	62	79	59
N	35	82	62
O	48	80	10
P	57	69	15
Q	96	64	59
R	58	86	67
S	86	90	68



3.7 Box Plot

The box plot is an exploratory graph. It is a standardized way of displaying the distribution of data based on the five summary statistics: minimum, first quartile, median, third quartile, and maximum. First and third quartile is called two hinges, first quartile is the lower hinge and the third quartile is the upper hinge. Minimum and the maximum are two whiskers. Minimum is the lower whisker and the maximum is the upper whisker. In other words we can say that box plot visualizes five summary statistics: the median, two hinges and two whiskers.

In the simplest box plot the central triangle spans the first quartile to the third quartile (inter quartile range IQR). A segment inside the rectangle shows the median and whiskers above and below the box show the locations of the minimum and maximum.

Box plot is useful for identifying outliers and for comparing distributions. In other words we can say that box plot gives us information about the location and variation in the data set. Particularly it helps us in detecting and illustrating location and variation changes between different groups of data.

3.7.1 Types of Box Plot

Commonly used types of box plot are single box plot and multiple box plot.

Single box plot

A single box plot can be drawn for one set of data with no distinct groups. In such a plot the width of the box is arbitrary.

Multiple box lot

Multiple box plots can be drawn together to compare multiple data sets or to compare groups in a single data set. In such a plot the width of the box plot can be set proportional to the number of points in the given group or sample.

The box plot provides answers to the following questions.

i) Is a factor significant?

ii) Does the location differ between subgroups or between different data sets?

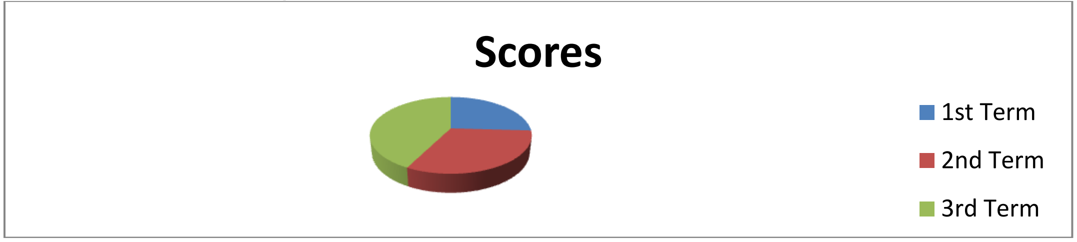
iii) Does the variation differ between subgroups or between different data sets?

iv) Are there any outliers?

A box-plot can tell whether a data set is symmetric (when the median is in the center of the box), but it can’t tell the shape of the symmetry the way a histogram can.

3.8 Pie Chart

A pie chart displays data in an easy pie-slice format with varying sizes. The size of a slice tells how much data exists in one element. The bigger the slice, the more of that particular data was gathered and vice versa. Pie charts are mainly used to show comparison among various segments of data. When items are presented on a pie chart, it is easy to see which item has maximum frequency and which is not or which item is the most popular and which is not. The main purpose of using a pie chart is to show part-whole relationship. These charts are used for displaying data that are classified into nominal or ordinal categories.



3.8.1 How to Read a Pie Chart?

It is easy to read and interpret a pie-chart. Usually, a pie-chart has several bits of data, and each is pictured on a pie-chart as a pie slice. Some data have larger slices than others. So it is easy to decide which data have maximum frequency and which have minimum.

3.8.2 When to Use the Pie Chart?

There are some simple criteria that can be used to determine whether a pie chart is right choice or not for a given data.

i) **Do the parts make up a meaningful whole?**
Pie charts should be used only if parts or slices can define the entire set of data in a way that makes a meaningful sense to the viewer.

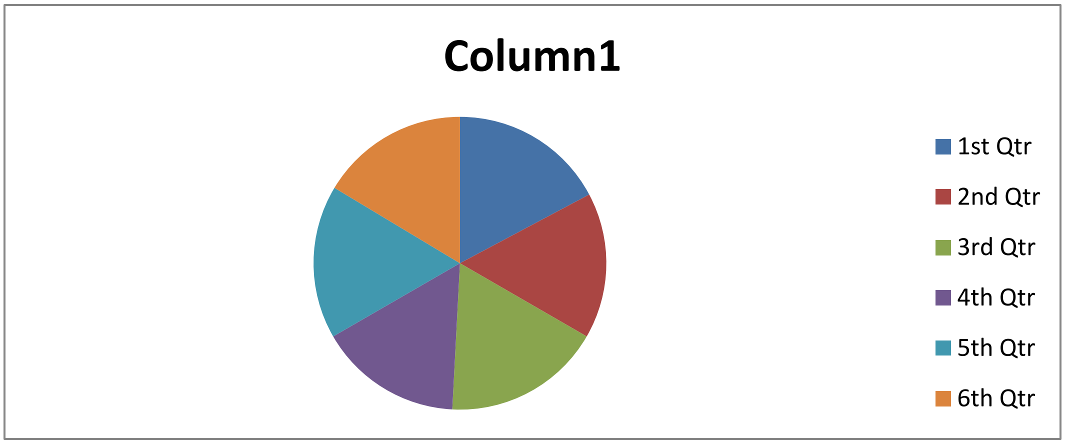
ii) **Are the parts mutually exclusive?**
If there is overlap between the parts, it is better to use any other chart.

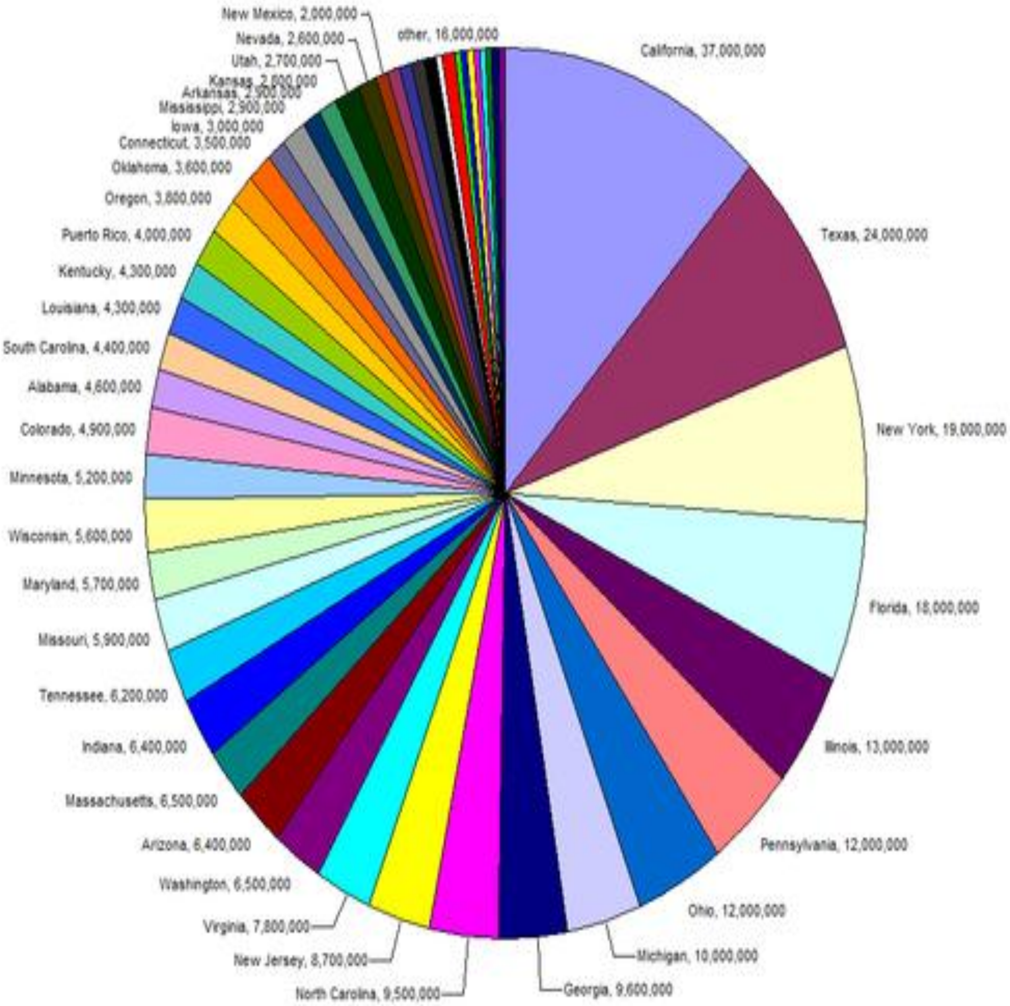
iii) **Do you want to compare the parts to each other or the parts to the whole?**
If the main purpose is to show part-whole relationship then pie chart is useful but if the main purpose is to show part-part relationship then pie chart is useless and wise to use another chart.

iv) **How many parts do you have?**
If there are more than five to seven parts it advisable to use a different chart. Pie charts with lots of slices of varying size are hard to read.

3.8.3 Draw Backs of Pie-Charts

There are two features that help us read the values on a pie chart: the angle a slice covers (compared to the entire circle) and the area of slice (compared to the entire circle). Generally, we are not very good at measuring angles. We only recognize angles of 90^o and 180^o with high degree of precision. Other angles are rather impossible to perceive with a high degree of precision. Look upon following two pie-graphs. In the first, which quarter is larger and which is smaller? And what information can we get from the second graph?





Source: https://eagereyes.org/techniques/pie-charts

3.9 Self Assessment Questions

Q. 1 What is a bar chart?
Q. 2 For what purpose bar carts are used?
Q. 3 What type of characteristics a pictogram should have to successfully convey the meaning?
Q. 4 Write down the advantages and drawbacks of using pictograms.
Q. 5 What is a histogram?
Q. 6 Draw a bell-shaped histogram.
Q. 7 Write down the methods for drawing cumulative frequency polygon.
Q. 8 Write down the rationale for using scatter plot.
Q. 9 Write down any four questions that can be answered using scatter plot.
Q. 10 Write down the types of box plot.
Q. 11 What is a pie-chart?
Q. 12 Write down the criteria to determine whether pie-chart is a right choice.

3.10 Activities

1. Make a list of advantages and disadvantages of bar chart.
2. Make a list of advantages and disadvantages of pictogram.
3. Make a list of the situations that provide rationale to use scatter plot.
4. Make a pie chart that shows the drawback of pie chart.

3.11 Bibliography

Gravetter, F. J., & Wallnau, L. B. (2002). *Essentials of Statistics for the Behavioral Sciences (4th Ed.)*. Wadsworth, California, USA.

https://eagereyes.org/techniques/pie-charts

http://www.pqsystems.com/qualityadvisor/DataAnalysisTools/histogram.php

UNIT-4

DESCRIPTIVE STATISTICS: MEASURES OF DISPERSION

Written By:
Miss Sumbal Asghar

Reviewed By:
Dr. Rizwan Akram Rana

Introduction

Measures of central tendency estimate normal or central value of a dataset, while measures of dispersion are important for describing the spread of the data, or its variation around a central value. Two distinct samples may have same mean or median, but completely different level of variability and vice versa. A proper description of a set of data should include both of these characteristics. There are various methods that can be used to measure the dispersion of a dataset. In this unit you will study range, quartiles, quartile deviation, mean or average deviation, standard deviation and variance. Two measures of shape and discussion about co-efficient of variation are also included in this unit.

Objectives

After reading this unit, you will be able to:

1. tell the basic purpose of measure of central tendency.
2. define Range.
3. determine range of a given data.
4. write down the formulas for determining quartiles.
5. define mean or average deviation.
6. determine variance and standard deviation.
7. define normal curve.
8. explain skewness and kurtosis.

4.1 Introduction to Measures of Dispersion

Measures of central tendency focus on what is an average or in the middle of the distribution of scores. Often the information provided by these measures does not give us clear picture of the data and we need something more. It means that knowing the mean, median, and mode of a distribution does allow us to differentiate between two or more than two distributions; and we need additional information about the distribution. This additional information is provided by a series of measures which are commonly known as measures of dispersion.

There is dispersion when there is dissimilarity among the data values. The greater the dissimilarity, the greater the degree of dispersion will be.

Measures of dispersion are needed for four basic purposes.

i) To determine the reliability of an average.
ii) To serve as a basis for the control of the variability.
iii) To compare two or more series with regard to their variability.
iv) To facilitate the use if other statistical measures.

Measure of dispersion enables us to compare two or more series with regards to their variability. It is also looked as a means of determining uniformity or consistency. A high degree would mean little consistency or uniformity whereas low degree of variation would mean greater uniformity or consistency among the data set. Commonly used measures of dispersion are range, quartile deviation, mean deviation, variance, and standard deviation.

4.1.1 Range

The range is the simplest measure of spread and is the difference between the highest and lowest scores in a data set. In other words we can say that range is the distance between largest score and the smallest score in the distribution. We can calculate range as:
Range = Highest value of the data – Lowest value of the data

For example, if lowest and highest marks scored in a test are 22 and 95 respectively, then
Range = 95 – 22 = 73

The range is the easiest measure of dispersion, and is useful when you wish to evaluate whole of a dataset. But it is not considered a good measure of dispersion as it does not utilize the other information related to the spread. The outliers, either extreme low or extreme high value, can considerably affect the range.

4.1.2 Quartiles

The values that divide the given set of data into four equal parts is called quartiles, and is denoted by Q_1, Q_2, and Q_3. Q_1 is called the lower quartile and Q_3 is called the upper quartile. 25% of scores are less than $Q_1$and 75% scores are less than Q_3. Q_2 is the median. The formulas for the quartiles are:

$$Q_1 = (N + \frac{1}{4})^{th} \text{ Score}$$

$$Q_2 = 2\,(N + \frac{1}{4})^{th} = (N + \frac{1}{2})^{th} \text{ Score}$$

$$Q^3 = 3(N + 1) / 4^{th} \text{ Score}$$

4.1.3 Quartile Deviation (QD)

Quartile deviation or semi inter-quartile range is one half the differences between first and the third quartile, i.e.

$$Q\,D = Q_3 – Q_1$$

Where Q_1 = the first quartile (lower quartile)
Q_3 = third quartile (upper quartile)

Calculating quartile deviation from ungrouped date:

In order to calculate quartile deviation from ungrouped data, following steps are used.

i) Arrange the test scores from highest to lowest

ii) Assign serial number to each score. The first serial number is assigned to the lowest score.

iii) Determine first quartile (Q_1) by using formula $Q_1 = \frac{N}{4}$. Use obtained value to locate the serial number of the score that falls under Q_1.

iv) Determine the third (Q_3), by using formula $Q_3 = \frac{3N}{4}$. Locate the serial number corresponding to the obtained answer. Opposite to this number is the test score corresponding to Q_3.

v) Subtract the Q1 from Q3, and divide the difference by 2.

4.1.4 Mean Deviation or Average Deviation

The mean or the average deviation is defined as the arithmetic mean of the deviations of the scores from the mean or the median. The deviations are taken as positive. Mathematically
For ungrouped data

$$M.D = \Sigma \left| X - \overline{X} \right| / N$$

For grouped data

$$M.D = \Sigma f \left| X - \overline{X} \right| / \Sigma f$$

4.1.5 Standard Deviation

Standard deviation is the most commonly used and the most important measure of variation. It determines whether the scores are generally near or far from the mean, i.e. are the scores clustered together or scattered. In simple words, standard deviation tells how tightly all the scores are clustered around the mean in a data set. When the scores are close to the mean, standard deviation is small. And large standard deviation tells that the scores are spread apart. Standard deviation is simply square root of variance, i.e.

Standard deviation $\sigma = \surd$ Variance

Or

$$\sigma = \surd\, \Sigma (X - X)^2 / n$$

σ is a Greek letter "Sigma"

4.1.6 Variance

The variance of a set of scores is denoted by $\sigma^2$and is defined as

$$\sigma^2 = \Sigma (X - \overline{X})^2 / n$$

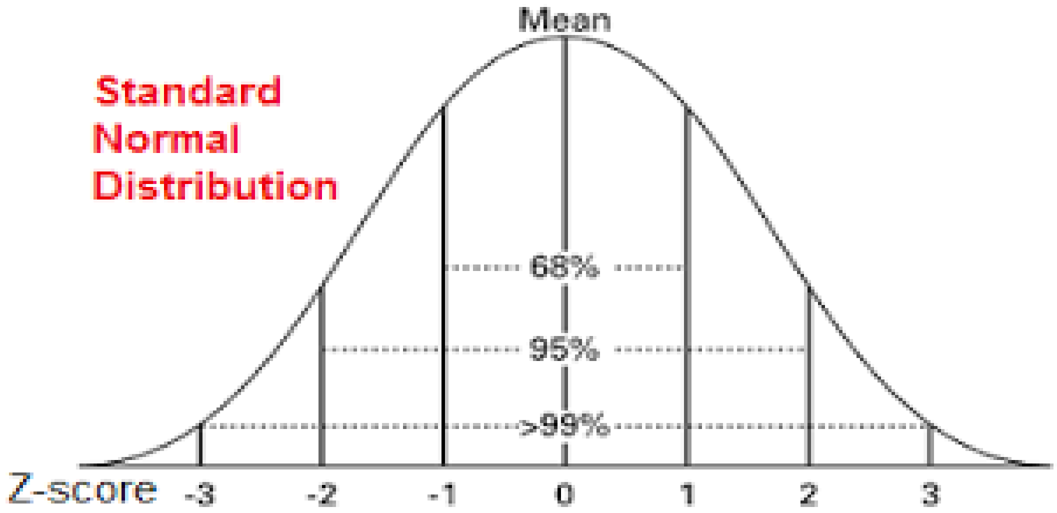
Where $\overline{X}$ is the mean, n is the number of data values and X stand for each of the scores, and Σ means add up all the values.

And alternate equivalent formula for variance is

$$\sigma^2 = (\Sigma X^2 / n) - X^2$$

4.2 Normal Curve

One way of presenting out how data are distributed is to plot them in a graph. If the data is evenly distributed, our graph will come across a curve. In statistics this curve is called a normal curve and in social sciences, it is called the bell curve. Normal or bell curved is distribution of data may naturally occur in several possible ways, with a number of possibilities for standard deviation (which could be from 1 to infinity). A standard normal curve has a mean of 0 and standard of 1. The larger the standard deviation, the flatter the curve will be and vice versa. A standard normal distribution is given below.



Source: Google Images

A normal curve has following properties.

i) The mean, median or mode are equal.
ii) The curve is symmetric at the center (i.e. around the mean).
iii) Exactly half of the values are to the left of the center and half to the right.
iv) The total area under the curve is 1.

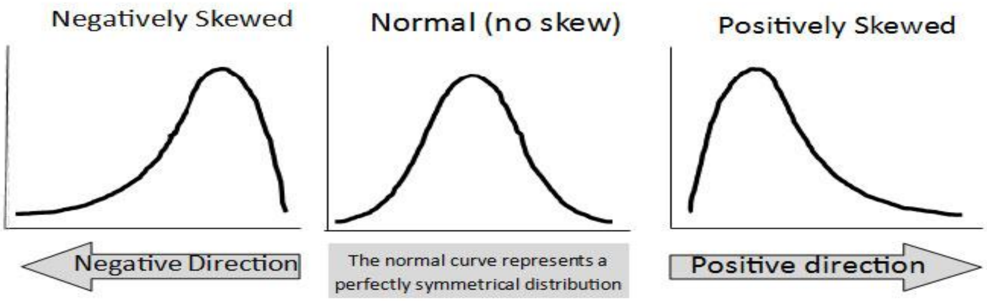
4.2.1 Numerical Measures of Shape

One of the fundamental tasks in any statistical analysis is to characterize the location and variability of a data set. Two important measures of shape, skewness and kurtosis, give us a more precise evaluation of the data. Measures of dispersion tell us about the variation of the data set, while skewness tells us about the direction of variation and kurtosis tells us the shape variation. Let us have a brief review of these measures of shape.

a) Skewness

Skewness tells us about the amount and direction of the variation of the data set. It is a measure of symmetry. A distribution or data set is symmetric if it looks the same to the left and right of the central point. If bulk of data is at the left i.e. the peak is towards left and the right tail is longer, we say that the distribution is skewed right or positively skewed.

On the other hand if the bulk of data is towards right or, in other words, the peak is towards right and the left tail is longer, we say that the distribution is skewed left or negatively skewed.If the skewness is equal to zero, the data are perfectly symmetrical. But it is quiet unlikely in real world.



Source: Google Images

Here are some rules of thumb:

i) If the skewness is less than – 1or greater than + 1, the distribution is highly skewed.

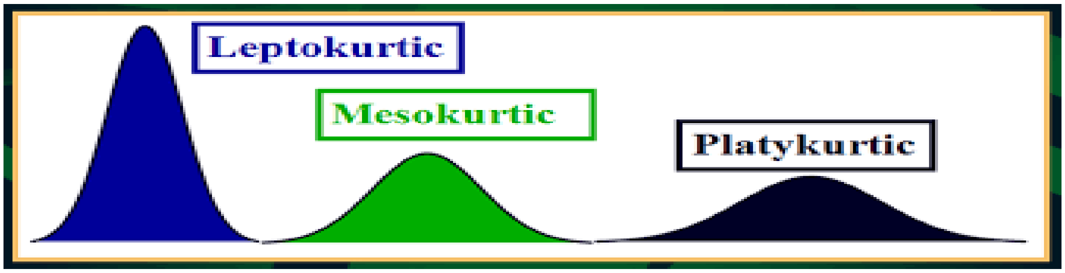
ii) If the skewness is between -1 and - $\frac{1}{2}$ or between + $\frac{1}{2}$ and + 1, the distribution is moderately skewed.

iii) If the skewness is between **-** $\frac{1}{2}$ and + $\frac{1}{2}$, the distribution is approximately skewed.

b) Kurtosis

Kurtosis is a parameter that describes the shape of variation. It is a measurement that tells us how the graph of the set of data is peaked and how high the graph is around the mean. In other words we can say that kurtosis measures the shape of the distribution, .i.e. the fatness of the tails, it focuses on how returns are arranged around the mean. A positive value means that too little data is in the tail and positive value means that too much data is in the tail. This heaviness or the lightness in the tail means that data looks more peaked of less peaked. Kurtosis is measured against the standard normal distribution. A standard normal distribution has a kurtosis of 3.

Kurtosis has three types, mesokurtic, platykurtic, and leptokurtic. If the distribution has kurtosis of zero, then the graph is nearly normal. This nearly normal distribution is called mesokurtic. If the distribution has negative kurtosis, it is called platykurtic. An example of platykurtic distribution is a uniform distribution, which has as much data in each tail as it does in the peak. If the distribution has positive kurtosis, it is called leptokurtic. Such distribution has bulk of data in the peak.



Source: Google Images

4.3 Co-Efficient of Variation

The coefficient of variation is another useful statistics for measuring dispersion of a data set. The coefficient of variation is

C.V = (s / $\overline{x}$) × 100

The coefficient of variation is invariant with respect to the scale of the data. On the other hand, standard deviation is not scale variant.

4.4 Self Assessment Questions

Q. 1 Write down the basic purpose of measure of central tendency.
Q. 2 Define range.
Q. 3 Write down the range of the following data.
12, 15, 35, 18, 21, 33, 18, 24, 48, 55, 36, 32, 17
Q. 4 What do you understand by mean deviation.
Q. 5 Define normal curve.
Q. 6 Write down the properties of normal curve.
Q. 7 Write down types of kurtosis

4.5 Activities

Take a cardboard. Cut it into 4x4 pieces, and:

i) Cut one piece into standard normal distribution shape and mention its name on it.
ii) Cut one piece into negatively skewed shape and mention its name on it.
iii) Cut one piece into positively skewed shape and mention its name on it.
iv) Cut one piece into no skewed shape and mention its name on it.
v) Cut one piece into mesokurtic shape and mention its name on it.
vi) Cut one piece into platykurtic shape and mention its name on it.
vii) Cut one piece into leptokurtic shape and mention its name on it.

4.6 Bibliography

Bartz, A. E. (1981). *Basic Statistical Concepts (2nd Ed.)*. Minnesota: Burgess Publishing Company

Deitz, T., & Kalof, L. (2009). *Introduction to Social Statistics*. UK: Wiley_-Blackwell

Gravetter, F. J., & Wallnau, L. B. (2002). *Essentials of Statistics for the Behavioral Sciences (4th Ed.)*. Wadsworth, California, USA.

UNIT-5

DESCRIPTIVE STATISTICS: MEASURES OF CENTRAL TENDENCY

Written By:
Salman Khalil Chaudhary

Reviewed By:
Dr. Rizwan Akram Rana

Introduction

In this unit you will study three main measures of central tendency – the mean, median and the mode. The main purpose of measures of central tendency is to identify the location of the center of various distributions. This helps us to get a better idea as to where the center of a distribution is located.

Merits and demerits of mean, median and mode are also discussed in the unit.

Objectives

After reading this unit, you will be able to:

1. write down the goals of measure of central tendency.
2. explain the characteristics of good measure of central tendency.
3. determine mean of a given set of data.
4. explain merits and demerits of mean.
5. define median.
6. explain procedures for determining median in case number of scores is even or odd.
7. explain merits and demerits of median.
8. calculate median of a given data.
9. define mode.
10. explain merits and demerits of mode.
11. calculate mode of a given data.

5.1 Introduction

Measures of central tendency (also referred as measures of center of central location) allow us to summarize data with a single value. It is a typical score among a group of scores (the midpoint). They give us an easy way to describe a set of data with a single number. This single number represents a value or score that generally is in the middle of the dataset.

The goal of the measure of central tendency is:

i) To condense data in a single value.
ii) To facilitate comparison between data.

Good measure of central tendency should be:

i) Be strictly defined.
ii) Be simple to understand and easy to calculate.
iii) Be capable of further mathematical treatment.
iv) Be based on all values of given data.
v) Have sampling stability.
vi) Not be unduly affected by extreme values.

Commonly used measures of central tendency are the mean, the median and the mode. Each of these indices is used with a different scale of measurement.

5.2 Mean

Mean is the most commonly used measure in educational research. It is appropriate for describing ratio or interval data. It can also be used for both continuous and discrete numeric data. It is the arithmetic average of the score. It is determined by adding up all the scores and then by the sum by the total number of scores. Suppose we have scores, 40, 85, 94, 62, 76, 66, 90, 59, 68, and 84. In order to find the mean of these scores we simply add all the scores, which comes to 724. Then divide this sum 10 (total number of scores). We will get 72.4, which is the mean score.

The formula for computing the mean is:
(Mean score) $X = \overline{\Sigma}X/n$

Where Σ represents "Sum of", X represents any raw score value, n represents total number of scores.

We can also define mean as mean is the amount each individual would get if the total (ΣX) were divided equally among all the individual members (n) in the distribution. In some other words we can say that the mean is the balance point for the distribution.

To interpret the as the "balance point or the center value", we can use the analogy of a seesaw. Its mean lies right at the center where the fulcrum keeps the board perfectly balanced. As the mean is based on every score or value of the dataset so it is influenced by outliers and skewed distribution. Also it cannot be calculated for categorical data as the values cannot be summed.

5.2.1 Merits of Mean

i) It is rigidly defined.
ii) It is easy to understand and calculate.
iii) It is used for further analysis and treatment.
iv) It is based upon all the values of the given data.
v) It is capable of further mathematical treatment.
vi) It is not much affected by sampling fluctuations.

5.2.2 Demerits of Mean

i) It cannot be calculated if any observation is missing.
ii) It cannot be calculated for data with open ended distribution.
iii) It may not lie in the middle of series, if series is skewed.
iv) It is affected by extreme values.
v) It cannot be located graphically.
vi) It may be number which is not present in the data.
vii) It can be calculated for the data representing qualitative values.

5.3 Median

Median is the middle value of rank order data. It divides the distribution in two halves (i.e. 50% of scores or observations on either side of median value). It means that this value separates higher half of the data set from the lower half. The goal of the median is to determine the precise midpoint of the distribution. Median is appropriate for describing ordinal data.

5.3.1 Procedure for Determining Median

When the number of scores is odd, simply arrange the scores in order (from lower to higher or from higher to lower). The median will be the middle score in the list. Consider the set of scores 2, 5, 7, 10, 12. The score “7”lies in the middle of the scores, so it is median.

When there is an even number of scores in the distribution, arrange the scores in order (from lower to higher or from higher to lower). The median will be the average of the middle two score in the list. Consider the set of scores 4, 6, 9, 14 16, 20. The average of the middle two scores 11.5 (i.e. 9+14/2 = 23/2 = 11.5) is the median of the distribution.

Median is less affected by outliers and skewed data and is usually preferred measure of central tendency when the distribution is not symmetrical. The median cannot be determined for categorical or nominal data.

5.3.2 Merits of Median

i) It is rigidly defined.
ii) It is easy to understand and calculate.
iii) It is not affected by extreme values.
iv) Even if the extreme values are not known median can be calculated.
v) It can be located just by inspection in many cases.
vi) It can be located graphically.
vii) It is not much affected by sampling fluctuations.
viii) It can be calculated by data based on ordinal scale.
ix) It is suitable for skewed distribution.
x) It is easily located in individual and discrete classes.

5.3.3 Demerits of Median

i) It is not based on all values of the given data.
ii) For larger data size the arrangements of the data in the increasing order is somewhat difficult process.
iii) It is not capable for further mathematical treatment.
iv) It is not sensitive to some change in the data value.
v) It cannot be used for further mathematical processing.

5.4 Mode

The mode is the most frequently occurring score in the distribution. Consider following data set.

25, 43, 39, 25, 82, 77, 25, 47.

The score 25 comes more frequently, so it is the mode. Sometimes there may be no single mode if no one value appears more than any other. There may be one mode (uni-modal), two modes (bi-model), three modes (tri-model), or more than three modes (multi-model).

Mode is useful when scores reflect a nominal scale of measurement. But along with mean and median it can also be used for ordinal, interval or ratio data. It can be located graphically by drawing histogram.

5.4.1 Merits of Mode

i) It is easy to understand and easy to calculate.
ii) It is not affected by extreme values.
iii) Even if the extreme values are not known mode can be calculated.
iv) It can be located just by inspection in many cases.
v) It can be located graphically.
vi) It is always present in the data.
vii) It is applicable for both quantitative and qualitative data.
viii) It is useful for methodological forecasts.

5.4.2 Demerits of Mode

i) It is not rigidly defined.
ii) It is not based upon all values of the given data.
iii) It is not capable of further mathematical calculation.
iv) There will be no mode if there is no common value in the data.
v) It cannot be used for further methodological processing.

5.5 Self-Assessment Question

Q. 1 What are the goals of measure of central tendency?
Q. 2 What are the characteristics of good measure of central tendency?
Q. 3 Define mean.
Q. 4 Calculate mean of a given set of data.
55, 45, 53, 62, 36
Q. 5 Write down merits and demerits of mean.
Q. 6 Define median.
Q. 7 Explain procedure for determining median if:
i) The number of scores is even.
ii) The number of scores is odd.
Q. 8 Write down merits and demerits of median.
Q. 9 Calculate median of the given data.
i) 42, 40, 51, 65, 82, 68, 77, 69, 80
ii) 38, 40, 61, 56, 90, 74, 72, 90, 49, 64
Q. 10 Define mode.
Q. 11 Write down merits and demerits of mode.
Q. 12 Calculate mode of the given data.
i) 65, 39, 66, 54, 33, 39. 55, 64, 38, 91, 72
ii) 38, 40, 61, 56, 90, 74, 72, 90, 49, 64
iii) 72, 74, 69, 68, 50, 56, 74, 42, 86, 44, 56, 72, 69

5.6 Activities

Discuss with your colleague and:
i) Try to make a list of merits and demerits of mean not given in the unit**.**
ii) Try to make a list of merits and demerits of median not given in the unit**.**
iii) Try to make a list of merits and demerits of mode not given in the unit**.**

5.7 Bibliography

Agresti, A. & Finlay, B. (1997). *Statistical Methods for Social Sciences*, (3rd *Ed.*). Prentice Hall.

Anderson, T. W., & Sclove, S. L. (1974). *Introductory Statistical Analysis*, Finland: Houghton Mifflin Company.

Argyrous, G. (2012). *Statistics for Research, with a guide to SPSS*. India: SAGE Publications.

Bartz, A. E. (1981). *Basic Statistical Concepts (2nd Ed.)*. Minnesota: Burgess Publishing Company.

Gravetter, F. J., & Wallnau, L. B. (2002). *Essentials of Statistics for the Behavioral Sciences (4th Ed.)*. Wadsworth, California, USA.

UNIT-6

INFERENTIAL STATISTICS

Written By:
Salman Khalil Chaudhary

Reviewed By:
Dr. Rizwan Akram Rana

Introduction

Inferential statistics is of vital importance in educational research. It is used to make inferences about the population on the bases of data obtained from the sample. It is also used to make judgments of the probability that an observed difference among groups is a dependable one or one that might have happened by chance in the study.

In this unit, you will study introduction, area, logic and importance of inferential statistics. Hypothesis testing, logic and process of hypothesis testing and errors in hypothesis are also discussed. In the last of the unit *t*-test, its types and general assumptions regarding the use of *t*-test are discussed.

Objectives

After reading this unit, you will be able to:

1. explain the term "Inferential Statistics".
2. explain the area of Inferential Statistics.
3. explain the logic of Inferential Statistics.
4. explain the Importance of Inferential Statistics in Educational Research.
5. tell, What Hypothesis Testing is.
6. explain the Logic of Hypothesis Testing.
7. explain the Uncertainty and errors in Hypothesis Testing.
8. explain *t*-test and its Types.

6.1 Introduction to inferential Statistics

Many statistical techniques have been developed to help researchers make sense of the data they have collected. These techniques are divided into two categories; descriptive and inferential. Descriptive statistics are the techniques that allow a researcher to quickly summarize the major characteristics of the data set. Inferential statistics, on the other hand, is set of techniques that allow a researcher to go a step further by helping a researcher uncover patterns or relationships in the data set, make judgment about data, or apply information about a smaller data set to a larger group. These techniques are part of the process of data analysis used by the researchers to analyze, interpret and make inferences about their results. In simple words we can say that inferential statistics helps researchers to make generalization about a population based on the data obtained from the sample. Since the sample is a small subset of the larger population, so the inferences made on the bases of the data obtained from sample cannot be free from errors. That is, we cannot say with 100% confidence that the characteristics of the sample accurately reflect the characteristics of the larger population. Hence only qualified inferences can be made, with a degree of certainty, which is often expressed in terms of probability (90% or 95% probability that the sample reflects the population).

Descriptive statistics only gives us the central values, dispersion or the variability of the data but inferential statistics leads us to take a decision about the whole population and in the end to any conclusion. Inferential statistics allows us to use what we have learnt from descriptive statistics. Inferential statistics enables us to infer from the data obtained the sample what the population might think.

6.1.1 Areas of Inferential Statistics

Inferential statistics has two broad areas

i) Estimating Parameter

This means taking a statistics from the sample data (e.g. the sample mean) and saying something about population parameter (e.g. the population mean).

ii) Hypothesis testing

This is where a researcher can use sample data to answer research questions.

Inferential statistics deals with two or more than two variables. If in an analysis there are two variables it is called bivariate analysis and if the variables are more than two it is called multivariate analysis. A number of different types of inferential statistics are in use. All of which depend of the type of variable i.e. nominal, ordinal, interval, and ratio. Although the type of statistical analysis is different for these variables, yet the main theme is the same we try to determine how one variable compare to another.

It should be noted that inferential statistics always talk in terms of probability. This can be made highly reliable by designing right experimental conditions. The inferences are always an estimate with a confidence interval. In some cases there is simply a rejection of hypothesis.

Several models are available in inferential statistics that help in the process of data analysis. A researcher should be careful while choosing any model. Because, choosing a wrong model may give wrong conclusions.

6.1.2 Logic of Inferential Statistics

Suppose a researcher wants to know the difference between the male and female students with respect to interest in learning English as a foreign language. He hypothesizes that the female students are more interested in learning English as a foreign language than the male students. To test the hypothesis he randomly selects 60 male students from a 1000 male students of English language course and 60 female students from a 1000 female students of English language course. All the students are given an attitude scale to complete. Now the researcher has two data sets: the attitude scores of male group and the attitude scores of female group. The design of the study is as shown:

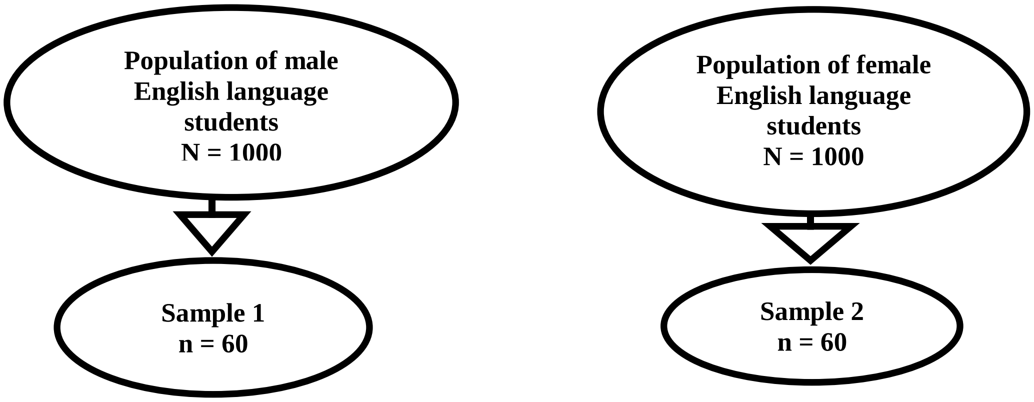

Population of male English language students N = 1000
Population of female English language students N = 1000
Sample 1 n = 60
Sample 2 n = 60

Fig: Selection of two samples from two different populations

The researcher wants to know whether the male population is different from female population – that is, will the mean score of the male group on attitude scale is different from the mean score of the female group? The researcher does not know the means of the two populations. He only has mean scores of two samples on which he has to rely on to provide information about the populations.

Now it comes in mind that is it reasonable to assume that each sample will give a fairly accurate picture of the whole population? It certainly is possible, because each sample was selected randomly from its population. On the other hand, the students in each sample are only a small portion of their respective population. It is only rare that a sample is absolutely identical to the population from which it is drawn, on given characteristics. The data the researcher obtains from two samples depends on the individual students selected to be in the sample. If another two samples were selected randomly their makeup would differ from previously selected samples. Their mean on the attitude scale would be different, and the researcher would end up with a different data set. How can the researcher be sure that any particular selected sample is a true representative of its population? Indeed he cannot. He needs some help to be sure that the sample is representative of the population and the results obtained from the sample data be generalized to whole population. Inferential statistics will help the researcher and allow him to make judgment about data and make generalization about a population based on the data obtained from the sample.

6.2 Importance of Inferential Statistics in Research

Inferential statistics is of vital importance in research in general and in educational research in particular. It allows us to use what we have learnt from descriptive statistics, and allow us to go beyond immediate data. Inferential statistics infers on the basis of sample data what the population might think. It helps us to make judgments about the probability that an observation is dependable or one that happened by chance in the study. It helps enables researchers to infer properties of a population based on data collected from a sample of individuals

Inferential statistics have larger value because these techniques offset problems associated with data collection. For example, time-cost factor associated with collection of data on the entire population may be prohibitive. The population may large and difficult to manage. In this case inferential statistics can prove to be invaluable to educational/social scientist.

6.3 Hypothesis Testing

It is usually impossible for a researcher to observe each individual in a population. Therefore, he selects some individual from the population as sample and collects data from the sample. He then uses the sample data to answer questions about the population. For this purpose, he uses some statistical techniques.

Hypothesis testing is a statistical method that uses sample data to evaluate a hypothesis about a population parameter (Gravetter & Wallnau, 2002).A hypothesis test is usually used in context of a research study. Depending on the type of research and the type of data, the details of the hypothesis test will change from on situation to another.

Hypothesis testing is a formalized procedure that follows a standard series of operations. In this way a researcher has a standardized method for evaluating the results of his research study. Other researchers will recognize and understand exactly how the data were evaluated and how conclusions were drawn.

6.3.1 Logic of Hypothesis Testing

According to Gravetter & Wallnau (2002) the logic underlying hypothesis testing is as follows:

i) First, a researcher states a hypothesis about a population. Usually, the hypothesis concerns the value of the population mean. For example, we might hypothesize that the mean IQ for the registered voters Pakistan is M = 100.

ii) Before a researcher actually selects a sample, he uses the hypothesis to predict the characteristics that the sample should have. For example, if he hypothesizes that the population mean IQ = 100, then he would predict that the sample should have a mean around 100. It should be kept in mind that the sample should be similar to the population but there is always a chance certain amount of error.

iii) Next, the researcher obtains a random sample from the population. For example, he might select a random sample of n = 200 registered voters to compute the mean IQ for the sample.

iv) Finally, he compares the obtained sample data with the prediction that was made from the hypothesis. If the sample mean is consistent with the prediction, he will conclude that the hypothesis is reasonable. But if there is big difference between the data and the prediction, he will decide that the hypothesis is wrong.

6.3.2 Four-Step Process for Hypothesis Testing

The process of hypothesis testing goes through following four steps.

i) Stating the Hypothesis

The process of hypothesis testing begins by stating a hypothesis about the unknown population. Usually, a researcher states two opposing hypotheses. And both hypotheses are stated in terms of population parameters.

The first and most important of two hypotheses is called *null hypothesis*. A null hypothesis states that the treatment has no effect. In general, null hypothesis states that there is no change, no effect, no difference – nothing happened. The null hypothesis is denoted by the symbol H_o (H stands for hypothesis and 0 denotes that this is zero effect).

The *null hypothesis* (H_o) states that in the general population there is no change, no difference, or no relationship. In an experimental study, null hypothesis (H_o) predicts that the independent variable (treatment) will have no effect on the dependent variable for the population.

The second hypothesis is simply the opposite of null hypothesis and it is called the *scientific or alternative hypothesis*. It is denoted by H_1. This hypothesis states that the treatment has an effect on the dependent variable.

The *alternative hypothesis* (H_1) states that there is a change, a difference, or a relationship for the general population. In an experiment, H_1 predicts that the independent variable (treatment) will have an effect on the dependent variable.

ii) Setting Criteria for the Decision

In a common practice, a researcher uses the data from the sample to evaluate the authority of null hypothesis. The data will either support or negate the null hypothesis. To formalize the decision process, a researcher will use null hypothesis to predict exactly what kind of sample should be obtained if the treatment has no effect. In particular, a researcher will examine all the possible sample means that could be obtained if the null hypothesis is true.

iii) Collecting data and computing sample statistics

The next step in hypothesis testing is to obtain the sample data. Then raw data are summarized with appropriate statistics such as mean, standard deviation etc. then it is possible for the researcher to compare the sample mean with the null hypothesis.

iv) Make a Decision

In the final step the researcher decides, in the light of analysis of data, whether to accept or reject the null hypothesis. If analysis of data supports the null hypothesis, he accepts it and vice versa.

6.3.3 Uncertainty and Error in Hypothesis Testing

Hypothesis testing is an inferential process. It means that it uses limited information obtained from the sample to reach general conclusions about the population. As a sample is a small subset of the population, it provides only limited or incomplete information about the whole population. Yet hypothesis test uses information obtained from the sample. In this situation, there is always the probability of reaching incorrect conclusion. Generally two kinds of errors can be made.

i) Type I Errors

A type I error occurs when a researcher rejects a null hypothesis that is actually true. It means that the researcher concludes that the treatment does have an effect when in fact the treatment has no effect.

Type I error is not a stupid mistake in the sense that the researcher is overlooking something that should be perfectly obvious. He is looking at the data obtained from the sample that appear to show a clear treatment effect. The researcher then makes a careful decision based on available information. He never knows whether a hypothesis is true or false.

The consequences of a type I error can be very serious because the researcher has rejected the null hypothesis and believed that the treatment had a real effect. it is likely that the researcher will report or publish the research results. Other researchers may try to build theories or develop other experiments based on false results.

ii) Type II Errors

A type II error occurs when a researcher fails to reject the null hypothesis that is really false. It means that a treatment effect really exists, but the hypothesis test has failed to detect it. This type of error occurs when the effect of the treatment is relatively small. That is the treatment does influence the sample but the magnitude of the effect is very small.

The consequences of Type II error are not very serious. In case of Type II error the research data do not show the results that the researcher had hoped to obtain. The researcher can accept this outcome and conclude that the treatment either has no effect or has a small effect that is not worth pursuing. Or the researcher can repeat the experiment with some improvement and try to demonstrate that the treatment does work. It is impossible to determine a single, exact probability value for a type II error.

Summarizing we can say that a hypothesis test always leads to one of two decisions.

i) The sample data provides sufficient evidence to reject the null hypothesis and the researcher concludes that the treatment has an effect.

ii) The sample data do not provide enough evidence to reject the null hypothesis. The researcher fails to reject the null hypothesis and concludes that the treatment does not appear to have an effect.

In either case, there is a chance that the data are misleading and the decision is wrong. The complete set of decision and outcome is shown in the following table.

Table: 6.1

Possible outcome of statistical decision

		Actual Situation	
		No effect, H_o true	Effect exists, H_o false
Experimenter's Decision	Reject H_o	Type I Error	Decision Correct
	Retain H_o	Decision Correct	Type II Error

Source: Gravetter & Wallnau, (2002)

6.4 T-Test

A t-test is a useful statistical technique used for comparing mean values of two data sets obtained from two groups. The comparison tells us whether these data sets are different from each other. It further tells us how significant the differences are and if these differences could have happened by chance. The statistical significance of t-test indicates whether or not the difference between the mean of two groups most likely reflects a real difference in the population from which the groups are selected.

t-tests are used when there are two groups (male and female) or two sets of data (before and after), and the researcher wishes to compare the mean score on some continuous variable.

6.4.1 Type of T-Test

There are a number of t-test available but two main types independent sample t-test and paired sample *t*-test are most commonly used. Let us deal with these types in some detail.

i) Independent sample t-test

Independent sample t-test is used when there are two different independent groups of people and the researcher is interested to compare their scores. In this case the researcher collects information from two different groups of people on only one occasion.

ii) Paired sample *t*-test

Paired sample *t*-test is also called repeated measures. It is used the researcher is interested in comparing changes in the scores of the same group tested at two different occasions.

Here at this level it is necessary to know some general assumptions regarding use of t-test. The first assumption regarding t-test concerns the scale of measurement. It means that it is assumed that the dependent variable is measured at interval or ratio scale. The second assumption made is that of a simple random sample, that the data is collected from a representative, randomly selected portion of the total population. The third assumption is that the data, when plotted, results in a normal distribution i.e. in bell-shaped distribution curve. The fourth assumption is that the observation that make up data must independent of one another. That is, each observation or measurement must not be influences by any other observation or measurement. The fifth assumption is that a reasonably large sample size is used. A large sample size means that the distribution of results should approach a normal bell-shaped curve. The final assumption is homogeneity of variance. Variance will be homogeneous or equal when the standard deviation of samples is approximately equal.

6.5 Self-Assessment Questions

Q. 1 What do you mean by inferential statistics?
Q. 2 Write down the area of inferential statistics.
Q. 3 What is the importance of inferential statistics in educational research?
Q. 4 What do mean by hypothesis testing?
Q. 5 Briefly state the logic behind hypothesis testing.
Q. 6 What are type I and type II errors?
Q. 7 In what situation will you use independent sample *t*-test for your data?
Q. 8 In what situation will you use paired sample *t*-test for your data?
Q. 9 What do you know about:
    a) An independent sample *t*-test.
    b) A paired sample *t*-test.

6.6 Activities

1. Suppose we exclude inferential statistics from our research. What will happen? Write down a few lines.
2. You have scores of two different groups of students and you have to compare the scores. Discuss with your colleague and select appropriate statistical test.

6.7 Bibliography

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to Design and Evaluate in Education*. (8th Ed.) McGraw-Hill, New York

Frey, L. R., Carl H. B., & Gary L. K. (2000). *Investigating Communication: An Introduction to Research Methods.*2nd Ed. Boston: Allyn and Bacon

Gravetter, F. J., & Wallnau, L. B. (2002). *Essentials of Statistics for the Behavioral Sciences (4th Ed.)*. Wadsworth, California, USA.

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Albany: Duxbury Press.

Pallant, J. (2005). *SPSS Survival Manual – A step by step guide to data analysis using SPSS for Windows (Version 12)*. Australia: Allen & Unwin.

UNIT-7

INFERENTIAL STATISTICS: CORRELATION AND REGRESSION

Written By:
Prof. Dr. Nasir Mahmood

Reviewed By:
Dr. Rizwan Akram Rana

Introduction

A correlation is a relationship between two variables. The purpose of using correlation in research is to determine the degree to which a relationship exists between two or more variables. Correlation is important in research because several hypotheses are stated in terms of correlation or lack of correlation between two variables, so correlational studies are directly related to such hypotheses.

Regression is used when the relationship includes a dependent variable and one or more independent variables. It helps us understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships.

Owing to the importance of correlation and regression in research, these are given in this unit in detail.

Objectives

After reading this unit, you will be able to:

1. explain correlation.
2. explain where and why to use correlation.
3. explain what considerations should be kept in mind while interpreting correlation.
4. explain Pearson and Spearman correlation method.
5. explain the situations in which Spearman correlation can be used.
6. explain Regression.
7. explain why we use regression analysis.
8. explain types of regression.
9. explain *p*-value.

7.1 Correlation

Correlation is a statistical technique used to measure and describe relationship between two variables. These variables are neither manipulated nor controlled, rather they simply are observed as they naturally exist in the environment. Suppose a researcher is interested in relationship between number of children in a family and IQ of the individual child. He would take a group of students coming from different families. Then he simply observe or record the number of children in a family and then measure IQ score of each individual student same group. He will neither manipulate nor control any variable. Correlation requires two separate scores for each individual (one score from each of two variables). These scores are normally identified as X and Y and can be presented in a table or in a graph.

7.1.2 Characteristics of Relationship that Correlation Measures

A correlation measures three characteristics of the relationship between X and Y. These are:

i) The Direction of the Relationship

The direction of the relationship can be classified into two basic categories: positive and negative.

In a positive correlation both variables tend to change into same direction. When variable X increases, the variable Y also increases. And if the variable X decreases, the variable Y also decreases. In other words we can say that both variables are directly proportional to each other.

In a negative correlation both variables do not tend to change into same direction. They go in opposite direction of each other. When the variable X increases, the variable Y decreases. And if the variable X decreases, the variable Y increases. In other words we can say that both variables are indirectly proportional to each other.

The direction of the relationship is identified by the sign of the correlation. A positive sign (+) indicates positive relationship. A negative sign (−) indicates negative relationship.

- 1---------------------- - .5 ---------------------- 0 ---------------------- .5 ---------------------- + 1

Strong negative   moderate negative   No   moderate positive

Strong positive

relationship   relationship   relationship   relationship   relationship

ii) The form of the Relationship

The form of correlation measures how well the data fit the specific form being considered. For example, a linear correlation measures how well the data points fit on a straight line

iii) The Degree of the Relationship

The degree of relationship is measured by the numerical value of the correlation. This value varies from 1.00 to – 1.00. A perfect correlation is always identified by a correlation of 1.00 and indicates a perfect fit. + 1.00 will indicate perfect positive correlation and –1.00 will indicate perfect negative correlation. A correlation of 0 indicates no correlation or no fit at all.

7.2 The Pearson Correlation

The most commonly used correlation is the Pearson Correlation. It is also known as Pearson product-moment Correlation. It measures the degree and the direction of linear relationship of between two variables. It is denoted by r, and r = degree to which X and Y vary together / degree to which X and Y vary separately = co-variability of X and Y / variability of X and Y vary separately

To calculate the Pearson correlation r we use the formula

$$r = \frac{\text{SP}}{\sqrt{\text{SSx SSy}}}$$

where SP is the sum of the product of deviation.

Two formulas (definitional and computational) are available to calculate the sum of square of product. Both formulas are given in the following box.

1.	The definitional formula is $\text{SP} = \sum (\text{X} - \text{X})(\text{Y} \overline{-} \text{Y})$
2.	The computational formula is $SP = \sum \text{XY} - \frac{\sum \text{X} \sum \text{Y}}{\text{n}}$

SS is sum of squares, SS_x is the sum of squares of the variable X and SS_y is the sum of squares of variable Y. In the following lines different formulas are given to calculate SS_x and SS_y. These formulas are categorized as definitional and computational. The definitional formulas for sum of squares of variable X are:

$$SS_x = \sum (\overline{X} - X)^2$$

The computational formulas for sum of squares of variable X are:

$$SS_x = \sum X^2 - \frac{(\sum \text{X})2}{\text{n}}$$

The definitional formulas for sum of squares of variable Y are:

$$SS_y = \sum (\overline{Y} - Y)^2$$

The computational formulas for sum of squares of variable Y are:

$$SS_y = \sum Y^2 - \frac{(\sum \text{Y})2}{\text{n}}$$

It should be kept in mind that whichever formula one uses, it will yield similar result.

7.2.2 Using and Interpreting Pearson Correlation

First let us have a brief discussion about where and why we use correlation. The discussion follows under following headings.

i) Prediction

If two variables are known to be related in some systematic way, it is possible to use one variable to make prediction about the other. For example, when a student seeks admission in a college, he is required to submit a great deal of personal information, including his scores in SSC annual/supplementary examination. The college officials want this information so that they can predict that student's chance of success in college.

ii) Validity

Suppose a researcher develops a new test for measuring intelligence. It is necessary that he should show that this new test valid and truly measures what it claims to measure. One common technique for demonstrating validity is to use correlation.

If newly constructed test actually measures intelligence, then the scores on this test should be related to other already established measures of intelligence – for example standardized IQ tests, performance on learning tasks, problem-solving ability, and so on. The newly constructed test can be correlated to each of these measures to demonstrate that the new test is valid.

iii) Reliability

Apart from determining validity, correlations are also used to determine reliability. A measurement procedure is reliable if it produces stable and consistent measurement. It means a reliable measurement procedure will produce the same (or nearly same) scores when the same individuals are measured under the same conditions. One common way to evaluate reliability is to use correlations to determine relationship between two sets of scores.

iv) Theory Verification

Many psychological theories make specific predictions about the relationship between two variables. For example, a theory may predict a relationship between brain size and learning ability; between the parent IQ and the child IQ etc. In each case, the prediction of the theory could be tested by determining the correlation between two variables.

Now let us have a few words on interpreting correlation. For interpreting correlation following consideration should be kept in mind.

i) Correlation simply describes a relationship between two variables. It does not explain why two variables are related. That is why correlation cannot be interpreted as a proof of cause and effect relationship between two variables.

ii) The value of the correlation cannot be affected by range of scores represented in the data.

iii) One or two extreme data points, often called outliers, can have a dramatic effect on the value of the correlation.

iv) When judging how good a relationship is, it is tempting to focus on the numerical value of the correlation. For example, a correlation of + 5 is halfway between 0 and 1.00 and therefore appears to represent a moderate degree of relationship. Here it should be noted that we cannot interpret correlation as a proportion. Although a correlation of 1.00 means that there is a 100% perfectly predictable relationship between variables X and Y; but a correlation of .5 does not mean that we can make a prediction with 50% accuracy. The appropriate process of describing how accurately one variable predicts the other is to square the correlation. Thus a correlation of $r = .5$ provides $r^2 = .5^2 = .25$, 25% accuracy. (The value r^2 is called coefficient of determination because it measures the proportion of variability in one variable that can be determined from the relationship with the other variable).

7.3 The Spearman Correlation

The most commonly used measure of relationship is the Pearson correlation. It measures the degree of linear relationship between two variables and is used with interval or ratio data. However other measures of correlation have been developed for non-linear relationship and for other type of data (or scale of measurement). One such measure is the Spearman Correlation. The Spearman correlation is used in two situations.

i) The Spearman correlation is designed to measure the relationship between variables measured on an ordinal scale of measurement.

ii) The Spearman correlation is used when the researcher wants to measure the consistency of a relationship between the variables X and Y. In this case the original scores are first converted into ranks, and then Spearman correlation is used to measure the relationship for the ranks. Incidentally, when there is consistently one-directional relationship between two variables, the relationship is said to be monotonic. Thus, the Spearman correlation can be used to measure the degree of monotonic relationship between two variables.

As the Pearson correlation measures the degree of linear relationship between two variables, the spearman correlation measures the consistency of relationship. It can be used as a valuable alternative of Pearson correlation even when the original raw scores are on an interval or ratio scale. Generally Spearman correlation is computed by using Pearson correlation formula, i.e.

$$\mathrm{r_s} = \frac{SP}{\sqrt{\mathrm{SSx\ SSy}}}$$

Another formula is also used for calculating Spearman correlation. It is:

$$\mathrm{r_s} = 1 - \frac{6\sum D^2}{\sqrt{\mathrm{SSx\ SSy}}}$$

where D is the difference between X rank and Y rank for each individual. Again this formula will yield the same result as Pearson correlation formula.

7.4 Regression

A correlation quantifies the degree and direction to which two variables are related. It does not fit a line through the data points. It does not have to think about the cause and effect. It does not natter which of the two variables is called dependent and which is called independent.

On the other hand regression finds the best line that predicts dependent variables from the independent variable. The decision of which variable is calls dependent and which calls independent is an important matter in regression, as it will get a different best-fit line if we exchange the two variables, i.e. dependent to independent and independent to dependent. The line that best predicts independent variable from dependent variable will not be the same as the line that predicts dependent variable from independent variable.

Let us start with the simple case of studying the relationship between two variables X and Y. The variable Y is dependent variable and the variable X is the independent variable. We are interested in seeing how various values of the independent variable X predict corresponding values of dependent Y. This statistical technique is called regression analysis. We can say that regression analysis is a technique that is used to model the dependency of one dependent variable upon one independent variable. Merriam-Webster online dictionary defines regression as a functional relationship between two or more correlated variables that is often empirically determined from data and is used especially to predict values of one variable when given variables of others. According to Gravetter & Wallnua (2002), regression is a statistical technique for finding the best-fitting straight line for a set of data is called regression, and the resulting straight line is called regression line.

7.4.1 Objectives of Regression Analysis

The regression analysis is used to explain variability in dependent variable by mean of one or more of independent variables and to analyze relationships among variables to answer the question of how much dependent variable changes with the changes in the independent variables and to forecast or predict the value of dependent variable based on the values of the independent variable.

The primary objective of the regression is to develop a relationship between a response variable and the explanatory variable for the purpose of prediction, assumes that a functional relationship exists, and alternative approaches are superior.

7.4.2 Why do we use Regression Analysis?

Regression analysis estimates the relationship between two or more variables and is used for forecasting or finding cause and effect relationship between the variables. There are multiple benefits of using regression analysis. These are as follows:

i) It indicates the significant relationships between dependent and the independent variables.

ii) It indicates the strength of impact of multiple independent variables on a dependent variable.

iii) It allows us to compare the effects of variables measured on different scales.

These benefits help a researcher to estimate and evaluate the best set of variables to be used for building productive models.

7.4.3 Types of Regression

Commonly used types of regression are:

i) Linear Regression

It is the most commonly used types of regression. In this technique the dependent variable is continuous and the independent variable can be continuous or discrete and the nature of regression line is linear. Linear regression establishes a relationship between dependent variable (Y) and one or more independent variables (X) using best fit straight line (also known as regression line).

ii) Logistic Regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with the dichotomous (binary) variable. Like all regression analysis, the logistic regression is a predictive analysis. It is used to describe and explain relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio level independent variables.

iii) Polynomial Regression

It is a form of regression analysis in which the relationship between independent variable X and dependent variable Y is modeled as an n^{th} degree polynomial in x. this type of regression fits a non-linear relationship between the values of X with the corresponding values of Y.

iv) Stepwise Regression

It is a method of fitting regression model in which the choice of predictive variables is carried out by an automatic procedure. In each step, a variable is considered for addition or subtraction from the set of explanatory variables based on some pre-specified criteria. The general idea behind this procedure is that we build our regression model from a set of predictor variable by entering and removing predictors in our model, in a stepwise manner, until there is no justifiable reason to enter or remove any more.

v) Ridge Regression

It is a technique for analyzing multiple regression data that suffer from multicollinearity (independent variables are highly correlated). When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so that they may be far from the true value. By adding the degree of bias to the regression estimates, ridge regression reduces the standard errors.

vi) LASSO Regression

LASSO or lasso stands for Least Absolute Shrinkage and Selection Operator. It is a method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. This type of regression uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean.

vii) Elastic Net Regression

This type of regression is a hybrid of lasso and ridge regression techniques. It is useful when there are multiple features which are correlated.

7.5 P-Value

The *p*-value is the level of marginal significance within a statistical hypothesis test representing the probability of occurrence of a given event. This value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected.

In other words we can say that p-value is the calculated probability or the probability of finding the observed or more extreme results when the null hypothesis is true. P-value is also described in terms of rejecting null hypothesis when it is actually true.

A p-value is used in hypothesis testing to help researcher support or reject the null hypothesis. It is evidence against the null hypothesis. The smaller p-value is the stronger the evidence to reject the null hypothesis.

In conducting tests of statistical significance (such as t-tests and ANOVA), a researcher answers this central question: if the null hypothesis was true in the population (that is, if there is really no difference between groups and no treatment effect), what is the probability of obtaining the results that we observed in our experiment? The key outcome of this type of inferential statistical tests is a p-value. This value is the probability of obtaining the same results as previously observed.

If the p-value gets lower (i.e. closer to 0% and farther away from 100), a researcher is more inclined to reject the null hypothesis and accept the research hypothesis.

A relatively simple way to interpret p-value is to think of them as representing how likely a result would occur by chance. For a calculated p-value of .01, we can say that the observed outcomes would be expected to occur by chance only 1 in 100 times in repeated tests on different samples of the population. Similarly a p-value of .05 would represent the expected outcome to occur by chance only 5 times out of 100 times in repeated tests and a p-value of .001 would represent the expected outcome to occur by chance only once if the same treatment is repeated for 1000 times on different samples of the population. In case of p-value .01, the researcher is 99% confident of getting similar results if same test is repeated for 100 times. Similarly in case of p-value .05, the researcher is 95% confident and in case of p-value .001, he is 999% confident of getting similar results if same test is repeated for 100 times and 1000 times respectively.

7.6 Self-Assessment Questions

Q. 1 Briefly explains what you understand by “correlation”.
Q. 2 Write down where and why to use correlation?
Q. 3 Write down the considerations that should be kept in mind while interpreting correlation.
Q. 4 Which formula is used to calculate Pearson correlation?
Q. 5 Which formula is used to calculate Spearman correlation?
Q. 6 What do you understand by “regression”?
Q. 7 Why do we use regression analysis?
Q. 8 Write down the types of regression.
Q. 9 Write down a brief note on *p*-value?

7.7 Activities

1. Think and make a list of using correlation.
2. Enlist the consideration that you will keep in mind while using correlation.
3. Think and write primary objective of regression analysis.

7.8 Bibliography

Argyrous, G. (2012). *Statistics for Research, with a guide to SPSS*. India: SAGE Publications.

Bartz, A. E. (1981). *Basic Statistical Concepts (2nd Ed.)*. Minnesota: Burgess Publishing Company

Deitz, T., & Kalof, L. (2009). *Introduction to Social Statistics*. UK: Wiley_-Blackwell

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to Design and Evaluate in Education*. (8th Ed.) McGraw-Hill, New York

Frey, L. R., Carl H. B., & Gary L. K. (2000). *Investigating Communication: An Introduction to Research Methods.*2nd Ed. Boston: Allyn and Bacon

Gravetter, F. J., & Wallnau, L. B. (2002). *Essentials of Statistics for the Behavioral Sciences (4th Ed.)*. Wadsworth, California, USA.

UNIT-8

INFERENTIAL STATISTICS: ANOVA

Written By:
Prof. Dr. Nasir Mahmood

Reviewed By:
Dr. Rizwan Akram Rana

Introduction

Analysis of Variance (ANOVA) is a statistical procedure used to test the degree to which two or more groups vary or differ in an experiment. This unit will give you an insight of ANOVA, its logic, one-way ANOVA, its assumptions, logic and procedure. F-distribution, interpretation of F-distribution and multiple procedures will also be discussed.

Objectives

After reading this unit you will be able to:

1. explain what ANOVA is.
2. write down the logic behind using ANOVA.
3. explain what F-distribution is.
4. explain logic behind one-way ANOVA.
5. explain the assumptions underlying one way ANOVA.
6. explain multiple comparison procedures.

8.1 Introduction to Analysis of Variance (ANOVA)

The t-tests have one very serious limitation – they are restricted to tests of the significance of the difference between only two groups. There are many times when we like to see if there are significant differences among three, four, or even more groups. For example we may want to investigate which of three teaching methods is best for teaching ninth class algebra. In such case, we cannot use t-test because more than two groups are involved. To deal with such type of cases one of the most useful techniques in statistics is analysis of variance (abbreviated as ANOVA). This technique was developed by a British Statistician Ronald A. Fisher (Dietz & Kalof, 2009; Bartz, 1981)

Analysis of Variance (ANOVA) is a hypothesis testing procedure that is used to evaluate mean differences between two or more treatments (or population). Like all other inferential procedures. ANOVA uses sample data to as a basis for drawing general conclusion about populations. Sometime, it may appear that ANOVA and *t*-test are two different ways of doing exactly same thing: testing for mean differences. In some cased this is true – both tests use sample data to test hypothesis about population mean. However, ANOVA has much more advantages over t-test. *t*-tests are used when we have compare only two groups or variables (one independent and one dependent). On the other hand ANOVA is used when we have two or more than two independent variables (treatment). Suppose we want to study the effects of three different models of teaching on the achievement of students. In this case we have three different samples to be treated using three different treatments. So ANOVA is the suitable technique to evaluate the difference.

8.1.1 Logic of ANOVA

Let us take a hypothetical data given in the table.

Table 8.1

Hypothetical Data from an Experiment examining learning performance under three Temperature condition

Treatment 1	Treatment 2	Treatment 3
50^o	70^o	90^o
Sample 1	Sample 2	Sample 3
0	4	1
1	3	2
3	6	2
1	3	0
0	4	0
$\overline{X} = 1$	$\overline{X} = 4$	$\overline{X} = 1$

There are three separate samples, with n = 5 in each sample. The dependent variable is the number of problems solved correctly

These data represent results of an independent-measure experiment comparing learning performance under three temperature conditions. The scores are variable and we want to measure the amount of variability (i.e. the size of difference) to explain where it comes from. To compare the total variability, we will combine all the scores from all the separate samples into one group and then obtain one general measure of variability for the complete experiment. Once we have measured the total variability, we can begin to break it into separate components. The word analysis means breaking into smaller parts. Because we are going to analyze the variability, the process is called analysis of variance (ANOVA). This analysis process divides the total variability into two basic components:

i) Between-Treatment Variance

Variance simply means difference and to calculate the variance is a process of measuring how big the differences are for a set of numbers. The between-treatment variance is measuring how much difference exists between the treatment conditions. In addition to measuring differences between treatments, the overall goal of ANOVA is to evaluate the differences between treatments. Specifically, the purpose for the analysis is to distinguish is to distinguish between two alternative explanations.

a) The differences between the treatments have been caused by the treatment effects.

b) The differences between the treatments are simply due to chance.

Thus, there are always two possible explanations for the variance (difference) that exists between treatments

1) **Treatment Effect:** The differences are caused by the treatments. For the data in table 8.1, the scores in sample 1 are obtained at room temperature of 50^o and that of

sample 2 at 70^o. It is possible that the difference between sample is caused by the difference in room temperature.

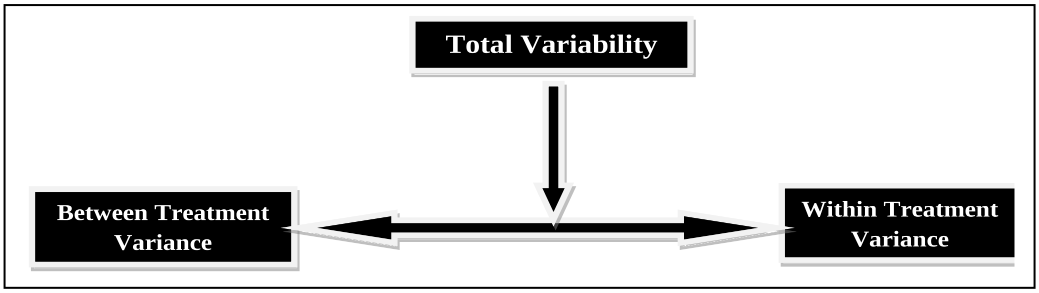
2) **Chance:** The differences are simply due to chance. It there is no treatment effect, even then we can expect some difference between samples. The chance differences are unplanned and unpredictable differences that are not caused or explained by any action of the researcher. Researchers commonly identify two primary sources for chance differences.

- ***Individual Differences***
  Each participant of the study has its own individual characteristics. Although it is reasonable to expect that different subjects will produce different scores, it is impossible to predict exactly what the difference will be.
- ***Experimental Error***
  In any measurement there is a chance of some degree of error. Thus, if a researcher measures the same individuals twice under same conditions, there is greater possibility to obtain two different measurements. Often these differences are unplanned and unpredictable, so they are considered to be by chance.

Thus, when we calculate the between-treatment variance, we are measuring differences that could be either by treatment effect or could simply be due to chance. In order to demonstrate that the difference is really a treatment effect, we must establish that the differences between treatments are bigger than would be expected by chance alone. To accomplish this goal, we will determine how big the differences is when there is no treatment effect involved. That is, we will measure how much difference (variance) occurred by chance. To measure chance differences, we compute the variance within treatments

ii) Within-Treatment Variance

Within each treatment condition, we have a set of individuals who are treated exactly the same and the researcher does not do anything that would cause these individual participants to have different scores. For example, in table 8.1 the data shows that five individuals were treated at a 70^o room temperature. Although, these five students were all treated exactly the same, there scores are different. Question is why are the score different? A plain answer is that it is due to chance. Figure 8.1 shows the overall analysis of variance and identifies the sources of variability that are measures by each of two basic components.



Measures Differences Measures Differences
due to: due to:

i. Treatment Effect i. Chance
ii. Chance

Fig: 8.1 The independent-measures analysis of variance partition or analyses, the total variability into two components: variance between treatment and variance within treatment.

8.2 The F-Distribution

After analyzing the total variability into two basic components (between treatment and within treatment), the next step is to compare them. The comparison is made by computing a statistics called f-ratio. For independent measure ANOVA, the F-ratio is calculated using the formula:

$$F = \frac{variance\ between\ treatment}{variance\ within\ treatment}$$

$$F = \frac{treatment\ effect + difference\ due\ to\ chance}{difference\ due\ to\ chance}$$

The value obtained for F-ratio will help determine whether or not any treatment effects exist. Consider above stated two possibilities.

1. When the treatment has no effect, then the difference between the treatments will be entirely due to chance. In this case the numerator and the denominator of F distribution are both measuring the same chance differences. Then F-ratio should have a value equal to 1.00. In terms of formula' we have

$$\mathrm{F} = \frac{0 + difference\ due\ to\ chance}{difference\ due\ to\ chance}$$

$$= \frac{difference\ due\ to\ chance}{difference\ due\ to\ chance}$$

$$= 1.00$$

The F-ratio equal to 1.00 indicates that the differences between treatments are about the same as the difference expect by chance. So, when F-ratio is equal to 1.00, we will conclude that there is no evidence to suggest that the treatment has any effect.

2. When the treatment does have an effect, then between-treatments differences (numerator) should be larger than chance (denominator). In this case numerator of F-ratio should be considerably larger than the denominator, and we should obtain F-ratio larger than 1.00. Thus, a large F-ratio indicates that the difference between are greater than chance; that is the treatment does have a significant effect.

8.2.1 Interpretation of the F-Statistic

The denominator in the F-statistic normalizes our estimate of the variance assuming that Ho is true. Hence, if F = 2, then our sample has two times as much variance as we would expect if Ho were true. If F = 10, then our sample has 10 times as much variance as we would expect if Ho were true. Ten times is quite a bit more variance than we would expect. In fact, for denominator degrees of freedom larger than 4 and any number of numerator degrees of freedom, we would reject Ho at the 5% level with an F-statistic of 10.

8.3 One Way ANOVA (Logic and Procedure)

The one way analysis of variance (ANOVA) is an extension of independent two-sample t-test. It is a statistical technique by which we can test if three or more means are equal. It tests if the value of a single variable differs significantly among three or more level of a factor. We can also say that one way ANOVA is a procedure of testing hypothesis that K population means are equal, where $K \geq 2$. It compares the means of the samples or groups in order to make inferences about the population means. Specifically, it tests the null hypothesis:
Ho : $\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$

Where μ = group mean and k = number of groups

If one way ANOVA yields statistically significant result, we accept the alternate hypothesis (HA), which states that there are two group means that are statistically significantly different from each other. Here it should be kept in mind that one way ANOVA cannot tell which specific groups were statistically significantly different from each other. To determine which specific groups are different from each other, a researcher will have to use post hoc test.

As there is only one independent variable or factor in one way ANOVA so it is also called single factor ANOVA. The independent variable has nominal levels or a few ordinal levels. Also, there is only one dependent variable and hypotheses are formulated about the means of the group on dependent variable. The dependent variable differentiates individuals on some quantitative dimension.

8.3.1 Assumptions Underlying the One Way ANOVA

There are three main assumptions

i) Assumption of Independence

According to this assumption the observations are random and independent samples from the populations. The null hypothesis actually states that the samples come from populations that have the same mean. The samples must be random and independent if they are to be representative of the populations. The value of one observation is not related to any other observation. In other words, one individual's score should not provide any clue as to how any of the other individual should score. That is, one event does not depend on another.

A lack of assumption of independence leads to most serious consequences. If this assumption is violated, one way ANOVA will be inappropriate to statistic,

ii) Assumption of Normality

The distributions of the population from which the samples are selected are normal. This assumption implies that the dependent variable is normally distributed in each of the groups.

One way ANOVA is considered a robust test against the assumption of normality and tolerates the violation of this assumption. As regards the normality of grouped data, the one way ANOVA can tolerate data that is normal (skewed or kurtotic distribution) with only a small effect on I error rate. However, platykurtosis can have profound effect when group sizes are small. This leaves a researcher with two options:

i) Transform data using various algorithms so that the shape of the distribution becomes normally distributed. Or

ii) Choose nonparametric Kruskal-Wallis H Test which does not require the assumption of normality. (This test is available is SPSS).

iii) Assumptions of Homogeneity of Variance

The variances of the distribution in the populations are equal. This assumption provides that the distribution in the population have the same shapes, means, and variances; that is, they are the same populations. In other words, the variances on the dependent variable are equal across the groups.

If assumption of homogeneity of variances has been violated then tow possible tests can be run.

i) Welch test, or

ii) Brown and Forsythe test

Alternatively, Kruskal-Wallis H Test can also be used. All these tests are available in SPSS.

8.3.2 Logic Behind One Way ANOVA

In order to test pair of sample means differ by more than would be expected by chance, we might conduct a series of t-tests on K sample means – however, this approach has a major problem, i.e.

When we use a t-test once, there is a chance of Type I error. The magnitude of this error is usually 5%. By running two tests on the same data we will have increased his chance of making error to 10%. For the third administration, it will be 15%, and so on. These are unacceptable errors. The number of t-tests needed to compare all possible means would be:

$$\frac{K(K-1)}{2}$$

Where K = Number of means

When more than one t-test is run, each at a specific level of significance such as α = .05, the probability of making one or more Type I error in a series of t-test is greater than α. The increased number of Type I error is determined as:

$1 – (1 - \alpha)^{c}$

Where	α	=	level of significance for each separate t-test
	c	=	number of independent t-test

An ANOVA controls the chance for these errors so that the type I error remains at 5% and a researcher can become more confident about the results.

8.3.3 Procedure for Using ANOVA

In using ANOVA manually we need first to compute a total sum of squares (SS_{total}) and then partition this value into two components: between treatments and within treatments. This analysis is outlined in Fig 8.2

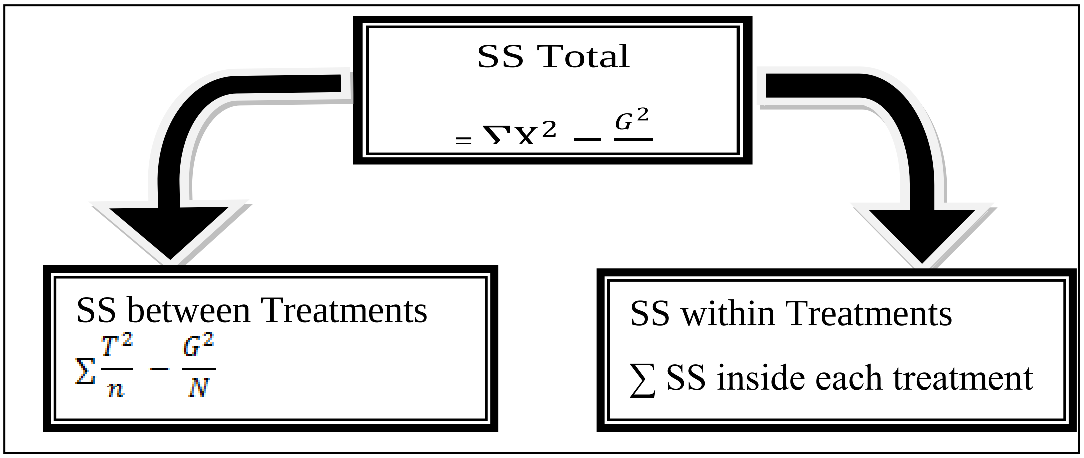

SS Total
$= \sum X^2 - \frac{G^2}{}$
SS between Treatments
$\sum \frac{T^2}{n} - \frac{G^2}{N}$
SS within Treatments
∑ SS inside each treatment

Fig: 3 partitioning the total sum of square (SS Total) for the independent measure ANOVA

1) **The Total Sum of Squares (SS $_{\textbf{Total}}$)**
It is the total sum of square for the entire set of N scores. It can be calculated using computational formula for SS:
SS $_{\text{Total}}$ = $\sum X^2 - \frac{(\sum X)^2}{N}$
But $(\sum X)^2 = G^2$ then
SS $_{\text{Total}}$ = $\sum X^2 - \frac{G^2}{N}$

2) **Sum of Squares within Treatments (SS $_{\textbf{Within}}$)**
The sum of square inside each treatment can be calculated as:
SS within = $SS_1 + SS_2 + \ldots + SS_n$
= ∑ SS $_{\text{Inside each treatment}}$

3) **Sum of Squares Between Treatments (SS $_{\textbf{Between}}$)**
The computational formula for SS $_{\text{Between}}$ is as:
SS $_{\text{Between}}$ = $\sum \frac{T^2}{n} - \frac{G^2}{N}$
Now
SS $_{\text{Total}}$ = SS $_{\text{Between}}$ + SS $_{\text{Within}}$

8.4 Multiple Comparison Procedure

In one-way ANOVA "R^{2}" measures the effect size, it suffers one possible limitation – it does not indicate which group may be the responsible for a significant effect. All that a significant R2 and F statistic say is that the means for the groups are unlikely to have been sampled from a single hat of means. Unfortunately, there is no simple, unequivocal statistical solution to the problem of comparing for different levels of an ANOVA factor. A number of statistical methods have been developed to test for the difference in means among the levels of an ANOVA factor. Collectively these are known as multiple comparison procedures (MCPs) or sometimes, as post hoc (i.e. after the fact) tests. These tests should be used regarded as an afterthought than a rigorous examination of pre-specified hypotheses.

Most of the multiple-comparisons methods are meant to pair-wise comparisons of group means, to determine which are significantly from which others. The main purpose of most multiple-comparison procedures is to control the overall significance level, for some set of interferences performed as a follow-up to ANOVA. This overall significance level is the probability, conditional on all the null hypotheses being tested being true, of rejecting at least one of them, or equivalently, of having at least one confidence interval not include the true value.

The various methods differ in how well they properly control the overall significance level and in their relative power. Commonly used method sand their relative power is given below.

- Bonferroni – It is extremely general and simple, but often not powerful.
- Tucky's – It is the best of all possible pair-wise comparisons when sample sizes are unequal or confidence intervals are needed. It is also very good even with equal sample sizes without confidence intervals.
- Stepdown – It is the most powerful for all possible pair-wise comparisons when sample sizes are equal.
- Dunnett's – It is suitable for comparing one sample to each of the others, but not comparing the others to each other.
- Hsu's MCB – It compares each mean to the best of the other means.
- Scheffè's – It is suitable for unplanned contrasts among sets of means.

8.5 Self Assessment Questions

Q. 1 When will you use ANOVA in your research?
Q. 2 Write down the logic behind using ANOVA.
Q. 3 Write a short note on one way ANOVA.
Q. 4 Write down main assumptions underlying one way ANOVA.
Q. 5 What are multiple comparison procedures?
Q. 6 What is the basic purpose of multiple comparison procedures?

8.6 Activities

1. Suppose you have to see the difference between three groups. Discuss with your colleague and select appropriate statistical test.
2. In your study, the treatment you used had no effect. What will be the F-ratio?

8.7 Bibliography

Deitz, T., & Kalof, L. (2009). *Introduction to Social Statistics*. UK: Wiley_-Blackwell

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to Design and Evaluate in Education*. (8th Ed.) McGraw-Hill, New York

Pallant, J. (2005). *SPSS Survival Manual – A Step by Step Guide to Data Analysis Using SPSS for Windows (Version 12)*. Australia: Allen & Unwin.

UNIT-9

INFERENTIAL STATISTICS: CHI-SQUARE(X^2)

Written By:
Prof. Dr. Nasir Mahmood

Reviewed By:
Dr. Rizwan Akram Rana

Introduction

The chi-square (χ^2) statistics is commonly used for testing relationship between categorical variables. It is intended to test how likely it is that an observed difference is due to chance. In most situations it can be used as a quick test of significance. In this unit you will study this important technique in detail.

Objectives

After reading this unit you will be able to

1. Explain chi-square (x^2) Distribution.
2. Describe uses of chi-square (x^2) distribution.
3. Explain types of Chi-square (x^2) distribution.

9.1 The Chi-Square Distribution

The Chi-Square (or the Chi-Squared - χ^2) distribution is a special case of the gamma distribution (the gamma distribution is family of right skewed, continuous probability distribution. These distributions are useful in real life where something has a natural minimum of 0.). a chi-square distribution with n degree of freedom is equal to a gamma distribution with $a = n/2$ and $b = 0.5$ (or $\beta = 2$).

Let us consider a random sample taken from a normal distribution. The chi-square distribution is the distribution of the sum of these random samples squared. The degrees of freedom (say k) are equal to the number of samples being summed. For example, if 10 samples are taken from the normal distribution, then degree of freedom df = 10. Chi-square distributions are always right skewed. The greater the degree of freedom, the more the chi-square distribution looks like a normal distribution.

9.1.1 Uses of Chi-Square (χ^2) Distribution

The chi-square distribution has many uses which include:

i) Confidence interval estimation for a population standard deviation of a normal distribution from a sample standard deviation.

ii) Independence of two criteria of classification of qualitative variables (contingency tables).

iii) Relationship between categorical variables.

iv) Sample variance study when the underlying distribution is normal.

v) Tests of deviations of differences between expected and observed frequencies (one-way table).

vi) The chi-square test (a goodness of fit test).

9.1.2 What is a Chi-Square Statistic?

A Chi-Square Statistic is one way to a relationship between two categorical (non-numerical) variables. The Chi-Square Statistic is a is a single number that tells us how much difference exists between the observed counts and the counts that one expects if there is no relationship in the population.

There are two different types of chi-square tests, both involve categorical data. These are:

a) A chi-square goodness of fit test, and

b) A chi-square test of independence.

In the coming lines these tests will be dealt in some details.

9.2 Chi-Square (χ^2) Goodness-of-Fit Test

The chi-square (χ^2) goodness of fit test (commonly referred to as one-sample chi-square) is the most commonly used goodness of fit test. It explores the proportion of cases that fall into the various categories of a single variable, and compares these with hypothesized values. In some simple words we can say that it is used to find out how the observed value of a given phenomena is significantly different from the expected value. Or we can also say that it is used to test if sample data fits a distribution from a certain population. In other words we can say that chi-square goodness of fit test tells us if the sample data represents the data we expect to find in the actual population. It tells us whether sample data are consistent with a hypothesized distribution. This is a variation of more general chi-square test. The setting for this test is a single categorical variable that can have many levels.

In chi-square goodness of fit test sample data is divided into intervals. Then, the numbers of points that fall into the intervals are compared with the expected numbers of points in each interval. . The null hypothesis for the chi-square goodness of fit test is that the data does not come from the specified distribution. The alternate hypothesis is that the data comes from the specified distribution. The formula for chi-square goodness of fit test is:

$$\chi^2 = \sum \frac{(\text{Observed Values} - \text{Expected Values})^2}{\text{Expected Values}}$$

$$= \sum \frac{(\text{O} - \text{E})^2}{\text{E}}$$

9.2.1 Procedure for Chi-Square (χ^2) Goodness of Fit Test

For using chi-square (χ^2) goodness of fit test we will have to set up null and alternate hypothesis. A null hypothesis assumes that there is no significance difference between observed and expected value. Then, alternate hypothesis will become, there is significant different difference between the observed and the expected value. Now compute the value of chi-square of fit test using formula:

$$\chi^2 = \sum \frac{(\text{Observed Values} - \text{Expected Values})^2}{\text{Expected Values}}$$

Two potential disadvantages of chi-square are:

a) The chi-square test can only be used to put data into classes. If there is data that have not been put into classes then it is necessary to make a frequency table of histogram before performing the test.

b) It requires sufficient sample size in order for chi-square approximation to be valid.

9.2.2 When to Use the Chi-Square Goodness of Fit Test?

The chi-square goodness of fit test is appropriate when the following conditions are met:

- The sampling method is simple random.
- The variable under study is categorical.
- The expected value of the number of sample observation in each level of the variable is at least 5.

For the chi-square goodness of fit test, the hypotheses take the form:

H_0 : The data are not consistent with a specified distribution.
H_a : The data are consistent with a specified distribution.

The null hypothesis (H_0) specifies the proportion of observations at each level of the categorical variable. The alternative hypothesis (H_a) is that a least one of the specified proportion is not true.

9.2.3 Basic Framework of Goodness of Fit Tests

The procedure for carrying out a goodness of fit test is as follows:

i) States the null hypothesis (H_0)

It might take the form:
The data are not consistent with a specified distribution.

ii) States the alternate hypothesis (H_a)

This is an opposite statement to the null hypothesis
The data are consistent with a specified distribution.

iii) Calculate the Test Statistic

The test statistic is calculated using the formula

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Where O and E represent the observed an d expected frequencies respectively.

iv) Find the *p*-value

The range of our p-value can be found by comparing test statistic to table values.

v) Reach a conclusion

We need a p-value less than the significance level, generally less than 5% ($p < .05$), to reject the null hypothesis. It is suitable to write a sentence in the context of the question, i.e. “the data appears to follow a normal distribution”

9.3 Chi-Square Independence Test

A chi-square (χ^2) test of independence is the second important form of chi-square tests. It is used to explore the relationship between two categorical variables. Each of these variables can have two of more categories.

It determines if there is a significant relationship between two nominal (categorical) variables. The frequency of one nominal variable is compared with different values of the second nominal variable. The data can be displayed in R*C contingency table, where R is the row and C is the column. For example, the researcher wants to examine the relationship between gender (male and female) and empathy (high vs. low). The researcher will use chi-square test of independence. If the null hypothesis is accepted there would be no relationship between gender and empathy. If the null hypothesis is rejected then the conclusion will be there is a relationship between gender and empathy (e.g. say females tent to score higher on empathy and males tend to score lower on empathy).

The chi-square test of independence being a non-parametric technique follow less strict assumptions, there are some general assumptions which should be taken care of:

i) Random Sample - Sample should be selected using simple random sampling method.
ii) Variables - Both variables under study should be categorical.
iii) Independent Observations – Each person or case should be counted only once and none should appear in more than one category of group. The data from one subject should not influence the data from another subject.
iv) If the data are displayed in a contingency table, the expected frequency count for each cell of the table is at least 5.

Both the chi-square tests are sometime confused but they are quite different from each other.

- The chi-square test for independence compares two sets of data to see if there is relationship.
- The chi-square goodness of fit test is to fit one categorical variable to a distribution.

9.4 Self-Assessment Questions

Q. 1 What is chi-square (χ^2) distribution?
Q. 2 What are the uses of chi-square (χ^2) distribution?
Q. 3 What is a chi-square (χ^2) statistics?
Q. 4 What do you know about chi-square (χ^2) goodness of fit test?
Q. 5 Write down the procedure for goodness of fit test.
Q. 6 When will you use chi-square (χ^2) goodness of fit test?
Q. 7 Write down the basic framework of goodness of fit test.
Q. 8 What is chi-square (χ^2) independence test?

9.5 Activities

1. Make a list of multiple comparison procedures.
2. Make a list of steps of using ANOVA.

9.6 Bibliography

Agresti, A. & Finlay, B. (1997). *Statistical Methods for Social Sciences*, (3rd *Ed.*). Prentice Hall.

Anderson, T. W., & Sclove, S. L. (1974). *Introductory Statistical Analysis*, Finland: Houghton Mifflin Company.

Bartz, A. E. (1981). *Basic Statistical Concepts (2nd Ed.)*. Minnesota: Burgess Publishing Company

Deitz, T., & Kalof, L. (2009). *Introduction to Social Statistics*. UK: Wiley_-Blackwell

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to Design and Evaluate in Education*. (8th Ed.) McGraw-Hill, New York

Gay, L. R., Mills, G. E., & Airasian, P. W. (2010). *Educational Research: Competencies for Analysis and Application, 10th Edition*. Pearson, New York USA.

Gravetter, F. J., & Wallnau, L. B. (2002). *Essentials of Statistics for the Behavioral Sciences (4th Ed.)*. Wadsworth, California, USA.

Pallant, J. (2005). *SPSS Survival Manual – A step by step guide to data analysis using SPSS for Windows (Version 12)*. Australia: Allen & Unwin.