Statistical Techniques

DPSY527

Edited by
Dr. Mohammad Amin Wani
Dr. Jotika judge

LOVELY PROFESSIONAL UNIVERSITY

Statistical Techniques

Edited By:
Dr. Mohammad Amin Wani
Dr. Jotika Judge

CONTENT

UNIT 01: Introduction to Statistics

Content

Objectives

- Understand about Statistics
- Importance of Statistics
- Basic tenets of Statistics
- Important role of Statistics in psychology

Introduction

Statistics, like many other sciences, is a developing discipline; it is not static. It has gradually developed during the last few centuries. Throughout history it has been defined in different manners. Some definitions from the past seem very strange today, but those definitions had their place in their own time. Defining a subject has always been a difficult task. A good definition from today may be discarded in the future. It is difficult to define statistics, and some of the definitions are discussed here.

The kings and rulers in ancient times were interested in their manpower. They conducted censuses of populations to gather data. They used this information to calculate their strength and ability for war. In those days, statistics was defined as:

This definition places the emphasis on counting only, and that common man considers statistics as nothing but counting. This used to be the situation a very long time ago; statistics today is not mere counting of people, counting of animals, counting of trees and counting of fighting forces. It has now developed into a rich method of data analysis and interpretation. This definition is very simple but it covers only some areas of statistics. Averages are very simple and important in statistics. Experts are interested in average deaths rates, average birth rates,

average increases in population, average increases in per capita income, average increases in standard of living and cost of living, average development rates, average inflation rates, average production of rice per acre, average literacy rates and many other averages from different fields of practical life. But statistics is not limited to averages only; there are many other statistical tools like measures of variation, measures of correlation, measures of independence, etc. Thus, this definition is weak and incomplete and is no longer applicable.

This definition covers a major part of statistics, and is close to modern statistics. But it is not complete because it stresses only probability. There are some areas of statistics in which probability is not used.
This definition is close to modern statistics, but it does not cover the entire scope of modern statistics.

Secrist has given a detailed definition of statistics in the plural sense and can be found in the previous post. He has not given any importance to statistics in the singular sense.

Statistics both in the singular and the plural sense have been combined in the following definition which is accepted as the modern definition of statistics:

1.1.Basic understanding about variables

Variable is any characteristics, number, or quantity that can be measured or counted. A variable may also be called a data item. Age, sex, business income and expenses, country of birth, capital expenditure, class grades, eyecolour and vehicle type are examples of variables. It is called a variable because the value may vary between data units in a population, and may change in value overtime.

For example; 'income' is a variable that can vary between data units in a population (i.e. the people or businesses being studied may not have the same incomes) and can also vary over time for each data unit (i.e. income can go up or down).There are different ways variables can be described according to the ways they can be studied, measured, and presented.

Numeric variables have values that describe a measurable quantity as a number, like 'how many' or 'how much'. Therefore numeric variables are quantitative variables.

Numeric variables may be further described as either continuous or discrete:

- A continuous variable is a numeric variable. Observations can take any value between a certain set of real numbers. The value given to an observation for a continuous variable can include values as small as the instrument of measurement allows. Examples of continuous variables include height, time, age, and temperature.

- A discrete variable is a numeric variable. Observations can take a value based on a count from a set of distinct whole values. A discrete variable cannot take the value of a fraction between one value and the next closest value. Examples of discrete variables include the number of registered cars, number of business locations, and number of children in a family, all of of which measured as whole units (i.e. 1, 2, 3 cars).

The data collected for a numeric variable are quantitative data. Categorical variables have values that describe a 'quality' or 'characteristic' of a data unit, like 'what type' or 'which category'. Categorical variables fall into mutually exclusive (in one category or in another) and exhaustive (include all possible options) categories. Therefore, categorical variables are qualitative variables and tend to be represented by a non-numeric value.

Categorical variables may be further described as ordinal or nominal:

- An ordinal variable is a categorical variable. Observations can take a value that can be logically ordered or ranked. The categories associated with ordinal variables can be ranked higher or lower than another, but do not necessarily establish a numeric difference between each category. Examples of ordinal categorical variables include academic grades (i.e. A, B, C), clothing size (i.e. small, medium, large, extra large) and attitudes (i.e. strongly agree, agree, disagree, strongly disagree).
- A nominal variable is a categorical variable. Observations can take a value that is not able to be organised in a logical sequence. Examples of nominal categorical variables include sex, business type, eye colour, religion and brand.

Common Types of Variables

- **Categorical variable**: variables than can be put into categories. For example, the category "Toothpaste Brands" might contain the variables *Colgate* and *Aquafresh*.
- **Confounding variable**: extra variables that have a hidden effect on your experimental results.
- **Continuous variable**: a variable with infinite number of values, like "time" or "weight".
- **Control variable**: a factor in an experiment which must be held constant. For example, in an experiment to determine whether light makes plants grow faster, you would have to control for soil quality and water.
- **Dependent variable**: the outcome of an experiment. As you change the independent variable, you watch what happens to the dependent variable.
- **Discrete variable**: a variable that can only take on a certain number of values. For example, "number of cars in a parking lot" is discrete because a car park can only hold so many cars.
- **Independent variable**: a variable that is not affected by anything that you, the researcher, does. Usually plotted on the x-axis.
- **Lurking variable**: a "hidden" variable the affects the relationship between the independent and dependent variables.

- A **measurement variable** has a number associated with it. It's an "amount" of something, or a"number" of something.
- **Nominal variable**: another name for categorical variable.
- **Ordinal variable**: similar to a categorical variable, but there is a clear order. For example, income levels of low, middle, and high could be considered ordinal.
- **Qualitative variable**: a broad category for any variable that can't be counted (i.e. has no numerical value). Nominal and ordinal variables fall under this umbrella term.
- **Quantitative variable:** A broad category that includes any variable that can be counted, or has a numerical value associated with it. Examples of variables that fall into this category include discrete variables and ratio variables.
- **Random variables** are associated with random processes and give numbers to outcomes of random events.
- A **ranked variable** is an ordinal variable; a variable where every data point can be put in order (1st, 2nd, 3rd, etc.).
- **Ratio variables**: similar to interval variables, but has a meaningful zero.

Less Common Types of Variables

- **Active Variable:** a variable that is manipulated by the researcher.
- **Antecedent Variable**: a variable that comes before the independent variable.
- **Attribute variable**: another name for a categorical variable (in statistical software) or a variable that isn't manipulated (in design of experiments).
- **Binary variable**: a variable that can only take on two values, usually 0/1. Could also be yes/no, tall/short or some other two-variable combination.
- **Collider Variable**: a variable represented by a node on a causal graph that has paths pointing in as well as out.
- **Covariate variable**: similar to an independent variable, it has an effect on the dependent variable but is usually not the variable of interest. See also: **Noncomitant variable**.
- **Criterion variable**: another name for a dependent variable, when the variable is used in non-experimental situations.
- **Dichotomous variable**: Another name for a binary variable.
- **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."
- **Endogenous variable**: similar to dependent variables, they are affected by other variables in the system. Used almost exclusively in econometrics.
- **Exogenous variable**: variables that affect others in the system.
- **Explanatory Variable**: a type of independent variable. When a variable is independent, it is not affected at all by any other variables. When a variable isn't independent for certain, it's an explanatory variable.
- **Extraneous variables** are any variables that you are not intentionally studying in your experiment or test.

- A **grouping variable** (also called a coding variable, group variable or by variable) sorts data within data files into categories or groups.
- **Identifier Variables:** variables used to uniquely identify situations.
- **Indicator variable**: another name for a dummy variable.
- **Interval variable**: a meaningful measurement between two variables. Also sometimes used as another name for a continuous variable.
- **Intervening variable**: a variable that is used to explain the relationship between variables.
- **Latent Variable:** a hidden variable that can't be measured or observed directly.
- **Manifest variable**: a variable that can be directly observed or measured.
- **Manipulated variable**: another name for independent variable.
- **Mediating variable** or intervening variable: variables that explain how the relationship between variables happens. For example, it could explain the difference between the predictor and criterion.
- **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.
- **Nuisance Variable**: an extraneous variable that increases variability overall.
- **Observed Variable**: a measured variable (usually used in SEM).
- **Outcome variable**: similar in meaning to a dependent variable, but used in a non-experimental study.
- **Polychotomous variables**: variables that can have more than two values.
- **Predictor variable**: similar in meaning to the independent variable, but used in regression and in non-experimental studies.
- **Responding variable**: an informal term for dependent variable, usually used in science fairs.
- **Scale Variable**: basically, another name for a measurement variable.
- **Study Variable (Research Variable)**: can mean any variable used in a study, but does have a more formal definition when used in a clinical trial.
- **Test Variable**: another name for the Dependent Variable.
- **Treatment variable**: another name for independent variable.

A lot of psychology students are surprised (and sometimes dismayed) to realize that statistics courses are required for graduation in their chosen major. Yes, statistics courses are a major part of virtually all psychology programs. You will also encounter the subject in many of your other classes, particularly those that involve experimental design or research methods.

To succeed in psychology, you not only need to be able to pass a statistics class. You need to be able to understand statistics, too.

1.2.The Importance of Statistics in Psychology

Statistics allow us to make sense of and interpret a great deal of information. Consider the sheer volume of data you encounter in a given day. How many hours did you sleep? How many students in your class ate breakfast this morning? How many people live within a one-mile radius of your home? By using statistics, we can organize and interpret all of this information in a meaningful way.

In psychology, we are also confronted with enormous amounts of data. How do changes in one variable impact other variables? Is there a way we can measure that relationship? What is the overall strength of that relationship and what does that mean? Statistics allow us to answer these kinds of questions. Statistics allow psychologists to:

- Organize data: When dealing with an enormous amount of information, it is all too easy to become overwhelmed. Statistics allow psychologists to present data in ways that are easier to comprehend. Visual displays such as graphs, pie charts, frequency distributions, and scatterplots allow researchers to get a better overview of data and look for patterns they might otherwise miss.
- Describe data: Think about what happens when researchers collect a great deal of information about a group of people (for example, the U.S. Census). Descriptive statistics provide a way to summarize facts, such as how many men and women there are, how many children there are, or how many people are currently employed.
- Make inferences based on data: By using what's known as inferential statistics, researchers can infer things about a given sample or population. Psychologists use the data they have collected to test a hypothesis. Using statistical analysis, researchers can determine the likelihood that a hypothesis should be either accepted or rejected.[1]

Having a solid understanding of statistical methods can help you excel in almost all other classes. Whether you are taking social psychology or human sexuality, you will be spending a great deal of time learning about research. Your foundation of statistical knowledge will allow you to make better sense of the research you'll find described in your other psychology courses.

Secondly, think about all the claims about psychology that you encounter on a daily basis outside of class. Magazines publish stories about the latest scientific findings, self-help books make proclamations about different ways to approach problems, and news reports interpret (or misinterpret) psychology research.

By understanding the research process, including the kinds of statistical analyses that are used, you will be able to become a wise consumer of psychology information and make better judgments of the information you come across. By understanding statistics, you can make better decisions about your health and well-being.

Many prospective psychology students assume that their chosen major will require very little math. After all, psychology is the science of the mind and behavior, so what does math have to do with it?

Quite a bit, actually. Math classes, and statistics in particular, are an important part of any psychology program. You will need to take math classes that fulfill your school's general education requirements as well as additional statistics requirements to fulfill your psychology program's core requirements.

In most cases, you will have to take at least two math classes, but in other cases, it might end up being between three and five. Check your school's graduation requirements as well as your psychology program's core requirements for more information.

Knowing why statistics are important might not help with that sense of dread you feel before stepping into your first stats course. But even if you don't consider yourself "good at math," you can still succeed in your stats classes. You might have to put in some extra effort, but help is available.

Start with your instructor. They might be able to recommend books, online tools, and on-campus resources. Many colleges and universities offer a math lab where students can go to receive extra help and tutoring with any type of math course, including statistics. Consider joining or forming a study group with classmates, too.

1.3. Summary

Statistics is the science concerned with developing and studying methods for collecting, analyzing, interpreting and presenting empirical data. Statistics is a highly interdisciplinary field; research in statistics finds applicability in virtually all scientific fields and research questions in the various scientific fields motivate the development of new statistical methods and theory. In developing methods and studying the theory that underlies the methods statisticians draw on a variety of mathematical and computational tools.Two fundamental ideas in the field of statistics are uncertainty and variation. There are many situations that we encounter in science (or more generally in life) in which the outcome is uncertain. In some cases the uncertainty is because the outcome in question is not determined yet (e.g., we may not know whether it will rain tomorrow) while in other cases the uncertainty is because although the outcome has been determined already we are not aware of it (e.g., we may not know whether we passed a particular exam). Probability is a mathematical language used to discuss uncertain events and probability plays a key role in statistics. Any measurement or data collection effort is subject to a number of sources of variation. By this we mean that if the same measurement were repeated, then the answer would likely change. Statisticians attempt to understand and control (where possible) the sources of variation in any situation.

1.4.Keywords

Variables, Moderating Variable, Nominal Variable, Statistics

1.5. Self-Assessment/Evaluation

1- Psychology is scientific in nature – true/ false

2- Psychology does involve experimentation. True/ false

3- Statistics is important true/ false

4- Psychological statistics is true science true/ false

5- Statistics is a type of school of thought in Psychology true/ false

6- Statistics is incomplete without Psychology true/ false

7- Statistics is philosophy. True/ false

8- Psychology differs from other arts disciplines true/ false

9- Wilhelm Wundt is not called father of statistics true/ false

10- Sigmund Freud was a mathematician true/ false

11- Statistics is all about numbers true/ false

12- Psychology is now known as study of soul true/ false

13- Statistics is a tool of mathematics true/ false

14- Psychology is incomplete without statistics true/ false

15- Psychology and statistics are two separate fields true/ false

1.6Review Questions

- Psychology needs statistics. Discuss
- What is statistics of psychology?
- Describe types of variables.
- Discuss some major definition of statistics.

Further/Suggested Readings

- Cohen, J. (1977), *Statistical Power Analysis for the Behavioural Sciences*. Academic Press: New York.
- Downie, N.M. and Heath, R.W. (1970), *Basic Statistical Methods*. Harper and Row Publishers: New York.
- https://www.stat.uci.edu/what-is-statistics/
- https://www.abs.gov.au/websitedbs/D3310114.nsf/home/statistical+language++what+are+variables#:~:text=A%20variable%20is%20any%20characteristics,type%20are%20examples%20of%20variables.

UNIT 02: Scales of Measurement

Content

Objectives

- Understand about measurement
- Importance of measurement Statistics
- Basic tenets of measurement
- Important role of measurement Statistics in psychology

Introduction

Before you can use statistics to analyze a problem, you must convert information about the problem into data. That is, you must establish or adopt a system of assigning values, most often numbers, to the objects or concepts that are central to the problem in question. This is not an esoteric process but something people do every day. For instance, when you buy something at the store, the price you pay is a measurement: it assigns a number signifying the amount of money that you must pay to buy the item. Similarly, when you step on the bathroom scale in the morning, the number you see is a measurement of your body weight. Depending on where you live, this number may be expressed in either pounds or kilograms, but the principle of assigning a number to a physical quantity (weight) holds true in either case.

Data need not be inherently numeric to be useful in an analysis. For instance, the categories *male* and *female* are commonly used in both science and everyday life to classify people, and there is nothing inherently numeric about these two categories. Similarly, we often speak of the colors of objects in broad classes such as *red* and *blue*, and there is nothing inherently numeric about these categories either. (Although you could make an argument about different wavelengths of light, it's not necessary to have this knowledge to classify objects by color.)

This kind of thinking in categories is a completely ordinary, everyday experience, and we are seldom bothered by the fact that different categories may be applied in different situations. For instance, an artist might differentiate among colors such as *carmine*, *crimson*, and *garnet*, whereas a layperson would be satisfied to refer to all of them as *red*. Similarly, a social scientist might be interested in collecting information about a person's marital status in terms such as *single – never married*, *single – divorced*, and *single – widowed*, whereas to someone else, a person in any of those three categories could simply be considered *single*. The point is that the level of detail used in a system of classification should be appropriate, based on the reasons for making the classification and the uses to which the information will be put.

Measurement is the process of systematically assigning numbers to objects and their properties to facilitate the use of mathematics in studying and describing objects and their relationships. Some types of measurement are fairly concrete: for instance, measuring a person's weight in pounds or kilograms or his height in feet and inches or in meters. Note that the particular system of measurement used is not as important as the fact that we apply a consistent set of rules: we can easily convert a weight expressed in kilograms to the equivalent weight in pounds, for instance. Although any system of units may seem arbitrary (try defending feet and inches to someone who grew up with the metric system!), as long as the system has a consistent relationship with the property being measured, we can use the results in calculations.

Measurement is not limited to physical qualities such as height and weight. Tests to measure abstract constructs such as intelligence or scholastic aptitude are commonly used in education and psychology, and the field of psychometrics is largely concerned with the development and refinement of methods to study these types of constructs. Establishing that a particular measurement is accurate and meaningful is more difficult when it can't be observed directly. Although you can test the accuracy of one scale by comparing results with those obtained from another scale known to be accurate, and you can see the obvious use of knowing the weight of an object, the situation is more complex if you are interested in measuring a construct such as intelligence. In this case, not only are there no universally accepted measures of intelligence against which you can compare a new measure, there is not even common agreement about what "intelligence" means. To put it another way, it's difficult to say with confidence what someone's actual intelligence is because there is no certain way to measure it, and in fact, there might not even be common agreement on what it is. These issues are particularly relevant to the social sciences and education, where a great deal of research focuses on just such abstract concepts.

2.2Levels of Measurement

Statisticians commonly distinguish four types or levels of measurement, and the same terms can refer to data measured at each level. The levels of measurement differ both in terms of the meaning of the numbers used in the measurement system and in the types of statistical procedures that can be applied appropriately to data measured at each level.

2.3 Nominal Data

With *nominal* data, as the name implies, the numbers function as a *name* or label and do not have numeric meaning. For instance, you might create a variable for gender, which takes the value 1 if the person is male and 0 if the person is female. The 0 and 1 have no numeric meaning but function simply as labels in the same way that you might record the values as M or F. However, researchers often prefer numeric coding systems for several reasons. First, it can simplify analyzing the data because some statistical packages will not accept nonnumeric values for use in certain procedures. (Hence, any data coded nonnumerically would have to be recoded before analysis.) Second, coding with numbers bypasses some issues in data entry, such as the conflict between upper- and lowercase letters (to a computer, *M* is a different value than *m*, but a person doing data entry might treat the two characters as equivalent).

Nominal data is not limited to two categories. For instance, if you were studying the relationship between years of experience and salary in baseball players, you might classify the players according to their primary position by using the traditional system whereby 1 is assigned to the pitchers, 2 to the catchers, 3 to first basemen, and so on.

If you can't decide whether your data is nominal or some other level of measurement, ask yourself this question: do the numbers assigned to this data represent some quality such that a higher value indicates that the object has more of that quality than a lower value? Consider the  of coding gender so 0 signifies a female and 1 signifies a male. Is there some quality of gender-ness of which men have more than women? Clearly not, and the coding scheme would work as well if women were coded as 1 and men as 0. The same principle applies in the baseball  : there is no quality of baseball-ness of which outfielders have more than pitchers. The numbers are merely a convenient way to label subjects in the study, and the most important point is that every position is assigned a distinct value. Another name for nominal data is *categorical* data, referring to the fact that the measurements place objects into categories (male or female, catcher or first baseman) rather than measuring some intrinsic quality in them. Chapter 5 discusses methods of analysis appropriate for this type of data, and some of the techniques covered in on nonparametric statistics are also appropriate for categorical data.

When data can take on only two values, as in the male/female example,  it can also be called *binary* data. This type of data is so common that special techniques have been developed to study it, including logistic regression which has applications in many fields. Many medical statistics, such as the odds ratio and the risk ratio were developed to describe the relationship between two binary variables because binary variables occur so frequently in medical research.

2.4 Ordinal Data

Ordinal data refers to data that has some meaningful *order*, so that higher values represent more of some characteristic than lower values. For instance, in medical practice, burns are commonly described by their degree, which describes the amount of tissue damage caused by the burn. A first-degree burn is characterized by redness of the skin, minor pain, and damage to the epidermis (outer layer of skin) only. A second-degree burn includes blistering and involves the superficial layer of the dermis (the layer of skin between the epidermis and the subcutaneous tissues), and a third-degree burn extends through the dermis and is characterized by charring of the skin and possibly destruction of nerve endings. These categories may be ranked in a logical order: first-degree burns are the least serious in terms of tissue damage, second-degree burns more serious, and third-degree burns the most serious. However, there is no metric analogous to a ruler or scale to quantify how great the distance between categories is, nor is it possible to determine whether the difference between first- and second-degree burns is the same as the difference between second- and third-degree burns.

Many ordinal scales involve ranks. For instance, candidates applying for a job may be ranked by the personnel department in order of desirability as a new hire. This ranking tells you who is the preferred candidate, the second most preferred, and so on, but does not tell you whether the first and second candidates are in fact very similar to each other or the first-ranked candidate is much more preferable than the second. You could also rank countries of the world in order of their population, creating a meaningful order without saying anything about whether, say, the difference between the 30th and 31st countries was similar to that between the 31st and 32nd countries. The numbers used for measurement with ordinal data carry more meaning than those used in nominal data, and many statistical techniques have been developed to make full use of the information carried in the ordering while not assuming any further properties of the scales. For instance, it is appropriate to calculate the median (central value) of ordinal data but not the mean because it assumes equal intervals and requires division, which requires ratio-level data.

3.5 Interval Data

Interval data has a meaningful order and has the quality of *equal intervals* between measurements, representing equal changes in the quantity of whatever is being measured. The most common example of the interval level of measurement is the Fahrenheit temperature scale. If you describe temperature using the Fahrenheit scale, the difference between 10 degrees and 25 degrees (a

difference of 15 degrees) represents the same amount of temperature change as the difference between 60 and 75 degrees. Addition and subtraction are appropriate with interval scales because a difference of 10 degrees represents the same amount of change in temperature over the entire scale. However, the Fahrenheit scale has no natural zero point because 0 on the Fahrenheit scale does not represent an absence of temperature but simply a location relative to other temperatures. Multiplication and division are not appropriate with interval data: there is no mathematical sense in the statement that 80 degrees is twice as hot as 40 degrees, for instance (although it is valid to say that 80 degrees is 40 degrees hotter than 40 degrees). Interval scales are a rarity, and it's difficult to think of a common example other than the Fahrenheit scale. For this reason, the term "interval data" is sometimes used to describe both interval and ratio data (discussed in the next section).

2.6 Ratio Data

Ratio data has all the qualities of interval data (meaningful order, equal intervals) and a natural zero point. Many physical measurements are ratio data: for instance, height, weight, and age all qualify. So does income: you can certainly earn 0 dollars in a year or have 0 dollars in your bank account, and this signifies an absence of money. With ratio-level data, it is appropriate to multiply and divide as well as add and subtract; it makes sense to say that someone with $100 has twice as much money as someone with $50 or that a person who is 30 years old is 3 times as old as someone who is 10.

It should be noted that although many physical measurements are ratio-level, most psychological measurements are ordinal. This is particularly true of measures of value or preference, which are often measured by a Likert scale. For instance, a person might be presented with a statement (e.g., "The federal government should increase aid to education") and asked to choose from an ordered set of responses (e.g., strongly agree, agree, no opinion, disagree, strongly disagree). These choices are sometimes assigned numbers (e.g., 1—strongly agree, 2—agree, etc.), and this sometimes gives people the impression that it is appropriate to apply interval or ratio techniques (e.g., computation of means, which involves division and is therefore a ratio technique) to such data. Is this correct? Not from the point of view of a statistician, but sometimes you do have to go with what the boss wants rather than what you believe to be true in absolute terms.

2.7 Continuous and Discrete Data

Another important distinction is that between *continuous* and *discrete* data. Continuous data can take any value or any value within a range. Most data measured by interval and ratio scales, other than that based on counting, is continuous: for instance, weight, height, distance, and income are all continuous.

In the course of data analysis and model building, researchers sometimes recode continuous data in categories or larger units. For instance, weight may be recorded in pounds but analyzed in 10-pound increments, or age recorded in years but analyzed in terms of the categories of *0–17*, *18–65*,

and *over 65*. From a statistical point of view, there is no absolute point at which data becomes continuous or discrete for the purposes of using particular analytic techniques (and it's worth remembering that if you record age in years, you are still imposing discrete categories on a continuous variable). Various rules of thumb have been proposed. For instance, some researchers say that when a variable has 10 or more categories (or, alternatively, 16 or more categories), it can safely be analyzed as continuous. This is a decision to be made based on the context, informed by the usual standards and practices of your particular discipline and the type of analysis proposed.

Discrete variables can take on only particular values, and there are clear boundaries between those values. As the old joke goes, you can have 2 children or 3 children but not 2.37 children, so "number of children" is a discrete variable. In fact, any variable based on counting is discrete, whether you are counting the number of books purchased in a year or the number of prenatal care visits made during a pregnancy. Data measured on the nominal scale is always discrete, as is binary and rank-ordered data.

2.8 Operationalization

People just starting out in a field of study often think that the difficulties of research rest primarily in statistical analysis, so they focus their efforts on learning mathematical formulas and computer programming techniques to carry out statistical calculations. However, one major problem in research has very little to do with either mathematics or statistics and everything to do with knowing your field of study and thinking carefully through practical problems of measurement. This is the problem of *operationalization*, which means the process of specifying how a concept will be defined and measured.

Operationalization is always necessary when a quality of interest cannot be measured directly. An obvious example is intelligence. There is no way to measure intelligence directly, so in the place of such a direct measurement, we accept something that we can measure, such as the score on an IQ test. Similarly, there is no direct way to measure "disaster preparedness" for a city, but we can operationalize the concept by creating a checklist of tasks that should be performed and giving each city a disaster-preparedness score based on the number of tasks completed and the quality or thoroughness of completion. For a third example, suppose you wish to measure the amount of physical activity performed by individual subjects in a study. If you do not have the capacity to monitor their exercise behavior directly, you can operationalize "amount of physical activity" as the amount indicated on a self-reported questionnaire or recorded in a diary.

Because many of the qualities studied in the social sciences are abstract, operationalization is a common topic of discussion in those fields. However, it is applicable to many other fields as well. For instance, the ultimate goals of the medical profession include reducing mortality (death) and reducing the burden of disease and suffering. Mortality is easily verified and quantified but is frequently too blunt an instrument to be useful since it is a thankfully rare outcome for most diseases. "Burden of disease" and "suffering," on the other hand, are concepts that could be used to

define appropriate outcomes for many studies but that have no direct means of measurement and must therefore be operationalized. Examples of operationalization of burden of disease include measurement of viral levels in the bloodstream for patients with AIDS and measurement of tumor size for people with cancer. Decreased levels of suffering or improved quality of life may be operationalized as a higher self-reported health state, a higher score on a survey instrument designed to measure quality of life, an improved mood state as measured through a personal interview, or reduction in the amount of morphine requested for pain relief.

Some argue that measurement of even physical quantities such as length require operationalization because there are different ways to measure even concrete properties such as length. (A ruler might be the appropriate instrument in some circumstances, a micrometer in others.) Even if you concede this point, it seems clear that the problem of operationalization is much greater in the human sciences, when the objects or qualities of interest often cannot be measured directly.

2.9 Proxy Measurement

The term *proxy measurement* refers to the process of substituting one measurement for another. Although deciding on proxy measurements can be considered as a subclass of operationalization, this book will consider it as a separate topic. The most common use of proxy measurement is that of substituting a measurement that is inexpensive and easily obtainable for a different measurement that would be more difficult or costly, if not impossible, to collect. Another example is collecting information about one person by asking another, for instance, by asking a parent to rate her child's mood state.

For a simple example of proxy measurement, consider some of the methods police officers use to evaluate the sobriety of individuals while in the field. Lacking a portable medical lab, an officer can't measure a driver's blood alcohol content directly to determine whether the driver is legally drunk. Instead, the officer might rely on observable signs associated with drunkenness, simple field tests that are believed to correlate well with blood alcohol content, a breath alcohol test, or all of these. Observational signs of alcohol intoxication include breath smelling of alcohol, slurred speech, and flushed skin. Field tests used to evaluate alcohol intoxication quickly generally require the subjects to perform tasks such as standing on one leg or tracking a moving object with their eyes. A Breathalyzer test measures the amount of alcohol in the breath. None of these evaluation methods provides a direct test of the amount of alcohol in the blood, but they are accepted as reasonable approximations that are quick and easy to administer in the field.

To look at another common use of proxy measurement, consider the various methods used in the United States to evaluate the quality of health care provided by hospitals and physicians. It is difficult to think of a direct way to measure quality of care, short of perhaps directly observing the care provided and evaluating it in relation to accepted standards (although you could also argue that the measurement involved in such an evaluation process would still be an operationalization of the abstract concept of "quality of care"). Implementing such an evaluation method would be

prohibitively expensive, would rely on training a large crew of evaluators and relying on their consistency, and would be an invasion of patients' right to privacy. A solution commonly adopted instead is to measure processes that are assumed to reflect higher quality of care: for instance, whether anti-tobacco counseling was appropriately provided in an office visit or whether appropriate medications were administered promptly after a patient was admitted to the hospital.

Proxy measurements are most useful if, in addition to being relatively easy to obtain, they are good indicators of the true focus of interest. For instance, if correct execution of prescribed processes of medical care for a particular treatment is closely related to good patient outcomes for that condition, and if poor or nonexistent execution of those processes is closely related to poor patient outcomes, then execution of these processes may be a useful proxy for quality. If that close relationship does not exist, then the usefulness of the proxy measurements is less certain. No mathematical test will tell you whether one measure is a good proxy for another, although computing statistics such as correlations or chi-squares between the measures might help evaluate this issue. In addition, proxy measurements can pose their own difficulties. To take the example of evaluating medical care in terms of procedures performed, this method assumes that it is possible to determine, without knowledge of individual cases, what constitutes appropriate treatment and that records are available that contain the information needed to determine what procedures were performed. Like many measurement issues, choosing good proxy measurements is a matter of judgment informed by knowledge of the subject area, usual practices in the field in question, and common sense.

2.10. Summary

Measurement is defined as the system or act of measuring. It can be understood as a process of defining physical items using numbers. For example, "this rod is bigger than that rod". This statement is serving a very limited purpose of comparison where we do not know anything about the individual attributes of the given rods. But if we say that the length of the first rod is 20 inches, and the length of the second rod is 15 inches, therefore 1st rod is bigger than the second one by 5 inches. This statement is making more sense mathematically and offers us a reason for our deduction.

In mathematics, measurement is often considered as a separate branch as it includes a wide range of knowledge including conversion, units, measuring length, mass, time, etc. It is associated with other branches as well like geometry, trigonometry, algebra, etc. We use the concept of measurement with the shapes (area, volume, etc), measuring heights and distances using trigonometric ratios is also a type of measurement (trigonometry), and measurement can also be done using unknown quantities or variables to establish a general relationship (algebra). Now, before learning about measurement units, let us learn the abbreviations that are generally used to represent units of measurement.

2.11. Keywords/Glossary

measurement, statistics, application,definition

2.12 Self-Assessment/Evaluation

1- Measurement in statistics is necessary – true/ false

2- Psychology does involve measurement. True/ false

3- Psychology deals with the scientific study of human nature true/ false

4- Psychology involve laboratory experiments too true/ false

5- Psychology uses statistics- true/ false

6- Human behaviour is subjective -true/ false

7- Measurement scales are multiple- true/ false

8- Measurement scale is only one true/ false

9- Measurement scale measures hear rate- True/ false

10- Statistics uses calculations-true/ false

11- Sigmund Freud is labeled as father of statistics-true/ false

12- Psychology is a branch of statistics-true/ false

13- Behaviour can be measured with the help of statistics True/ false

14- Human behaviour is considered to be very dynamic True/ false

15- Study area Psychology and Philosophy were separated earlier True/ false

2.13 Review Questions

- Psychology needs measurement. Discuss
- What is the importance of measurement in psychology?
- Describe types of measurement scales.

Further/Suggested Readings

Fallix, F. and Brown, B. Bruce (1983), *Statistics for Behavioural Sciences*. The Dorsey Press: Illinois.

6. Ferguson, G.A. (1980), *Statistical Analysis in Psychology and Education*. McGraw Hill Book Co.: New York.

https://www.sciencelearn.org.nz/resources/1851-measurement-introduction

https://www.cuemath.com/measurement/

UNIT 03: Representation of Data

Content

Objectives/Expected Learning Outcomes

Understand aboutRepresentation of data

Importance of Representation of data

Application of Representation of data

3.1. Frequency Tabulations

The frequency tabulation is a very popular method for summarizing data because even very large data sets can be condensed to a manageable form without substantial loss of information. Consider the data on the lengths of shoots of *Banksiaericifolia* shown in part below.

36.4 31.8 31.0 39.4 28.8 31.8 28.7 37.0 25.5 19.3 44.0 38.0 28.6
29.1 21.1 30.4 31.2 38.0 39.0 19.3 27.6 19.1 32.5 26.8 39.9 36.1
33.2 26.5 38.1 14.9 33.2 27.8 24.7 24.9 25.0 33.1 24.1 19.7 19.1
26.9 22.5 25.5 33.0 19.4 26.8 24.6 37.5 19.8 43.7 38.1 30.8 34.5

.
.
.
.
.
.

16.8 43.9 27.9 44.4 29.7 23.0 26.8 43.4 29.4 26.7 16.5 22.1 23.0

27.8 33.1 34.9 20.5 25.4 10.0 28.2 31.0 10.6 28.4 16.5 22.3 17.6
21.9 27.0 26.5 29.2 24.9 18.4 24.1 28.3 29.0 29.1 18.8 36.7 24.7
26.2 32.6 22.3 31.7 37.1 35.6 19.5 26.9 24.8 19.2 25.1 22.1 37.9
29.9 42.1 36.6 25.5 34.2 22.4 40.5 21.2 32.3 31.5 34.2 34.5 39.2

There are 500 measurements, quite a formidable data set. By inspection, the minimum shoot length can be determined as 10.0 cm and the maximum as 44.9 cm. These values define the sample range. We now subdivide the range into intervals or classes, each of equal size. It is generally advisable to round the minimum value down and the maximum value up to appropriate values when deciding on class intervals. In this case it seems sensible to divide the range 10 to 45 cm into seven intervals each 5 cm wide.

If we count the number of shoots that lie in each of the seven intervals, we have the basis for a frequency tabulation. Such a tabulation is shown below.

LENGTH	Frequency	Percent	Cumul. Frequency	Cumul. Percent
10<x<15	6	1.2	6	1.2
15<x<20	35	7.0	41	8.2
20<x<25	93	18.6	134	26.8
25<x<30	155	31.0	289	57.8
30<x<35	130	26.0	419	83.8
35<x<40	57	11.4	476	95.2
40<x<45	24	4.8	500	100.0

The frequency column was obtained by counting the number of measurements that lie within each class. The percent frequency column was obtained by representing each count as a percentage of the total count. The cumulative frequency and cumulative percentage frequencies were obtained by progressively summing the corresponding frequencies.

Frequency tabulations provide summaries of data sets without substantial loss of information. In this case, there has been minimal information lost -- for example, the average shoot length calculated directly from the frequency tabulation (using the class midpoints) of 28.85 is in close agreement with the figure of 28.97 calculated from the raw data. A reader of a paper containing a frequency tabulation would have access to almost as much information as if the entire data set had been published, yet the table takes up far less space and would take up little more room if based on 5 million measurements rather than only 500.

PROC FREQ will produce frequency tabulations for discrete data, and also for continuous data provided there have been some preliminary manipulations to get it into discrete form. PROC TABULATE is used if more sophisticated tabular reports are required.

3.2. Line Diagram?

The definition of a single-line diagram or SLD is an electrical diagram or drawing that represents the components of an electrical installation system represented by symbols, and describes how the components are related. Sometimes aa single line drawing or diagram of an electrical installation is also called a one-line diagram. In this article, we will briefly discuss what an electrical SLD is, types of electrical diagrams, the importance and benefits of a single line diagram. It will also discuss the importance and needed to regularly update or update electrical installation drawing documents for the purposes of reliability, operation and electrical safety.

– – – – – – – – – – – – – – –

Types of Electrical Diagrams

In the field, personnel often refer to single-line diagrams as "electrical drawings". Even though there are several types of diagrams or drawings in the electrical system. Each type of electrical diagram has a unique function. The types of electrical diagrams include:

- Ladder Diagram
- Wiring Diagram
- One-line Diagram

Ladder Diagram

Usually drawn like a ladder so it is called a ladder diagram. Ladder diagram is a diagram that shows the function of an electric circuit using electrical symbols. The ladder diagram does not show the actual location of the components. The ladder diagram allows one to understand and solve problems in a circuit quickly. Ladder diagrams can also be referred to as line diagrams, elementary diagrams, or electrical schematic diagrams.

Wiring Diagram

Wiring diagrams use electrical symbols like ladder diagrams but they try to show the actual location of the components. Wiring diagrams are also referred to as connection diagrams. The wiring diagram helps you to identify cables and components such as those found on equipment.

One-line Diagram

A one line diagram or single-line diagram is a simplified way of representing a three-phase power system. Single line diagrams do not show exact electrical-circuit connections. As the name suggests, a one-line chart uses a single line to represent all three phases. This is the most basic type of electrical installation blueprint. A single line diagram shows the rating and capacity of electrical equipment and circuit conductors and protection devices.

– – – – – – – – – – – – – – – –

Scopes of Single-Line Diagrams

Single line diagram information typically includes:

- Incoming line (nominal voltage and amount – capacity and value)
- Main circuit breaker, main fuse, cut-outs (CTO), switch, and bus-tie
- Power transformer (rating, twist connection and earthing method)
- Feeder circuit breaker
- Fused switches relays (function, use, and type)
- Current/potential transformers (size, type and ratio)
- Transformer for control system
- All mains and load cables
- All substations, including integral relays and main panels and load properties at each feeder and at each substation
- Critical equipment voltages and sizes (UPS, batteries, generators, power distribution, transfer switches, computer room air conditioners)

Benefit of One-Line Diagrams

- Helps identify when to perform troubleshooting and simplifies the troubleshooting process

- Accurate single line diagram will further ensure the safety of personnel work
- Meets compliance with applicable regulations and standards
- Ensure a safer and more reliable operation of the facility

An electrical single-line diagram is a blueprint of the electrical system. Creating a one-line diagram is the first step in preparing a critical response plan, enabling electrical personnel to fully understand the layout and design of the facility's electrical distribution system.

Whether it's a new or an existing facility, the single line diagram is the road map for all future testing, service and maintenance activities. An effective single line diagram will clearly show how the main components of an electrical system are connected. It shows the correct power distribution path from the incoming power source to each downstream load – including the rating and size of each electrical appliance, its circuit conductors, and protection devices.

Often decision makers feel that they do not need to update electrical installation diagrams, or do not even consider them important. Many industrial and commercial facilities operate without accurate single line diagrams. These conditions may be considered important until they encounter real problems or losses due to not updating or inaccurate electrical installation diagrams.

Then what is the importance of updating a single line diagram so that we need to update it? From the electrical engineering and safety aspect, electrical SLD is the main resource for calculating short-circuit currents, determining selective protection coordination and ultimately calculating incident energy – making it one of the most important safety documents available at the facility. Safe operation of facilities is primary, while SLDs often do not get the attention they need. This is ironic.

NFPA 70E 2015 (Article 205.2) mentions "A single line diagram, where provided for the electrical system, shall be maintained in a legible condition and shall be kept current".

Updated one-line diagram provide brief maps of equipment, redundancy, and protection. Regular updates with every change needed, no matter how small.

These documents form the basis for the work of many other related functions. As an example:

- Safety management personnel and electrical maintenance personnel use SLD in the context of hazardous energy control programs and LOTO practices (log-out take-out)
- An accurate single-line diagram is needed at project tender for bidding to be accurate
- Regulations require it
- The latest SLD documents are needed when there are plans to expand factories or building facilities

Modifications to the electrical system can present new hazards. For example, changing a motor or transformer can create a greater fault current than before. Over current protection devices that have been set at a certain level can fail to work without warning.

One-line diagram documentation is also used for efficient maintenance scheduling, safety evaluation and more.

– – – – – – – – – – – – – – – –

In electrical safety, based on the NFPA 70E on Electrical Safety Standards at Work, there are several studies, assessments and evaluations that require the need for us to update single-line diagrams so that these activities can be carried out. The studies, assessments and evaluations related to electrical safety include:

- Short-circuit study

- Protection coordination
- Arc flash study
- Log-out and Take-Out (LOTO) program
- Electrical safety studies and evaluations
- Electrical safety procedures
- etc

3.3. Histogram

A histogram can be defined as a set of rectangles with bases along with the intervals between class boundaries. Each rectangle bar depicts some sort of data and all the rectangles are adjacent. The heights of rectangles are proportional to corresponding frequencies of similar as well as for different classes. Let's learn about histograms more in detail.

A **histogram** is the graphical representation of data where data is grouped into continuous number ranges and each range corresponds to a vertical bar.

- The horizontal axis displays the number range.
- The vertical axis (frequency) represents the amount of data that is present in each range.

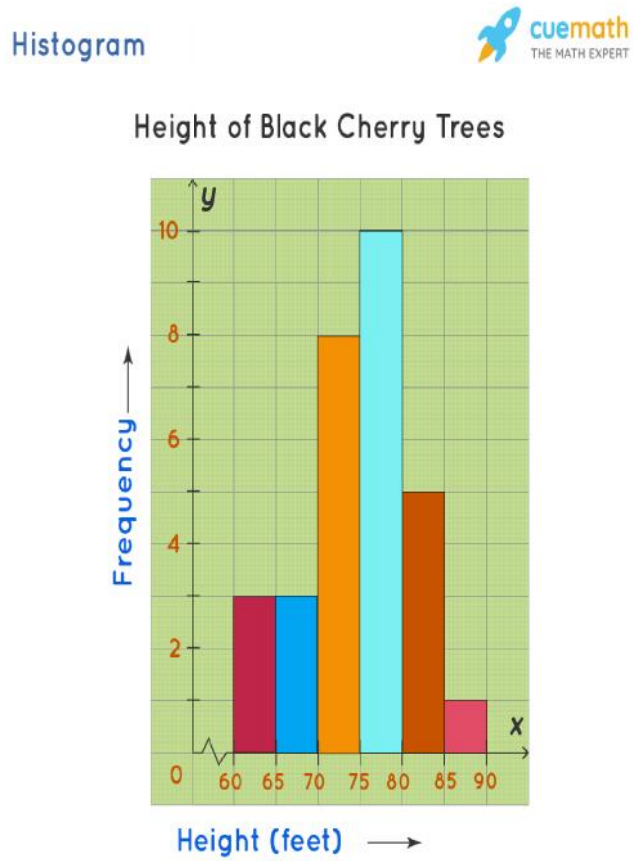
The number ranges depend upon the data that is being used.

A histogram graph is a bar graph representation of data. It is a representation of a range of outcomes into columns formation along the x-axis. in the same histogram, the number count or multiple occurrences in the data for each column is represented by the y-axis. It is the easiest manner that can be used to visualize data distributions. Let us understand the **histogram graph** by plotting one for the given below example.

Uncle Bruno owns a garden with 30 black cherry trees. Each tree is of a different height. The height of the trees (in inches): 61, 63, 64, 66, 68, 69, 71, 71.5, 72, 72.5, 73, 73.5, 74, 74.5, 76, 76.2, 76.5, 77, 77.5, 78, 78.5, 79, 79.2, 80, 81, 82, 83, 84, 85, 87. We can group the data as follows in a frequency distribution table by setting a range:

Height Range (ft)	Number of Trees (Frequency)
60 - 75	3
66 - 70	3
71 - 75	8
76 - 80	10
81 - 85	5
86 - 90	1

This data can be now shown using a histogram. We need to make sure that while plotting a histogram, there shouldn't be any gaps between the bars.



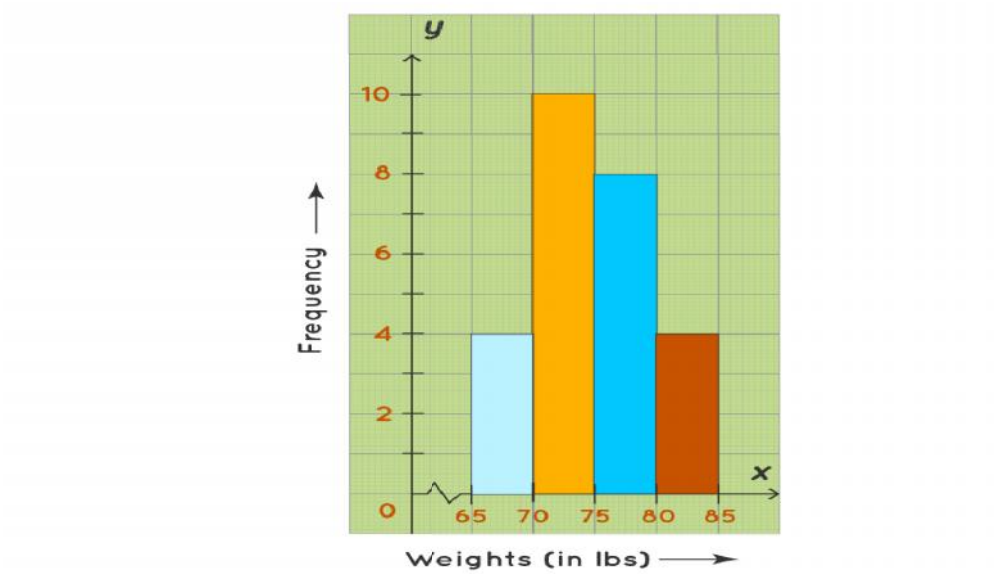
The process of making a histogram using the given data is described below:

- **Step 1:** Choose a suitable scale to represent weights on the horizontal axis.
- **Step 2:** Choose a suitable scale to represent the frequencies on the vertical axis.
- **Step 3:** Then draw the bars corresponding to each of the given weights using their frequencies.

Example: Construct a histogram for the following frequency distribution table that describes the frequencies of weights of 25 students in a class.

Weights (in lbs)	**Frequency (Number of students)**
65 - 70	4
70 - 75	10
75 - 80	8
80 - 85	4

- **Step 1:** On the horizontal axis, we can choose the scale to be 1 unit = 11 lb. Since the weights in the table start from 65, not from 0, we give a break/kink on the X-axis.
- **Step 2:** On the vertical axis, the frequencies are varying from 4 to 10. Thus, we choose the scale to be 1 unit = 2.
- **Step 3:** Then draw the bars corresponding to each of the given weights using their frequencies.

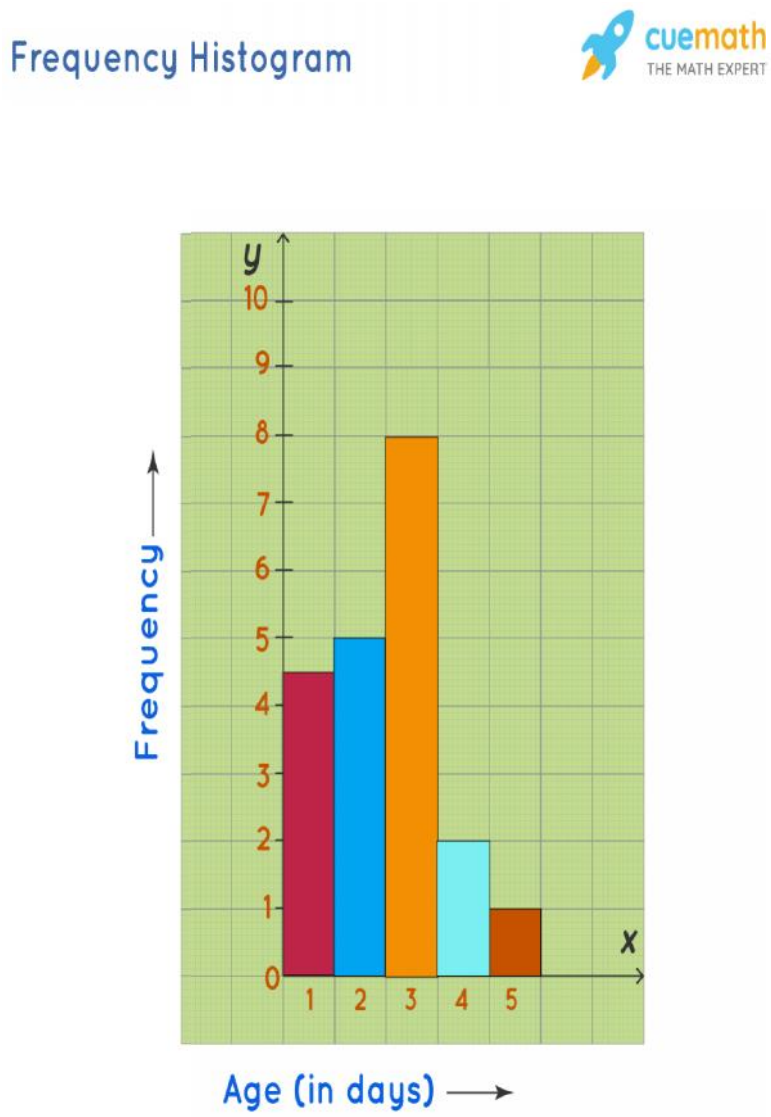


Frequency Histogram

A frequency histogram is a histogram that shows the frequencies (the number of occurrences) of the given data items. For example, in a hospital, there are 20 newborn babies whose ages in increasing order are as follows: 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 5. This information can be shown in a frequency distribution table as follows:

Age (in days)	Frequency
1	4
2	5
3	8
4	2
5	1

This data can be now shown using a frequency histogram.



Histogram Shapes

The histogram can be classified into different types based on the frequency distribution of the data. There are different types of distributions, such as normal distribution, skewed distribution, bimodal distribution, multimodal distribution, comb distribution, edge peak distribution, dog food distribution, heart cut distribution, and so on. The histogram can be used to represent these different types of distributions. We have mainly 5 types of histogram shapes. They are listed below:

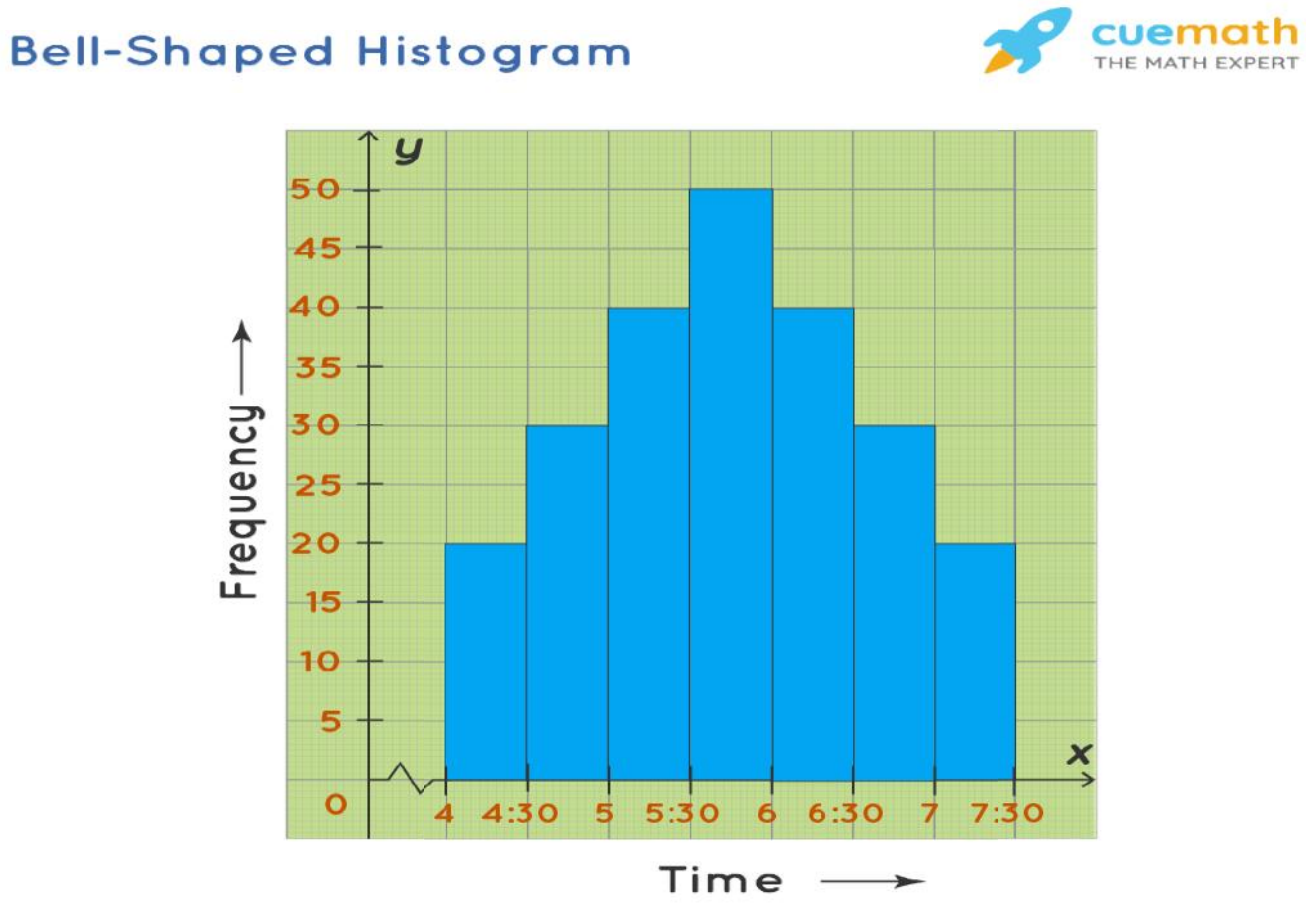
1. Bell Shaped Histogram
2. Bimodal Histogram
3. Skewed Right Histogram
4. Skewed Left Histogram
5. Uniform Histogram

Let us discuss the above-mentioned types of histogram or histogram shapes in detail with the help of practical illustrations.

Bell-Shaped Histogram

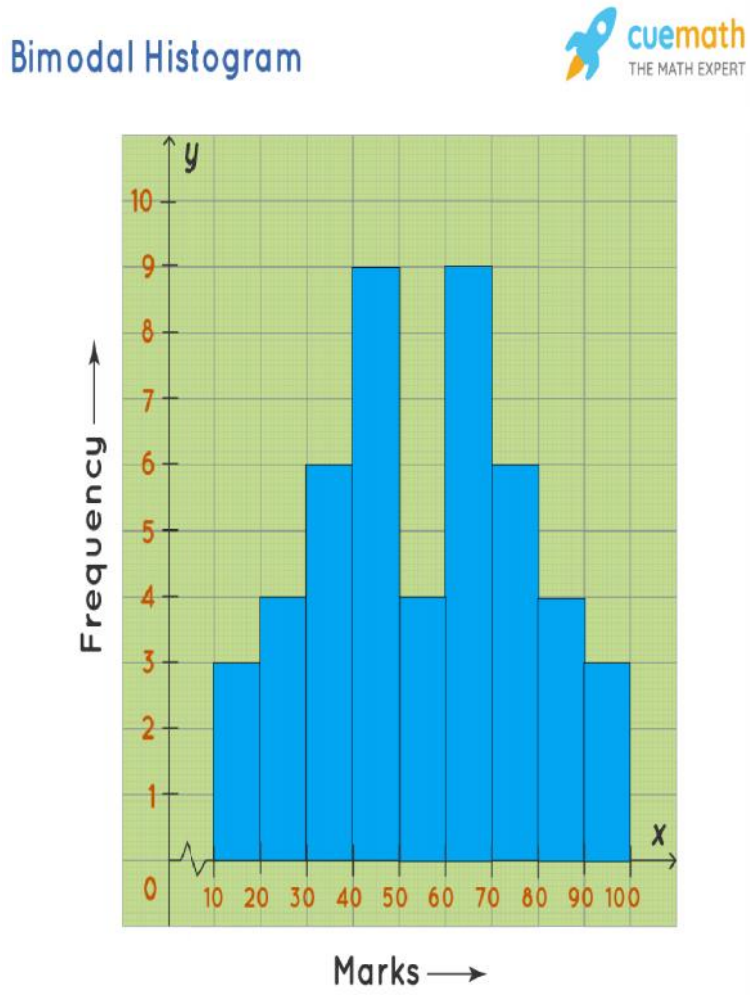
A bell-shaped histogram has a single peak. The histogram has just one peak at this time interval and hence it is a **bell-shaped histogram**. For example**,** the following histogram shows the number

of children visiting a park at different time intervals. This histogram has only one peak. The maximum number of children who visit the park is between 5.30 PM to 6 PM.



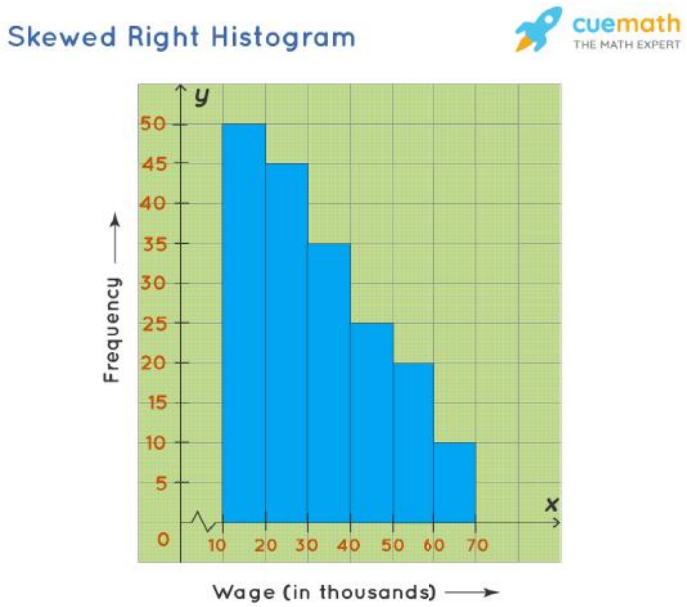
Bimodal Histogram

A bimodal histogram has two peaks and it looks like the graph given below. For example, the following histogram shows the marks obtained by the 48 students of Class 8 of St.Mary's School. The maximum number of students have scored either between 40 to 50 marks OR between 60 to 70 marks. This histogram has two peaks (between 40 to 50 and between 60 to 70) and hence it is a **bimodal histogram**.



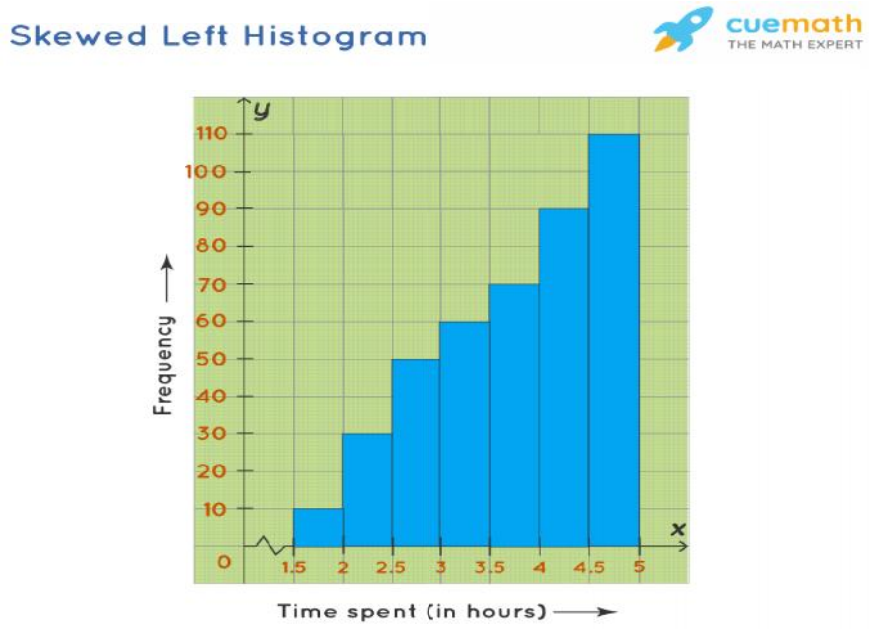
Skewed Right Histogram

A skewed right histogram is a histogram that is skewed to the right. In this histogram, the bars of the histogram are skewed to the right, hence called a **skewed right histogram**. For example, the following histogram shows the number of people corresponding to different wage ranges. The histogram is skewed to the right. For the maximum number of people, wages ranged from 10-20(thousands)



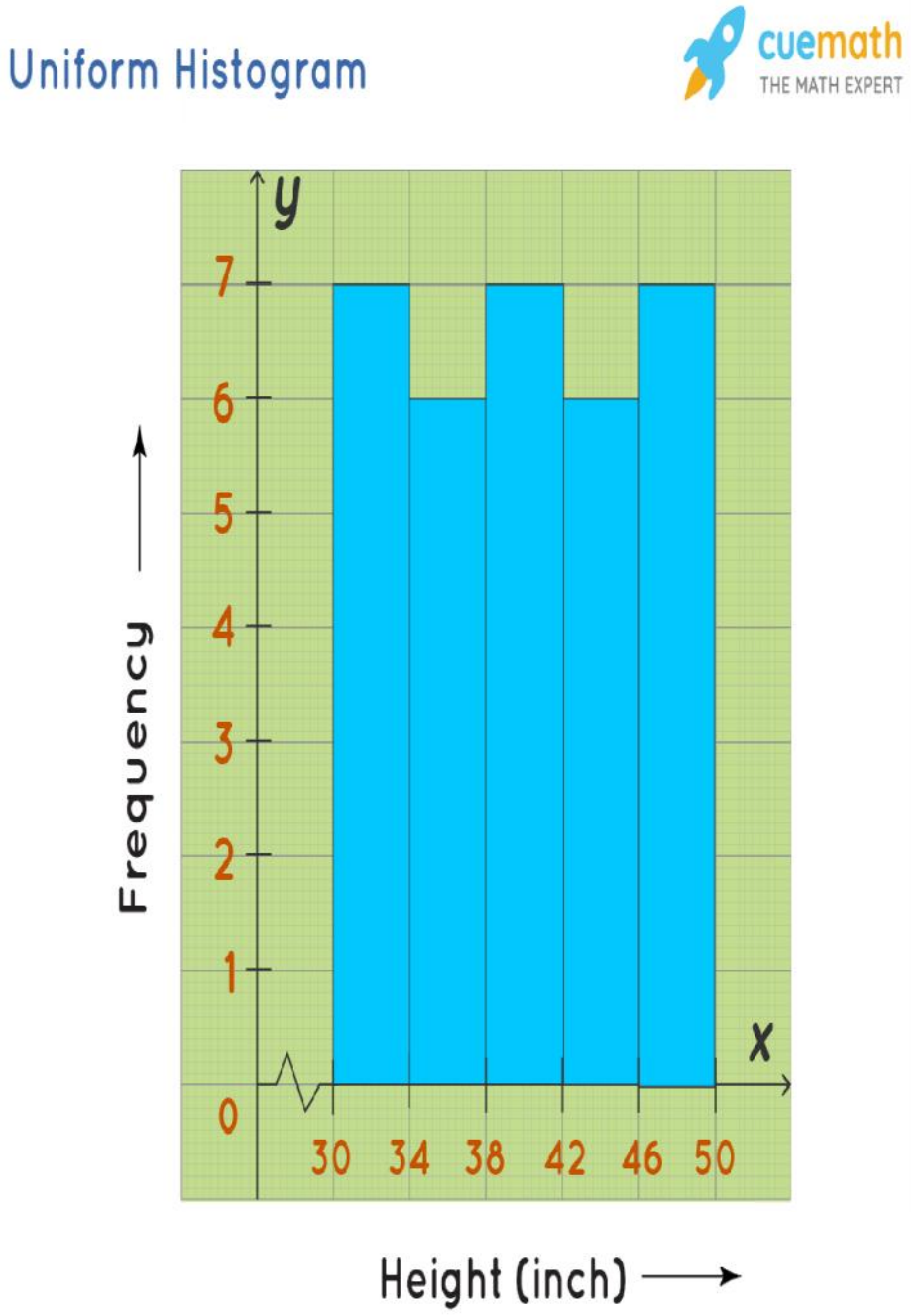
Skewed Left Histogram

A skewed left histogram is a histogram that is skewed to the left. In this histogram, the bars of the histogram are skewed to the left side, hence, called a skewed left histogram. For example, the following histogram shows the number of students of Class 10 of Greenwood High School according to the amount of time they spent on their studies on a daily basis. The maximum number of students study 4.5-5(hours) on daily basis.



Uniform Histogram

A uniform histogram is a histogram where all the bars are more or less of the same height. In this histogram, the lengths of all the bars are more or less the same. Hence, it is a uniform histogram. For example, Ma'am Lucy, the Principal of Little Lilly Playschool, wanted to record the heights of her students. The following histogram shows the number of students and their varying heights. The height of the students ranges between 30 inches to 50 inches.



3.4. Bar Graph

- A **bar graph** is the graphical representation of categorical data using rectangular bars where the length of each bar is proportional to the value they represent.

The main differences between a bar chart and a histogram are as follows:

Bar Graph	Histogram
Equal space between every two consecutive bars.	No space between two consecutive bars. They should be attached to each other.

X-axis can represent anything.	X-axis should represent only continuous data that is in terms of numbers.

A bar graph is a specific way of representing data using rectangular bars where the length of each bar is proportional to the value they represent. It is a graphical representation of data using bars of different heights. In real life, bar graphs are commonly used to represent business data.

A **bar graph** is a graph that shows complete data with rectangular bars and the heights of bars are proportional to the values that they represent. The bars in the graph can be shown vertically or horizontally. Bar graphs are also known as bar charts and it is a pictorial representation of grouped data. It is one of the ways of data handling. Bar graph is an excellent tool to represent data that are independent of one another and that do not need to be in any specific order while being represented. The bars give a visual display for comparing quantities in different categories. The bar graphs have two lines, horizontal and vertical axis, also called the x and y-axis along with the title, labels, and scale range.

Some properties that make a bar graph unique and different from other types of graphs are given below:

- All rectangular bars should have equal width and should have equal space between them.
- The rectangular bars can be drawn horizontally or vertically.
- The height of the rectangular bar is equivalent to the data they represent.
- The rectangular bars must be on a common base.

A bar graph is mostly used in mathematics and statistics. Some of the uses of the bar graph are as follows:

- The comparisons between different variables are easy and convenient.
- It is the easiest diagram to prepare and does not require too much effort.
- It is the most widely used method of data representation. Therefore, it is used by various industries.
- It is used to compare data sets. Data sets are independent of one another.
- It helps in studying patterns over long periods of time.

Types of Bar Graphs

Bar Graphs are mainly classified into two types:

- **Vertical Bar Graph**
- **Horizontal Bar Graph**

The bars in bar graphs can be plotted horizontally or vertically, but the most commonly used bar graph is the vertical bar graph. Apart from the vertical and horizontal bar graphs, there are two more types of bar graphs, which are given below:

- **Grouped Bar Graph**
- **Stacked Bar Graph**

Let us understand all the types of bar graphs in detail.

When the given data is represented vertically in a graph or chart with the help of rectangular bars that show the measure of data, such graphs are known as vertical bar graphs. The rectangular bars are vertically drawn on the x-axis, and the y-axis shows the value of the height of the rectangular bars which represents the quantity of the variables written on the x-axis.

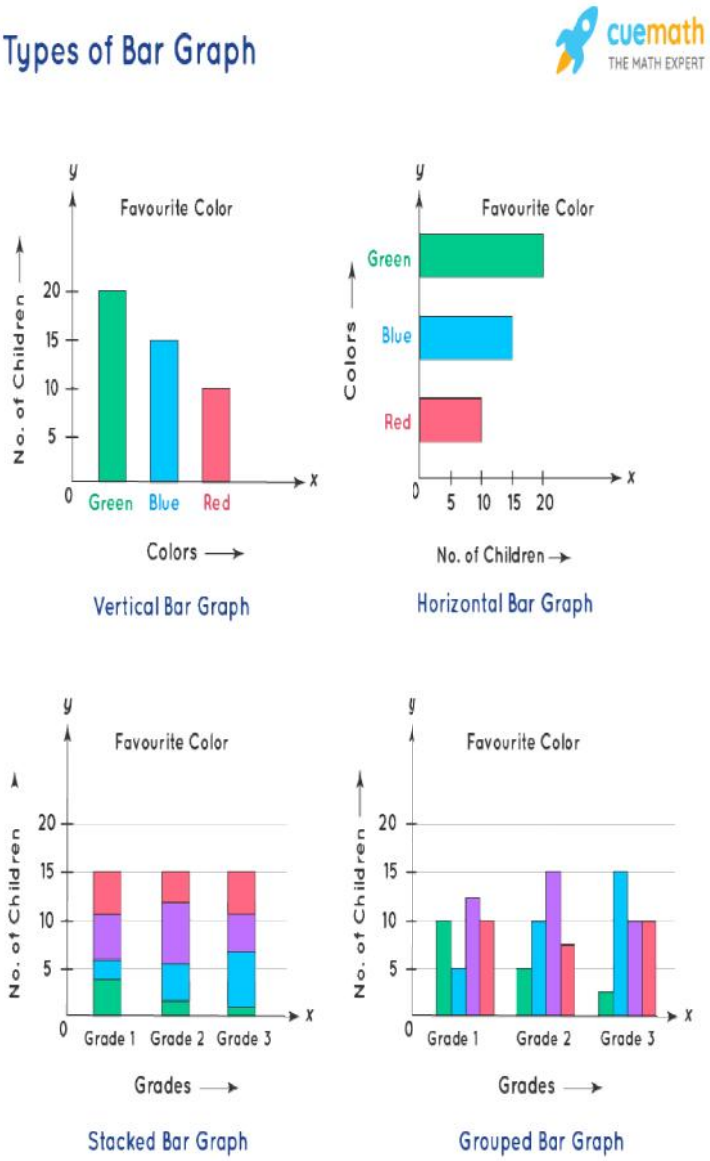
When the given data is represented horizontally by using rectangular bars that show the measure of data, such graphs are known as horizontal bar graphs. In this type, the variables or the categories of the data have to be written and then the rectangular bars are horizontally drawn on the y-axis and the x-axis shows the length of the bars equal to the values of different variables present in the data.

The stacked bar graph is also referred to as the composite bar graph. It divides the whole bar into different parts. In this, each part of a bar is represented using different colors to easily identify the different categories. It requires specific labeling to indicate the different parts of the bar. Thus, in a stacked bar graph every rectangular bar represents the whole, and each segment in the rectangular bar shows the different parts of the whole. It can be shown vertically or horizontally.

Grouped Bar Graph

The grouped bar graph is also referred the clustered bar graph. It is used to show the discrete value for two or more categorical data. In this, rectangular bars are grouped by position for levels of one categorical variable, with the same colors showing the secondary category level within each group. It can be shown both vertically and horizontally.

Observe the figure given below which shows different types of bar graphs.



Let us understand how to draw a bar graph with help of an example. Liza went to the market for buying different types of fruits in different quantities of each- 5 apples, 3 mangoes, 2 watermelons, 3 strawberries, 6 oranges. She wants to display the data by making a bar graph so that she can visually understand which type of fruits she buys the most.

Let us use the following steps to make a bar graph of the most bought fruit.

- Step 1: Take a graph paper and give the title of the bar graph like "Most Bought Fruit".
- Step 2: Draw the horizontal axis (x-axis) and vertical axis (y-axis) on a plane.
- Step 3: Now label the horizontal axis as "Types of Fruits" which is an independent category and the vertical axis as "Number of Fruits" which is a dependent category.
- Step 4: Label the fruits' names such as apples, mangoes, watermelon, strawberries, oranges and give an equal gap or leave equal space between each fruit on the horizontal axis.
- Step 5: Give the scale of the graph which shows the way in which numbers are used in the data. It is a system of marks at fixed intervals which helps in measuring objects. For example, the scale of a graph can be written as 1 unit = 1 fruit.
- Step 6: Now start making rectangular bars with equal gaps for each fruit and give height to their respective numbers.

- Step 7: The bar graph is ready, observe the height of rectangular bars of each fruit and find out the most bought fruit.

While drawing a bar graph it is very important to mention four things - labels on axes, title, scale, and name of the axes.

3.5. What is a Bar Chart?

A bar chart is a representation of numerical data in pictorial form of rectangles (or bars) having uniform width and varying heights." They are also known as bar graphs. Bar charts are one of the means of data handling in statistics. The bar charts have three major characteristics such as:

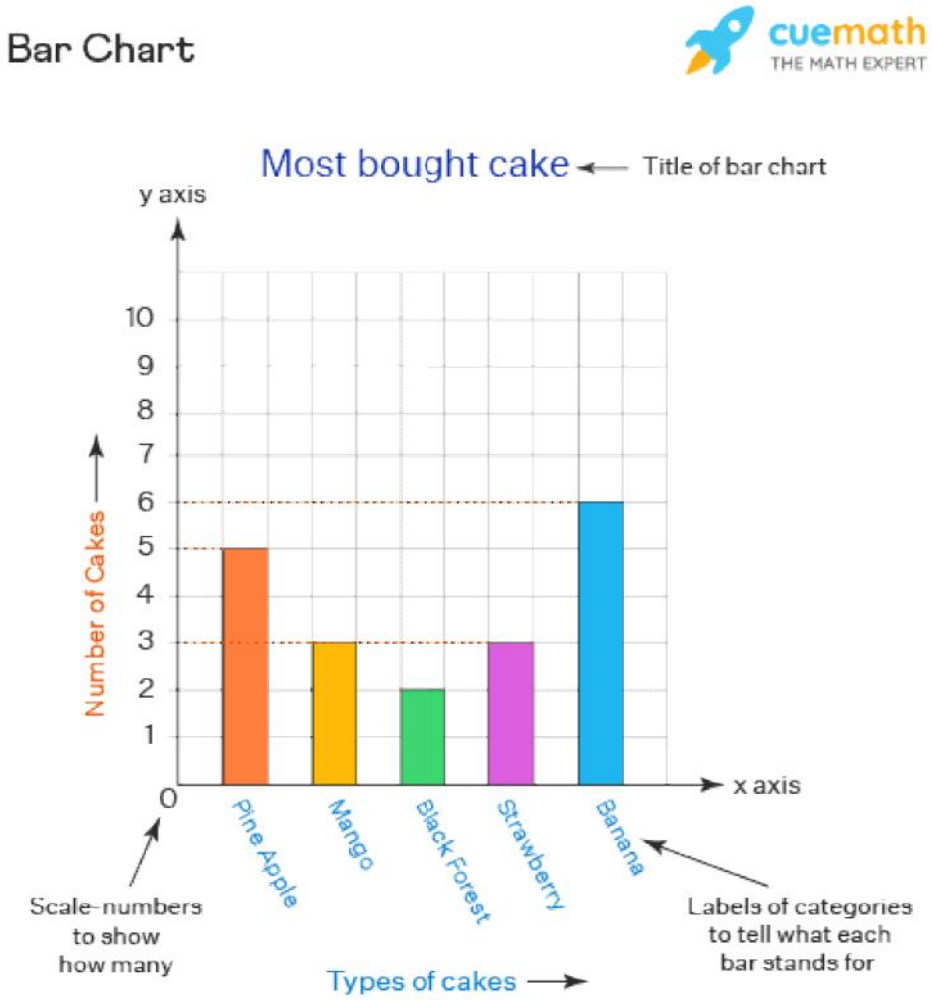
- The bar charts are used to compare the different data among different groups.
- Bar charts show the relationship with the help of two axes. On one axis it represents the categories and on another axis, it represents the discrete values.
- Over a period of time bar charts shows the major changes in available data.

Let's understand how to draw a bar chart with help of an example. Rose went to the market for buying different types of cakes in different quantities of each- 5 Pine-apple cakes, 3 Mango flavored, 2- Black forest, 3 Strawberry flavored 6 Banana flavored cakes. She wants to display the data of cakes by representing the given information on a bar chart so that she can visually understand which type of cake she purchased the most.

Let us follow the below-mentioned steps to make a bar chart.

- Step 1: Take a graph chart and give the title of the bar chart like "Most Bought Cake".
- Step 2: Draw the horizontal axis (x-axis) and vertical axis (y-axis) on graph paper or chart.
- Step 3: Now label the horizontal axis as "Types of Cakes" and the vertical axis as "Number of Cakes".
- Step 4: Label the flavor of cake names such as pineapple, mango, black forest, strawberry, and, banana. Leave equal space between each flavor name on the horizontal axis.
- Step 5: Fix the scale parameter on the vertical axis for the given data such as 1,2,3,4,5,6,7,8,9,10.
- Step 6: Now start making rectangular bars with equal gaps for each fruit-flavored cake and raise the bar to a certain height according to their respective numbers.
- Step 7: The bar chart is ready, observe the height of the rectangular bars of each cake and find out the most bought cake easily with help of visual representation.

While drawing a bar chart it is necessary to mark four important parameters to have an easy read- labels on axes, the title of a bar chart, scale, and name of the axes.



From the above bar chart, we can easily say that banana flavored cake is the most bought fruit cake that Rose purchased.

Bar Charts are mainly classified into two types:

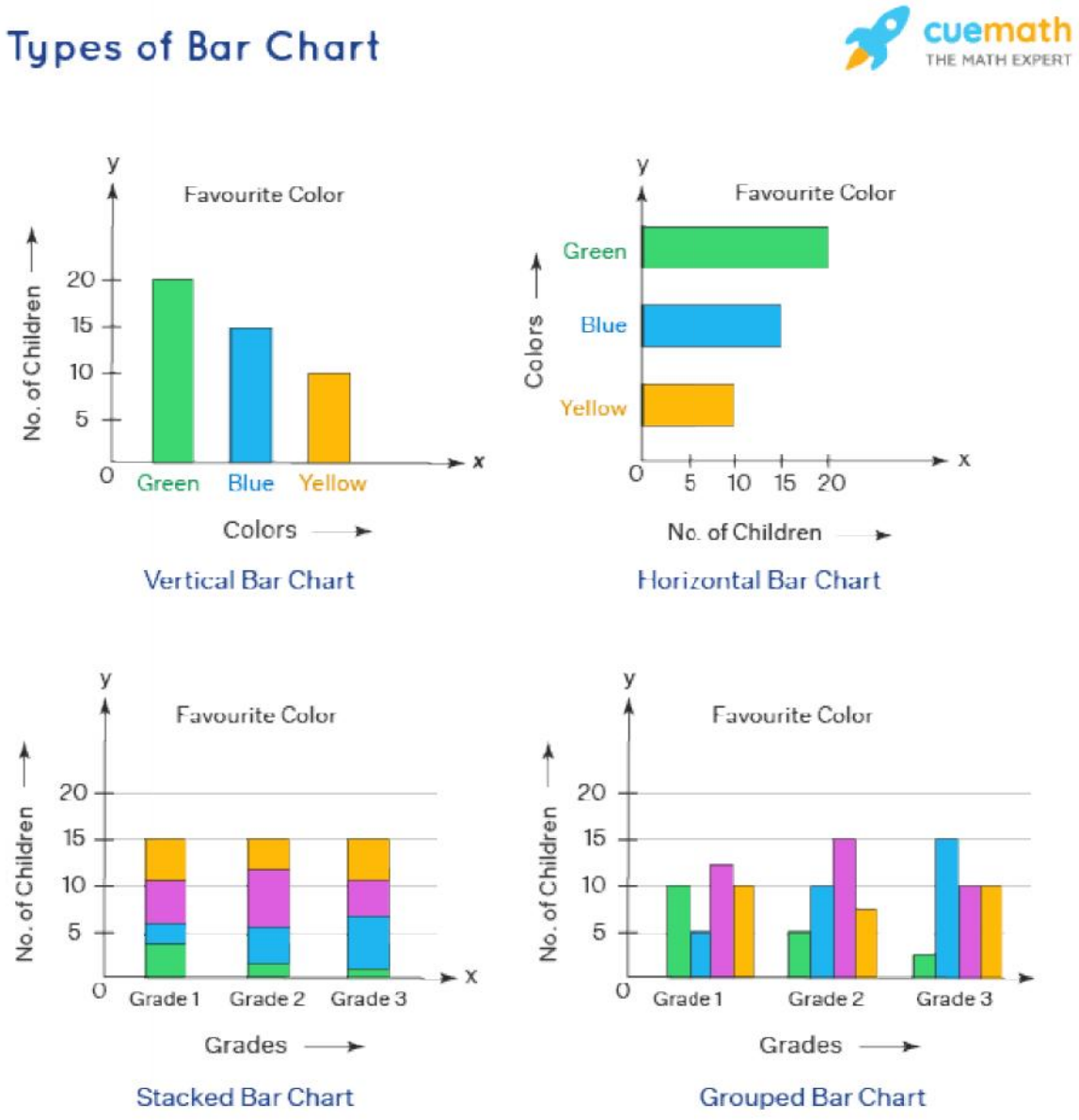
- **Horizontal Bar Charts:** When the given data is represented via horizontal bars on a graph (chart) paper such graphs are known as horizontal bar charts. These horizontal rectangular bars show the measures of the given data. In this type, the categories of the data are marked on the x-axis and y-axis. The y-axis category shows the horizontal representation of the bar chart.
- **Vertical Bar Charts:** When the given data is represented via vertical bars on a graph (chart) paper it is known as a vertical bar chart. These vertical rectangular bars represent the measure of data. The rectangular bars are vertically drawn on the x-axis and the y-axis. These rectangular bars represent the quantity of the variables written on the x-axis.

The bars in bar charts can be plotted horizontally or vertically. Amongst the two types, the most used bar chart type is the vertical bar chart. Further, the vertical and horizontal bar charts are classified as:

- **Vertical (or Horizontal) Grouped Bar Charts:** The grouped bar chart is also referred the clustered bar chart (graph). It represents the discrete value for two or more categorical data. In this, horizontal (or vertical) bars are grouped by position. For example, if the bar chart is representing 3 groups with many variables (one group with 4 data values) then each value will be represented by different colors. The color coding for each group will be the same for similar data values.

- **Vertical (or Horizontal) Stacked Bar Charts:** The stacked bar chart is also known as the composite bar chart. It shows the division of the whole bar chart into different parts. To easily identify the category we use different colors bars and specific labeling. Thus, in a stacked bar chart one rectangular bar represents the whole parameter. In one bar we show multiple segments with different colors. Each segment in the bar shows the different parts of that respective label. It can be drawn vertically or horizontally.

Look at the image given below showing different types of bar charts.



Few properties are listed below that shows how bar charts are unique and different from other types of graphs:

- The rectangular bars in a bar chart can be drawn horizontally or vertically.
- In a bar chart, horizontal (or vertical) rectangular bars should have equal width and space between them.
- The height of the rectangular bars in a bar chart is equivalent to the given data it represents.
- The bar chart has two axes, the x-axis, and the y-axis.
- The bar chart has labels on axes, the title of a bar chart, the scale, and the name of the axes.

Uses of Bar Chart

A bar chart is mostly used to represent the data in statistics formation. Some of the uses of the bar chart are listed below:

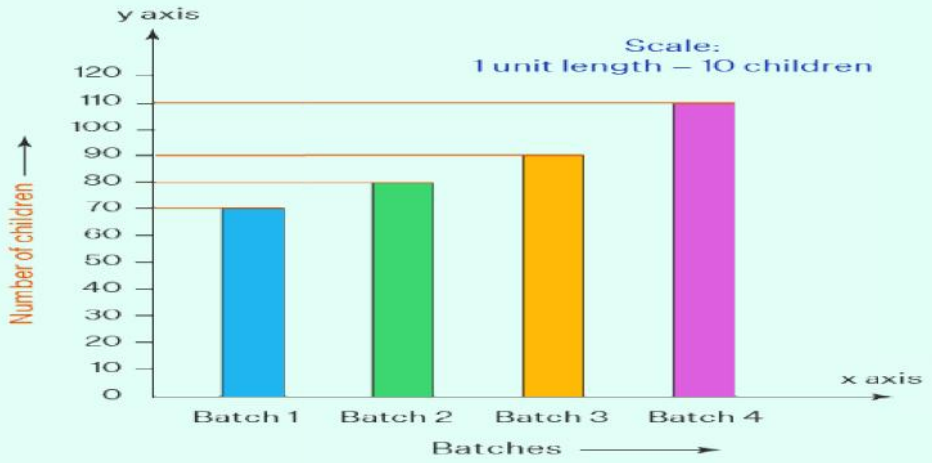
- The bar charts are used to show the comparisons between different variables. It is very easy and convenient to visualize the parameters in a pictorial form.
- The bar charts are the easiest method to show the heavy data and 9is time-saving.
- Most widely used method of data representation. Therefore, it is used by various industries.
- It helps in studying patterns over long periods of time.

Bar Chart Examples

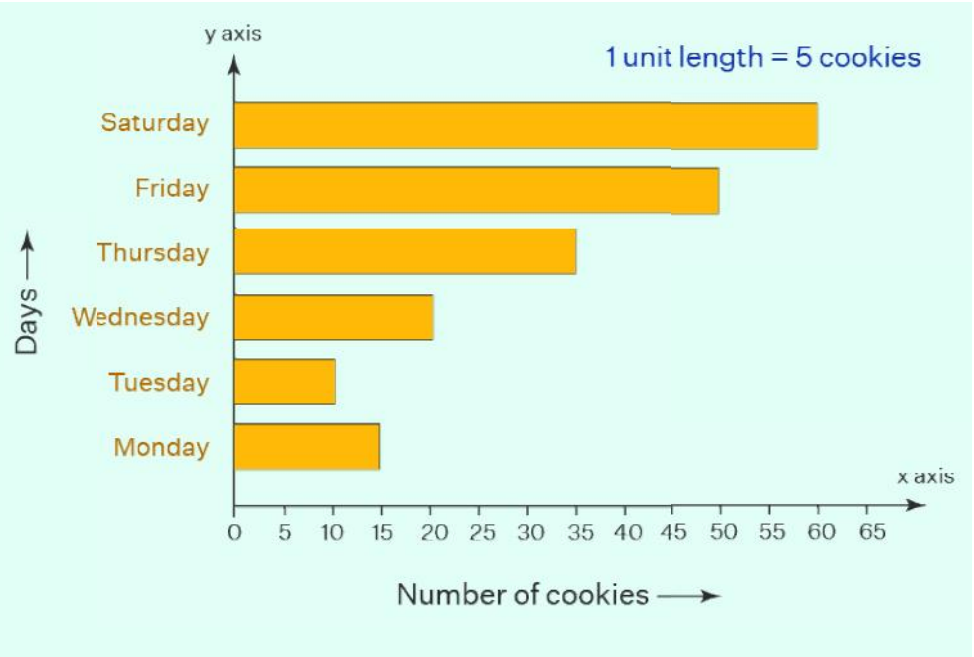
1. **Example 1: The number of children in 4 different batches of an educational institute is given below. Represent the data on a bar chart.**

Batches	Number of Children
Batch 1	70
Batch 2	80
Batch 3	90
Batch 4	110

2. **Solution:** The data is represented by the bar graph as follows:



3.
4. **Example 2: Observe the given horizontal bar chart which is showing the baking of cookies in a bakery from Monday to Saturday. Find out on which day the maximum number of cookies were baked and the number of cookies baked on that day.**



Solution: From the above bar chart, it is clearly visible that on Saturday the maximum number of cookies were baked as the length of the bar on Saturday is the maximum raised. The value of the bar is raised to 60. Hence 60 cookies were baked on that day.

The "**pie chart"** is also known as a "circle chart", dividing the circular statistical graphic into sectors or sections to illustrate the numerical problems. Each sector denotes a proportionate part of the whole. To find out the composition of something, Pie-chart works the best at that time. In most cases, pie charts replace other graphs like the bar graph, line plots, histograms, etc.

Formula

The pie chart is an important type of data representation. It contains different segments and sectors in which each segment and sector of a pie chart forms a specific portion of the total(percentage). The sum of all the data is equal to 360°.

The total value of the pie is always 100%.

To work out with the percentage for a pie chart, follow the steps given below:

- Categorize the data
- Calculate the total
- Divide the categories
- Convert into percentages
- Finally, calculate the degrees

Therefore, the pie chart formula is given as

(Given Data/Total value of Data) × 360°

Note: It is not mandatory to convert the given data into percentages until it is specified. We can directly calculate the degrees for given data values and draw the pie chart accordingly.

Imagine a teacher surveys her class on the basis of favourite Sports of students:

Football	Hockey	Cricket	Basketball	Badminton
10	5	5	10	10

The data above can be represented by a pie chart as following and by using the circle graph formula, i.e. the pie chart formula given below. It makes the size of the portion easy to understand.

Step 1: First, Enter the data into the table.

Football	Hockey	Cricket	Basketball	Badminton
10	5	5	10	10

Step 2: Add all the values in the table to get the total.

I.e. Total students are 40 in this case.

Step 3: Next, divide each value by the total and multiply by 100 to get a per cent:

Football	**Hockey**	**Cricket**	**Basketball**	**Badminton**
(10/40) × 100	(5/ 40) × 100	(5/40) ×100	(10/ 40) ×100	(10/40)× 100
=25%	=12.5%	=12.5%	=25%	=25%

Step 4: Next to know how many degrees for each "pie sector" we need, we will take a full circle of 360° and follow the calculations below:

The central angle of each component = (Value of each component/sum of values of all the components)×360°

Football	**Hockey**	**Cricket**	**Basketball**	**Badminton**
(10/ 40)× 360°	(5 / 40) × 360°	(5/40) × 360°	(10/ 40)× 360°	(10/ 40) × 360°
=90°	=45°	=45°	=90°	=90°

Now you can draw a pie chart.

Step 5: Draw a circle and use the protractor to measure the degree of each sector.

Let us take an example for a pie chart with an explanation here to understand the concept in a better way.

Question: The percentages of various cops cultivated in a village of particular distinct are given in the following table.

Items	**Wheat**	**Pulses**	**Jowar**	**Groundnuts**	**Vegetables**	**Total**

Percentage of cops	**125/3**	**125/6**	**25/2**	**50/3**	**25/3**	**100**

Represent this information using a pie-chart.

Solution:

The central angle = (component value/100) × 360°

The central angle for each category is calculated as follows

Items	Percentage of cops	Central angle
Wheat	125/3	[(125/3)/100] × 360° = 150°
Pulses	125/6	[(125/6)/100] × 360° = 75°
Jowar	25/2	[(25/2)/100] × 360° = 45°
Groundnuts	50/3	[(50/3)/100] × 360° = 60°
Vegetables	25/3	[(25/3)/100] × 360° = 30°
Total	100	360°

3.6. Summary

Data representation is another way of analysing numerical data. A graph is a sort of chart through which statistical data are represented in the form of lines or curves drawn across the coordinated points plotted on its surface. Graphs enable us in studying the cause and effect relationship between two variables. Graphs help to measure the extent of change in one variable when another variable changes by a certain amount.

3.7. Keywords

Histogram, Bar Graph , Bar Chart, Line Diagram

3.8. Self-Assessment

1-Data representation is important– true/ false

2- Data representation is used in Psychology true/ false

3- Data representation is used in psychology only true/ false

4- Psychology makes use of data true/ false

5- Statistics is science true/ false

6- Data representation is philosophy true/ false

7- Psychology was philosophically related to study of soul. True/ false

8- Psychology differs from other arts disciplines True/ false

9- Data representation is ambiguous True/ false

10- Sigmund Freud did the basic experimentation in Psychology and they were scientific in nature True/ false

11- Sigmund Freud is labeled as father of statistics True/ false

12- Statistics is now known as study of soul True/ false

13- Changing Human behaviour is the subject matter of Psychology True/ false

14- Human behaviour is considered to be very dynamic True/ false

15- Study area Psychology and Philosophy were separated earlier True/ false

3.9. Review Questions

- What's data representation? Discuss its relevance
- What is the importance of data representation in psychology?
- Describe types of data representation with its methods.

Further Readings

Books 

- Ferguson, G.A. (1980), *Statistical Analysis in Psychology and Education*. McGraw Hill Book Co.: New York.
- Fisher, R.A. and Yates, F. (1963), *Statistical Tables for Biological, Agricultural and Medical Research*. Oliver Boyd: Edinburgh.

- https://www.sciencelearn.org.nz/resources/1851-measurement-introduction
- https://www.cuemath.com/measurement/

UNIT 04: Normal Probability Curve

Content

Objectives/Expected Learning Outcomes

Understand about Normal Probability Curve

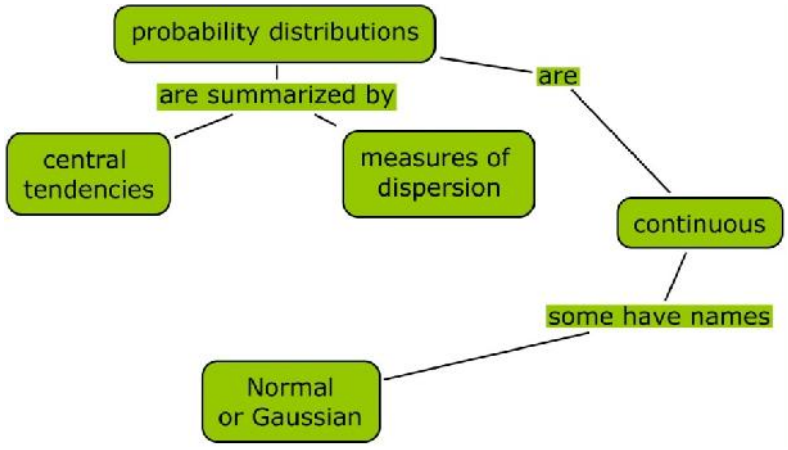
Importance of Normal Probability Curve

Basic tenets of Normal Probability Curve

Important role of Normal Probability Curvein psychology

Introduction

This module introduces one of the most important distributions in statistics: the normal distribution.



After completion of this module the student will be able to

- recognize the normal distribution
- recognize the standard normal distribution
- perform calculations with the normal distribution using Excel
- convert a normal distribution to a standard normal distribution

Knowledge and Skills

- normal distribution
- standard normal distribution

Prerequisites

- sample mean, average
- sample standard deviation
- sample variance
- percentile

4.1. Characteristics

The U.S. Department of Health of Human Services (HHS) is a U.S. government agency with responsibility to "protect the health of all Americans and providing essential human services" (http://www.hhs.gov/about/). One of HHS's agencies is the Centers for Disease Control and Prevention (CDC), which oversees a number of coordinating centers/offices and the National Institutes of Health including the Coordinating Center for Health Information and Service (CCHIS). The National Center for Health Statistics (NCHS) is an office within the CCHIS. According to NCHS's website, its mission "is to provide statistical information that will guide actions and policies to improve the health of the American people. As the Nation's principal health statistics agency, NCHS leads the way with accurate, relevant, and timely data"(http://www.cdc.gov/nchs/). Data on the health and nutritional status of adults and children in the U.S. is collected by the National Health and Nutrition Examination Survey (NHANES). We will be looking at one of NHANES' data sets, namely the growth curves for infants. The data files can be found at http://www.cdc.gov/growthcharts/html_charts/lenageinf.htm.

Growth charts that provide percentiles for length and weight at various ages are important screening tool to monitor the growth of infants and assess their nutritional status and health risks caused by, for instance, obesity. The following data[1] lists selected percentiles for height (cm) for boys at 0 months:

Percentile	3	5	10	25	50	75	90	95	97
Height [cm]	44.93	45.57	46.55	48.19	49.99	51.77	53.36	54.31	54.92

The distribution of the length follows a certain pattern that is described by the **normal distribution**.

We say that a random variable is normally distributed with mean μ and standard deviation σ if the **probability density function** is given by

$$f(x) = \frac{1}{\pm\sqrt{2®}} e^{-\frac{(x-«)^2}{2\pm^2}} \text{ for } -\infty < x < \infty$$

Because of its importance in statistics, it is worthwhile to memorize the form of the density function. The cumulative distribution function $F(x) = P(X \leq x)$ needs to be calculated by a computer or looked up in tables.

Read Subsections 6.1.1 and 6.1.2

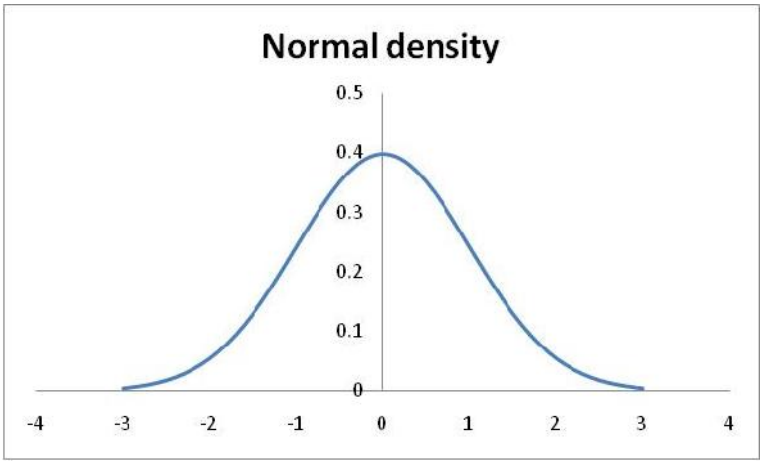


Figure 1: The probability density function of a normal distribution with mean 0 and standard deviation 1.

The probability density function of a normal distribution with mean 0 and standard deviation 1 is displayed in Figure 1. Note the bell-shaped curve.

EXCEL has a function that returns the cumulative distribution function and the density function for a normal distribution with a specified mean and standard deviation. The syntax is

NORMDIST(x,mean,standard_dev,cumulative)

Where **x** is the value at which you want to evaluate the distribution, **mean** is the mean, **standard_dev** is the standard deviation, and **cumulative** is a logical value (TRUE for the cumulative distribution function and FALSE for the probability density function).

In-class Activity 1

(a) The height for boys at age 0 is normally distributed with mean 49.99cm and standard deviation 2.66cm. Use EXCEL to confirm the percentiles given in the table above.

Instruction: To determine the percentile of the value 44.93cm, you need to calculate the probability that the height is less than or equal to the value 44.93cm. The following function in EXCEL calculates this value

=NORMDIST(44.93,49.99,2.66, TRUE)

Rounded to two decimal places, the answer is 0.03, which is the 3rd percentile, thus confirming the result in the table. Repeat this for the other values given in the table

Computation of Normal Probability Curve:

If a coin is tossed unbiased it will fall either head (H) or tail (T). This the probability of appearing a head is one chance in two. So the probability ratio of H is ½ and T is ½.

$$\text{Thus, } (H + T)^1 = H\frac{1}{2} + T\frac{1}{2} = 1.00$$

Likewise of we shall toss two coins, coin x and coin y there are four possible ways of falling.

1		2		3		4	
x	y	x	y	x	y	x	y
H	H	T	H	H	T	T	T

Thus the four possible ways are-both x and y may fall H, x may fall T and y H, x may fall H and yT or both may fall T.

Expressed in ratios

Probability of two heads = ¼

Probability of two tails = ¼

Probability of one H and one T = ¼

Probability of one T and one H = ¼

Thus the ratio is ¼ + ½ + ¼ = 1.00

ADVERTISEMENTS:

The expected appearance of heads and tails of two coins can be expressed as:

(H + T)2 = H2 + 2HT + T2

If we shall increase the number of coins to three i.e. x, y and Z, there can be eight possible arrangements.

1	2	3	4	5	6	7	8
x y z	x y z	x y z	x y z	x y z	x y z	x y z	x y z
H H H	H H T	H T H	T H H	H T T	T H T	T T H	T T T

The expected appearance of heads and tails of coins can be expressed as:

$$(H + T)^3 = H^3 + 3H^2T + 3HT^2 + T^3$$

where

1 H^3 = Three heads; one out of 8; Probability ratio = $\frac{1}{8}$

1 H^3 = Three heads; one out of 8; Probability ratio = $\frac{1}{8}$

3 H^2T = Two heads 1 tail; 3 out of 8; Probability ratio = $\frac{3}{8}$

3 HT^2 = 1 heads 2 tails; 3 out of 8; Probability ratio = $\frac{1}{8}$

Total = 1

In this way we can determine the probability of different combinations of heads and tails of any number of coins. We can obtain probability of any number of coins by binomial expansion. An expression containing two terms is called a binomial expres¬sion. Binomial theorem is an algebraic formula which expands the power of a binomial expression in the form of a series.

The formula reads like this:

(H + T)n = C (n, 0) Hn + C (n, 1) Hn-1 T + C (n, 2) H(n-2) T2 ….

ADVERTISEMENTS:

… +C(n,r) Hn-r Tr + …. + C(n,n)Tn… (11.1)

Where C = Possible combinations.

C(n,r) = n!/r! (n – r)!

n! means 1 x 2 x 3 x …. x n

n = Total number of observations or persons.

r = Number of observations or persons taken at a time.

Thus binomial expansion of

$$(T + H)^{10} = H^{10} + 10H^9 T^1 + 45 H^8 T^2 + 120 H^7 T^3 + 210 H^6 T^4 + 252 H^5 T^5 + 210 H^4 T^6 + 120 H^3 T^7 + 45 H^2 T^8 + 10 H^1 T^{10} + T^{10}$$

Where

$1\ H^{10}$ = 10 heads : 1 out of 1024 :

$$\text{Probability ratio} = \frac{1}{1024}$$

$10\ H^9T^1$ = 9 heads 1 tail : 10 out of 1024

$$\text{probability ratio} = \frac{10}{1024}$$

$45\ H^8T^2$ = 8 heads 2 tail : 45 out of 1024

$$\text{probability ratio} = \frac{45}{1024}$$

$120\ H^7T^3$ = 7 heads 3 tail : 120 out of 1024

$$\text{probability ratio} = \frac{120}{1024}$$

$210\ H^6T^4$ = 6 heads 4 tails : 210 out of 1024

$$\text{probability ratio} = \frac{210}{1024}$$

$252\ H^5T^5$ = 5 heads 5 tails : 252 out of 1024

$$\text{probability ratio} = \frac{252}{1024}$$

$210\ H^4T^6$ = 4 heads 6 tails : 210 out of 1024

$$\text{probability ratio} = \frac{210}{1024}$$

$120\ H^{35}T^7$ = 3 heads 7 tails : 120 out of 1024

$$\text{probability ratio} = \frac{120}{1024}$$

$45\ H^2T^8$ = 2 heads 8 tails ; 45 out of 1024

$$\text{probability ratio} = \frac{45}{1024}$$

$10\ H^1T^9$ = 1 heads 9 tails : 10 out of 1024

$$\text{probability ratio} = \frac{10}{1024}$$

$1\ T^{10}$ = 10 tails : 1 out of 1024 probability ratio $= \frac{1}{1024}$

Total = 1

If the above data are plotted on a graph as histogram and frequency polygon it will be as below (fig. 11.1)

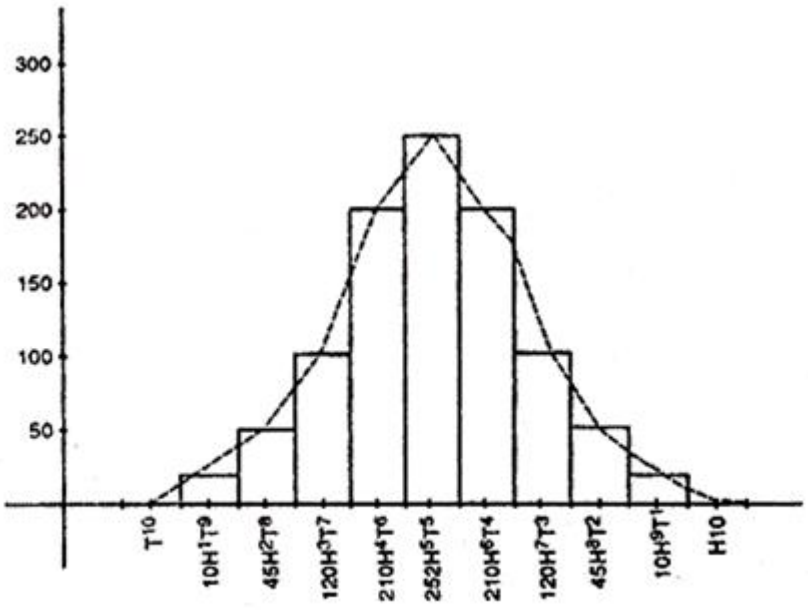


Fig. 11.1 Probability curve obtained from the expansion of $(H + T)^{10}$

Thus the figure we obtained from toss of 10 coins (H + T)10 is a symmetrical many sided polygon.

And if we shall go on in¬creasing the number of coins, with each increase the polygon would exhibit a perfectly smooth surface line the figure-11.2 given below:

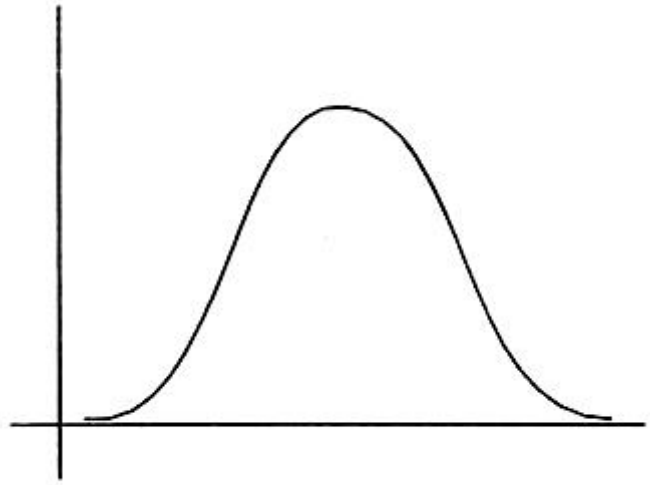


Fig. 11.2 Normal Probability Curve

This bell shaped curve is called as the 'Normal Probability Curve'. Thus the "graph of the probability density function of the normal distribution is a continuous bell shaped curve, symmetrical about the mean" is called normal probability curve.

In statistics it is important because:

(a) It is the distribution of many naturally occurring variables, such as intelligence of 8th grade students, height of the 10th grade students etc.

(b) The distribution of the means of samples drawn from most parent populations is normal or approximately so when the samples are sufficiently large.

Therefore normal curve has great significance in social sciences and behavioural sciences. In behavioural measurement most of the aspects approximates to the normal distribution. So that Nor¬mal Probability Curve or most popularly known as NPC is used as a reference curve. In order to understand the utility of the NPC we must have to understand the properties of the NPC.

Characteristics of Normal Probability Curve:

Some of the major characteristics of normal probability curve are as follows:

1. The curve is bilaterally symmetrical.

The curve is symmetrical to its ordinate of the central point of the curve. It means the size, shape and slope of the curve on one side of the curve is identical to the other side of the curve. If the curve is bisected then its right hand side completely matches to the left hand side.

2. The curve is asymptotic:

The Normal Probability Curve approaches the horizontal axis and extends from-∞ to + ∞. Means the extreme ends of the curve tends to touch the base line but never touch it.

It is depicted in figure (11.3) given below:

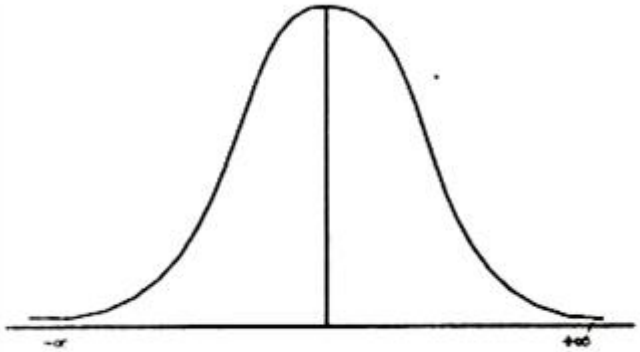


Fig. 11.3

3. The Mean, Median and Mode:

The mean, Median and mode fall at the middle point and they are numerically equal.

4. The Points of inflection occur at ± 1 Standard deviation unit:

The points of influx in a NPC occur at ± 1σ to unit above and below the mean. Thus at this point the curve changes from convex to concave in relation to the horizontal axis.

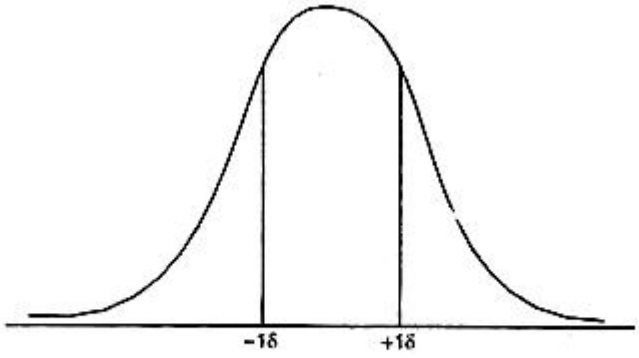


Fig. 11.4

5. The total area of NPC is divided in to ± standard deviations:

The total of NPC is divided into six standard deviation units. From the center it is divided in to three +ve′ standard deviation units and three —ve′ standard deviation units.

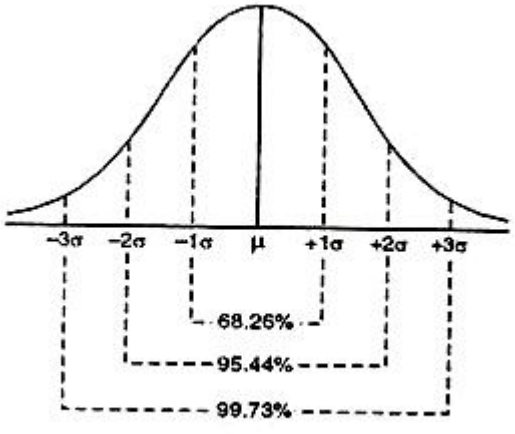


Fig. 11.5

Thus ± 3σ of NPC include different number of cases separately. Between ± 1σ lie the middle 2/3rd cases or 68.26%, between ± 2σ lie 95.44% cases and between ± 3σ lie 99.73% cases and beyond + 3σ only 0.37% cases fall.

6. The Y ordinate represents the height of the Normal Probability Curve:

The Y ordinate of the NPC represents the height of the curve. At the center the maximum ordinate occurs. The height of the curve at the mean or mid point is denoted as Y0.

In order to determine the height of the curve at any point we use the following formula:

$$Y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

Where Y = Ordinate, the height of curve abvoe the horizontal axis.

x = deviation from the mean.

σ = Standard deviation

π = a constant whose value is 3.1416

e = a constant, the value is 2.7183

7. It is unimodal:

The curve is having only one peak point. Because the maximum frequency occurs only at one point.

8. The height of the curve symmetrically declines:

The height of the curve decline to both the direction sym¬metrically from the central point. Means the M + σ and M — σ are equal if the distance from the mean is equal.

9. The Mean of NPC is μ and the standard deviation is σ:

As the mean of the NPC represent the population mean so it is represented by the μ (Meu). The standard deviation of the curve is represented by the Greek Letter, σ.

10. In Normal Probability Curve the Standard deviation is the 50% larger than the Q:

In NPC the Q is generally called the probable error or PE.

The relationship between PE and a can be stated as following:

1 PE = .6745σ

1σ = 1.4826PE.

11. Q can be used as a unit of measurement in determining the area within a given part:

12. The Average Deviation about the mean of NPC is .798σ:

There is a constant relationship between standard deviation and average deviation in a NPC.

The Average Deviation

$$= \sigma \times \sqrt{\frac{2}{\pi}} = .798\,\sigma$$

13. The model ordinate varies increasingly to the standard deviation:

In a Normal Probability curve the modal ordinate varies in¬creasingly to the standard deviation. The standard deviation of the Normal Probability Curve increases, the modal ordinate decreases and vice-versa.

4.3. Applications of Normal Probability Curve

Some of the most important applications of normal probability curve are as follows:

The principles of Normal Probability Curve are applied in the behavioural sciences in many different areas.

1. NPC is used to determine the percentage of cases in a normal distribution within given limits:

The Normal Probability Curve helps us to determine:

i. What percent of cases fall between two scores of a distribution.

ii. What percent of scores lie above a particular score of a distribution.

iii. What percent of scores lie below a particular score of a distribution.

Example:

Given a distribution of scores with a mean of 24 and σ of 8. Assuming normality what percentage of the cases will fall between 16 and 32.

Solution:

Here first of all we have to convert both the scores 16 and 32 into a standard score.

$$Z = \frac{x}{\sigma}$$

Where Z = Standard Score

x = deviation of the score (X—M)

σ = Standard deviation.

$$\text{Z Score for 16} = \frac{x}{\sigma} = \frac{16-24}{8} = \frac{-8}{8}$$

$$= -1\sigma$$

$$\text{Z score for 32} = \frac{x}{\sigma} = \frac{32-24}{\sigma} = \frac{8}{8} = +1\sigma$$

Entering in to the Table-A, the table area under NPC, it is found that 34.13 cases fall between mean and – 1σ and 34.13 cases fall between mean and + 1σ. So ± σ covers 68.26% of cases. So that 68.25% cases will fall between 16 and 32.

Example: Given a distribution of scores with a mean of 40 and σ of 8. Assuming normality what percentage of cases will lie above and below the score 36.

Solution:

First of all we have to convert the raw score 36 into standard score.

$$\text{The Z score of } 36 = Z = \frac{x}{\sigma} = \frac{36-40}{8} = \frac{-4}{8}$$
$$= -0.5\sigma$$

Entering into the Table-A, the table area under the NPC it is found that 19.15% cases fall between Mean and -.5σ. Therefore the total percentage of cases above the score 36 is 50 + 19.15 = 69.15% and below the score 36 is 50-19.15 = 30.85%. So in the distribution 69.15% cases are above the score 36 and 30.85% scores are below the score 36.

2. NPC is used to determine the value of a score whose percentile rank is given:

By using NPC table we can determine the raw score of the individual if the percentile rank is given.

Example: In a distribution of scores of a doss Pinky's percentile rank in statistics is 65. The mean of the distribution is 55 with a standard deviation of 10. Find but the raw score of Pinky in Statistics.

Solution:

As Pinky's percentile rank is 65 so in a normal distribution her position is 35% above the mean. By entering in to the table 'A' we found that 35% from the mean is + 1.04 σ.

$$\frac{X-M}{\sigma} = Z$$
$$\Rightarrow \quad \frac{X-55}{10} = 1.04$$
$$\Rightarrow \quad X - 55 = 10 \times 1.04$$
$$\Rightarrow \quad X - 55 = 10.4$$
$$\Rightarrow \quad X = 10.4 + 55 = 65.4 \text{ or } 65.$$

So Pinky's raw score in statistics is 65.

By putting the value in 'Z' score.

3. NPC is used to find the limits in a normal distribution which include a given percentage of cases:

When a distribution is normally distributed and what we know about the distribution is Mean and the Standard deviation at that time by using the table area under NPC we can determine the limits which include a given percentage of cases.

Example: Given a distribution of scores with a mean of 20 and σ of 5. If we assume normality what limits will include the middle 75% of cases.

Solution:

In a normal distribution the middle 75% cases include 37.5% cases above the mean and 37.5% cases below the mean. From the Table-A we can say that 37.5% cases covers 1.15 σ units. Therefore the middle 75% cases lie between mean and ± 1.15 σ units.

As $\sigma = 5$

So, $5 \times 1.15\,\sigma = \pm 5.57\,\sigma$ units.

Adding the value to mean we can get

$20 + 5.75 = 25.75$

$20 - 5.75 = 14.25$

So in this distribution middle 75% cases will include the limits 14.25 to 25.75.

4. It is used to compare two distributions in terms of- overlapping:

If scores of two groups on a particular variable are normally distributed. What we know about the group is the mean and standard deviation of both the groups. And we want to know how much the first group over-laps the second group or vice-versa at that time we can determine this by using the table area under NPC.

5. NPC helps us in dividing a group into sub-groups according to certain ability and assigning the grades:

When we want to divide a large group in to certain sub-groups according to some specified ability at that time we use the standard deviation units of a NPC as units of scale.

Example: An achievement test was administered to the 600 8th grade students. The teacher wants to assign these students in to 4 grades namely A, B, C and D according to their performance in the test. Assuming the normality of the distribution of scores calculate the number of students can be placed in each group.

Solution:

The area under a NPC is divided in to ± 3σ units or 6σ units.

Here we have to divide the students in to 4 sections.

So each section has $\frac{6\sigma}{4} = 1.5\ \sigma$ uits.

So if we shall distribute the section in order of merit.

The section-A will be within 1.5σ to 3σ

Section B will be within Mean to 1.5σ

Section C will be within Mean to -1.5σ

and Section D will be with in −1.5σ to – 3σ.

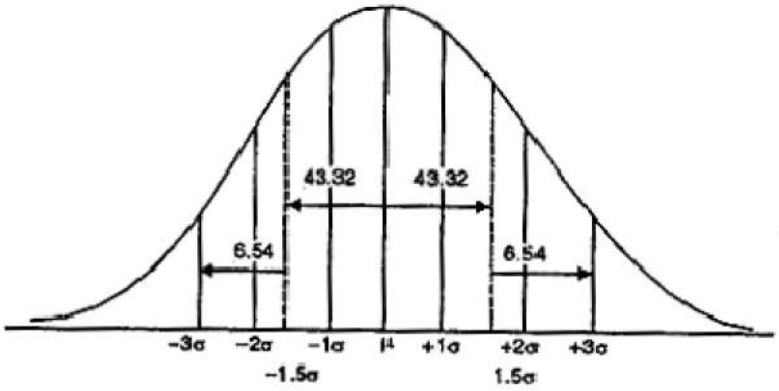


Fig. 11.9

Entering into the Table 'A' we found that
3σ covers the 49.86% cases
1.5σ covers the 43.32% cases
Grade 'A' will cover = 49.86—43.32 = 6.54%
Table area with in μ to + 1.5σ covers 43.32% cases
So grade B will cover 43.32% cases.
Table area with in μ and —1.5σ covers 43.32% cases.
So grade C will cover 43.32% cases.
Table are with in μ and —3σ covers 49.86% and —1.5σ covers 43.32% cases. So between —1.5σ and —3σ, 49.86 — 43.32 = 6.54% cases lie.

6. NPC helps to determine the relative difficulty of test items or problems:

When it is known that what percentage of students successfully solved a problem we can determine the difficulty level of the item or problem by using table area under NPC.

7. NPC is useful to normalize a frequency distribution:

In order to normalize a frequency distribution we use Normal Probability Curve. For the process of standardizing a psychological test this process is very much necessary.

8. To test the significance of observations of experiments we use NPC:

In an experiment we test the relationship among variables whether these are due to chance fluctuations or errors of sampling procedure or it is real relationship. This is done with the help of table area under NPC.

9. NPC is used to generalize about population from the sample:

We compute standard error of mean, standard error of standard deviation and other statistics to generalize about the population from which the sample are drawn. For this computation we use the table area under NPC.

4.3 Summary

A normal distribution, sometimes called the bell curve (or De Moivre distribution [1]), is a distribution that occurs naturally in many situations. For example, the bell curve is seen in tests like the SAT and GRE. The bulk of students will score the average (C), while smaller numbers of

students will score a B or D. An even smaller percentage of students score an F or an A. This creates a distribution that resembles a bell (hence the nickname). The bell curve is symmetrical. Half of the data will fall to the left of the mean; half will fall to the right.

4.4 Keywords/Glossary

NPC, statistics, normal distribution,Computation of Normal Probability Curve

4.5 Self-Assessment/Evaluation

1. The purpose of inferential statistics is to simplify and organize the data from a study. (True/False)
2. Individual differences can be expected for psychological variables such as intelligence, anxiety, and athletic ability. (True/False)
3. Frequency distributions are a subset of inferential statistics. (True/False)
4. Summary statistics are a subset of descriptive statistics. (True/False)
5. Mean scores are helpful in the interpretation of nominal data. (True/False)
6. Cross-tabulation is a useful way to describe the relationship between two nominal variables. (True/False)
7. It is possible to convert a grouped frequency distribution to a frequency distribution without referring back to the original data. (True/False)
8. It is possible to convert a frequency distribution to a grouped frequency distribution without referring back to the original data. (True/False)
9. When you are using a continuous variable, it is best to use a frequency distribution instead of a grouped frequency distribution. (True/False)
10. A frequency polygon can be used to compare distributions of scores from more than one group. (True/False)
11. Traditionally, the frequency of each score (or interval) is shown on the *Y* axis or ordinate of a frequency polygon. (True/False)
12. Skewed distributions usually show the familiar bell shape. (True/False)
13. The normal distribution is actually defined by a mathematical equation. (True/False)
14. A classroom test that is too difficult will produce a distribution that is negatively skewed. (True/False)
15. The range is a frequently used measure of central tendency. (True/False)

4.6 Review Questions

- What's NPC? Discuss its importance
- Discuss uses of NPC?
- Discuss applications of NPC

4.7 Further/Suggested Readings

Arkin, H. and Coltan, R. (1950), *Tables for Statistician*. Vamis & Novel Inc.: New York.

Cohen, J. (1977), *Statistical Power Analysis for the Behavioural Sciences*. Academic Press: New York.

Downie, N.M. and Heath, R.W. (1970), *Basic Statistical Methods*. Harper and Row Publishers: New York.

Fallix, F. and Brown, B. Bruce (1983), *Statistics for Behavioural Sciences*. The Dorsey Press: Illinois.

https://www.stat.uci.edu/what-is-statistics/

https://www.abs.gov.au/websitedbs/D3310114.nsf/home/statistical+language++what+are+variables#:~:text=A%20variable%20is%20any%20characteristics,type%20are%20examples%20of%20variables.

UNIT 05: Measures of Central tendency

Content

Objectives/Expected Learning Outcomes

Understand Measures of Central tendency

Importance of Measures of Central tendency

Measures of Central tendency and its importance

Understand Mean, mode, median

Introduction

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics. The mean (often called the average) is most likely the measure of central tendency that you are most familiar with, but there are others, such as the median and the mode.

The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others. In the following sections, we will look at the mean, mode and median, and learn how to calculate them and under what conditions they are most appropriate to be used.

5.1.Mean (Arithmetic)

The mean (or average) is the most popular and well known measure of central tendency. It can be used with both discrete and continuous data, although its use is most often with continuous data (see our Types of Variable guide for data types). The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. So, if we have

values in a data set and they have values …, the sample mean, usually denoted by

(pronounced "x bar"), is:

This formula is usually written in a slightly different manner using the Greek capitol letter,

, pronounced "sigma", which means "sum of...":

You may have noticed that the above formula refers to the sample mean. So, why have we called it a sample mean? This is because, in statistics, samples and populations have very different meanings and these differences are very important, even if, in the case of the mean, they are calculated in the same way. To acknowledge that we are calculating the population mean and not the sample mean, we use the Greek lower case letter "mu", denoted as

:

The mean is essentially a model of your data set. It is the value that is most common. You will notice, however, that the mean is not often one of the actual values that you have observed in your data set. However, one of its important properties is that it minimises error in the prediction of any one value in your data set. That is, it is the value that produces the lowest amount of error from all other values in the data set.

An important property of the mean is that it includes every value in your data set as part of the calculation. In addition, the mean is the only measure of central tendency where the sum of the deviations of each value from the mean is always zero.

5.2.When not to use the mean

The mean has one main disadvantage: it is particularly susceptible to the influence of outliers. These are values that are unusual compared to the rest of the data set by being especially small or large in numerical value. For example, consider the wages of staff at a factory below:

Staff	1	2	3	4	5	6	7	8	9	10
Salary	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k

The mean salary for these ten staff is $30.7k. However, inspecting the raw data suggests that this mean value might not be the best way to accurately reflect the typical salary of a worker, as most workers have salaries in the $12k to 18k range. The mean is being skewed by the two large salaries. Therefore, in this situation, we would like to have a better measure of central tendency. As we will find out later, taking the median would be a better measure of central tendency in this situation.

Another time when we usually prefer the median over the mean (or mode) is when our data is skewed (i.e., the frequency distribution for our data is skewed). If we consider the normal distribution - as this is the most frequently assessed in statistics - when the data is perfectly normal, the mean, median and mode are identical. Moreover, they all represent the most typical value in the data set. However, as the data becomes skewed the mean loses its ability to provide the best central location for the data because the skewed data is dragging it away from the typical value. However,

the median best retains this position and is not as strongly influenced by the skewed values. This is explained in more detail in the skewed distribution section later in this guide.

5.3.Median

The median is the middle score for a set of data that has been arranged in order of magnitude. The median is less affected by outliers and skewed data. In order to calculate the median, suppose we have the data below:

65	55	89	56	35	14	56	55	87	45	92

We first need to rearrange that data into order of magnitude (smallest first):

14	35	45	55	55	**56**	56	65	87	89	92

Our median mark is the middle mark - in this case, 56 (highlighted in bold). It is the middle mark because there are 5 scores before it and 5 scores after it. This works fine when you have an odd number of scores, but what happens when you have an even number of scores? What if you had only 10 scores? Well, you simply have to take the middle two scores and average the result. So, if we look at the example below:

65	55	89	56	35	14	56	55	87	45

We again rearrange that data into order of magnitude (smallest first):

14	35	45	55	**55**	**56**	56	65	87	89

Only now we have to take the 5th and 6th score in our data set and average them to get a median of 55.5.

5.4.Mode

The mode is the most frequent score in our data set. On a histogram it represents the highest bar in a bar chart or histogram. You can, therefore, sometimes consider the mode as being the most popular option. An example of a mode is presented below:

Normally, the mode is used for categorical data where we wish to know which is the most common category, as illustrated below:

We can see above that the most common form of transport, in this particular data set, is the bus. However, one of the problems with the mode is that it is not unique, so it leaves us with problems when we have two or more values that share the highest frequency, such as below:

We are now stuck as to which mode best describes the central tendency of the data. This is particularly problematic when we have continuous data because we are more likely not to have any one value that is more frequent than the other. For example, consider measuring 30 peoples' weight (to the nearest 0.1 kg). How likely is it that we will find two or more people with **exactly** the same weight (e.g., 67.4 kg)? The answer, is probably very unlikely - many people might be close, but with

such a small sample (30 people) and a large range of possible weights, you are unlikely to find two people with exactly the same weight; that is, to the nearest 0.1 kg. This is why the mode is very rarely used with continuous data.

Another problem with the mode is that it will not provide us with a very good measure of central tendency when the most common mark is far away from the rest of the data in the data set, as depicted in the diagram below:

In the above diagram the mode has a value of 2. We can clearly see, however, that the mode is not representative of the data, which is mostly concentrated around the 20 to 30 value range. To use the mode to describe the central tendency of this data set would be misleading.

5.5. Skewed Distributions and the Mean and Median

We often test whether our data is normally distributed because this is a common assumption underlying many statistical tests. An example of a normally distributed set of data is presented below:

When you have a normally distributed sample you can legitimately use both the mean or the median as your measure of central tendency. In fact, in any symmetrical distribution the mean, median and mode are equal. However, in this situation, the mean is widely preferred as the best measure of central tendency because it is the measure that includes all the values in the data set for its calculation, and any change in any of the scores will affect the value of the mean. This is not the case with the median or mode.

However, when our data is skewed, for example, as with the right-skewed data set below:

We find that the mean is being dragged in the direct of the skew. In these situations, the median is generally considered to be the best representative of the central location of the data. The more skewed the distribution, the greater the difference between the median and mean, and the greater emphasis should be placed on using the median as opposed to the mean. A classic example of the above right-skewed distribution is income (salary), where higher-earners provide a false representation of the typical income if expressed as a mean and not a median.

If dealing with a normal distribution, and tests of normality show that the data is non-normal, it is customary to use the median instead of the mean. However, this is more a rule of thumb than a strict guideline. Sometimes, researchers wish to report the mean of a skewed distribution if the median and mean are not appreciably different (a subjective assessment), and if it allows easier comparisons to previous research to be made.

5.6.Summary of when to use the mean, median and mode

Please use the following summary table to know what the best measure of central tendency is with respect to the different types of variable.

Type of Variable	**Best measure of central tendency**
Nominal	Mode
Ordinal	Median

Interval/Ratio (not skewed)	Mean
Interval/Ratio (skewed)	Median

5.7.Summary

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics. The mean (often called the average) is most likely the measure of central tendency that you are most familiar with, but there are others, such as the median and the mode. The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others. In the following sections, we will look at the mean, mode and median, and learn how to calculate them and under what conditions they are most appropriate to be used.

5.8.Keywords/Glossary

Statistics, Mean, Mode, Median,

5.9.Self-Assessment/Evaluation

1. The mean is the score at the 50th percentile. (True/False)
2. Mean, median and mode can be the same for some data. (True/False)
3. The variance, unlike the range, uses all the scores in its computation. (True/False)
4. If all of the scores in a distribution are increased by exactly five points, the range will increase by five points. (True/False)
5. A product-moment correlation of -1.00 means that there is no linear relationship between the variables. (True/False)
6. With ordinal data, the appropriate correlation to use is the Spearman rank-order correlation. (True/False)
7. Cross-tabulation for nominal data is the conceptual equivalent of scatter plots for score data. (True/False)
8. A nonlinear relationship is best indexed with a product-moment correlation. (True/False)
9. A circular scatter plot usually indicates a moderately strong positive relationship between the variables. (True/False)
10. A sample is a subset of the people or objects in a population. (True/False)
11. The null hypothesis states that there are no individual differences within the groups. (True/False)
12. If we are using a *t*-test and we reject the null hypothesis, we can be sure that the population means are different. (True/False)
13. Alpha is the level of Type II error that we can expect. (True/False)

14. If nothing else changes, decreasing alpha from .05 to .01 will increase the probability of a Type II error. (True/False)
15. If we have two groups and want to compare the means of the groups, we can use either a *t*-test or an analysis of variance. (True/False)
16. The sensitivity of a statistical procedure to the differences being sought is called the reliability of the procedure. (True/False)

5.10.Review Questions

- What is the mean of the following numbers? 1, 2, 3, 5, 5, 5, 7, 9, 11, 12
- What is the median of the following numbers? 1, 2, 3, 5, 5, 5, 7, 9, 11, 12
- What is the mode for the following numbers? 1, 2, 3, 5, 5, 5, 7, 9, 11, 12
- What is the range of the following numbers? 1, 2, 3, 5, 5, 5, 7, 9, 11, 12

Further/Suggested Readings

- Coltan, R. (1950), *Tables for Statistician*. Vamis& Novel Inc.: New York.
- Cohen, J. (1977), *Statistical Power Analysis for the Behavioural Sciences*. Academic Press: New York.
- Downie, N.M. and Heath, R.W. (1970), *Basic Statistical Methods*. Harper and Row Publishers: New York.
- Fallix, F. and Brown, B. Bruce (1983), *Statistics for Behavioural Sciences*. The Dorsey Press: Illinois.
- https://www.stat.uci.edu/what-is-statistics/
- https://www.abs.gov.au/websitedbs/D3310114.nsf/home/statistical+language++what+are+variables#:~:text=A%20variable%20is%20any%20characteristics,type%20are%20examples%20of%20variables.

Unit6: Measures of Dispersion

Content

Objectives

Understand Standard deviation,

Importance of Quartile deviation

Uses of Measures of Dispersion

Understand the importance of Measures of Dispersion

Introduction

Dispersion is referred to as the spread or dispersion of the data. Statistical dispersion means the extent to which a numerical data is likely to vary about an average value. In other words, dispersion helps to understand the distribution of the data.

In statistics, the measures of dispersion help to interpret the variability of data i.e. to know how much homogenous or heterogeneous the data is. In simple terms, it shows how squeezed or scattered the variable is.

There are two main types of dispersion methods in statistics which are:

- Absolute Measure of Dispersion
- Relative Measure of Dispersion

6.1. Standard Deviation

The standard deviation is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance. The standard deviation is calculated as the square root of variance by determining each data point's deviation relative to the mean.

If the data points are further from the mean, there is a higher deviation within the data set; thus, the more spread out the data, the higher the standard deviation.As a downside, the standard deviation calculates all uncertainty as risk, even when it's in the investor's favour—such as above-average returns.

Formula of standard deviation =

$n-1\sum i=1n(xi-x)2/$ n-1

where:xi=Value of the *ith* point in the data set

x= The mean of the value set

n=number of the points in data s

Calculating the Standard Deviation

Standard deviation is calculated as follows:

1. The mean value is calculated by adding all the data points and dividing them by the number of data points.
2. The variance for each data point is calculated by subtracting the mean from the value of the data point. Each of those resulting values is then squared and the results summed. The result is then divided by the number of data points less one.
3. The square root of the variance—result from no. 2—is then used to find the standard deviation.

Using the Standard Deviation

Standard deviation is an especially useful tool in investing and trading strategies as it helps measure market and security volatility—and predict performance trends. As it relates to investing, for example, an index fund is likely to have a low standard deviation versus its benchmark index, as the fund's goal is to replicate the index.

On the other hand, one can expect aggressive growth funds to have a high standard deviation from relative stock indices, as their portfolio managers make aggressive bets to generate higher-than-average returns.

A lower standard deviation isn't necessarily preferable. It all depends on the investments and the investor's willingness to assume risk. When dealing with the amount of deviation in their portfolios, investors should consider their tolerance for volatility and their overall investment objectives. More aggressive investors may be comfortable with an investment strategy that opts for vehicles with higher-than-average volatility, while more conservative investors may not.

Standard deviation is one of the key fundamental risk measures that analysts, portfolio managers, advisors use. Investment firms report the standard deviation of their mutual funds and other products. A large dispersion shows how much the return on the fund is deviating from the expected normal returns. Because it is easy to understand, this statistic is regularly reported to the end clients and investors.

Example of Standard Deviation

Say we have the data points 5, 7, 3, and 7, which total 22. You would then divide 22 by the number of data points, in this case, four—resulting in a mean of 5.5. This leads to the following determinations: $\bar{x} = 5.5$ and N = 4.

The variance is determined by subtracting the mean's value from each data point, resulting in -0.5, 1.5, -2.5, and 1.5. Each of those values is then squared, resulting in 0.25, 2.25, 6.25, and 2.25. The square values are then added together, giving a total of 11, which is then divided by the value of N minus 1, which is 3, resulting in a variance of approximately 3.67.

The square root of the variance is then calculated, which results in a standard deviation measure of approximately 1.915.

6.2. Quartile Deviation

The Quartile Deviation can be defined mathematically as half of the difference between the upper and lower quartile. Here, quartile deviation can be represented as QD; Q_3 denotes the upper quartile and Q_1 indicates the lower quartile.Quartile Deviation is also known as the Semi Interquartile range.

Quartile Deviation Formula

Suppose Q_1 is the lower quartile, Q_2 is the median, and Q_3 is the upper quartile for the given data set, then its quartile deviation can be calculated using the following formula.

$QD = (Q_3 - Q_1)/2$

Quartile Deviation for Ungrouped Data

For an ungrouped data, quartiles can be obtained using the following formulas,

Q_1 = [(n+1)/4]th item

Q_2 = [(n+1)/2]th item

Q_3 = [3(n+1)/4]th item

Where n represents the total number of observations in the given data set.

Also, Q_2 is the median of the given data set, Q_1 is the median of the lower half of the data set and Q_3 is the median of the upper half of the data set.

Before, estimating the quartiles, we have to arrange the given data values in ascending order. If the value of n is even, we can follow the similar procedure of finding the median.

Quartile Deviation Example

Find the quartiles and quartile deviation of the following data:

17, 2, 7, 27, 15, 5, 14, 8, 10, 24, 48, 10, 8, 7, 18, 28

Solution:

Given data:

17, 2, 7, 27, 15, 5, 14, 8, 10, 24, 48, 10, 8, 7, 18, 28

Ascending order of the given data is:

2, 5, 7, 7, 8, 8, 10, 10, 14, 15, 17, 18, 24, 27, 28, 48

Number of data values = n = 16

Q_2 = Median of the given data set

n is even, median = (1/2) [(n/2)th observation and (n/2 + 1)th observation]

= (1/2)[8th observation + 9th observation]

= (10 + 14)/2

= 24/2

= 12

Q_2 = 12

Now, lower half of the data is:

2, 5, 7, 7, 8, 8, 10, 10 (even number of observations)

Q_1 = Median of lower half of the data

= (1/2)[4th observation + 5th observation]

= (7 + 8)/2

= 15/2

= 7.5

Also, the upper half of the data is:

14, 15, 17, 18, 24, 27, 28, 48 (even number of observations)

Q_3= Median of upper half of the data

= (1/2)[4th observation + 5th observation]

= (18 + 24)/2

= 42/2

= 21

Quartile deviation = $(Q_3 - Q_1)/2$

= (21 – 7.5)/2

= 13.5/2

= 6.75

Therefore, the quartile deviation for the given data set is 6.75

Uses of Quartile Deviation

The quartile deviation helps to examine the spread of a distribution about a measure of its central tendency, usually the mean or the average. Hence, it is in use to give you an idea about the range within which the central 50% of your sample data lies.

6.3.Range

The range in statistics for a given data set is the difference between the highest and lowest values. For example, if the given data set is {2,5,8,10,3}, then the range will be 10 – 2 = 8.

Thus, the range could also be defined as the difference between the highest observation and lowest observation. The obtained result is called the range of observation. The range in statistics represents the spread of observations.

Range Formula

The formula of the range in statistics, can simply be given by the difference between highest and lowest value.

Range = Highest Value – Lowest Value

Or

Range = Highest observation – Lowest observation

Or

Range = Maximum value – Minimum Value

Example of Range

Find the range of given observations: 32, 41, 28, 54, 35, 26, 23, 33, 38, 40.

Solution: Let us first arrange the given values in ascending order.

23, 26, 28, 32, 33, 35, 38, 40, 41, 54

Since, 23 is the lowest value and 54 is the highest value, therefore, the range of the observations will be;

Range (X) = Max (X) – Min (X)

= 54 – 23

= 31

Hence, is the required answer.

Uses of the Range

The range generally gives you a good indicator of variability when you have a distribution without extreme values. When paired with measures of central tendency, the range can tell you about the span of the distribution. But the range can be misleading when you have outliers in your data set.

6.4.Percentile

A percentile (or a centile) is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall. For example, the 20th percentile is the value (or score) below which 20% of the observations may be found.

The term percentile and the related term percentile rank are often used in the reporting of scores from norm-referenced tests. For example, if a score is at the 86th percentile, where 86 is the percentile rank, it is equal to the value below which 86% of the observations may be found. In contrast, if it is in the 86th percentile, the score is at or below the value of which 86% of the observations may be found. Every score is in the 100th percentile.
The 25th percentile is also known as the first quartile (Q1), the 50th percentile as the median or second quartile (Q2), and the 75th percentile as the third quartile (Q3). In general, percentiles and quartiles are specific types of quantiles.

6.5.Summary

Interquartile range is defined as the difference between the 25th and 75th percentile (also called the first and third quartile). Hence the interquartile range describes the middle 50% of observations. If the interquartile range is large it means that the middle 50% of observations are spaced wide apart. The important advantage of interquartile range is that it can be used as a measure of variability if the extreme values are not being recorded exactly (as in case of open-ended class intervals in the frequency distribution). Other advantageous feature is that it is not affected by extreme values. The main disadvantage in using interquartile range as a measure of dispersion is that it is not amenable to mathematical manipulation.

6.6. Keywords

Standard deviation, Quartile deviation, Range, Percentile

6.7 Self-assessment

1. Correlation refers to the dispersion of data.
    a. True
    b. False
2. There are two main types of dispersion
    a. True
    b. False
3. Correlation can explain the extent of relationship between two variables
    a. True
    b. False
4. Dispersion is the measurement of variability in a data
    a. True
    b. False
5. 25th percentile is known as q2.
    a. True
    b. False
6. If you know the percentile you can understand quartile of a data
    a. True
    b. False
7. Standard deviation refers to the deviation from mean
    a. True
    b. False
8. Pearson coefficient cannot explain the strength of linear relationship between two variables
    a. True
    b. False

9. Spearman's rank order correlation is a non-parametric method of Pearson correlation coefficient
    a. True
    b. False
10. Percentile is often used to determine the variability in norm referenced test.
    a. True
    b. False
11. We can use range if there are extreme values
    a. True
    b. False
12. Median is important while calculating quartile deviation
    a. True
    b. False
13. An increase in one variable results in the decrease in other variable is positive correlation
    a. True
    b. False
14. Spearman's correlation is powerful to determine the monotonic relationship
    a. True
    b. False
15. Cause- effect relationship can be explained using correlation
    a. True
    b. False

6.8. Review questions

16. Explain measures of dispersion
17. what is range?
18. Explain the relation between quartile deviations and percentile

Further/Suggested Readings

-  Arkin, H. and Coltan, R. (1950), *Tables for Statistician*. Vamis& Novel Inc.: New York.
- Cohen, J. (1977), *Statistical Power Analysis for the Behavioural Sciences*. Academic Press: New York.
- Downie, N.M. and Heath, R.W. (1970), *Basic Statistical Methods*. Harper and Row Publishers: New York.
- Fallix, F. and Brown, B. Bruce (1983), *Statistics for Behavioural Sciences*. The Dorsey Press: Illinois.

- https://www.stat.uci.edu/what-is-statistics/
- https://www.abs.gov.au/websitedbs/D3310114.nsf/home/statistical+language++what+are+variables#:~:text=A%20variable%20is%20any%20characteristics,type%20are%20examples%20of%20variables.

Unit7: Relationship between variables

Content

Objectives

Understand correlation

Importance of correlation

Uses of Measures of Dispersion

Understand the importance of types of correlation

Introduction

Correlation is a statistical term describing the degree to which two variables move in coordination with one another. If the two variables move in the same direction, then those variables are said to have a positive correlation. If they move in opposite directions, then they have a negative correlation.

Correlation coefficients are indicators of the strength of the linear relationship between two different variables, x and y. A linear correlation coefficient that is greater than zero indicates a positive relationship. A value that is less than zero signifies a negative relationship. Finally, a value of zero indicates no relationship between the two variables x and y.

7.1. Relationship between variables

The statistical relationship between two variables is referred to as their correlation. A correlation could be positive, meaning both variables move in the same direction, or negative, meaning that when one variable's value increases, the other variables' values decrease.

7.2. Pearson's Product Moment Correlation

The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by *r*. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

The Pearson correlation coefficient, *r*, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

The stronger the association of the two variables, the closer the Pearson correlation coefficient, *r*, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. Achieving a value of +1 or -1 means that all your data points are included on the line of best fit – there are no data points that show any variation away from this line. Values for *r* between +1 and -1 (for example*r* = 0.8 or -0.4) indicate that there is variation around the line of best fit. The closer the value of *r* to 0 the greater the variation around the line of best fit.

Pearson Correlation Coefficient Formula

where cov is the covariance and (cov(X,Y)= $\sum Ni=1(Xi-\bar{}X)(Yi-\bar{}Y)N \sum i = 1 N (X i - X \bar{}) (Y i - Y \bar{}$) N , σ_X is standard deviation of X and σ_Y is standard deviation of Y. Given X and Y are two random variables.

7.3. Spearman's Rank Order Correlation

The Spearman's rank-order correlation is the nonparametric version of the Pearson product-moment correlation. Spearman's correlation coefficient, (ρ, also signified by r_s) measures the strength and direction of association between two ranked variables. In order to administer Spearman's Rank Order Correlation, you need wo variables that are either ordinal, interval or ratio (see our Types of Variable guide if you need clarification).

Although you would normally hope to use a Pearson product-moment correlation on interval or ratio data, the Spearman correlation can be used when the assumptions of the Pearson correlation are markedly violated. However, Spearman's correlation determines the strength and direction of the monotonic relationship between your two variables rather than the strength and direction of the linear relationship between your two variables, which is what Pearson's correlation determines.

A monotonic relationship is a relationship that does one of the following: (1) as the value of one variable increases, so does the value of the other variable; or (2) as the value of one variable increases, the other variable value decreases.Spearman's correlation measures the strength and direction of monotonic association between two variables. Monotonicity is "less restrictive" than that of a linear relationship. For example, the middle image above shows a relationship that is monotonic, but not linear.

Spearman's Rank Order Correlation formula:

$$R_s = 1 - \left(\frac{6 \Sigma d^2}{n^3 - n} \right)$$

7.4. Limitations of Correlation

1.Correlation can't look at the presence or effect of other variables outside of the two being explored. Importantly, correlation doesn't tell us about cause and effect.

2. Correlation also cannot accurately describe curvilinear relationships.

3. It won't determine what variables have the most influence.

7.5. Summary

Interquartile range is defined as the difference between the 25th and 75th percentile (also called the first and third quartile). Hence the interquartile range describes the middle 50% of observations. If the interquartile range is large it means that the middle 50% of observations are spaced wide apart. The important advantage of interquartile range is that it can be used as a measure of variability if the extreme values are not being recorded exactly (as in case of open-ended class intervals in the frequency distribution).[2] Other advantageous feature is that it is not affected by extreme values. The main disadvantage in using interquartile range as a measure of dispersion is that it is not amenable to mathematical manipulation."Correlation is not causation" means that just because two variables are related it does not necessarily mean that one causes the other.

A correlation identifies variables and looks for a relationship between them. An experiment tests the effect that an independent variable has upon a dependent variable but a correlation looks for a relationship between two variables.

This means that the experiment can predict cause and effect (causation) but a correlation can only predict a relationship, as another extraneous variable may be involved that it not known about.

7.6. Keywords

Correlation, Pearson Product Moment, Spearman Rank Order

7.7. Self-assessment

1. Pearson coefficient cannot explain the strength of linear relationship between two variables
    a. True
    b. False

2. Spearman's rank order correlation is a non-parametric method of Pearson correlation coefficient
    a. True
    b. False
3. Percentile is often used to determine the variability in norm referenced test.
    a. True
    b. False
4. We can use range if there are extreme values
    a. True
    b. False

5. Median is important while calculating quartile deviation
    a. True
    b. False
6. An increase in one variable results in the decrease in other variable is positive correlation
    a. True
    b. False
7. Spearman's correlation is powerful to determine the monotonic relationship
    a. True
    b. False
8. Cause- effect relationship can be explained using correlation
    a. True
    b. False
9. Correlation refers to the dispersion of data.
    a. True
    b. False
10. There are two main types of dispersion
    a. True
    b. False
11. Correlation can explain the extent of relationship between two variables
    a. True
    b. False
12. Dispersion is the measurement of variability in a data
    a. True
    b. False
13. 25^{th} percentile is known as q2.
    a. True
    b. False
14. If you know the percentile you can understand quartile of a data
    a. True
    b. False
15. Standard deviation refers to the deviation from mean
    a. True
    b. False

7.8. Review questions

16. Explain measures of dispersion
17. what is range?
18. Explain the relation between quartile deviations and percentile
19. What are the limitations of correlation?
20. Differentiate between Spearman's correlation and Pearson's correlation.

Further/Suggested Readings

- Books

- Arkin, H. and Coltan, R. (1950), *Tables for Statistician*. Vamis& Novel Inc.: New York.
- Cohen, J. (1977), *Statistical Power Analysis for the Behavioural Sciences*. Academic Press: New York.
- Downie, N.M. and Heath, R.W. (1970), *Basic Statistical Methods*. Harper and Row Publishers: New York.
- Fallix, F. and Brown, B. Bruce (1983), *Statistics for Behavioural Sciences*. The Dorsey Press: Illinois.

- https://www.stat.uci.edu/what-is-statistics/
- https://www.abs.gov.au/websitedbs/D3310114.nsf/home/statistical+language++what+are+variables#:~:text=A%20variable%20is%20any%20characteristics,type%20are%20examples%20of%20variables.

UNIT 8 – Hypothesis

Contents

Objectives

This unit will enable you to understand

Meaning and nature of hypothesis

Uses of hypotheses

Types of hypotheses

Functions of Hypotheses

Introduction

A hypothesis is a concept or idea that you test through research and experiments. In other words, it is a prediction that is can be tested by research. Most researchers come up with a hypothesis statement at the beginning of the study. Thus basically, you make a prediction about the outcome at the start of the study and conduct experiments to test whether this predictipon is true and to what extent.

8.1.Meaning

We know that research begins with a problem or a felt need or difficulty. The purpose of research is to find a solution to the difficulty. It is desirable that the researcher should propose a set of suggested solutions or explanations of the difficulty which the research proposes to solve. Such tentative solutions formulated as a proposition are called hypotheses. To understand the meaning

of a hypothesis, let us see some definitions: "A hypothesis is a tentative generalization, the validity of which remains to be tested. In its most elementary stage the hypothesis may be any guess, hunch, imaginative idea, which becomes the basis for action or investigation". (G.A.Lundberg)

"It is a proposition which can be put to test to determine validity". (Goode and Hatt). "A hypothesis is a question put in such a way that an answer of some kind can be forth coming" - (Rummel and Ballaine). These definitions lead us to conclude that a hypothesis is a tentative solution or explanation or a guess or assumption or a proposition or a statement to the problem facing the researcher, adopted on a cursory observation of known and available data, as a basis of investigation, whose validity is to be tested or verified.

8.2.Nature of Hypothesis

A hypothesis controls and directs the research study. When a problem is felt, we require the hypothesis to explain it. Generally, there is more than one hypothesis which aims at explaining the same fact. But all of them cannot be equally good. Therefore, how can we judge a hypothesis to be true or false, good or bad? Agreement with facts is the sole and sufficient test of a true hypothesis. Therefore, certain conditions can be laid down for distinguishing a good hypothesis from bad ones. The formal conditions laid down by thinkers provide the criteria for judging a hypothesis as good or valid. These conditions are as follows:

i) A hypothesis should be empirically verifiable: The most important condition for a valid hypothesis is that it should be empirically verifiable. A hypothesis is said to be verifiable, if it can be shown to be either true or false by comparing with the facts of experience directly or indirectly. A hypothesis is true if it conforms to facts and it is false if it does not. Empirical verification is the characteristic of the scientific method. ii) A hypothesis should be relevant: The purpose of formulating a hypothesis is always to explain some facts. It must provide an answer to the problem which initiated the enquiry. A hypothesis is called relevant if it can explain the facts of enquiry.

iii) A hypothesis must have predictive and explanatory power: Explanatory power means that a good hypothesis, over and above the facts it proposes to explain, must also explain some other facts which are beyond its original scope. We must be able to deduce a wide range of observable facts which can be deduced from a hypothesis. The wider the range, the greater is its explanatory power.

iv) A hypothesis must furnish a base for deductive inference on consequences: In the process of investigation, we always pass from the known to the unknown. It is impossible to infer any thing from the absolutely unknown. We can only infer what would happen under supposed conditions by applying the knowledge of nature we possess. Hence, our hypothesis must be in accordance with our previous knowledge.

v) A hypothesis does not go against the traditionally established knowledge: As far as possible, a new hypothesis should not go against any previously established law or knowledge. The new hypothesis is expected to be consistent with the established knowledge.

8.3.Functions of hypothesis testing

If a clear scientific hypothesis has been formulated, half of the research work is already done. The advantages/utility of having a hypothesis are summarized here underneath:

i) It is a starting point for many a research work.

ii) It helps in deciding the direction in which to proceed.

iii) It helps in selecting and collecting pertinent facts.

iv) It is an aid to explanation.

v) It helps in drawing specific conclusions.

vi) It helps in testing theories.

vii) It works as a basis for future knowledge.

Give real life examples of Hypothesis testing

8.3Types of Hypothesis

Hypotheses can be classified in a variety of ways into different types or kinds. The following are some of the types of hypotheses:

i) **Explanatory Hypothesis:** The purpose of this hypothesis is to explain a certain fact. All hypotheses are in a way explanatory for a hypothesis is advanced only when we try to explain the observed fact. A large number of hypotheses are advanced to explain the individual facts in life. A theft, a murder, an accident are examples.

ii) **Descriptive Hypothesis:** Some times a researcher comes across a complex phenomenon. He/ she does not understand the relations among the observed facts. But how to account for these facts? The answer is a descriptive hypothesis. A hypothesis is descriptive when it is based upon the points of resemblance of some thing. It describes the **cause** and **effect** relationship of a phenomenon e.g., the current unemployment rate of a state exceeds 25% of the work force. Similarly, the consumers of local made products constitute a significant market segment.

iii) **Analogical Hypothesis:** When we formulate a hypothesis on the basis of similarities (analogy), it is called an analogical hypothesis e.g., families with higher earnings invest more surplus income on long term investments.

iv) **Working hypothesis:** Sometimes certain facts cannot be explained adequately by existing hypotheses, and no new hypothesis comes up. Thus, the investigation is held up. In this situation, a researcher formulates a hypothesis which enables to continue investigation. Such a hypothesis, though inadequate and formulated for the purpose of further investigation only, is called a working hypothesis. It is simply accepted as a starting point in the process of investigation.

v) **Null Hypothesis:** It is an important concept that is used widely in the sampling theory. It forms the basis of many tests of significance. Under this type, the hypothesis is stated negatively. It is null because it may be nullified, if the evidence of a random sample is unfavourable to the hypothesis. It is a hypothesis being tested (H0). If the calculated value of the test is less than the permissible value, Null hypothesis is accepted, otherwise it is rejected. The rejection of a null hypothesis implies that the difference could not have arisen due to chance or sampling fluctuations.

vi)**Statistical Hypothesis:** Statistical hypotheses are the statements derived from a sample. These are quantitative in nature and are numerically measurable. For example, the market share of product X is 70%, the average life of a tube light is 2000 hours etc.

8.5. Summary

A hypothesis (plural hypotheses) is a precise, testable statement of what the researcher(s) predict will be the outcome of the study. It is stated at the start of the study. This usually involves proposing a possible relationship between two variables: the independent variable (what the researcher changes) and the dependent variable (what the research measures). In research, there is a convention that the hypothesis is written in two forms, the null hypothesis, and the alternative hypothesis (called the experimental hypothesis when the method of investigation is an experiment).

8.6 Key Words

Working or Research Hypothesis: A research hypothesis is a specific, clear prediction about the possible outcome of a scientific research study based on specific factors of the population.

Null Hypothesis: A null hypothesis is a general statement which states no relationship between two variables or two phenomena. It is usually **denoted by H_0.**

Alternative Hypothesis: An alternative hypothesis is a statement which states some statistical significance between two phenomena. It is usually **denoted by H_1 or H_A.**

8.7. Self-Assessment

1 -An alternative hypothesis is a statement which states some statistical significance between two phenomena (T/F)

2 -What is true about hypotheses, Hypotheses is/ are

A-Precise

B- Testable

C- Both

D-none

3- A hypothesis must have ----------& ---------- power

A-Predictive

B-Explanatory

C- Both

D-none

4- Hypotheses help in selecting and collecting pertinent facts(T/F)

5- A hypothesis is aabsolute generalization, the validity of which remains to be theorized (T/ F)

6- A----------works as a basis for future knowledge.

7- hypothesis should be relevant enough (T/F)

8- A hypothesis is a question put in such a way that an answer of some kindcan be forth coming (T/F)

9- hypothesis is a tentative solution orexplanation or a guess or assumption or a proposition or a statement to theproblem facing the researcher (T/F)

10- Which among these is/ are type of hypothesis/es

A-Null

B-Alternative

C- Both

D-none

11- Null Hypotheses is always hypothesis of vital differences (T/F)

12- We formulate a ---------- hypothesis on the basis of similarities (analogy),

13- The purpose of ----------hypothesis is to explain a certain fact

14- Hypotheses help in drawing specific conclusions. (T/F)

15- Hypotheses usually weakens conclusion drawing tendency (T/F)

Answers				
1	2	3	4	5
T	C	C	T	F
6	7	8	9	10
Hypotheses	T	T	T	C
11	12	13	14	15
F	Analogical	Explanatory	T	F

8.8. Review Questions

1. Discus the various types hypotheses.
2. How does social research need Hypotheses
3. What are various Functions of Hypotheses
4. What role do Null Hypotheses play is scientific research

Further Readings

 Kazdin, A. E. (2003). Research design in clinical psychology(4th ed.). Boston: Allyn and Bacon.

UNIT 9- Hypothesis testing

Contents

Objectives

Identify the four steps of hypothesis testing.

Define null hypothesis, alternative hypothesis,level of significance, test statistic, p value, and statistical significance.

Define Type I error and Type II error, and identify the type of error that researchers control.

Introduction

We use inferential statistics because it allows us to measure behavior in samples to learn more about the behavior in populations that are often too large or inaccessible. We use samples because we know how they are related to populations. For example, suppose the average score on a standardized exam in a given population is 1,000. if we selected a random sample from a population, then on average the value of the sample mean will equal the population mean. In our example, if we select a random sample from this population with a mean of 1,000, then on average, the value of a sample mean will equal 1,000. On the basis of the central limit theorem, we know that the probability of selecting any other sample mean value from this population is normally distributed.

In behavioral research, we select samples to learn more about populations of interest to us. In terms of the mean, we measure a sample mean to learn more about the mean in a population. Therefore, we will use the sample mean to describe the population mean. We begin by stating the value of a population mean, and then we select a sample and measure the mean in that sample. On average, the value of the sample mean will equal the population mean. The larger the difference or discrepancy between the sample mean and population mean, the less likely it is that we could have selected that sample mean, if the value of the population mean is correct. This type of experimental situation, using the example of standardized exam scores,

9.1. Testing of Hypothesis

When the hypothesis has been framed in the research study, it must be verified as true or false. Verifiability is one of the important conditions of a goodhypothesis.Verification of hypothesis means testing of the truth of the hypothesis in the light of facts. If the hypothesis agrees with the facts, it is said to be true and may be accepted as the explanation of the facts. But if it doesnot agree it is said to be false. Such a false hypothesis is either totally rejectedor modified. Verification is of two types viz., Direct verification and Indirectverification.

Direct verification may be either by observation or by experiments. When direct observation shows that the supposed cause exists where it was thought to exist, we have a direct verification. When a hypothesis is verified by an experiment in a laboratory it is called direct verification by experiment. Whenthe hypothesis is not amenable for direct verification, we have to depend on

9.2.Standard Error

Standard error is a measure of the average, or standard, distance between a sample statistic and the corresponding population parameter. The advantage of computing the standard error is that it provides a measure of how much difference it is reasonable to expect between a statistic and a parameter. Notice that this distance is a measure of the natural discrepancy that occurs just by chance. Samples are intended to represent their populations but they are not expected to be perfect. Typically, there is some discrepancy between a sample statistic and the population parameter, and the standard error tells you how much discrepancy to expect.

9.3. Level of significance

The alpha level, or level of significance, for a hypothesis test is the maximum probability that the research result was obtained simply by chance. A hypothesis test with an alpha level of .01, for example, means that the test demands that there is less than a 1% (.01) probability that the results are caused only by chance. The alpha level provides a criterion for interpreting the test statistic. As we noted earlier, a test statistic with a value greater than 1.00 usually indicates that the obtained result is greater than would be expected from chance. However, researchers typically demand research results that are not just greater than chance but significantly greater than chance. The alpha level provides a criterion for significance. Remember, the goal of a hypothesis test is to rule out chance as a plausible explanation for the results. To achieve this, researchers determine which results are reasonable to expect just by chance (without any treatment effect), and which results are extremely unlikely to be obtained by chance alone. The alpha level is a probability value that defines what is extremely unlikely. That is, the alpha level is the probability that the sample results would be obtained even if the null hypothesis were true. By convention, alpha levels are very small probabilities, usually .05, .01, or .001. An alpha level of .01, for example, means that a sample result is considered to be extremely unlikely to occur by chance (without any treatment effect) if it has a probability that is less than .01.Such a sample results in rejection of the null hypothesis and the conclusion thata real treatment effect does exist.

9.4 Confidence interval

Confidence interval is a technique for estimating the magnitude of an unknown population value such as a mean difference or a correlation. The logic behind a confidence interval is that a sample statistic should provide a reasonably accurate estimate of the corresponding population parameter.Therefore, the value of the parameter should be located in an interval, or arange of values, centered around the sample statistic.

The level of significance in hypothesis testing is the criterion we use to decide whether the value stated in the null hypothesis is likely to be true.

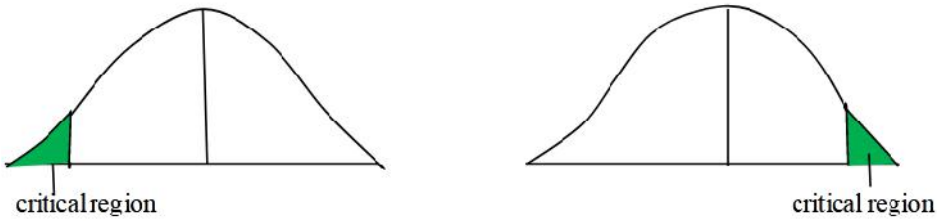
9.5 t-test

In many research situations, the data are numerical scores, so it is possible to compute sample means. All of the hypothesis tests covered in this section use the means obtained from sample data as the basis for testing hypotheses about population means. The goal of each test is to determine whether the observed sample mean differences are more than would be expected by chance alone; that is, if the sample data provide enough evidence for the conclusion that some factor other than chance (for example, a treatment effect) has caused the means to be different. In many research situations, the data are numerical scores, so it is possible to compute sample means. All of the hypothesis tests covered in this section use the means obtained from sample data as the basis for testing hypotheses about population means. The goal of each test is to determine whether the observed sample mean differences are more than would be expected by chance alone; that is, if the sample data provide enough evidence for the conclusion that some factor other than chance (for example, a treatment effect) has caused the means to be different.

9.6 One Tailed Versus Two Tailed tests

In Statistics hypothesistesting, we need to judge whether it is a one-tailed or a two-tailed test so that we can findthe critical values in tables such as StandardNormal z Distribution Table and t Distribution Table. And then, by comparingtest statistic value withthe critical value or whether statistic value falls in the critical region, we make a conclusion either to reject the null hypothesis or to fail to reject the null hypothesis. How can we tell whether it is a one-tailed or a two-tailedtest? It depends on the originalclaim in the question. A one-tailedtest looks for an "increase" or "decrease" in the parameter whereas a two-tailed test looks for a "change" (could be increase or decrease) in the parameter. Therefore, if we see words suchas "increased, greater, larger, improved andso on", or "decreased, less, smaller andso on" inthe original claim of a question (>,< are usedin H1), a one-tail test is applied. If words such as "change, the same, different/difference and soon" are used in the claim of the question($\neq$ is used in H1), a two-tailed test is applied.

In a one-tailed test, the critical regionhas just one part (the green area below). It can be a left- tailed test or a right-tailed test. Left-tailed test: The critical regionis in the extreme left region (tail)under the curve. Right-tailed test: The critical regionis in the extreme right region (tail) under the curve.



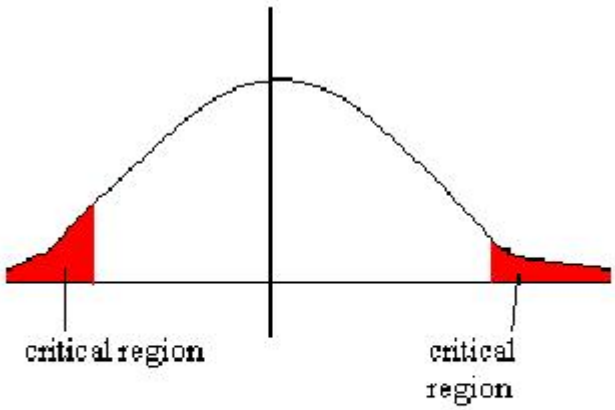
Left-tailed test

Sign used in H_1: <

Right-tailed test

Sign used in H_1: >

In two-tailed test, the critical region has two parts (thered areas below) which are in the two extreme gerions (tails) under thecurve.



Two-tailed test

Sign used in H_1: ≠

9.7 Errors in Hypothesis Testing

Because a hypothesis test is an inferential process (using limited information to reach a general conclusion), there is always a possibility that the process will lead to an error. Specifically, a sample always provides limited and incomplete information about its population. In addition, some samples are not good representatives of the population and can provide misleading information. If a researcher is misled by the results from the sample, it is likely that the researcher will reach an incorrect conclusion. Two kinds of errors can be made in hypothesis testing.

Type I Errors

One possibility for error occurs when the sample data appear to show a signify cant effect but, in fact, there is no effect in the population. By chance, the researcher has selected an unusual or extreme sample. Because the sample appears to how that the treatment has an effect, the researcher incorrectly concludes that there is a significant effect. This kind of mistake is called a Type I error.

Type II Errors

A Type II error, or beta (b) error, is the probability of incorrectly retaining the null hypothesis.

The second possibility for error occurs when the sample data do not show a significant effect when, in fact, there is a real effect in the population. This often occurs when an effect is very small and does not produce sample data that are sufficiently extreme to reject the null hypothesis. In this case, the researcher concludes that there is no significant effect when a real effect actually exists. This is a Type II error. The consequence of a Type II error is that a researcher fails to detect a real effect. Whenever research results do not show a significant effect, the researcher may choose to abandon the research project under the assumption that either there is no effect or the effect is too small to be of any consequence. On the other hand, the researcher may be convinced that an effect really exists but failed to show up in the current study. In this case, the researcher may choose to repeat the study, often using a larger sample, a stronger version of the treatment, or some other refinement that might increase the likelihood of obtaining a significant result

9.8. Summary

Hypothesis testing or significance testing is a method for testing a claim or hypothesis about a parameter in a population, using data measured in a sample. In this method, we test some hypothesis by determining the likelihood that a sample statistic could have been selected, if the hypothesis regarding the population parameter were true.

9.9. Key Words

Null hypothesis, which we presume is true.

Level of significance, or significance level, refers to a criterion of judgment upon which a decision is made regarding the value stated in a null hypothesis.

Test statistic is a mathematical formula that allows researchers to determine the likelihood of obtaining sample outcomes if the null hypothesis were true.

A ***p***value is the probability of obtaining a sample outcome, given that the value stated in the null hypothesis is true. The p value of a sample outcome is compared to the level of significance.

9.10. Self-Assessment

1- Null Hypotheses testing is always hypothesis of stating, there will be no significant difference (T/F)

2- We formulate a null hypothesis on the basis of review (T/F)

3- The purpose of hypothesis testing is to explain a certain fact in marketResearch strategy (T/F)

4- Hypotheses testing help us in drawing specific conclusions. (T/F)

5- Hypotheses testing usually weakens conclusion drawing tendency (T/F)

6- A statement made about a population for testing purpose is called?
a) Statistic
b) Hypothesis
c) Level of Significance
d) Test-Statistic

7. If the assumed hypothesis is tested for rejection considering it to be true is called?
a) Null Hypothesis
b) Statistical Hypothesis
c) Simple Hypothesis
d) Composite Hypothesis

8- A statement whose validity is tested on the basis of a sample is called?
a) Null Hypothesis
b) Statistical Hypothesis
c) Simple Hypothesis
d) Composite Hypothesis

9- A hypothesis which defines the population distribution is called?
a) Null Hypothesis
b) Statistical Hypothesis
c) Simple Hypothesis
d) Composite Hypothesis

10- If the null hypothesis is false then which of the following is accepted?
a) Null Hypothesis
b) Positive Hypothesis
c) Negative Hypothesis
d) Alternative Hypothesis.

11- The rejection probability of Null Hypothesis when it is true is called as?
a) Level of Confidence
b) Level of Significance
c) Level of Margin
d) Level of Rejection

12- The point where the Null Hypothesis gets rejected is called as?
a) Significant Value
b) Rejection Value
c) Acceptance Value
d) Critical Value

13- If the Critical region is evenly distributed then the test is referred as?
a) Two tailed
b) One tailed
c) Three tailed
d) Zero tailed

14- The type of test is defined by which of the following?
a) Null Hypothesis
b) Simple Hypothesis
c) Alternative Hypothesis
d) Composite Hypothesis

15- Consider a hypothesis H_0 where $\phi_0 = 5$ against H_1 where $\phi_1 > 5$. The test is?
a) Right tailed
b) Left tailed
c) Center tailed
d) Cross tailed

Answers				
1	2	3	4	5
T	T	T	T	F
6	7	8	9	10
B	A	B	C	D
11	12	13	14	15
B	D	A	C	A

9.11. Review Questions

1. State the four steps of hypothesis testing.

2. What are two decisions that a researcher makes in hypothesis testing?

3. What is a Type I error (a)?

4. What is a Type II error (b)?

5. What is the power in hypothesis testing?

6. What are the critical values for a one–independent sample nondirectional (two-tailed) z test at a .05 level of significance?

Further Readings

 Clover, Vernon T. &Balsley, Howard L. (1984). Business ResearchMethods, Grid, Inc: Columbus.

Ghosh, B.N. (1992). Scientific Methods and Social Research, Sterling Publishers Pvt. Ltd : New Delhi.

UNIT 10- Analysis of Variance

Contents

Objectives

After studying this unit you will be able to

Analyze the structure of data

Understand one way ANOVA

Two way ANOVA

F- Ratio

Introduction

An **ANOVA** test is a way to find out if survey or experiment results are significant. In other words, they help you to figure out if you need to reject the null hypothesis or accept the alternate hypothesis. Basically, **you're testing groups to see if there's a difference between them.** Examples of when you might want to test different groups:A group of psychiatric patients are trying three different therapies: counseling, medication and biofeedback. You want to see if one therapy is better than the others. A manufacturer has two different processes to make light bulbs. They want to know if one process is better than the other. Students from different colleges take the same exam. You want to see if one college outperforms the other.

10.1 ANOVA

When a research study obtains means from more than two groups or more than two treatment conditions, the appropriate hypothesis test is an analysis ofvariance, commonly referred to as an ANOVA. When the groups are defined by a single factor with more than two levels (such as three age groups or threetemperature conditions), the test is called a single-factor analysis of variance, ora one-way ANOVA .The mean is computed for each group of participants or for each treatment condition, and the differences among the means from the sample data are used to evaluate a hypothesis about the differencesamong the corresponding population means. The null hypothesis states thatthere are no differences among the population means. The test statistic for ANOVA is an Fratio, which has the same basicstructure as the t statistic. The numerator of the ratio measures the size of theobtained mean differences and the denominator measures how much differenceis

reasonable to expect between the sample means if the null hypothesisis true. However, both the numerator and denominator of the F-ratio are variancesthat provide an overall measure of the mean differences among severaldifferent sample means. The denominator of the F-ratio is often called theerror variance because it measures mean differences that are not caused by thetreatments but rather are simply the result of chance or error. A large F-ratioindicates that the sample mean differences are greater than would be expectedif there were no corresponding mean differences in the population. The appropriatemeasure of effect size is 2, which measures the percentage of varianceaccounted for by the treatment effect. In general, 2 is computed by dividingthe sum of squared deviations (SS) for the treatment by the sum of squared deviations for the error variance.

10.2 Variance Ratio Test

Variance has always been used as a measure of statistical analysis. Sir RA Fisher, the great statistician of the twentieth century, introduced the term "variance" in 1920 for the analysis of statistical data. The technique for the analysis of variance of two or more samples was developed by Fisher himself. The "variance-ratio test" is also known as "*F*-ratio test" or *F*-test. The *F*-test demonstrates that whether the variance of two populations from which the samples have been drawn is equal or not, whereas the "analysis of variance" ascertains the difference of variance among more than two samples.

The *F*-test demonstrates that whether the variance of two populations from which the samples have been drawn is equal or not, whereas the "analysis of variance" ascertains the difference of variance among more than two samples.

10.3 ANOVA for correlated scores

Repeated measures ANOVA is the equivalent of the one-way ANOVA, but for related, not independent groups, and is the extension of the dependent t-test. A repeated measures ANOVA is also referred to as a within-subjects ANOVA or ANOVA for correlated samples. All these names imply the nature of the repeated measures ANOVA, that of a test to detect any overall differences between related means. There are many complex designs that can make use of repeated measures, but throughout this guide, we will be referring to the most simple case, that of a one-way repeated measures ANOVA. This particular test requires one independent variable and one dependent variable. The dependent variable needs tobe continuous (interval or ratio) and the independent variable categorical (either nominal or ordinal).

10.3.1 When to use a Repeated Measures ANOVA

We can analyse data using a repeated measures ANOVA for two types of study design. Studies that investigate either (1) changes in mean scores over three or more time points, or (2) differences in mean scores under three or more different conditions. For example, for (1), you might be investigating the effect of a 6-month exercise training programme on blood pressure and want to measure blood pressure at 3 separate time points (pre-, midway and post-exercise intervention), which would allow you to develop a time-course for any exercise effect. For (2), you might get the same subjects to eat different types of cake (chocolate, caramel and lemon) and rate each one for taste, rather than having different people taste each different cake. The important point with these two study designs is that the same people are being measured more than once on the same dependent variable (i.e., why it is called repeated measures).

10.4. Two way ANOVA

When a research design includes more than one factor (a factorial design), youmust use a hypothesis test to evaluate the significance of the mean differences. The simplest case, a two-factor design, requires a two factor analysis of variance, or two-way ANOVA. The two-factor ANOVAconsists of three separate hypothesis tests. One test evaluates the main effectsfor the fi rst factor, a second test evaluates the main effects for the second factor,and a third test evaluates the interaction. The significance of any one test has no relationship to the significance of any other test. The analysis also produces three separate values of 2 to measure the effect size for each of the maineffects and for the interaction.

The data from a two-factor design can be displayed as a matrix with the levels of one factor defining the rows and the levels of the second factor defining the columns. The mean is computed for each cell in the matrix, and the overall mean is computed for each row and for each column. The differences among the sample means are used to evaluate the three hypotheses about the differences among the corresponding population means. For all three tests, the null hypothesis states that there are no differences among the population means.

10.5. Summary

A common approach to figure out a reliable treatment method would be to analyse the days it took the patients to be cured. We can use a statistical technique which can compare these three treatment samples and depict how different these samples are from one another. Such a technique, which compares the samples on the basis of their means, is called ANOVA.

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

10.6. Key Words

ANOVA is a statistical technique that is used to check if the means of two or more groups are significantly different from each other

"***F*-ratio test**; The *F*-test demonstrates that whether the variance of two populations from which the samples have been drawn is equal or not, whereas the "analysis of variance" ascertains the difference of variance among more than two samples.

10.7. Self-Assessment

1. Analysis of variance is a statistical method of comparing the of several populations.
a. Standard Deviations
b. Variances
c. Means
d. Proportions
e. None Of The Above
2. The sum of squares measures the variability of the observed values around their respective treatment means.
a. Treatment
b. Error
c. Interaction

d. Total

3. The sum of squares measures the variability of the sample treatment means around the overall mean.

a. Treatment
b. Error
c. Interaction
d. Total

4. If the true means of the k populations are equal, then MSTR/MSE should be:

a. more than 1.00
b. close to 1.00
c. close to 0.00
d. close to -1.00
e. a negative value between 0 and - 1
f. not enough information to make a decision

5. If the MSE of an ANOVA for six treatment groups is known, you can compute

a. df1
b. the standard deviation of each treatment group
c. the pooled standard deviation
d. b and c
e. all answers are correct

6. To determine whether the test statistic of ANOVA is statistically significant, it can be compared to a critical value. What two pieces of information are needed to determine the critical value?

a. sample size, number of groups
b. mean, sample standard deviation
c. expected frequency, obtained frequency
d. MSTR, MSE

7. Which of the following is an assumption of one-way ANOVA comparing samples from three or more experimental treatments?

a. All the response variables within the k populations follow a normal distributions.
b. The samples associated with each population are randomly selected and are independent from all other samples.
c. The response variable within each of the k populations have equal variances.
d. All of the above.

8. The error deviations within the SSE statistic measure distances:

a. within groups
b. between groups
c. both (a) and (b)
d. none of the above
e. between each value and the grand mean

9. When the *k* population means are truly different from each other, it is likely that the average error deviation:

a. is relatively large compared to the average treatment deviations
b. is relatively small compared to the average treatment deviations
c. is about equal to the average treatment deviation
d. none of the above
e. differ significantly between at least two of the populations

10. As variability due to chance decreases, the value of *F* will

a. Increase
b. Stay The Same
c. Decrease
d. Can't Tell From The Given Information

11. In a study, subjects are randomly assigned to one of three groups: control, experimental A, or experimental B. After treatment, the mean scores for the three groups are compared. The appropriate statistical test for comparing these means is:

a. The Correlation Coefficient
b. Chi Square
c. The T-Test

d. The Analysis Of Variance
12. In one-way ANOVA, which of the following is used within the *F*-ratio as a measurement of the variance of individual observations?
a. SSTR
b. MSTR
c. SSE
d. MSE
e. none of the above
13. When conducting an ANOVA, *F*DATA will always fall within what range?
a. between negative infinity and infinity
b. between 0 and 1
c. between 0 and infinity
d. between 1 and infinity
14. If *F*DATA = 5, the result is statistically significant
a. Always
b. Sometimes
c. Never

15. If *F*DATA= 0.9, the result is statistically significant
a. Always
b. Sometimes
c. Never

Answers				
1	2	3	4	5
C	B	A	B	C
6	7	8	9	10
C	D	A	B	B
11	12	13	14	15
D	C	C	C	C

10.8. Review Questions

Define ANOVA?
What do you mean by one way ANOVA?
Discuss need and importance of ANOVA in social science research?

Further Readings

Kazdin, A. E. (2003). Research design in clinical psychology (4th ed.). Boston: Allyn and Bacon.

UNIT 11- Advanced Statistics

Contents

Objectives

After studying this unit you will be able to

Understand different types of correlation

Regression line

Uses and Principles of factor analysis

Introduction

The most commonly used techniques for investigating the relationship between two quantitative variables are correlation and linear regression. Correlation quantifies the strength of the linear relationship between a pair of variables, whereas regression expresses the relationship in the form of an equation. For example, in patients attending an accident and emergency unit (A&E), we could use correlation and regression to determine whether there is a relationship between age and urea level, and whether the level of urea can be predicted for a given age.

11.1 Partial correlation

Partial correlation is a method used to describe the relationship between two variables whilst taking away the effects of another variable, or several other variables, on this relationship. Partial correlation is best thought of in terms of multiple regression; StatsDirect shows the partial correlation coefficient r with its main results from multiple linear regression. A different way to calculate partial correlation coefficients, which does not require a full multiple regression, is show below for the sake of further explanation of the principles: Consider a correlation matrix for variables A, B and C (note that the multiple line regression function in Stats Direct will output correlation matrices for you as one of its options):

The partial correlation of A and B adjusted for C is:

$$r_{ABC} = \frac{r_{AB} - r_{AC}r_{BC}}{\sqrt{(1 - r_{AC}^2)(1 - r_{BC}^2)}}$$

11.2Multiple correlations

The multiple correlation coefficient denoting a correlation of one variable with multiple other variables. The multiple correlation coefficients is denoted as $\mathbf{R}_{A.BCD...k}$ which denotes that A is correlated with B, C, D, up to k variables. For example, we want to compute multiple correlation between A with B and C then it is expressed as $R_{A.BC}$. In this case we create a linear combination of the B and C which is correlated with A. We continue with the same example which we have discussed for partial and semipartial correlations. This example has academic achievement, anxiety and intelligence as three variables. The correlation between academic achievement with the linear combination of anxiety and intelligence is multiple correlation. This denotes the proportion of variance in academic achievement explained by intelligence and anxiety. We denote this as R (Academic Achievement. Intelligence, Anxiety), which is a multiple correlation. Often, it is used in the context of regression, where academic achievement is a criterion variable and intelligence and anxiety are called as predictors. We are not using regression equation since you have not learned it. The Multiple R can be calculated for two predictor variable as follows

$$R_{A.BC} = \sqrt{\frac{r_{AB}^2 + r_{AC}^2 - 2r_{AB}r_{AC}r_{BC}}{1 - r_{BC}^2}}$$

Where,

$R_{A.BC}$ = is multiple correlation between A and linear combination of B and C.

r_{AB} = is correlation between A and B

r_{AC} = is correlation between A and C

r_{BC} = is correlation between B and C

11.3 Regression

The Pearson correlation describes the linear relationship between two variablesWhenever a linear relationship exists, it is possible to compute the equation for the straight line that provides the best fi t for the data points. The process of fi nding the linear equation is called regression, and the resulting equation is called the regression equation. All linear equations have the same general structure and can be expressed as

$$Y = bX + a$$

where b and a are fi xed constants. The value of b is called the slope constant because it describes the slope of the line (how much Y changes when X is increased by 1 point). The value of a is called the Y-intercept because it is the point at which the line intersects the Y axis. The process of regression identifyes the specifi c values for b and a that produce the most accurate predictions for Y. That is, regression identifies the specifi c equation that results in the smallest possible error between the predicted Y values on the line and the actual Y values in the data.

11.4 factor analysis

Factor analysis is a useful tool for investigating variable relationships for complex concepts such as socioeconomic status, dietary patterns, or it allows researchers to investigate concepts they cannot measure directly. It does this by using a large number of variables to esimate a few interpretable underlying factors. The key concept of factor analysis is that multiple observed variables have similar patterns of responses because they are all associated with a latent variable (i.e. not directly measured). For example, people may respond similarly to questions about income, education, and

occupation, which are all associated with the latent variable socioeconomic status. In every factor analysis, there are one fewer factors than there are variables. Each factor captures a certain amount of the overall variance in the observed variables, and the factors are always listed in order of how much variation they explain.

11.5. Summary

A basic model of factor analysis is employed in the estimation of multiple correlation coefficients and partial regression weights. Estimators are derived for situations in which some or all of the independent variates are subject to errors in measurement. The effect of the errors is indicated and the problem of bias in the estimators is considered. In one special case it is shown how a best subset of the independent variates of any size can readily be found for data under analysis.A different way to calculate partial correlation coefficients, which does not require a full multiple regression, is show below for the sake of further explanation of the principles: Consider a correlation matrix for variables A, B and C (note that the multiple line regression function in Stats Direct will output correlation matrices for you as one of its options):

11.6. Key Words

Factor analysis:-A research design that includes two or more factors.

Regression:-A statistical technique used for predicting one variable from another. The statistical process of finding the linear equation that produces the most accurate predicted values for Y using one predictor variable, X.

11.7. Self-Assessment

1. Which of the following are types of correlation?
    a. Positive and Negative
    b. Simple, Partial and Multiple
    c. Linear and Nonlinear
    d. All of the above
2. Which of the following is true for the coefficient of correlation?
    a. The coefficient of correlation is not dependent on the change of scale
    b. The coefficient of correlation is not dependent on the change of origin
    c. The coefficient of correlation is not dependent on both the change of scale and change of origin
    d. None of the above
3. Which of the following statements is true for correlation analysis?
    a. It is a bivariate analysis
    b. It is a multivariate analysis
    c. It is a univariate analysis
    d. Both a and c
4. If the values of two variables move in the same direction, _____
    a. The correlation is said to be non-linear
    b. The correlation is said to be linear
    c. The correlation is said to be negative
    d. The correlation is said to be positive

5. If the values of two variables move in the opposite direction, _____
    a. The correlation is said to be linear
    b. The correlation is said to be non-linear
    c. The correlation is said to be positive
    d. The correlation is said to be negative
6. Which of the following techniques is an analysis of the relationship between two variables to help provide the prediction mechanism?
    a. Standard error
    b. Correlation
    c. Regression
    d. None of the above
7. Which of the following statements is true about the arithmetic mean of two regression coefficients?
    a. It is less than the correlation coefficient
    b. It is equal to the correlation coefficient
    c. It is greater than or equal to the correlation coefficient
    d. It is greater than the correlation coefficient
8. What is the meaning of the testing of the hypothesis?
    a. It is a significant estimation of the problem
    b. It is a rule for acceptance or rejection of the hypothesis of the research problem
    c. It is a method of making a significant statement
    d. None of the above
9. Which of the following statements is true about the null hypothesis?
    a. Any wrong decision related to the null hypothesis results in two types of errors
    b. Any wrong decision related to the null hypothesis results in one type of an error
    c. Any wrong decision related to the null hypothesis results in four types of errors
    d. Any wrong decision related to the null hypothesis results in three types of errors
10. Which of the following statements is true about the type two error?
    a. Type two error means to accept an incorrect hypothesis
    b. Type two error means to reject an incorrect hypothesis
    c. Type two error means to accept a correct hypothesis
    d. Type two error means to reject a correct hypothesis
11. Which of the following statements is true about the level of significance?
    a. In testing a hypothesis, we take the level of significance as 2% if it is not mentioned earlier
    b. In testing a hypothesis, we take the level of significance as 1% if it is not mentioned earlier
    c. In testing a hypothesis, we take the level of significance as 10% if it is not mentioned earlier
    d. In testing a hypothesis, we take the level of significance as 5% if it is not mentioned earlier
12. The independent variable is used to explain the dependent variable in _____.
    a. Linear regression analysis

b. Multiple regression analysis

c. Non-linear regression analysis

d. None of the above

13. Which of the following statements is true about the regression line?

a. A regression line is also known as the line of the average relationship

b. A regression line is also known as the estimating equation

c. A regression line is also known as the prediction equation

d. All of the above

14. Which of the following statements is true about the correlational analysis between two sets of data?

a. The correlational analysis between two sets of data is known as a simple correlation

b. The correlational analysis between two sets of data is known as multiple correlation

c. The correlational analysis between two sets of data is known as partial correlation

d. None of the above

15. The original hypothesis is known as _____.

a. Alternate hypothesis

b. Null hypothesis

c. Both a and b are incorrect

d. Both a and b are correct

Answers				
1	2	3	4	5
D	C	C	D	D
6	7	8	9	10
C	D	B	A	A
11	12	13	14	15
A	A	D	A	B

11.8. Review Questions

1. Discuss partial correlation in detail
2. Define Regression
3. Describe need and importance of Factor analysis

Further Readings

 Anderson, T.W., An Introduction to Multivariate Analysis, New York: John Wiley & Sons, 1958.

Bailey, Kenneth D., "Methods of Social Research," New York, 1978.

Baker, R.P., and Howell, A.C., The Preparation of Reports, New York: Ronald Press, 1938.

12. Non- Parametric Tests

Content

Objectives

This unit will enable you: -

To understand the concept and importance of non-parametric tests

To understand about the nature and assumption of non-parametric test

To understand the concept of distribution free statistic

To understand the concept and importance of chi-square, contingency coefficient, median and sign test, Friedman test

Introduction

Statistics is classified into two types based on its population distribution. There will be instances where the population is normally distributed and, in those instances, **parametric statistics** can be used for hypothesis testing. There will be certain instances where the population will not be normally distributed or researcher is conducting studies on rare or unique phenomenon where probability sampling is not possible. Under such circumstances **non-parametric statistics** can be used for hypothesis testing. Thus, non-parametric statistics is not expecting any characteristic structure to parameters from the population. Unlike parameter tests, which require an underlying population to compare non-parametric tests does not require an underlying population to compare data obtained.

We use non-parametric tests everywhere, for example, checking the fanbase for IPL teams in your classroom by head counting.

12.1Non-parametric tests

1. Non- parametric tests are statistical tests which does not require any underlying population with parameters: non-parametric does not go with hypothesis testing based on population distribution and data can even be skewed.
2. Non-parametric statistics can be used in a very small population: When the data size is very small non-parametric tests remains the only option.
3. Non-parametric tests deal with data that are in ordinal scales: non-parametric data does not require data in ratio and interval scale. It can be administered to be data which is distributed in ordinal scale.
4. Non-parametric tests do not have assumption regarding sampling method: There is no strict rules of probability sampling for non-parametric statistics. Data collected through non-probability sampling can be used for non-parametric analysis.
5. Non-parametric tests can only make few numbers of assumptions compared to parametric tests: it is kind of robust test but does not depend on any distribution.

12.2. Assumptions of non-parametric tests

1. Non-parametric tests are based on central value of median, where the analysis is based on the differences between medians.
2. The data should either be in nominal, ordinal, ratio or interview scale for non-parametric analysis.
3. Independence of data and test is the main assumption of non-parametric test.

4. Randomness in terms of sampling is considered to be an assumption of non-parametric tests.
5. Non-parametric test does not analyze raw data like parametric test. It requires data in rank order form.
6. To use non-parametric tests obvious non-normality or possible non-normality must be there.


Try to read some articles which uses non-parametric tests and try to find out the nature and assumptions of non-parametric tests

12.3.Distribution free statistics

If the population does not have any parametric related to normality and sampling it is considered to be distribution free statistic. Non-parametric tests are known as distribution free statistic because it does not require any underlying assumptions about the population. The examples of distribution free statistics are: Chi-square test, Mann Whitney U test, Kruskal Wallis test, Wilcoxon sum rank test, and the Sign test.

12.4. Chi-square test

A chi-square test is a distribution free statistic which tests hypothesis based on the difference between observed values and expected values. Karl Pearson introduced the chi-square test to analyze data categorically. It is basically developed to analyze the observations of set of variables that are distributed randomly. So, the assumption of chi-square test is that data must be distributed categorically like three categories under gender: male, female and other gender.

Properties of chi-square test

1. Degrees of freedom is very important in chi-square test, where two times the number of df must be equivalent to variance.
2. Degrees of freedom must be equal to the mean distribution.
3. As the degrees of freedom increases the chi-square distribution reach closer to normal distribution.

Formula of Ch-square test

The formula of chi-square test =

$$\chi c2=\sum Ei(Oi-Ei)2$$

where:c=Degrees of freedom$O$=Observed value(s)E=Expected value(s)

There are three main types of chi-square test to find out the goodness of fit. Test for homogeneity and independence tests. The same formula is used for all the three types.

Goodness of fit is used when a randomly collected sample has only one category and has to be compared with population. If we need to find out the association between two categorical variables, we can use independence of test. Test of homogeneity can be used to understand whether the distribution of the variable is same in every population.

How to do Chi-square test?

A survey was conducted in 2018 and found that 60% of those who own a house have only one house and 28% has 2 house and 12% has three or more houses. You want to verify the data and conducts a survey to cross check and collected data from 129 house owners and found that 73 had one house and 38 had 2 houses and 18 had three or more houses

Step 1: Formulating null hypothesis, H0: the proportion of house owners are 0.6, .28 and .12 respectively.
Step 2: Finding out the appropriate test, here goodness of fit can be used.
Step 3: Tabulate the information

No of house	Observed	Expected	0i-Ei	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
1	73	.60X 129=77.4	-4.4	19.36	0.25
2	38	.28X129=36.1	1.9	3.61	0.1
3 or more	18	.12X129=15.5	2.5	6.25	0.40
Total	129				0.75

Step 4: degrees of freedom is 2, and the table value is 5.99 and the chi-square value is .75 which is lesser so we cannot reject the null hypothesis.

12.5.Contingency coefficient

Contingency coefficient is calculated by calculating the square of chi-square statistics. It is used to find out the association between different categories. If the coefficient is 0 or anywhere near to 0 it is assumed that there is no association between the groups.

https://accendoreliability.com/contingencycoefficient/#:~:text=The%20contingency%20coefficient%20is%20a,coefficient%20with%20the%20following%20formula.&text=and%2C%20n%20%3D%20total%20sample%20size.

12.6 Sign and Median test

12.6.1 Sign test

Sign test is a one sample test. Sign test is named so because it uses signs of + and _. It can be used for both one sample and two samples when quantitative measurement is not possible but the data can be ranked.

12.6.2 Properties of sign test

1. It can be used for the data that is distribution free.

2. When assumptions about population cannot be made, we can use sign test.

2. Sign test is not concerned about the skewness and symmetry of the data.

12.6.3 How to do a sign test?

The marks obtained by the students on 10 are given below

5

7

9

8

5

4

3

7

8

4

6

7

8

3

4

Step 1: Formulating null hypothesis that H0: The obtained median value is equal to the assumed median value.

Step 2: Finding out the median (after arranging the data in ascending order it is found that median is 6).

Step 3: Assigning + or – signs to the scores below median (-) and above median (+).

5-

7+

9+

8+

5-

4-

3-

7+

8+

4-

6 0

7+

8+

3-

4-

Step 4: compare the difference and found that there is 7+ and 7- . Thus, there is no difference thus we cannot reject null hypothesis.

12.6.4 Median test

Median test is also a one sample non-parametric test that determines whether the median of the single set of data is same as the assumed median. It can be used to find out whether the obtained median is equal to the assumed median. This test is not considered to be efficient with large sample size.

12.6.5 Properties of median test

1. It requires very few assumptions about the data.

2. It has limited power compared to there tests like Mann Whitney U test.

3. Sample size can be considerably low for median test.

12.6.6 How to do a median test?

Suppose you manage a children's park and you hypothesize that the median age of the visitors is 9. This is your null hypothesis thus H0 is "the median age of the visitors is 9".

Step 1: Assign a value of 0 to those below the median age.

2-0

3-0

4-0

5-0

6-0

7-0

9-1

10-1

11-1

12-1

Step 2: Find out the frequency

It is found that 6 children are below the median and 4 children are above the median.

Step 3; Find out the expected frequency

We expect 5 children below 8 and 5 above 8.

Step 4: Calculating the chi-square statistics

$X^2 = \Sigma [(of - ef)^2/ef]$ which is $(6 - 5)^2/5 + (4 - 5)^2/5 = .40$

Step 5: Finding the degrees of freedom and it is 1.

Step 6: Finding the chi-square value from the table and it is 3.84, and our value is .40 which is lesser than chi-square value. Thus, we cannot reject the null hypothesis.

12.7 Friedman test

Friedman's test is considered to be parallel to the parametric test of one-way anova which is developed by economist Milton Friedman using repeated measures. It assesses the difference between the groups. You can administer Friedmann test if at least one dependent variable is in ordinal scale and it can be seen as an extension of sign test.

Properties of Friedman test

1. You can use Friedman's test if there is no interaction is found between the groups.

2. It can be used if data is at least in ordinal scale.

3. Assumptions of normality is not a concern for Friedmann's test.

How to do a Friedman test?

Step 1: Formulating the null hypothesis

Step 2: Ranking the data together and independently of the rows

Step 3: Sum the ranks

Step 4: Compute the statistic using the formula:

Step 5: Compare it with the table value.

 https://www.statology.org/friedman-test/

12.8 Summary

The scope of statistics in psychology is devastating and it helps in quantifying the psychological attributes. Despite the nature of the data and assumptions of the population and sampling methods we can use a range of statistical methods to test hypothesis.

Generally statistical methods are divided into parametric and non-parametric statistical methods. Parametric statistics have large number of assumptions regarding the population and is concerned about normality and probability sampling methods.

Non-parametric statistical method has fewer assumptions regarding the population like normality, skewness, sample size and sampling methods. Non-parametric statistics deals with data that are distributed in nominal and ordinal scales and they are statistical methods in non-parametric which can be used as an alternative for parametric statistics.

Non-parametric tests are also known as distribution free statistics because it does not have concerns related to the assumptions of the population. The examples are: Mann-Whitney U test, Kruskal Wallis test, sign test etc.

Chi-square test is a distribution free statistic which focuses on the difference between the observed and expected frequencies. There are three types of chi-square test which is goodness of fit, independence test and test for homogeneity.

Sign test and median test are one sample non-parametric tests. Sign test got its name because of the use of + and _ signs in data tabulation. Both sign and median test is based on the median and finds out the difference between assumed median and observed median.

Friedman test is an alternative for one-way anova which can be administered on large sample size. It follows the assumption that at least one dependent variable must be in ordinal scale.

12.9. Keywords:

Non-Parametric statistics: Statistical methods that are used to find out the difference or associations between categorical data or samples that does not meet the criteria for normality or assumptions of probability of sampling or data distribution.

Distribution free statistics: Methods that are used for hypothesis testing n data that does not meet the norms of normality or assumptions of population. For example, Mann Whitney U-test, Kruskal Wallis test and Chi-square test.

Chi-square test: An example of non-parametric test that is used to find out the difference between observed frequencies and expected frequencies.

Sign test:A non-parametric one sample test which uses signs of + and – for data tabulation and which uses median as a criterion for statistic.

Median test:A non-parametric test which is based on the difference between assumed median and obtained median.

Freidman test: An alternative of one-way anova, which can be used to find out the differences between the groups.

12.10 Self-Assessment

1. ……. refers to the methods in statistics which has fewer assumptions about the data

a. Parametric statistics

b. non-parametric statistics

c. scientific measurement

d. geometrical analysis

2. …… is another name for non-parametric statistics

a. non-parametric statistics

b. parametric statistics

c. mean deviation

d. median test

3. ……. Test computes the difference between observed and expected frequency

a. Friedman test

b. median test

c. sign test

d. chi-square test

4. sign test uses ……. While tabulating data

a. codes

b. numbers

c. signs

d. lines

5. …. Test computes the difference between observed and expected median

a. chi-square

b. Kruskal Wallis

c. Median test

d. Mann Whitney U test

6. Mann Whitney U test and Kruskal Wallis test are examples of

a. Parametric tests

b. distribution free statistics

c. anova

d.one sample test

7. …….. developed Friedman test

a. Spearman

b. Pearson

C. Friedman

d. Anderson

8. Goodness of fit is a type of……

a. chi-square test

b. median test

c. Kruskal Wallis test

d. Mann Whitney U test

9. Which test is not powerful if sample size is large

a. Sign test

b. chi-square test

c. Friedman test

d. median test

10. Oi-Ei is found in ……

a. sign test

b. median test

c. chi-square test

d. Friedman test

11. If there are 3 categories of data then the degrees of freedom is…….

a. 2

b.3

c.4

d.0

12. …… calculates the square of chi-square statistic

a. contingency coefficient

b. chi-square test

c. median test

d. Friedman test

13. If a data is skewed you can use…….

a. parametric

b. graphs

c. mathematical models

d. non-parametric tests

14. …… requires at least one dependent variable in ordinal scale

a. median test

b. Freidman test

c. chi-square test

d. sign test

15 if there are two or three categories and you need to see the difference you can use

a. Goodness of fit
b. Independence of tests
c. Test of homogeneity
d. Median test

Answers

1	2	3	4	5
b	a	d	c	c
6	7	8	9	10
b	c	a	d	c
11	12	13	14	15
a	a	d	a	a

12.11 Review Questions

1. Explain the nature of non-parametric tests

2. What is the difference between parametric and non-parametric test?

3. What are the assumptions of non-parametric test?

4. Explain chi-square test and its properties

5. Explain Friedman's test with the steps to administer the test.

Further readings

Kaila, H. L. (2017). Textbook of Parametric and Nonparametric Statistics. *Journal of Psychosocial Research*, *12*(1).

13. Computational Technique: Data coding, entry, and checking

Content

Objectives

This unit will enable you: -

To understand the concept and importance of computational technique

To understand the data entry

To understand data coding

Introduction

Computational methods play an important role in data analysis due to the large data size and complexity of the data. and various underlying assumptions of the given data that must be statistically validated.Computational statistics or statistical computing is an interdisciplinary exchange that incorporates the techniques from statistics and computer science. Computational statistics is growing at a tremendous pace and there are a large number of advancements being made in it.

Computational methods are used even by Instagram and YouTube for data analytics to understand the dynamics of the users.

Computational methods use mathematical programs like algorithms, mathematical problems, and solutions. These methods have the ability to:

1. To analyze the complexities of the world, in a simple manner.
2. To extract the information from huge data sources.
3. To identify the important elements from a dataset.
4. To draw assumptions and conclusions from data.
5. To prove whether predictions made by researchersare true or not.

Try to read some articles about computational techniques and try to find out the nature of computational techniques.

1.2 Data Coding

Data coding is the process of driving codes from the observed data. In qualitative research the data is either obtained from observations, interviews or from questionnaires. The purpose of data coding is to bring out the essence and meaning of the data that respondents have provided. The data coder extract preliminary codes from the observed data, the preliminary codes are further filtered and refined to obtain more accurate precise and concise codes. Later, in the evaluation of data the researcher assigns values, percentages or other numerical quantities to these codes to draw inferences. It should be kept in mind that the purpose of data coding is not to just to eliminate excessive data but to summarize it meaningfully. The data coder should ascertain that none of the important points of the data have been lost in data coding.

1.3Data Entry

Data entry is considered a non–core process for most organizations and is usually performed on data forms such as spreadsheets, handwritten or scanned documents, audio or video. Addition, modification, and deletion are the three modes of operation in data entry.

Data entry jobs do not require any special qualifications, knowledge or talent, and only require accuracy and fast turnaround. As such, data entry jobs are frequently outsourced in order to lower costs. Computers are also used in automated data entry, as they are highly accurate and can be programmed to fetch and transcribe data into the required medium.

Accurately keyed data is the base upon which the organization can perform analyses and make plans

1.4- Data Checking

Data checking is an activity through which the correctness conditions of the data are verified. Context: It also includes the specification of the type of the error or condition not met, and the qualification of the data and its division into the "error free" and "erroneous data".

While quality control activities (detection/monitoring and action) occur during and after data collection, the details should be carefully documented in the procedures manual. A clearly defined communication structure is a necessary pre-condition for establishing monitoring systems. There should not be any uncertainty about the flow of information between principal investigators and staff members following the detection of errors in data collection. A poorly developed communication structure encourages lax monitoring and limits opportunities for detecting errors.

Data checking, Detection or monitoring can take the form of direct staff observation during site visits, conference calls, or regular and frequent reviews of data reports to identify inconsistencies, extreme values or invalid codes. While site visits may not be appropriate for all disciplines, failure to regularly audit records, whether quantitative or quantitative, will make it difficult for investigators to verify that data collection is proceeding according to procedures established in the manual. In addition, if the structure of communication is not clearly delineated in the procedures manual, transmission of any change in procedures to staff members can be compromised.

Quality control also identifies the required responses, or 'actions' necessary to correct faulty data collection practices and also minimize future occurrences. These actions are less likely to occur if data collection procedures are vaguely written and the necessary steps to minimize recurrence are not implemented through feedback and education (Knatterud, et al, 1998)

Examples of data collection problems that require prompt action include:

- errors in individual data items
- systematic errors
- violation of protocol
- problems with individual staff or site performance
- fraud or scientific misconduct

1.5. Summary

Coding is the analytic task of assigning codes to non-numeric data. Coding language data is a technique used in a variety of research traditions. In traditional content analysis, coding falls under the heading of "human coding" and makes use of a codebook which, according to Neuendorf (2016) should be set up in advance of coding and should be "so complete and unambiguous as to almost eliminate the individual differences among coders" (Chapter 5, Section on Codebooks and Coding Forms, para. 1). In qualitative analysis, coding is treated as an activity that creates and assigns a word or phrase to symbolize, summarize, or otherwise capture some attribute of "a portion of

language-based or visual data," often in interaction with that data. Finally, in-text mining,especially that using supervised machine learning, language data is often coded in the first stage of work to create a corpus from which a machine can then "learn".

1.6. Keywords

Coding is the analytic task of assigning codes to non-numeric data.
Data checking is an activity through which the correctness conditions of the data are verified

1.7. Self-Assessment

1. Coding is the analytic task of assigning codes to non-numeric data
    a. True
    b. False
2. Data checking does not ensure the correctness of data
    a. True
    b. False
3. Addition, modification, and deletion are the three modes of operation in data checking.
    a. True
    b. False
4. Data checking is an inevitable process in computational techniques.
    a. True
    b. False
5. Computational methods play an important role in data analysis
    a. True
    b. False
6. Data entry can only be done in spreadsheets.
    a. True
    b. False
7. Language coding is the first step in machine coding, which allows the machine to learn.
    a. True
    b. False
8. Computational techniques are used to make the data analysis complicated.
    a. True
    b. False
9. Data coding is not possible with qualitative data
    a. True
    b. False
10. Data entry requires expertise and skills
    a. True
    b. False

1.8. Review questions

1. How data checking is important for research in social science
2. What do you mean by Data Entry?
3. How data coding has its relevance with social science research

Further Readings

- Knatterud.,G.L., Rockhold, F.W., George, S.L., Barton, F.B., Davis, C.E., Fairweather, W.R., Honohan, T., Mowery, R, O'Neill, R. (1998). Guidelines for quality assurance in multicenter trials: a position paper. *Controlled Clinical Trials*, 19:477-493.

- Most, .M.M., Craddick, S., Crawford, S., Redican, S., Rhodes, D., Rukenbrod, F., Laws, R. (2003). Dietary quality assurance processes of the DASH-Sodium controlled diet study. *Journal of the American Dietetic Association*, 103(10): 1339-1346.

- Whitney, C.W., Lind, B.K., Wahl, P.W. (1998). Quality assurance and quality control in longitudinal studies. Epidemiologic Reviews, 20(1): 71-80

14. Advance Computational Technique

Content

Objectives

This unit will enable you to:

Know about different types of advance computational techniques

Understand measurement through SPSS

Understand how to calculate descriptive statistics through SPSS

Acquire knowledge about N-vivo and R

Introduction

14.1. Advance Computational Technique

In statistics, computer techniques are applied to make data tabulation, analysis, and computation easier. There are several software and package that are available for this purpose. Basic programs like Excel sheets and sophisticated packages like SPSS, AMOS, N-Vivo, and R. These packages are made for the use of researchers and they can use any of these for meeting their research objectives.

Advantages of Computational techniques in statistics

1. It makes data tabulation easy
2. It ensures hand free data classification and paperless mathematical techniques
3. Accuracy of data analysis can be ensured.
4. Scope for applying multiple statistical analyses and graphical representations is possible with advanced computational techniques.
5. A comprehensive analysis can be done using computational techniques in statistics.

14.2. Measurement through SPSS

Statistical Package for Social Sciences, known as SPSS is a useful tool for educationalists, researchers, scientists, and healthcare practitioners.To define the label of measurement in a variable, we need to understand the property of measure in SPSS. In statistics, we have 4 levels of measurement namely nominal, ordinal, ratio, and interval scale. But in SPSS both ratio and interval levels of measurement are treated as scales and thus it has only three scales.

1. **Nominal scale:**
   Nominal variables are also known as categorical variables were ranking the items is not possible. For example, gender, where we cannot assign ranks rather, we can just categorize them.
2. **Ordinal scale:**
   Ordinal level of measurement is applied when there is a possibility for intrinsic ranking of items. For example, if you are assessing the workplace satisfaction on a 3-point scale where:
   1= High satisfaction
   2=Average satisfaction
   3=Below satisfaction

Where I, 2, and 3 indicate order, and thus ranking is possible.

3. **Scale:**
   Continuous scale that is interval and ratio are together known as scales in SPSS for operational purposes. The scale incorporates the variables that can be categorically organized, has equal intervals, and has zero.

14.3. Descriptive statistics through SPSS

Descriptive statistics help you to provide the summary and description of the data. For example, the average age of the distribution, the average income of the population, and the demographic details of the population. Descriptive statistics include:

1. Mean
2. Mode
3. Standard deviation
4. Skewness
5. Kurtosis
6. Standard Error of the mean
7. Frequencies
8. Quartiles
9. Percentiles
10. Minimum
11. Maximum
12. Outliers
13. Range
14. N valid responses

How to run descriptive analysis on SPSS?

To run the descriptive statistics on spss, we need to select Analyse then after that, we need to select descriptive statistics, and then we need to select descriptive. The descriptive list of the variables will showon the left side of the

column then we need to move the variables to the right side. After that, a dialogue box will appear where you need to select the descriptive statistics you want to apply like mean, sum, standard deviation, variance, S.E, Range, Kurtosis, and skewness.

14.4. Uses of N-Vivo

Analysis of qualitative data is very difficult earlier because it was done manually. Now researchers are using the N-Vivo program which helps in analyzing the data collected using qualitative methods like interviews, focus group discussions, social media, and videos. The uses of the N-Vivo package is:

1. It helps to analyze and organize the data that is unstructured in nature like texts, audio, images, videos, and audio.
2. It helps to transcribe the data in N-Vivo which enables playback ability.
3. N-Vivoplugins help to take data from Facebook, Twitter, Instagram, etc. which helps to analyze the content.
4. It helps to import notes that are essential for qualitative analysis.
5. Import citations from various sources are available which can be used for review of literature and references.
6. Simple text analysis can be performed using N-Vivo.


Before using N-Vivo software, the researcher has to use a pile of papers for data analysis

14.5. Uses of R

R is a programming language that helps to do quantitative analysis. It does a broad variety of statistics like traditional tests, time arrangement analysis, grouping, bunching, and statistical techniques.

1. It can be used in financial management and financial data analysis. It helps to graphical plots which is a part of simple financial tasks.
2. In the banking sector R is used for financial reporting. It also does risk analytics in the banking sector.
3. In the healthcare sector R software is used for pre-clinical trials. It helps in statistical modeling in the epidemiology field.
4. R is used for data mining from social media. Google uses R as an important statistical tool for data analytics.
5. R proves to be an effective tool in the E-commerce field for linear modeling and test analysis.
6. In the manufacturer sector R is used to analyze the sentiments of customers. They also use R to cut off the costs for production.
7. R can be used both for descriptive analysis and exploratory data analysis.
8. It allows hypothetical testing in statistical models.
9. It can provide predictive models for data.
10. It provides an interactive web application page called RShiny. It creates interactive visualizations using RShiny.

14.6. Keywords

Computational technique: In statistics, computer techniques are applied to make data tabulation, analysis, and computation easier.

SPSS: Statistical Package for Social Sciences, known as SPSS is a useful tool for educationalists, researchers, scientists, and healthcare practitioners.

Descriptive statistics on SPSS: To select Analyse then after that, we need to select descriptive statistics, and then we need to select descriptive, we need to move the variables to the right side. After that, a dialogue box will appear where you need to select the descriptive statistics you want to apply.
N-Vivo: It's a software which uses for qualitative analysis by researchers mainly for coding the data obtained through interviews, focus group discussions, videos and audio.

R: R is a programming language that helps to do quantitative analysis. It does a broad variety of statistics like traditional tests, time arrangement analysis, grouping, bunching, and statistical techniques.

14.7. Summary

Computational techniques in statistics, are applied to make data tabulation, analysis, and computation easier.SPSS is a Statistical Package for Social Sciences, known as SPSS is a useful tool for educationalists, researchers, scientists, and healthcare practitioners.

N-Vivo is another software which uses for qualitative analysis by researchers mainly for coding the data obtained through interviews, focus group discussions, videos and audio.
R is a programming language that helps to do quantitative analysis. It does a broad variety of statistics like traditional tests, time arrangement analysis, grouping, bunching, and statistical techniques.

14.8. Self-Assessment

1. Computational techniques make data analysis easy for researchers
   a. True
   b. False
2. Comprehensive analysis is not possible using computational techniques
   a. True
   b. False
3. SPSS stands for Statistical Package for Social Sciences
   a. True
   b. False
4. N-Vivo is programming software for both quantitative and qualitative analysis
   a. True
   b. False
5. Simple text analysis can be done using N-Vivo package
   a. True
   b. False
6. R is a programming language
   a. True
   b. False

7. R does not allow hypothetical testing
   a. True
   b. False
8. R can provide predictive models
   a. True
   b. False

9. R interacts with RShiny to provide interactive visualizations
   a. True
   b. False
10. Computational techniques cannot be used for data tabulation
    a. True
    b. False
11. Time arrangement analysis is possible with R

a. True
b. False

12. R can be used for descriptive statistics
    a. True
    b. False
13. Ordinal scale of measurement has absolute zero
    a. True
    b. False
14. Nominal scales deal with categorical variables
    a. True
    b. False
15. Ratio and interval scale are together defined as scale in SPSS for operational purposes
    a. True
    b. False

14.9. Review Questions

1. What are the advantages of computational techniques?
2. What is SPSS?
3. Explain about descriptive statistics on SPSS
4. What are the uses of N-Vivo?
5. What are the uses of R?

Further Readings

https://worldsustainable.org/descriptive-statistics-on-spss/
https://libguides.library.kent.edu/statconsulting/NVivo#:~:text=What%20is%20NVivo%3F,social%20media%2C%20and%20journal%20articles.

LOVELY PROFESSIONAL UNIVERSITY
Jalandhar-Delhi G.T. Road (NH-1)
Phagwara, Punjab (India)-144411
For Enquiry: +91-1824-521360
Fax.: +91-1824-506111
Email: odl@lpu.co.in