Lesson one: Introduction to Statistics

Meaning of Statistics

The term statistics can be used in two senses: in singular and in plural. When used in plural, the term statistics refer to a set of numerical data or numerical values. When used in singular, the term statistics refers to an area of study that deals with statistical methods that are used to transform data into information.

Statistics in Plural

Data

Data are recorded facts relating to objects of study and their attributes. For example, the ages of 5 employees of a certain firm in years are 34, 62, 54, 28, and 39.

Objects of study are entities on which data are collected. They are the entities which are observed to generate data. They go by different names depending on the situation. Objects of study are also called subjects, participants, cases, items, experimental units or elements. Examples of objects of study include people, businesses, religious organizations, governments, public institutions, activities, events, regions, objects etc.

Attributes

Data relate to attributes of an object or objects. Attributes are also called features, variables or factors. An attribute is a characteristic of interest for an object of study.

Attributes of human beings include: age, height, weight, color of the eye, intelligence, religious affiliation, heart rate, blood pressure, temperature, motivation, quantity produced, marks obtained in a test, salary.

Attributes of a business include: startup capital, share price ratio, sales, purchases, electricity expenses, salaries and wages expenses, tax paid, capital employed, liabilities as at a particular date, number of employees, size of the company, net profit, creditors balance, debtors balance, bank balance, value of assets. etc.

Numerical Data and Non-Numerical Data

Data may be numerical or non-numerical. Numerical data are data that are expressed in numbers and represents quantities that can be measured or counted. Non-numerical data are data that are expressed in words or categories and not in numbers. Numerical data are also called quantitative data while non-numerical data are also called qualitative data. Statistical data or statistics are numerical data.

Statistical data are usually obtained by observation, counting, measurement or performing mathematical operations. Examples of statistics include:

- Figures relating to the population such as number of births, deaths, marriages, divorces, age groups, size of population in different regions.
- Figures relating to prices of commodities
- Data on gross domestic product
- Data on company sales, production, expenditure, inventories, capital employed, gross profit, net profit, purchases etc.
- Incomes of people in a particular region.
- Number of students enrolled in different programs in a college.
- Salaries of workers in a particular industry
- Data on unemployment rates in a county.
- Data on inflation rates in a county.
- Data on interest rates charged by various commercial banks on loans.
- Crimes statistics
- Accident statistics.
- Data on prices of securities such as shares.
- Data on exchange rates.
- Data on health care.

Variables

A variable is a characteristic of an object that can assume different values. The key aspect of a variable is the idea that objects differ. Examples of variables include weight, height, gender, favorite color, age, number of lessons attended in a semester, field of study, shoe size, religious affliction, ethnicity, occupation and number of customers entering a supermarket each day. A characteristic that retains the same value from element to element is called a constant.

Data are the values that a variable can assume. The values of a variable can be called observations, measurements, scores, or just data. Values of a variable are obtained by observation, counting, measurement, classification, or some kind of mathematical operation. For example, weight of a person is obtained by measurement while number of students in a given room is obtained by counting. The rate of economic growth in a given year is obtained by performing a mathematical operation.

A collection of data values forms a data set. Each value in the data set is called datum. A data set may consist of one or more variables. A data set that has one variable is called univariate data set. A data set with two variables is a bivariate data set. A data set that has three or more variables is called a Multivariate data set.

Data Set	**Variables**	**Examples**
Univariate	One	Income of employees
Bivariate	Two	Income and age
Multivariate	Three or more	Income, age, gender, religions, affliction, grade, tribe, height and weight.

Types of Variables

A variable is something whose value can change. A constant is something whose value does not change. The number of minutes in an hour is an example of a constant. Variables can take on various types of values, some of them are numbers and some are categories. Consider the following variables and possible values.

Weight: 50kg, 52kg, 49kgs, 70.9kgs,
Height: 150cm, 156.2cm, 140cm 175.4cm
Number of lessons attended: 5, 12, 1, and 10
Marks obtained: 14, 29, 0, 28
Gender: Male, female, female, male
Favorite color: grey, white, orange, yellow, red
Religious affiliation: Islam, Christian, Christian, Judaism
Place of birth: home, hospital, roadside, elsewhere

Variables can be classified as qualitative variables or quantitative variables.

Qualitative Variables are variables which assume non-numeric values.

A qualitative variable is also called a categorical variable. A qualitative variable is a variable whose values are non-numerical. The values of a qualitative variable are described by words rather than numbers e.g. gender, favorite color, ethnicity, religious affiliation, make of a car, geographical location, satisfaction with a product (very unsatisfied, unsatisfied, neutral, satisfied very satisfied. Values of gender are male and female. Values of favourite colour are red, blue white black etc.

Quantitative Variables

Quantitative Variables are also called numeric variables. These are variables which naturally assume numerical values e.g. height, weight, age, income, number of students, body temperature, heart rate etc. Meaningful arithmetic operations (eg addition, subtraction) can be performed on values of quantitative variables. Quantitative variables are further classified into Discrete variables and continuous variables.

Discrete Variables

A discrete variable is a variable which can assume only certain values along an interval, with the possible values having gaps between them e.g. number of children in a family, number of students in a classroom, heart rate and number of customers. Thus, a discrete variable is a variable whose values change by steps.

Discrete variables usually assume positive integer vales i.e. 0, 1, 2, 3 etc. However, fractional values are also possible e.g. the possible values of the variable quantity of milk bought by customers in a supermarket are 0.25L, 0.50L, 0.75L 1L, 1.25L, 1.50L etc. A customer will not be able to purchase 0.30578 litres. The distinguishing feature of discrete variables is that gaps exist between the possible values.

Continuous Variables

A continuous variable is a variable which can assume any value within an interval of numbers i.e. between the lowest and highest points on the scale. They are usually obtained by measuring e.g. height, weight, body temperature, blood pressure. There is an infinite number of values between any two specific values of a continuous variable. Therefore, a continuous variable can assume an infinite number of values between any two specific values.

Measurement Scales

Also called Scales of measurement or Levels of measurement.

Measurement is the process of assigning values to characteristics of elements such as weight, height, gender, motivation, job satisfaction etc.

Scales of measurement refer to the system used to define data so that it can be examined meaningfully. There are four measurement scales: nominal scale, ordinal scale, interval scale and ratio scale. Data measured on nominal scale, ordinal scale, interval scale and ratio scale are called nominal data, ordinal data, interval data and ratio data respectively. Variables measured on nominal and ordinal scale are qualitative or categorical variables while variables measured on interval and ratio scale are called quantitative or numeric variables. Measurement scales determines which statistical techniques are appropriate for data analysis.

Numbers used in statistics

Numbers used in statistics may be either meaningful numbers or numbers that are merely used as codes. Meaningful numbers are numbers that directly represent the measured or observed amount of some attribute of an element. E.g number of employees in a company, cost of production in a firm, weight of a product etc. Data in the form of meaningful numbers is called quantitative data and may be measured on interval or ratio scale. A ratio scale has a true zero point while an interval scale does not possess a true zero point.

Numbers that are merely used as codes represent categorical data or qualitative data. The numerical codes are not meant for mathematical computation. Categorical data may be measured on nominal scale or ordinal scale. If categories do not have a natural, meaningful order, then the data is measured on nominal scale e.g. the variable gender: the values male and female have no natural order; none is greater than the other. If categories have a meaningful order, the variables are measured on an ordinal scale e.g. categorization of employees as: 1) Most productive, 2) Productive, and 3) Least productive.

Properties of Numbers Used in Measurement

1. **Identity** – each value on the measurement scale has a unique meaning. Numbers are used to identify and classify elements.
2. **Order**. Numbers or values on the measurement scale have an ordered relationship with one another. That is, one value is greater than or less than another. Numbers indicate the relative positions of elements but the difference between any pair of numbers is not meaningful. We cannot determine, exact difference between values.
3. **Equal Intervals.** The difference between any pair of numbers is meaningful. This means that numerically equal differences between numbers represent equal differences in the characteristic being measured. Differences between numbers on a measurement scale are meaningful if the scale has a unit of measurement. A difference of 1 is the same amount throughout the entire scale. For example, the difference between 4 and 5 would be equal to the difference between 10 and 11. A Temperature of 20^0 is more than 40^0, 60^0 is more than 40. The difference between both examples is 20 degrees. The scale does not contain

a true zero value. Zero value does not mean that there is no temperature. It means it is very cold.

4. **Origin or True Zero Point**. This is an absolute and meaningful zero point. Assigning a zero indicate an absence of the characteristic being measured. If a scale has a true zero point, no values exist below zero. Therefore, ratios are meaningful.

Nominal Scale of Measurement

A nominal scale is a scale that classifies objects into mutually exclusive and exhaustive categories in which no ranking is implied i.e. the order in which the categories are presented is not important. The categories cannot be ranked because they have no underlying numeric value. e.g. the variable gender is measured on nominal scale: the values male and female have no natural order; none is greater than the other.

Ordinal Scale

Ordinal scale classifies elements into distinct categories in which ranking is implied but differences between the ranks are not meaningful. In ordinal scale, numbers represent greater than or less than relationships. For example, five products may be ranked by a customer as 1, 2, 3, 4, and 5 where 5 is the best and 1 is the worst. In this scale, we do not know how much better one product is than others, we only know that it is better i.e. differences between numbers are not meaningful because ordinal scale has no unit of measurement. Numbers are used to place objects in ordered categories but differences between consecutive numbers along the scale are not equal.

Ordinal scale measures non-numeric concepts such as happiness, satisfaction, preference etc.

Interval Scale: This is a scale on which equal differences between numbers represent equal differences in the characteristic being measured. The difference between numbers is meaningful because the scale has a constant unit of measurement. However, the ratio of two numbers is not meaningful because the scale does not possess a true zero point. Zero is just another point on the scale and does not mean the absence of the characteristic being measured.

An example of interval scale is temperature on Fahrenheit thermometer. If two objects have temperatures of 20^0 and 35^0 respectively, it implies that the second object is hotter than the first by 15^0 . Also, the difference between 20^0 and 35^0 is the same as the difference between 40^0 and 55^0. A temperature of 0^0 does not mean the absence of heat. However, 10^0F is not twice as hot as 5^0F. Other examples of data measured on interval scale are time on 24 or 12 our clock, calendar years and shoe size are based on interval scale. The difference between 7.00am and 8.00am is the same as the difference between 9.00am and 10.00am. The interval scale has a unit of measurement that permits us to describe how much more or less one object possesses than another.

Ratio Scale

A ratio scale possesses a unit of measurement and there exists a true zero point. Zero indicates complete absence of the characteristic being measured. Because the numbers begin at an absolute zero-point, true ratios of data values exist when the same variable is measured on two different members of the population. Examples of variables measured on ratio scale include: Weight, height, length of time, units produced, salaries etc. Suppose that two people earn sh. 3000 and sh. 6000 respectively. $6000/3000 = 2$ means that the second person earns twice as much as the first person.

Properties Possessed by Different Scales of Measurement

Scale of measurement/Property	**Nominal Scale**	**Ordinal Scale**	**Interval Scale**	**Ratio Scale**
Identity	✓	✓	✓	✓
Order		✓	✓	✓
Equal intervals			✓	✓
True zero present				✓

Each property can be described by answering the following questions.

1. Does a number identify or categorize elements?
2. Does a larger number indicate a greater value than a smaller number?
3. Does subtracting two numbers represent some meaningful value?
4. Does dividing (or taking the ratio of) two numbers represent some meaningful value.

Meaning of Statistics in Singular

Statistics is a science that focusses on the methods involved in collecting, organizing, summarizing, analyzing, interpreting and presenting a set of data; and making inferences about a population based on information contained in a sample taken from that population. Statistical methods are used to convert data into useful information. Business statistics is a collection of methods that are used to convert data into meaningful information in a business environment.

Branches of Statistics

There are **two** branches of statistics: Descriptive Statistics and Inferential Statistics

Definition of terms used in Descriptive Statistics and Inferential Statistics

Population

The term population may refer to things as well as people. A population is a set of objects with a common characteristic which consists of all objects that a researcher is interested in studying. Examples of populations include: all employees in a given company, all students in a university, all registered voters in a county, all potential buyers of a new product, all households in a country, all items manufactured in an assembly line during a certain period, all securities traded at a stock exchange market, all fish in Indian Ocean, all birds of the air etc.

When researchers gather data from the whole population for a given variable of interest, they call it a census study. Most of the time due to the expense, time, size of the population, medical concern etc, it is not possible to use the entire population for a statistical study; therefore, researchers use samples.

Sample

A sample is a group of elements selected from a population to represent the entire population. It is a portion of the whole population that is examined in order to make conclusions about the population from which the sample is drawn e.g. we can use a sample of 2000 people to estimate the voting behavior of 30 million voters, 200 students can be selected to represent a university of 10,000 students.

A Parameter – is a numerical measure that describes a specific characteristic of a population e.g. average age of all TUM students, percentage of all registered voters in a country who are likely to vote for a certain candidate. Percentage of all potential buyers of a new product who have actually bought the product. A parameter is a numerical property of a population that describes population data.

A Statistic – is a numerical measure that describes a specific characteristic of a sample e.g. average age of 100 students selected from TUM students. It is used to estimate the value of the corresponding population parameter. A statistic is a quantity that is calculated from sample data.

For large or infinite populations, a parameter is an unknown value for the entire population. Therefore, it has to be estimated using a sample that is representative of the entire population.

Suppose we want to know the average weight of people living in Kenya. We can't use the weights of all people in Kenya. Instead, we use a sample that is representative of the entire population to estimate the average weight of Kenyans.

Descriptive Statistics

This is a branch of statistics that consists of methods that are used to describe the characteristics of a data set. The methods can be applied to a set of population data or a set of sample data without making a specific distinction as to which is involved. The characteristics of a data set that we seek to describe using descriptive statistics include:

i. Central tendency or location or centre of the data set. Central tendency is the property of a data set having a central point around which most data values tend to cluster. A measure of central tendency is the value that is located at the centre of a data set. It is the value around which most observations in the data set are clustered.

ii. Spread or dispersion or the variability of the values in the data set. This is the extent to which values in the data set are clustered or scattered. Variability tells whether the values of a data set are spread over a wide range or are clustered together.

iii. Shape of the distribution or data set. There are two aspects of shape of a distribution:

  a. Symmetry Vs. Skewness. This is the extent to which the right-hand side of a distribution is a mirror image of the left-hand side of the distribution if the data has been arranged in ascending order.

  b. Kurtosis – this is the extent to which a distribution, given its spread, is flat versus peaked.

iv. Relationship between variables in a data set.

Methods used to describe data include tabular methods, graphical methods and numerical methods.

1) **Graphical Methods** – Graphs are used to describe the observed data.
2) **Tabular Methods** – Tables are used to describe the observed data.
3) **Numerical Methods** – Numerical measures are computed from the observed data to describe the data. These measures include:

   I. Measures of central tendency such as the mean, median and mode.

II. Measures of variability such as the range, interquartile range, standard deviations and variance.

III. Measure of skewness such as Pearson coefficient of skewness.

IV. Measures of Kurtosis

V. Measures of correlation such as covariance, Pearson coefficient of correlation, and spearman rank correlation coefficient.

Inferential Statistics

This is a branch of statistics that uses sample data (data collected from a small group) to make inferences or conclusions about the characteristics of a population (a large group). In inferential statistics we use the information obtained from the descriptive analysis of sample data to infer the properties of the population. That is, inferential statistics deals with methods that are used to find out something about a population, based on information obtained from a sample. Inferential statistics uses probability to make inferences from sample to population.