PROBABILIY AND STATISTICS I

LECTURE ONE

Introduction to Statistics

Lecturer: Dr. Emily Roche

INTRODUCTION

This lecture will focus on definition of terms, measurement Scales, data types, data collection methods and sampling methods.

Intended learning outcomes

At the end of this lecture, you will be able to explain the meaning of statistical terms, distinguish different types of data and measurement scales and illustrate the methods of data collection.

References

These lecture notes should be supplemented with relevant topics from the book listed in the Bibliography at the end of the lecture.

Statistics

Statistics is the science of data. It deals in collection of data, organization of data, presentation of data, analyzing data and interpretation of data.

Statistics may also be defined as numerical data which has been collected from a given source for a particular purpose

Statistics can also be considered as the science of "good" decision making under uncertainties based on some numerical and measurable scales. Decision making processes must be based on data, not on personal opinion nor on belief

- Data are the facts and figures that are collected, summarized, analyzed, and interpreted.

Objectives of statistics

The main objective of Statistics is to make inferences (e.g., prediction, making decisions) about certain characteristics of population based on information contained in a random sample from the entire population. The condition for randomness is essential to make sure the sample is representative of the population.

Application of statistics

- **Quality control** – Usually there is a quality control institutions in every government which is charged with the responsibility of ensuring that the manufactured products meet the customers standards. These institutions, together with other control departments have developed quality control charts which they use to check whether the products are up to standards or not.

- **Forecasting –** Statistics is very important when predicting the future in a particular situation. For instance if a given situation involves a dependent and independent variables one can develop an equation which can be used to predict the output under certain given conditions.

- **Human resource management** – Statistics may be used efficiently to understand the human capital in an organization thus creating a conducive work environment.

- **Accounting –** Public accounting firms use statistical sampling procedures when conducting audits for their clients.

- **Finance –** Financial analysts use a variety of statistical information, including price-earnings ratios (The P/E looks at the relationship between the stock price and the company's earnings. The P/E is the most popular metric of stock analysis, although it is far from the only one you should consider) to guide their investment recommendations.

- **Marketing –** Electronic point-of-sale scanners at retail checkout counters are being used to collect data for a variety of marketing research applications
- **Production –** A variety of statistical quality control charts are used to monitor the output of a production process
- **Economics –** Economists use statistical information in making forecasts about the future of the economy or some aspect of it

Definition of basic terms

Population is a set of all elements of interest in a particular study.

A **sample** is a subset of the population.

Statistic: A statistic is a quantity that is calculated from a sample of data (a descriptive measure of a sample).

Parameter: A descriptive measure of a population. It is an unknown value, and therefore it has to be estimated from a statistic.

Types of statistics

Statistics can be sub-divided into two basic areas:

Descriptive statistics – is the discipline of quantitatively describing the main features collected data. It aims at looking for the patterns of data and to present the information in convenient forms like tables, graphs, and any other numerical methods used to summarize data

Inferential statistics - is the process of using information obtained from analyzing a sample to make estimates about characteristics of the entire population. It mainly aims at drawing conclusions and/or making decisions concerning a population based on sample results. It includes estimation and hypothesis testing

The main difference between descriptive statistics and inferential statistics is that descriptive statistics aims at summarizing a data set, rather than use the data to learn about the population that the data are thought to represent.

Classification of data

Data are the facts and figures collected, summarized, analyzed, and interpreted. A collection of data is known as **Data set.**

Data can be classified into two distinct categories namely:

1. Based on source
2. Based on nature

1. Based on Sources of data

Data based on source is further classified as:

a. **Primary Data**

This is data that has been collected from first-hand-experience. Primary data has not been published yet and is more reliable, authentic and objective. Primary data has not been changed or altered by human beings, therefore its validity is greater than secondary data.
Examples include data collected through experiments and surveys among others.

Importance of Primary Data:
Validity: Validity is one of the major concerns in a research. Validity is the quality of a research that makes it trustworthy and scientific. It is the use of scientific methods in research to make it logical and acceptable. Using primary data in research can improve the validity of research. First hand information obtained from a sample that is representative of the target population yields data that is valid for the entire target population.
Authenticity: Authenticity is the genuineness of the research. Authenticity can be at stake if the researcher invests personal biases or uses misleading information in the

research. Primary research tools and data become more authentic if the methods chosen to analyze and interpret data are valid and reasonably suitable for the data type. Primary sources are more authentic because the facts have not been overdone. Primary source can be less authentic if the source hides information or alters facts due to some personal reasons. There are methods that can be employed to ensure factual yielding of data from the source.

Reliability: Reliability is the certainty that the research is true enough to be trusted. For example, if a research study concludes that junk food consumption does not increase the risk of cancer and heart diseases. This conclusion should have to be drawn from a sample whose size, sampling technique and variability is not questionable. Reliability with the use of primary data. In the similar research mentioned above if the researcher uses experimental method and questionnaires the results will be highly reliable. On the other hand, if he relies on the data available in books and on internet he will collect information that does not represent the real facts.

b. **Secondary Data**

This is data collected from a source that has already been published in any form. A very common example is the review of literature in many researches which is based on secondary data, mostly from books, journals and periodicals. Some sources of secondary data include Published Printed/electronic Sources (journals, books, magazines), websites and weblogs e.t c

Importance of Secondary Data:

Secondary data can be less valid but its importance still exists as listed below:

1. When it is difficult to obtain primary data in these cases getting information from secondary sources is easier and possible.
2. When primary data does not exist in such situation one has to confine the research on secondary data.
3. When primary data is present but the respondents are not willing to reveal it in such case too secondary data can suffice: for example, if the research is on the

psychology of transsexuals first it is difficult to find out transsexuals and second they may not be willing to give information needed for the research, so data must be collected from books or other published sources.

2. **Based on nature data**

Data based on nature is also further subdivided into qualitative or quantitative data.

a. **Qualitative data (categorical) –** Data that can be put in distinct partition or categories according to some defining the characteristics. The categories have no superiority attached to them.
   Examples: Eye color (blue, black), Gender (Male, female), Political affiliation, Blood group
b. **Quantitative data (Numerical) –** these are measures that take numerical values. They can be put into an order and further divided into two groups: discrete data or continuous data.
   - ***Discrete data*** are countable data and are collected by counting, for example, the number of defective items produced during a day's production, family size etc.
   - ***Continuous data*** – data that can be assigned an infinite number of values between whole numbers. They are collected by measuring and are expressed on a continuous scale. For example, measuring the height of a person, temperature etc.

Levels of Measurements (measures of scale)

Measurement scales are used to categorize and/or quantify variables

Properties of Measurement Scales

Each scale of measurement satisfies one or more of the following properties of measurement.

- **Identity –** Each value on the measurement scale has a unique meaning.

- **Magnitude –** Values on the measurement scale have an ordered relationship to one another. That is, some values are larger and some are smaller.
- **Equal intervals –** Scale units along the scale are equal to one another. This means, for example, that the difference between 1 and 2 would be equal to the difference between 19 and 20.
- **Absolute zero –** The scale has a true zero point, below which no values exist.

The measurement scales include:

Nominal Scale of Measurement

The nominal scale of measurement only satisfies the identity property of measurement. Values assigned to variables represent a descriptive category, but have no inherent numerical value with respect to magnitude.

Gender is an example of a variable that is measured on a nominal scale. Individuals may be classified as "male" or "female", but neither value represents more or less "gender" than the other. Religion and political affiliation are other examples of variables that are normally measured on a nominal scale.

Ordinal Scale of Measurement

The ordinal scale has the property of both identity and magnitude. Each value on the ordinal scale has a unique meaning, and it has an ordered relationship to every other value on the scale.

An example of an ordinal scale in action would be the results of a horse race, reported as "win", "place", and "show". We know the rank order in which horses finished the race. The horse that won finished ahead of the horse that placed, and the horse that placed finished ahead of the horse that showed. However, we cannot tell from this ordinal scale whether it was a close race or whether the winning horse won by a mile.

Interval Scale of Measurement

The interval scale of measurement has the properties of identity, magnitude, and equal intervals.

A perfect example of an interval scale is the Fahrenheit scale to measure temperature. The scale is made up of equal temperature units, so that the difference between 40 and 50 degrees Fahrenheit is equal to the difference between 50 and 60 degrees Fahrenheit.

With an interval scale, you know not only whether different values are bigger or smaller, you also know how much bigger or smaller they are. For example, suppose it is 60 degrees Fahrenheit on Monday and 70 degrees on Tuesday. You know not only that it was hotter on Tuesday, you also know that it was 10 degrees hotter.

Ratio Scale of Measurement

The ratio scale of measurement satisfies all four of the properties of measurement: identity, magnitude, equal intervals, and an absolute zero.

The weight of an object would be an example of a ratio scale. Each value on the weight scale has a unique meaning, weights can be rank ordered, units along the weight scale are equal to one another, and there is an absolute zero.

Absolute zero is a property of the weight scale because objects at rest can be weightless, but they cannot have negative weight.

Methods of collecting data

There are three main methods of data collections namely experiments, surveys and direct observation.

Experiments:

An experiment is defined as manipulation (changing value/ situations) of one or more independent variables to see how the dependent variable is affected.

Independent variables are those variables over which the researcher has control and wishes to manipulate.

Dependent variables are those over which the researcher has little or no direct control but has a string interest in testing.

Direct observations:

This is the systematic process of recording the behavioral patterns of people, objects and occurrences without questioning or communicating with them.

Observation is the main source of information in the field research. The researcher goes into the field and observes the conditions in their natural state.

Survey:

Surveys involve the selection and study of a sample of items from a population. Survey is the most commonly used method in social sciences, management, marketing and psychology to some extent. Surveys can be conducted in different methods.

- Questionnaire: is the most commonly used method in survey. Questionnaires are a list of questions open-ended or close -ended for which the respondents give answers. Questionnaire can be conducted via telephone, mail, live in a public area, or in an institute, through electronic mail or through fax and other methods.
- Interview: Interview is a face-to-face conversation with the respondent. In interview the main problem arises when the respondent deliberately hides information otherwise it is an in depth source of information. The interviewer can not only record the statements the interviewee speaks but he can observe the body language, expressions and other reactions to the questions too. This enables the interviewer to draw conclusions easily.
- Mail survey:

Bibliography

Gupta, SP (Dr.), (2014). *Statistical methods* (43rd Ed.). Sultan Chand & Sons.

S. C. Gupta and V. K. Kapoor, (2020). *Fundamentals of mathematical Statistics* (12th Ed). Sultan Chand & Sons.