Credits: Contents presented here have been taken from the book Introduction to Statistical Theory Part 1 by Professor Sher Muhammad Chaudhry

1 Introduction to Statistics

1.1 What is Statistics?

Statistics is the science that includes procedures and techniques used to collect, process and analyse numerical data to make inferences and to reach decisions in the face of uncertainty.

1.2 Why study statistics?

Because it assists in:

- summarizing larger sets of data in a form that is easily understandable.
- efficient design of laboratory and field experiments as well as surveys.
- sound and effective planning in any field of inquiry
- drawing general conclusions and in making predictions under given conditions.

Statistical techniques are used in almost every branch of learning, both in computer science related fields such as Data Science, Artificial Intelligence and other sciences such as Astronomy, Physics, Geology etc.

1.3 Branches of Statistics

1.3.1 Descriptive Statistics:

is the branch of statistics which deals with concepts and methods concerned with summarization and description of the important aspects of numerical data. This area of study consists of the condensation of data, their graphical displays, and the computation of a few numerical quantities that provide information about the center of the data and indicate the spread of the observations.

1.3.2 Inferential Statistics:

deals with procedures for making inferences about the characteristics that describe the larger group of data or the whole called the population from the knowledge derived from only a part of data, known as sample. This area includes the estimation of population parameters and testing of statistical hypothesis.

1.3.3 Populations and Samples:

A population is a collection or set of all possible observations whether finite or infinite, relevant to some characteristics of interest.

A statistical population may be real such as the height of all college students or hypothetical such as all the possible outcomes from the toss of a coin.

The number of observations in a finite population is called the size of the population and is denoted by the letter N. Numerical quantities describing a population are called parameters, represented by Greek letters.

A sample is a part or a subset of a population. Generally it consists of some of the observations but in certain situations, it may include the whole of the population.

The number of observations included in a sample is called the size of the sample and is denoted by the letter n.

A numerical quantity computed from a sample, is called a statistic, which is represented by Latin letter.

The information derived from the sample is used to draw conclusions about the population.

1.3.4 Observations and Variables

In statistics, an observation often means any sort of numerically recording of information, whether it is physical measurement such as height or weight; a classification such as heads or tails, or an answer to a question such as yes or no.

Variables: A characteristic that varies with an individual or an object, is called a variable. For example, age is a variable as it varies from person to person. A variable can assume a number of values.

Domain: The given set of all possible values from which the variable takes on a value is called its domain.

Constant: If the domain of a variable contains only one value, then the variable is referred to as a constant.

Variables may be classified into quantitative and qualitative according to the form of the characteristics of interest.

Quantitative: when a characteristic can be expressed numerically e.g. age, weight, income, or number of children.

Qualitative: if the characteristic is non numeric e.g. education, gender, eye color, intelligence, etc.

1.3.5 Discrete and Continuous Variables

A quantitative variable may be classified as discrete or continuous.

Discrete Variable: A discrete variable is one that can take only a discrete set of integers or whole numbers. It represents count data e.g. number of persons in a family, number of rooms in a house.

Continuous Variable: A variable is called a continuous variable if it can take on any value, fractional or integral within a given interval. i.e. its domain is an interval with all possible values without gaps. e.g. age of a person, height of a plant, temperature at a place, etc.

1.3.6 Variable Related Notations

A variable whether countable or measurable, is generally denoted by some symbol such as X or Y.

Xi, or Yj represents the ith or jth value of the variable.

1.4 Errors of Measurement

Experience has shown that a continuous variable can never be measured with perfect fineness because of certain habits and practices, methods of measurements, instruments used, etc. The measurements are thus always recorded correct to the nearest units and hence are of limited accuracy.

For example, if a student's weight is recorded as 60kg (correct to the nearest kilogram), his true weight in fact lies between 59.5 kg and 60.5 kg. Thus there is a difference between the measured value and the true value. This sort of departure from true value is technically known as the error of measurement.

If the observed value and the true value of a variable are denoted by x and $x + \epsilon$, then the difference $(x + \epsilon) - x$ is the error. This error involves the unit of measurement of x and is therefore called an absolute error.

An absolute error divided by the true value is called the relative error.

Thus $relativeerror = \frac{\epsilon}{x+\epsilon}$, which when multiplied by zero is percentage error.

Relative and percentage errors are independent of the units of measurement of x.

An error is said to be biased when the observed value is consistently and constantly higher or lower than the true value. Biased errors arise from the personal limitations of the observer, the imperfection in the instruments used or some other conditions which control the measurements.

These errors are not revealed by repeating the measurements.They are cumulative in nature,that is,the greater the number of measurements,the greater would be the magnitude of error. They are thus more troublesome.These errors are also called cumulative or systematic errors.

An error is said to be unbiased when the deviations,i.e.the excesses and defects,from the true value tend to occur equally often. Unbiased errors are revealed when measurements are repeated and they tend to cancel out in

the long run. These errors are therefore compensating and are also known as random errors or accidental errors.

A measurement free from all classes of errors is considered as an accurate measurement. This is why efforts are made to reduce the magnitude of errors to a minimum so that the level of accuracy at which the measurements are recorded,is increased. To achieve this end,a clear understanding of the meaning of significant digits and the process of rounding off the numbers is very important in statistical computations.

1.5 Significant Digits

Accuracy in measurements is related to significant digits.The significant digits in a number,are those that represent accurate and meaningful information. For instance,the number 35 representing a continuous variable has two significant digits. In recorded measurements,all digits except zeros are always significant For zeros,we may state as:

- Zeros are significant if they follow a decimal point and conclude a number,e.g.the measurement 2.500 has four significant digits.
- Zeros are non-significant when they follow a decimal point but commence a number,e.g. the measurements.04 and .000237 contain only 1 and 3 significant digits respectively.
- Zeros may or may not be significant when they lie entirely to the left of the decimal point,where they may not represent measurement but may be used to simply locate the decimal point. In such a case,a definite specification such as standard notation,becomes necessary. When any number is expressed as a product of a power of 10 and a number between 1 and 10,it is said to be written in standard notation.For example,the number 75400 can have 3 significant digits when written in standard notation as 7.54 x 104, It can also have 5 significant digits if written as 7.5400 x 104.
- Zeros are always significant when they occur within a series of significant digits,e.g.the numbers 20.3..1001,4.00507,etc.,have 3.4 and 6 significant digits respectively.

It should be remembered that:

(a) significant digits in a number are not disturbed by the location of the decimal point, e.g. the measurements recorded as 269., 26.9, .269 or .000269 have only 3 significant digits;

(b) in case of discrete data which are generated by the process of counting, the number of significant digits is considered indefinite because

the level of accuracy cannot be improved, e.g. the number 15700 has indefinite significant digits;

(c) the rules regarding the determination of the number of significant digits, are applicable to continuous variables;

(d) in the operations of addition and subtraction, all digit positions which are not significant in any of the values being added or subtracted, are not significant in the total or difference;

(e) in the operations of multiplication and division, the number of significant digits in the result is determined by the value with the smallest number of significant digits that enters into the calculations.

1.6 Rounding off a Number

The process of rounding off or simply rounding a number means that a certain number of digits counted from the left, are to be retained and the last few digits are to be (i)dropped in a decimal number or (ii) replaced with zeros in a whole number.

The rules generally used for rounding decimal numbers are as follows:

(i) The last significant digit is increased by 1, if the first digit of the remainder to be dropped is more than 5 or is 5 followed by digits not all of which are zero, e.g. the numbers 2.145001 and 5.3772 are rounded off to three significant digits as 2.15 and 5.38 respectively.

(ii) The last significant digit remains unaltered, if the first digit of the remainder to be dropped is 4 or less, e.g. the numbers 2.154 and 7.3627 are rounded off to three significant digits as 2.15 and 7.36 respectively.

(iii) When the digit to be dropped is exactly 5, the accepted practice is to increase the last significant digit by 1, if it is odd and to leave unaltered if it is even, e.g. the number 4.535 and 2.745 are rounded off to three significant digits as 4.54 and 2.74 respectively.

For rounding whole numbers, we can change the word "the first digit to be dropped" to "the first digit to be replaced by zero" in the rules stated above.

The point to be made here is that the rules for identifying significant digits and the process of rounding the numbers should be applied to final calculations and not to the intermediate results.

1.7 COLLECTION OF DATA

The most important part of statistical work is perhaps the collection of data. Statistical data are collected either by a complete enumeration of the whole

field, called census, which in many cases would be too costly and too time consuming as it requires large number of enumerators and supervisory staff, or by a partial enumeration associated with a sample which saves much time and money. The sampling methods explained at length in later chapters, are increasingly employed both in official and in private inquiries to collect data.

When data are classified according to source, it is customary to make he following distinction.

Data that have been originally collected (raw data) and have not undergone any sort of statistical treatment, are called Primary data, while data that have undergone any sort of treatment by statistical methods at least once, i.e. the data have been collected, classified, tabulated or presented in some form for a certain purpose, are called Secondary data.

A brief description of the methods generally adopted either on census basis or on sample basis for collecting data, is given below.

1.7.1 Collection of Primary Data

One or more of the following methods are employed to collect primary data:

(i) Direct Personal Investigation. In this method, an investigator collects the information personally from the individual concerned. Since he interviews the informants himself, the information collected is generally considered quite accurate and complete. This method may prove very costly and time consuming when the area to be covered is vast. However it is useful for laboratory experiments or localized inquires. Errors are likely to enter the results due to personal bias of the investigator.

(ii) Indirect Investigation or Personal Interviews. Sometimes the direct sources do not exist or the informants hesitate to respond for some reasons or other. In such a case, third parties or witnesses having information are interviewed. As some of the informants are likely to deliberately give wrong information, so the reliance is not placed on the evidence of one witness only. Moreover, due allowance is to be made for the personal bias. This method is useful when the information desired is complex or there is reluctance or indifference on the part of the informants. It can be adopted for extensive inquires.

(iii) Collection through Questionnaires. A questionnaire is an inquiry form comprising of a number of pertinent questions with space for entering information asked. The questionnaires are usually sent by mail and the informants are requested to return the questionnaires to the investigator, after doing the needful within a certain period. This method is cheap, fairly expeditious and good for extensive inquiries. But the difficulty is that the majority of respondents (persons who are required to

answer the questions) does not care to fill the questionnaires in and to return them to the investigators. Sometimes, the questionnaires are returned incomplete and full of errors. In spite of these drawbacks, the method is considered as the standard method for routine business and administrative inquiries. The answers to the questionnaires are very often recorded by trained enumerators to overcome the difficulties these days. It is important to note that the questions should be few, brief, very simple, easy for, all respondents to answer, clearly worded and not offensive to certain respondents.

(iv) Collection through Enumerators. Under this method, the information is gathered by employing trained enumerators who assist the informants in making the entries in the schedules or questionnaires correctly. This method gives the most reliable' information if the enumerator is well trained, experienced and tactful. It is considered the best method when a large scale governmental inquiry is to be conducted. This method cannot be adopted by private individual or institution as its cost would be prohibitive to them.

(v) Collection through Local Sources. In this method, there is no formal collection of data but the agents or local correspondents are directed to collect and to send the required information, using their own judgment as to the best way of obtaining it. This method is cheap and expeditious, but gives only the estimates.

1.7.2 Editing of Data

The primary data should be intensively checked at an early stage in order to locate incomplete or inconsistent entries. If possible, the incomplete and defective questionnaires should be returned to the respondents for amendments.

1.7.3 Collection of Secondary Data.

The secondary data may be obtained from the following sources:

Official, e.g. the publications of the Statistical Division, Ministry of Finance, the Federal and Provincial Bureaus of Statistics, Ministries of Food, Agriculture , Industry, Labor, etc. Semi-Official, e.g. State Bank of Pakistan, Railway Board, Central Cotton Commission, Boards of Economic Inquiry, District Councils, Municipalities, etc. Publications of Trade Associations, Chambers of Commerce, etc. Technical and Trade journals and newspapers. Research organisations such as universities, and other institutions. In order to accept the secondary data as authoritative, one should critically examine the reliability of the compiler and the suitability of the

data. The scope and object of the inquiry, sources of information and the degree of accuracy should also be carefully scrutinized.

link to related video lecture