Statistics

Tushar B. Kute,
http://tusharkute.com

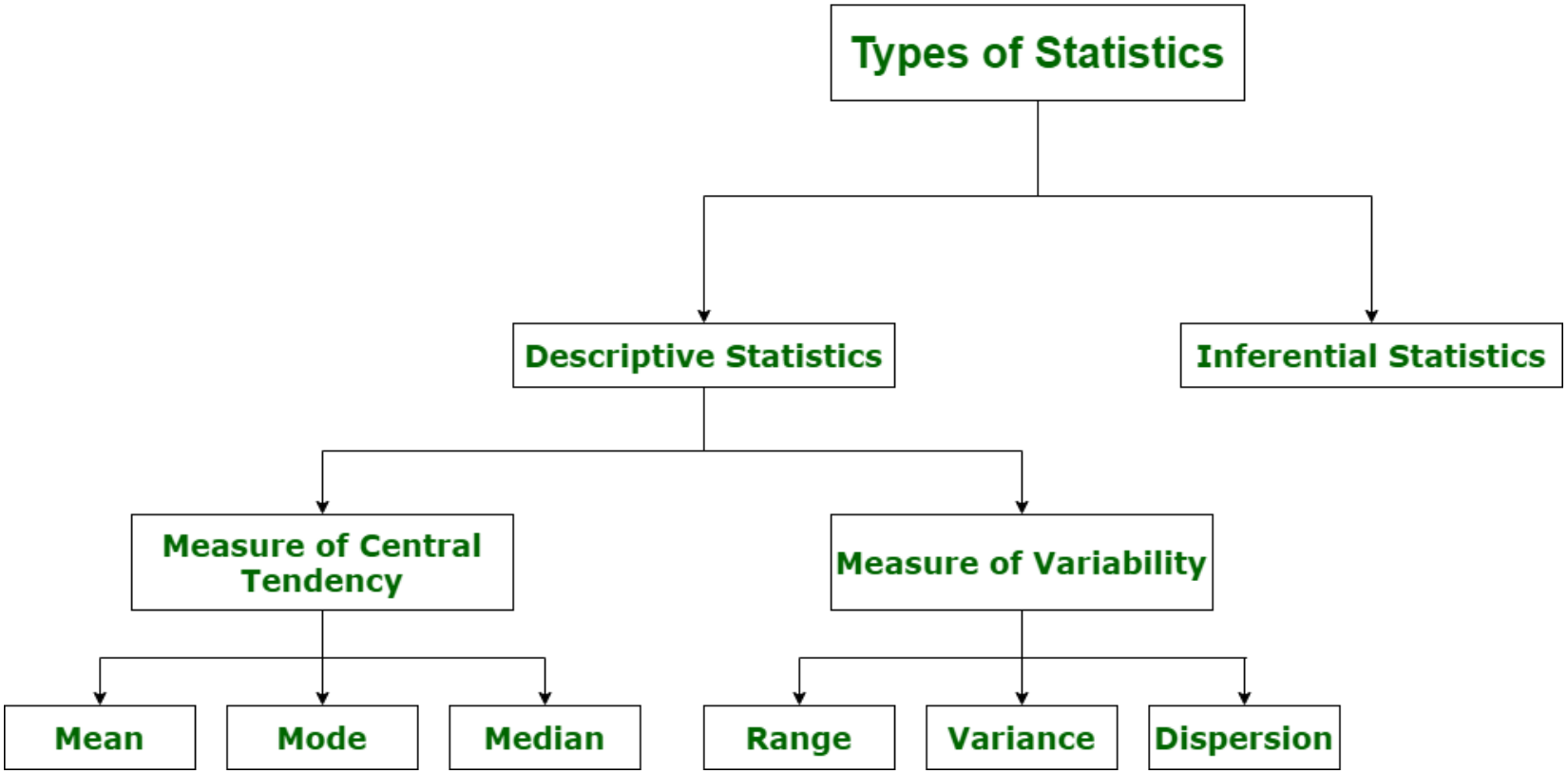
What is statistics?

- Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.
- In applying statistics to a scientific, industrial, or social problem, it is conventional to begin with a statistical population or a statistical model to be studied.
- Populations can be diverse groups of people or objects such as "all people living in a country" or "every atom composing a crystal". Statistics deals with every aspect of data, including the planning of data collection in terms of the design of surveys and experiments

What is statistics?

- When census data cannot be collected, statisticians collect data by developing specific experiment designs and survey samples.
- Representative sampling assures that inferences and conclusions can reasonably extend from the sample to the population as a whole.
- An experimental study involves taking measurements of the system under study, manipulating the system, and then taking additional measurements using the same procedure to determine if the manipulation has modified the values of the measurements.
- In contrast, an observational study does not involve experimental manipulation.

Types of statistics?



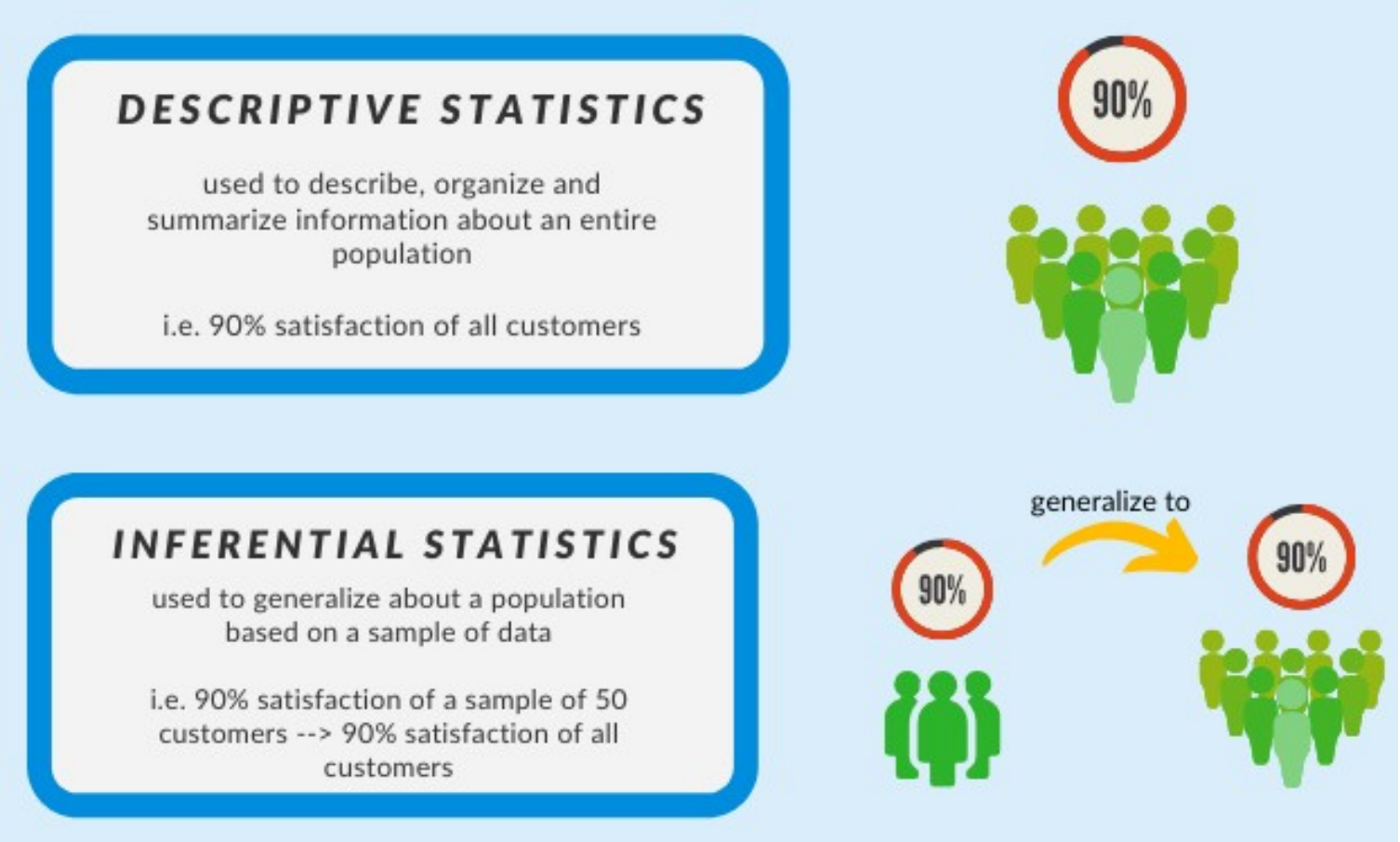
Descriptive Statistics

- A descriptive statistic (in the count noun sense) is a summary statistic that quantitatively describes or summarizes features of a collection of information, while descriptive statistics in the mass noun sense is the process of using and analyzing those statistics.
- Descriptive statistics is distinguished from inferential statistics (or inductive statistics), in that descriptive statistics aims to summarize a sample, rather than use the data to learn about the population that the sample of data is thought to represent.

Inferential Statistics

- Statistical inference is the process of using data analysis to deduce properties of an underlying probability distribution.
- Inferential statistical analysis infers properties of a population, for example by testing hypotheses and deriving estimates.
- It is assumed that the observed data set is sampled from a larger population. Inferential statistics can be contrasted with descriptive statistics.
- Descriptive statistics is solely concerned with properties of the observed data, and it does not rest on the assumption that the data come from a larger population.

Comparing


DESCRIPTIVE STATISTICS
used to describe, organize and summarize information about an entire population
i.e. 90% satisfaction of all customers
90%
INFERENTIAL STATISTICS
used to generalize about a population based on a sample of data
i.e. 90% satisfaction of a sample of 50 customers --> 90% satisfaction of all customers
generalize to
90%
90%

Why Statistics?

- What features are the most important?
- How should we design the experiment to develop our product strategy?
- What performance metrics should we measure?
- What is the most common and expected outcome?
- How do we differentiate between noise and valid data?

From data to knowledge

- In isolation, raw observations are just data. We use descriptive statistics to transform these observations into insights that make sense.
- Then we can use inferential statistics to study small samples of data and extrapolate our findings to the entire population.

Terminologies of statistics

- Population: It is an entire pool of data from where a statistical sample is extracted. It can be visualized as a complete data set of items that are similar in nature.
- Sample: It is a subset of the population, i.e. it is an integral part of the population that has been collected for analysis.
- Variable: A value whose characteristics such as quantity can be measured, it can also be addressed as a data point, or a data item.

Terminologies of statistics

- Distribution: The sample data that is spread over a specific range of values.
- Parameter: It is a value that is used to describe the attributes of a complete data set (also known as ‘population’). Example: Average, Percentage
- Quantitative analysis: It deals with specific characteristics of data- summarizing some part of data, such as its mean, variance, and so on.
- Qualitative analysis: This deals with generic information about the type of data, and how clean or structured it is.

Statistical Inference

"The use of a sample of data to draw inferences or conclusions about some aspect of the situation from which the data were taken."

Statistical Inference

- Statistical inference is the process of analysing the result and making conclusions from data subject to random variation.
- It is also called inferential statistics. Hypothesis testing and confidence intervals are the applications of the statistical inference.
- Statistical inference is a method of making decisions about the parameters of a population, based on random sampling.
- It helps to assess the relationship between the dependent and independent variables.
- The purpose of statistical inference to estimate the uncertainty or sample to sample variation.

Statistical Inference

- It allows us to provide a probable range of values for the true values of something in the population.
- The components used for making statistical inference are:
  - Sample Size
  - Variability in the sample
  - Size of the observed differences

Statistical Inference Procedure

- The procedure involved in inferential statistics are:
  - Begin with a theory
  - Create a research hypothesis
  - Operationalize the variables
  - Recognize the population to which the study results should apply
  - Formulate a null hypothesis for this population
  - Accumulate a sample from the population and continue the study
  - Conduct statistical tests to see if the collected sample properties are adequately different from what would be expected under the null hypothesis to be able to reject the null hypothesis

Statistical Inference Solution

- Statistical inference solutions produce efficient use of statistical data relating to groups of individuals or trials.
- It deals with all characters, including the collection, investigation and analysis of data and organizing the collected data.
- By statistical inference solution, people can acquire knowledge after starting their work in diverse fields. Some statistical inference solution facts are:
  - It is a common way to assume that the observed sample is of independent observations from a population type like Poisson or normal
  - Statistical inference solution is used to evaluate the parameter(s) of the expected model like normal mean or binomial proportion

Statistical Inference Solution

- An example of statistical inference is given below.

Suit	Spade	Clubs	Hearts	Diamonds
No.of times drawn	90	100	120	90

- Question: From the shuffled pack of cards, a card is drawn. This trial is repeated for 400 times, and the suits are given below:
- While a card is tried at random, then what is the probability of getting a
  - Diamond cards
  - Black cards
  - Except for spade

Statistical Inference Solution

- By statistical inference solution,

  Total number of events = 400

  i.e.,90+100+120+90=400

- (1) The probability of getting diamond cards:
  - Number of trials in which diamond card is drawn = 90
  - Therefore, P(diamond card) = 90/400 = 0.225

Statistical Inference Solution

- (2) The probability of getting black cards:

  Number of trials in which black card showed up = 90+100 =190

  Therefore, P(black card) = 190/400 = 0.475

- (3) Except for spade

  Number of trials other than spade showed up = 90+100+120 =310

  Therefore, P(except spade) = 310/400 = 0.775

Statistical Machine Learning

- The methods used in statistics are important to train and test the data that is used as input to the machine learning model. Some of these include outlier/anomaly detection, sampling of data, data scaling, variable encoding, dealing with missing values, and so on.
- Statistics is also essential to evaluate the model that has been used, i.e. see how well the machine learning model performs on a test dataset, or on data that it has never seen before.
- Statistics is essential in selecting the final and appropriate model to deal with that specific data in a predictive modelling situation.
- It is also needed to show how well the model has performed, by taking various metrics and showing how the model has fared.

Descriptive Statistics

- Descriptive statistics summarize and organize characteristics of a data set.
- A data set is a collection of responses or observations from a sample or entire population.
- In quantitative research, after collecting data, the first step of statistical analysis is to describe characteristics of the responses, such as the average of one variable (e.g., age), or the relation between two variables (e.g., age and creativity).

Descriptive Statistics – Types

- There are 3 main types of descriptive statistics:
  - The distribution concerns the frequency of each value.
  - The central tendency concerns the averages of the values.
  - The variability or dispersion concerns how spread out the values are.
- You can apply these to assess only one variable at a time, in univariate analysis, or to compare two or more, in bivariate and multivariate analysis.

Descriptive Statistics – Example

- You want to study the popularity of different leisure activities by gender. You distribute a survey and ask participants how many times they did each of the following in the past year:
  - Go to a library
  - Watch a movie at a theater
  - Visit a national park
- Your data set is the collection of responses to the survey.
- Now you can use descriptive statistics to find out the overall frequency of each activity (distribution), the averages for each activity (central tendency), and the spread of responses for each activity (variability).

Reference: https://www.scribbr.com

Thank you

This presentation is created using LibreOffice Impress 5.1.6.2, can be used freely as per GNU General Public License

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com