AI Machine Learning L1 -Applied Statistics

Cutting Edge Program-AA Team

Anilkumar T G

Principal Consultant

9th Jun 2017

Agenda

1. Introduction to Applied Statistics
2. Descriptive Statistics
3. **Inferential Statistics**
4. Probability and Statistics Practice with Python
5. Class room Exercise

Introduction to Applied Statistics

Introduction

- Statistics deals with study of numerical data.
  - Involves in collection, calculation, presentation and analysis of vast numerical data
  - Helps to draw conclusion from various analysis methods

- Statistics can be broadly classified into
  - Descriptive Statistics
  - Inferential Statistics

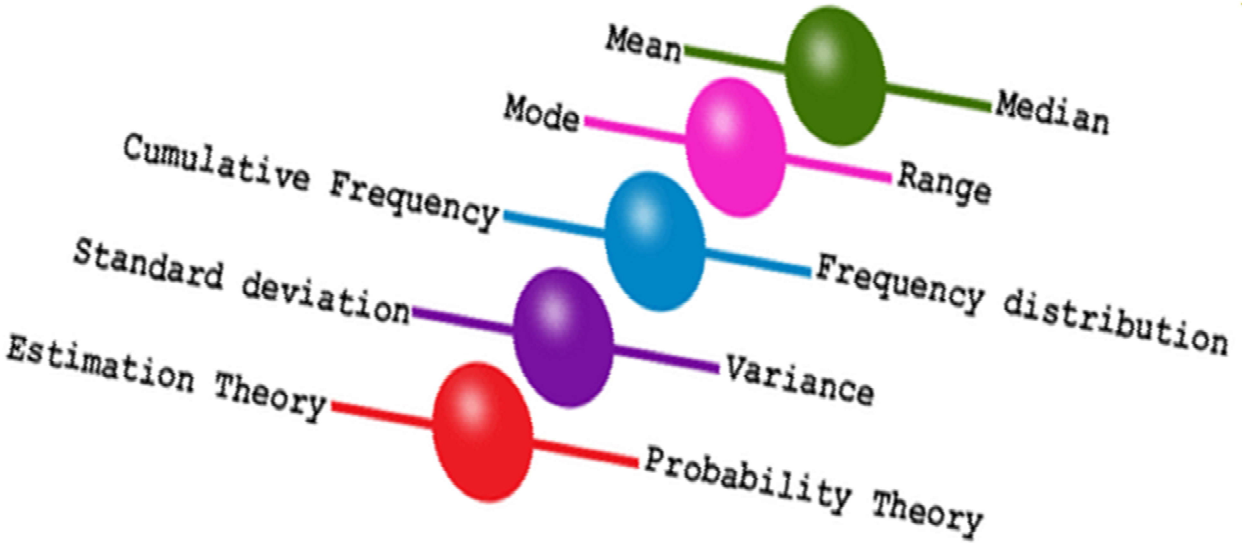
- Descriptive Statistics provides the features quantitatively on the sample data

- Inferential Statistics provides inferences for the total sampled data

- Although both of these methods give different views of the given data set but together provide strong result on the data analysis

Important Statistics Topics

- Statistics is not just plots or graphs or tabulation of data
  - Need to understand the following important topics
  - Mean & Median, Mode & Range, Cumulative Frequency & Frequency distribution, Standard deviation & Variance, Estimation Theory & Probability Theory



Mean & Median

- **Sample Mean Formula**
  - **If x1,x2,.....xn are the N values of a variate X,**
  - **the Arithmetic Mean ($\bar{X}$) can be defined as the sum of all the N values divided by the total numbers**

$$\bar{X} = \frac{1}{N}[x_1 + x_2 + \ldots\ldots + x_n] = \frac{\sum X}{N}$$

$$\text{Mean} = \frac{\sum X}{N}$$

Where,

X = Values of the variables

$\sum$ = Total of the values and read as summations

N = Number of items.

Mean & Median

- **Weighted Mean**
  - **Assign various weights to some individual values.**
- **If X = x1,x2,x3,.....,xn,**

  W = w1,w2,w3,.....,wn Is a set of non negative weights, derived from formula

Weighted Mean

$$\bar{x}_w = \frac{w_1 x_1 + w_2 x_2 + \ldots\ldots\ldots\ldots + w_n x_n}{w_1 + w_2 + \ldots\ldots + w_n}$$

$$\bar{x}_w = \frac{\sum WX}{\sum W}$$

Where,

$\bar{x}_w$ = Weighted arithmetic mean

x = Values of the items

w = Weight of the item .

- **If all the weights are equal, then the weighted mean is the same as the arithmetic mean**

Mean & Median

- Harmonic Mean
  - It is ratio of total number of given values to the summation of the reciprocals of the given set of values
  - It is also called as lowest mean

In a sample set of size N, where X1,X2,X3,...,Xn are the values collected. The harmonic mean of the sample can be given as

Harmonic Mean

$$H = \frac{N}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Mean & Median

- **Geometric Mean**
  - **The nth root of the product of a set of data gives the Geometric mean.**

The formula for geometric mean for a set of values a1,a2,.....,an is given below

$$\text{Geometric Mean} = \left(\prod_{i=1}^{n} a_i\right)^{\frac{1}{n}} = \sqrt[n]{a_1 a_2 \ldots\ldots\ldots\ldots a_n}$$

Mean & Median

- **Mean Deviation**
  - **It is a measure of difference between the actual mean value and the observed value.**
  - **The mean deviation is the mean of the absolute deviations that are observations of some desired average. And this desired average may be from arithmetic mean, the median or the mode.**
  - **The formula for mean deviation (MD) is as follows:**
  - **Mean Deviation = $\sum |X - \bar{X}| \; N$**

Mean & Median

- **Median**
  - **Median is a value that divides the distribution in two halves**
  - **Or it is considered as the middle value of a given set or distribution.**
  - **The formula for Median is defined as**

$$M = L + \frac{\frac{N}{2} - cf}{f} \times h$$

Here,

L = Lower limit of median class

cf = Cumulative frequency of class prior to median class.

f = Frequency of median class.

h = Class size.

Mean & Median

- Geometric Median
  - Generalizes the median
  - It has the property of minimizing the sum of distances for one-dimensional data
  - Gives a central tendency in higher dimensions
  - Also known as the Fermat–Weber point

For a given set of k points x1,x2,x3,.....,xn with each xi belongs to real field, the geometric median is defined as

$$\text{Geometric Median} = \arg\min \sum_{i=1}^{k} ||x_i - y||$$

Mode & Range

- **Mode**
  - **It is particular value that has occurred maximum times in the list**
  - **How to find the Mode?**
    - **1: Sort the given data set from least to greatest**
    - **2: Find out the value /values that has frequently occurred**

Mode Formula for Grouped Data:

$$\text{Mode} = L + \frac{(f_m - f_1)h}{2f_m - f_1 - f_2}$$

Where,

L = Lower limit of modal class

f_m = Frequency of modal class

f_1 = Frequency of class preceding the modal class

f_2 = Frequency of class succeeding the modal class

h = Size of class interval.

Mode & Range

- Range
  - It is the set of all possible output values when we substitute for possible values of x that is domain of a given function.
  - This provides the output of the function

- If we define Ra as the relation from a set X to the set Y .
- The range of Ra is n the set of all second components or coordinates of the ordered pairs belongs to Ra.

- For example : If X = { 1, 3, 5, 7 }, Y = { 2, 4, 6, 8, 10 } and Ra = { (1, 8), (3, 6), (5, 2), (1, 4) } is a relation from A to B then,

- The range of Ra = { 8, 6, 2, 4 }

Frequency Distribution

- It is the tabulation of the values with one or more variables
  - The orderly arrangement of data grouped as per the magnitude of the observations

- Types of frequency distribution
  - Relative frequency type
  - Relative cumulative frequency type
  - Grouped frequency type
  - Ungrouped frequency
  - Cumulative frequency

Standard deviation and Variance

- Standard deviation
- It is the measure of statistical data dispersion
  - Dispersion or variation is the property of the data to spread over a field
  - Dispersion measures the deviation of the data from its average or mean position.
  - The degree of dispersion or variation is calculated by the means of measures of variation

- There are many different measures of variation most common of those are listed below
  - Range
  - Mean deviation
  - Standard Deviation
  - Quartile deviation.

Standard deviation and Variance

Method to calculate the Standard Deviation:

- **1. First find out the arithmetic mean.**
- **2. Calculate the deviation of every item from the mean.**
- **3. Square the deviations and sum up.**
- **4. We obtain $\sum(x-\bar{x})2$**
- **5. The sum is divided by the total no., of items.**
- **6. Standard deviation is the under root of sum calculated in the above step.**

$$\Rightarrow S = \sqrt{\frac{\sum(X-M)^2}{N-1}}$$

Standard deviation and Variance

Population Variance:

- Population variance is calculated as

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

where n = Number of observations (population size)

$\bar{x}$ = Population mean

xi = Values of observations

Sample Variance:

- Sample variance is calculated as

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Where n = Number of observations (sample size)

$\bar{x}$ = Sample mean

xi = Values in data set

Estimation and Probability Theory

- Estimation is the process by which one makes inferences about a population, based on information obtained from a sample

- An estimate of a population parameter expressed in two ways
  - Point estimate of a population parameter is a single value of a statistic.
    - For example, the sample mean x is a point estimate of the population mean μ.
    - Similarly, the sample proportion p is a point estimate of the population proportion P
  - Interval estimate is defined by two numbers, between which a population parameter is said to lie.
    - For example, $a < x < b$ is an interval estimate of the population mean μ.
    - It indicates that the population mean is greater than a but less than b.

Estimation and Probability Theory

- **Probability Theory: It is a number expressing the chances for a specific event will occur**

Theoretical Probability Formula

$$P(\text{Event}) = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}}$$

$$\text{Experimental probability} = \frac{(Frequency\ of\ an\ outcome)}{(Total\ number\ of\ trials)}$$

- **Types of Probability:**
  - **Marginal type**
  - **Conditional type**
  - **Joint type**
  - **Union type**

Estimation and Probability Theory

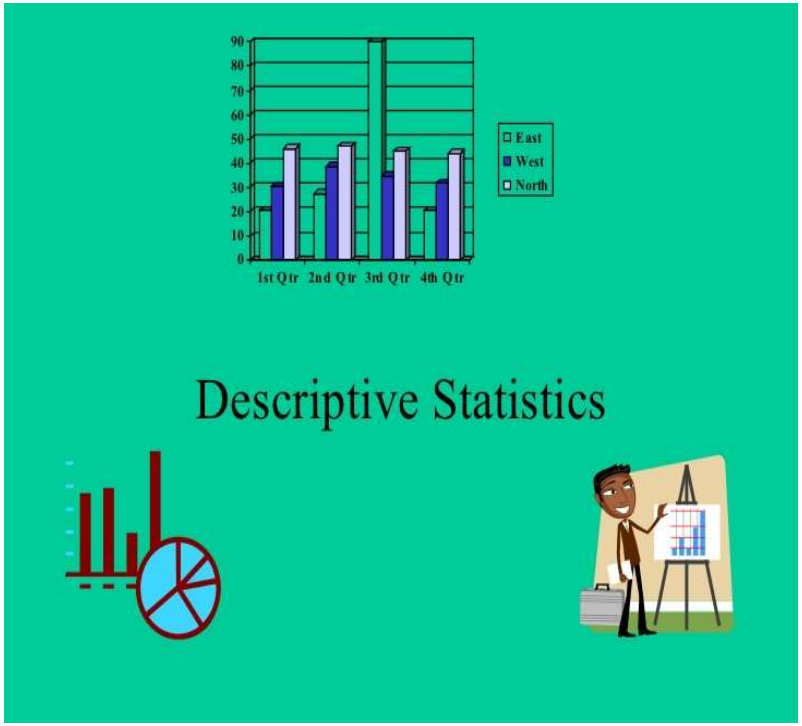
Terms associated with Probability:

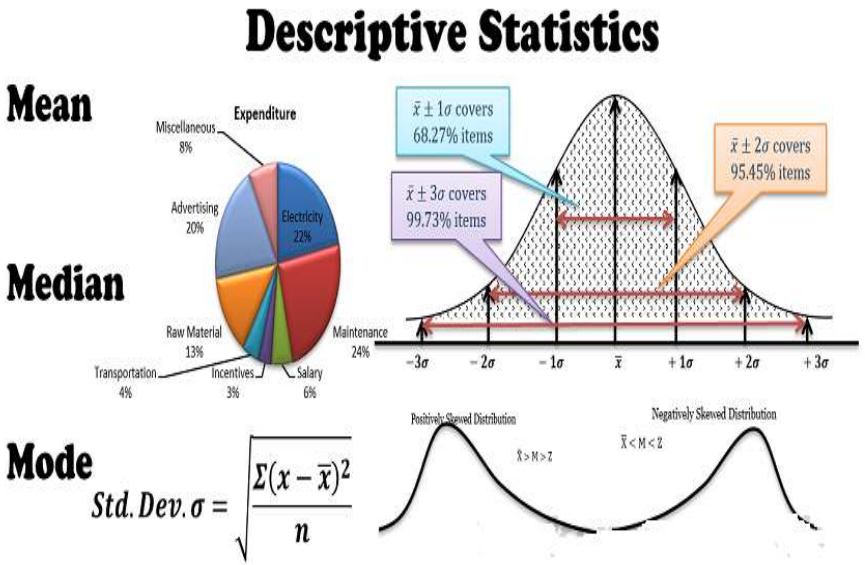
- **Experiment**
  - **A process to get an outcome**
    - **Example: Tossing a dice**
- **Sample Space**
  - **It is the set S of all possible outcomes from an experiment**
    - **Example: For a dice toss, S = 1,2,3,4,5,6**
- **Trial**
  - **Every observation made in experiment.**
- **Event**
  - **The subset of the sample space.**
- **Dependent Events**
  - **One event produces an effect on next event with certain degree of probability**
- **Independent Events**
  - **Two events which are mutually exclusive and no probable dependency occurs**
- **Outcome**
  - **The result of a every trial.**

Descriptive Statistics

Descriptive Statistics

- **Descriptive Statistics helps to describe in graphical or quantitative approach**
  - **Example:**
    - **The individual performance of each player in cricket**
    - **Performance of student in the examination**





Descriptive Statistics

- **Method uses to measure the Descriptive Statistics**
  - **Measuring Central Tendency**
    - It involves in measuring the data that lies in the center of a given frequency distribution
    - The primary measures are mean, median and mode
  - **Measuring Spread**
    - Measures of **spread** describe how similar or varied the set of observed values are for a particular variable (data item).
    - Measures of **spread** include the range, quartiles and the interquartile range, variance and standard deviation.
  - **Association or Graphical Representation**
    - Graphs of various types used for describing statistical data
    - histogram, bar graph, box and whisker plot, line graph, scatter plot, ogive, pie chart etc.

Inferential Statistics

Inferential Statistics

- Used for arriving at conclusion
- Analyses samples & infers about the prediction for the population
- It's a procedure for prediction & conclusion of data that are of random variation type
- Observational and sampling errors can be detected and predicted
- Works as test for hypothesis of given data and set the direction for "*what to do next*"
- Two type of Inferential Statistics
  - Confidence Interval
    - Its an interval that provides a range for the parameter for a given population
  - Hypothesis Test
    - It is tests of significance and provides evidence or claim for the population by analyzing sample.

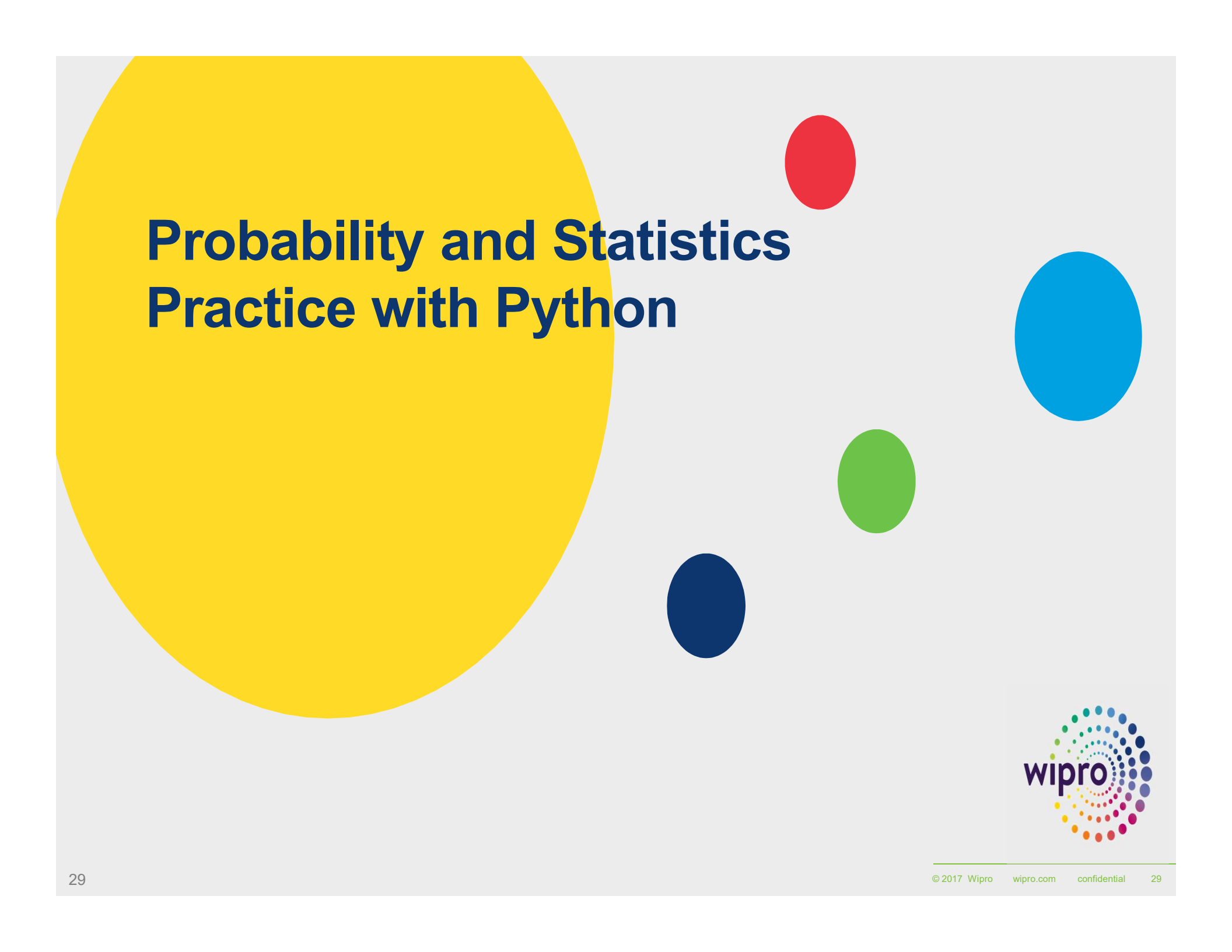
Comparison of Descriptive & Inferential Statistics

Descriptive	Inferential Statistics
• Description about the sample	• Predicts and Infers about larger data
• Describes some characteristics about a data	• Analyzes deeply and observes the statistical data
• Deal with central tendency & spread of distribution frequency	• Details about hypothesis tests and confidence interval
• Measures of this type are numbers	• Measures are not of exact numbers
• Deals with small samples resulting in higher accuracy	• Deals with larger sample of data to draw conclusion & hence lesser accuracy
• The conclusion cannot be made outside the given data	• Predictions and extrapolation can be made outside the sample data

Examples of Descriptive & Inferential Statistics

Descriptive	Inferential Statistics
• Number of students in a school	• Average marks obtained by all the students
• Population of a particular nation or city	• Floating population of a city or nation
• Estimation of cavity teeth by a dentist	• Prediction on number of teeth likely to go on cavity in future
• Average rainfall at particular place over an year	• Prediction of estimated rainfall in the coming year

Probability and Statistics Practice with Python



Thank you