Module 2

Data Analytics with Python – Statistics

Section: Math for Data Science – basic statistics

Basic Statistics & Linear Algebra

Statistics

The branch of mathematics that deals with the collection, analysis, interpretation, and presentation of masses of numerical data is called statistics.

The discipline that utilizes data samples to support claims about populations (a larger data set).Utilizes models, representations, and synopses for a set of experimental data and applies the results to a larger population.

Descriptive Statistics

It is the set of methods to describe the collected data. It aims to describe the characteristics of a sample data set and understanding the specific set of observations.

It summarizes and represents data using graphs, charts, and tables.

Descriptive statistics Involves following methods

- Measurement of central tendency
- Measurement of dispersion/spread
- Measurement of skewness and kurtosis
- Exploration of relationships of paired data

Inferential Statistics

It is a set of methods used to make a generalization, estimate, prediction, or decision.It aims to test a hypothesis and derive conclusions about a population based on the sample.

Its analysis results are generalized from a sample to a larger population.

Inferential statistics Includes the following analysis tools:

- Hypothesis tests
- Confidence Intervals
- Regression analysis

Linear Algebra

Linear Algebra is a branch of mathematics that is extremely useful in data science. Linear algebra is the most important math skill in machine learning. It is used in data preprocessing, data transformation, and model evaluation.

Most ML models can be expressed in matrix form. A dataset itself is often represented as a matrix.

Types Of Variables

For training machine to develop artificial intelligence model are relied on data. Data are categorized in different types as follows:

Data can be broadly divided into two types:

- Categorical Variable
- Numerical (Continuous) Variable

Categorical Variable

Provides information about the category of an object or information which cannot be measured. Variables that can be added into categories according to their characteristics.

Example: Performance of a student in terms of Good, Average and Poor, it falls under category of categorical data Also, name, roll no of students are information that cannot be measured using some scale of measurement. So, they would fall under categorical data.

Categorical data is also called qualitative data. It can be further divided into two types:

A. Nominal data
B. Ordinal data

a. Nominal Data

Data which has no numeric value but a named value.

- It is used for assigning named values to attributes
- Nominal values cannot be quantified

Examples

- Blood group: A, B, O, AB etc.
- Nationality: Indian, British, American etc.
- Gender: Male, Female, Other

b. Ordinal Data

In addition to possessing all properties of nominal data, can also be naturally ordered it's called ordinal data. Ordinal data assigns named values to attributes but unlike nominal data, they can be arranged in a sequence of increasing or decreasing value so that we can say whether a value is better or greater than another value.

Examples are:

- Customer satisfaction: Very Happy, Happy, Unhappy
- Grades: A, B, C etc.
- Hardness of materials; Very Hard, Hard, Soft etc.

Numerical Variable

It relates to information about the quantity of an object. It can be measured. For example, if we consider the attribute marks, it can be measured using a scale of measurement.

There are two types of Numerical data:

- Interval data
- Ratio data

Interval Data

Interval Data also called an integer, is defined as a data type which is measured along a scale, in which each point is placed at equal distance from one another.

- Interval data always appears in the form of numbers or numerical values where the distance between the two points is standardized and equal.
- Interval data is measured on an interval scale.
- Celsius and Fahrenheit are examples of interval scales. Each point on these scales differs from neighboring points by intervals of exactly one degree.
- The difference between 20 and 21 degrees is identical to the difference between 225 and 226 degrees
- Interval data doesn't have a defined absolute zero point. Because there's no true zero, you can't multiply or divide scores on interval scales. 30°C is not twice as hot as 15°C. Similarly, -5°F is not half as cold as -10°F.

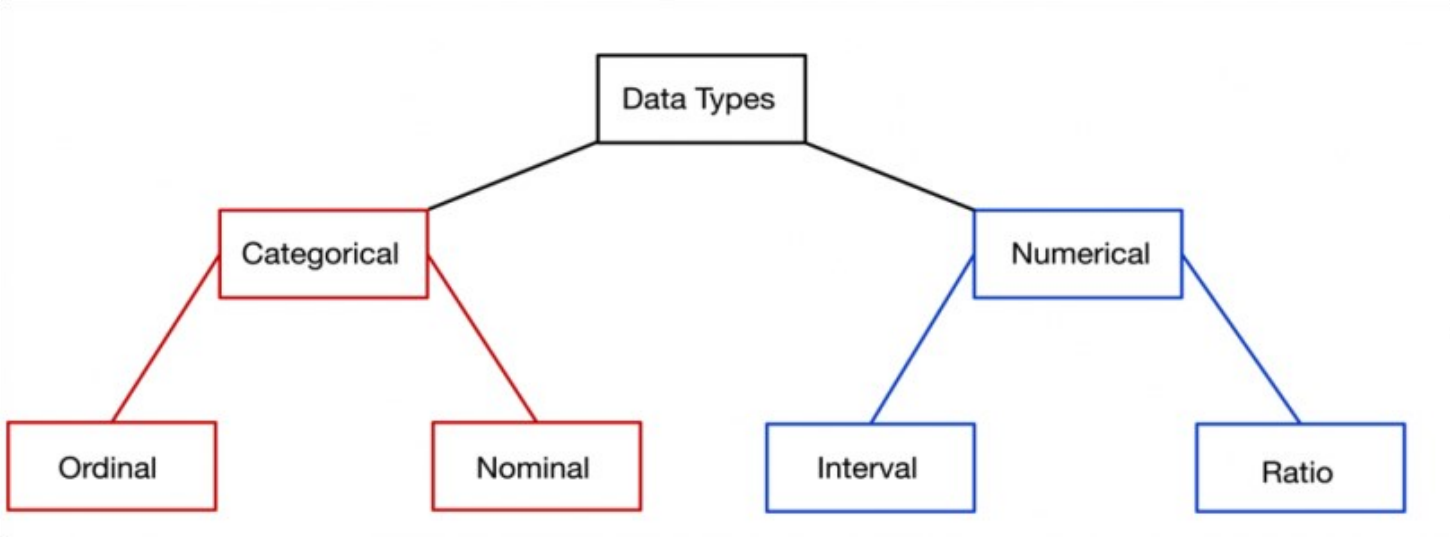
The lack of absolute point zero makes comparisons of direct magnitudes impossible. For example, Object A is twice as large as Object B is not a possibility in interval data. Central tendency (center or location of the distribution) can be measured by mean, median or mode.

Ratio data

- Represents numeric data for which exact value can be measured.
- Absolute zero is available for ratio data.
- These variables can be added, subtracted, multiplied, or divided.
- The central tendency can be measured by mean, median or mode and methods of dispersion such as standard deviation.

Examples: height, weight, age, salary etc.

- Attributes can also be categorized as discrete or continuous
- Discrete attributes can assume a finite or countably infinite number of values.
- Nominal attributes such as roll number, street number, pin code can have a finite number of values whereas numeric attributes such as count, rank of students can have countably infinite values.
- Special type of discrete attribute which can assume two values only is called binary attribute. Ex: male/female, positive/negative etc.
- Continuous attributes can assume any possible value which is a real number. Example: length, height, weight etc.



Terms Used in Statistics

Population: A collection of all persons, objects, or items.

Ex: All Automobiles, All employees of Microsoft

Census: When researchers gather data from whole population for a given measurement of interest, they call it census.

Ex: US Population Census taken every 10 years

Sample: A portion of whole or representative of the whole population.

Ex: 75 samples of Dairy milks for testing quality

Generally Statistical Methods are classified in two types

Descriptive Statistics

Measures of Central Tendency

The measure of central tendency is the commonly used method to identify the centre of the data set. Used to describe a data set by identifying the central position within that data set.

Measures of central tendency

- Mean
- Median
- Mode

Mean

The most popular measure of central tendency. It is also referred to as the average value in the data set. Defined as the sum of all values in a data set, divided by the total number of values in a data set.

Example: Let x1, x2, x3, ……. Xn be the values in the data set, and there are 'n' number of values in the data set

Mean=

$$\frac{x_1 + x_2 + x_3 + \ldots\ldots + x_n}{n} \quad \text{or} \quad \frac{\sum x}{n}$$

Median

Middle value in a data set, after it is arranged in the order of magnitude.Values are first arranged from the smallest to the largest. Median is the middle value of the ordered data set

If the total number of values is an odd number, then median is average of the middle two values.

Median = [(n+1)/2]th value if there are odd number of values.

Example: For dataset {3, 13, 7, 5, 21, 23, 39, 23, 40, 23, 14, 12, 56, 23, 29}

we put those numbers in order we have: 3, 5, 7, 12, 13, 14, 21, 23, 23, 23, 23, 29, 39, 40, 56

There are **fifteen** numbers so n=15

Median = (15+1)/2 = 8 - Our middle is the **eighth** number

The median value of this set of numbers is **23**.

If the total number of values is an even number.

Median = average of the [n/2]th and the [(n/2) +1]th values, if there are an even number of values.

Example: For dataset {3, 13, 7, 5, 21, 23, 23, 40, 23, 14, 12, 56, 23, 29}

When we put those numbers in order we have: 3, 5, 7, 12, 13, 14, 21, 23, 23, 23, 23, 29, 40, 56

There are now **fourteen** numbers and so we don't have just one middle number, we have a **pair of middle numbers**:

n=14

n/2= 7^{th} number

(n/2) +1= 8^{th} number

In this example the middle numbers are **21 and 23**.

To find the value halfway between them, add them together and divide by 2:

21 + 23 = 44
then **44 ÷ 2 = 22**

So, the **Median** in this example is **22**.

Mode

Most frequently occurring value in the data set, i.e., value of the random sample that occurs with the greatest frequency.
It is applicable to all levels of data measurement, i.e., nominal, ordinal, interval, and ratio scales.

Example: In data set {6, 9, 3, 6, 6, 5, 2, 3}, the Mode is 6 as it occurs most often.

Measures of Spread

A measure of how the data is dispersed or spread around the mean, i.e., the average. It is used in quantitative data, as the variables can be arranged in a logical order, with a low and high value.

Spread can be measured in

- Range
- Quartiles and Interquartile range
- Percentile
- Variance
- Standard deviation

Range

- It is the simplest measure of variability.
- The distance between the smallest value and the largest value in a dataset.
- Represents the width of the smallest interval that contains all the data.
- The range of a list is calculating the difference between the largest value and the smallest value.
- Affected by outliers, as variance may either be too low or too high because of outliers.

Quartiles Range

Divide an ordered dataset into four equal parts, and refer to the values of the point between the quarters

- The lower quartile (Q1) is the point between the lowest 25% of values and the highest 75% of values. It is also called the 25th percentile.
- The second quartile (Q2) is the middle of the data set. It is also called the 50th percentile, or the median.
- The upper quartile (Q3) is the point between the lowest 75% and highest 25% of values. It is also called the 75th percentile.
- The interquartile range (IQR) is the difference between the upper (Q3) and lower (Q1) quartiles and describes the middle 50% of values when ordered from lowest to highest.

Percentile

- The percentile usually indicates that a certain percentage falls below that percentile. For example, if you score in the 25th percentile, then 25% of test takers are below your score. The “25” is called the **percentile rank**.

- The *n*th percentile is the smallest score that is greater than **or equal to** a certain percentage of the scores. To rephrase this, it’s the percentage of data that falls at or below a certain observation

Variance

- Variance is a measure of how the spread-out a data set is.
- Calculated as the average squared deviation of each number from the mean of a data set.
- The formula for calculating variance (S^2) of a data set is:

$$S^2 = \frac{\sum (x - \bar{x})^2}{n}$$

(Where x is the individual value, $\bar{x}$ is the mean of the sample set, n is the number of values in the distribution.)

Standard Deviation

- The mean distance of all values from the overall mean.
- The square root of variance.
- Standard Deviation of a sample data set can be calculated using the formula:

$$S = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

(Where x is the individual value, $\bar{x}$ is the mean of the sample set, n is the number of values in the distribution.)

Summary

- Statistics is the field in which data samples are utilize support claims about populations (a larger data set).Utilizes models, representations, and synopses for a set of experimental data and applies the results to a larger population.

- Data can be broadly divided into two types: Categorical Variable and Numerical (Continuous) Variable.

- Population is a collection of all persons, objects, or items. and Sample is a portion of whole or representative of the whole population.

- Mean is the average value of data, Median is the centre value of ordered data set, Mode is the Most frequently occurring value in the data set.

- Variance is a measure of how the spread-out a data set is and Standard Deviation is the mean distance of all values from the overall mean.