Data Science with Python Programming

- Presentation By Uplatz
- Contact us: https://training.uplatz.com
- Email: info@uplatz.com
- Phone: +44 7836 212635

Uplatz

Introduction to Statistical Analysis

Learning outcomes:

- What is Statistical Analysis?
- Introduction to Math And Statistics For Data Science
- Terminologies In Statistics – Statistics For Data Science
- Categories In Statistics
- Correlation
- Mean, Median, and Mode
- Quartile

What is Statistical Analysis?

Statistical analysis is the process of generating statistics from stored data and analysing the results to deduce or infer meaning about the underlying dataset or the reality that it attempts to describe. It's the science of collecting, exploring and presenting large amounts of data to discover underlying patterns and trends. Statistics are applied every day – in research, industry and government – to become more scientific about decisions that need to be made.

What is Statistical Analysis

For example:

- Manufacturers use statistics to weave quality into beautiful fabrics, to bring lift to the airline industry and to help guitarists make beautiful music.
- Researchers keep children healthy by using statistics to analyse data from the production of viral vaccines, which ensures consistency and safety.
- Communication companies use statistics to optimize network resources, improve service and reduce customer churn by gaining greater insight into subscriber requirements.
- Government agencies around the world rely on statistics for a clear understanding of their countries, their businesses and their people.

Introduction Math And Statistics For Data Science

As Josh Wills once said,

“Data Scientist is a person who is better at statistics than any programmer and better at programming than any statistician.”

Math and Statistics for Data Science are essential because these disciples form the basic foundation of all the Machine Learning Algorithms. In fact, Mathematics is behind everything around us, from shapes, patterns and colours, to the count of petals in a flower. Mathematics is embedded in each and every aspect of our lives.

Uplatz

Introduction Math And Statistics For Data Science

Although having a good understanding of programming languages, Machine Learning algorithms and following a data-driven approach is necessary to become a Data Scientist.

Introduction Math And Statistics For Data Science

Introduction To Statistics:

To become a successful Data Scientist you must know your basics. Math and Stats are the building blocks of Machine Learning algorithms. It is important to know the techniques behind various Machine Learning algorithms in order to know how and when to use them. Now the question arises, what exactly is Statistics?

Introduction Math And Statistics For Data Science

Introduction To Statistics:

Statistics is a Mathematical Science pertaining to data collection, analysis, interpretation and presentation. Statistics is used to process complex problems in the real world so that Data Scientists and Analysts can look for meaningful trends and changes in Data. In simple words, Statistics can be used to derive meaningful insights from data by performing mathematical computations on it.

Terminologies In Statistics – Statistics For Data Science

One should be aware of a few key statistical terminologies while dealing with Statistics for Data Science.

- ***Population*** *is the set of sources from which data has to be collected.*
- *A* ***Sample*** *is a subset of the Population*
- *A* ***Variable*** *is any characteristics, number, or quantity that can be measured or counted.*
- *A variable may also be called a data item.*

Terminologies In Statistics – Statistics For Data Science

- *Also known as a statistical model, A statistical **Parameter** or population parameter is a quantity that indexes a family of probability distributions. For example, the mean, median, etc of a population.*

Categories In Statistics

There are two main categories in Statistics, namely:

Descriptive Statistics

Inferential Statistics

Descriptive Statistics:

Descriptive Statistics uses the data to provide descriptions of the population, either through numerical calculations or graphs or tables. Descriptive Statistics helps organize data and focuses on the characteristics of data providing parameters.

Categories In Statistics

Descriptive Statistics:

It deals with the quantitative description of data through numerical representations or graphs.

Suppose you want to study the average height of students in a classroom, in descriptive statistics you would record the heights of all students in the class and then you would find out the maximum, minimum and average height of the class.

Categories In Statistics

Inferential Statistics:

Inferential Statistics makes inferences and predictions about a population based on a sample of data taken from the population in question. Inferential statistics generalizes a large data set and applies probability to arrive at a conclusion. It allows you to infer parameters of the population based on sample stats and build models on it.

Categories In Statistics

Inferential Statistics:

Example: A company is thinking about buying 50,000 electric batteries from a manufacturer. It will buy the batteries if no more that 1% of the batteries are defective. It is not possible to test each battery in the population of 50,000 batteries since it takes time and costs money. Instead, it will select few samples of 500 batteries each and test them for defects. The results of these tests will then be used to estimate the percentage of defective batteries in the population.

Correlation

Correlation is a statistical technique for measuring the relationship between the two variables. It is of three types- Positive Correlation, Negative Correlation, and Zero Correlation.

In a **positive correlation**, both variables increase and decrease together. Whereas, in a **negative correlation**, one variable increases and the other decreases. And, finally, in the case of zero correlation, there is no relation between the variables.

Correlation

Positive Correlation Examples:

1) The more time you spend running on a treadmill, the more calories you will burn.
2) The longer your hair grows, the more shampoo you will need.

Negative Correlation Example:

1) A student who has many absences has a decrease in grades.
2) If a train increases speed, the time to get to the final point decreases.

Mean, Median, and Mode

What can we learn from looking at a group of numbers?

In Machine Learning (and in mathematics & statistics) there are often three values that interests us:

Mean - The average value

Median - The mid point value

Mode - The most common value

Mean, Median, and Mode

Mean:

The mean value is the average value.

Example: We have registered the speed of 13 cars:

speed =

[99,86,87,88,111,86,103,87,94,78,77,85,86]

To calculate the mean, find the sum of all values, and divide the sum by the number of values:

(99+86+87+88+111+86+103+87+94+78+77+85+86) / 13 = 89.77

The NumPy module has a method for this.

Mean, Median, and Mode

Mean:

Example:

```
import numpy as np
speed =
[99,86,87,88,111,86,103,87,94,78,77,85,86]

x = np.mean(speed)
print(x)
```

Mean, Median, and Mode

Median:

The median value is the value in the middle, after you have sorted all the values:

77, 78, 85, 86, 86, 86, 87, 87, 88, 94, 99, 103, 111

It is important that the numbers are sorted before you can find the median.

If there are two numbers in the middle, divide the sum of those numbers by two.

77, 78, 85, 86, 86, 86, 87, 87, 94, 98, 99, 103

(86 + 87) / 2 = 86.5

The NumPy module has a method for this:

Mean, Median, and Mode

Median :

Example:

```
import numpy as np
speed =
[99,86,87,88,111,86,103,87,94,78,77,85,86]

x = np.median(speed)

print(x)
```

Mean, Median, and Mode

Mode:

The Mode value is the value that appears the most number of times:

99, 86, 87, 88, 111, 86, 103, 87, 94, 78, 77, 85, 86

Use the **SciPy mode()** method to find the number that appears the most. SciPy is a free and open-source Python library used for scientific computing and technical computing. SciPy contains modules for optimization, linear algebra, integration, interpolation, special functions, signal and image processing, and other tasks common in science and engineering.

Mean, Median, and Mode

Mode:

Example:

```
from scipy import stats
speed =
[99,86,87,88,111,86,103,87,94,78,77,85,86]

x = stats.mode(speed)

print(x)
```

Quartile

A quartile is a statistical term describing a division of observations into four defined intervals based upon the values of the data and how they compare to the entire set of observations.
The quartile measures the spread of values above and below the mean by dividing the distribution into four groups. A quartile divides data into three points – a lower quartile, median, and upper quartile – to form four groups of the data set. Quartiles are used to calculate the interquartile range, which is a measure of variability around the median.

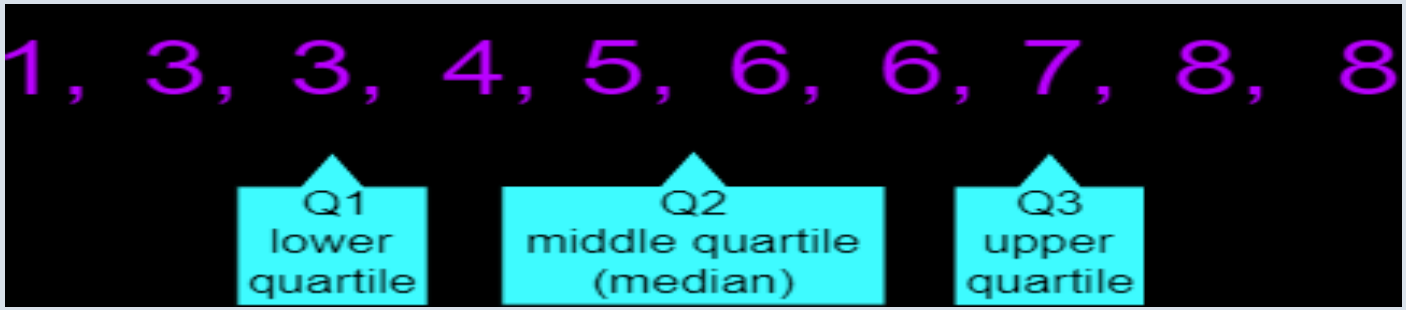
How Quartile work?

How Quartile work?

A quartile divides data into three points – a lower quartile, median, and upper quartile – to form four groups of the data set. The lower quartile or first quartile is denoted as Q1 and is the middle number that falls between the smallest value of the data set and the median. The second quartile, Q2, is also the median. The upper or third quartile, denoted as Q3, is the central point that lies between the median and the highest number of the distribution.

How Quartile work?

Example: 1, 3, 3, 4, 5, 6, 6, 7, 8, 8
The numbers are already in order
Cut the list into quarters:



In this case Quartile 2 is half way between 5 and 6:
Q2 = (5+6)/2 = **5.5**
And the result is:
Quartile 1 (Q1) = **3** Quartile 2 (Q2) = **5.5**
Quartile 3 (Q3) = **7**
Interquartile Range is: Q3-Q1 = 4

Uplatz