

The Basics of Statistics for Data Science By Statisticians



INSIDE THE GUIDE

TOPICS AND HIGHLIGHTS

Overview

Introduction to Statistics

Terminologies in Statistics

Types of Analysis

Data Types

Measures of Central Tendency

Measurements of Relationships between Variables

Probability Distribution Functions

Continuous Data Distributions

Discrete Data Distributions

Moments

Probability

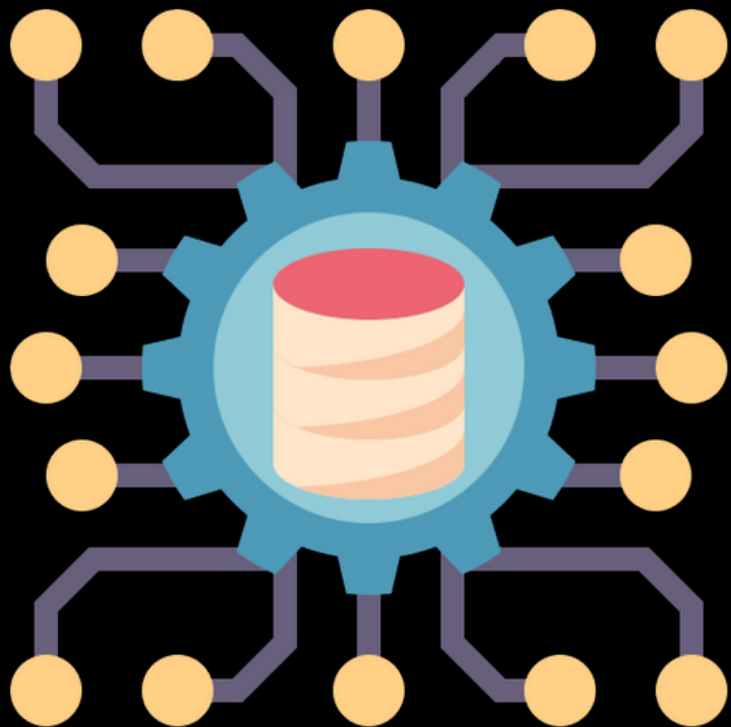
Accuracy

Conclusion



OVERVIEW

- Statistics is the building block of the machine learning algorithms.
- But most of the students don't know how much statistics they need to know to start data science.



INTRODUCTION TO STATISTICS



- Statistics has various methods that are helpful to solve the most complex problems of real life.
- Statistics has the power to drive meaningful insight from the data.
- Statistics offers a variety of functions, principles, and algorithms which is helpful to analyze raw data, build a Statistical Model and infer or predict the result.

TYPES OF ANALYSIS

STATISTICS HAS TWO TYPES OF ANALYSIS

QUANTITATIVE ANALYSIS

Quantitative Analysis is also known as statistical analysis. It is the science or an art of collecting and interpreting data with numbers and graphs. We also use it to identify patterns and trends.

QUALITATIVE ANALYSIS

Qualitative is also known as Non-Statistical Analysis. It gives generic information. It also uses text, sound and other forms of media.



DATA TYPES

STATISTICS HAS TWO TYPES OF DATA TYPES

NUMERICAL

Numerical data types are those data types which are expressed with digits. These data types are measurable. There are two major types of data types i.e. **discrete** and **continuous**.

CATEGORICAL

Categorical data types are qualitative data and it is classified into categories. There are two types of major categorical data types i.e. **nominal** (no order) or **ordinal** (ordered data).



MEASURES OF CENTRAL TENDENCY

MEAN

Means stands for the average of the given dataset.

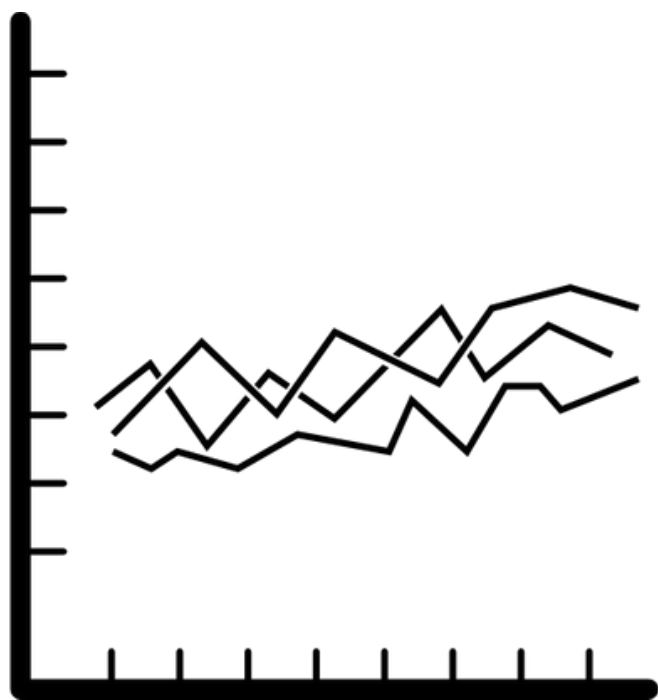
MEDIAN

Median is the middle of the given ordered dataset.

MODE

Mode is the most common value in a given dataset.

MEASURES OF VARIABILITY



RANGE

Range is the difference between the maximum and minimum value in a given dataset.

VARIANCE (Σ^2)

Variance measures how spread out a set of the given data is relative to the mean.

STANDARD DEVIATION (Σ)

It is also a measurement of how spread out numbers are in the given data set. Square root of variance is also known as standard deviation.

Probability

CONDITIONAL PROBABILITY

In this probability $P(A|B)$ is the likelihood of an event occurring. The event occurring is based on the occurrence of an event that occurred previously

BAYES' THEOREM

The Bayes' theorem is the most popular mathematical formula. It is used to determine the conditional probability. It is based on the methodology that the probability of A given B is equal to the probability of B given A times the probability of A over the probability of B".

ACCURACY

TRUE POSITIVE

It detects the condition, if the condition is present.

TRUE NEGATIVE

It does not detect the condition, if the condition is not present.

FALSE-POSITIVE

It automatically detects the condition if the condition is absent.


FALSE-NEGATIVE

It does not detect the condition if the condition is present.


SENSITIVITY

It measures the ability of a test to detect the condition. If the condition is present. The sensitivity = $TP / (TP + FN)$





		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative



ACCURACY



SPECIFICITY

It measures the ability of a test to correctly exclude the condition if the condition is absent. Its specificity = $TN / (TN + FP)$

PREDICTIVE VALUE POSITIVE

Predictive value positive is also called as precision. In this the proportion of positives that correspond to the presence of the condition. Here is the formula $PVP = TP / (TP + FP)$

PREDICTIVE VALUE NEGATIVE

In this the proportion of negatives. It also corresponds to the absence of the condition. Here is the formula $PVN = TN / (TN + FN)$

CONCLUSION

Now we have gone through all the basic concepts of statistics for data science. If you are going to start with data science then you should try to have a good command over all these statistics concepts. It will help you a lot when you start learning data science. With the help of these concepts you will be able to understand the data science concepts. So what are you waiting for? Grab the best statistics books and start learning these concepts.