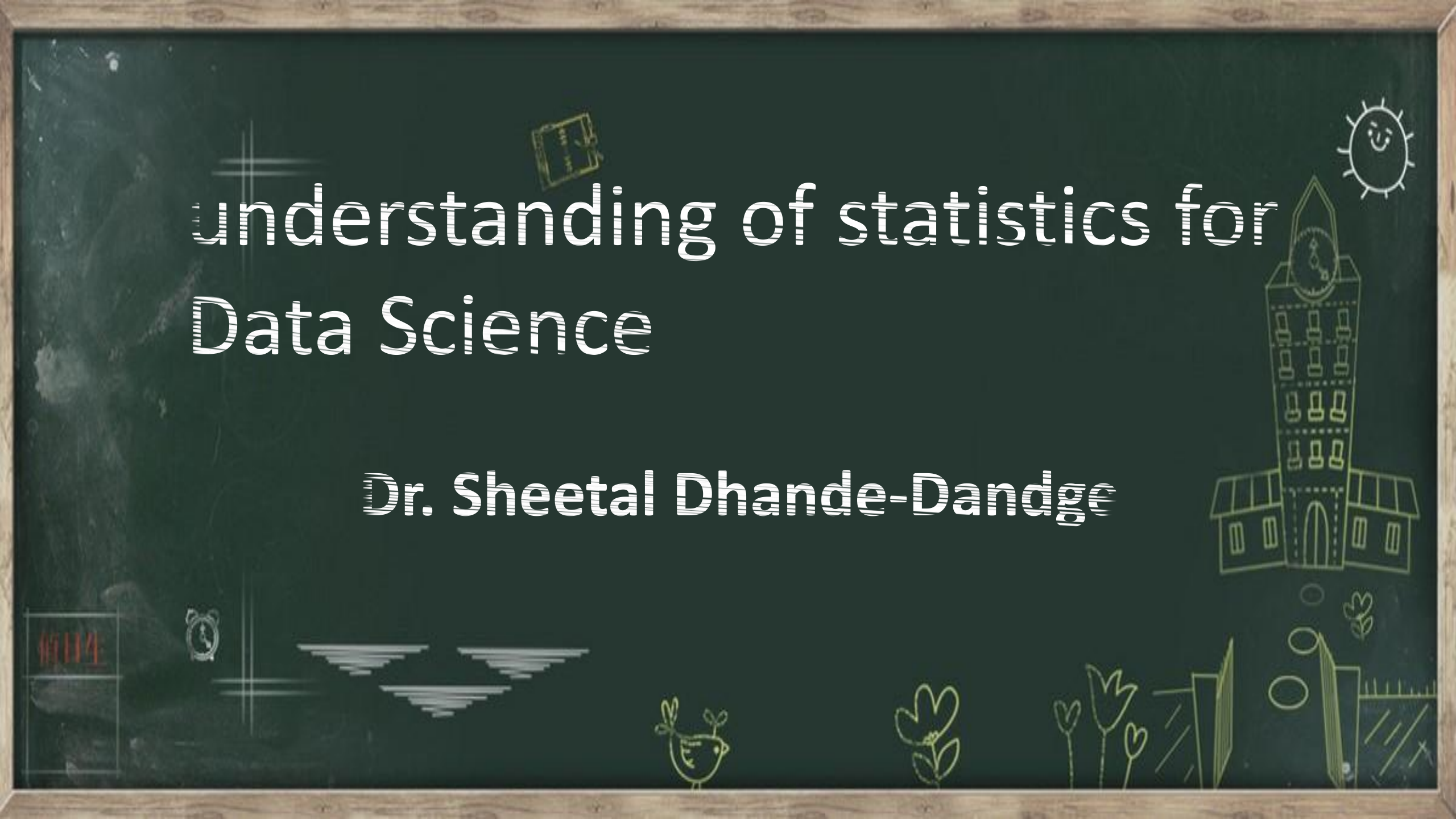
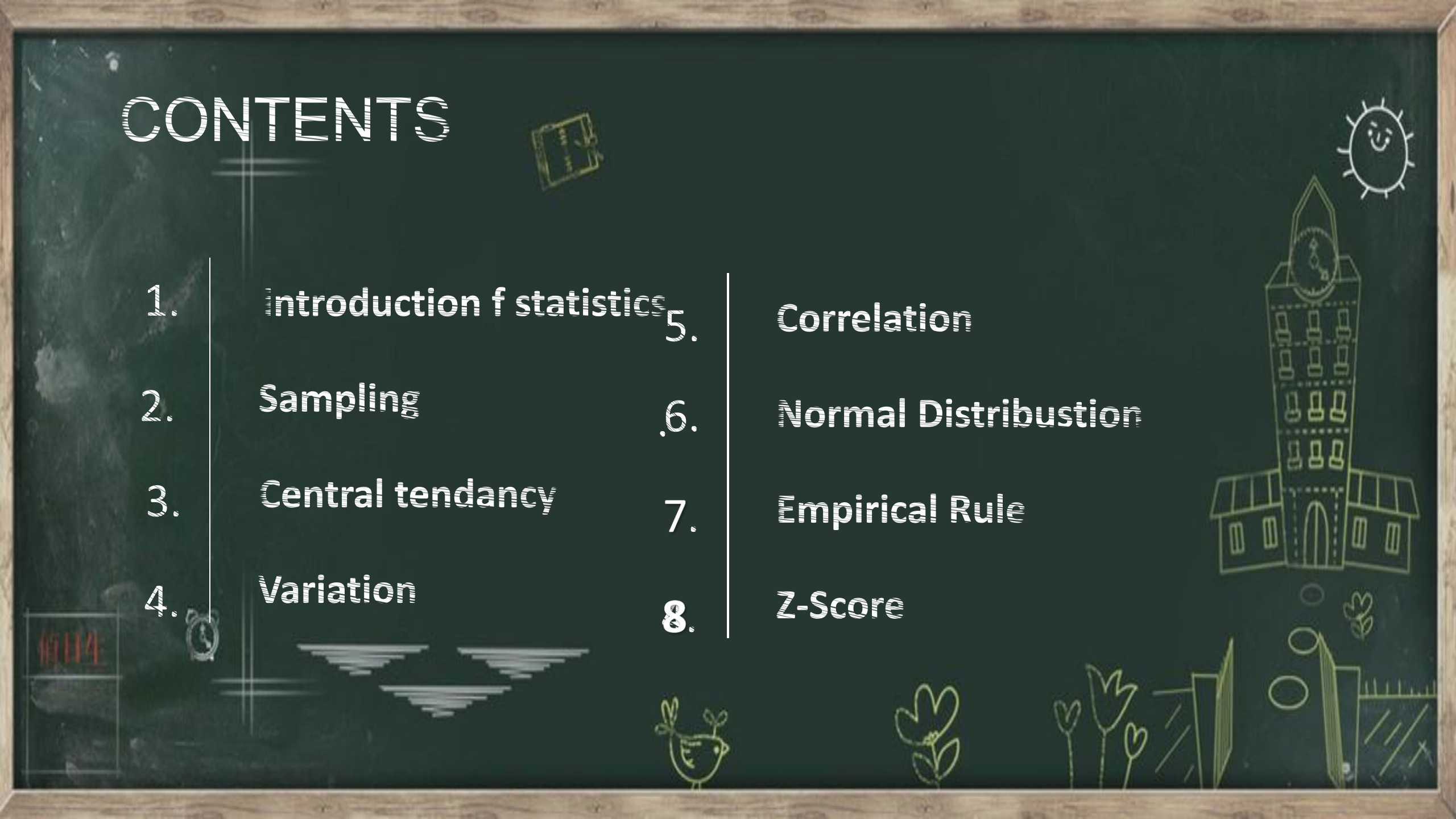
understanding of statistics for Data Science

Dr. Sheetal Dhande-Dandge



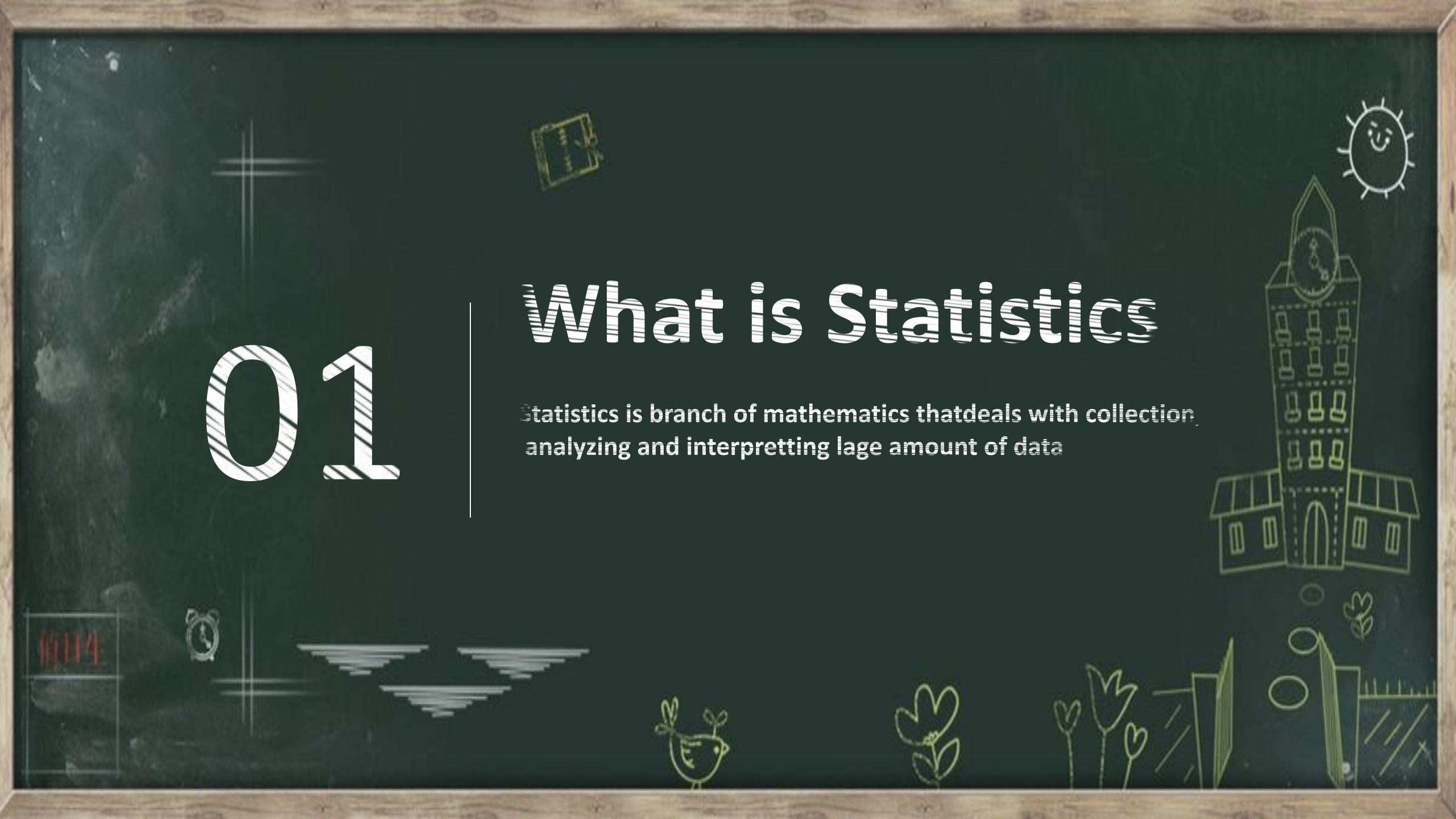
CONTENTS



01

What is Statistics

Statistics is branch of mathematics thatdeals with collection, analyzing and interpretting lage amount of data



01

Why Statistics is important

Statistics allows us to drive knowledge from large dataset and this knowledge cn then used to make prediction ,decision, classification,etc

01

Where Statistics is used

Statistics used in various field, some of them are

1. Medical Research
2. Stock Market
3. Sales Projection
4. WEather forecating

02

Sampling

Sampling is the process of collecting data to perform analysis on

02

Sampling

Sample Vs Population

Sampling Frame is List From (Population) Which Sample is Selected.

02

Sampling Error

Sample Error is an erroe that leads to our Sample not acuurately representing our population

02

Random Sampling

Random Sampling is process of seleting subset/sample from a population in such a way that every data pint is euqally likely to be inclued in sample

02

Systemaic Sampling

Systematic Sampling is process of seleting your sample by picking every kth element in your population , you don’t need to list this.

03

Central Tendancy

Is used to indicate where does the middle or center of distribution of our data lies

- Mean
- Mode
- Median

03

Central Tendancy: MODE

is used to indicate the most frequent data point , in other words , the one which occur most number of times.

03

Central Tendancy: Median

is the middle of data , if the dataset arrange is asending order then the data element which occurs right at center , is the median

03

Central Tendancy: Mean

is average of data ,in simpler terms its sum of values divided by total number of values , its represented by Greek letter ‘sigma’

Central Tendancy: Mean

03

Trimmed Mean : is used to deal with outliers by trimmimg or removing same data from both ends so as to get rid of outliers.

Weighted Mean: is used to certain values are supposed to count more in same context.
Eg. Calculating average grade of student based on their grade distribution

04

Variation

variation in stastics is used to show how data isdispressed, or spred out , several measures of variation are used in stastics

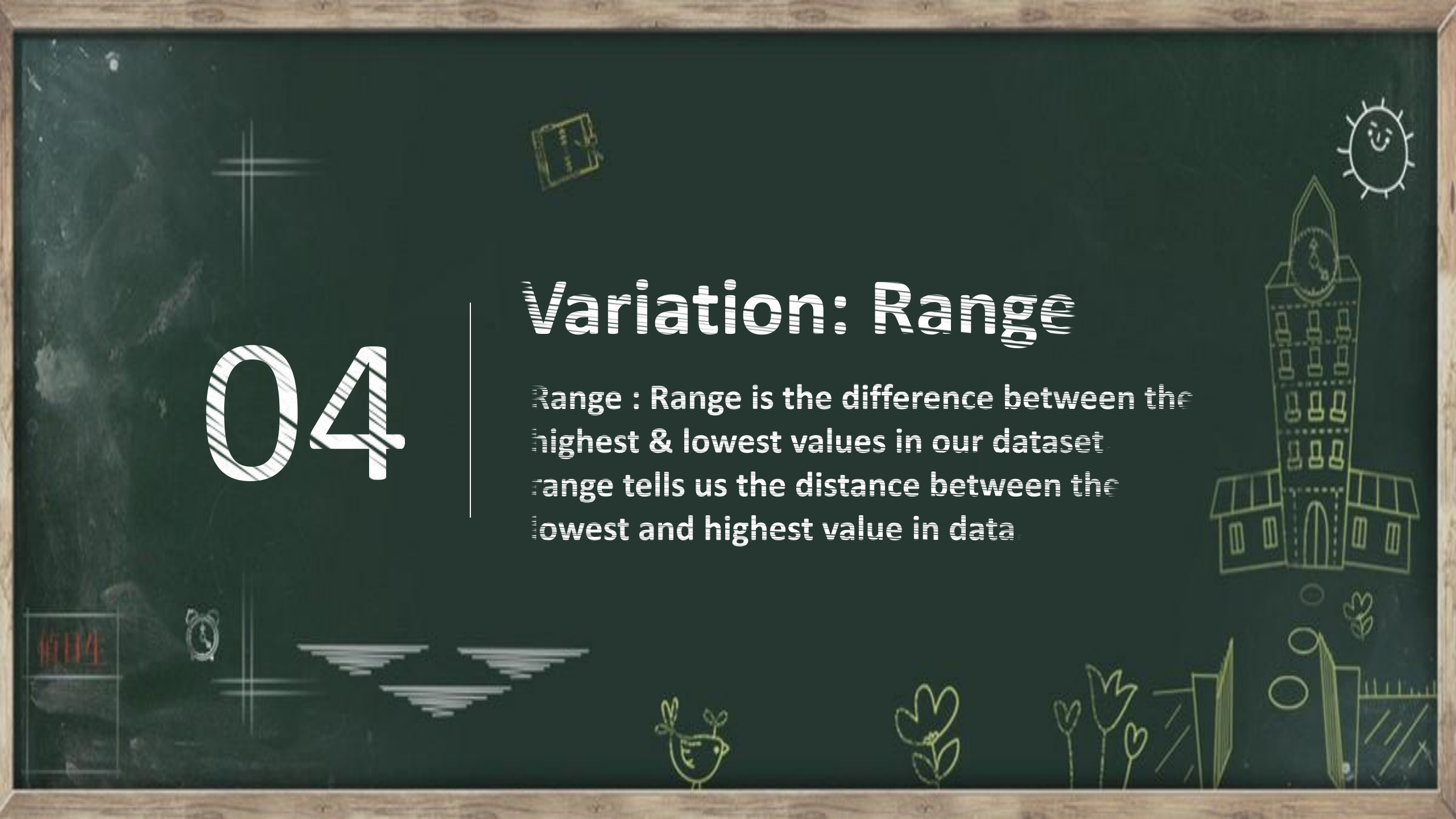
- Range
- Quartile
- Varience

04

Variation: Range

Range : Range is the difference between the highest & lowest values in our dataset

range tells us the distance between the lowest and highest value in data



04

Variation: Percetile

Percetile: Percetile are score that are used to describe a value below which some observations fall .

IF X is at 70th percentile it mean 70% of other data points from our sample are below X.

04

Variation: Quartile

Quartile : are used to break the data into four parts , so asto better find the spread of data in a way that is less influenced by outlier

Quartile are expressed in percentile, 1st Quartile is 25th percentile, 2nd quartile is 50th percentile(median) & 3rd Quartile is 75th percentile.

04

Variation: interQuartile Range (IQR)

InterQuartile Range (IQR) : is the difference between the lower and upper quartile. this gives us a better idea of the range of data

Standard Variance

01

Standard varience measures how far a set of number are spread out from their average value

Standard Deviation

02

Standard Deviation is used to express the Magnitute by which the members of a group differ from the mean value for the group

Sample Variance

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

Sample Standard Deviation

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

correlation

01

Correlation is a term that is measure of the strength of linear relationship between two Quantative Variable..

$$r = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable
$\overline{Y}$ = mean of Y variable

Positive Correlation : is the term that is used a term that is used to describe a positive linear relationship between two quantitative variable

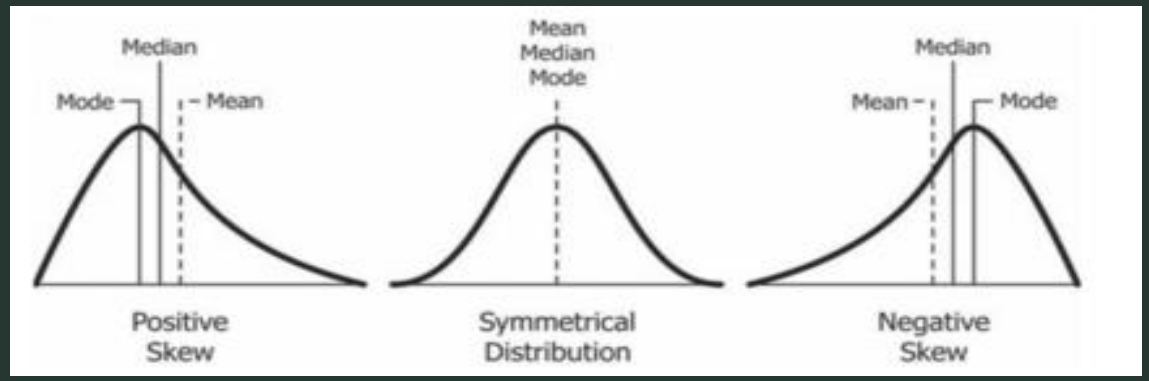
No Correlation : is the term that is used a term that is used to describe a No linear relationship between two quantitative variable

Negative Correlation : is the term that is used a term that is used to describe a Negative linear relationship between two quantitative variable

Normal Distribution

01

is the term used to describe a distribution which when plotted gives us ashape of bell curve.It has mean of zero and std deviation of 1.



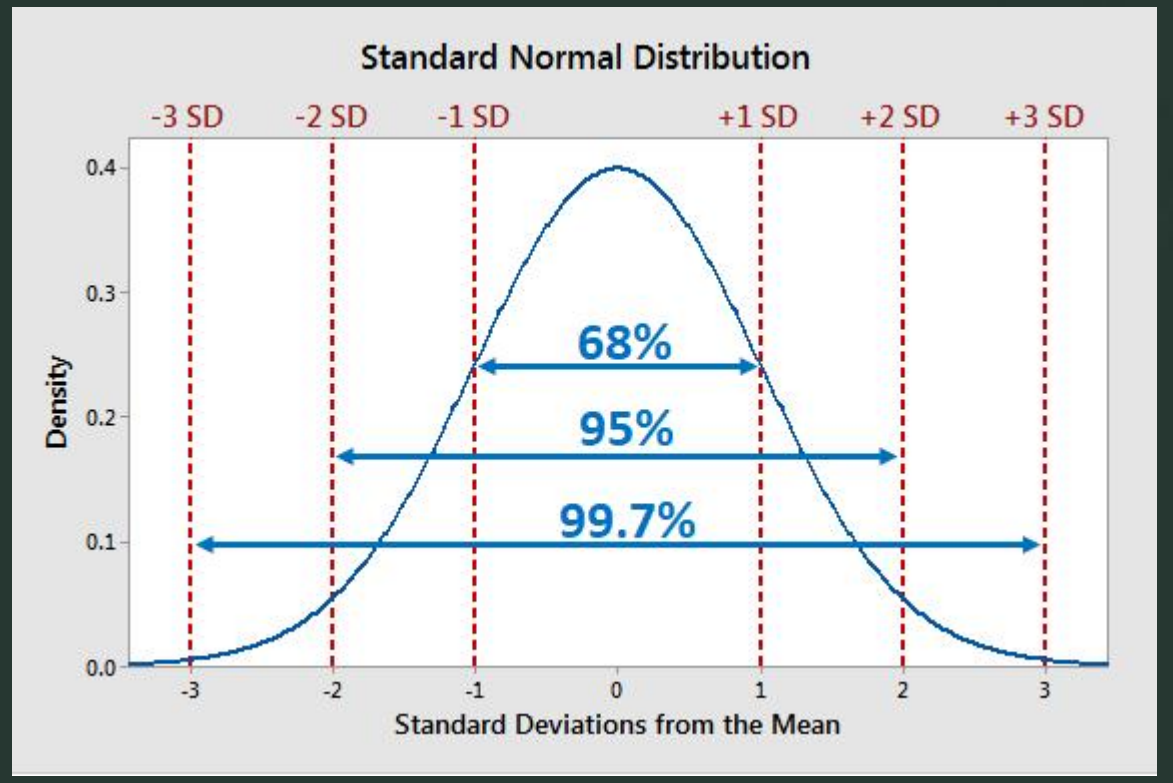
Empirical Rule

01

is used to remember the Percentage of values that lie within a bandaround the mean in Normal Distribution with a width of two,four six stndard deviation

The empirical rule, also referred to as the three-sigma rule or 68-95-99.7 rule, is a statistical rule which states that for a normal distribution, almost all observed data will fall within three standard deviations (denoted by σ) of the mean or average (denoted by μ).

In particular, the empirical rule predicts that 68% of observations falls within the first standard deviation (μ ± σ), 95% within the first two standard deviations (μ ± 2σ), and 99.7% within the first three standard deviations (μ ± 3σ).



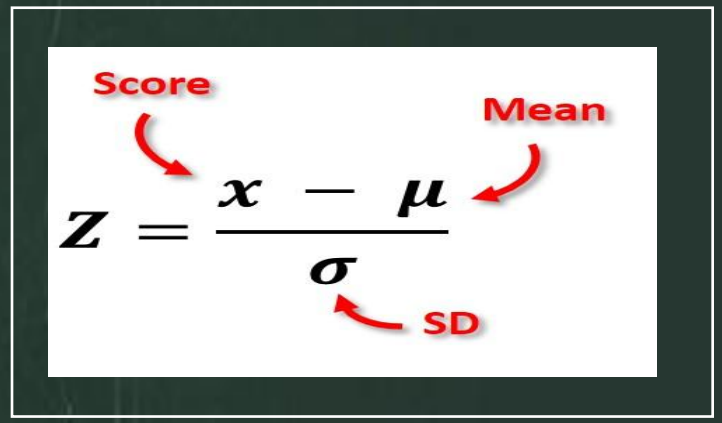
Z-score

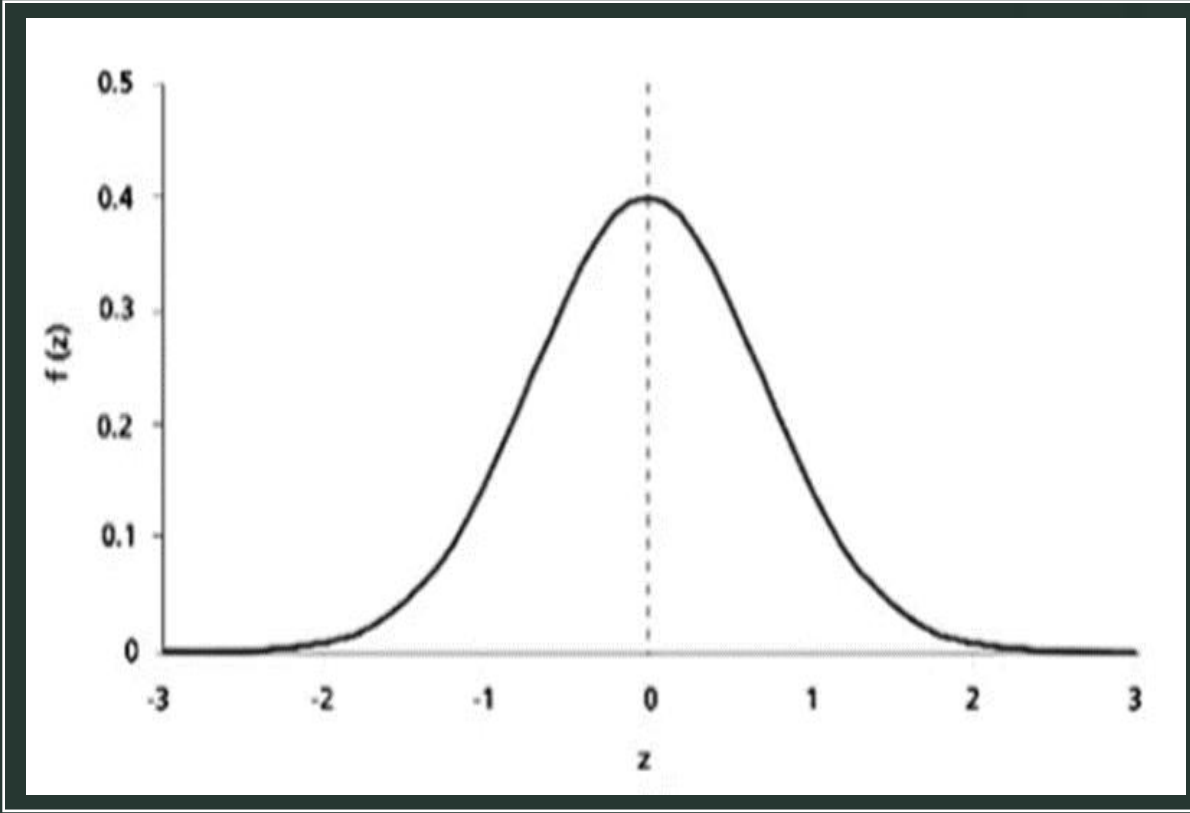
01

Z-score is a statistical measurement that describes a value's relationship to the mean of a group of values Z-score is measured in terms of standard deviations from the mean. If a Z-score is 0, it indicates that the data point's score is identical to the mean score

Z-Score is Measure of how many standard deviation below or above the population mean a raw score

Z-score can be placed on a normal distribution curve.






THANK YOU