Theory

Refer to : https://roadmap.sh/data-analyst

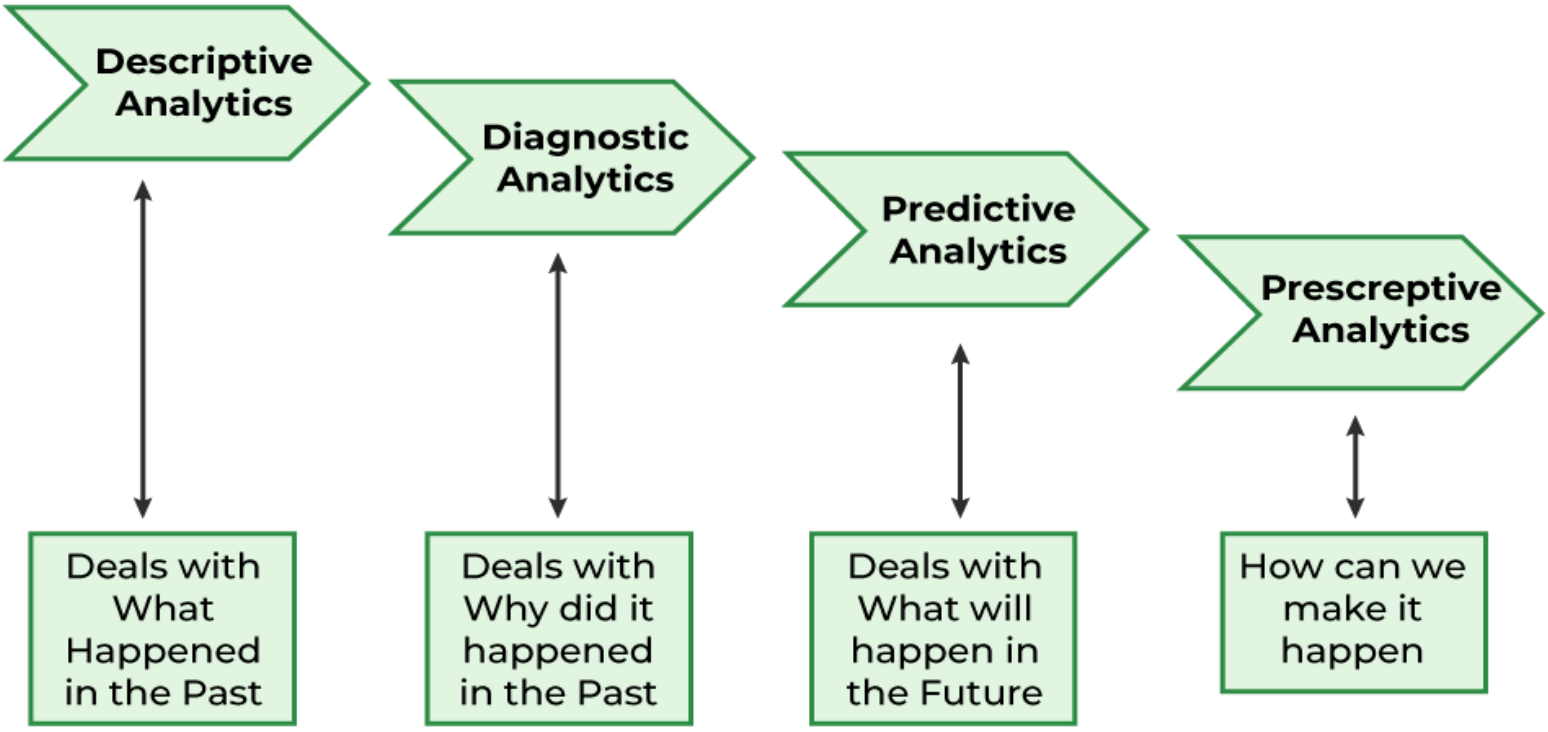
What is Data Science?

Data science is an in-demand career path for people with an aptitude for research, programming, math, and computers.

What is Data Analytics?

Data analytics is the process of analyzing raw data to find trends and answer questions

Types of Data Analytics: https://www.geeksforgeeks.org/data-analytics-and-its-type/



Stats

Types of Stats

1. **Descriptive Stats:** Organising and Summarizing Data.
2. **Inferential Stats:** To Draw inferences about the population data from sample data.

Sample & Population

Types of Sample:

1. **Random** - Eg: Exit Poll during Elections

2. **Stratified Sampling**
   - No Overlapping of groups (starts)
   - Eg: Separate Male Female survey
   - Dividing group based on something
   - Not to be confused with Convenience Sampling

3. **Systematic Sampling** - Eg: Selecting every n^{th} individual for survey

4. **Convenience / Non-probability Sampling**
   - Only selected (specific) people will take part in the survey.
   - Based on particular Domain
   - select participants based on their ease of access
   - Eg: if survey is about train, stand outside railway and ask
   - Not to be confused with Stratified Sampling

Variables

Types of Variables:

1. **Quantitative Variables** - Can be measured - Height, Age etc

   Types of Quantitative Variables**:**

   i. **Discrete** (Bar Graph) - Fixed number - My number of eyes

   ii. **Continuous** (Histogram) - Fractional values - My weight

2. **Qualitative / Categorical Variables** - Based on some characteristics - Gender, Blood Group etc

   Types:

   i. **Nominal** - Categories without order - Gender, Colors

   ii. **Ordinal** - Categories with order - Performance / Level

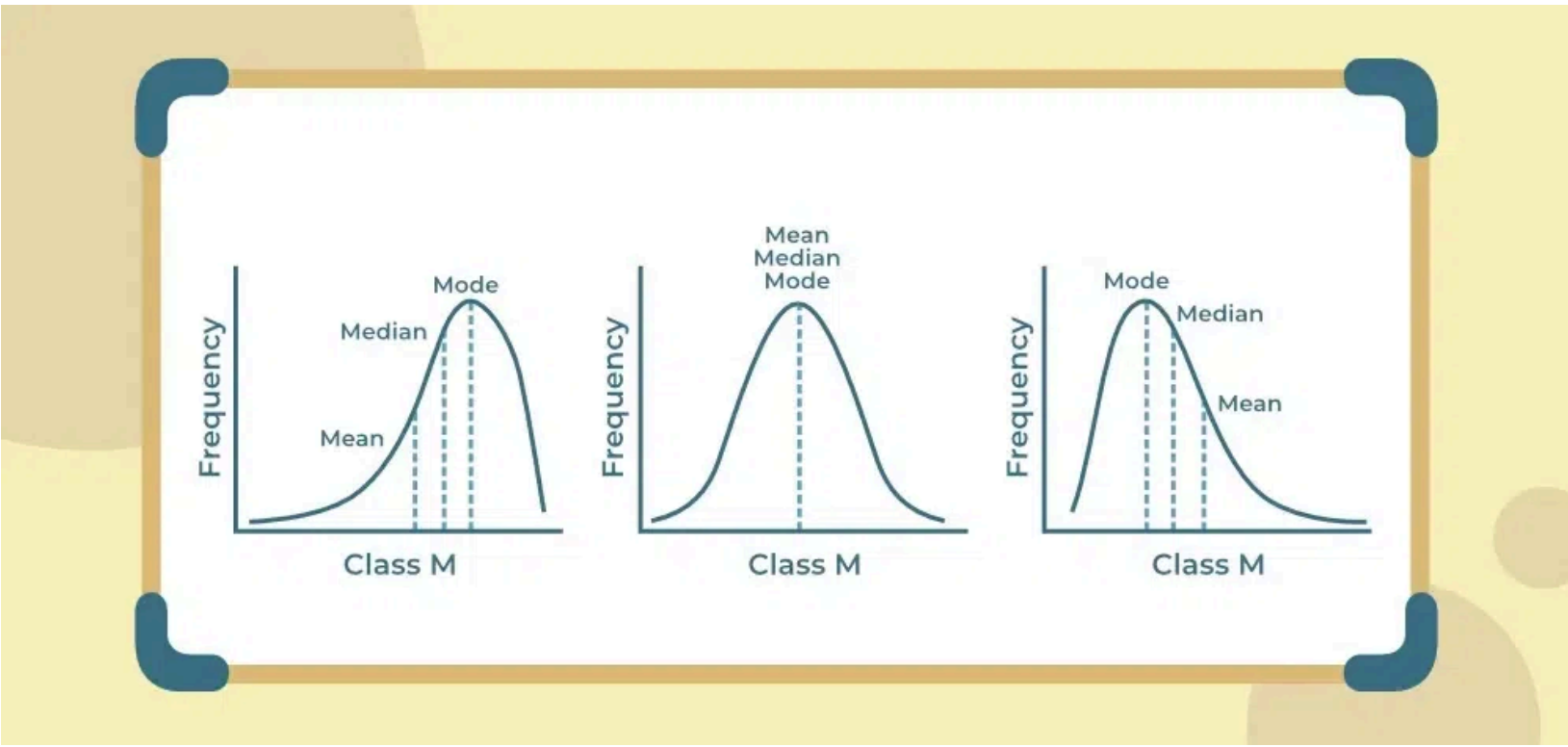
   iii. **Binary** - Only two options - Pass/Fail, Yes/No

Types of Variable Measurement Scales:

1. **Nominal** - Categorical, Classes - Colors
2. **Ordinal** - Order does NOT matters - Rank basis on marks
3. **Interval** - Order & Value BOTH matters - Distance etc
4. **Ratio** - Simple Ratio - Ratio

Frequency Distribution: Bar Graph(Discrete) & Histogram(Continuous).

Central Tendency: Measure of Central Tendency of a dataset represents a central value or a typical value for a dataset which can be used to do further analysis on the data.



Mean, Median, Mode ---------- 2 × Mean + Mode = 3 × Median

Types of Mean: AM, GM, HM

Median: Middle number after sorting in ASC.

How can mean be misleading?
The mean can be misleading if the dataset contains extreme values (outliers), as it does not represent the majority of the data.

How can mode be misleading?
The mode can be misleading in datasets with multiple modes, as it may not accurately represent the central tendency.

Measure Of Dispersion (Spread) : Measures of Dispersion are used to represent the scattering of data. These are the numbers that show the various aspects of the data spread across various parameters.

https://www.geeksforgeeks.org/measures-of-dispersion/

Variance: Sum of squares of differences between all numbers and means.
Deviation for above example. First, calculate the deviations of each data point from the mean, and square the result of each:

$$2 \qquad \sum(xi - \bar{x})^2$$

(2−5)^2 = 9
+
(5−5)^2 = 0

Standard Deviation: Square of Variance. It is a measure of the extent to which data varies from the mean.

Coefficient of Variation: SD / Mean

These statistical measures are crucial for :

- understanding the distribution of data.
- mean provides a central value
- variance and standard deviation give insights into the data's variability or spread, indicating the consistency or volatility of the dataset.

SD and Variance cannot be -ve bcoz both are square and square root values.
Outliers affect both of them hugely.

- In Population, we divide by N,
- In Sample, we divide by n-1 (also called **degree of freedom**) Why?
- The deviation from the sample mean always sum to 0.

Direct method :

$$\bar{X} = \frac{\sum fx}{\sum f}$$

Assumed mean method :

$$\bar{X} = A + \frac{\sum fd}{\sum f}$$

$$d = x - A$$

Step deviation method :

$$\bar{X} = A + \left[\frac{\sum fd}{\sum f} \times c\right], \text{ where } d = \frac{x - A}{c}$$

$$M_m = l + \left(\frac{\frac{n}{2} - cf}{f}\right)h$$

Where
l = lower limit of median class,
n = number of observations,
cf = cumulative frequency of class preceding the median class,
f = frequency of median class,
h = class size (assuming class size to be equal)

Percentiles: A value below which a certain percentage of observation lies.

Eg: Dataset: 1, 2, 2, 3, 4
Q1: What is the percentile ranking of 3?
A: n=5
percentile ranking of 3 = no. of values less than 3 / n = 3 / 5 = 0.6 = 60%
Q2: What value exists at percentile ranking 25% ?
A: **Percentile * (n+1) / 100**
25 * 6 / 100 = 6/4 = 1.5 (Index Position!)

Avg of value between 1 and 2 = 1+2/2 = 1.5 (value)

& Quartiles:

Five Number Summary:

1. Min **2.** 1st Q **3.** Median **4.** 3rd Q **5.** Max

Removing the outlier:

{ 1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27 }, n = 19

[Lower Fence <----------> Higher Fence]

Lower Fence = Q1 - 1.5 (IQR)
Upper Fence = Q3 + 1.5 (IQR)

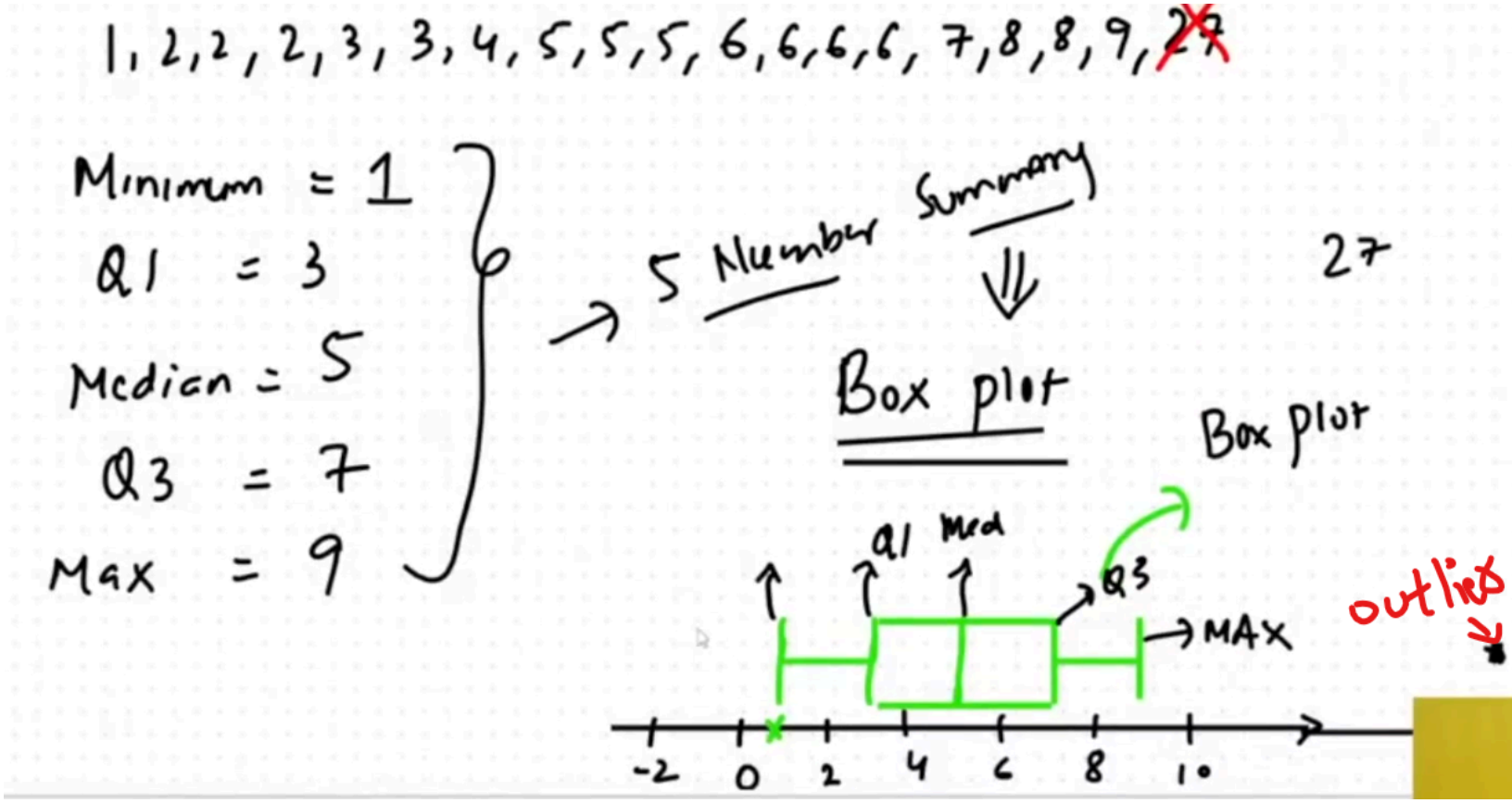
Q3 = 75% = 75 * (n-1) / 100 = 75 * 20 / 100 = 15th element = **7**
Q1 = 25% = 25 * (n-1) / 100 = 25 * 20 / 100 = 5th element = **3**

IQR = Q3 - Q1 = 7-3 = 4

Lower Fence = 3 - 1.5 (4) = -3
Upper Fence = 7 + 1.5 (4)= 13

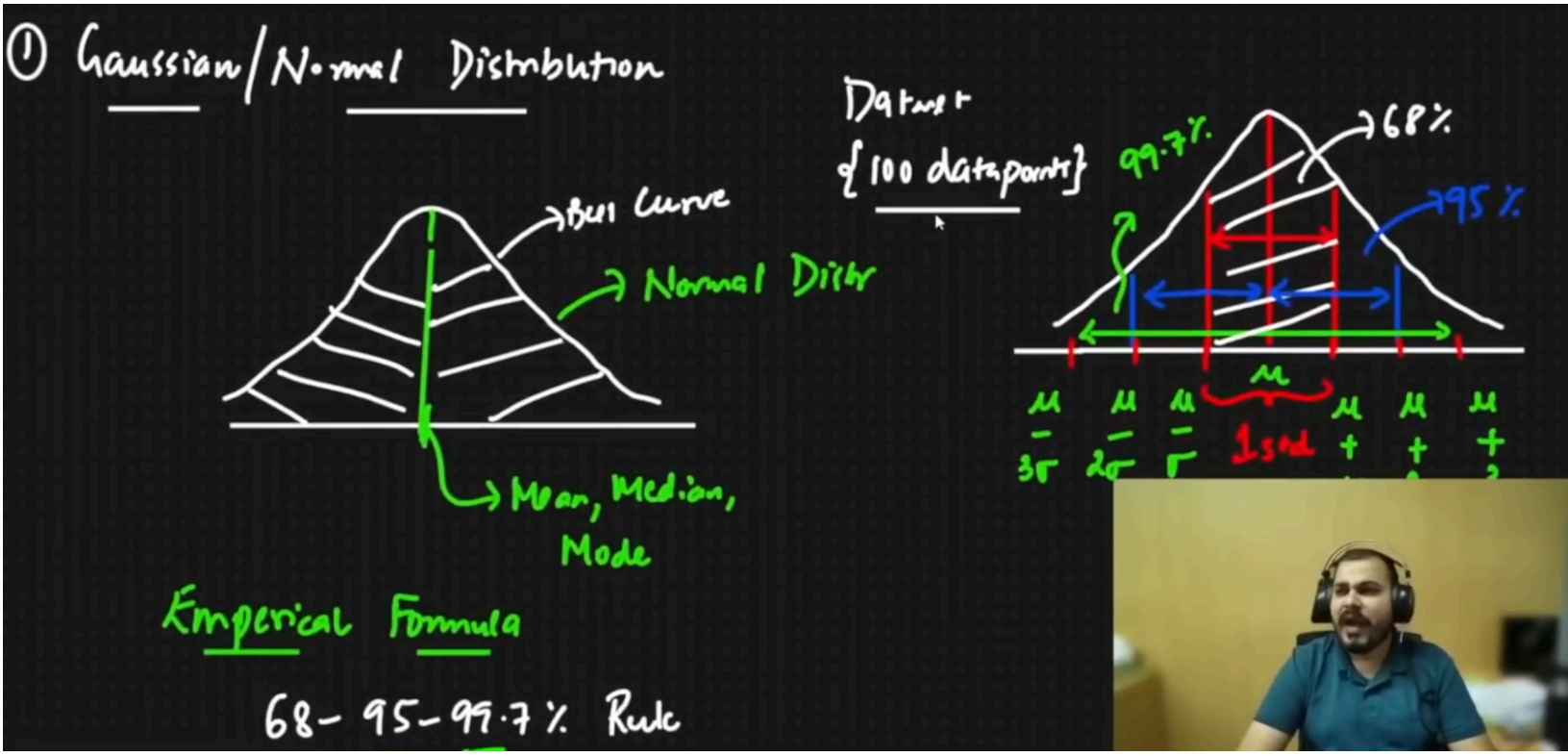
[-3 <—> 13]
Anything greater than 13 is an outlier, Anything lesser than -3 is an outlier.

So, remove 27.



Applications of Boxplot:
Detection of outlier

Distribution:
68-95-99.7 Rule : Empirical Formula



Z-Score: To calculate how many standard deviations to the left or right the given value is.
z = (x-μ)/σ
x = raw score
μ = mean
σ = standard deviation

1 2 3 4 5 6 7

$\{-3, -2, -1, 0, 1, 2, 3\}$

$Z(1) = \frac{1-4}{1} = -3$ $Z(3) = \frac{3-4}{1} = -1$

$Z(2) = \frac{2-4}{1} = -2$

$Y \sim SND(\mu=0, \sigma=1)$

$\{1, 2, 3, 4, 5, 6, 7\}$ → Normal Distr

⇓

Z score

⇓

Standard Normal Distr

$(\mu=0, \sigma=1)$

$\{-3, -2, -1, 0, 1, 2, 3\}$ ← Satisfying this