Date 1 March 2025
DELTA Pg No.

Introduction to fundamental statistical Concepts

* Mean (the average) np.mean(value) @ arithmetic average

The mean (@average) is a measure of central tendancy that represents the typical value in a dataset. It is calculated by, Mean = $\frac{\Sigma \text{All values}}{\text{No. of values}} = \frac{\Sigma N}{\Sigma N_{num.}}$

most common type of mean is arithmetic mean.

* Median (middle value) np.median(value)

The median is the middle value of a dataset when arranged in ascending or descending order. It is a measure of central tendancy that helps identify the center of the data.

How to calculate? → Sort the data.

a) Arrange the numbers in a order.

b) If the variable is odd, middle number is the median → $\frac{N+1}{2}$

If the variable is even, average of two middle number is median

→ $\frac{N}{2}$ or $\frac{N+1}{2}$

* Correlation

Correlation is a statistical measure that describes the relationship between two variables. It tells us how strongly and in what direction the two variables move together.

Types of Correlation

a) Positive

one variable increases, other

b) Negative

one variable increases, other

c) No correlation

No relation between varia

* Mode, is the elements with the common value in the data.

Unimodal - distribution of values with only one mode
bimodal - with two mode
multimodel - more than 2 mode
Mode can be both Categorical/numerical

Python -
```
from statistics import multimode
x = multimode(value)
```

Using R -
```
mode <- function (n) {
  Unique_values <- unique(n)
  table <- tabulate(match(n, unique_values))
  Unique_values[table == max(table)]
}
```

* Correlation

Correlation is a statistica
relationship between two variables.
and in what direction the two var

Types of Correlation

a) Positive

one variable increases, other also increases.

b) Negative

one variable increases, other decreases.

c) No correlation

No relation between variable.

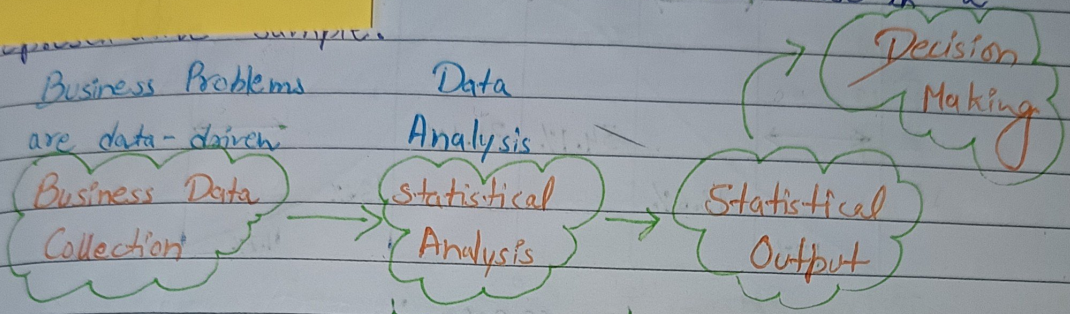
* Mode, is the elements with the highest frequency or the most common value in the data.

Shape of Distribution
- Symmetry & skewness
  - Positive Skew (right-skewed) tail extends to right, mean > median.
  - Negative skew (left-skewed) tail extends to left, mean < median
- Kurtosis
  Describes the 'tailedness' of the distribution. It tells us how many outliers exists.

Date 27 March 2025
DELTA Pg No.

→ statistical Analysis
things → Non-statiscal Analysis.

...ce of collecting and analysing the ...rge quantities, especially for the purpose ...s in a whole from those in a ...



Categories in Statistics (two major)

(a) Descriptive Statistics - organising/summarizing the data, calculation
Describe the basic features of the data in a study.
They provide simple summaries about the sample and the measures. {Average/%} "Numerical Analysis/output"

(b) Inferential statistics - for conclusions of the data, estimation
Helps in generalising about the population by using various analytical tests and tools.
"Relative Analysis/output".

Statistical Measures (Three major)

(a) Measure of frequency - frequency table → Histogram, Frequency Distribution
Describe basic features of data in a study. They provide simple summaries about the sample and the measures.

(b) Measure of Central Tendency (mean/median/mode.)
Helps in generalising about the population by using various analytical tests and tools. (center of distribution of Data)

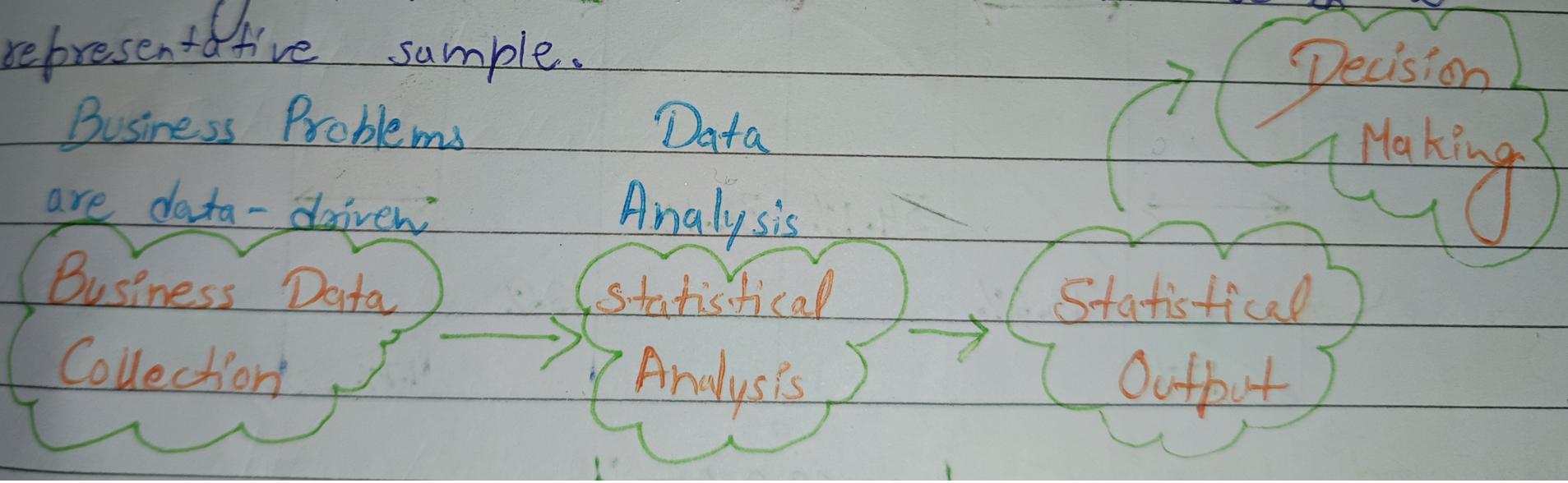
(c) Measure of Spread (standard deviation/ variance/ IQR / Range)
Helps in generalising about the population by using various analytical tests and tools.

Introduction to statistics

→ statistical Analysis
→ Non-statiscal Analysis

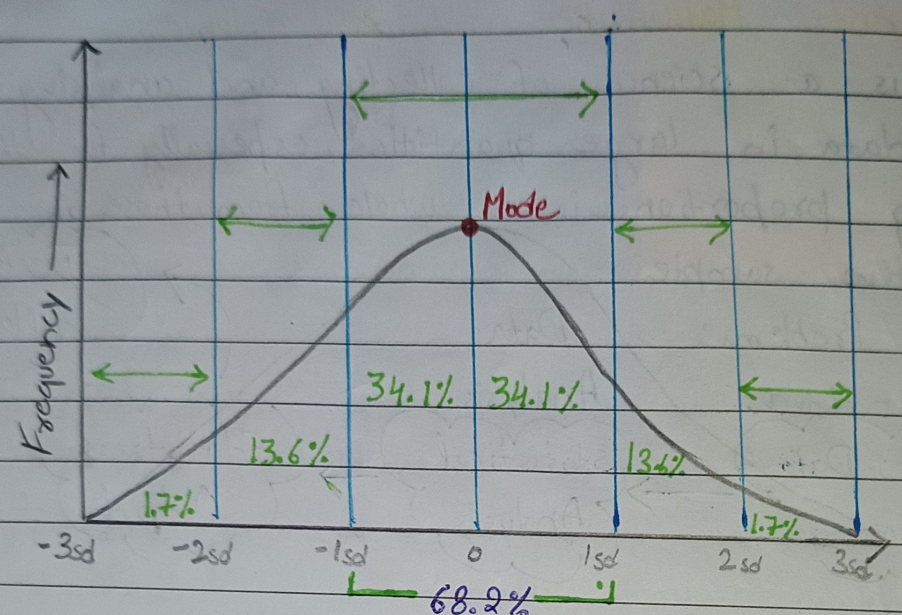
Statistics is a science of collecting and analysing the numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.

Business Problems are data-driven



Date
DELTA Pg No.

• Distribution Curve (bell-curve)



↳ most of the datapoints are accumulated here.

range : distance between minimum and maximum values

frequency : No° of repetitions of values in X-axis.

Central Tendency : accumulation of data points towards the center.

Null Hypothesis (H_0)

It is commonly accepted fact. In ML, it usually represents the conclusion that the model will not work. It is the opposite of the alternative hypothesis (H_A). Researchers attempt to reject the null Hypothesis.

$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ and $t = \frac{\bar{x} - u}{s/\sqrt{n}}$

Date 03/04/2025
DELTA Pg No.

Week 1 - (Thursday)

(1.1) Fundamental Mathematics for data Science. (1.1)

Key Mathematical Concepts for Data Science

- Linear Algebra and Probability
  understand vectors, matrices and distributions
- Essential Operations
  Performs matrix manipulation and probability calculations
- Descriptive statistics
  Summarise data and interpret visualisations
- Probability Applications
  Apply distributions to predictive modelling.

Linear Algebra: The backbone of Data Science. (1.2)

Essential Tools for complex Problem solving

- focuses on vectors, matrices and linear transformations
- Powers data science, machine learning, physics and engineering
- Provides a structured approach to multi-variable problem-solving.
- Enhances data analysis, algorithm optimisation and model development.

Key Components in Linear Algebra.

The building blocks of Data magic

- Vectors Lists of numbers that show direction or represent the data points
- Matrices Grids of numbers used to store and organise data
- Linear Transformation ways to change or move data, like scaling or rotating.

Applications of Linear Algebra in Data Science

Data Representation	structures data with vectors and matrices
Dimensionality Reduction	Simplifies datasets with retaining insights
Machine Learning Algorithms	Powers algorithms like regression and SVM
Transformation & Projections	Converts data to reveal hidden patterns
Optimisation Problems	Solves tasks like minimising loss
Big Data Operations	Enables fast Processing of large datasets
Neural Networks	Supports deep learning transformations
Feature Engineering	Create impactful features for models
Graph Theory	Models networks and connections
Broad Applications	Extends to AI, NLP and bio informatics.

Scalars Made Simple (1.3).

Single Numerical value with no Direction

* Types

- **Real Scalars**

  Whole numbers, fractions, and irrational numbers (e.g., 5, -3.2, 0.75, ...)

- **Complex Scalars**

  Numbers in the form $a+bi$, where i represents the imaginary unit

* Operations

There are four operations that can be done with scalars :-

- Addition
- subtraction
- Multiplication
- Division

* Properties

Scalars are subjected to below properties :-

- Commutative
- Associative
- Distributive.

Date
DELTA Pg No.

What are vectors?

The Power of size and direction.

- A vector is a mathematical object that has both magnitude (size) and direction.
- A vector is typically represented as an arrow. The length shows the magnitude and the head shows the direction of the vector.
- Vectors are widely used in fields such as physics, engineering & data science for various applications.
- They are essential tools for accurate analysis, modelling and solving complex problems in these areas.

Magnitude

The length of the vector, showing the size or strength of the quantity

Direction

The vector's orientation in space, indicated by an arrowhead.

Notation Conventions.

Scalars	Vectors
• Represented by italicised letters (a, b, c)	• Represented by bold letter ($\mathbf{v}$) or arrows ($\vec{v}$)
• Indicate single values without direction.	• Indicate quantities with magnitude and direction.
• Usually written in plain typeface.	• Can be expressed as column vectors $\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}$

Vector Magnitude

The distance from the vector's starting point to its endpoint in space. Used in distance measurements and in normalising vectors to unit length. $\|\mathbf{v}\| = \sqrt{(v_1^2 + v_2^2 + v_3^2)}$

Date
DELTA Pg No.

Classification of vectors.

vectors types and their unique characteristics

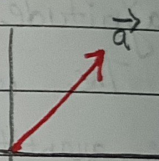
Position
- represent a point in space relative to an origin
- describe spatial relationships within a coordinate system

Displacement
- Indicate the change in position from one point to another
- Capture both distance and direction of movement

Zero
- A vector with all components equal to zero
- Represents no direction or magnitude

Unit
- vectors with a magnitude of one, used to show direction
- Represent direction without considering magnitude.

Visual Representation of vectors



- vectors are visually represented by arrows, showing both magnitude and direction.
- Its magnitude is represented by the arrow's length.
- Its direction is indicated by the arrow's orientation.

Same for 3-dimensional.

Applications of vector Operations. (1.4)

Harnessing the power of vector operations

Machine Learning
- Data point as vector
- Measure similarity
- Grap similar points

Computer Graphics
- Enable 3D rendering
- Simulate lighting
- Life-like visuals

Physical simulations
- Simulate motion
- Give direction + value
- Precise simulation

Date
DELTA Pg No.

The role of matrices in Data Science.

rows - m
Columns - n

Data Representation
- Organises data into rows and columns
- simplifies processing

Transformation
- PCA reduces features, keeps variance
- Scaling standardises data.

Machine Learning
- Regression finds co-efficients
- PCA extracts key components.

Types of Probability.

Theoretical. Based on logic and models — Dice
Empirical From observations and data — weather forecasting
Subjective Based on personal judgement — Estimating team victory.

Types of measuring Data.

a) Nominal Scale - categories without order
b) ordinal scale categories with order
c) Interval scale equal intervals, no true zero
d) Ratio scale Equal intervals, with a true zero

median
mean
mode

Types of Statistics.

- measures of Central Tendency
- Measures of dispersion (spread of data)
- Data visualisation

Descriptive statistics
- It summarises & describes the key characteristics of dataset, providing an overview of it main features
- Uses measures like mean, median and mode.

Inferential statistics
- It makes prediction about a population based on sample data, allowing for generalisations and conclusions about larger groups.
- Involves hypothesis testing & estimation for informed decision making.

Date
DELTA Pg No.

Applications of Central Tendency and Dispersion.

(A) Descriptive Analysis
Summarises key dataset characteristics highlighting central tendency and spread.

(B) Hypothesis Testing
Determines if observed differences are statistically significant

(C) Risk Assessment
Assesses data variability and the potential for extreme values, crucial for predicting risk.

Types of Matrix.

Square matrix
Identity matrix
Zero matrix
Diagonal matrix
Transpose
Determinant

Principal Component Analysis (PCA)

It is a statistical technique used to simplify large datasets by reducing their dimensions while retaining most of the original information. It transforms correlated variables into a smaller set of uncorrelated variables called principal components. This method is widely used in data preprocessing for machine learning, pattern recognition, and image processing.

In dimensionality reduction, techniques like principal components analysis (PCA) rely on matrices to reduce the number of features while preserving variance. Similarly, feature scaling uses matrix operations to standardise and normalise data, ensuring features are on a comparable scale.

Date 6/4/25
DELTA Pg No.

In machine learning, matrices underpin key calculations. In linear regression, they help solve for model coefficients. PCA uses matrix decomposition to identify critical components, while neural networks depend on matrix operations to manage weights and activations across network layers, enabling fast computation of large datasets.

LIVE SESSION

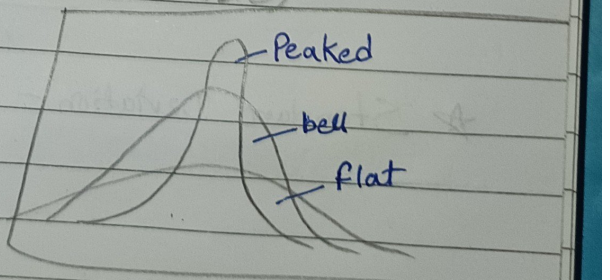
Math/stats for DS

Week 1

* Agenda
  a) Prob. distribution
  b) Basics of statistics → Discriptive, Inferential.
  c) Simple discriptive analysis.
     Houseprice → R. (as a tool)

* Descriptive Statistics.
  Describing the data → Measure of numbers (frequency)
  → Measure of central Tendency
  (Mean, median, mode.)
  Avg, Central value, Highest frequency

* Data (Prob. Distribution)
  a) Quantitative (Measured)
     ↙ Discrete
     ↘ Continuous.
  b) Qualitative. (non-measurable)



ND
1) Bell shaped Curve
2) Mean & std dev
   ↳ the curve is symmetric about its mean
   ↳ Std dev impacts the peakedness.
3) Area under the curve = prob. associated within the range.
4) Asymptotic → Tails extend to the infinity
5) Total Area Under Curve = 1

DELTA Pg No.

Scales of Measurement.

A) Nominal Scale.

Names, Labels, qualities.

No ranking / orders / Numbers

Mode / frequency / chi-square.

B) Ordinal Scale

same as nominal but there is rank / order.

Median / Mode.

C) Interval Scale

- Equal interval
- No True zero
- Equal difference b/w intervals.

Mean, SD, ANOVA, Regression.

D) Ratio Scale

- true zero
- "+, −, ×, ÷"

Mean, Median, Mode, SD, Geometric mean, Harmon Mean.

- Discrete Data. (Whole number) (Bar Graph)

a) Bernoulli Distribution — a process with 2 possible outcome
b) Uniform Distribution
c) Binomial Distribution —
d) Poisson Distribution — Models the prob. of events occuring in fixed intervals.

(Columns / variables / attributes)

Central Limit Theorem

Random Variable

a random process, where each outcome is associated with some numbers.

Add N numbers of samples of this variable:

$x_1 + x_2 + x_3 \ldots\ldots + x_n$

The distribution of this sum looks more like a bell curve as $N \to \infty$

Mean, $\mu = E[x] = \sum_x P(X = x) \cdot x$

Variance, $Var(x) = E[(X - \mu)^2]$

Standard deviation, $\sigma = \sqrt{Var[x]}$

Normal Distribution

$$Pdf = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}$$

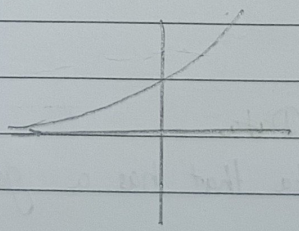
mean = distribution is symmetric about mean

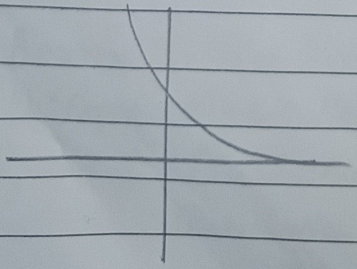
std = flatness or peakness

a) Bell shaped
b) depends on mean and std.
c) Area under curve ⇒ prob. associate with range
d) Asymptotic curve ⇒ tails extends till ∞
e) Total area under the curve = 1

Exponential Growth

e^x





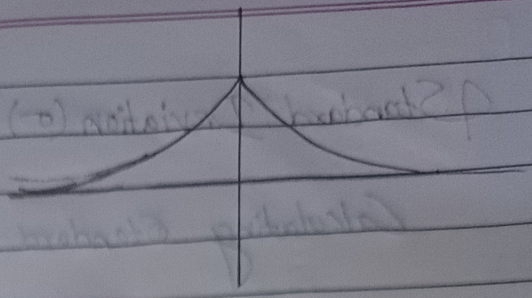
e^{-x}

Exponential decay.

Exponential distribution.

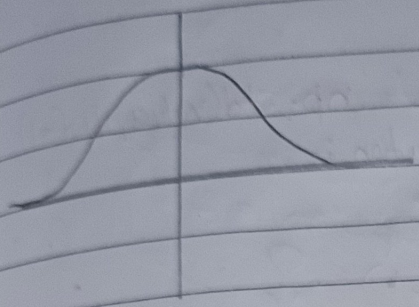
Date
DELTA Pg No.

$e^{-|x|}$



e^{-x^2} bell curve.

normal distribution



Percentile, a value below which a certain percentage of observation lie.

$$\text{Percentile} = \frac{\text{value}}{n} \times 100$$

$$\text{Value} = \frac{\text{Percentile}}{100} \times (n+1)$$

↳ Index position

Removing Outliers (Generic)

Lowerfence ⟷ Upperfence

$Lf = Q_1 - 1.5(IQR)$

$Uf = Q_3 + 1.5(IQR)$

$IQR = Q_3 - Q_1$

Variation

There are different measur

commonly used are :

- Range (largest value − s
- Quartiles and Percentile
- Interquartile Range
- Standard Deviation.

4 parts

Quartiles.

$q = [0, 0.25, 0.5, 0.75, 1]$

np.quantile(values, q)

Between Q_0 and Q_1 are 25% lowest values in data.

b/w Q_1 & Q_2 are next 25% and so on.

Five number Summary

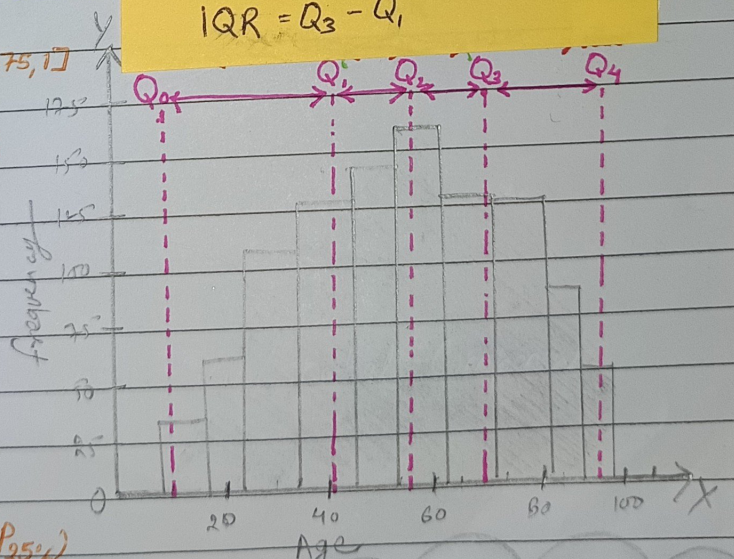
Q_0 - smallest value.

Q_1 - separating value ($P_{25\%}$)

Q_2 - middle value (median) ($P_{50\%}$)

Q_3 - separating value. ($P_{75\%}$)

Q_4 - Largest value ($P_{100\%}$)



np.percentile(values, 65)

↑ 65th percentile.

Outliers — extreme & unequal values in data.

Variation

There are different measures of variation. The most commonly used are:

- Range (largest value − smallest value) `np.ptp(values)`
- Quartiles and Percentiles ← 4 parts, → 100 parts
- Interquartile Range (IQR) $[Q_1 - Q_3]$ `stats.iqr(values)`
- Standard Deviation.

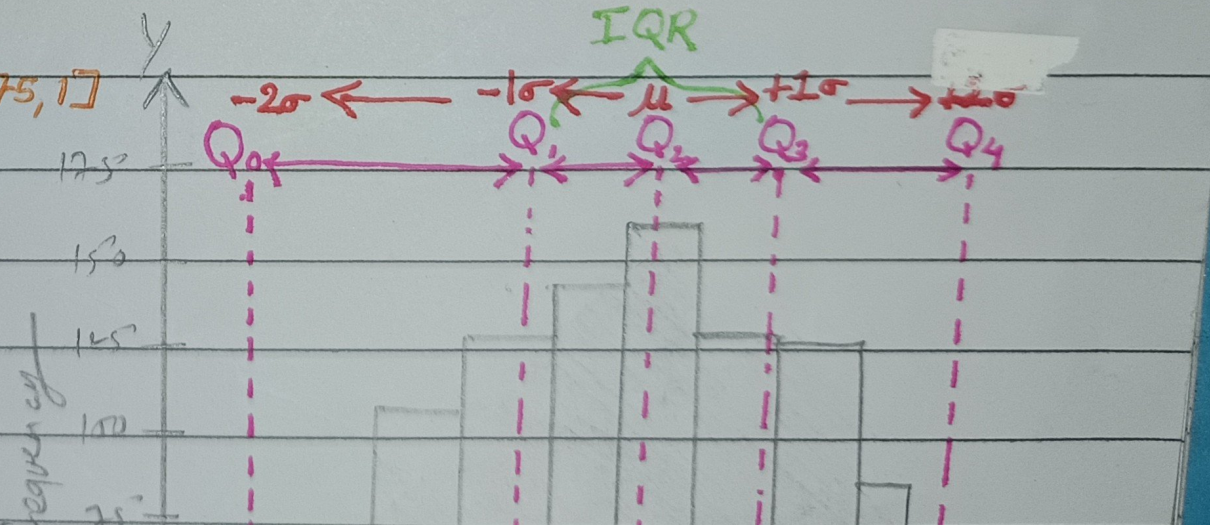
Quartiles.

`np.quantile(values, x)`

$x = [0, 0.25, 0.5, 0.75, 1]$

Between Q_0 and Q_1 are 25% lowest values in data.

b/w Q_1 & Q_2 are next



Date
DELTA Pg No.

Standard Deviation (σ)

$\sqrt{\text{variance}}$

np.std(values)

Calculating Standard deviation :-

$$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{n}}$$

np.std(values, ddof=

Calculating sample standard deviation :-

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

Why sample variance is $(n-1)$?

random variable, $X \longrightarrow Z$ (statistics)

Z = deviation of x from its mean in no° of its std.

$$Z = \frac{x - \mu}{\sigma}$$

Always = mean : 0
std : 1

Prob. density function

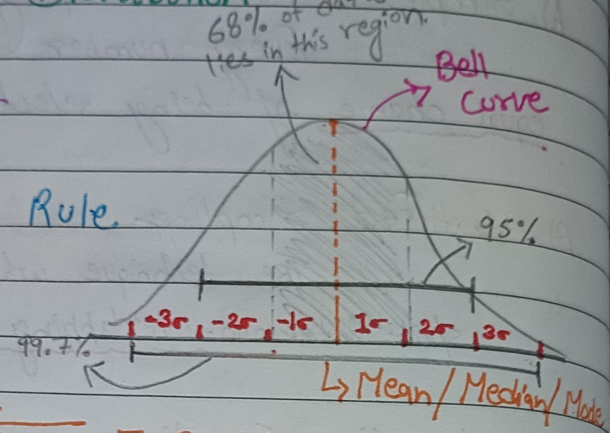
$$PDF = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2}$$

Distribution.

(mean) $\bar{z} = 0$

(std) $\bar{z} = 1$

A) Gaussian / Normal Distribution



Emperical formula :-

68 - 95 - 99.7% Rule

$$Z\ score = \frac{x_i - \mu}{\sigma}$$

ML model or data

Standardization ← Z-score

{1, 2, 3, 4, 5, 6, 7} → Normal Distribution

⇓ Z-score ⇓

{-3, -2, -1, 0, 1, 2, 3} → Standard Normal Distribution. {μ = 0, σ = 1}

* y ~ std (μ = 0, σ = 1) *

It is the process of rescaling the data so that it has :-

mean = 0

STD = 1.

$$x_{nom} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Bounded data

NN, KNN, CNN

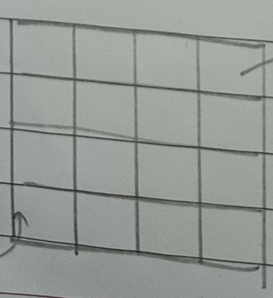
Normalization ← Scaling

It refers to scaling the data to fit within a specific range — usually it's [0 to 1].

Minmax Scalar → (0 to 1)

CNN → Image classification.

Pixels

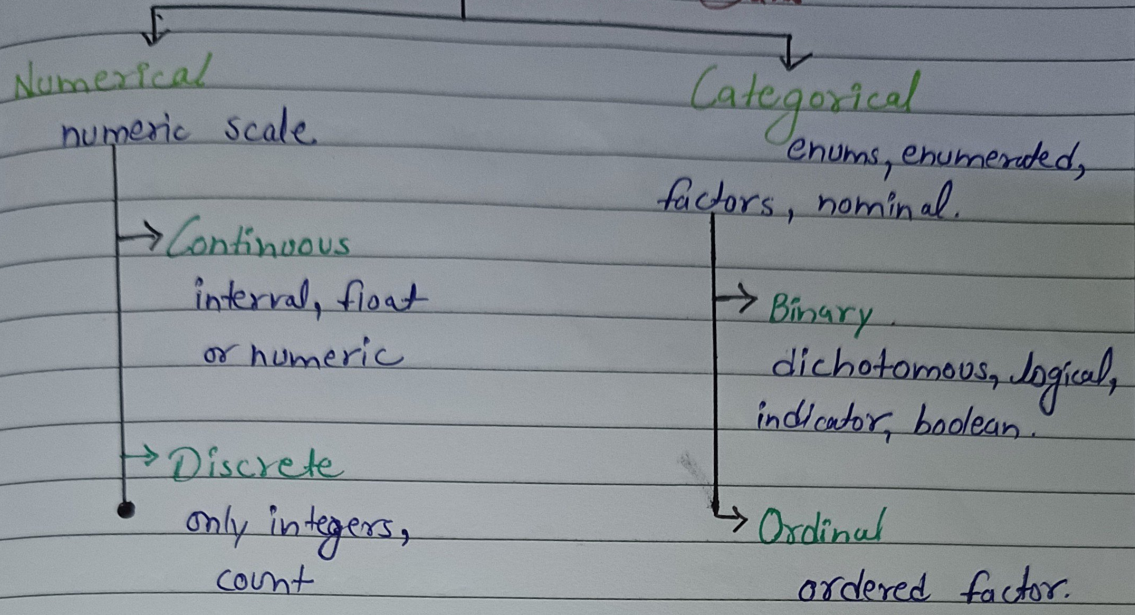


0-255 ⇓ Minmax Scalar → 0-1

Normalization

Date
DELTA Pg No.

Elements of structured Data



Estimation of Location

- Mean (average)
  the sum of all values divided by number of values.
- weighted mean (weighted average)
  sum of all values times a weight divided by sum of weights.
- Median (50th percentile)
  value such that one-half of data lies above & below.
- Percentile (quantile)
  value such that p percent of data lies below.
- weighted median
  value such that one-half of sum of weights lies above & below the sorted data.
- Trimmed mean (truncated mean)
  average of all values after dropping a fixed number of extreme values.
- Robust (resistant)
  Not sensative to extreme values.
- Outlier (extreme value)
  a data value that is too differ from other values.