* What is statistics?

→ Statistics is the science of collecting, oragamizing, analyzing, interpreting, and presenting data.

* Use case in Data science:-

→ Data Analysis :- understand trends and patterns
→ predictive Modeling :- Build ml models
→ Decision Making :- optimize business strategies
→ Data cleaning :- Detect outliers, missing value
→ A/B Testing :- Evaluate the impect of changes.

* What is Data in statistics?

→ Data refers to raw facts, figures or information collected for analysis.

* Applications of statistics:-

→ Healthcare :- patient analysis
→ Finance :- Risk assessment, fraud detection
→ Marketing :- customer behavior analysis
→ sports :- performance tracking, game strategy
→ Goverment :- unemployment rates, economic indicators.

* Types of Data :-

1. Numerical Data :-

→ Measurable quantities

→ Examples :-

→ Age, salary, temperature.

(i) Discrete :- countable numbers

→ Example :- Number of children

(ii) Categorical :-

(iii) continuous :- measurable quantities with decimals

→ Example :- Height, weight.

2. Categorical Data:-

→ Descriptive lables or names.

→ Examples:-

→ Gender, color, country.

(i) Nominal :- No specific order

→ Example :- male / Female, Red / Blue.

(ii) Ordinal :- ordered categories

→ Example :- Low / Medium / High, Ratings.

* Representation of Data - Different Graphs and patterns.

→ Bar chart :- categorical data
→ Pie chart :- percentage distribution
→ Histogram :- Frequency of numerical data.
→ Line Graph :- Trends over time
→ Box plot :- spread and outliers.
→ Scatter plot :- correlation between two variables.

* population:-

→ The entire group you want to study

→ Example:-

→ All students in India.

* Sample:-

→ A smaller part selected from the population.

→ Example:-

→ 500 students selected from Gujarat.

* Types of statistics:-

(i) Descriptive statistics
(ii) Inferential statistics

(i) Descriptive statistics:-

→ To summarize and describe the main features of a dataset.

(ii) Inferential statistics:-

→ To make predictions or inferences about a population based on a sample.

(i) Descriptive statistics:-

→ central Tendency
→ Dispersion
→ Shape of Distribution
→ Data presentation.

_/_/_

* Measures of central Tendency :-

→ Measures of central Tendency are statistical metrics that represent the center point or typical value of a dataset.

(i) MEAN (Average) :-

→ The mean is the sum of all values divided by the number of values.

→ Formula :-

→ Mean = Sum of all values / number of values.

→ Example :-

→ data = [10, 20, 30, 40, 50]
mean = sum(data) / len(data)
print("Mean:", mean)

→ Output :-

→ Mean :- 30

(ii) MEDIAN (Middle value) :-

→ The median is the middle value in a sorted list.

→ If even number of values, it's the average of the two middle values.

→ Example :-

(i) odd count :-

→
```
data = [10, 20, 30, 40, 50]
data.sort()
median = data[len(data) // 2]
print("Median:", median)
```

→ output :-

→ median :- 30

(ii) Even count :-

→
```
data = [10, 20, 30, 40]
mid = len(data) // 2
median = (data[mid-1] + data[mid]) / 2
print("median:", median)
```

→ output :-

→ median :- 25

* MODE (Most Frequent value) :-

→ The mode is the value that appears most frequently.

→ Example :-

→ from statistics import mode
data = [10, 20, 30, 20, 30, 30, 30, 40]
print("Mode:", mode(data))

→ output :-

→ Mode :- 30

* Bonus with Pandas :-

→ import pandas as pd

df = pd.DataFrame({'score': [85, 90, 75, 90, 60, 90]})

print("Mean:", df['score'].mean())
print("Median:", df['score'].median())
print("Mode:", df['score'].mode()[0])

* output :-

→ Mean :- 81.66666
Median :- 87.5
Mode :- 90

* Measure of Dispersion :-

→ Measures of dispersion tell us how spread out or scattered the data values are in a dataset.

(i) Range :-

→ The difference between the maximum and minimum value in the dataset.

→ Formula :-

→ Range = Maximum value - Minimum value

→ Example :-

→ data = [10, 20, 30, 40, 50]
range_val = max(data) - min(data)
print("Range:", range_val)

→ Output :-

→ Range: 40

(ii) variance:-

→ variance shows how far each value in the dataset is from the mean, on average.

→ It's the average of the squared differences from the mean.

→ Formula:-

→ $\text{Variance} = \frac{\Sigma(x_i - \bar{x})^2}{n}$

→ x_i = each value

→ $\bar{x}$ = mean of values

→ n = total number of values.

→ Example:-

→
```
import statistics
data = [10, 20, 30, 40, 50]
variance = statistics.variance(data)
print("variance:", variance)
```

→ output :-

→ Variance :- 250

(iii) Standard Deviation :-

→ The square root of the variance.

→ It shows how much, on average, the values deviate from the mean, in original units.

→ Formula :-

→ Standard Deviation = $\sqrt{\text{variance}}$

→ Example :-

→
```
import statistics
data = [10, 20, 30, 40, 50]
std_dev = statistics.stdev(data)
print("Standard Deviation:", std_dev)
```

→ Output :-

→ Standard Deviation: 15.811388300841896

* Gaussian (Normal) Distribution:-

→ A Gaussian distribution (Normal distribution) is a bell-shaped curve that shows how data is distributed.

→ Most of the data points are concentrated arround the mean, with fewer values farther away.

* Characteristics :-

→ Shape :- Bell-shaped

→ Mean = Median = Mode :- All three are at the center

→ Spread :- controlled by standard deviation.

→ Tails :- Extend infinitely in both directions.

* Real-life Examples :-

→ Heights of people

→ Marks of students

→ Blood pressure

→ IQ scores

* Normal Distribution Formula :-

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

→ μ = mean

→ σ = standard deviation

→ e = Euler's number.

→ $\sqrt{2\pi}$ = constant factor

→ x = Data point

→ $(x-\mu)$ = Distance from mean

→ $\frac{x-\mu}{\sigma}$ = standardized value

→ $\left(\frac{x-\mu}{\sigma}\right)^2$ = squared z-score

→ $e^{-\frac{1}{2}(\ldots)}$ = Exponential decay.

* Skewness :-

→ Skewness is a statistical measure that describes the asymmetry of a data distribution around its mean.

→ It helps Identify if the data is symmetrically distributed or tited towards one side.

* Types of skewness :-

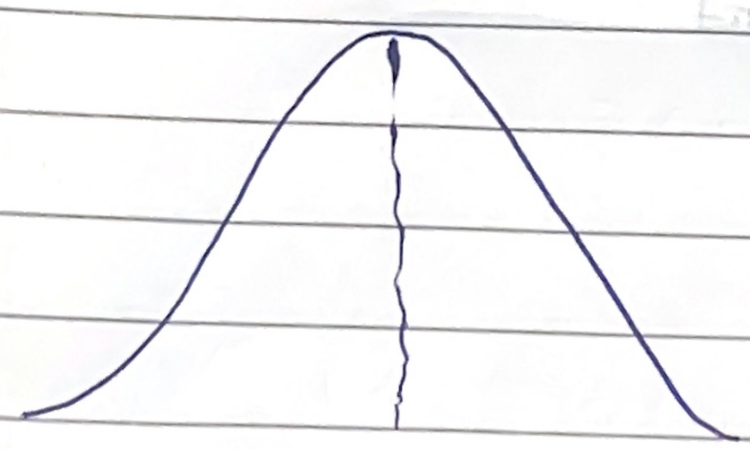
(i) Symetrical distribution :- (zero - skewness)

→ Data is evenly distributed

→ Bell curve

→ Mean ≈ Median

→



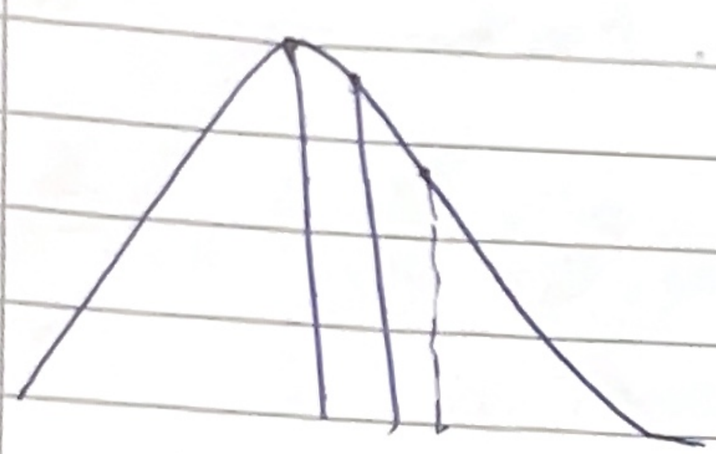
(ii) positive skewed :- (Right - skewed)

→ Tail on the right is longer

→ skewed to the right

→ mean > median

→



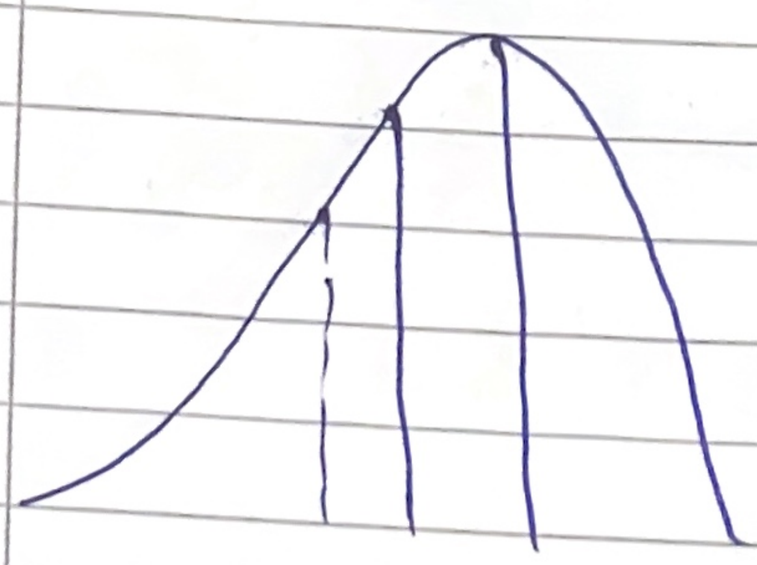
(iii) Negative skew :- (left - skewed)

→ Tail on the left is longer

→ skewed to the left

→ mean < median

→



* How to calculate skewness :-

→ Formula :-

→ $\text{Skewness} = \dfrac{3(\text{Mean} - \text{median})}{\text{Standard Deviation}}$

→ using python

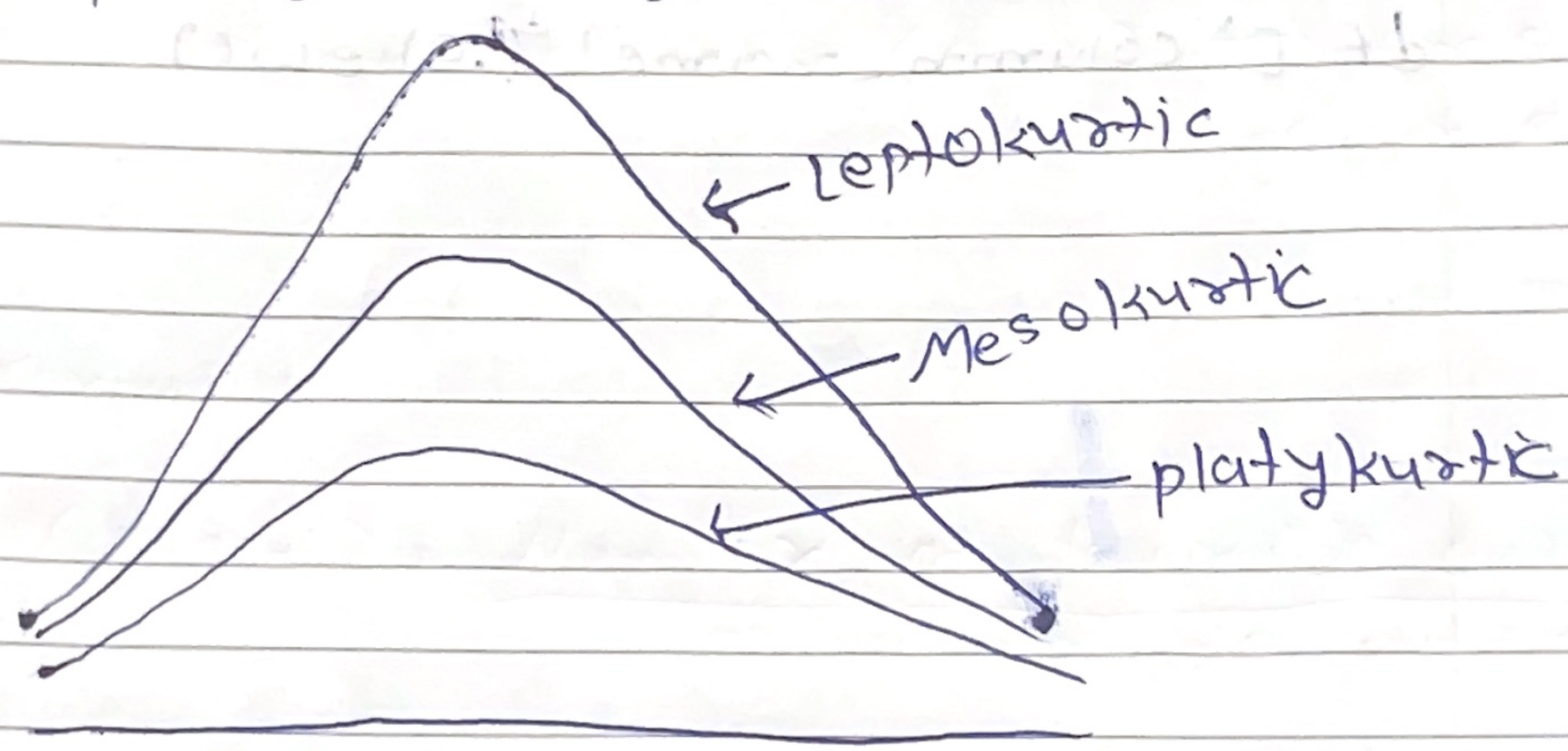
→ import pandas as pd
df ['column_name'].skew()

* Kurtosis :-

→ Kurtosis measures the "tailedness" of a distribution - how heavy or light the tails are compared to a normal distribution.

→ It tells us how likely it is to get outliers.

* Types of Kurtosis :-



(i) platykurtic :-

→ Flatter peak, lighter tails.

→ Broad peak, thin tails shape

→ outliers :- Low

(ii) Mesokurtic :-

→ Normal distribution

→ Shape :- Moderate peak and tails

→ outliers :- Average

(iii) Leptokurtic :-

→ More peaked, heavier tails shape

→ Shape :- high peak, fat tails

→ outliers :- High

* How to calculate kurtosis in python.

→ from scipy.stats import kurtosis

```
kurtosis(data)
kurtosis(data, fisher = False)
```