University of Sulaimani
College of Science
Computer Department
4th Stage

Data Science Management

Statistics for Data Science

Class 4

Theoretical and practical lectures

سؤال مهم في الامتحان :- لماذا نستخدم (linear - Algebra) في Data Science ؟

Assist. Prof. Dr. Miran Taha Abdullah
2025-2026

Class Agenda

- Introduction to Statistics in Data Science
- Descriptive statistics
- Inferential statistics
- Python Demonstrations

Algorithms (linear, statistics....)

لازم تعرف الفرق بين Tools و model و module ؟

وحدة فيك

هي اللغة التي تستخدمها (python)

Introduction to Statistics in Data Science

Why Statistics Matter in Data Science:

- In data science, **statistics** are essential for **analyzing** and **interpreting** data. From **cleaning** data to building **predictive** models, statistical techniques form the backbone of data-driven decisions.

Examples:

يهتم به

تقييم

- **AI Models** (training data, performance evaluation),

- **Machine Learning** (feature selection, model validation),

سؤال مهم لماذا نسوي ؟

- **Big Data** (trend analysis, anomaly detection).

مهم

- **Descriptive statistics** help summarize data for immediate insights.

مثل min و max و averge

مثل خاتمة بعد الفهم (conclusion)

- **Inferential statistics** enable us to make predictions and test hypotheses beyond the sample.

فرضيات

- Python libraries like **Pandas**, **SciPy**, **Statsmodels**, and **Scikit-learn** are essential for **performing statistical analysis** in data science.

- Understanding both branches of statistics is crucial for building **accurate**, **reliable models** in computer science and machine learning.

Pandas: Used for **data manipulation and exploration**. It provides powerful data structures like DataFrames to clean, organize, and summarize data before analysis.

Example: Calculating mean, median, or grouping data by categories.

SciPY: Provides **advanced scientific and statistical functions**. It’s useful for probability distributions, hypothesis testing, and numerical integration.

Example: Performing t-tests, chi-square tests, or computing correlation coefficients.

Statsmodels – Specialized for **statistical modeling and inference**. It supports regression analysis, time series analysis, and statistical tests.

Example: Building linear regression models or performing ANOVA tests.

Scikit-learn: Primarily used for **machine learning**, but also supports statistical tasks like regression, classification, and clustering. It’s great for predictive modeling using statistical techniques.

Example: Predicting outcomes with linear regression or detecting anomalies using clustering algorithms.

Key Concepts: Descriptive Statistics

في الامتحان يجيبلك سؤال وهو مسألة ويقلك ثلوث
تحلها؟ هات لازم تستخدم

1- Measures of Central Tendency:

- متوسط **Mean**: The average of all data points.
- وسيط **Median**: The middle value when data is sorted.
- منوال **Mode**: The most frequently occurring value.

2- Measures of Dispersion:

- **Range**: Difference between the maximum and minimum values.
- **Variance**: Average of the squared differences from the mean.
- **Standard Deviation**: Square root of the variance; measures data spread.

3- Skewness and Kurtosis:

- الانحراف **Skewness**: Describes the asymmetry of the data distribution.
- **Kurtosis**: Measures the "tailedness" of the data distribution.

دورته بنظام (الاستدلالي)

Key Concepts: Inferential Statistics

لماذا نستخدم Sample بوجود Population ؟

Sampling (Benefit) → easy to manage
→ Time-consuption
→ cost-effective

1- Sampling and Sampling Distribution:

- Importance of sample size and randomness.
- Central Limit Theorem

فرضيه اختبار

2- Hypothesis Testing:

اذا ما تغير الجديد عن القديم

هذا يعني

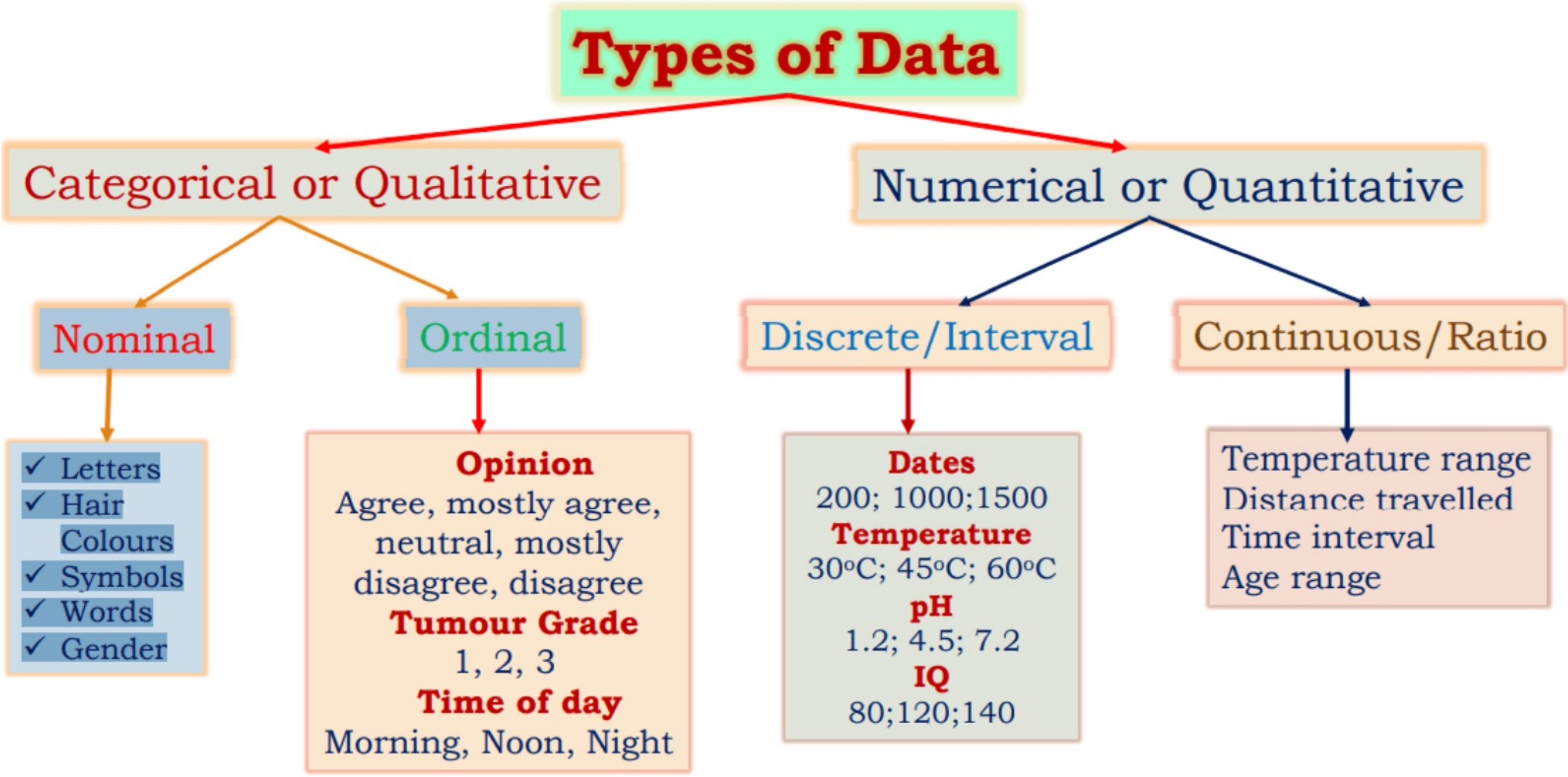
ام اذا تغير

- **Null Hypothesis (H_0)**
- **Alternative Hypothesis (H_1)**
- **p-value**
- **Types of Tests: t-tests** (one-sample, independent, paired). **ANOVA**: Comparing means across multiple groups. **Chi-square Test**: Testing categorical variables.

3- Confidence Intervals:

4- Regression Analysis:

- **Linear Regression**: Predicting continuous outcomes.
- **Logistic Regression**: Predicting binary outcomes.



Types of Data

Categorical or Qualitative

Nominal

- ✓ Letters
- ✓ Hair Colours
- ✓ Symbols
- ✓ Words
- ✓ Gender

Ordinal

Opinion
Agree, mostly agree, neutral, mostly disagree, disagree

Tumour Grade
1, 2, 3

Time of day
Morning, Noon, Night

Numerical or Quantitative

Discrete/Interval

Dates
200; 1000;1500

Temperature
30°C; 45°C; 60°C

pH
1.2; 4.5; 7.2

IQ
80;120;140

Continuous/Ratio

Temperature range
Distance travelled
Time interval
Age range

1- Measures of Central Tendency:

These help to describe the center of the data.

Mean: The average of all data points.

- *Example*: The average exam score of 50 students is 80 out of 100.

Median: The middle value when the data points are ordered.

- *Example*: For the exam scores, if the sorted values are [50, 60, 70, 80, 90], the median score is 70.

Mode: The most frequent value in the dataset.

- *Example*: In a set of [1, 2, 2, 3, 3, 3, 4], the mode is 3 because it occurs most frequently

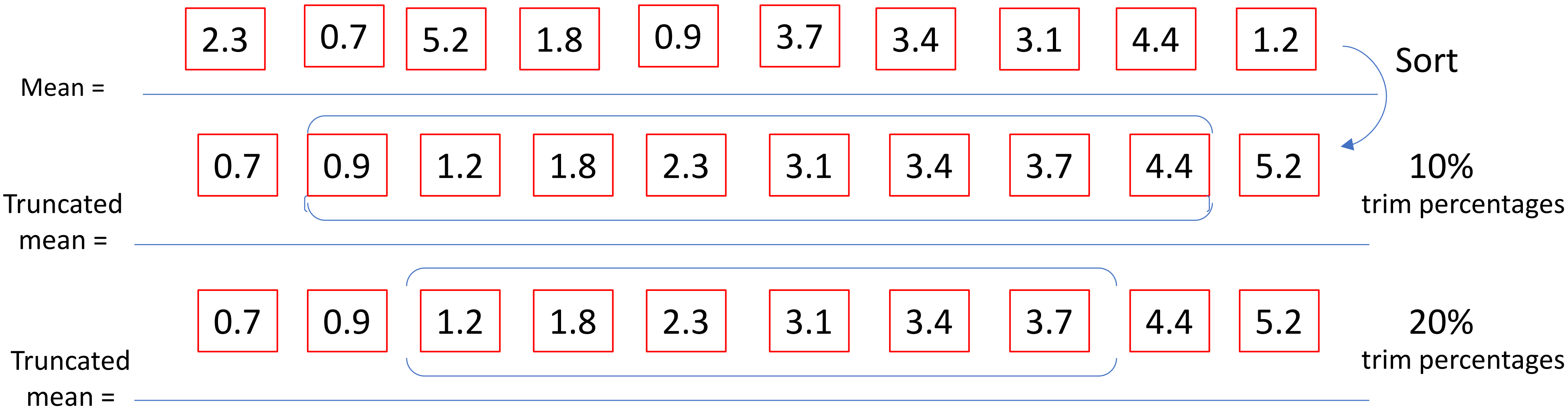
The Mean

- The mean is a common and intuitive way to summarize a set of numbers it might simply called the “average”. The mean is the sum of all of the data elements divided by how many elements there are

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} : \frac{\sum_{i=1}^{n} x_i}{n}$$

← مهم انتبه في الأمتحان

A trimmed mean (similar to an adjusted mean) is a method of averaging that removes a small designated percentage of the largest and smallest values before calculating the mean. After removing the specified **outlier** observations(switch to next slide) , the trimmed mean is found using a standard arithmetic averaging formula. The use of a trimmed mean helps eliminate the influence of outliers or data points on the tails that may unfairly affect the traditional or arithmetic mean.



2.3 0.7 5.2 1.8 0.9 3.7 3.4 3.1 4.4 1.2

Mean =

Sort

0.7 0.9 1.2 1.8 2.3 3.1 3.4 3.7 4.4 5.2

10%
trim percentages

Truncated mean =

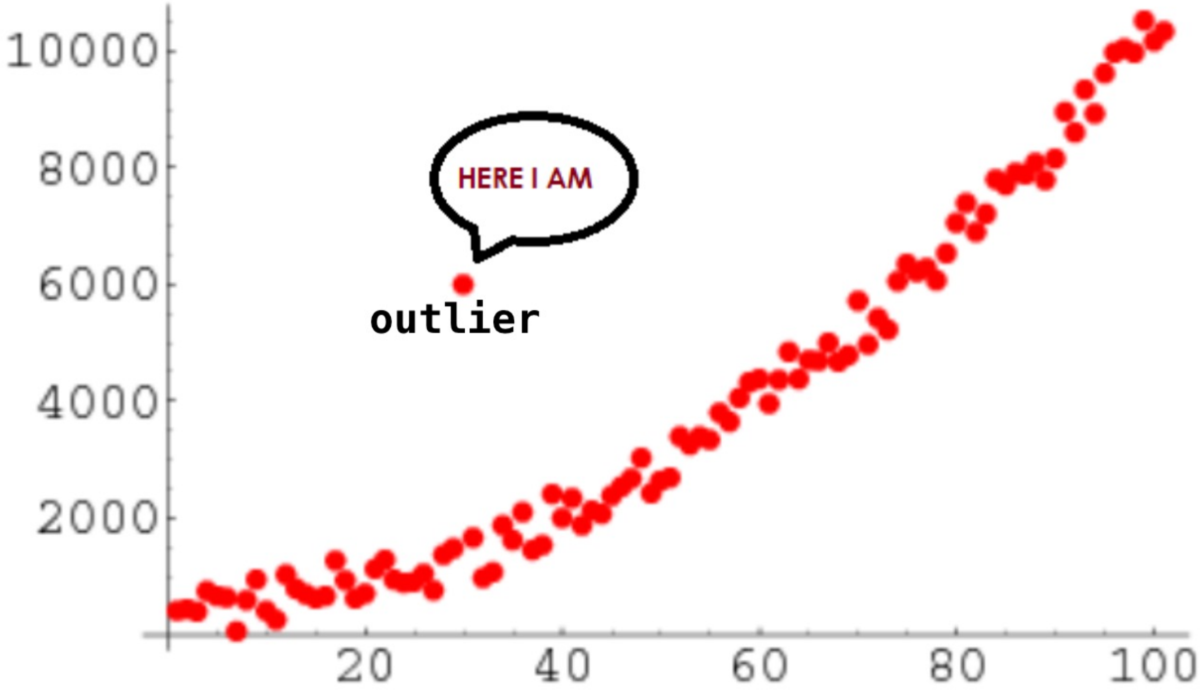
0.7 0.9 1.2 1.8 2.3 3.1 3.4 3.7 4.4 5.2

20%
trim percentages

Truncated mean =

$$\text{Trimmed mean} = \overline{x} = \frac{\sum_{i=p+1}^{n-p} x_{(i)}}{n-2p}$$

An **outlier** is a data point in a data set that is distant from all other observations. A data point that lies outside the overall distribution of the dataset

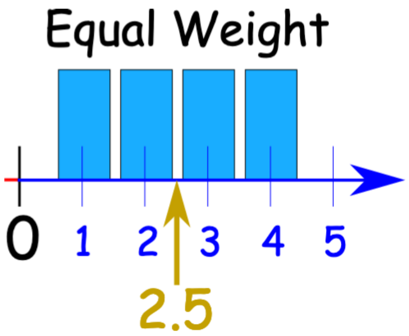


[illegible] من [illegible]

- What causes the outliers?
- Impact of the outlier
- Methods to Identify outliers

The weighted mean is a type of mean that is calculated by multiplying the weight (or probability) associated with a particular event or outcome with its associated quantitative outcome and then summing all the products together.

$$\text{Weighted mean} = \bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i}^{n} w_i}$$



$$\text{Mean} = \frac{1 + 2 + 3 + 4}{4} = \frac{10}{4} = 2.5$$

Weights

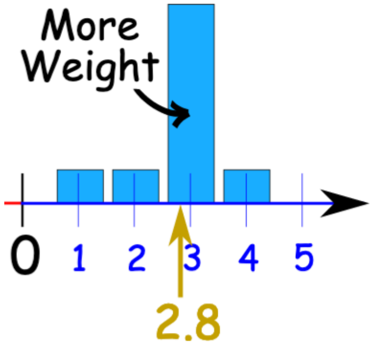
Each of those numbers has a "weight" of ¼”

$$\text{Mean} = \tfrac{1}{4} \times 1 + \tfrac{1}{4} \times 2 + \tfrac{1}{4} \times 3 + \tfrac{1}{4} \times 4$$

$$= 0.25 + 0.5 + 0.75 + 1 = \mathbf{2.5}$$

Now let's change the weight of **3** to 0.7, and the weights of the other numbers to 0.1

$$\text{Mean} = 0.1 \times 1 + 0.1 \times 2 + 0.7 \times 3 + 0.1 \times 4$$
$$= 0.1 + 0.2 + 2.1 + 0.4 = \mathbf{2.8}$$



Example: Sam wants to buy a new camera, and decides on the following rating system:

- Image Quality **50%**
- Battery Life **30%**
- Zoom Range **20%**

The Sonu camera gets 8 (out of 10) for Image Quality, 6 for Battery Life and 7 for Zoom Range

The **Conan camera** gets 9 for Image Quality, 4 for Battery Life and 6 for Zoom Range

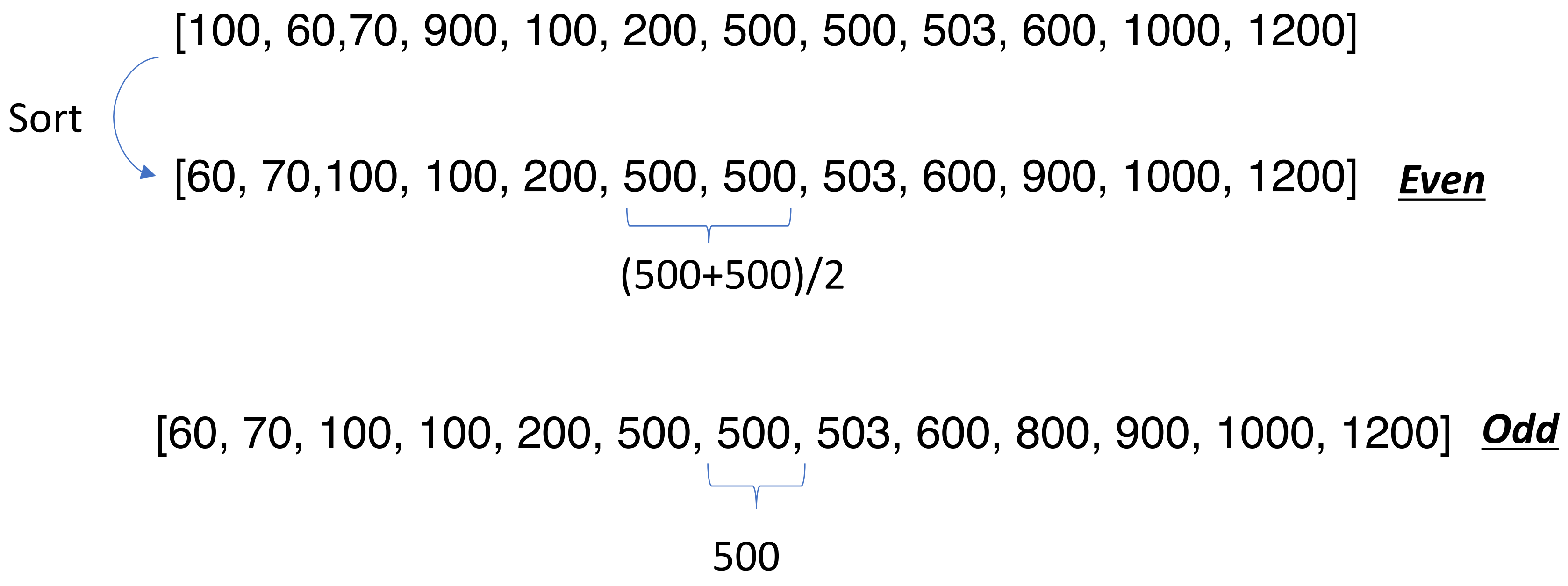
Which camera is best?

Sonu: 0.5 × 8 + 0.3 × 6 + 0.2 × 7 = 4 + 1.8 + 1.4 = **7.2**
Conan: 0.5 × 9 + 0.3 × 4 + 0.2 × 6 = 4.5 + 1.2 + 1.2 = **6.9**

Sam decides to buy the Sonu.

Median

- The **median** of a collection of numbers is another kind of average. To find the median, we sort the numbers in ascending order. If the length of the list of numbers is odd, the number in the middle of the list is the median. If the length of the list of numbers is even, we get the median by taking the mean of the two middle numbers.



[100, 60,70, 900, 100, 200, 500, 500, 503, 600, 1000, 1200]

Sort

$[60, 70,100, 100, 200, \underbrace{500, 500}_{(500+500)/2}, 503, 600, 900, 1000, 1200]$ ***Even***

$[60, 70, 100, 100, 200, 500, \underbrace{500,}_{500} 503, 600, 800, 900, 1000, 1200]$ ***Odd***

2- Measuring the Dispersion

Measuring the dispersion tells us how far away the numbers in a set of data are from the mean of the data set. We'll learn to calculate three different measurements of dispersion: range, variance, and standard deviation.

Range: The difference between the maximum and minimum values.

Variance: The average of the squared differences from the mean.

Standard Deviation: The square root of the variance. It indicates how spread out the data points are.

Coefficient of Variation (CV) is a standardized measure of dispersion that shows how much variability exists in relation to the mean of the dataset.

Finding the Range of a Set of Numbers

The **Range** is the **difference between the largest and smallest values** in a dataset.

$$\text{Range} = \text{Maximum Value} - \text{Minimum Value}$$

Benefit:
It gives a **quick idea of how spread out the data** is.

Example:
If students' test scores range from **40 to 90**, → **Range** = 90 − 40 = **50**

This means there is a **50-point spread** between the lowest and highest score.

Interpretation:
A **larger range** = more variation;
a **smaller range** = data values are more consistent.

Finding the Variance

Variance measures **how far each data point is from the mean**, on average, by taking the **average of squared differences** from the mean.

$$\sigma^2 = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}{N}$$

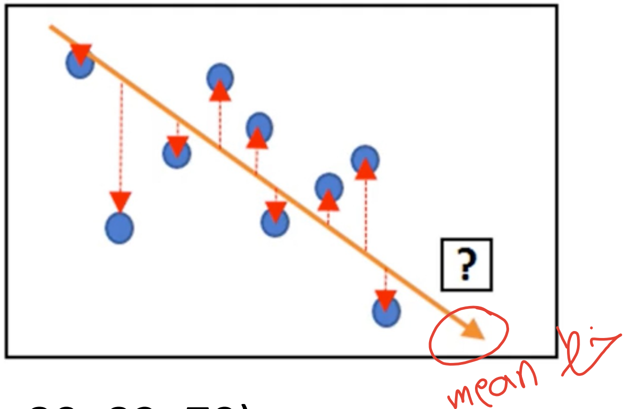
x_i = Each value in the data set

$\overline{x}$ = Mean of all values in the data set

N = Number of values in the data set

Benefit:

Shows **data spread** and **consistency** of data points.



Example:

Suppose two teachers give weekly quizzes.

In Class A, students' scores are close to each other (e.g., 78, 80, 82, 79).

In Class B, scores vary widely (e.g., 50, 70, 90, 100).

Class A has a **smaller variance**, meaning students' performance is **more consistent.**

Interpretation:

High variance → data points vary widely (more fluctuation).

Low variance → data points are close to the mean (Consistent data).

Finding the Standard deviation (SD)

The standard deviation is calculated as the square root of variance by determining each data point's deviation relative to the mean.

Benefit:
Easier to interpret than variance because it’s in the **same units** as the original data.

$$\text{Standard Deviation} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$$

where:

$x_i =$ Value of the i^{th} point in the data set

$\overline{x} =$ The mean value of the data set

$n =$ The number of data points in the data

Example:
In exam scores, if SD is **low**, most students performed close to the average.

If SD is **high**, performance varied a lot.

Interpretation:
Small SD → consistent performance.
Large SD → large differences among values.

Finding the Co-efficient of Variation (CV)

The coefficient of variation (relative standard deviation) is commonly used to compare the data dispersion between distinct series of data

$$\mathrm{CV} = \frac{\sigma}{\mu}$$

where:

$\sigma = \text{standard deviation}$

$\mu = \text{mean}$

Benefit:

Used to **compare variability** across datasets with different scales or units.

Example:

Machine A: Mean output = 100 kg, SD = 5 → CV = 5%

Machine B: Mean output = 200 kg, SD = 40 → CV = 20%

Machine A is **more consistent**.

Interpretation:

A **smaller CV** means **more stability** and **less risk**.

Calculating the Correlation Between Two Data Sets

Correlation measures the **strength and direction of a relationship** between two variables.
It ranges from **–1 to +1**.

+1 → Perfect positive correlation

0 → No correlation

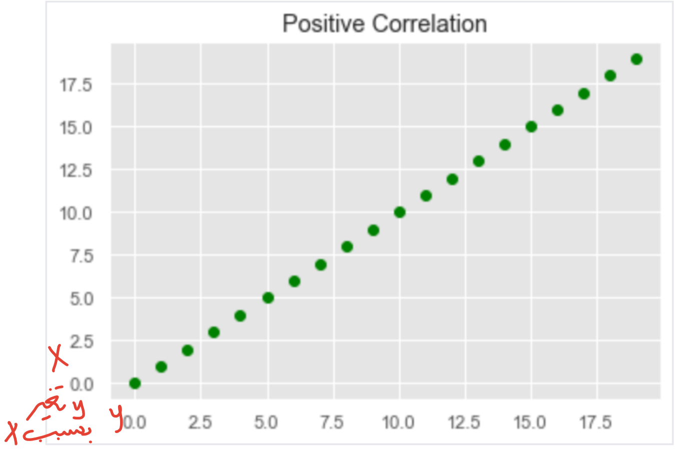
–1 → Perfect negative correlation

يجب سؤال وانت لازم ترجع وتبين
العلاقة بين x و y ؟

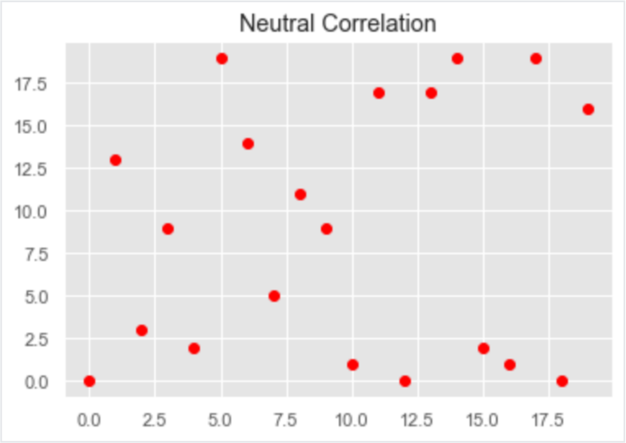
Benefit:
Helps in **predicting** one variable based on another.

Example:

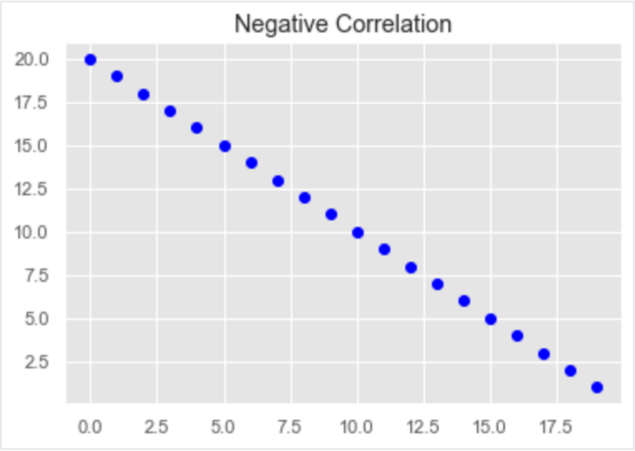
r = 0.9 → strong positive relation



r = 0 → no linear relationship



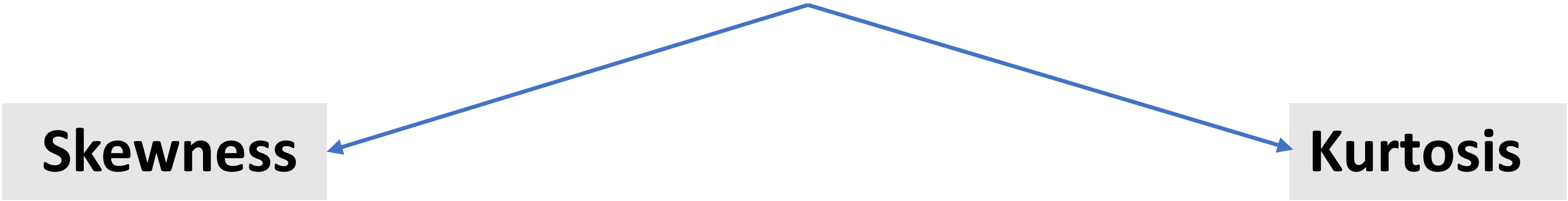
r = –0.8 → strong negative relation



$$\text{correlation} = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left(n\sum x^2 - \left(\sum x\right)^2\right)\left(n\sum y^2 - \left(\sum y\right)^2\right)}}.$$

Σxy Sum of the products of the individual elements of the two sets of numbers, x and y
Σx Sum of the numbers in set x
Σy Sum of the numbers in set y
(Σx)2 Square of the sum of the numbers in set x
(Σy)2 Square of the sum of the numbers in set y
Σx2 Sum of the squares of the numbers in set x
Σy2 Sum of the squares of the numbers in set y

3- Skewness and Kurtosis:



Skewness

Kurtosis

Skewness tells us if a dataset is symmetric or if it leans to one side.

Importance: Helps understand if most values are low or high, and detect bias in data.

Example:

تزايد ثابت

Dataset 1: [2, 3, 4, 5, 6] → symmetric → skewness ≈ 0

Dataset 2: [1, 2, 2, 3, 10] → right-skewed → skewness > 0

صار كو تزايد قوي و مفاجئ

Interpretation: Most numbers are small, but one very high number (10) pulls the tail to the right.

Kurtosis shows how heavy or light the tails of a distribution are.

Importance: Helps detect extreme values (outliers) that can affect analysis.

Example:

Dataset 1: [3, 3, 4, 4, 5] → low kurtosis (flat, few extremes)
Dataset 2: [1, 3, 4, 5, 10] → high kurtosis (peaked, extreme values present)

Interpretation: Dataset 2 has more extreme values (1 and 10) than Dataset 1.

Python (scipy.stats.skew, scipy.stats.kurtosis) to calculate these values quickly.

Problem:

A teacher recorded the scores of 7 students in a quiz:

Scores: 5, 6, 7, 8, 8, 9, 15

We want to analyze the **shape of the distribution** using Skewness and Kurtosis.

Step 1: Calculate basic statistics:

Mean (average) = (5+6+7+8+8+9+15)/7 = 8

Median = middle value = 8

Standard deviation (SD) ≈ 3.13

Step 2: Skewness:

Skewness formula

$$\boxed{\text{Skewness} = \frac{\text{Mean} - \text{Median}}{\text{SD}}}$$

Skewness : (8 – 8)/3.13 ≈ 0 → approximately symmetric but notice the high value 15 pulls the tail slightly right.

Interpretation:

Skewness ≈ 0 → fairly symmetric distribution

High value 15 → small right skew

Step 3: Kurtosis

Kurtosis measures how heavy the tails are.

High kurtosis → more extreme values.
most scores are 5–9, but 15 is an extreme value → heavy right tail.

So kurtosis is positive → indicates the presence of outlier.

Step 4: Conclusion / Result

Skewness: Slightly positive → tail on the right
Kurtosis: Positive → one extreme high score (outlier)

Skewness tells which side the data is stretched.
Kurtosis tells whether there are extreme values (outliers)

Descriptive Statistics

- numpy.sum() - Computes the sum of array elements.
- numpy.mean() : Computes the arithmetic mean of the array elements.
- numpy.median() : Computes the median of the array elements.
- numpy.average() : Computes the weighted average of the array elements.
- numpy.std() : Computes the standard deviation of the array elements.
- numpy.var() : Computes the variance of the array elements.
- numpy.ptp() : Computes the range (max - min) of values in the array.
- numpy.min() : Finds the minimum value in the array.
- numpy.max() : Finds the maximum value in the array.
- numpy.percentile() : Computes the nth percentile of the array elements.
- numpy.quantile() : Computes the quantile of the array elements.

Frequency and Histogram

- numpy.histogram() : Computes the histogram of the data.
- numpy.bincount() : Counts the number of occurrences of each value in an array of non-negative integers.

Correlation and Covariance

- numpy.corrcoef() : Computes the Pearson correlation coefficient matrix.
- numpy.cov() : Computes the covariance matrix.

Statistical Testing

- numpy.random.normal() : Generates random samples from a normal distribution (useful for statistical simulations).
- numpy.random.seed() : Sets the random seed for reproducibility.

Random Number Generation

- numpy.random.seed() : Set random seed.
- numpy.random.rand() : Random numbers in [0, 1).
- numpy.random.randn() : Samples from a standard normal distribution.
- numpy.random.randint() : Random integers in a range.
- numpy.random.choice() : Random selection from an array.
- numpy.random.shuffle() : Shuffle elements of an array.
- numpy.random.normal() : Samples from a normal distribution.
- numpy.random.uniform() : Samples from a uniform distribution.

Key Parameters and Units

Parameter	**Definition**	**Unit**	**Purpose/Reason**
Population Mean (μ)	Average of all population values	Same as data (e.g., points, cm, $)	Represents true average of the population
Sample Mean ($\bar{x}$)	Average of sample values	Same as data	Used to estimate μ
Population Variance (σ^2)	Measure of population spread	Square of data unit (e.g., points^2)	True variability in population
Sample Variance (s^2)	Measure of sample spread	Square of data unit	Estimates σ^2 when population data unknown
Standard Deviation (σ / s)	Square root of variance	Same as data	Easier interpretation of spread
Confidence Interval (CI)	Range where population parameter likely lies	Same as data	Quantifies uncertainty in estimation
Significance Level (α)	Probability of Type I error	None (0–1)	Controls risk of rejecting true hypothesis
p-value	Probability of observing data if H_0 is true	None (0–1)	Helps decide whether to reject H_0

Types of Inferential Statistics

Estimation

Point Estimation: Single value estimate (e.g., sample mean = 72 points).

Interval Estimation (Confidence Intervals): Range estimate (e.g., μ = 72 ± 5 points, 95% CI).

Reason to Use: Gives measure of certainty about parameter.

Hypothesis Testing

Null Hypothesis (H_0): Statement of no effect (e.g., mean score = 70).

Alternative Hypothesis (H_1): Statement of effect (e.g., mean score ≠ 70).

t-test / z-test: Compare sample mean with population mean.

Chi-square test: Test association between categorical variables.

Reason to Use: Decide if observed effect is statistically significant.

Regression and Correlation

Correlation: Measures strength of relationship (unitless, -1 to 1).

Simple Linear Regression: Predict Y from X ($Y = a + bX$).

Reason to Use: Model relationships and make predictions.

ANOVA (Analysis of Variance)

Compare means of multiple groups.

F-statistic: Ratio of between-group to within-group variance.

Reason to Use: Detect if group differences are significant.

1) Point Estimation

Definition: Point estimation refers to the process of using data from a sample to estimate a population parameter, such as the **mean** or **variance**.

```
import numpy as np

# Sample data: Heights of 10 students in cm
heights = [150, 160, 155, 165, 170, 175, 160, 158, 162, 168]

# Calculate sample mean
sample_mean = np.mean(heights)

print(f"Estimated population mean height: {sample_mean:.2f} cm")
```

```
Estimated population mean height:    161.30 cm
```

2) Confidence Intervals

Definition: Confidence Interval is a range of values that is used to estimate a population parameter. We use the sample mean and standard deviation to calculate a range in which the true population parameter is likely to lie.

Confidence Level: The probability that the confidence interval contains the true parameter. Common confidence levels are **90%**, **95%**, and **99%**.

$$\mathrm{CI} = \hat{\mu} \pm Z \times \frac{\sigma}{\sqrt{n}}$$

$\hat{\mu}$ = sample mean

Z = Z-score corresponding to the confidence level

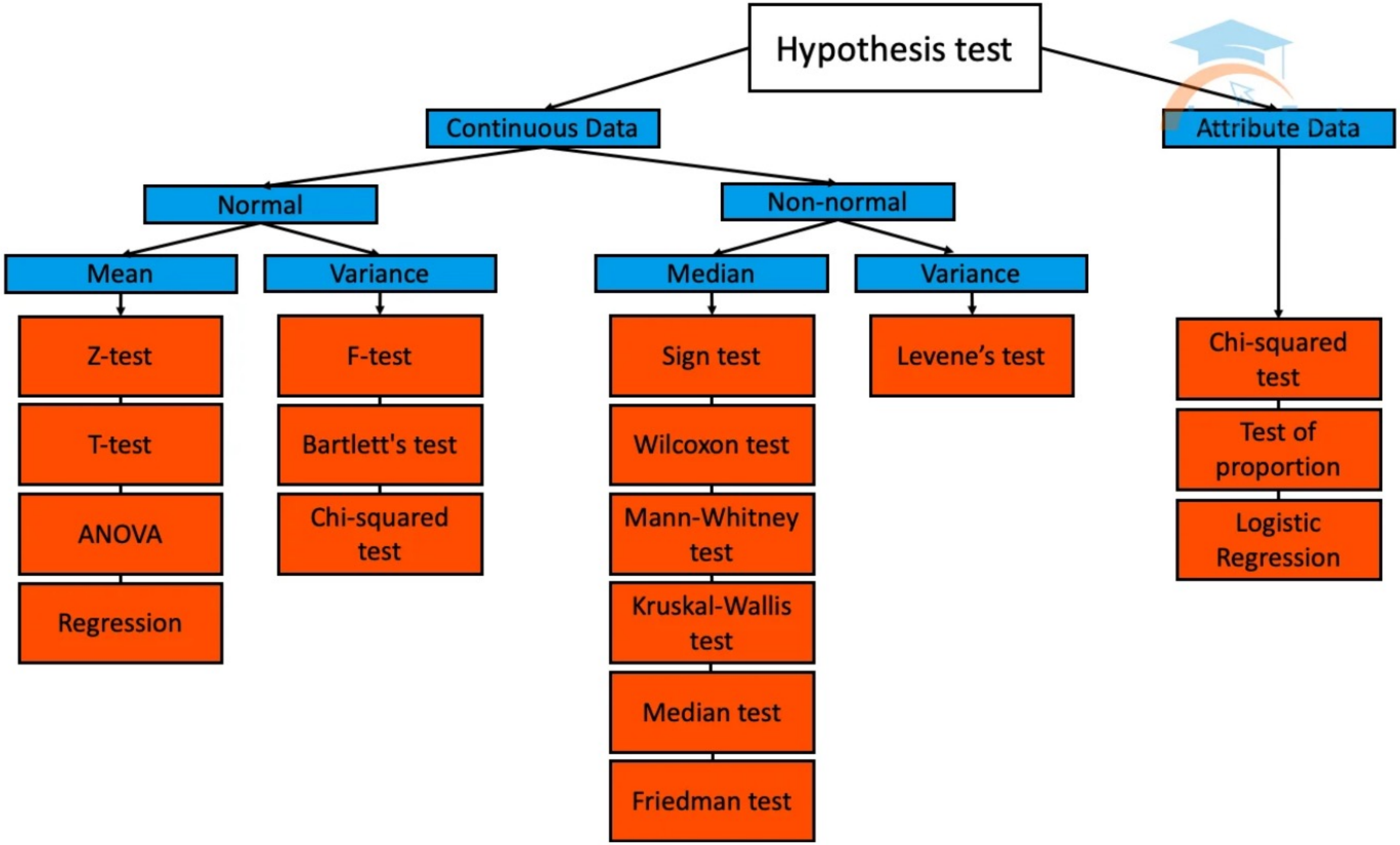
σ = sample standard deviation

n = sample size

```
z_score = stats.norm.ppf(1 - (1 -
confidence_level) / 2)
```

3) Hypothesis Testing

Definition: Hypothesis testing allows us to test an assumption or claim about a population based on sample data. We test the **null hypothesis** (H_0) against the **alternative hypothesis** (H_1).



Hypothesis Testing

Definition: Hypothesis testing allows us to test an assumption or claim about a population based on sample data. We test the **null hypothesis** (H_0) against the **alternative hypothesis** (H_1).

Hypothesis Testing:

1. State the Hypotheses:

1. **Null Hypothesis (H_0)**: The population parameter is equal to a specific value.
2. **Alternative Hypothesis (H_1)**: The population parameter is not equal to that value (or different in a specific way).

2. Set the Significance Level (α):

1. The significance level, usually set at 0.05 (5%), is the threshold for deciding whether the p-value is small enough to reject the null hypothesis.
2. If **p-value < α**, we reject H_0.
3. If **p-value ≥ α**, we fail to reject H_0.

pes

3. Collect the Sample Data:

1. You gather the data from a sample, which is a smaller group taken from the entire population.

4. Calculate the Test Statistic:

1. Based on the sample data, you calculate a test statistic (like a **t-statistic** or **z-score**) that measures how much the sample data deviates from the null hypothesis.

5. Determine the p-value:

1. Use statistical methods or software (like Python) to calculate the p-value.

6. Make a Decision:

1. If the p-value is **less than the significance level (α)**, we reject the null hypothesis and conclude that there is enough evidence to support the alternative hypothesis.
2. If the p-value is **greater than or equal to α**, we fail to reject the null hypothesis, meaning there isn't enough evidence to support the alternative hypothesis.

A teacher claims that the average test score in her class is **70** ($H_0 : \mu = 70$).

You collect scores from **5 students**: [65, 70, 75, 80, 85].

You want to test if the data supports the claim ($H_1 : \mu \neq 70$).

1. State the Hypotheses

- Null Hypothesis (H_0): The average score is 70 ($\mu = 70$).
- Alternative Hypothesis (H_1): The average score is not 70 ($\mu \neq 70$).

2. Collect Data

- Sample data: [65, 70, 75, 80, 85].
- Sample mean ($\bar{x}$) = $\frac{65+70+75+80+85}{5} = 75$.
- Sample size (n) = 5.
- Sample standard deviation (s) = 7.91 (calculated as the spread of the data).

3. Calculate the t-Statistic

The t-statistic formula is:

$$\boxed{t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}}$$

Substitute the values:

- $\bar{x} = 75$, $\mu = 70$, $s = 7.91$, $n = 5$,

$$t = \frac{75 - 70}{\frac{7.91}{\sqrt{5}}} = \frac{5}{3.54} \approx 1.41$$

4. Determine the p-Value

- Degrees of freedom ($df = n - 1 = 4$).
- Use a t-distribution table or Python (`scipy.stats`) to find the p-value.

Key Functions

- `stats.ttest_1samp()` : Perform a one-sample t-test.
- `stats.ttest_ind()` : Conduct an independent two-sample t-test.
- `stats.ttest_rel()` : Perform a paired t-test.
- `stats.norm.ppf()` : Get the Z-score for a given confidence level.
- `stats.chi2_contingency()` : Perform a Chi-square test.

2) Example: All employees of the company are listed in alphabetical order. From the first 10 numbers, you randomly select a starting point (**Systematic Sampling**): number 6. From number 6 onwards, every 10th person on the list is selected (6, 16, 26, 36, and so on), and you end up with a sample of 100 people.

```
population = 1000

step = 10

sample = [element for element in range(6, population, step)]

print (sample)
```

$$\text{Count} = \frac{\text{population} - \text{start}}{\text{step}} = \frac{1000 - 6}{10} = 99$$

creates a list of numbers:

- Starting at 6 (start=6).
- Ending before population (stop=1000).
- Incrementing by step=10.

End of Class 4