Contents

[STATISTICS 2](#_Toc210301766)

[DESCRIPTIVE STATISTICS 2](#_Toc210301767)

[INFERENTIAL STATISTICS 2](#_Toc210301768)

[MEASURES OF DESCRIPTIVE STATISTICS 2](#_Toc210301769)

[MEASURES OF CENTRAL TENDENCY - FINDING THE “MIDDLE” OF DATA 2](#_Toc210301770)

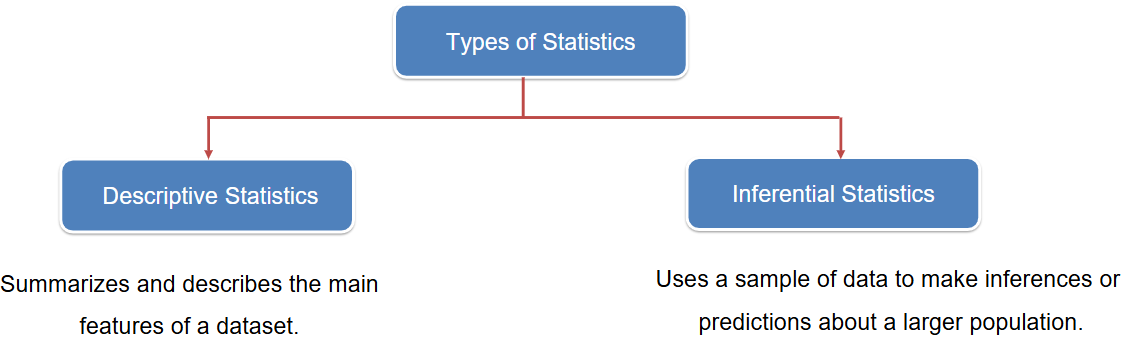
[MEASURES OF DISPERSION (VARIABILITY) - HOW DATA IS SPREAD OUT 3](#_Toc210301771)

[MEASURES OF DISTRIBUTION SHAPES 6](#_Toc210301772)

[TYPES OF DATA IN STATISTICS 6](#_Toc210301773)

# STATISTICS

* The science of collecting, organizing, analyzing and interpreting data to make decisions
* It goal is to make sense of information and draw meaningful conclusions.



# DESCRIPTIVE STATISTICS

* Describes the key feature of dataset.
* In descriptive statistics
  + We find the central value (called central tendency) of the dataset. The different measures of central measures central tendencies Mean, Median & Mode
  + **Or:** How dispersed is the dataset's data?
* **Examples**:
* Calculating the **average** score of students in a class.
* Creating a **bar chart** to show the number of employees in different departments.
* Finding the **median** income in a city.
* Summarizing data using **mean, mode, range, standard deviation**, etc.

# INFERENTIAL STATISTICS

**Definition**: Inferential statistics use data from a sample to make generalizations or predictions about a larger population.

**Purpose**: To draw conclusions beyond the immediate data.

**Examples**:

A white background with black text

AI-generated content may be incorrect.

# MEASURES OF DESCRIPTIVE STATISTICS

**Measures of Descriptive Statistics** are tools used to summarize and describe the main features of a dataset. They help in understanding the **distribution, central tendency, and variability of data**.

## MEASURES OF CENTRAL TENDENCY - FINDING THE “MIDDLE” OF DATA

These describe the center or average of a dataset.

## MEAN(AVERAGE)

|  |  |
| --- | --- |
| Sum of all values divided by the number of values.  *Example*: Mean of [2, 4, 6] is (2+4+6)/3 = 4 | import numpy as np  nums = [1, 2, 3, 4, 5]  avg = np.mean(nums)  print(f"Average of nums: {avg}") |

## MEDIAN

A math equations and formulas

AI-generated content may be incorrect.

|  |  |
| --- | --- |
| calculate the **median** of the dataset: **7, 3, 9, 5, 4**  Follow these steps:   * **Step 1: Arrange the data in ascending order**   Sorted list: **3, 4, 5, 7, 9**   * **Step 2: Identify the number of values**   There are **5 values**, which is an **odd number**.   * **Step 3: Find the middle value**   Since the number of values is odd, the **median** is the middle value.    The **3rd value** in the sorted list is **5 hence the Median = 5** | If the number of values (**n**) is **even**, the **median** is calculated slightly differently than when **n is odd**.  **Steps to Calculate Median When n is Even**  Let’s take an example dataset:**7, 3, 9, 5, 4, 6**   * **Step 1: Arrange the data in ascending order**   Sorted list: **3, 4, 5, 6, 7, 9**   * **Step 2: Identify the middle two values**   There are **6 values**, so the middle two are:   * + **3rd value** = 5   + **4th value** = 6 * **Step 3: Calculate the average of the middle two**     **Median = 5.5** |

## MODE

|  |  |
| --- | --- |
| * **Mode**: The value that appears most frequently. * *Example*: Mode of [2, 2, 3, 4] is 2 * There may be no mode, one mode or multiple mode * Numpy lacks a mode function, so use the **scipy library** instead. | import numpy as np  from scipy import stats  nums = [4, 4, 3, 5, 6, 7, 8, 9, 4, 3, 2, 1, 4, 5, 6, 7, 8, 9, 4, 3, 2]  data = np.array(nums)  mode\_result = stats.mode(data, keepdims=True)  print("Mode:", mode\_result.mode[0])  print("Frequency:", mode\_result.count[0])  Mode: 4  Frequency: 5 |

## WHEN TO USE MEAN OR MEDIAN?

We use “mean”- when the data is symmetrical and ”mode ”- when the data is skewed

|  |  |
| --- | --- |
| **Product Prices** | * Mean: 39 * Mode : 14   In this example – the data is skewed due a outliers in the data- hence the mean is relatively higher . In this use case is mode is right measure to calculate the product price(measure of central tendency) |
| 21 |
| 4 |
| 9 |
| 14 |
| 29 |
| 2 |
| 194 |

## SKEWNESS

# MEASURES OF DISPERSION (VARIABILITY) - HOW DATA IS SPREAD OUT

**These describe how spread out the data i.e. describes the spread of variability of a dataset – i.e. how far the data points are from the center (mean or median)**

**A diagram of a flowchart

AI-generated content may be incorrect.**

* **Variance**: Average of squared differences from the mean.
* **Standard Deviation**: Square root of the variance; shows how much data deviates from the mean.
  + *Example*: A low standard deviation means data is close to the mean.

## RANGE

|  |  |
| --- | --- |
| * **Range**: Difference between the highest and lowest values. * *Example*: Range of [3, 7, 10] is 10 - 3 = 7 | import numpy as np  nums = [4, 4, 3, 5, 6, 7, 8, 2]  max = np.max(nums)  min = np.min(nums)  range = max - min  print("Range is: ", range) |

## IQR- INTERQUARTILE RANGE

**IQR**, or **Interquartile Range**, is a measure of statistical dispersion — it tells us how spread out the middle 50% of your data is. It’s especially useful for identifying outliers and understanding the variability in a dataset.

**🔍 What is IQR?**

The IQR is the difference between the **third quartile (Q3)** and the **first quartile (Q1)**:IQR = Q3 - Q1 $

* **Q1 (25th percentile)**: The value below which 25% of the data falls.
* **Q3 (75th percentile)**: The value below which 75% of the data falls.

How to Calculate IQR (Step-by-Step)

**Example: When number of elements is odd**

**Dataset:**[6, 7, 15, 36, 39, 41, 42, 43, 47]

This dataset has **9 elements**, which is an **odd number**.

**Step 1: Sort the Data: (**Already sorted)- [6, 7, 15, 36, 39, 41, 42, 43, 47]

**Step 2: Find the Median (Q2):** Middle value (5th element) = **39 .**So, **Q2 = 39**

**Step 3: Split the Data (Exclude the Median)**

* **Lower half** (before median): [6, 7, 15, 36]
* **Upper half** (after median): [41, 42, 43, 47]

**Step 4: Find Q1 (Median of Lower Half)**

* Lower half: [6, 7, 15, 36]
* Median of lower half = (7 + 15)/2 = 11.0
* So, **Q1 = 11.0**

**Step 5: Find Q3 (Median of Upper Half)**

* Upper half: [41, 42, 43, 47]
* Median of upper half = (42 + 43)/2 = 42.5
* So, **Q3 = 42.5**

**Step 6: Calculate IQR**

* IQR = Q3 - Q1 = 42.5 - 11.0 = 31.5

**✅ Final Results:**

* **Q1** = 11.0
* **Q2 (Median)** = 39
* **Q3** = 42.5
* **IQR** = **31.5**

**Example: When number of elements is even**

* Dataset : [6, 7, 15, 36, 39, 41, 42, 43]
* Sort the data (already sorted here).
* **Find the median (Q2):**
  + There are 8 numbers → median is the average of the 4th and 5th values:
  + Median = (36+39)/2=37.5(36 + 39)/2 = 37.5(36+39)/2=37.5
* **Split the data into two halves:**
  + Lower half: [6, 7, 15, 36]
  + Upper half: [39, 41, 42, 43]
* **Find Q1 (median of lower half):** Q1 = (7+15)/2=11(7 + 15)/2 = 11(7+15)/2=11
* **Find Q3 (median of upper half):** Q3 = (41+42)/2=41.5(41 + 42)/2 = 41.5(41+42)/2=41.5
* **Calculate IQR**: IQR = Q3 - Q1 = 41.5−11=30.541.5 - 11 = 30.541.5−11=**30.5**

## SAMPLE AND POPULATION

**Population**

* The **entire group** of individuals or items you're interested in studying.
* Examples:
  + All people in a country.
  + All leaves on a tree.
* Has **parameters**:
  + **Mean (μ)**: Average of the population.
  + **Size (N)**: Total number of items.
  + **Variance (σ²)**: Spread of data in the population.

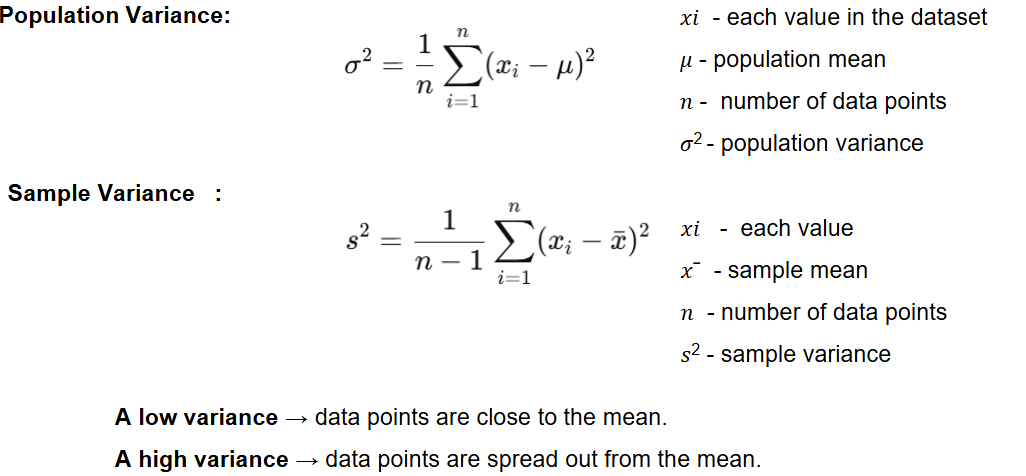
**Sample**

* A **subset** of the population, selected for analysis.
* Used when studying the whole population is impractical.
* Examples:
  + 1,000 people from a country.
  + 100 leaves from a tree.
* Has **statistics**:
  + **Mean (x̄)**: Average of the sample.
  + **Size (n)**: Number of items in the sample.
  + **Variance (s²)**: Spread of data in the sample.

## VARIANCE

Variance measures **how far data points are spread out from the mean**.

* It’s the **average of squared deviations** from the mean.
* Always **non-negative** (because of squaring).
* The **square root of variance = Standard Deviation**.



NOTE : The formula of population and sample variance is slightly different

**Example**:

* Class A test scores = [50, 51, 52, 49, 50] → low variance (students perform similarly).
* Class B test scores = [10, 30, 70, 90, 100] → high variance (big differences in performance).

## CALCULATION- SAMPLE VARIANCE

**Step-by-step calculation of the variance for the dataset:X = [10, 12, 23, 23, 16]**

**Step 1: Calculate the Mean**: Mean= (10 + 12 + 23 + 23 + 16)/5 = 84/5 = 16.8

**Step 2: Calculate the Squared Differences from the Mean**

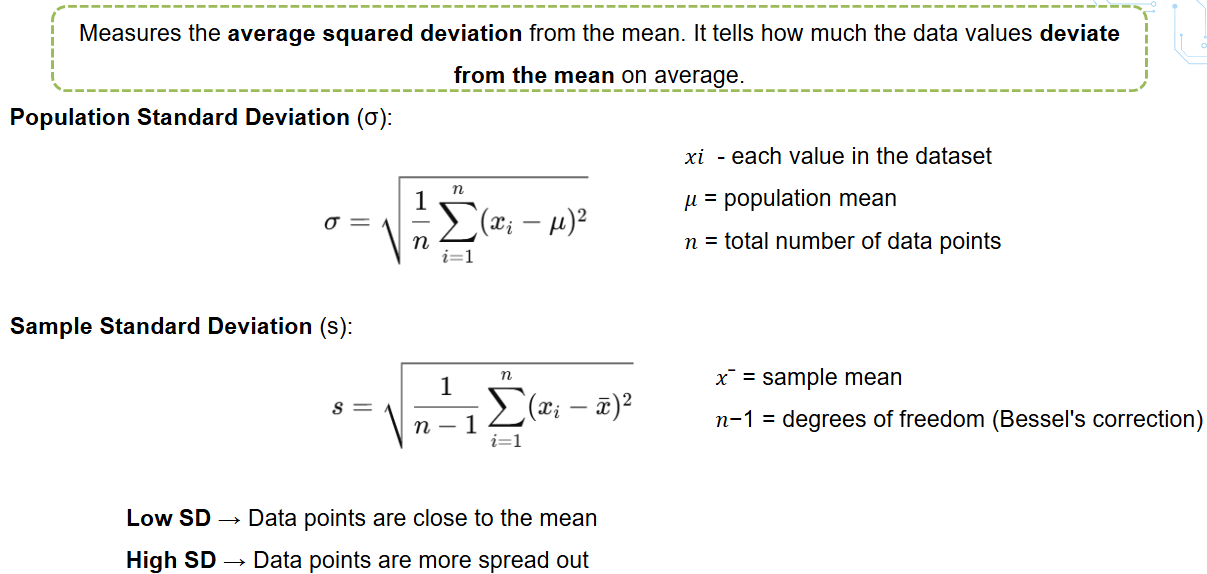
**A math equation with numbers

AI-generated content may be incorrect.**

**Step 3: Calculate the Sample Variance**

****

## STANDARD DEVIATION



Here’s a **step-by-step explanation** of how to calculate the **standard deviation** for the dataset: **X = [10, 12, 23, 23, 16]**

**Step 1: Calculate the Mean :** Mean = (10 + 12 + 23 + 23 + 16)/5 = 84/5 = 16.8

**Step 2: Find the Squared Differences from the Mean**

A screenshot of a computer

AI-generated content may be incorrect.

**Step 3: Calculate the Sample Variance**

A black text with a white background

AI-generated content may be incorrect.

**Step 4: Calculate the Standard Deviation**



## CALCULATION

|  |  |
| --- | --- |
| import numpy as np  nums = [10, 12, 23, 23, 16]  np\_arr = np.array(nums)  variance = np.var(np\_arr, ddof=1)  sd = np.std(np\_arr, ddof=1)  print("Variance:", variance)  print("Standard Deviation:", sd) | **Variance: 36.7**  **Standard Deviation: 6.058052492344384**   * ddof= 1: For Sample variance |

# MEASURES OF DISTRIBUTION SHAPES

# TYPES OF DATA IN STATISTICS

A diagram of data flow

AI-generated content may be incorrect.

Categorical (Qualitative) Data

These represent characteristics or attributes that describe categories.

* **Nominal Data**:
  + Categories with no inherent order.
  + **Examples**: Gender (Male, Female), Hair Color (Black, Brown, Blonde), Nationality (Indian, American, Japanese)
* **Ordinal Data**
  + Categories with a meaningful order, but the intervals between them are not uniform.
  + **Examples**: Education Level (High School, Bachelor's, Master's), Customer Satisfaction (Satisfied, Neutral, Dissatisfied)

Numerical (Quantitative) Data

These represent measurable quantities and can be expressed in numbers.

* **Discrete Data**
  + Countable values, often whole numbers.
  + **Examples**: Number of students in a class, Number of cars in a parking lot
* **Continuous Data**
  + Measurable values that can take any value within a range.
  + **Examples**: Height (in cm), Weight (in kg), Temperature (in °C)

# DISTRIBUTION

## NORMAL DISTRIBUTION

# PROBABLITY THEORY