Contents

[KEY DEFINATIONS 2](#_Toc209964501)

[POPULATION 2](#_Toc209964502)

[SAMPLE 2](#_Toc209964503)

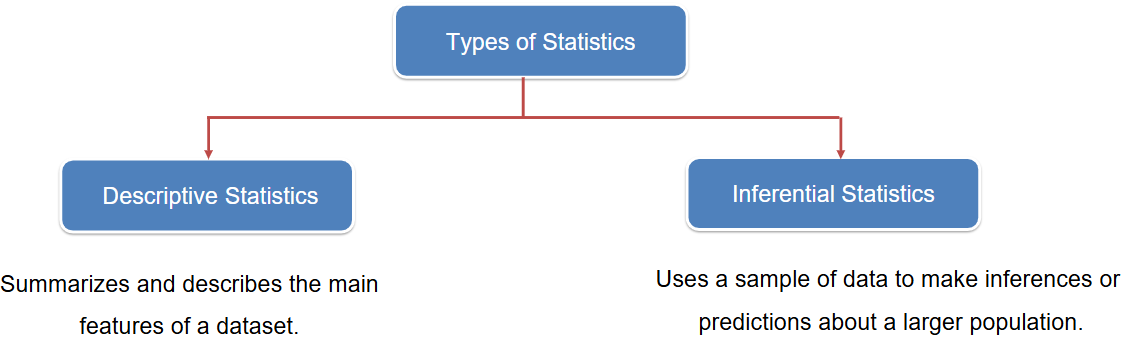
[PARAMETERS 2](#_Toc209964504)

[DISTRIBUTION 3](#_Toc209964505)

[UNIFORM DISTRIBUTION 3](#_Toc209964506)

# WHAT IS 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

**Definition**: Descriptive statistics involve methods for organizing, summarizing, and presenting data in an informative way.

**Purpose**: To describe the main features of a dataset.

**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 |

## 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.

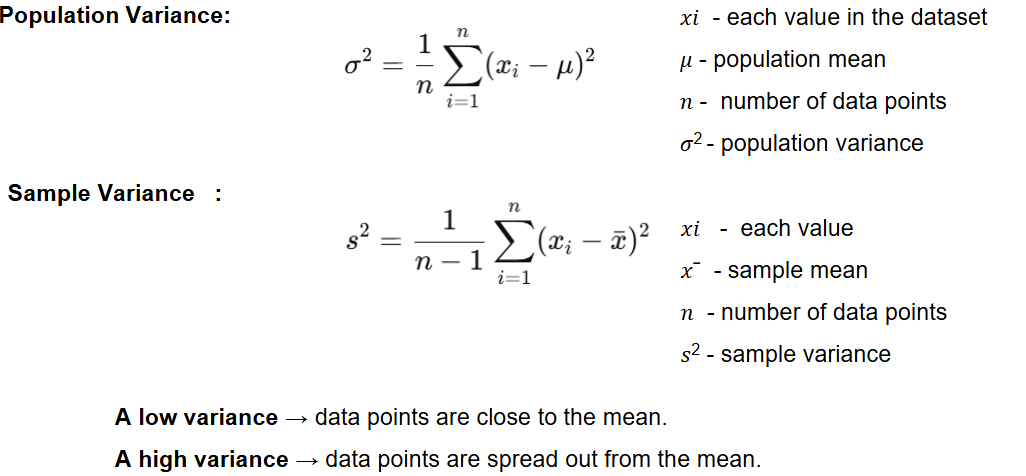
MEAN

|  |  |
| --- | --- |
| * **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) |

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**.



**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

**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**

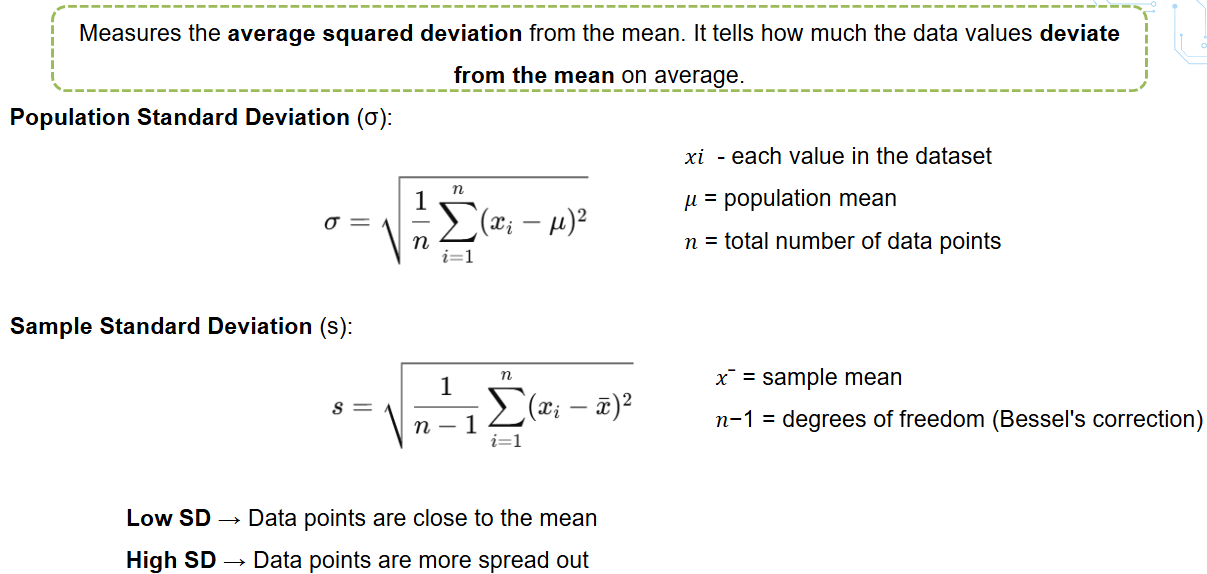
**A math equation with numbers

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)