# Statistics lecture 1

*This lecture is created and assembled by "Sulaiman Ahmed". Do not share without proper credit.*
*Email: sulaimanahmed013@gmail.com*
*Website: sulaimanahmed013.wixsite.com/my-site*
*LinkedIn: linkedin.com/in/sulaimanahmed*

## 1. Statistics and Its Applications

Statistics is a field that deals with:

- Collection of data
- Organization of data
- Analysis of data
- Interpretation of data
- Presentation of data

The ultimate goal of statistics is to enable effective decision-making based on data analysis.

### Example: Online Shopping Age Data

Consider an age feature for online shopping:

```
24, 27, 14, 13, 28, 29, 31, 32
```

With this data, we can calculate:

- **Mean age**: The average age of the customers.
- **Median age**: The middle value when the ages are ordered.
- **Distribution of age**: Understanding how ages are spread out across the dataset.

We can also create visualizations such as:

- **Histograms**: To see the frequency distribution of ages.
- **Probability Density Function (PDF)**: To understand the likelihood of different age ranges.
- **Cumulative Density Function (CDF)**: To understand the cumulative probability up to a certain age.

### Applications of Statistics

Statistics is widely used in various fields. Some key applications include:

1. **Machine Learning and Data Science**: Used for model building, validation, and prediction.

    - **Example**: Predicting housing prices based on historical data.
2. **Data Analysis**: Helps in extracting insights from data.
    - **Example**: Analyzing customer feedback to improve products.
3. **Business Intelligence and Analytics**: Assists in making informed business decisions.
    - **Example**: Determining the best marketing strategy based on sales data.
4. **Risk Analysis**: Used in finance and insurance to assess risks.
    - **Example**: Calculating the risk of loan defaults.
5. **Everyday life decisions**: Helps in making informed personal decisions.
    - **Example**: Analyzing budget and expenses to manage finances.
6. **Medical research (e.g., vaccine trials)**: Used to validate the effectiveness and safety of treatments.
    - **Example**: Determining the efficacy of a new drug through clinical trials.

# 2. Types of Statistics

There are two main types of statistics:

## 2.1 Descriptive Statistics

Descriptive statistics involves organizing and summarizing data. It provides simple summaries and visualizations of the data.

Techniques include:

1. **Measure of Central Tendency**
    - **Mean**: The average value.
    - **Median**: The middle value when data is sorted.
    - **Mode**: The most frequently occurring value.
2. **Measure of Dispersion**
    - **Variance**: Measures how far data points are from the mean.
    - **Standard Deviation**: The square root of the variance, representing the average distance from the mean.

## Examples of Descriptive Statistics

- Suppose we have a dataset of exam scores: `85, 88, 92, 91, 87, 90, 89`.
    - **Mean**: (85 + 88 + 92 + 91 + 87 + 90 + 89) / 7 = 88.86
    - **Median**: The middle value is 89.
    - **Mode**: There is no mode as no value repeats.
    - **Variance** and **Standard Deviation**: These would be calculated to understand the spread of scores.

## Real-life Usage

- **Sports**: Analyzing player performance data to improve strategies.
- **Education**: Summarizing student test scores to evaluate teaching effectiveness.
- **Healthcare**: Summarizing patient data to track disease outbreaks.

## 2.2 Inferential Statistics

Inferential statistics involves making conclusions or inferences about a population based on a sample of data. It allows us to make predictions and generalizations.

Techniques include:

- **Z-test**: Used to determine if there is a significant difference between sample and population means.
- **T-test**: Used to compare the means of two groups.
- **Chi-square test**: Used to examine the association between categorical variables.

## Examples of Inferential Statistics

- **Medical Trials**: Testing the effectiveness of a new drug by experimenting on a sample group and inferring the results to the broader population.
- **Market Research**: Using a sample survey to infer the preferences of the entire market.
- **Manufacturing**: Quality control using sample inspections to infer the quality of the entire production.

## Real-life Usage

- **Politics**: Predicting election outcomes based on exit polls.
- **Public Health**: Estimating the spread of diseases using sample data.
- **Economics**: Making economic forecasts based on sample data from surveys.

# 3. Population vs Sample Data

## Population

- Represents the entire group being studied.
- Denoted by capital N.
- Example: All 100,000 people on an island.

## Sample

- A subset of the population.
- Denoted by lowercase n.
- Example: 10,000 people selected from the island population.

## Importance of Sampling

Sampling is used when it's impractical or impossible to study the entire population. It helps in:

- Reducing costs and time.
- Making studies feasible.
- Providing results that can be generalized to the population if the sample is representative.

## Example

- **Exit Polls**: During elections, pollsters use samples to predict the outcome of the entire election.

## Real-life Usage

- **Market Research**: Conducting surveys with a sample to understand consumer preferences.
- **Healthcare**: Clinical trials conducted on a sample of patients to infer the effects on the entire population.
- **Environmental Studies**: Sampling water from different locations to assess overall pollution levels.

# 4. Measure of Central Tendency

## 4.1 Mean (Average)

The mean is the sum of all data points divided by the number of points. It gives an overall average.

For a population:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

μ = N∑i=1N xi

For a sample:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

x¯ = n∑i=1n xi

Where:

- $\mu$μ is the population mean.
- $\bar{x}$x¯ is the sample mean.
- $x_i$xi are individual data points.
- $N$N is the population size.
- $n$n is the sample size.

## Example

Consider the dataset: `5, 10, 15, 20, 25`.

- Population Mean: $\mu = \frac{5 + 10 + 15 + 20 + 25}{5} = 15$μ = 55+10+15+20+25 = 15
- Sample Mean (if we consider the sample `5, 10, 15`): $\bar{x} = \frac{5 + 10 + 15}{3} = 10$x¯ = 35+10+15 = 10

## Real-life Usage

- **Business**: Calculating the average sales per month to inform inventory decisions.
- **Education**: Determining the average score of students to assess overall performance.
- **Healthcare**: Finding the average heart rate in a study to draw health conclusions.

## 4.2 Median

The median is the middle value when the data is arranged in order. If the number of observations is even, it is the average of the two middle numbers.

## Example

For the dataset `4, 8, 15, 16, 23`:

- Median: 15 (middle value)

For the dataset `4, 8, 15, 16, 23, 42`:

- Median: (15 + 16) / 2 = 15.5

## Real-life Usage

- **Income Data**: Median income is often used instead of mean income to avoid skewing by extremely high values.
- **Real Estate**: Median home prices are used to understand the market without the influence of extreme values.
- **Healthcare**: Median survival times in clinical trials to provide a clearer picture of typical outcomes.

## 4.3 Mode

The mode is the value that appears most frequently in the dataset. There can be more than one mode if multiple values have the same highest frequency.

## Example

For the dataset `1, 2, 2, 3, 4`:

- Mode: 2

For the dataset `1, 1, 2, 2, 3`:

- Mode: 1 and 2 (bimodal)

## Real-life Usage

- **Retail**: Determining the most sold product in a store.
- **Education**: Identifying the most common grade received by students.
- **Healthcare**: Finding the most common symptom in a patient group.

# 5. Measure of Dispersion

## 5.1 Variance

Variance measures the spread of data points around the mean. It is the average of the squared differences from the mean.

For a population:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

σ2 = N∑i=1N (xi − μ)2

For a sample:

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

s2 = n − 1∑i=1n(xi − x¯)2

Where:

- $\sigma^2$σ2 is the population variance.
- $s^2$s2 is the sample variance.

## Example

Consider the dataset `2, 4, 4, 4, 5, 5, 7, 9`:

- Mean ($\bar{x}$x¯) = 5
- Population Variance ($\sigma^2$σ2) = $\frac{(2-5)^2 + (4-5)^2 + (4-5)^2 + (4-5)^2 + (5-5)^2 + (5-5)^2 + (7-5)^2 + (9-5)^2}{8}$ = 48(2−5)2+(4−5)2+(4−5)2+(4−5)2+(5−5)2+(5−5)2+(7−5)2+(9−5)2 = 4
- Sample Variance ($s^2$s2) = $\frac{(2-5)^2 + (4-5)^2 + (4-5)^2 + (4-5)^2 + (5-5)^2 + (5-5)^2 + (7-5)^2 + (9-5)^2}{7}$ = 4.577(2−5)2+(4−5)2+(4−5)2+(4−5)2+(5−5)2+(5−5)2+(7−5)2+(9−5)2 = 4.57

## Real-life Usage

- **Finance**: Calculating variance in investment returns to assess risk.
- **Manufacturing**: Measuring variance in product weights to maintain quality control.
- **Healthcare**: Analyzing variance in patient response times to treatments.

## 5.2 Standard Deviation

Standard deviation is the square root of the variance and provides a measure of the average distance between each data point and the mean.

For a population:

$$\sigma = \sqrt{\sigma^2}$$

σ =  σ2

√

For a sample:

$$s = \sqrt{s^2}$$

s = s2
√

## Example

Using the previous dataset `2, 4, 4, 4, 5, 5, 7, 9`:

- Population Standard Deviation ($\sigma$σ) = $\sqrt{4} = 2$ 4
  √
  = 2
- Sample Standard Deviation ($s$s) = $\sqrt{4.57} \approx 2.14$ 4.57
  √
  ≈ 2.14

## Real-life Usage

- **Finance**: Assessing the volatility of stock prices.
- **Education**: Understanding the spread of test scores among students.
- **Healthcare**: Evaluating the consistency of medical test results.


*Email: sulaimanahmed013@gmail.com*
*Website: sulaimanahmed013.wixsite.com/my-site*
*LinkedIn: linkedin.com/in/sulaimanahmed*