#UpSkillWithKalpesh

Day 05

Data Science Unlocked

From Zero to Data Hero

Statistics Basics for Data Science

Kalpesh Pathade
@DataSimplified

Statistics(1/3) - Basics

I. Introduction to Statistics

What is Statistics?

Statistics is the branch of mathematics that deals with collecting, analyzing, interpreting, presenting, and organizing data. It provides methods for drawing meaningful conclusions from data and allows researchers and data scientists to make informed decisions based on data.

There are two main types of statistics:

- **Descriptive Statistics**: Summarizes and describes the characteristics of a dataset.
- **Inferential Statistics**: Makes predictions or inferences about a population based on a sample.

Key Concepts in Statistics:

- **Population**: The entire set of items or individuals of interest.
- **Sample**: A subset of the population used to estimate the characteristics of the population.
- **Data**: Raw facts and figures collected for analysis. Data can be qualitative or quantitative.
- **Variables**: Characteristics or properties that can be measured or categorized (e.g., height, weight, age, gender).
- **Parameters**: Numerical values that describe a population (e.g., population mean, population variance).
- **Statistics**: Numerical values that describe a sample (e.g., sample mean, sample variance).

Types of Statistics: Descriptive vs. Inferential

Descriptive Statistics

Descriptive statistics is the branch of statistics that involves organizing, summarizing, and presenting data in a meaningful way. It is used to describe and understand the basic features of data.

Key Tools in Descriptive Statistics:

1. **Measures of Central Tendency: Includes mean (arithmetic average), median (middle value), and mode (most frequent value)**
2. **Measures of Dispersion: Includes range (difference between largest and smallest values), variance (average squared deviation from mean), and standard deviation (square root of variance)**
3. **Data Visualization: Includes histograms (distribution representation), bar plots (categorical data comparison), and box plots (quartile distribution)**

Inferential Statistics

Inferential statistics allows data scientists to make generalizations about a population based on sample data. It helps answer questions such as: "What is the likelihood that the observed data is due to chance?" and "What are the probable outcomes in future scenarios?"

Key Tools in Inferential Statistics:

1. **Hypothesis Testing**: Involves making assumptions (hypotheses) about a population parameter and testing these assumptions using sample data. Common tests include the **t-test**, **chi-square test**, and **ANOVA**.
2. **Confidence Intervals**: A range of values, derived from the sample data, that is likely to contain the population parameter.
3. **Regression Analysis**: A method for modeling the relationship between variables. This includes **linear regression** for predicting numeric values based on independent variables.

Applications of Statistics in Data Science

Statistics plays a crucial role in data science. Data scientists use statistical methods to interpret large datasets, identify trends, and make predictions. Some of the key applications of statistics in data science include:

1. **Exploratory Data Analysis (EDA)**: Statistics help summarize and visualize datasets to uncover patterns and relationships within the data.

2. **Machine Learning**: Statistical techniques such as regression, classification, and clustering are fundamental in building predictive models. Machine learning algorithms are built on statistical principles.
3. **A/B Testing**: Inferential statistics is used to compare two or more variants to determine which one performs better in terms of user behavior, conversions, or any other metric.
4. **Sampling and Estimation**: Since collecting data from an entire population is often impractical, statistics helps select representative samples and make inferences about the population.
5. **Data Cleaning and Preprocessing**: Statistics aids in identifying anomalies, missing values, and outliers in datasets, enabling effective data preparation.
6. **Predictive Analytics**: By applying regression models and statistical tests, data scientists can forecast future trends, sales, customer behavior, and other variables.

II. Data Types and Scales of Measurement

In statistics, understanding the different types of data and the scales of measurement is crucial as they determine how data can be analyzed and which statistical methods can be applied.

1. Scales of Measurement

The scale of measurement refers to how data is categorized and quantified. There are four primary scales of measurement: **Nominal**, **Ordinal**, **Interval**, and **Ratio**. These scales determine the operations you can perform on the data (e.g., addition, subtraction, comparison).

Nominal Scale

- The nominal scale is the **simplest** level of measurement. It involves data that can be categorized into distinct groups or categories but cannot be ordered or ranked.
- **Examples**: Gender (male, female), Blood type (A, B, O, AB), Marital status (single, married, divorced)
- **Key Points**:
  - Categories are **mutually exclusive**.

- No inherent order between categories.
- Only the **mode** (most frequent category) can be calculated.

Ordinal Scale

- The ordinal scale allows for data that can be **ranked** or ordered. However, the intervals between the ranks are not necessarily equal or meaningful.
- **Examples**: Education level (high school, bachelor's, master's), Customer satisfaction (poor, fair, good, excellent), Ranks in a competition (1st, 2nd, 3rd)
- **Key Points**:
  - Data can be ordered or ranked.
  - **Median** and **mode** can be calculated.
  - Differences between ranks are **not meaningful** (e.g., the difference between "poor" and "fair" may not be the same as between "good" and "excellent").

Interval Scale

- The interval scale provides data with both **ordered** values and **equal intervals** between them. However, it does not have a true zero point, meaning you cannot make statements about absolute quantities.
- **Examples**: Temperature (Celsius, Fahrenheit), Calendar years (2020, 2021, 2022)
- **Key Points**:
  - Equal intervals between values (e.g., 10°C to 20°C is the same as 30°C to 40°C).
  - There is **no absolute zero**, so ratios are not meaningful (e.g., 20°C is not "twice" as hot as 10°C).
  - **Mean**, **median**, and **mode** can be calculated.

Ratio Scale

- The ratio scale is the highest level of measurement. It has all the properties of the interval scale, plus a **true zero point**, allowing for meaningful ratios between values.
- **Examples**: Height, Weight, Distance, Time, Income
- **Key Points**:
  - **True zero point** means the absence of the measured quantity (e.g., zero height means no height).

- Ratios are meaningful (e.g., a person weighing 80 kg is twice as heavy as a person weighing 40 kg).
- **Mean**, **median**, and **mode** can be calculated.

2. Data Types

Data can be categorized into two main types based on whether the data points are countable or measurable: **Discrete** and **Continuous** data.

Discrete Data

- Discrete data consists of **distinct, separate values** that can only take specific, individual values, usually **integers**. These are often counts or quantities.
- **Examples**: Number of students in a class, Number of cars in a parking lot, Number of goals scored in a match
- **Key Points**:
  - Can only take specific, countable values (e.g., 1, 2, 3, ...).
  - **Cannot** be subdivided into smaller parts (e.g., you can't have 2.5 students).
  - Often represented by **whole numbers**.

Continuous Data

- Continuous data can take any value within a given range and can be measured with great precision. It is often associated with physical quantities.
- **Examples**: Height, Weight, Temperature, Distance, Time
- **Key Points**:
  - Can take an infinite number of values within a range.
  - Can be **measured** to an arbitrary level of precision (e.g., 1.2 meters, 1.25 meters, 1.258 meters).
  - **Can be subdivided** into smaller intervals (e.g., time can be measured in seconds, milliseconds, microseconds).

III. Descriptive Statistics

Descriptive statistics involves summarizing and organizing data in a way that provides useful insights and makes it easier to understand. This can be achieved through **measures of central tendency**, **measures of dispersion**, and **percentiles/quartiles**. These statistics help describe the main features of a dataset.

1. Measures of Central Tendency

Measures of central tendency describe the center or typical value in a data set. These include **Mean**, **Median**, and **Mode**.

Mean

- The **mean** is the arithmetic average of a set of values. It is calculated by summing all the values in the dataset and dividing by the number of values.
- **Formula**:

$$Mean = \frac{\sum x_i}{n}$$

  Where xi represents each data point, and n is the total number of data points.

- **Example**: Given the dataset ⌫4, 5, 6, 7, 8↑, the mean is:

$$\frac{4 + 5 + 6 + 7 + 8}{5} = 6$$

- **Key Points**:
  - Sensitive to **outliers** (extremely high or low values can skew the mean).
  - Used with **interval** and **ratio** data.

Median

- The **median** is the middle value in a dataset when the values are arranged in ascending or descending order.
- If there is an odd number of values, the median is the middle value. If there is an even number, the median is the average of the two middle values.
- **Example**: For the dataset ⌫4, 5, 6, 7, 8↑, the median is **6** (the middle value).
- **Key Points**:
  - Less sensitive to outliers than the mean.

- Used with **ordinal**, **interval**, and **ratio** data.

Mode

- The **mode** is the value that appears most frequently in a dataset.
- **Example**: For the dataset ⌫4, 5, 6, 6, 7, 8↑, the mode is **6**.
- **Key Points**:
  - A dataset can have **one mode** (unimodal), **two modes** (bimodal), or **more than two modes** (multimodal).
  - It can be used with **nominal**, **ordinal**, **interval**, and **ratio** data.

---

2. Measures of Dispersion

Measures of dispersion describe the spread or variability of a dataset. These include **Range**, **Variance**, **Standard Deviation**, and **Interquartile Range**.

Range

- The **range** is the difference between the highest and lowest values in a dataset.
- **Formula**:

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

- **Example**: Given the dataset ⌫4, 5, 6, 7, 8↑, the range is:

  8–4 › 4

- **Key Points**:
  - Simple to calculate but can be heavily influenced by outliers.

Variance

- **Variance** measures how far each data point is from the mean of the dataset. It is the average of the squared differences from the mean.
- **Formula**:

$$\text{Variance} = \frac{\sum(x_i - \mu)^2}{n}$$

Where xi represents each data point, μ is the mean, and n is the number of data points.

- **Example**: Given the dataset ⌫4, 5, 6, 7, 8↑ with a mean of 6, the variance is:

$$\frac{(4-6)^2+(5-6)^2+(6-6)^2+(7-6)^2+(8-6)^2}{5} = \frac{4+1+0+1+4}{5} = 2$$

- **Key Points**:
  - Variance is in **squared units**, which can be difficult to interpret.
  - It is used with **interval** and **ratio** data.

Standard Deviation

- The **standard deviation** is the square root of the variance. It represents the average amount of deviation from the mean.
- **Formula**:

$$\text{Standard Deviation} = \sqrt{\text{Variance}}$$

- **Example**: For the dataset ⌫4, 5, 6, 7, 8↑ with a variance of 2, the standard deviation is:

$$\sqrt{2} \approx 1.41$$

- **Key Points**:
  - In the **same units** as the data, which makes it easier to interpret than variance.
  - It is commonly used in **interval** and **ratio** data.

Interquartile Range (IQR)

- The **interquartile range** is the difference between the third quartile (Q3) and the first quartile (Q1). It measures the spread of the middle 50% of the data.
- **Formula**:

$$IQR = Q3 - Q1$$

- **Key Points**:
  - Less sensitive to **outliers** than the range.
  - Used to detect **outliers** by defining boundaries outside of 1.5 * IQR from the quartiles.
  - Suitable for **ordinal**, **interval**, and **ratio** data.

---

3. Percentiles and Quartiles

- **Percentiles** are values that divide a dataset into 100 equal parts. The pth percentile is the value below which p% of the data falls.
  - **Example**: The 50th percentile (median) divides the data in half.
- **Quartiles** are special percentiles that divide the dataset into four equal parts. The **first quartile (Q1)** is the 25th percentile, the **second quartile (Q2)** is the 50th percentile (median), and the **third quartile (Q3)** is the 75th percentile.
  - **Q1**: 25th percentile, the point where 25% of the data is below.
  - **Q2**: 50th percentile, the median.
  - **Q3**: 75th percentile, the point where 75% of the data is below.

Key Points:

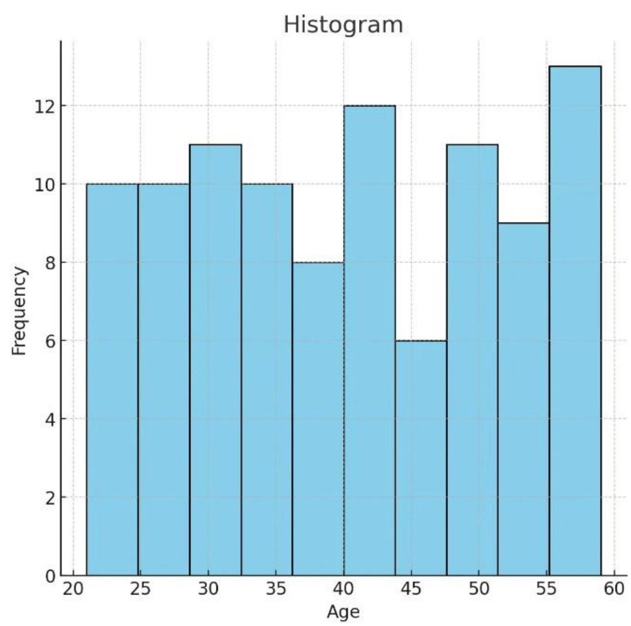
- **Box plots** often represent quartiles visually, showing the median, IQR, and potential outliers.
- The **IQR** is based on quartiles and is a measure of the spread of the middle 50% of the data.

IV. Data Visualization

Data visualization helps to represent and communicate data insights effectively. Various types of charts and plots are used to analyze data patterns, trends, and relationships.

1. Histograms

- **Purpose**: To show the distribution of a dataset.
- **Description**: A histogram is used to display the frequency distribution of a continuous variable. It divides the data into bins and shows how many data points fall within each bin.
- **Example**: A histogram for the ages in a dataset, which could be divided into bins like 20-30, 30-40, etc.
- **Key Points**:
  - Helps to understand the **distribution** and **spread** of the data.
  - Too many bins may lead to overfitting; too few bins may oversimplify the data.
- **Chart**:

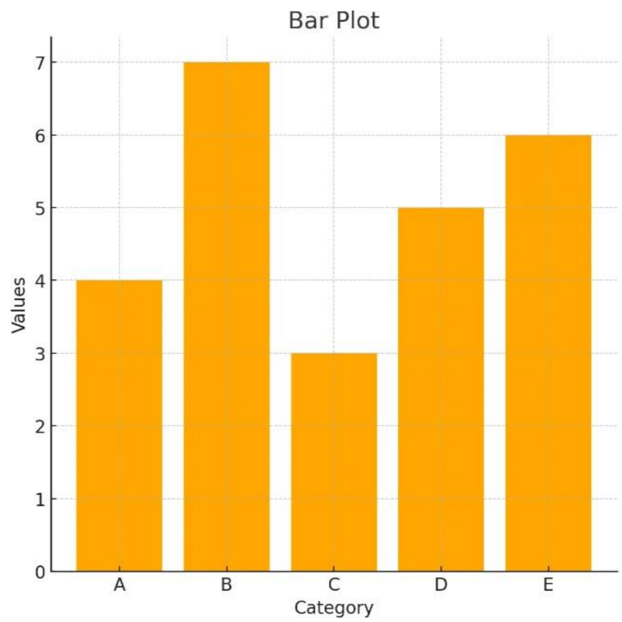


2. Bar Plots

- **Purpose**: To compare values across discrete categories.
- **Description**: Bar plots display categories along the x-axis and their corresponding values on the y-axis. Each bar represents one category.
- **Example**: A bar plot showing the number of sales for each product category.
- **Key Points**:
  - Best used for **categorical** data.

- Useful to compare values across different categories.

- **Chart**:

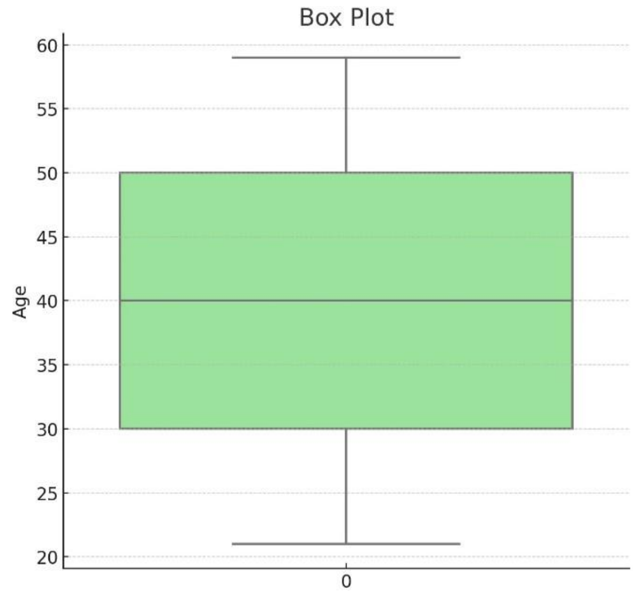


3. Pie Charts

- **Purpose**: To represent parts of a whole.
- **Description**: Pie charts display data as slices of a pie, where each slice represents a proportion of the total.
- **Example**: A pie chart showing the percentage distribution of sales across different regions.
- **Key Points**:
    - Suitable for **proportional** data.
    - Overuse of pie charts can be misleading, especially with many categories.
- **Chart**:

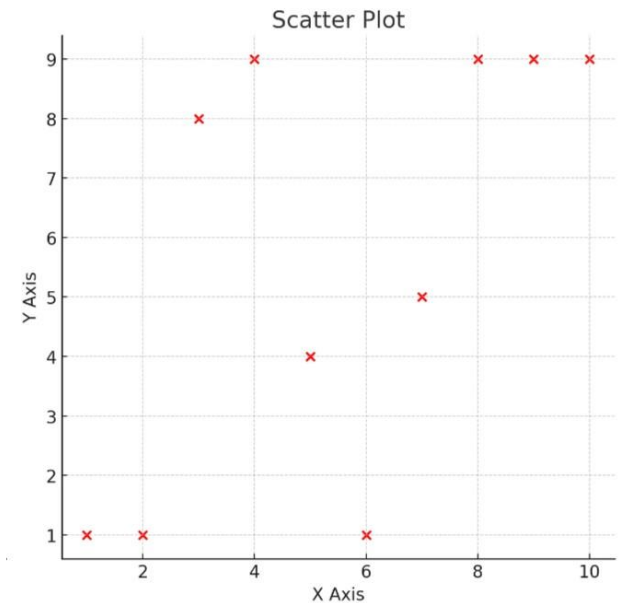
4. Box Plots

- **Purpose**: To visualize the distribution of a dataset and identify outliers.
- **Description**: Box plots display the minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum of a dataset. They also show outliers.
- **Example**: A box plot showing the distribution of test scores in a class.
- **Key Points**:
  - Useful for identifying **outliers** and understanding the **spread** of data.
  - Also used for comparing **distributions** between multiple datasets.
- **Chart**:



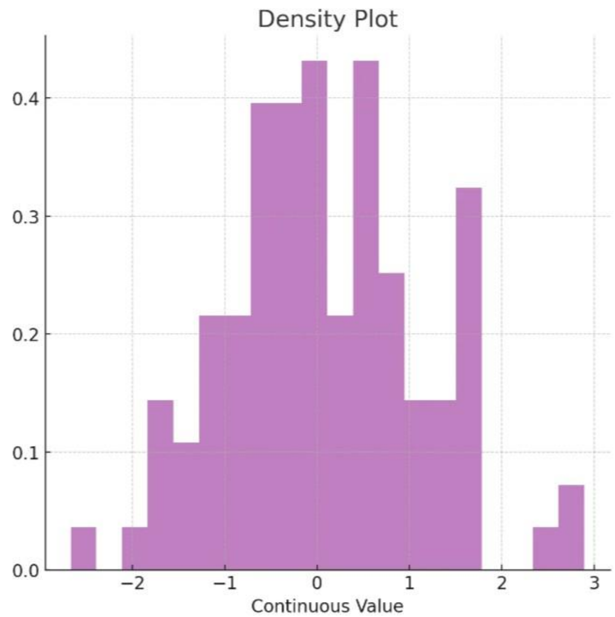
5. Scatter Plots

- **Purpose**: To explore the relationship between two continuous variables.
- **Description**: Scatter plots display individual data points on a two-dimensional plane, where each point represents a pair of values from two different variables.
- **Example**: A scatter plot showing the relationship between hours studied and exam scores.
- **Key Points**:
    - Great for detecting correlations or trends between two variables.
    - Outliers are clearly visible.
- **Chart**:



6. Density Plots

- **Purpose**: To estimate the distribution of a continuous variable.
- **Description**: A density plot is a smoothed version of a histogram. It shows the probability density function of a continuous variable.
- **Example**: A density plot showing the distribution of test scores for a group of students.
- **Key Points**:
  - Provides a smoother visualization of the distribution compared to histograms.
  - Useful for comparing multiple distributions.
- **Chart**:



V. Probability Basics

Probability is a branch of mathematics that deals with calculating the likelihood of an event occurring. In data science, probability theory is used to model uncertainty, predict outcomes, and understand the randomness in data.

1. Probability Theory

Probability theory is the mathematical framework used to describe and quantify uncertainty. It assigns a probability to each possible outcome of a random event. The probability of an event EE is a number between 0 and 1, where 0 means the event will not occur and 1 means the event will certainly occur.

- **Probability Formula**:

$$P(E) = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}}$$

- **Key Properties of Probability**:
  - The probability of any event is between 0 and 1: .

$$0 \leq P(E) \leq 1$$

- The sum of the probabilities of all possible outcomes of a random experiment equals 1.

- **Example**:
  When rolling a fair six-sided die, the probability of rolling a 3 is:

$$P(3) = \frac{1}{6}$$

---

2. Sample Spaces, Events

- **Sample Space**:
  The
  sample space is the set of all possible outcomes of a random experiment. For example, when flipping a coin, the sample space is ‹Heads, Tails}.

- **Event**:
  An
  event is a subset of the sample space, which may consist of one or more outcomes. For example, in the case of a die roll, an event could be the outcome of rolling an even number (i.e., ‹2, 4, 6›).

- **Example**:
  - For a die roll, the **sample space** is ‹1, 2, 3, 4, 5, 6›.
  - An event might be "rolling an even number," which is the set ‹2, 4, 6›.

- **Key Points**:
  - The total number of possible events is the number of subsets of the sample space.
  - The probability of an event is the sum of the probabilities of the outcomes within the event.

---

3. Conditional Probability

- **Conditional Probability** is the probability of an event occurring given that another event has already occurred. It is denoted as $P(A|B)$, meaning the probability of event A occurring given event B has occurred.

- **Formula**:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Where:

- P(A∩B) is the probability that both events A and B occur.
- P(B) is the probability of event B.

- **Example**:
  If you draw a card from a deck of 52 cards, the probability of drawing a red card (event A) given that the card is a heart (event B) is:

$$P(\text{Red}|\text{Heart}) = \frac{P(\text{Red} \cap \text{Heart})}{P(\text{Heart})} = \frac{1}{4} \div \frac{1}{4} = 1$$

- **Key Points**:
  - Conditional probability helps in situations where the outcome of one event affects the probability of another.
  - **Independence** of events impacts conditional probability.

---

4. Independent and Dependent Events

- **Independent Events**:
  Two events are
  independent if the occurrence of one event does not affect the occurrence of the other. Mathematically, two events A and B are independent if:

$$P(A \cap B) = P(A) \times P(B)$$

  - **Example**: Tossing a coin and rolling a die are independent events, as the result of one does not affect the result of the other.

- **Dependent Events**:
  Two events are
  dependent if the occurrence of one event affects the probability of the other. Mathematically, if A and B are dependent, then:

$$P(A \cap B) \neq P(A) \times P(B)$$

- **Example**: Drawing cards from a deck without replacement is an example of dependent events, as the first card drawn affects the probabilities of subsequent draws.

- **Key Points**:
  - Independence simplifies probability calculations.
  - Dependent events require more complex calculations, often involving conditional probability.

---

5. Bayes' Theorem

- **Bayes' Theorem** is a powerful tool in probability theory that allows you to update the probability of an event based on new evidence. It is particularly useful when dealing with conditional probabilities.
- **Formula**:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

  Where:
  - P(A|B) is the probability of event A given that event B has occurred.
  - P(B|A) is the probability of event B given that event A has occurred.
  - P(A) and P(B) are the prior probabilities of events A and B, respectively.

- **Example**:
  Suppose 1% of the population has a disease (event A), and a diagnostic test for the disease is 95% accurate (event B). If a person tests positive, what is the probability they actually have the disease? Using Bayes' Theorem, you can calculate this probability based on the accuracy of the test and the prevalence of the disease in the population.
- **Key Points**:
  - Bayes' Theorem is essential for **updating beliefs** in light of new data.
  - It plays a central role in **machine learning**, especially in classification tasks (e.g., Naive Bayes classifiers).