82. What is Statistics And its Application

Introduction to Statistics and Its Applications

Commands

- No technical commands were mentioned in this lecture.

Summary

- **Statistics** is defined as the field dealing with the **collection**, **organization**, **analysis**, **interpretation**, and **presentation** of data.
- The primary goal of statistics is to utilize data to understand behaviors and factors that lead to effective **decision-making** and business growth.
- Statistical analysis involves calculating metrics like **mean** and **median** and analyzing data **distributions** (e.g., Gaussian, Log-Normal).
- Visualization tools such as **Histograms**, **Probability Density Functions (PDF)**, and **Cumulative Density Functions (CDF)** are used to interpret data patterns.
- Statistics is fundamental to various roles, including **Machine Learning**, **Data Science**, **Data Analysis**, **Business Intelligence**, and **Risk Analysis**.
- Real-world applications range from business decisions (e.g., ATM placement) to scientific validation (e.g., **COVID-19 vaccination** safety).

What is Statistics?

Statistics is a field that deals with the following key processes regarding data:

- **Collection**
- **Organization**
- **Analysis**
- **Interpretation**
- **Presentation**

The Purpose of Statistics

The ultimate goal of performing these statistical processes is **decision-making**. By analyzing data, organizations can:

- Observe **customer behavior**.
- Identify important factors influencing outcomes.
- Make informed decisions to ensure **business profitability**.

Statistical Analysis Techniques

To make decisions, raw data must be analyzed using specific statistical tools and concepts.

Data Features and Metrics

Using a feature such as **Age** in an online shopping dataset, analysts can determine target demographics for promotional offers by calculating:

- **Mean**: The average value.

- **Median**: The middle value.

Distributions

Understanding the **distribution** of data is crucial. Common distribution types include:

- **Gaussian Distribution** (Normal Distribution)
- **Standard Normal Distribution**
- **Log-Normal Distribution**

Visualization

Statistics involves creating charts and graphs to understand data patterns:

- **Histogram**: Vertical bar charts used to represent data frequency.
- **PDF (Probability Density Function)**: A smoothed version of a histogram used to understand distribution.
- **CDF (Cumulative Density Function)**: Used for cumulative probability analysis.

Real-World Application Examples

Business Decision: ATM Placement

A bank uses statistics to decide whether to open a new ATM in **Location B**, five kilometers away from an existing ATM in **Location A**.

- **Process**: Analyze historical data from Location A (e.g., **mean transactions** per month, electricity costs, user traffic).
- **Outcome**: Make a **statistical decision** on whether Location B will be efficient and profitable based on the patterns observed in Location A.

Scientific Validation: Vaccination Safety

Statistics played a critical role during the **COVID-19 pandemic** to determine vaccine safety.

- **Process**: Select a sample group of people, administer the vaccine, and perform **statistical analysis** on the results.
- **Outcome**: Conclude whether the vaccination is safe for the general population based on experimental data.

Domains Using Statistics

Statistics is extensively used across various fields and roles, including:

- **Machine Learning** and **Data Science**
- **Data Analysis**
- **Business Intelligence** (BI) Developers and **Business Analytics**
- **Risk Analysis**
- Everyday activities and general decision-making.

83. Types Of Statistics

Types of Statistics in Data Science

Commands

- No technical commands were mentioned in this lecture.

Summary

- **Statistics** is broadly categorized into two main types: **Descriptive Statistics** and **Inferential Statistics**.
- **Descriptive Statistics** focuses on **organizing** and **summarizing** data to understand its features.
- Key techniques in descriptive statistics include **Measure of Central Tendency** (Mean, Median, Mode) and **Measure of Dispersion** (Variance, Standard Deviation) .
- **Inferential Statistics** involves collecting **sample data** to make **conclusions** or **inferences** about a larger **population data** set .
- Inferential statistics utilizes experiments and tests, such as **Z-test** and **T-test**, to derive conclusions.
- The distinction between **sample data** (subset) and **population data** (entirety) is fundamental to inferential statistics .

Exam Notes

Interview Question: Types of Statistics

Question: What are the two different types of statistics? Explain them with examples.

Answer: The two main types are **Descriptive Statistics** and **Inferential Statistics**. Descriptive statistics organizes and summarizes data (e.g., calculating the average height of a class), while inferential statistics uses sample data to make conclusions about a larger population (e.g., estimating the average height of all students in a college based on one class).

Descriptive Statistics

Descriptive Statistics is the branch dealing with the **organizing** and **summarizing** of data . It uses specific techniques to analyze the characteristics of a dataset.

Techniques Used

1. **Measure of Central Tendency**: This involves calculating metrics that represent the center point of a dataset.
    - **Mean**
    - **Median**
    - **Mode**
2. **Measure of Dispersion**: This helps in understanding the spread or variability of the data.
    - **Variance**
    - **Standard Deviation**

Inferential Statistics

Inferential Statistics deals with collecting data and using it to form **conclusions** or **inferences** through experiments .

Key Concepts

- **Process**:
    1. Collect **Sample Data**.
    2. Perform experiments (e.g., **Z-test**, **T-test**).
    3. Derive conclusions regarding the **Population Data**.
- **Population vs. Sample**:
    - **Sample Data**: A smaller subset of data collected for analysis.
    - **Population Data**: The larger, total dataset about which conclusions are made. The size of population data is always greater than sample data.

Practical Example: College Student Heights

To illustrate the difference between the two types, consider a scenario involving a college (College A) with **1000 students** .

Scenario Setup

- **Population**: The entire college consisting of 1000 students.
- **Sample**: A specific class of statistics students selected from the college.
- **Data Collected**: The heights of students in the sample class (e.g., 180cm, 170cm, 162cm, 150cm, 160cm) .

Applying Descriptive Statistics

In this context, descriptive statistics would involve calculating exact metrics for the **sample** itself.

- **Action**: Calculating the **mean (average) height** or median height of the specific students in the sample class.
- **Result**: Stating "The average height of this class is 165cm." This summarizes the data effectively for the group measured.

Applying Inferential Statistics

Inferential statistics uses the sample data to estimate characteristics of the entire **population**.

- **Action**: Using the height data from the sample class to reach a conclusion about the entire college.
- **Question**: "Based on this sample, what is the average height of all 1000 students?".
- **Result**: Making an inference or conclusion about the height of the entire population of 1000 students based on the experiments performed on the sample.

84. Population Vs Sample Data

Population and Sample

Commands

- No technical commands were mentioned in this lecture.

Summary

- **Population** refers to the entire set of data or individuals being studied (e.g., all people on an island).
- **Sample** is a subset selected from the population used to represent the whole (e.g., 10,000 people selected from 100,000).
- Sampling is necessary when it is logistically difficult or impossible to collect data from every individual in a population.
- **Notation**: Population size is denoted by **Capital N** (N), and Sample size is denoted by **small n** (n).
- **Inferential Statistics** involves using sample data to make conclusions or predictions about the population, such as in **exit polls**.

Key Concepts: Population vs. Sample

Before diving into measures of tendency or dispersion, it is crucial to understand two fundamental concepts in statistics: **Population** and **Sample**.

Population (N)

- **Definition**: The total number of individuals or data points in the specific group being studied.
- **Symbol**: Denoted by the capital letter **N**.

- **Example**: Consider an island where the total number of people living there is **100,000**. This entire group of 100,000 people represents the **Population**.

Sample (n)

- **Definition**: A smaller, manageable subset selected from the population.
- **Symbol**: Denoted by the small letter **n**.
- **Example**: From the island's population of 100,000, if you select **10,000 people** to study, this specific group is called the **Sample**.

Why Do We Use Samples?

Collecting data from an entire population is often impractical.

The Island Scenario

Imagine you are tasked with collecting the **weight** of every person on the island:

- **The Challenge**: Visiting 100,000 people individually to record their weight (e.g., 100kg, 70kg) is extremely difficult.
- **Logistical Issues**:
  - It is hard to locate everyone.
  - Some people might be absent or off the island.
  - The time and effort required are prohibitive.

The Solution

Instead of measuring everyone, you select a **Sample** (e.g., 10,000 people) to represent the population. You collect data from this subset to make estimates about the whole.

Applications: Inferential Statistics

Sampling is the foundation of **Inferential Statistics**, where we perform experiments on a sample to infer conclusions about the population.

Real-World Example: Exit Polls

- **Scenario**: During an election, it is impossible to ask every single voter who they voted for immediately.
- **Method**: News channels collect data from a **sample** of voters as they leave polling stations.
- **Outcome**: Based on this sample data, they predict (infer) which candidate or party is likely to win the election with the majority of votes.

Mathematical Notation

Understanding these symbols is essential for future topics like **Population Mean** vs. **Sample Mean**:

- **Population Size**: N (Capital N)
- **Sample Size**: n (Small n)

85. Measure Of Central Tendency

Measure of Central Tendency

Summary

- **Measure of Central Tendency** consists of three important sub-topics: **Mean**, **Median**, and **Mode**.
- The formula for **Mean** differs slightly depending on whether the data represents a **Population** or a **Sample**.
- **Mean** is sensitive to **outliers** (large or small values that differ significantly from other data points), which can drastically skew the average.
- **Median** is the central element of a sorted dataset and is robust against **outliers**, providing a better representation of the center in skewed distributions.
- **Mode** identifies the element with the **maximum frequency** and is also useful for handling distributions with outliers .
- Understanding the specific notations (e.g., μ for population mean, $\bar{x}$ for sample mean) is crucial for future topics like Measure of Dispersion.

Mean (Average)

The **Mean** represents the average of a dataset. The notation and formula change based on whether the dataset is a **Population** or a **Sample**.

Population Mean

For a **Population** of size N, the mean is denoted by the symbol μ (mu).

The formula is:

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

- x_i: Data points present in the population.
- N: Population size.

Sample Mean

For a **Sample** of size n, the mean is denoted by the symbol $\bar{x}$ (x-bar).

The formula is:

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

- n: Sample size.

Calculation Example

Consider a variable **Age** with values: `{1, 3, 4, 5}`.

To find the mean:

1. Sum the values: $1 + 3 + 4 + 5 = 13$.

2. Divide by the count (4): $13/4 = 3.25$.

The **Mean** is **3.25**, representing the central tendency of this distribution.

Median

The **Median** is the central value of a dataset. It is particularly useful for overcoming the impact of **outliers**.

The Impact of Outliers

If an **outlier** (a very large number, e.g., 100) is added to the previous dataset `{1, 3, 4, 5}`, the new set becomes `{1, 3, 4, 5, 100}`.

- **New Mean Calculation**: $(1 + 3 + 4 + 5 + 100)/5 = 113/5 = 22.6$.
- **Observation**: The mean jumped from **3.25** to **22.6** solely due to the outlier. This drastic change suggests the mean may no longer accurately represent the central tendency of the data.

Calculating Median

To calculate the median, you must first **sort the numbers**.

Odd Number of Elements

Dataset: `{1, 3, 4, 5, 100}` (Sorted).

- Count (n) = 5 (Odd).
- Select the **central element**.
- The 3rd element is **4**.
- **Median = 4**.

Compared to the mean of 22.6, the median of 4 is much closer to the original average (3.25) and is not heavily impacted by the outlier.

Even Number of Elements

If another outlier (e.g., 200) is added: `{1, 3, 4, 5, 100, 200}`.

- Count (n) = 6 (Even).
- Identify the two central elements: **4** and **5**.
- Calculate the average of these two elements: $(4 + 5)/2 = 4.5$.
- **Median = 4.5**.

Even with two large outliers, the median remains stable.

Mode

Mode is another technique used to measure central tendency that is also robust against **outliers**.

Definition

The **Mode** is defined as the element with the **maximum frequency** (the value that appears most often).

Calculation Example

Consider the dataset: `{4, 3, 2, 1, 1, 4, 4, 5, 2, 100}`.

- Frequency analysis:
  - 1: 2 times
  - **4: 3 times**
  - (Other numbers appear less frequently)

- The element **4** has the highest frequency.
- **Mode = 4**.

The mode focuses on the most frequent element, ignoring the magnitude of outliers like 100.

86. Measure Of Dispersion

Measure of Dispersion: Variance

Summary

- **Measure of Dispersion** is used to differentiate distributions that may have the same **mean** but different **spreads** of data.
- The two main components of dispersion discussed are **Variance** and **Standard Deviation**.
- **Variance** measures how far a set of numbers is spread out from their average value.
- Formulas for variance differ depending on whether the data represents a **Population** or a **Sample**.
- **Population Variance** (σ^2) is calculated by dividing the sum of squared differences from the mean by the total number of elements (N).
- **Sample Variance** (s^2) is calculated by dividing the sum of squared differences from the sample mean by $n - 1$.

Exam Notes

Sample Variance Denominator

Question: Why do we divide the sample variance by $n - 1$ instead of N?

Answer: This is a very important **interview question** regarding **Sample Variance**. While **Population Variance** divides by N, **Sample Variance** uses $n - 1$ (known as Bessel's correction) to provide an unbiased estimator of the population variance. This specific distinction is critical when working with sample data versus population data.

Introduction to Measure of Dispersion

The **Measure of Dispersion** is a statistical concept used to describe how spread out or scattered a dataset is. It is essential because calculating the **Mean** (average) alone is often insufficient to understand the nature of a distribution.

Two different datasets can have the exact same **Mean**, yet their data points can be distributed very differently. **Variance** and **Standard Deviation** are the tools used to quantify this spread.

Variance Calculation Example

To understand why variance is necessary, consider two distinct distributions of ages with the same number of elements ($n = 4$).

Comparing Two Distributions

Distribution 1: `{2, 2, 4, 4}`

- **Mean Calculation**: $(2 + 2 + 4 + 4)/4 = 3$.
- **Observation**: The data points (2 and 4) are very close to the mean (3).

Distribution 2: `{1, 1, 5, 5}`

- **Mean Calculation**: $(1 + 1 + 5 + 5)/4 = 3$.
- **Observation**: The data points (1 and 5) are further away from the mean (3) compared to the first distribution.

Although both have a **Mean** of **3**, Distribution 2 has a higher **spread** or **dispersion**. Variance allows us to calculate a specific number to represent this spread.

Population Variance

When calculating variance for **Population Data** (denoted by capital N), the formula uses the symbol σ^2 (Sigma Square)
.

Formula

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

- x_i: Individual data points.
- μ: Population Mean.
- N: Population size.

Calculation Steps

Using the previous examples as **Population Data**:

1. **For Distribution 1 (`2, 2, 4, 4`):**
    - Calculate squared differences from mean (3): $(2-3)^2 = 1$, $(2-3)^2 = 1$, $(4-3)^2 = 1$, $(4-3)^2 = 1$.
    - Sum of squares: $1+1+1+1 = 4$.
    - Divide by N (4): $4/4 = 1$.
    - **Variance (σ^2) = 1**.
2. **For Distribution 2 (`1, 1, 5, 5`):**
    - Calculate squared differences from mean (3): $(1-3)^2 = 4$, $(1-3)^2 = 4$, $(5-3)^2 = 4$, $(5-3)^2 = 4$.
    - Sum of squares: $4+4+4+4 = 16$.
    - Divide by N (4): $16/4 = 4$.
    - **Variance (σ^2) = 4**.

Conclusion: The higher variance in Distribution 2 (4 vs 1) mathematically confirms that its data is more dispersed.

Sample Variance

When working with **Sample Data** (denoted by small n), the formula changes slightly to provide a more accurate estimate. The symbol used is s^2.

Formula

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

- $\bar{x}$: Sample Mean (used instead of μ).
- n: Sample size.
- **Denominator**: The division is by $n-1$ rather than N.

The reason for using $n-1$ is a key concept in statistics and is often a topic of discussion in technical interviews.

87. Why Sample Variance Is Divided By n-1?

Measure of Dispersion: Sample Variance

Commands

- No commands were used in this lesson.

Summary

- **Sample Variance** (s^2) is calculated using a specific formula that divides by $n-1$ instead of n.
- **Population Variance** (σ^2) is calculated by dividing by the total population size (N).
- The adjustment of dividing by $n-1$ is known as **Bessel's correction**.
- Using $n-1$ ensures the calculation provides an **unbiased estimation** of the **true population variance**.
- Dividing by n when working with sample data typically leads to **underestimating** the variance.
- The term $n-1$ is also referred to as the **Degrees of Freedom (DOF)**.

Exam Notes

Sample Variance Denominator

Question: Why do we divide the sample variance by $n-1$ instead of n?

Answer: This is a frequent and **important interview question**. When we select a sample, the data points are naturally closer to the **sample mean** ($\bar{x}$) than they are to the **population mean** (μ). If we divide by n, the result tends to be smaller than the actual variance, meaning we are **underestimating the true population variance**. Dividing by $n-1$ (a smaller number) increases the result slightly, correcting this bias and providing an **unbiased estimation**.

Sample Variance Formula

The formula for **Sample Variance**, denoted as s^2, differs slightly from the population variance formula.

The Formula

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

- s^2: Sample Variance
- n: Sample data size
- x_i: Individual data points
- $\bar{x}$: Sample Mean
- $n-1$: The divisor used for **Bessel's correction**

Comparison with Population Variance

For context, the **Population Variance** (σ^2) uses the total population size N in the denominator:

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

- μ: Population Mean
- N: Population data size

Bessel's Correction and Unbiased Estimation

The primary reason for the difference in formulas lies in the goal of making accurate **inferences** about a population based on a sample.

Underestimation Problem

- When collecting **sample data**, we calculate a **sample mean** ($\bar{x}$).
- While the sample mean is often close to the **population mean** (μ), specific samples might have data points that are clustered or skewed.
- If we calculate the distance of sample points from the **sample mean**, the variance calculated with n will often be much smaller than the variance calculated from the **true population mean**.
- Using n creates a biased result that **underestimates** the true spread of the population.

The Solution

- By dividing by $n-1$ instead of n, we are dividing by a smaller number.
- Mathematically, this increases the value of the variance, compensating for the underestimation.
- This adjustment makes the sample variance an **unbiased estimator**, meaning it is a more accurate reflection of the **true population variance**.

Degrees of Freedom

- In statistics, the term $n-1$ is technically referred to as the **Degree of Freedom (DOF)**.
- This concept is specific to calculations involving **sample data**.
- Mentioning **Degrees of Freedom** is a valid and technical way to explain the concept during an interview.

88. Standard Deviation

Standard Deviation and Formula Revision

Commands

- No commands were used in this lesson.

Summary

- **Population Statistics** utilize capital N for size, μ for mean, and σ for standard deviation.
- **Population Variance** (σ^2) is calculated by dividing the sum of squared differences by N.
- **Standard Deviation** is the **square root** of the variance.
- While **Variance** represents the overall **spread** or **dispersion** of data, **Standard Deviation** quantifies **how far a specific data point is away from the mean**.
- **Sample Statistics** utilize small n for size, $\bar{x}$ for mean, and s for standard deviation.
- **Sample Variance** (s^2) differs from population variance by dividing by $n-1$ (Bessel's correction) instead of n.

Exam Notes

Distinguishing Terminologies

Question: How do you distinguish between Population and Sample statistics in calculations?

Answer: It is critical to distinguish between the terminologies and formulas for **Population** and **Sample** data.

- **Population**: Uses μ (mean), σ^2 (variance), and divides by N.
- **Sample**: Uses $\bar{x}$ (mean), s^2 (variance), and divides by $n-1$.

Population Statistics Formulas

When dealing with the entire group (Population), specific symbols and formulas are used.

Population Mean (μ)

The population mean is the sum of all data points divided by the total population size (N).

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

- N: Population size.
- x_i: Individual data points.

Population Variance (σ^2)

This measures the dispersion of the population.

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

Population Standard Deviation (σ)

The standard deviation is derived directly from the variance.

$$\sigma = \sqrt{\sigma^2} = \quad \text{Population Variance}$$

Understanding Standard Deviation

While variance provides a measure of spread, **Standard Deviation** offers a more interpretable metric regarding the distance of data points from the mean .

- **Definition**: It indicates **how far a data point is away from the mean**.
- **Usage**: It is used as a unit of measurement to describe the position of data points relative to the center.

Example Scenario

Consider a dataset with a **Mean of 3** and a **Standard Deviation of 1**.

- **Data Point 4**: This point is **one standard deviation to the right** of the mean ($3 + 1 = 4$).
- **Data Point 2**: This point is **one standard deviation to the left** of the mean ($3 - 1 = 2$).
- **Data Point 4.5**: This point would be **1.5 standard deviations** away from the mean.

Sample Statistics Formulas

When working with a subset of data (Sample), the formulas adjust to provide unbiased estimates.

Sample Mean ($\bar{x}$)

The sample mean uses small n for the number of items.

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

Sample Variance (s^2)

The sample variance includes **Bessel's correction**, dividing by $n - 1$.

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

Sample Standard Deviation (s)

Like the population metric, this is the square root of the variance.

$$s = \sqrt{s^2} = \quad \text{Sample Variance}$$

89. What Are Variables?

Statistics: Variables

Commands

- No commands were used in this lesson.

Summary

- A **Variable** is a property that can take on any value (e.g., Age, Gender, Height).
- It is distinct from a fixed list of values; a variable represents the attribute itself.
- **Quantitative Variables** deal with numerical values and are split into two types:
  - **Discrete Quantitative Variables**: Must be whole numbers (integers); cannot be fractions (e.g., number of children).
  - **Continuous Quantitative Variables**: Can take on any value within a range, including decimals and fractions (e.g., height, weight).
- **Qualitative (Categorical) Variables** deal with non-numerical categories or labels (e.g., gender, colors).

Exam Notes

Variable Types and Examples

Question: How do you differentiate between discrete, continuous, and categorical variables? Can you provide examples?

Answer: This is a common **interview question**. You must be able to define the types and give specific examples:

- **Discrete**: Finite, whole numbers (e.g., **Number of students** in a class).
- **Continuous**: Infinite possibilities including decimals (e.g., **Height** or **Weight**).
- **Categorical**: Non-numeric groups (e.g., **Gender** or **Colors**).

Definition of a Variable

In statistics, it is crucial to understand what a variable is before analyzing data.

- **Definition**: A **variable** is a property that can take up any value.
- **Concept**: It is an attribute where the data varies.
  - **Example**: **Age** is a variable because it can be assigned different values like 25 or 30.

- **Counter-Example**: A static list of ages (e.g., `{20, 25, 22}`) is a collection of data, not the variable itself. The variable is the "container" or property named **Age**.

- **Common Examples**:
  - **Gender**: Can be Male or Female.
  - **Height**: Can be 172 cm, 180 cm, etc.

Types of Variables

Variables are broadly classified into two main categories: **Quantitative** and **Qualitative**.

1. Quantitative Variables

These variables represent numerical data. They are further divided into two sub-types:

A. Discrete Quantitative Variable

- **Definition**: Variables that can only take on specific, distinct values, typically **whole numbers**. They cannot be fractions or decimals.
- **Key Characteristic**: You count these values.
- **Examples**:
  - **Number of children**: A person can have 3 children, but not 2.5 or 4.5 children.
  - **Number of houses**: Someone can own 5 houses, not 5.5.
  - **Number of bank accounts**: You can have 5 accounts, but not 5.5.
  - **Number of students in a class**: A class can have 50 students, not 45.5.
  - **Number of workers in a company**: There can be 100 workers, not 99.5.

B. Continuous Quantitative Variable

- **Definition**: Variables that can take on any value within a range, including **decimals** and **fractions**.
- **Key Characteristic**: You measure these values.
- **Examples**:
  - **Height**: Can be 175.5 cm, 182 cm, etc.
  - **Weight**: Can be 180 lbs, 90 lbs, 72.5 kg, 72.7 kg.
  - **Age**: While often treated as whole numbers, age is technically continuous (e.g., 25.5 years old).

2. Qualitative (Categorical) Variables

- **Definition**: Variables that represent types, qualities, or categories rather than numerical amounts. They do not have logical mathematical order or magnitude in the same way numbers do.
- **Examples**:
  - **Gender**: Categories like Male, Female.
  - **Colors**: Categories like Red, Green, Blue.
  - **Locations**: Categories like States, Cities, Places.

90. What are Random Variables

Statistics: Random Variables

Commands

- No specific commands were used in this lesson.

Summary

- A **Random Variable** (denoted by X) is a function whose values are derived from a random process or experiment.
- It quantifies the outcomes of a random phenomenon by assigning numerical values to them.
- There are two main types of random variables:
  - **Discrete Random Variables**: Typically represent countable outcomes, often whole numbers (e.g., tossing a coin, rolling a die).
  - **Continuous Random Variables**: Can take on any value within a range, including fractions and decimals (e.g., amount of rainfall, height of people).
- Understanding random variables is crucial for fields like **machine learning** and **deep learning**.

Introduction to Random Variables

A **Random Variable** is a fundamental concept in statistics, used extensively in data science and machine learning.

- **Notation**: It is typically denoted by a capital letter, such as X.
- **Definition**: A random variable is a **function** that assigns values derived from different processes or experiments.

To understand the concept, consider a simple algebraic equation: $y = 5x + 2$. In this equation, x acts as a variable that can take different inputs to produce different outputs (y). Similarly, a random variable takes the outcomes of a random process and maps them to numerical values.

Example: Tossing a Coin

Consider the experiment of **tossing a coin**.

- **Process**: Tossing the coin.
- **Possible Outcomes**: Head or Tail.
- **Random Variable Assignment**: We can define a function where we assign specific values to these outcomes:
  - If **Head**: Assign value **0**.
  - If **Tail**: Assign value **1**.

This assignment makes "Tossing a Coin" a process where the random variable derives specific values based on the outcome.

Example: Rolling a Fair Die

Consider the experiment of **rolling a fair die**.

- **Possible Outcomes**: The values can be **1, 2, 3, 4, 5, or 6**.
- Each roll produces one of these specific values derived from the experiment.

Types of Random Variables

Random variables are categorized into two distinct types based on the nature of the values they can assume.

1. Discrete Random Variable

A **Discrete Random Variable** derives values from processes that result in distinct, countable outcomes.

- **Characteristics**: The values are usually **whole numbers** or specific categorical values mapped to numbers.
- **Examples**:
  - **Tossing a Coin**: Results in 0 or 1.
  - **Rolling a Die**: Results in specific integers {1, 2, 3, 4, 5, 6}.

2. Continuous Random Variable

A **Continuous Random Variable** derives values from processes that can take on any value within a continuum or range.

- **Characteristics**: These variables can assume **infinite possibilities**, including **fractions** and **decimals**.

- **Examples**:
  - **Rainfall**: If predicting how many inches of rain will fall tomorrow, the value could be **1.1 inches**, **5.5 inches**, or **10.75 inches**. It is not restricted to whole numbers.
  - **Height of People**: Measuring the height of attendees at an event can yield values like **150 cm**, **160 cm**, or **160.1 cm**.

Comparison

Feature	Discrete Random Variable	Continuous Random Variable
Values	Countable, distinct values (often whole numbers)	Infinite values within a range (includes decimals)
Example Process	Counting items, Tossing coins	Measuring physical quantities
Example Data	0, 1, 2, 3	1.5, 2.75, 10.1

91. Histograms- Descriptive Statistics

Statistics: Histograms

Commands

- No specific commands were used in this lesson.

Summary

- **Histograms** are a fundamental statistical tool used to visualize the distribution of data.
- They serve as the foundation for deriving the **Probability Density Function (PDF)**.
- A histogram is constructed by creating **bins** (intervals) and counting the **frequency** of data points within those bins .
- **Kernel Density Estimation (KDE)** is a technique used to **smoothen** a histogram to create a continuous probability density curve.
- Histograms can represent both **continuous** and **discrete** data, though the visualization may differ slightly.

Introduction to Histograms

Histograms are a critical concept in statistics, primarily used to visualize how data is distributed. They are particularly important because they enable the derivation of the **Probability Density Function (PDF)** using techniques like **Kernel Density Estimation (KDE)**.

Constructing a Histogram: Step-by-Step

To understand the construction of a histogram, consider a random variable representing **Age**.

1. The Dataset

Consider the following set of values for the random variable **Age**:

`{23, 24, 25, 30, 34, 36, 40, 50, 60, 75, 80}`.

2. Defining Bins and Axis

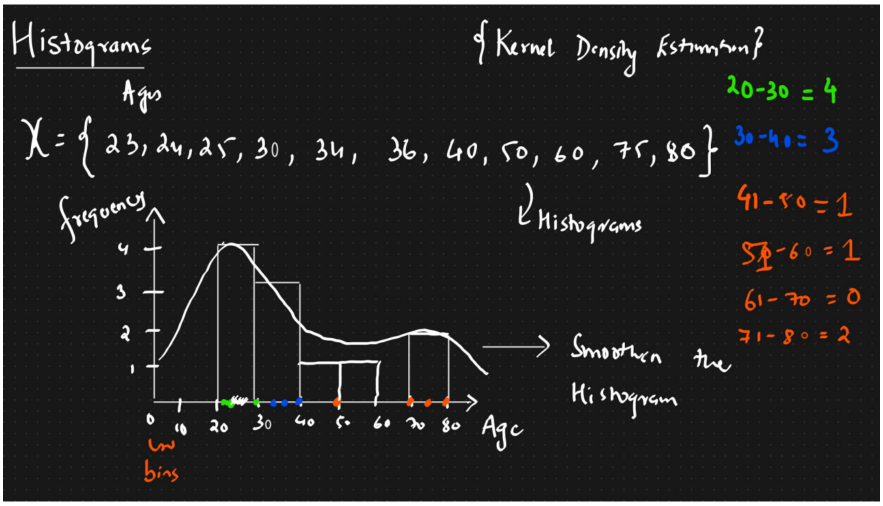
- Create a single-dimensional axis representing the data range (e.g., 0 to 80).
- Divide this axis into equal intervals known as **bins**. In this example, the **bin size** is set to **10** (e.g., 0-10, 10-20, 20-30, etc.).

- The bin size is customizable and can be adjusted based on the specific analysis requirements or code parameters.

3. Calculating Frequency

The core task is to count the number of data points (frequency) that fall into each specific bin.

- **20 to 30**: Contains values `{23, 24, 25, 30}`.
    - **Count**: 4.
- **30 to 40**: Contains values `{34, 36, 40}`.
    - **Count**: 3.
- **40 to 50**: Contains value `{50}`.
    - **Count**: 1.
- **50 to 60**: Contains value `{60}`.
    - **Count**: 1.
- **60 to 70**: No values exist in this range.
    - **Count**: 0.
- **70 to 80**: Contains values `{75, 80}`.
    - **Count**: 2.



4. Visualizing the Structure

The histogram is plotted with the **bins on the X-axis** and the **frequency (count) on the Y-axis**.

- **Building Blocks**: For each bin, a rectangular block is drawn with a height corresponding to its frequency.
    - The bin **20-30** has a block height of **4**.
    - The bin **30-40** has a block height of **3**.
    - The bin **60-70** has no block (height 0).

This structure visually represents the frequency of elements ranging between specific bins.

Probability Density Function (PDF) and KDE

One of the most powerful applications of a histogram is its ability to approximate the **Probability Density Function (PDF)**.

- **Smoothening**: By "smoothening" the blocky structure of a histogram, you can derive a continuous curve that represents the PDF.
- **Kernel Density Estimation (KDE)**: The mathematical concept used to perform this smoothening is called the **Kernel Density Estimator**. It helps derive the smoothened curve from the histogram data.

92. Percentile And Quartiles- Descriptive Statistics

Statistics: Percentiles and Quartiles

Commands

- No commands were used in this lesson.

Summary

- **Percentage** is a mathematical ratio representing a fraction of 100, while **Percentile** is a statistical measure indicating relative standing.

- A **Percentile** is defined as a value below which a certain percentage of observations lie.
- **Percentile Ranking** is calculated by determining the percentage of values in a distribution that are less than a specific value (x).
- To find the value corresponding to a specific percentile, the formula involves $(n + 1)$.
- **Quartiles** divide a distribution into four equal parts:
  - **1st Quartile (Q1)**: 25th Percentile.
  - **2nd Quartile (Q2)**: 50th Percentile (Median).
  - **3rd Quartile (Q3)**: 75th Percentile.

Exam Notes

Percentile vs. Percentage

Question: What is the practical difference between Percentage and Percentile in exams like CAT or GATE?

Answer: While **Percentage** calculates a score based on total marks (e.g., getting 3 out of 6 odd numbers is 50%), **Percentile** represents a **ranking** relative to other participants. For example, a **99th percentile** score means the candidate performed better than **99%** of all other test-takers.

Understanding Percentiles

The concept of percentiles is distinct from percentages. To illustrate, consider a simple list of numbers: `{1, 2, 3, 4, 5, 6}`.

- **Percentage**: Calculating the percentage of odd numbers involves counting them (3) and dividing by the total count (6), resulting in 50%.
- **Percentile**: This measures the position of a value relative to the rest of the dataset.

Definition

A **Percentile** is a value below which a certain percentage of observations in a distribution lie. It is frequently used in competitive exams and data analysis to understand the distribution of data.

Calculating Percentile Ranking

To find the percentile rank of a specific value (x) in a dataset, use the following formula:

$$\text{Percentile of } x = \left(\frac{\text{Number of values below } x}{n}\right) \times 100$$

- x: The value being evaluated.
- n: The total sample size (total number of values).

Example Calculation

Consider the sorted dataset: `{2, 2, 3, 4, 5, 5, 6, 7, 8, 8, 8, 9, 9, 10}`

- **Goal**: Find the percentile rank of the value **9**.
- **Step 1**: Count the total number of values (n). Here, $n = 14$.
- **Step 2**: Count the number of values strictly **less than 9**.
  - Values: `{2, 2, 3, 4, 5, 5, 6, 7, 8, 8, 8}`
  - Count = 11.
- **Step 3**: Apply the formula.
  - $\text{Percentile} = (\frac{11}{14}) \times 100$
  - $\text{Percentile} \approx 78.57\%$

Interpretation: **78.57%** of the entire distribution is less than the value 9.

Calculating Value from Percentile

Conversely, you may need to find which value in a dataset corresponds to a specific percentile (e.g., finding the 25th percentile).

Formula

$$\text{Value Index} = \frac{\text{Percentile}}{100} \times (n+1)$$

- **Percentile**: The target percentile (e.g., 25, 50, 75).
- n: The total sample size.

Example Calculation

Using the same dataset ($n = 14$), calculate the **25th Percentile**.

1. **Calculate the Index**:
    - $\text{Index} = \frac{25}{100} \times (14+1)$
    - $\text{Index} = 0.25 \times 15 = 3.75$
2. **Determine the Value**:
    - The index **3.75** is not a whole number, meaning the value lies between the **3rd** and **4th** positions in the sorted list.
    - **3rd Value**: 3
    - **4th Value**: 4
    - To find the exact percentile value, take the average (mean) of these two values.
    - $\text{Average} = \frac{3+4}{2} = 3.5$

Result: The 25th percentile of the distribution is **3.5**. This indicates that 25% of the distribution is less than 3.5.

Quartiles

Quartiles are specific percentiles that divide the data into four distinct quarters.

- **1st Quartile (Q1)**: Represents the **25th Percentile**. It is the value below which 25% of the data lies.
- **2nd Quartile (Q2)**: Represents the **50th Percentile**. This is also known as the **Median**.
- **3rd Quartile (Q3)**: Represents the **75th Percentile**. It is the value below which 75% of the data lies.

93. 5 Number Summary-Descriptive Statistics

Statistics: Five Number Summary and Box Plot

Commands

- No commands were used in this lesson.

Summary

- The **Five Number Summary** is a set of descriptive statistics that provides information about a dataset's range and distribution.
- It consists of five key values: **Minimum**, **First Quartile (Q1)**, **Median**, **Third Quartile (Q3)**, and **Maximum**.

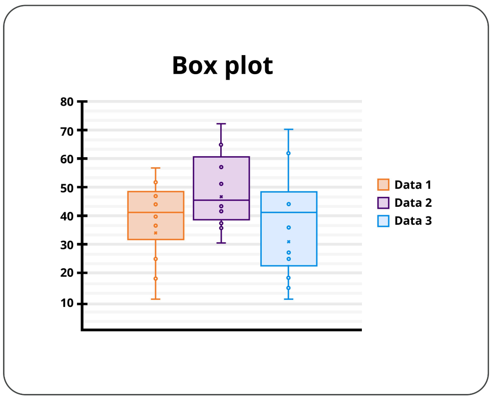
- These values are essential for data preprocessing steps in machine learning, particularly for **removing outliers**.
- The **Box Plot** (or Whisker Plot) is the primary visualization tool used to display the five-number summary and identify outliers graphically.
- **Interquartile Range (IQR)** is calculated as $Q3 - Q1$ and is used to determine the lower and upper fences for outlier detection.

Exam Notes

Visualizing Outliers

Question: What kind of plot should you use to visualize outliers in a dataset?

Answer: You should definitely answer **Box Plot**. It is explicitly designed to show the distribution of data and highlight points that fall outside the expected range (outliers) using "whiskers" or fences.



Shutterstock

The Five Number Summary

The Five Number Summary consists of the following elements, which divide the dataset into four equal parts:

1. **Minimum**: The lowest value in the dataset (excluding outliers).
2. **First Quartile (Q1)**: The 25th percentile.
3. **Median**: The 50th percentile (the middle value).
4. **Third Quartile (Q3)**: The 75th percentile.
5. **Maximum**: The highest value in the dataset (excluding outliers).

Calculating Outliers: Step-by-Step Example

To understand how to use the five-number summary to find outliers, consider the following dataset:

Dataset: `{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27}`

- **Total Count (n)**: 19 elements.
- **Sorted**: The data is already sorted.

Step 1: Calculate Quartiles

Using the percentile formula: $\text{Value Index} = \frac{\text{Percentile}}{100} \times (n+1)$

1. First Quartile (Q1 - 25th Percentile)

- $\text{Index} = \frac{25}{100} \times (19+1) = 0.25 \times 20 = 5$
- The 5th value in the sorted list is **3**.
- $Q1 = 3$

2. Third Quartile (Q3 - 75th Percentile)

- $\text{Index} = \frac{75}{100} \times (19+1) = 0.75 \times 20 = 15$

- The 15th value in the sorted list is **7**.
- $Q3 = 7$

Step 2: Calculate Interquartile Range (IQR)

The IQR represents the spread of the middle 50% of the data.

$$IQR = Q3 - Q1$$

$$IQR = 7 - 3 = 4$$

Step 3: Determine Fences (Boundaries)

To identify outliers, we calculate "fences." Any data point lying outside these fences is considered an outlier.

Lower Fence

- $\text{Formula} = Q1 - 1.5 \times (IQR)$
- $\text{Calculation} = 3 - 1.5(4)$
- $\text{Calculation} = 3 - 6 = -3$
- **Lower Fence = -3**

Higher Fence

- $\text{Formula} = Q3 + 1.5 \times (IQR)$
- $\text{Calculation} = 7 + 1.5(4)$
- $\text{Calculation} = 7 + 6 = 13$
- **Higher Fence = 13**

Step 4: Identify Outliers

We now check the dataset against the range $[-3, 13]$.

- The dataset contains the value **27**.
- Since $27 > 13$, the value **27 is an outlier**.

Box Plot (Whisker Plot)

The **Box Plot** visually encapsulates this information.

- **Box**: Represents the data between Q1 and Q3 (the IQR).
- **Line inside Box**: Represents the **Median**.
- **Whiskers**: Lines extending from the box to the **Minimum** and **Maximum** values that are *within* the calculated fences.
- **Points outside Whiskers**: Individual dots representing **outliers** (like the value 27 in the example above).

This visualization allows for immediate identification of data symmetry, skewness, and anomalies.

94. Correlation And Covariance

Covariance and Correlation

Commands

- No commands were used in this lesson.

Summary

- **Covariance** and **Correlation** are statistical measures used to quantify the relationship between two variables.
- **Covariance** indicates the direction of the linear relationship between variables (positive or negative) but does not have a standardized limit.
- **Correlation** is a standardized measure that limits the values between **-1** and **+1**, allowing for easier comparison of relationship strength.
- **Pearson Correlation Coefficient** is used for linear relationships.
- **Spearman Rank Correlation** is used for non-linear, monotonic relationships by utilizing the **rank** of the data points.
- These concepts are crucial in **Feature Selection** for data science, helping to identify which features significantly impact the target variable.

Exam Notes

Covariance of a Variable with Itself

Question: What is the covariance of a variable X with itself ($Cov(X, X)$)? **Answer**: The covariance of a variable with itself is equal to its **Variance** ($Var(X)$). This is mathematically derived from the formula, where the term $(x_i - \bar{x})(y_i - \bar{y})$ becomes $(x_i - \bar{x})^2$ when $Y = X$.

Disadvantage of Covariance

Question: What is the major disadvantage of using Covariance? **Answer**: Covariance does not have a specific limit value; it can range from $-\infty$ to $+\infty$. This makes it difficult to compare the strength of relationships across different pairs of variables because the magnitude depends on the scale of the variables.

Pearson vs. Spearman

Question: When should you use Spearman Rank Correlation over Pearson Correlation? **Answer**: You should use **Spearman Rank Correlation** when the relationship between variables is **non-linear** (monotonic). Pearson Correlation only captures linear relationships accurately. For non-linear data, Pearson might underestimate the strength of the relationship (e.g., giving 0.88 instead of 1), whereas Spearman will correctly identify a perfect monotonic relationship as 1.

Covariance

Covariance is a measure that quantifies the relationship between two random variables, such as X and Y (e.g., features in a dataset).

Definition

- If variables increase and decrease together, the covariance is **positive**.
- If one variable increases while the other decreases, the covariance is **negative**.

Formula

The formula for sample covariance is:

$$Cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

- x_i, y_i: Individual data points.
- $\bar{x}, \bar{y}$: Sample means of X and Y.
- n: Sample size.

Example Calculation

Consider data regarding hours studied (X) and exam scores (Y):

$$Cov(x,y) = \frac{(2-4)(50-70) + (3-4)(60-70) + (4-4)(70-70) + (5-4)(80-70) + (6-4)(90-70)}{4}$$

$$Cov(x,y) = 20$$

- **X**: {2, 3, 4, 5, 6} -> Mean $\bar{x} = 4$
- **Y**: {50, 60, 70, 80, 90} -> Mean $\bar{y} = 70$

Calculating the differences and products results in a covariance of **20**. The positive value indicates that as hours studied increase, exam scores also increase.

Advantages and Disadvantages

- **Advantage**: It quantifies the direction of the relationship between variables.
- **Disadvantage**: It lacks a standardized scale. A covariance of 300 is not necessarily "stronger" than a covariance of 20 if the scales differ. It provides no specific limit for comparison.

Correlation

Correlation solves the limitation of covariance by restricting the values to a specific range, typically **-1 to +1**.

1. Pearson Correlation Coefficient

This coefficient limits the covariance values by dividing by the product of the standard deviations of the variables.

Formula:

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y}$$

- **Range**: -1 to +1.
- **Interpretation**:
  - **+1**: Perfect positive linear correlation.
  - **-1**: Perfect negative linear correlation.
  - **0**: No linear correlation.
- **Limitation**: It only captures **linear** relationships properly. For non-linear data (e.g., exponential growth), it may fail to represent the true strength of the relationship.

2. Spearman Rank Correlation

Spearman correlation is used when the data follows a monotonic relationship but is not necessarily linear. It replaces the raw data values with their **ranks** before calculating the correlation.

Formula:

$$r_s = \frac{Cov(Rank(x), Rank(y))}{\sigma_{Rank(x)} \cdot \sigma_{Rank(y)}}$$

Key Difference:

- For non-linear data where X increases and Y increases (but not at a constant rate), **Pearson** might give a value like **0.88**.
- **Spearman** will give a value of **1.0** because the *ranks* perfectly align, capturing the monotonic nature of the relationship.

Practical Application: Feature Selection

In Data Science, correlation is vital for **Feature Selection**—deciding which variables to keep in a model.

Example: Housing Prices Dataset

- **Size of House vs. Price**: Highly positive correlation. As size increases, price increases. **Keep this feature**.
- **Haunted Status vs. Price**: Negative correlation. If a house is haunted, price decreases. **Keep this feature**.

- **Number of People Staying vs. Price**: Likely zero correlation. The number of occupants does not dictate the market value of a house. **Drop this feature** as it adds no predictive value.