Statistics Lecture 2

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

Table of Contents

Why Sample Variance is Divided by $n-1$n − 1

Introduction

We are going to discuss a very important interview question: **Why is sample variance divided by $n-1$n − 1?** This question is commonly asked in interviews and is crucial for understanding statistical concepts related to population and sample data.

Population vs. Sample

To begin, let's differentiate between population and sample:

- **Population**: The entire group that you want to draw conclusions about. It is usually denoted by NN.
- **Sample**: A subset of the population used to represent the population. It is denoted by nn.

Mean Calculation

Population Mean

The mean of a population can be calculated using the formula:

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

μ = N∑i=1N xi

Sample Mean

The mean of a sample, denoted by $\bar{x}$x¯, is calculated as:

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

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

Variance Calculation

Population Variance

The variance of a population, denoted by σ^2σ2, is given by:

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

σ2 = N∑i=1N (xi − μ)2

Sample Variance

The variance of a sample, denoted by s^2s2, is calculated as:

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

s2 = n − 1∑i=1n(xi − x¯)2

Why Divide by $n-1$n − 1?

The question arises: **Why is the sample variance divided by $n-1$n − 1 instead of nn?** This is a fundamental concept in statistics, often discussed in terms of *unbiased estimation*.

Understanding the Concept

1. **Bias in Estimation**: When we calculate sample statistics, we aim to estimate population parameters. However, directly using nn in the denominator tends to underestimate the population variance. This underestimation occurs because the sample mean $\bar{x}$x¯ is used as an estimate of the population mean μμ.

2. **Degrees of Freedom**: Dividing by $n-1$n − 1 instead of nn corrects this bias. The term $n-1$n − 1 is known as the *degrees of freedom*. It accounts for the fact that one degree of freedom is lost because the sample mean $\bar{x}$x¯ is used to compute each deviation $(x_i-\bar{x})$(xi − x¯).

Example to Illustrate

Let's consider a simple example to understand this better:

Population Data: Suppose we have the ages of individuals in a population:

$$\text{Population: } 20, 22, 24, 26, 28$$

Population: 20, 22, 24, 26, 28

Sample Data: We randomly select a sample of 3 individuals:

$$\text{Sample: } 20, 24, 28$$

Sample: 20, 24, 28

Population Mean and Variance

1. **Population Mean μμ**:

$$\mu = \frac{20+22+24+26+28}{5} = 24$$

μ = 520 + 22 + 24 + 26 + 28 = 24

2. **Population Variance σ^2σ2**:

$$\sigma^2 = \frac{(20-24)^2+(22-24)^2+(24-24)^2+(26-24)^2+(28-24)^2}{5} = 8$$

σ2 = 5(20 − 24)2 + (22 − 24)2 + (24 − 24)2 + (26 − 24)2 + (28 − 24)2 = 8

Sample Mean and Variance

1. **Sample Mean $\bar{x}$x¯**:

$$\bar{x} = \frac{20 + 24 + 28}{3} = 24$$

x¯ = 320 + 24 + 28 = 24

2. **Sample Variance s^2s2**:

$$s^2 = \frac{(20 - 24)^2 + (24 - 24)^2 + (28 - 24)^2}{2} = 16$$

s2 = 2(20 − 24)2 + (24 − 24)2 + (28 − 24)2 = 16

If we had divided by nn (which is 3), the variance would have been:

$$s^2 = \frac{(20 - 24)^2 + (24 - 24)^2 + (28 - 24)^2}{3} = \frac{16 + 0 + 16}{3} = 10.67$$

s2 = 3(20 − 24)2 + (24 − 24)2 + (28 − 24)2 = 316 + 0 + 16 = 10.67

This clearly underestimates the population variance.

Experimental Validation

Researchers have validated this concept through numerous experiments. They calculated sample variances by dividing by $n - 1$n − 1, $n - 2$n − 2, and so on. They found that dividing by $n - 1$n − 1 provides the best unbiased estimate of the population variance.

To summarize,

Dividing the sum of squared deviations by $n - 1$n − 1 instead of nn corrects the bias and provides an unbiased estimate of the population variance. This is crucial for accurate statistical analysis and is a fundamental concept in inferential statistics.

Key Takeaways

- **Population vs. Sample**: Understand the difference and the notations.
- **Mean and Variance Calculations**: Know the formulas for both population and sample.
- **Unbiased Estimation**: Learn why dividing by $n - 1$n − 1 is important for unbiased estimation of population variance.

Variables And Their Types

Introduction to Variables

Variables are fundamental building blocks in statistics and data analysis. Understanding variables allows us to effectively organize, analyze and derive insights from data. Let's explore variables in more depth.

Definition of a Variable

A variable is a characteristic or attribute that can take on different values. For example:

- Height: 170 cm, 185 cm, 162 cm, etc.
- Weight: 65 kg, 80 kg, 72 kg, etc.
- Age: 25 years, 40 years, 18 years, etc.
- Income: $50,000,$50,000,75,000, $100,000, etc.

Variables allow us to quantify and compare different observations or measurements.

Types of Variables

There are two main types of variables:

1. Quantitative Variables
2. Qualitative (Categorical) Variables

Quantitative Variables

Quantitative variables represent numerical quantities that can be measured and compared mathematically. They can be further divided into:

- Discrete Variables: These take on distinct, countable values.
  - Examples:
    - Number of children in a family (0, 1, 2, 3, etc.)
    - Number of cars owned (1, 2, 3, etc.)
    - Number of employees in a company
- Continuous Variables: These can take any value within a range.
  - Examples:
    - Height (172.5 cm, 180.3 cm, etc.)
    - Weight (68.7 kg, 75.2 kg, etc.)
    - Time (3.45 seconds, 2.78 hours, etc.)

Real-life Usage:

- In manufacturing, discrete variables like defect counts are used for quality control.
- In healthcare, continuous variables like blood pressure and cholesterol levels are crucial for patient health assessment.

Qualitative (Categorical) Variables

Qualitative variables represent categories or groups. They describe qualities that can't be measured numerically. Examples include:

- Gender: Male, Female, Non-binary
- Blood Type: A, B, AB, O
- Marital Status: Single, Married, Divorced, Widowed
- Education Level: High School, Bachelor's, Master's, PhD

Real-life Usage:

- In market research, categorical variables like customer preferences are used to segment audiences.
- In epidemiology, variables like disease status (infected/not infected) are crucial for studying disease spread.

Variables Measurement Scales

Types of Measurement Scales

Variables can be measured using four different scales, each with increasing levels of measurement precision:

1. Nominal Scale
2. Ordinal Scale
3. Interval Scale
4. Ratio Scale

Nominal Scale

The nominal scale is used for labeling variables without any quantitative value. Categories are mutually exclusive and have no inherent order.

Examples:

- Colors: Red, Blue, Green
- Professions: Teacher, Doctor, Engineer
- Types of Pets: Dog, Cat, Bird

Real-life Usage: In customer surveys, nominal scales are used to categorize responses like preferred brands or shopping locations.

Ordinal Scale

The ordinal scale ranks variables in order, but the intervals between ranks are not necessarily equal.

Examples:

- Education Levels: High School, Bachelor's, Master's, PhD
- Customer Satisfaction: Very Unsatisfied, Unsatisfied, Neutral, Satisfied, Very Satisfied
- Movie Ratings: 1 star, 2 stars, 3 stars, 4 stars, 5 stars

Real-life Usage: In product development, ordinal scales are used to prioritize features based on customer feedback.

Interval Scale

The interval scale has equal intervals between values, but lacks a true zero point.

Examples:

- Temperature in Celsius or Fahrenheit
- Calendar Years
- IQ Scores

Real-life Usage: In climate science, temperature measurements (often in Celsius) use the interval scale to track and compare temperatures over time.

Ratio Scale

The ratio scale has all the properties of the interval scale plus a true zero point.

Examples:

- Height
- Weight
- Income
- Age

Real-life Usage: In finance, ratio scales are used for metrics like Return on Investment (ROI), where a true zero (no return) is meaningful.

Frequency Distribution

Frequency distribution organizes data into categories and shows how often each category occurs.

Example:
Let's say we surveyed 50 people about their favorite color:

Color	Frequency
Red	15
Blue	20
Green	10
Yellow	5

Cumulative Frequency

Cumulative frequency shows the accumulation of frequencies up to each category.

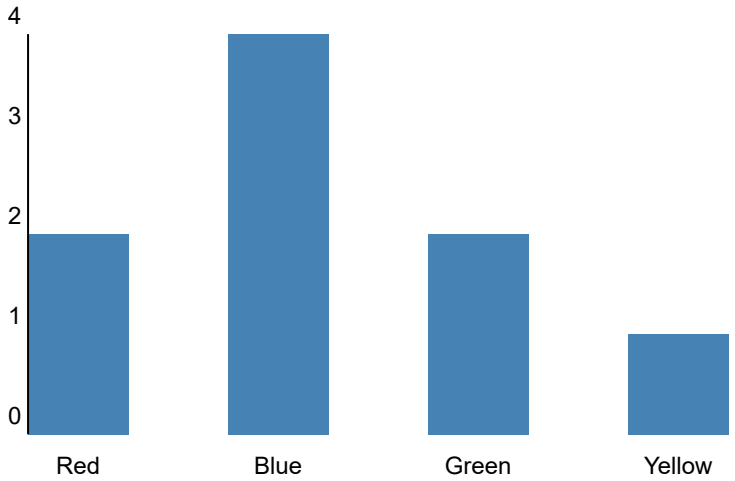
Color	Frequency	Cumulative Frequency
Red	15	15
Blue	20	35
Green	10	45
Yellow	5	50

Real-life Usage: In project management, cumulative frequency is used in burn-down charts to track progress over time.

Bar Graph

Bar graphs visually represent categorical data using rectangular bars. The height of each bar corresponds to the frequency of that category.

Here's an example bar graph for the color preference data:



Real-life Usage: In sales, bar graphs are used to compare revenue across different products or regions.

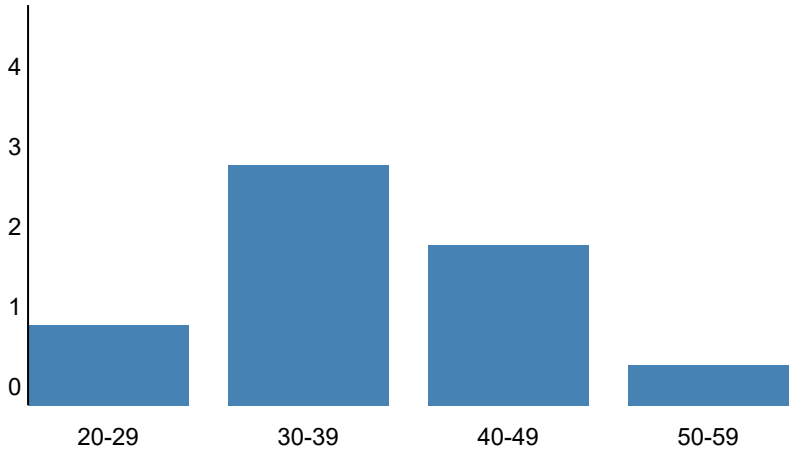
Histograms

Histograms represent the distribution of continuous data. Data is grouped into bins, and the height of each bar represents the frequency of data points in that bin.

Example: Ages of 50 participants in a study:

```
22, 25, 28, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46,
47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66,
67, 68, 69, 70, 71, 72, 73, 74, 75, 76
```

Grouped into bins of 10 years:



Real-life Usage: In finance, histograms are used to visualize the distribution of stock returns, helping investors understand risk and volatility.

Importance for Data Scientists and Analysts

Understanding these fundamental concepts is crucial for aspiring data scientists and analysts for several reasons:

1. **Data Understanding:** These concepts form the foundation for understanding the nature and structure of data. Knowing variable types helps in choosing appropriate analysis methods.
2. **Data Preprocessing:** Understanding measurement scales is crucial for data preprocessing. For example, categorical variables often need to be encoded differently than numerical variables for machine learning algorithms.

3. **Statistical Analysis:** Many statistical tests and models assume certain properties of variables. Understanding these properties helps in selecting the right analytical tools.
4. **Data Visualization:** Knowing when to use bar graphs vs histograms is key to effectively communicating insights through data visualization.
5. **Feature Engineering:** In machine learning, understanding variable types and scales is crucial for feature engineering and selection.
6. **Interpretation of Results:** The ability to interpret frequency distributions and various graphs is essential for deriving meaningful insights from data analysis.
7. **Communication:** These concepts provide a common language for communicating about data with team members and stakeholders.

Left Skewed and Right Skewed Distribution and Relation with Mean, Median, and Mode

Understanding Skewness

Skewness refers to the asymmetry in the distribution of data. A distribution can be either:

1. **Right Skewed (Positively Skewed)**
2. **Left Skewed (Negatively Skewed)**

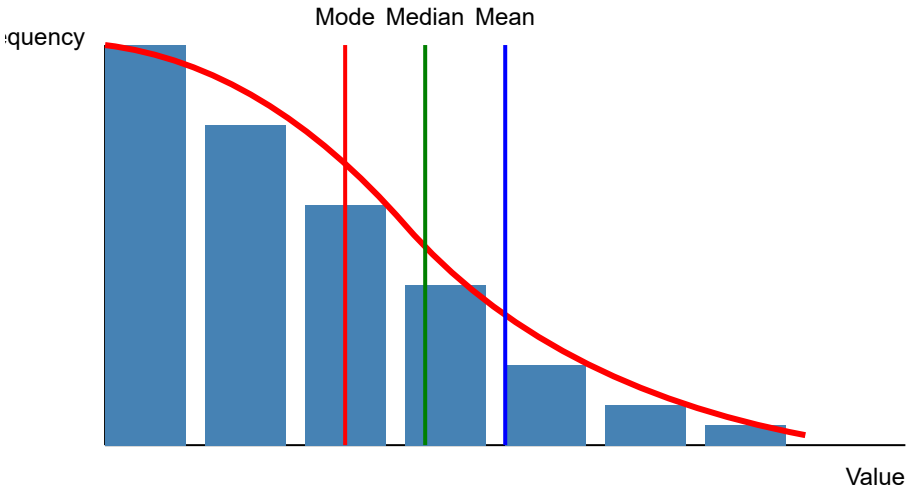
Right Skewed Distribution

A right skewed distribution, also known as a positively skewed distribution, has a long tail on the right side. This means the right side of the distribution is elongated compared to the left side.

Examples:

1. **Wealth Distribution:** A few people have very high wealth, while most have moderate or low wealth.
2. **Length of Comments on a Video:** A few comments are very long, while most are short or moderate in length.

Graph:



Relationship with Mean, Median, and Mode:

- **Mean > Median > Mode**

In a right skewed distribution:

- The mean is greater than the median.
- The median is greater than the mode.

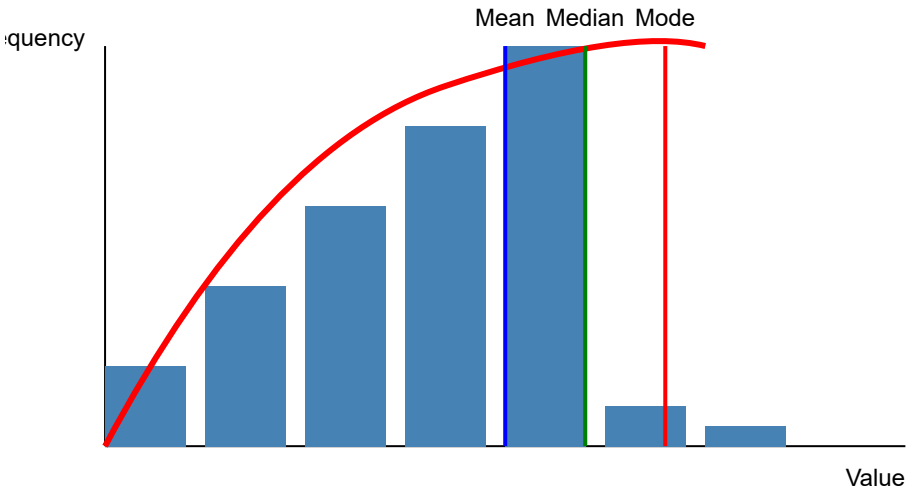
Left Skewed Distribution

A left skewed distribution, also known as a negatively skewed distribution, has a long tail on the left side. This means the left side of the distribution is elongated compared to the right side.

Examples:

1. **Lifespan of Human Beings:** Most people live around the average lifespan, but a few live significantly shorter lives.
2. **Test Scores:** In some cases, a few students may score very low, while most score around the average or high.

Graph:



Relationship with Mean, Median, and Mode:

- **Mean < Median < Mode**

In a left skewed distribution:

- The mean is less than the median.
- The median is less than the mode.

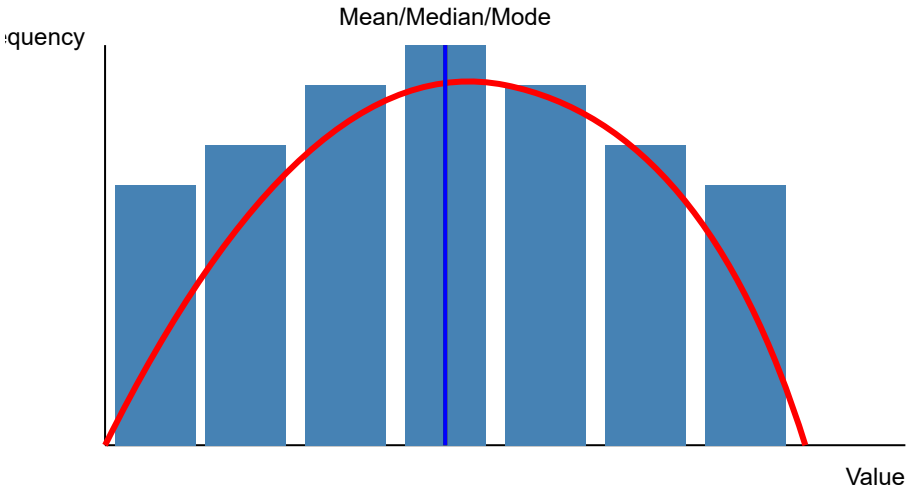
Symmetrical Distribution

A symmetrical distribution, also known as a normal distribution, has both sides of the distribution mirror each other. The data is evenly distributed around the center.

Examples:

1. **Height Distribution:** Most people's heights are around the average, with fewer people being very short or very tall.
2. **IQ Scores:** Most people have average IQ scores, with fewer people having very low or very high scores.

Graph:



Relationship with Mean, Median, and Mode:

- **Mean ≈ Median ≈ Mode**

In a symmetrical distribution:

- The mean, median, and mode are approximately equal.

To summarize:

- **Right Skewed Distribution**: Mean > Median > Mode
- **Left Skewed Distribution**: Mean < Median < Mode
- **Symmetrical Distribution**: Mean ≈ Median ≈ Mode

These concepts are fundamental in statistics and are essential for interpreting data distributions accurately. Keep practicing with real-life examples to solidify your understanding and be well-prepared for any data analysis tasks or interviews.

Certainly! I'll create a detailed lecture on "5 Number Summary and How to Handle Outliers Using IQR-Statistics" in Markdown format, based on the provided transcript. I'll also include references to the attached picture where appropriate.

5 Number Summary and How to Handle Outliers Using IQR-Statistics

The Five Number Summary

The Five Number Summary is a set of descriptive statistics that provides a quick overview of a dataset's distribution. It consists of five key values:

1. Minimum value
2. First quartile (Q1) or 25th percentile
3. Median
4. Third quartile (Q3) or 75th percentile
5. Maximum value

These values are crucial for understanding the spread and central tendency of your data.

Understanding Percentiles

Before we calculate the Five Number Summary, let's understand what percentiles are.

A percentile is a measure that indicates the value below which a given percentage of observations falls. For example, the 25th percentile (Q1) is the value below which 25% of the observations in a dataset are found.

Calculating Percentiles

To calculate the nth percentile:

1. Sort the data in ascending order.
2. Use the formula: `index = (n/100) * (N + 1)`
   Where N is the number of values in the dataset.
3. If the index is not a whole number, interpolate between the two nearest values.

Calculating the Five Number Summary

Let's use this dataset as an example:

```
3, 4, 5, 5, 6, 7, 8, 9, 9, 10
```

1. **Minimum**: 3
2. **Q1 (25th percentile)**:
   - Index = (25/100) * (10 + 1) = 2.75
   - Value = 5
3. **Median**: 6.5 (average of 6 and 7)
4. **Q3 (75th percentile)**:
   - Index = (75/100) * (10 + 1) = 8.25
   - Value = 9
5. **Maximum**: 10

Interquartile Range (IQR)

The IQR is a measure of statistical dispersion and is calculated as:

```
IQR = Q3 - Q1
```

In our example:

```
IQR = 9 - 5 = 4
```

Handling Outliers Using IQR

Outliers are data points that significantly differ from other observations. The IQR method is commonly used to identify and handle outliers.

Steps to Identify Outliers:

1. Calculate Q1, Q3, and IQR
2. Define the lower and upper bounds:
    - Lower bound = Q1 - 1.5 * IQR
    - Upper bound = Q3 + 1.5 * IQR
3. Any data point below the lower bound or above the upper bound is considered an outlier.

Using our example:

- Lower bound = 5 - 1.5 * 4 = -1
- Upper bound = 9 + 1.5 * 4 = 15

If we had a value of 27 in our dataset, it would be considered an outlier as it's above the upper bound.

Visualizing with a Box Plot

A box plot is an excellent way to visualize the Five Number Summary and identify outliers.

 alt text

In a box plot:

- The box represents the IQR
- The line inside the box is the median
- The whiskers extend to the minimum and maximum values (excluding outliers)
- Outliers are plotted as individual points beyond the whiskers

To summarize,

Understanding the Five Number Summary and how to handle outliers using IQR is crucial for:

1. Getting a quick overview of your data's distribution
2. Identifying potential data quality issues
3. Making informed decisions about data preprocessing

Practical Implementation

Python Code for Five Number Summary and Outlier Detection

Let's demonstrate how to perform the five number summary and outlier detection analysis using Python. Here's the code along with explanations:

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# Generate fictional data
np.random.seed(42)
data = np.concatenate([
    np.random.normal(100, 10, 95),  # 95 normal distributed values
    np.random.uniform(150, 200, 5)  # 5 potential outliers
])

# Create a DataFrame
df = pd.DataFrame(data, columns=['Value'])

# Calculate Five Number Summary
five_num_summary = df['Value'].describe()

print("Five Number Summary:")
print(five_num_summary)

# Calculate IQR
Q1 = df['Value'].quantile(0.25)
Q3 = df['Value'].quantile(0.75)
IQR = Q3 - Q1

# Define bounds for outliers
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Identify outliers
outliers = df[(df['Value'] < lower_bound) | (df['Value'] > upper_bound)]
```

```
print("\nNumber of outliers:", len(outliers))
print("Outliers:")
print(outliers)

# Create a box plot
plt.figure(figsize=(10, 6))
df['Value'].plot(kind='box')
plt.title('Box Plot of Values')
plt.ylabel('Value')
plt.show()

# Histogram with outlier boundaries
plt.figure(figsize=(10, 6))
df['Value'].hist(bins=30, edgecolor='black')
plt.axvline(lower_bound, color='r', linestyle='dashed', linewidth=2)
plt.axvline(upper_bound, color='r', linestyle='dashed', linewidth=2)
plt.title('Histogram of Values with Outlier Boundaries')
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.show()
```

This script does the following:

1. Generates 100 data points: 95 normally distributed around 100, and 5 potential outliers between 150 and 200.
2. Creates a pandas DataFrame with the data.
3. Calculates and prints the Five Number Summary using `describe()`.
4. Calculates the IQR and determines the bounds for outliers.
5. Identifies and prints the outliers.
6. Creates a box plot to visualize the distribution and outliers.

7. Creates a histogram with vertical lines indicating the outlier boundaries.

Running this script will provide you with the Five Number Summary, identified outliers, a box plot, and a histogram. This gives a comprehensive view of the data distribution and outliers, demonstrating the concepts discussed in the lecture.

R Code for Five Number Summary and Outlier Detection

Similarly, an equivalent R script that performs the same five number summary and outlier detection analysis on a fictional dataset:

```r
library(ggplot2)

# Generate fictional data
set.seed(42)
normal_data <- rnorm(95, mean = 100, sd = 10)
outlier_data <- runif(5, min = 150, max = 200)
data <- c(normal_data, outlier_data)

# Create a data frame
df <- data.frame(Value = data)

# Calculate Five Number Summary
five_num_summary <- summary(df$Value)
print("Five Number Summary:")
print(five_num_summary)

# Calculate IQR
Q1 <- quantile(df$Value, 0.25)
Q3 <- quantile(df$Value, 0.75)
IQR <- Q3 - Q1

# Define bounds for outliers
lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR

# Identify outliers
outliers <- df[df$Value < lower_bound | df$Value > upper_bound, ]

print(paste("\nNumber of outliers:", nrow(outliers)))
print("Outliers:")
print(outliers)
```

```r
# Create a box plot
boxplot <- ggplot(df, aes(y = Value)) +
  geom_boxplot() +
  ggtitle("Box Plot of Values") +
  ylab("Value")
print(boxplot)

# Histogram with outlier boundaries
histogram <- ggplot(df, aes(x = Value)) +
  geom_histogram(binwidth = 5, fill = "lightblue", color = "black") +
  geom_vline(xintercept = lower_bound, color = "red", linetype = "dashed", size = 1) +
  geom_vline(xintercept = upper_bound, color = "red", linetype = "dashed", size = 1) +
  ggtitle("Histogram of Values with Outlier Boundaries") +
  xlab("Value") +
  ylab("Frequency")
print(histogram)
```

This R script performs the same steps as the Python version:

1. Generates 100 data points: 95 normally distributed around 100, and 5 potential outliers between 150 and 200.
2. Creates a data frame with the generated data.
3. Calculates and prints the Five Number Summary using `summary()`.
4. Calculates the IQR and determines the bounds for outliers.
5. Identifies and prints the outliers.
6. Creates a box plot using ggplot2 to visualize the distribution and outliers.
7. Creates a histogram with vertical lines indicating the outlier boundaries, also using ggplot2.

To run this script, make sure you have the ggplot2 package installed. You can install it using `install.packages("ggplot2")` if you haven't already.

This R code provides the same comprehensive view of the data distribution and outliers as the Python version, allowing you to visualize and understand the concepts discussed in the lecture.

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