4COSC002W Mathematics for Computing

Lecture 9

Elementary Statistics

UNIVERSITY OF
WESTMINSTER

What is Statistical Analysis?

A Tool for Insight: Utilising methods to collect, analyse, and interpret data.

Purpose: To uncover patterns, trends, and insights in data.

Why It Matters in Computing?

- **Informs Decisions:** Guides data-driven decisions in software development and system design.
- **Enhances Performance:** Crucial for evaluating and improving technology performance and user experience.

Software Tools: R, Python, MATLAB, Excel.

“Good decisions come from good data analysis.”

Sample Dataset: Bugs Found in Software Projects

Project	Bugs Found
A	12
B	7
C	9
D	15
E	5
F	8
G	11

*check the Excel file on Blackboard

Basic Statistical Analysis for the Sample Dataset

Mean	9.250
Median	8.500
Mode	7
Standard Deviation	3.240
Variance	10
Interquartile Range (IQR)	4.750

Mean (Arithmetic Average)

The mean is the average value of a dataset, representing the central point of the data.

General Formula: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, where x_i are the values in the dataset and n is the number of values.

Excel Formula: =AVERAGE(B2:B9)

Application: In computing, the mean is used to calculate average performance metrics, such as average response time or average number of bugs in software projects.

Example: Sample Dataset Calculations

Mean (Arithmetic Average)

Sum of all values: $12 + 7 + 9 + 15 + 7 + 5 + 8 + 11 = 74$

Number of values: 8

Mean: $\frac{74}{8} = 9.25$

Result: Mean = 9.25

Median

The median is the middle value in a sorted list of numbers, effectively separating the dataset into two halves.

General Approach: If the number of observations is odd, the median is the middle number. If even, it's the average of the two middle numbers.

Excel Formula: =MEDIAN(B2:B9)

Application: The median is useful in computing for analyzing skewed data, such as median load times or median server response times.

Example: Sample Dataset Calculations

Median

Sorted Data: [5, 7, 7, 8, 9, 11, 12, 15]

Number of values: 8 (even)

Median is average of 4th and 5th values: $\frac{8+9}{2} = 8.5$

Result: Median = 8.5

Mode

The mode is the most frequently occurring value in the dataset. It can be used for both numerical and categorical data.

General Approach: Identify the value that appears most frequently.

Excel Formula: =MODE.SNGL(B2:B9)

Application: In computing, the mode can help identify the most common error types, the most used software feature, or the most common number of daily user logins.

Example: Sample Dataset Calculations

Mode

Frequency of each value: 5 (once), 7 (twice), 8 (once), 9 (once), 11 (once), 12 (once), 15 (once)

Most frequent value: 7

Result: Mode = 7

Range

The range is the difference between the highest and lowest values in a dataset.

General Approach: Range = Maximum value - Minimum value.

Excel Formula: =MAX(B2:B8) - MIN(B2:B8)

Application: Range is used in computing to understand the spread of data, such as the range of response times or the range in the number of daily active users.

Standard Deviation

Standard deviation quantifies the amount of variation or dispersion of a set of data values from the mean.

General Formula (for a sample): $s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$.

Excel Formula: =STDEV.S(B2:B9) for a sample

Application: Standard deviation is used in computing to measure variability in system performance, such as the consistency of response times or variability in daily website traffic.

Example: Sample Dataset Calculations

Standard Deviation

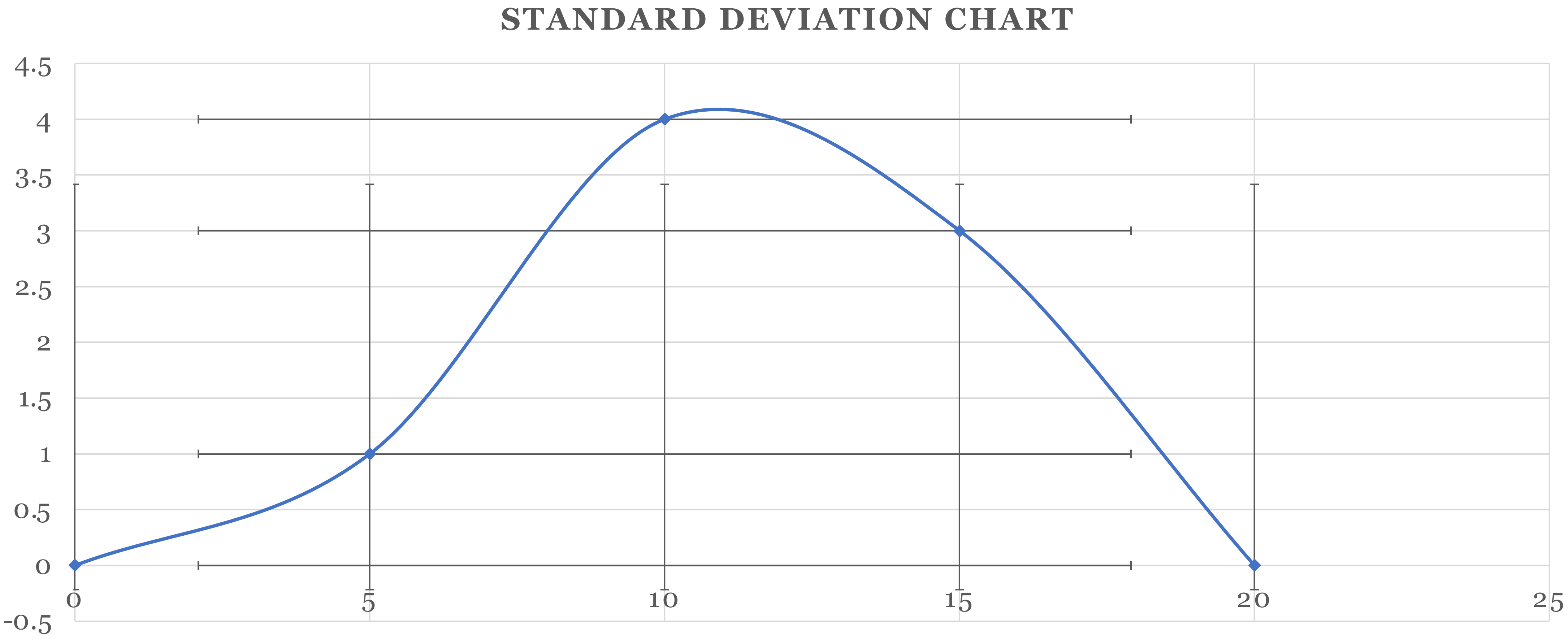
$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}.$$

Sum of squared differences from the mean: $(12 - 9.25)^2 + (7 - 9.25)^2 + \cdots + (11 - 9.25)^2$

Calculating and then taking the square root.

Result: Standard Deviation ≈ 3.24

Example: Sample Dataset Plot



Variance

Variance measures how much each number in the set is different from the mean. It's the square of the standard deviation.

General Formula (for a sample): $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$.

Excel Formula: =VAR.S(B2:B9) for a sample

Application: Variance is important in computing for identifying the consistency of data, like the consistency in the number of bugs found across different projects.

Example: Sample Dataset Calculations

Variance

Variance is the square of the standard deviation.

$\text{Variance} = (\text{Standard Deviation})^2$

Result: Variance = 10.5

Interquartile Range (IQR)

IQR is the range between the first quartile (25th percentile) and the third quartile (75th percentile). It shows the middle 50% of the data.

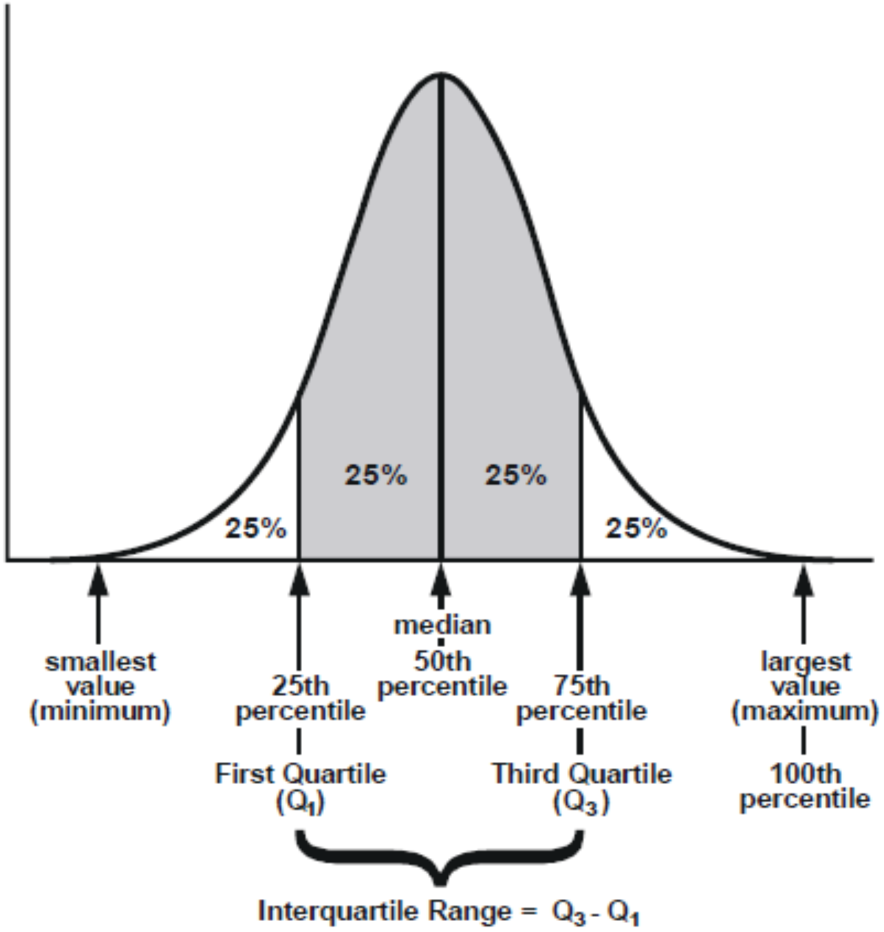
Percentiles are measures indicating the value below which a given percentage of observations fall.

General Approach: IQR = Q3 - Q1, where Q3 and Q1 are the third and first quartiles, respectively.

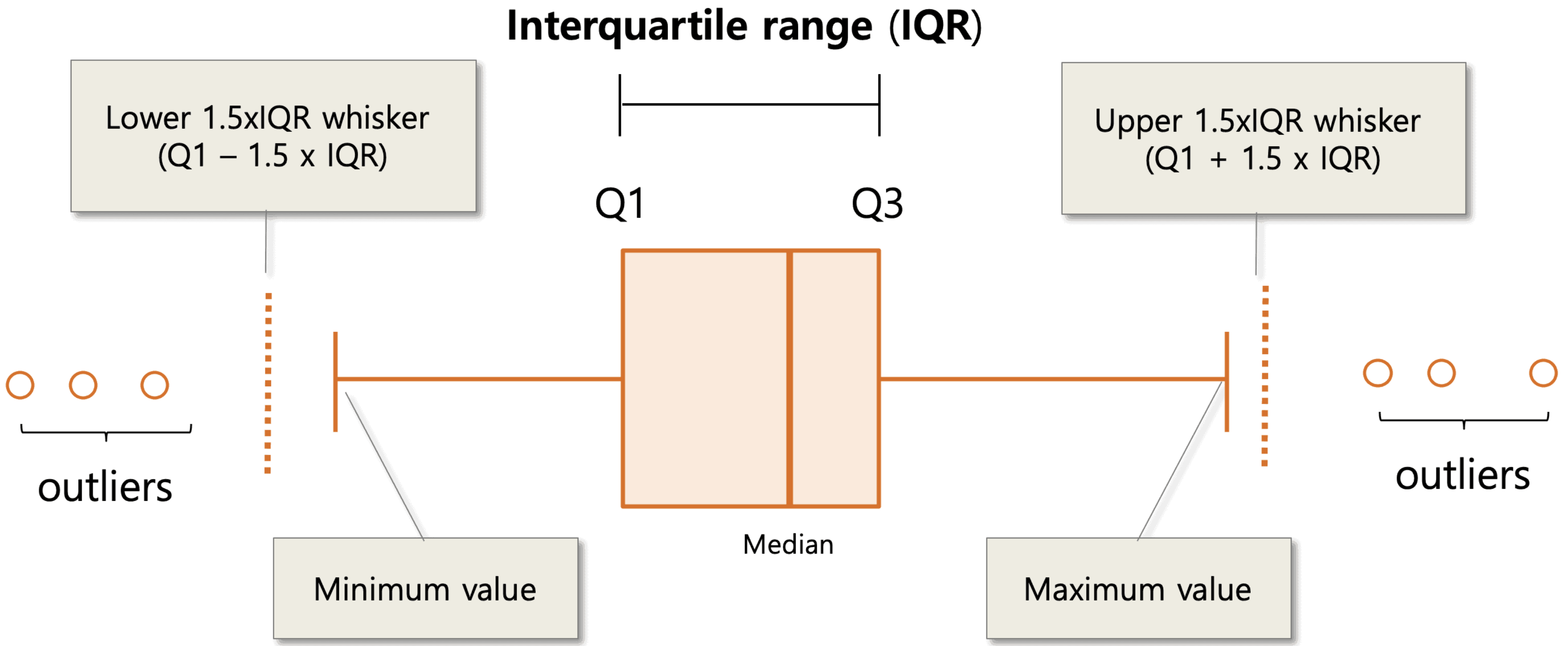
Excel Formula: =QUARTILE.EXC(B2:B9, 3) - QUARTILE.EXC(B2:B9, 1)

Application: In computing, IQR is used to measure the spread of the middle half of the data, which is less affected by outliers. It can be used to analyse time spent on different tasks in project management or user interaction times.

IQR vs Normal Distribution



IQR Boxplot



Interquartile range (IQR)

Lower 1.5xIQR whisker
(Q1 – 1.5 x IQR)

Q1

Q3

Upper 1.5xIQR whisker
(Q1 + 1.5 x IQR)

outliers

outliers

Median

Minimum value

Maximum value

Example: Sample Dataset Calculations

IQR Calculation

Step 1. Determine the Positions for Q1 and Q3:

Total Number of Observations (n): 8

Position of Q1 (25th percentile): $\frac{25}{100}\times(n+1)=\frac{25}{100}\times(8+1)=2.25$

Position of Q3 (75th percentile): $\frac{75}{100}\times(n+1)=\frac{75}{100}\times(8+1)=6.75$

Example: Sample Dataset Calculations

Step 2. Interpolate to Find Q1 and Q3:

Interpolation for Q1: Since 2.25 is not a whole number, interpolate between the 2nd and 3rd values in the sorted list. Values: 7 (2nd) and 7 (3rd)

Interpolation: $7 + (7 - 7) \times (2.25 - 2) = 7$

Interpolation for Q3: Since 6.75 is not a whole number, interpolate between the 6th and 7th values in the sorted list. Values: 11 (6th) and 12 (7th)

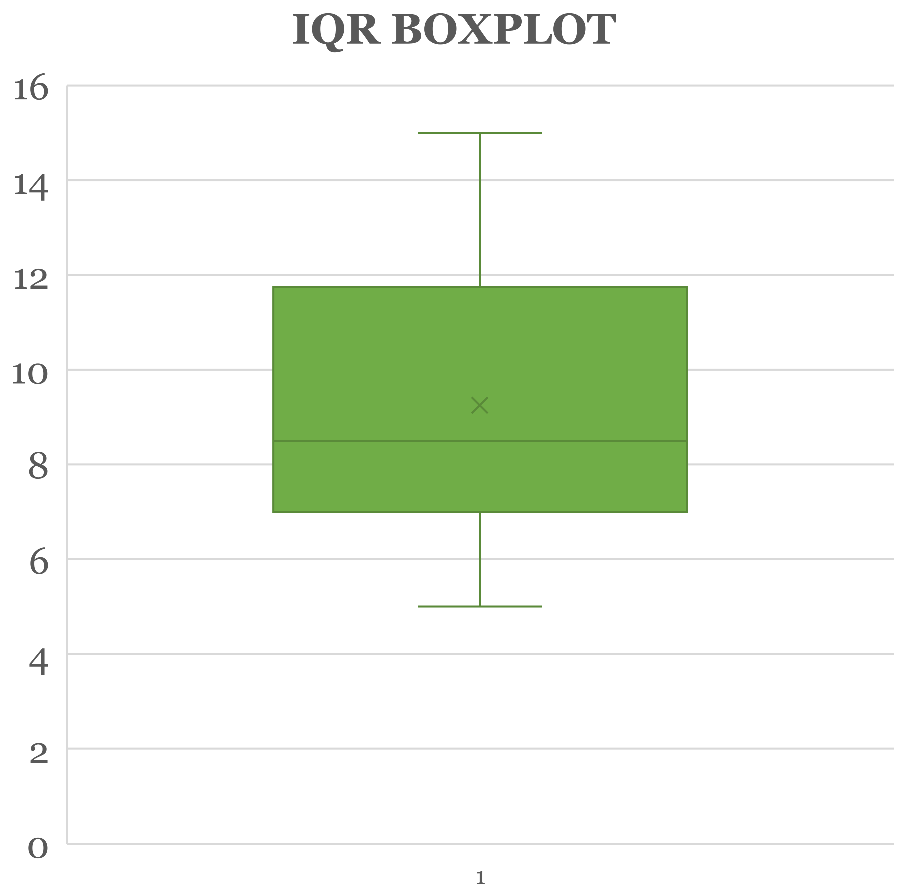
Interpolation: $11 + (12 - 11) \times (6.75 - 6) = 11.75$

Example: Sample Dataset Calculations

Step 3. Calculate the IQR:

$$\text{IQR} = Q3 - Q1 = 11.75 - 7 = 4.75$$

Example: Sample Dataset IQR Boxplot Plot


IQR BOXPLOT