Five Number Summary and Outlier Detection

Statistics Course Notes

1 Introduction

The **five number summary** is a statistical tool that describes data distribution using five key values:

1. Minimum (Min)
2. First Quartile (Q1, 25th percentile)
3. Median (Q2)
4. Third Quartile (Q3, 75th percentile)
5. Maximum (Max)

It helps identify **outliers** and is used in machine learning for feature engineering and data cleaning.

2 Key Concepts

2.1 Quartiles

- **Q1 (25th percentile):** Value below which 25% of data lies.
- **Q3 (75th percentile):** Value below which 75% of data lies.

2.2 Interquartile Range (IQR)

$$\text{IQR} = Q3 - Q1$$

IQR measures the spread of the middle 50% of data.

2.3 Outlier Boundaries

Outliers lie outside these ”fences”:

$$\text{Lower Fence} = Q1 - 1.5 \times \text{IQR}$$
$$\text{Upper Fence} = Q3 + 1.5 \times \text{IQR}$$

3 Example: Identifying Outliers

Given dataset:

$$\{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27\}$$

Step 1: Calculate five number summary

Statistic	Value
Min	1
Q1 (25th percentile)	3
Median	5
Q3 (75th percentile)	7
Max	27

Calculations:

- **Q1**: Position $= \frac{25}{100} \times (19+1) = 5^{th}$ element $= 3$
- **Median**: Average of 9^{th} and 10^{th} elements $= \frac{5+5}{2} = 5$
- **Q3**: Position $= \frac{75}{100} \times 20 = 15^{th}$ element $= 7$

Step 2: Compute IQR

$$\text{IQR} = Q3 - Q1 = 7 - 3 = 4$$

Step 3: Determine outlier fences

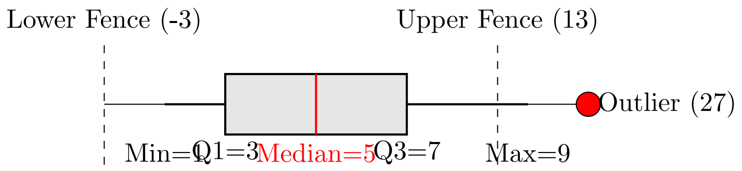
$$\text{Lower Fence} = 3 - 1.5 \times 4 = -3$$
$$\text{Upper Fence} = 7 + 1.5 \times 4 = 13$$

Conclusion: $27 > 13 \implies 27$ is an outlier.

4 Box Plot Visualization

After removing the outlier (27), the five number summary is:

Min	1
Q1	3
Median	5
Q3	7
Max	9



5 Key Takeaways

- Five number summary describes **center, spread, and skewness** of data.
- IQR-based outlier detection is robust for skewed distributions.
- Box plots visually summarize:
  - Central 50% of data (the box)
  - Median (red line)
  - Potential outliers (points beyond whiskers)
- Applications: Feature scaling, anomaly detection, data preprocessing.