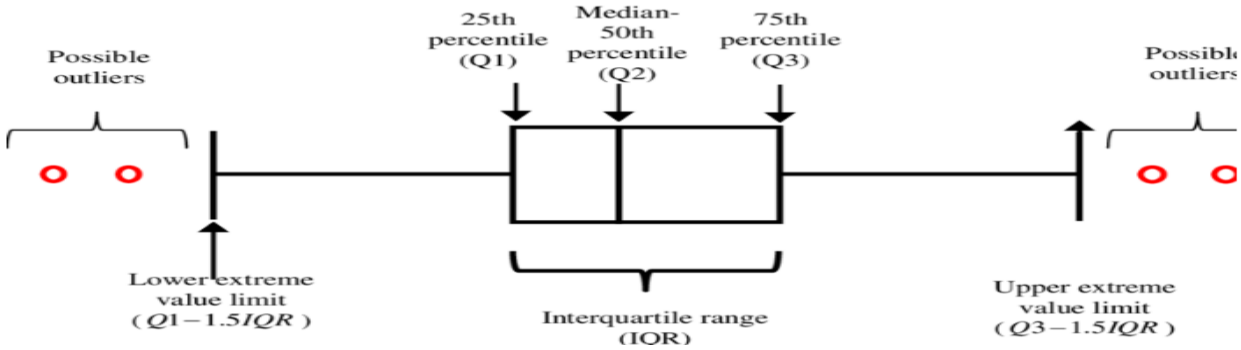
Chapter 9: Outlier Detection


Possible outliers
25th percentile (Q1)
Median-50th percentile (Q2)
75th percentile (Q3)
Possibl outliers
Lower extreme value limit ($Q1-1.5IQR$)
Interquartile range (IQR)
Upper extreme value limit ($Q3-1.5IQR$)

Outliers are data points that differ significantly from the majority of the data. Detecting and handling outliers is crucial as they can skew results, mislead analysis, and affect the performance of statistical models or machine learning algorithms.

Characteristics of Outliers

- Extremely high or low values compared to other data points.
- Points that fall outside the overall trend or pattern.
- May indicate data entry errors, variability in measurements, or rare events.

Importance of Outlier Detection

1. **Improves Data Quality**: Identifies errors or anomalies in data.
2. **Enhances Model Accuracy**: Prevents skewing of results in statistical or machine learning models.
3. **Supports Better Insights**: Differentiates between noise and meaningful deviations.

Methods for Outlier Detection

1. Statistical Methods

A. Z-Score Method

- Measures how many standard deviations a data point is from the mean.

  Formula: $Z=(X-\mu)/\sigma$

Where :
X = data point, μ = population mean, σ = standard deviation.

- Threshold: If $|Z|>3$, the point is considered an outlier.

B. Interquartile Range (IQR) Method

- Identifies outliers based on the spread of the middle 50% of the data.
- Formula: $IQR=Q3-Q1$
- Outlier thresholds:
  - Lower bound=$(Q1-1.5)\times IQR$,
  - Upper bound=$Q3+1.5\times IQR$
  - Points outside these bounds are considered outliers.

C. Grubbs' Test

- Tests whether the extreme value in a dataset is an outlier.
- Commonly used for small datasets.

D. Dixon's Q Test

- A statistical test for identifying a single outlier in small datasets.

Handling Outliers

1. **Investigate the Cause**:
   - Determine whether the outlier is due to an error or is a legitimate data point.
2. **Transformation**:
   - Apply log, square root, or other transformations to reduce the impact of outliers.
3. **Capping or Flooring**:
   - Replace extreme values with threshold values (e.g., Q1−1.5×IQRQ1 - 1.5 \times IQRQ1−1.5×IQR).
4. **Remove Outliers**:
   - Only if they are errors or irrelevant to the analysis.
5. **Imputation**:
   - Replace outliers with mean, median, or other calculated values.