Statistical Methods for Anomaly Detection

Assume data follows a known distribution (usually Gaussian)

Z-Score Method

$|z| > \text{threshold}$
(e.g., $\pm 3\sigma$)

Points beyond threshold are anomalies

Box Plot Method

$Q1 - 1.5 \times IQR$
$Q3 + 1.5 \times IQR$

Points beyond 1.5×IQR from quartiles

Mahalanobis Distance

$D^2 = (x-\mu)^T \Sigma^{-1} (x-\mu)$

For multi-feature data

Strengths

- ✓ Simple and interpretable
- ✓ Fast computation
- ✓ Well-understood theory
- ✓ Easy to implement

Limitations

- ✗ Assumes specific distribution
- ✗ Sensitive to outliers in training
- ✗ May not work for complex patterns
- ✗ Requires distribution knowledge

Works Well For

Low-dimensional data with known distribution

Preprocessing Tip

Remove known anomalies before fitting distribution

Z-Score Example

Student Test Score Analysis

Box Plot Example

Employee Salary Data Analysis

Mahalanobis Example

Customer Purchase Pattern (Amount,

```
Scores: 85, 90, 88, 92, 87, 89,
150 Mean(μ): 94.4 Std Dev(σ):
22.9 Z-score(150) = 2.43
```

150 detected as outlier ($|z| > 2$)

```
Salary($10k): 300, 320, 310, 330,
340, 325, 800 Q1 = 310, Q3 = 340
IQR = 30 Upper Fence = 340 +
1.5×30 = 385
```

$8M detected as outlier

Frequency)

```
Normal: ($500k, 10 times)
Suspicious: ($2M, 2 times)
Mahalanobis Distance considering
correlation = 5.8
```

Abnormal pattern detected ($D^2 > 4$)