4



数据预处理

异常值处理（Outlier Handling）

授课教师：[李阳源]

SL01:课程概述



- 课程主题: 异常值处理 (Outlier Handling)
- 学习目标:
  - 理解异常值的概念及影响
  - 掌握异常值检测方法
  - 学习异常值处理技术
- 主要内容:
  - 异常值的定义与类型
  - 异常值检测技术
  - 异常值处理方法
  - 异常值处理案例分析

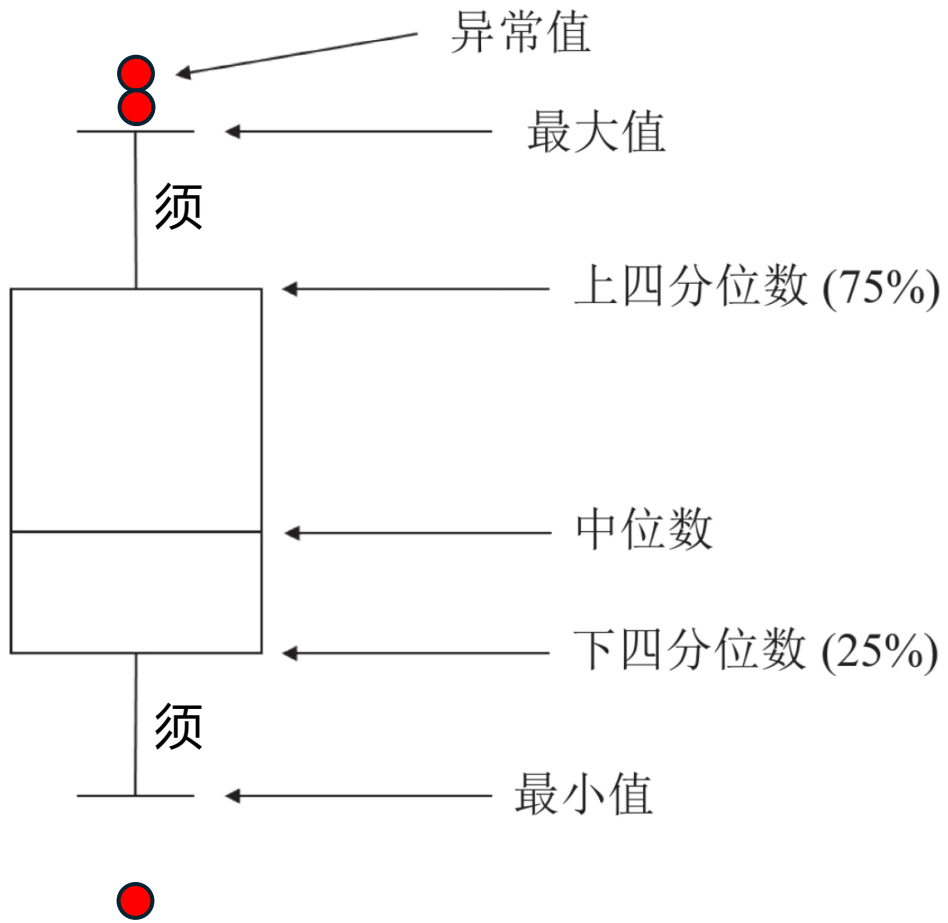
SL02:什么是异常值？



异常值(Outlier)的定义:

- 显著偏离整体数据模式(pattern)的值
- 超出正常范围的观测值
- 与其他观测值特性明显不同的数据点

箱线图



SL03: 异常值的重要性



异常值对分析的影响:

- 扭曲统计摘要值(zhāi yào zhí) (平均值、标准差等)
- 降低模型性能
- 可能导致错误结论
- 扭曲可视化结果

异常值处理的重要性:

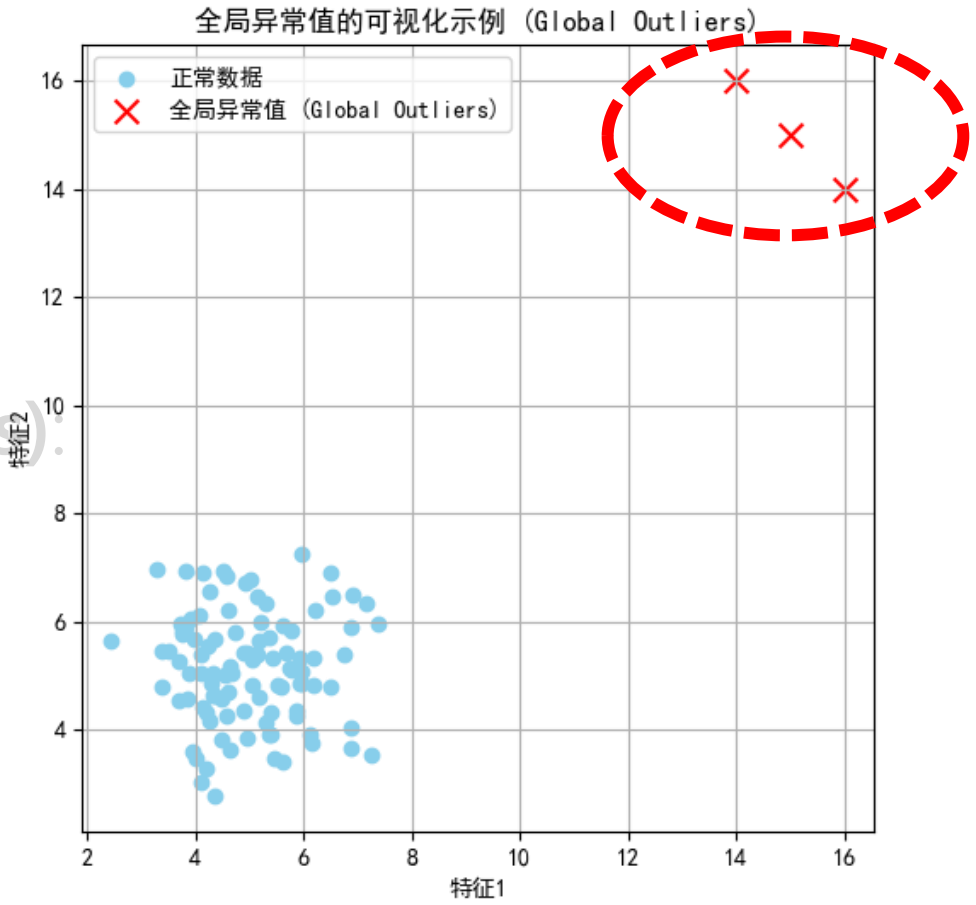
- 提高数据质量
- 改善模型准确性
- 确保分析结果可靠性(reliability)

SL04: 异常值的类型



异常值的主要类型:

1. **全局异常值(Global Outliers)**:
   1. 偏离整个数据集的值
   2. 例：收入数据中异常高的值
2. **上下文异常值(Contextual Outliers)**:
   1. 仅在特定上下文中被视为异常的值
   2. 例：冬季异常高温
3. **集体异常值(Collective Outliers)**:
   1. 单独看正常但一起出现时异常的值组
   2. 例：心电图中的异常模式



2:03 99

서비스 알림 결제 메시지

月上茶颜
17/04 12:35 PM

결제 완료
¥3.60
查看详情 >

결제 방법
UnionPay(8708)
거래 상대
月上茶颜

月上茶颜
17/04 12:31 PM

결제 완료
¥6.60
查看详情 >

결제 방법
UnionPay(8708)
거래 상대
月上茶颜

许昌好食穗餐饮管理有限公司
17/04 12:07 PM

결제 완료
¥22.00
查看详情 >

결제 방법
UnionPay(8708)

본 페이지 번역

2:03 99

서비스 알림 결제 메시지

许昌好食穗餐饮管理有限公司
18/04 6:21 PM

결제 완료
¥12.00
查看详情 >

결제 방법
UnionPay(8708)
거래 상대
许昌好食穗餐饮管理有限公司

许昌好食穗餐饮管理有限公司
18/04 6:12 PM

결제 완료
¥0.13
查看详情 >

결제 방법
UnionPay(8708)
거래 상대
许昌好食穗餐饮管理有限公司

月上茶颜
18/04 12:27 PM

결제 완료
¥6.60
查看详情 >

결제 방법

본 페이지 번역

2:03 99

서비스 알림 결제 메시지



许昌好食穗餐饮管理有限公司
18/04 12:06 PM

결제 완료
¥23.00
查看详情 >

결제 방법
UnionPay(8708)
거래 상대
许昌好食穗餐饮管理有限公司

许昌好食穗餐饮管理有限公司
18/04 7:04 AM

결제 완료
¥9.50
查看详情 >

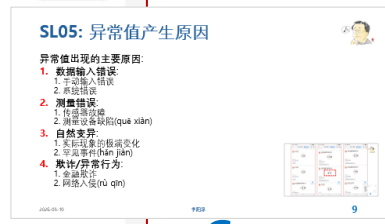
결제 방법
UnionPay(8708)
거래 상대
许昌好食穗餐饮管理有限公司

许昌职业技术学院
17/04 6:16 PM

결제 완료
¥12.50
查看详情 >

결제 방법
UnionPay(8708)

본 페이지 번역



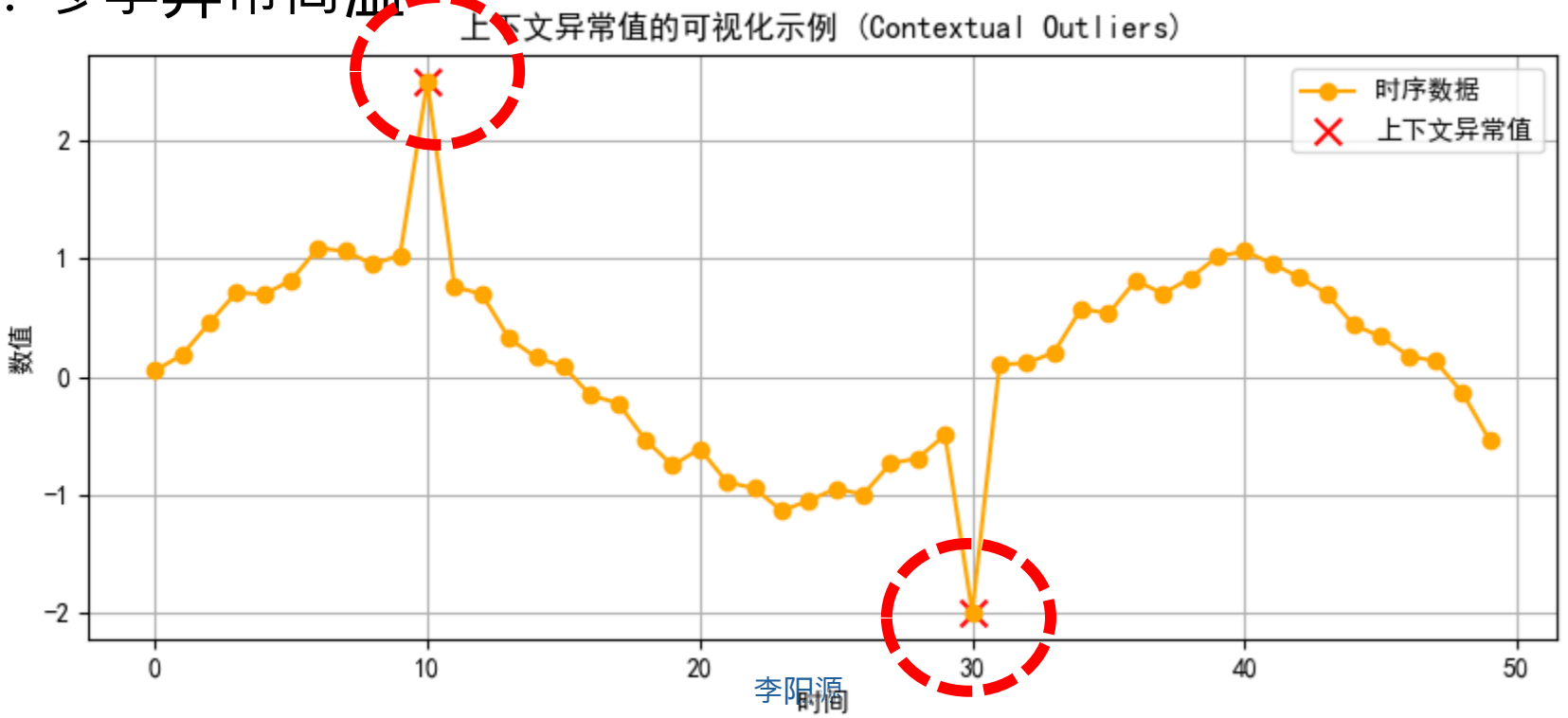
SL04: 异常值的类型



异常值的主要类型:

2. 上下文异常值(Contextual Outliers):

1. 仅在特定上下文中被视为异常的值
2. 例：冬季异常高温



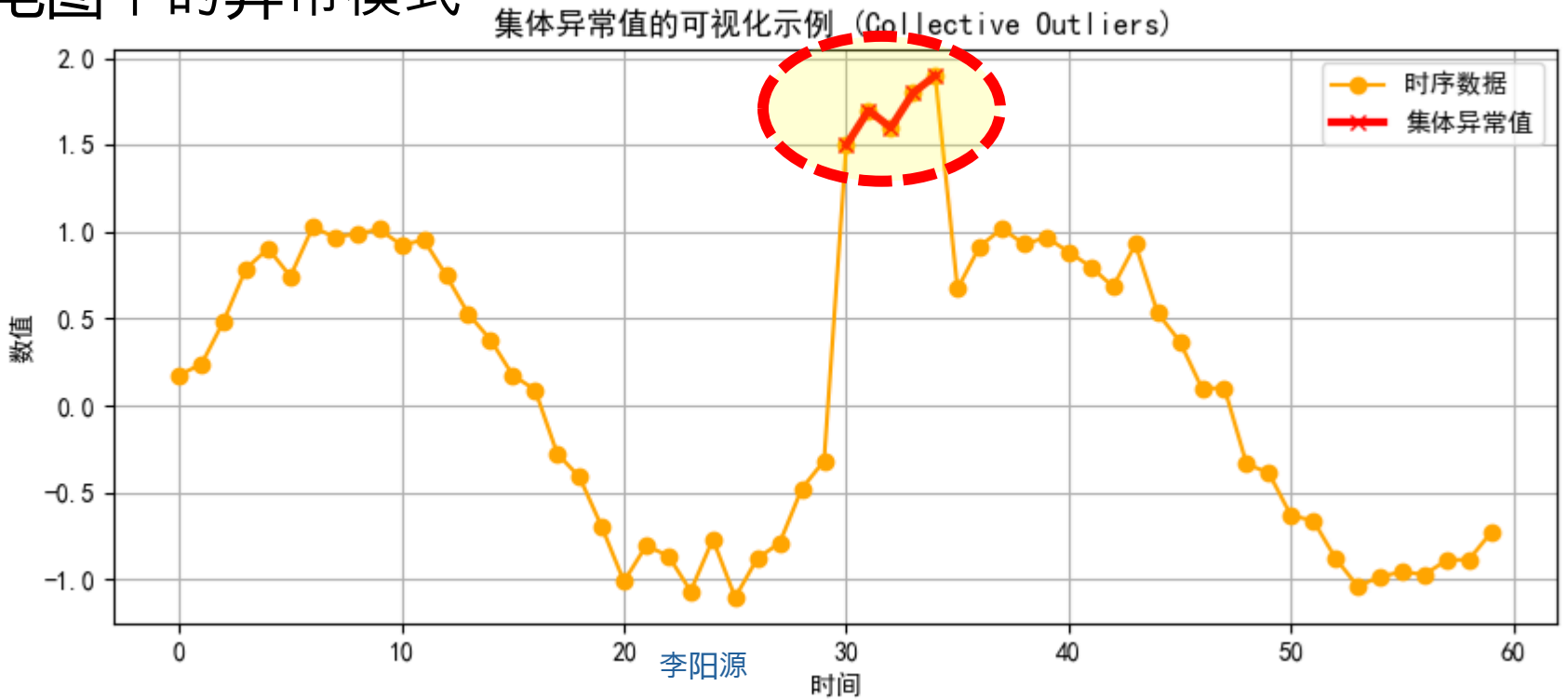
SL04: 异常值的类型



异常值的主要类型:

3. 集体异常值(Collective Outliers):

1. 单独看正常但一起出现时异常的值组
2. 例：心电图中的异常模式

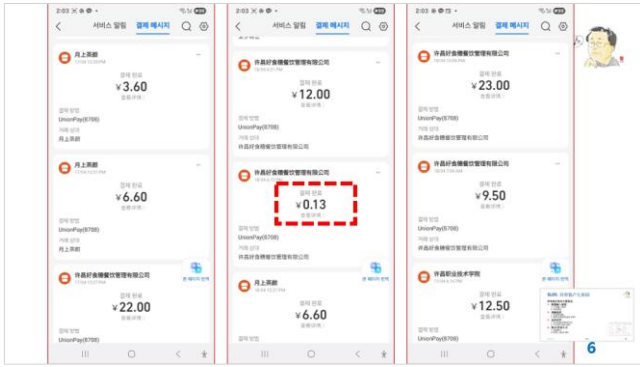


SL05: 异常值产生原因



异常值出现的主要原因:

1. **数据输入错误:**
   1. 手动输入错误
   2. 系统错误
2. **测量错误:**
   1. 传感器故障
   2. 测量设备缺陷(quē xiàn)
3. **自然变异:**
   1. 实际现象的极端变化
   2. 罕见事件(hǎn jiàn)
4. **欺诈/异常行为:**
   1. 金融欺诈
   2. 网络入侵(rù qīn)



SL06:异常值检测方法概述



主要异常值检测方法:

1. **统计方法:**
   1. Z-分数 (Z-score)
   2. 四分位距 (IQR)
   3. 马哈拉诺比斯距离 (Mahalanobis Distance)
2. **可视化方法:**
   1. 直方图(Histogram)
   2. 箱线图(**Box Plot** 或者 **Box-and-Whisker Plot**)
   3. 散点图（Scatter Plot）
3. **基于机器学习的方法:**
   1. 基于密度的方法(Density-Based Methods)
   2. 基于聚类的方法(Clustering-Based Methods)
   3. 集成方法(Ensemble Methods)

SL07: Z-分数(Z-score)方法



Z-分数概念:

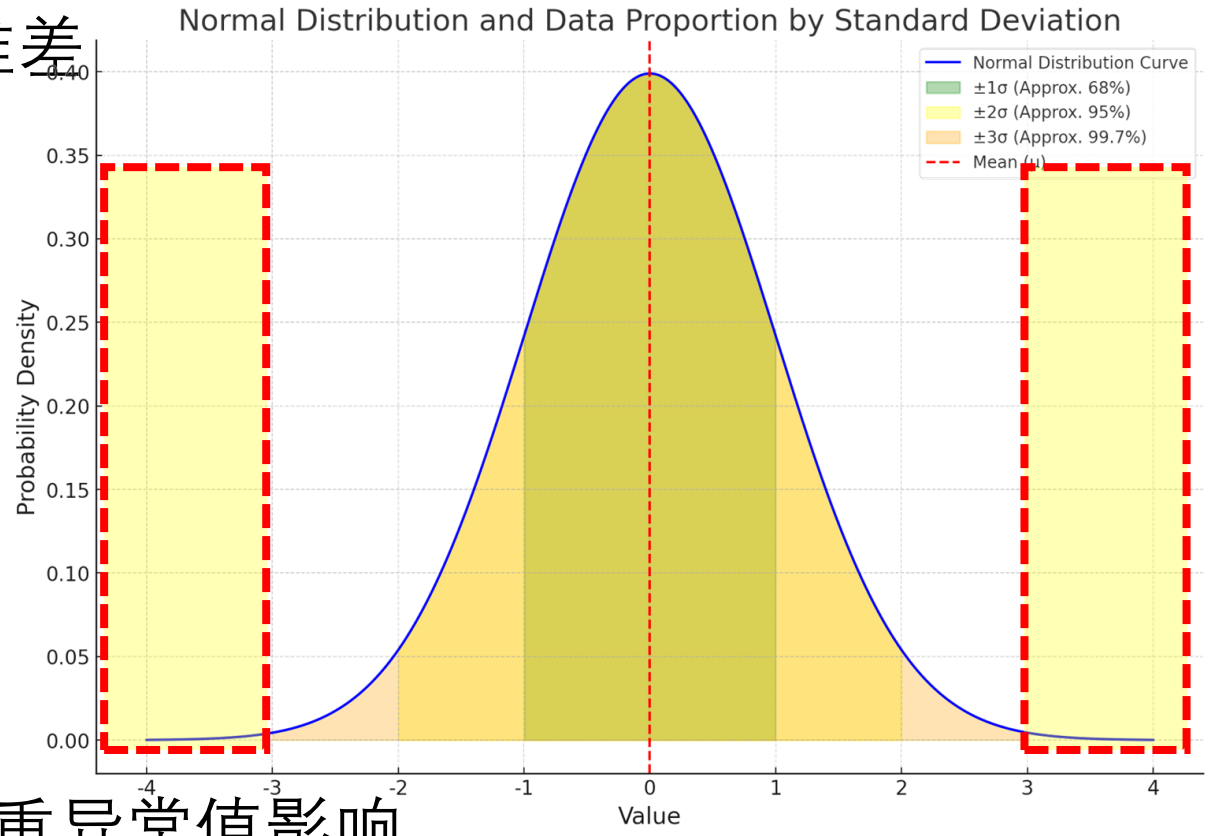
- 测量数据点偏离平均值多少个标准差
- $Z = (X - \mu) / \sigma$
  - X: 数据点
  - μ: 平均值
  - σ: 标准差

基于Z-分数的异常值判定:

- 通常$|Z| > 3$被视为异常值
- 需要正态分布假设

优点和局限性:

- 优点: 计算简单，易于理解
- 局限性: 需要正态分布假设，受多重异常值影响



SL08:Z-分数示例



示例数据: [10, 12, 15, 18, 20, 22, 25, 28, 30, 100]

计算过程:

1. 平均值(μ) = 28
2. 标准差(σ) = 26.85 $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$
3. Z-分数计算:

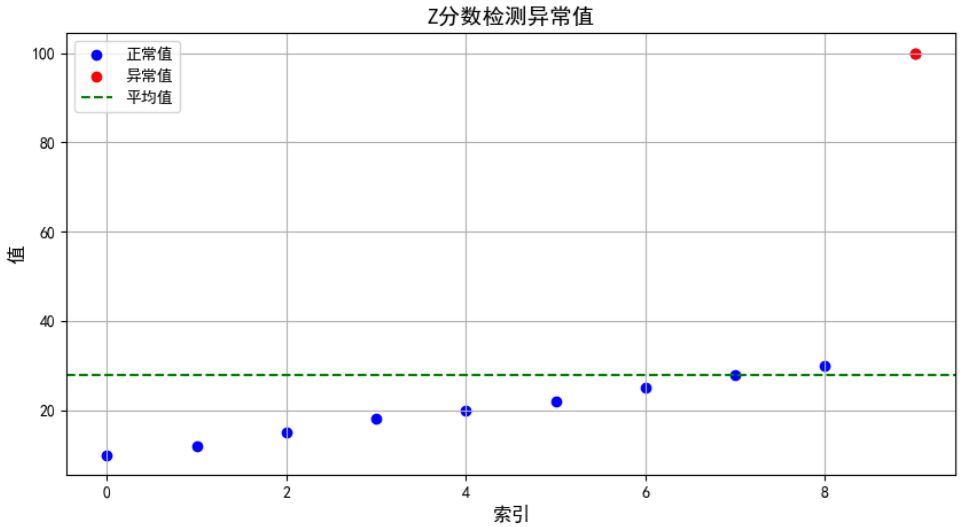
SL08:Z-分数示例



示例数据: [10, 12, 15, 18, 20, 22, 25, 28, 30, 100]

计算过程:

1. 样本数量：10
2. 计算平均值



$$\mu = \frac{10 + 12 + 15 + 18 + 20 + 22 + 25 + 28 + 30 + 100}{10} = 28.0$$

SL08:Z-分数示例



示例数据: [10, 12, 15, 18, 20, 22, 25, 28, 30, 100]

计算过程:

1. 样本数量：10
2. 平均值(μ) = 28
3. 标准差(σ) = 24.78

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

3. 计算每个数值与平均值的差及其平方 $(x_i - \mu)^2$



数据 x_i	$x_i - \mu$	$(x_i - \mu)^2$
10	-18	324

3. 计算每个数值与平均值的差及其平方 $(x_i - \mu)^2$



数据 x_i	$x_i - \mu$	$(x_i - \mu)^2$
10	-18	324
12	-16	256

3. 计算每个数值与平均值的差及其平方 $(x_i - \mu)^2$



数据 x_i	$x_i - \mu$	$(x_i - \mu)^2$
10	-18	324
12	-16	256
15	-13	169

3. 计算每个数值与平均值的差及其平方 $(x_i - \mu)^2$



数据 x_i	$x_i - \mu$	$(x_i - \mu)^2$
10	-18	324
12	-16	256
15	-13	169
18	-10	100

3. 计算每个数值与平均值的差及其平方 $(x_i - \mu)^2$



数据 x_i	$x_i - \mu$	$(x_i - \mu)^2$
10	-18	324
12	-16	256
15	-13	169
18	-10	100
20	-8	64

3. 计算每个数值与平均值的差及其平方 $(x_i - \mu)^2$



数据 x_i	$x_i - \mu$	$(x_i - \mu)^2$
10	-18	324
12	-16	256
15	-13	169
18	-10	100
20	-8	64
22	-6	36

3. 计算每个数值与平均值的差及其平方 $(x_i - \mu)^2$



数据 x_i	$x_i - \mu$	$(x_i - \mu)^2$
10	-18	324
12	-16	256
15	-13	169
18	-10	100
20	-8	64
22	-6	36
25	-3	9

3. 计算每个数值与平均值的差及其平方 $(x_i - \mu)^2$



数据 x_i	$x_i - \mu$	$(x_i - \mu)^2$
10	-18	324
12	-16	256
15	-13	169
18	-10	100
20	-8	64
22	-6	36
25	-3	9
28	0	0

3. 计算每个数值与平均值的差及其平方 $(x_i - \mu)^2$



数据 x_i	$x_i - \mu$	$(x_i - \mu)^2$
10	-18	324
12	-16	256
15	-13	169
18	-10	100
20	-8	64
22	-6	36
25	-3	9
28	0	0
30	2	4

3. 计算每个数值与平均值的差及其平方 $(x_i - \mu)^2$



数据 x_i	$x_i - \mu$	$(x_i - \mu)^2$
10	-18	324
12	-16	256
15	-13	169
18	-10	100
20	-8	64
22	-6	36
25	-3	9
28	0	0
30	2	4
100	72	5184

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

4. 求平方和（离均差平方和）

$$\sum (x_i - \mu)^2 = 6146$$



5. 计算总体方差（σ^2）

（分母是 N，因为是总体。）

$$\sigma^2 = \frac{6146}{10} = 614.6$$

6. 计算总体标准差（σ）

$$\sigma = \sqrt{614.6} \approx 24.78$$

✅ 最终结果

- 总体平均值（μ）：28.0
- 总体标准差（σ）：约 **24.78**

SL08:Z-分数示例



示例数据: [10, 12, 15, 18, 20, 22, 25, 28, 30, 100]

计算过程:

1. 平均值(μ) = 28
2. 标准差(σ) = 26.85
3. Z-分数计算: Z = (X - μ) / σ

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

2. 具体数据计算



数据 x_i	差值 $x_i - \mu$	Z-分数 Z
10	-18	$\frac{-18}{24.78} \approx -0.726$

2. 具体数据计算



数据 x_i	差值 $x_i - \mu$	Z-分数 Z
10	-18	$\frac{-18}{24.78} \approx -0.726$
12	-16	$\frac{-16}{24.78} \approx -0.646$

2. 具体数据计算



数据 x_i	差值 $x_i - \mu$	Z-分数 Z
10	-18	$\frac{-18}{24.78} \approx -0.726$
12	-16	$\frac{-16}{24.78} \approx -0.646$
15	-13	$\frac{-13}{24.78} \approx -0.525$

2. 具体数据计算



数据 x_i	差值 $x_i - \mu$	Z-分数 Z
10	-18	$\frac{-18}{24.78} \approx -0.726$
12	-16	$\frac{-16}{24.78} \approx -0.646$
15	-13	$\frac{-13}{24.78} \approx -0.525$
18	-10	$\frac{-10}{24.78} \approx -0.403$

2. 具体数据计算



数据 x_i	差值 $x_i - \mu$	Z-分数 Z
10	-18	$\frac{-18}{24.78} \approx -0.726$
12	-16	$\frac{-16}{24.78} \approx -0.646$
15	-13	$\frac{-13}{24.78} \approx -0.525$
18	-10	$\frac{-10}{24.78} \approx -0.403$
20	-8	$\frac{-8}{24.78} \approx -0.323$

2. 具体数据计算



数据 x_i	差值 $x_i - \mu$	Z-分数 Z
10	-18	$\frac{-18}{24.78} \approx -0.726$
12	-16	$\frac{-16}{24.78} \approx -0.646$
15	-13	$\frac{-13}{24.78} \approx -0.525$
18	-10	$\frac{-10}{24.78} \approx -0.403$
20	-8	$\frac{-8}{24.78} \approx -0.323$
22	-6	$\frac{-6}{24.78} \approx -0.242$

2. 具体数据计算



数据 x_i	差值 $x_i - \mu$	Z-分数 Z
10	-18	$\frac{-18}{24.78} \approx -0.726$
12	-16	$\frac{-16}{24.78} \approx -0.646$
15	-13	$\frac{-13}{24.78} \approx -0.525$
18	-10	$\frac{-10}{24.78} \approx -0.403$
20	-8	$\frac{-8}{24.78} \approx -0.323$
22	-6	$\frac{-6}{24.78} \approx -0.242$
25	-3	$\frac{-3}{24.78} \approx -0.121$
		0

2. 具体数据计算



数据 x_i	差值 $x_i - \mu$	Z-分数 Z
10	-18	$\frac{-18}{24.78} \approx -0.726$
12	-16	$\frac{-16}{24.78} \approx -0.646$
15	-13	$\frac{-13}{24.78} \approx -0.525$
18	-10	$\frac{-10}{24.78} \approx -0.403$
20	-8	$\frac{-8}{24.78} \approx -0.323$
22	-6	$\frac{-6}{24.78} \approx -0.242$
25	-3	$\frac{-3}{24.78} \approx -0.121$
28	0	$\frac{0}{24.78} = 0.000$

2. 具体数据计算

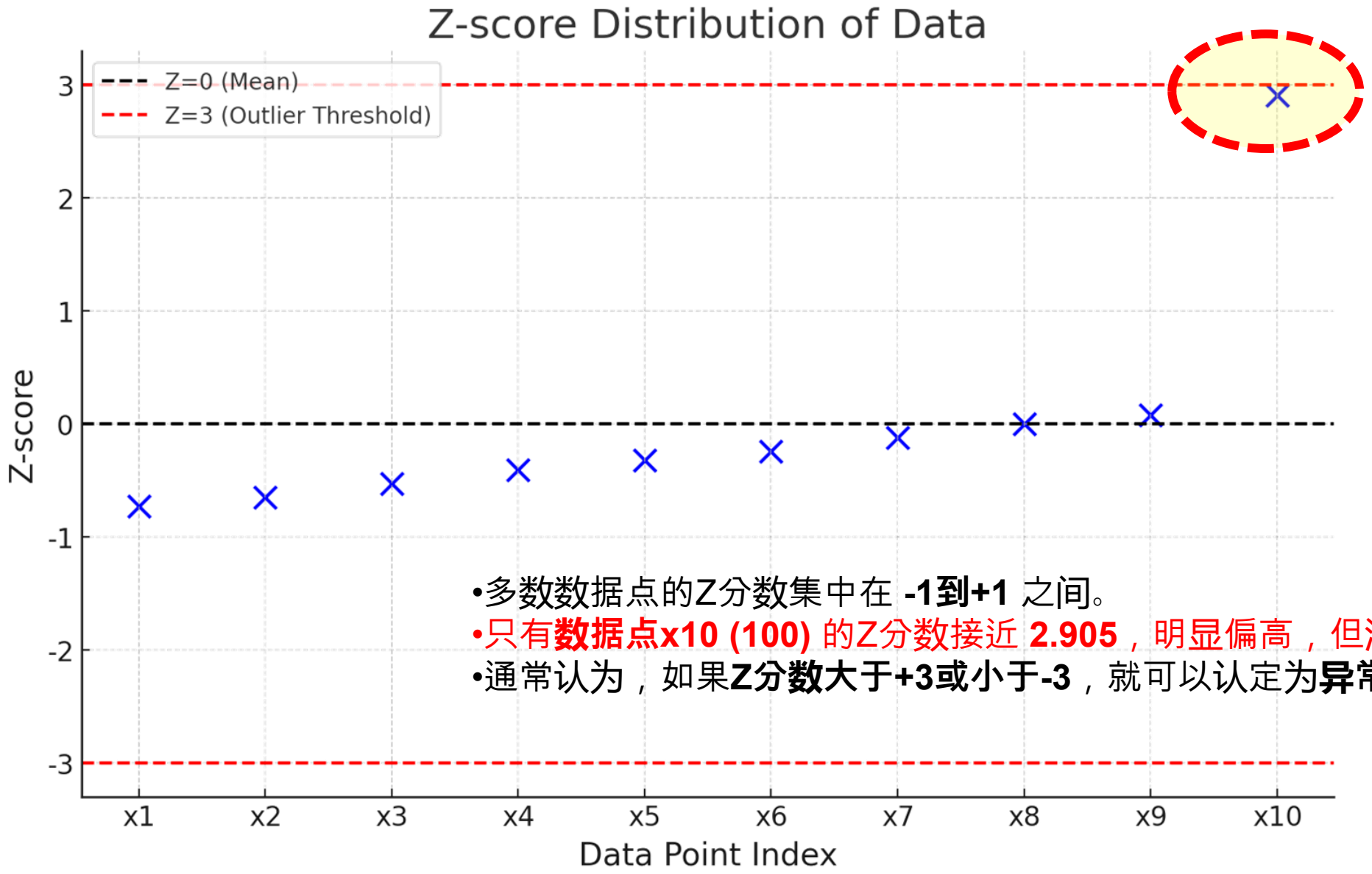


数据 x_i	差值 $x_i - \mu$	Z-分数 Z
10	-18	$\frac{-18}{24.78} \approx -0.726$
12	-16	$\frac{-16}{24.78} \approx -0.646$
15	-13	$\frac{-13}{24.78} \approx -0.525$
18	-10	$\frac{-10}{24.78} \approx -0.403$
20	-8	$\frac{-8}{24.78} \approx -0.323$
22	-6	$\frac{-6}{24.78} \approx -0.242$
25	-3	$\frac{-3}{24.78} \approx -0.121$
28	0	$\frac{0}{24.78} = 0.000$
30	2	$\frac{2}{24.78} \approx 0.081$

2. 具体数据计算



数据 x_i	差值 $x_i - \mu$	Z-分数 Z
10	-18	$\frac{-18}{24.78} \approx -0.726$
12	-16	$\frac{-16}{24.78} \approx -0.646$
15	-13	$\frac{-13}{24.78} \approx -0.525$
18	-10	$\frac{-10}{24.78} \approx -0.403$
20	-8	$\frac{-8}{24.78} \approx -0.323$
22	-6	$\frac{-6}{24.78} \approx -0.242$
25	-3	$\frac{-3}{24.78} \approx -0.121$
28	0	$\frac{0}{24.78} = 0.000$
30	2	$\frac{2}{24.78} \approx 0.081$
100	72	$\frac{72}{24.78} \approx 2.905$





•多数数据点的Z分数集中在 **-1到+1** 之间。
•只有**数据点x10 (100)** 的Z分数接近 **2.905**，明显偏高，但没有超过3。
•通常认为，如果**Z分数大于+3或小于-3**，就可以认定为**异常值（Outlier）**

SL06:异常值检测方法概述



主要异常值检测方法:

1. **统计方法:**
   1. Z-分数 (Z-score)
   2. 四分位距 (IQR)
   3. 马哈拉诺比斯距离 (Mahalanobis Distance)
2. **可视化方法:**
   1. 直方图(Histogram)
   2. 箱线图(**Box Plot** 或者 **Box-and-Whisker Plot**)
   3. 散点图（Scatter Plot）
3. **基于机器学习的方法:**
   1. 基于密度的方法(Density-Based Methods)
   2. 基于聚类的方法(Clustering-Based Methods)
   3. 集成方法(Ensemble Methods)

SL09: 四分位距(IQR)方法



IQR概念:

- IQR = Q3 - Q1 (第三四分位数 - 第一四分位数)
- 包含数据中间50%的范围

基于IQR的异常值判定:

- 下限边界: Q1 - 1.5 × IQR
- 上限边界: Q3 + 1.5 × IQR
- 超出边界的值被视为异常值

优点和局限性:

- 优点: 不需要分布假设，具有鲁棒性
- 局限性: 在小型数据集中可靠性较低``

SL10: IQR方法示例



示例数据: [10, 12, 15, 18, 20, 22, 25, 28, 30, 100]

计算过程:

步骤 2：计算四分位数（Q1, Q2, Q3）

- **中位数 Q2：**
  - 数据总数为 10（偶数个）
  - 中间两个数为 20 和 22
  - Q2=(20+22)/2=21
- **第一个四分位数 Q1（中间左侧数据的中位数）：**
  - 左侧五个数：[10, 12, 15, 18, 20] → 中位数是 **15**
- **第三个四分位数 Q3（中间右侧数据的中位数）：**
  - 右侧五个数：[22, 25, 28, 30, 100] → 中位数是 **28**

SL10: IQR方法示例



步骤 3：计算四分位距 IQR

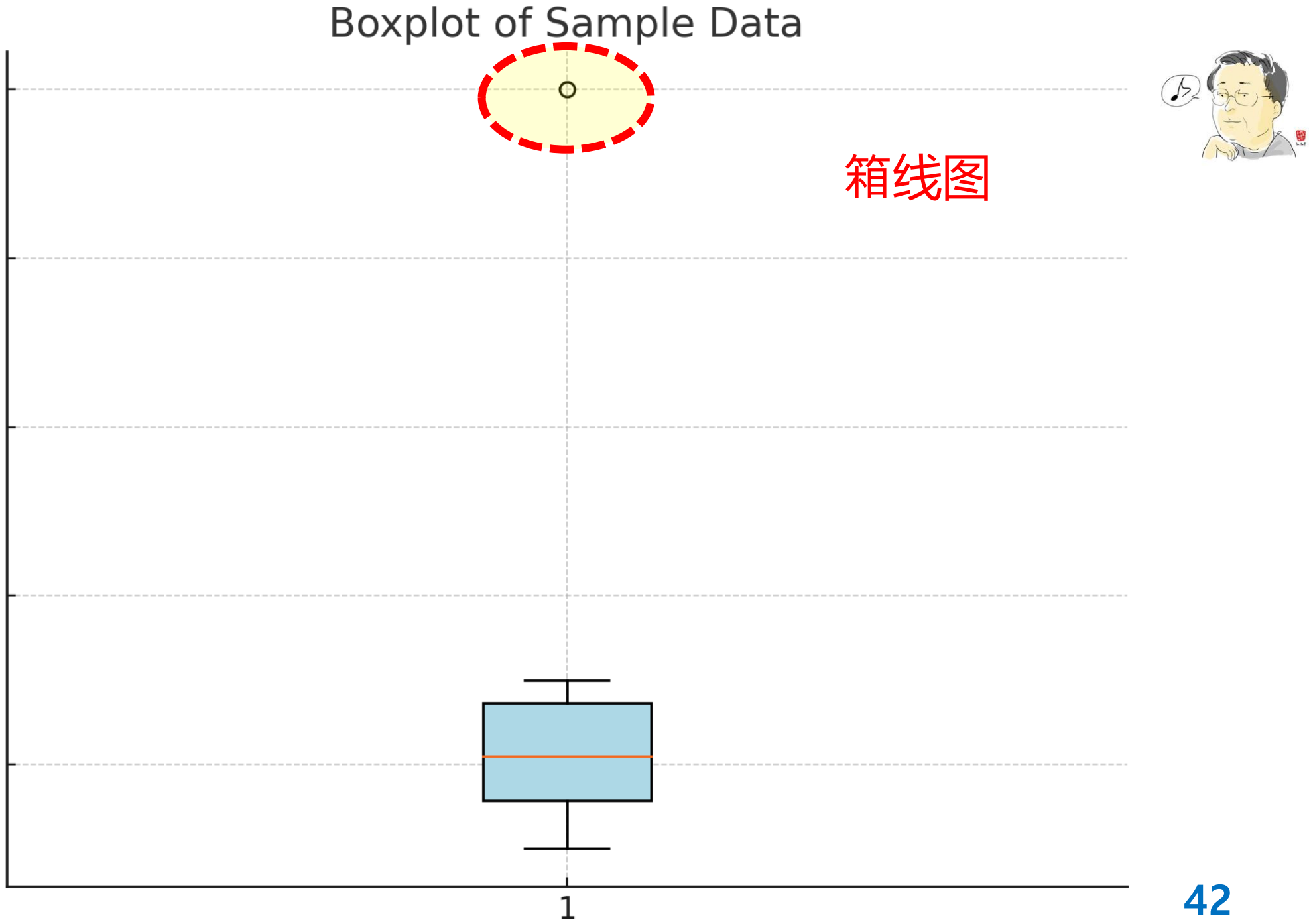
$$IQR = Q3 - Q1 = 28 - 15 = 13$$

步骤 4：计算异常值判断边界

- **下界：$Q1 - 1.5 \times IQR = 15 - 1.5 \times 13 = 15 - 19.5 = -4.5$**
- **上界：$Q3 + 1.5 \times IQR = 28 + 1.5 \times 13 = 28 + 19.5 = 47.5$**

步骤 5：判断异常值

- **在本例中，100 > 47.5，所以 100 是异常值**
- **所有小于 -4.5 或大于 47.5 的数据都视为异常值**


Boxplot of Sample Data
箱线图
1

SL06:异常值检测方法概述



主要异常值检测方法:

1. **统计方法:**
   1. Z-分数 (Z-score)
   2. 四分位距 (IQR)
   3. 马哈拉诺比斯距离 (Mahalanobis Distance)
2. **可视化方法:**
   1. 直方图(Histogram)
   2. 箱线图(**Box Plot** 或者 **Box-and-Whisker Plot**)
   3. 散点图（Scatter Plot）
3. **基于机器学习的方法:**
   1. 基于密度的方法(Density-Based Methods)
   2. 基于聚类的方法(Clustering-Based Methods)
   3. 集成方法(Ensemble Methods)

SL13: 多变量异常值检测

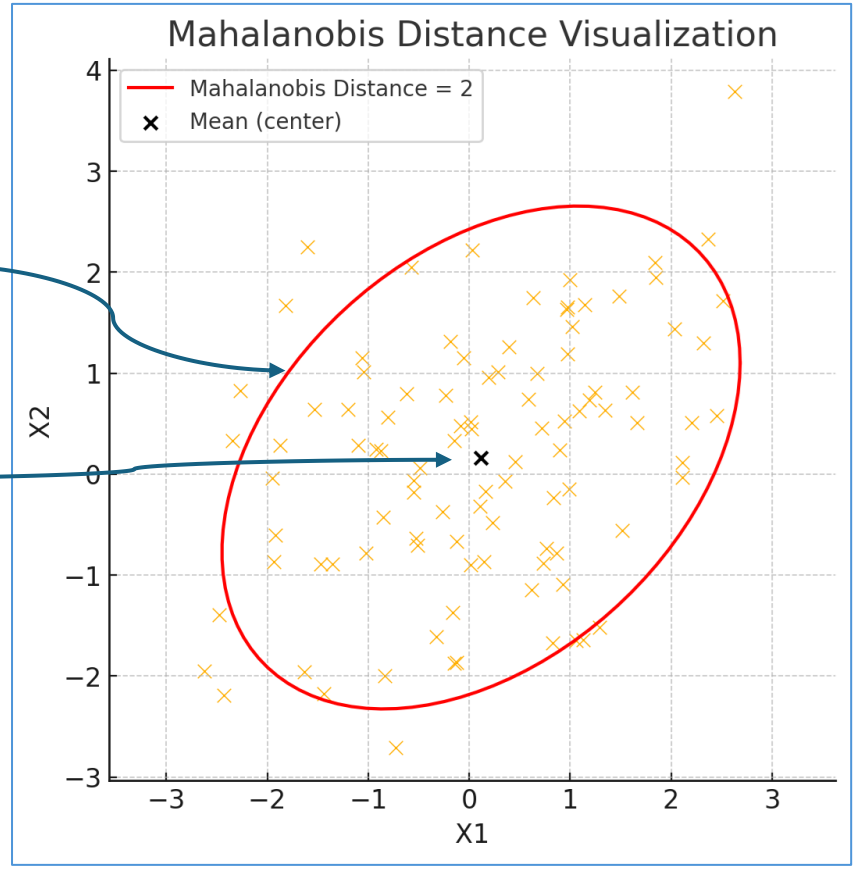


1. 马哈拉诺比斯距离(Mahalanobis Distance):

- 多维空间中数据点与分布中心的距离
- 考虑变量间相关性

红色椭圆代表**马哈拉诺比斯距离 = 2**的等距边界

黑色的“X”表示**数据的平均值（中心）**



SL06:异常值检测方法概述



主要异常值检测方法:

1. **统计方法:**
   1. Z-分数 (Z-score)
   2. 四分位距 (IQR)
   3. 马哈拉诺比斯距离 (Mahalanobis Distance)
2. **可视化方法:**
   1. 直方图(Histogram)
   2. 箱线图(**Box Plot** 或者 **Box-and-Whisker Plot**)
   3. 散点图 (Scatter Plot)
3. **基于机器学习的方法:**
   1. 基于密度的方法(Density-Based Methods)
   2. 基于聚类的方法(Clustering-Based Methods)
   3. 集成方法(Ensemble Methods)

SL12: 可视化异常值检测方法



1. 利用直方图:

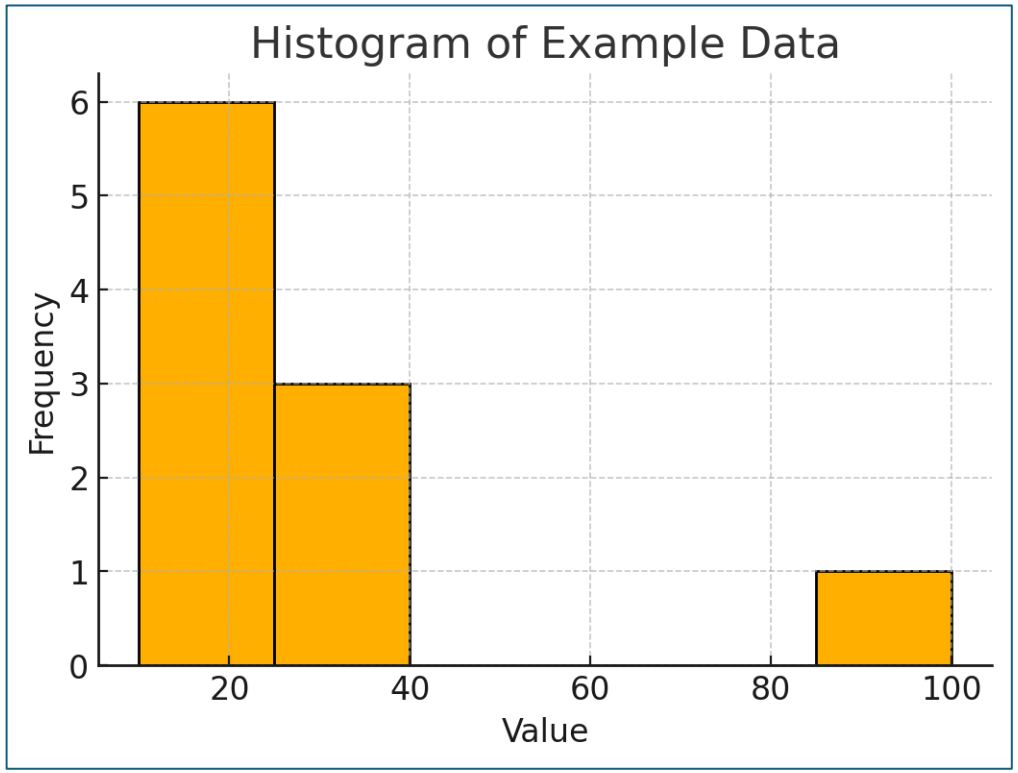
- 可视化数据分布
- 识别独立峰值
- 识别极端值

2. 利用散点图:

- 识别二维关系中偏离模式的点
- 识别远离聚类的点

3. 利用时间序列图:

- 检测随时间变化的急剧变化
- 考虑季节性的异常值识别



SL06:异常值检测方法概述



主要异常值检测方法:

1. **统计方法:**
   1. Z-分数 (Z-score)
   2. 四分位距 (IQR)
   3. 马哈拉诺比斯距离 (Mahalanobis Distance)
2. **可视化方法:**
   1. 直方图(Histogram)
   2. 箱线图(**Box Plot** 或者 **Box-and-Whisker Plot**)
   3. 散点图 （Scatter Plot）
3. **基于机器学习的方法:**
   1. 基于密度的方法(Density-Based Methods)
   2. 基于聚类的方法(Clustering-Based Methods)
   3. 集成方法(Ensemble Methods)

SL11: 使用箱线图(Box Plot)可视化异常值

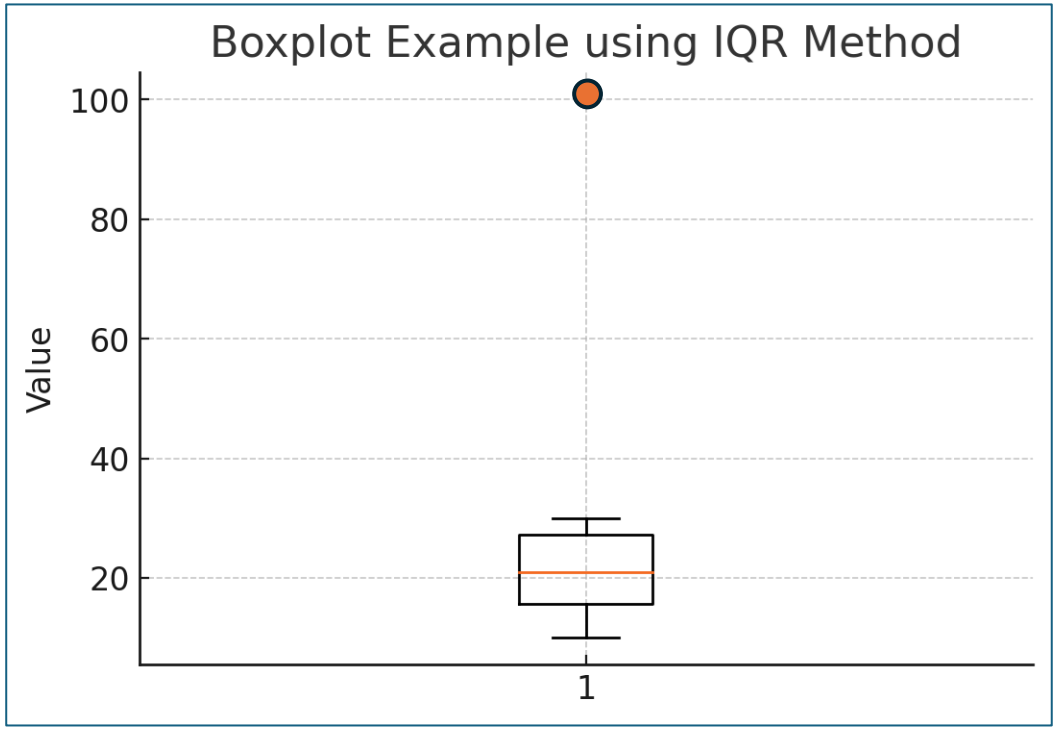


箱线图组成:

- 盒子: Q1到Q3 (IQR)
- 中线: 中位数(Q2)
- 须: 延伸至上/下限边界
- 点: 异常值

箱线图解读:

- 了解分布形态
- 确认异常值位置
- 评估数据对称性



SL06:异常值检测方法概述



主要异常值检测方法:

1. **统计方法:**
   1. Z-分数 (Z-score)
   2. 四分位距 (IQR)
   3. 马哈拉诺比斯距离 (Mahalanobis Distance)
2. **可视化方法:**
   1. 直方图(Histogram)
   2. 箱线图(**Box Plot** 或者 **Box-and-Whisker Plot**)
   3. 散点图 （Scatter Plot）
3. **基于机器学习的方法:**
   1. 基于密度的方法(Density-Based Methods)
   2. 基于聚类的方法(Clustering-Based Methods)
   3. 集成方法(Ensemble Methods)

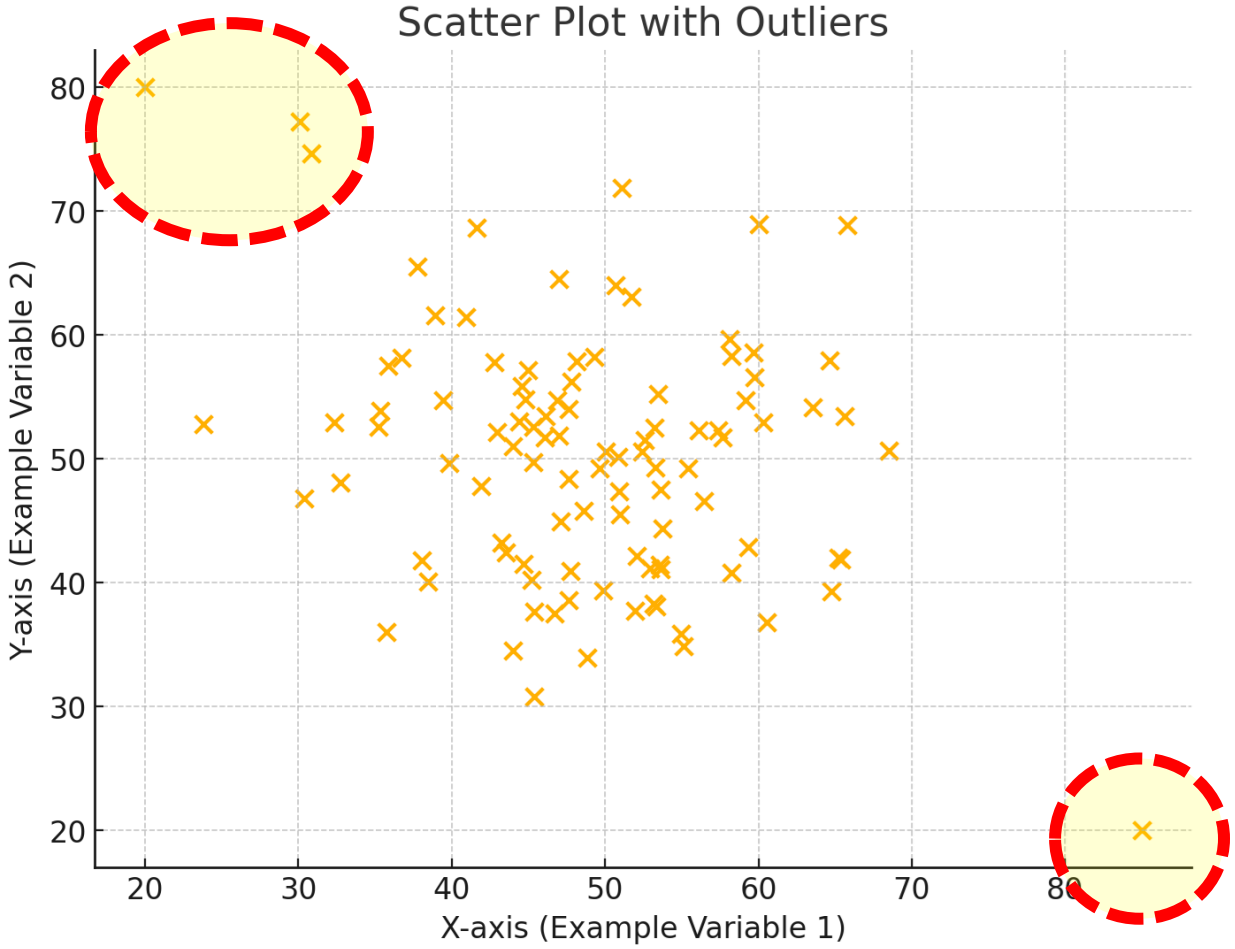
SL12: 可视化异常值检测方法



1. 利用直方图:
- 可视化数据分布
- 识别独立峰值
- 识别极端值

2. 利用**散点图**:
- 识别二维关系中偏离模式的点
- 识别远离聚类的点

3. 利用时间序列图:
- 检测随时间变化的急剧变化
- 考虑季节性的异常值识别

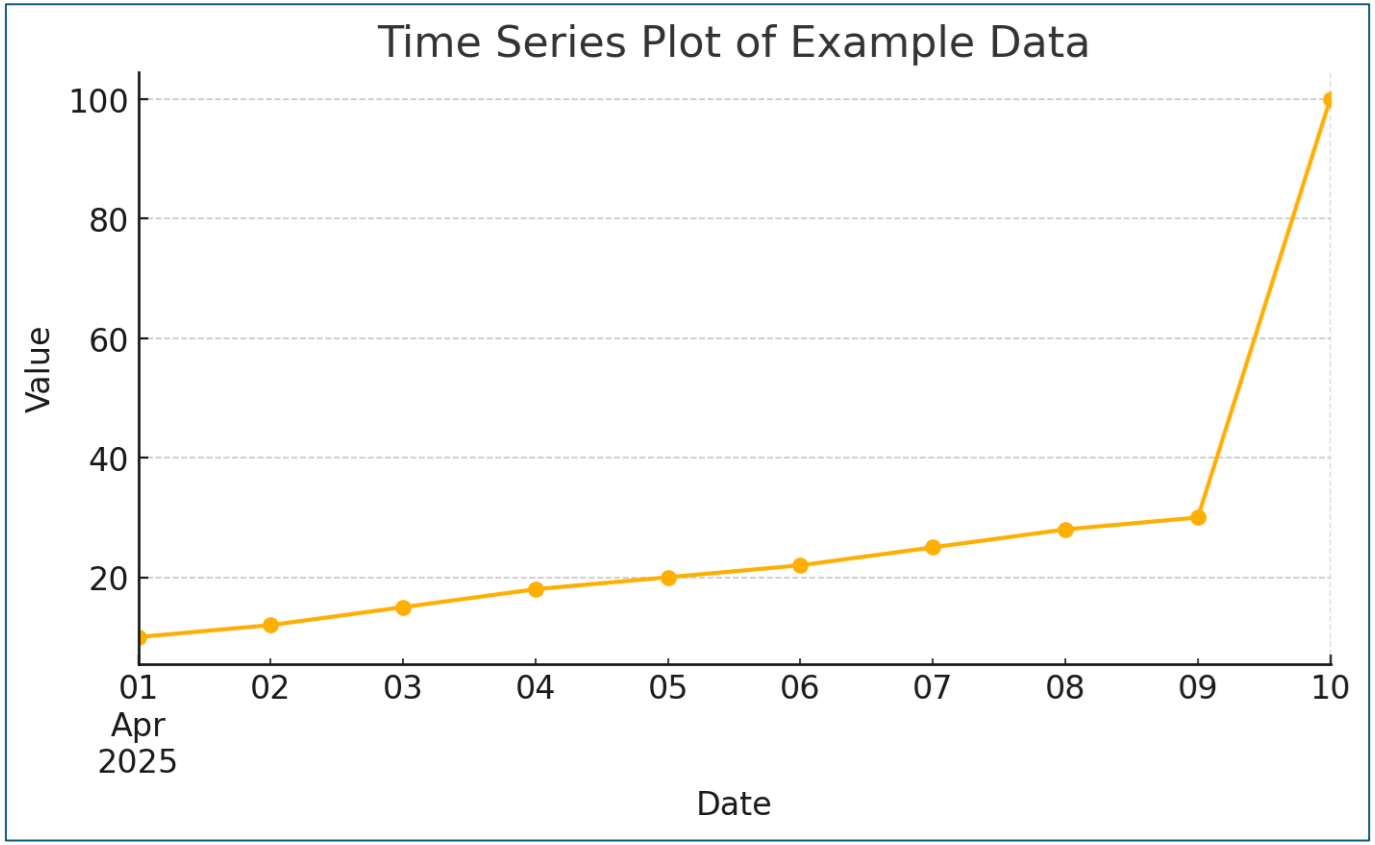


SL12: 可视化异常值检测方法



1. 利用直方图:
- 可视化数据分布
- 识别独立峰值
- 识别极端值
2. 利用散点图:
- 识别二维关系中偏离模式的点
- 识别远离聚类的点

3. 利用时间序列图:
- 检测随时间变化的急剧变化
- 考虑季节性的异常值识别



SL06:异常值检测方法概述



主要异常值检测方法:

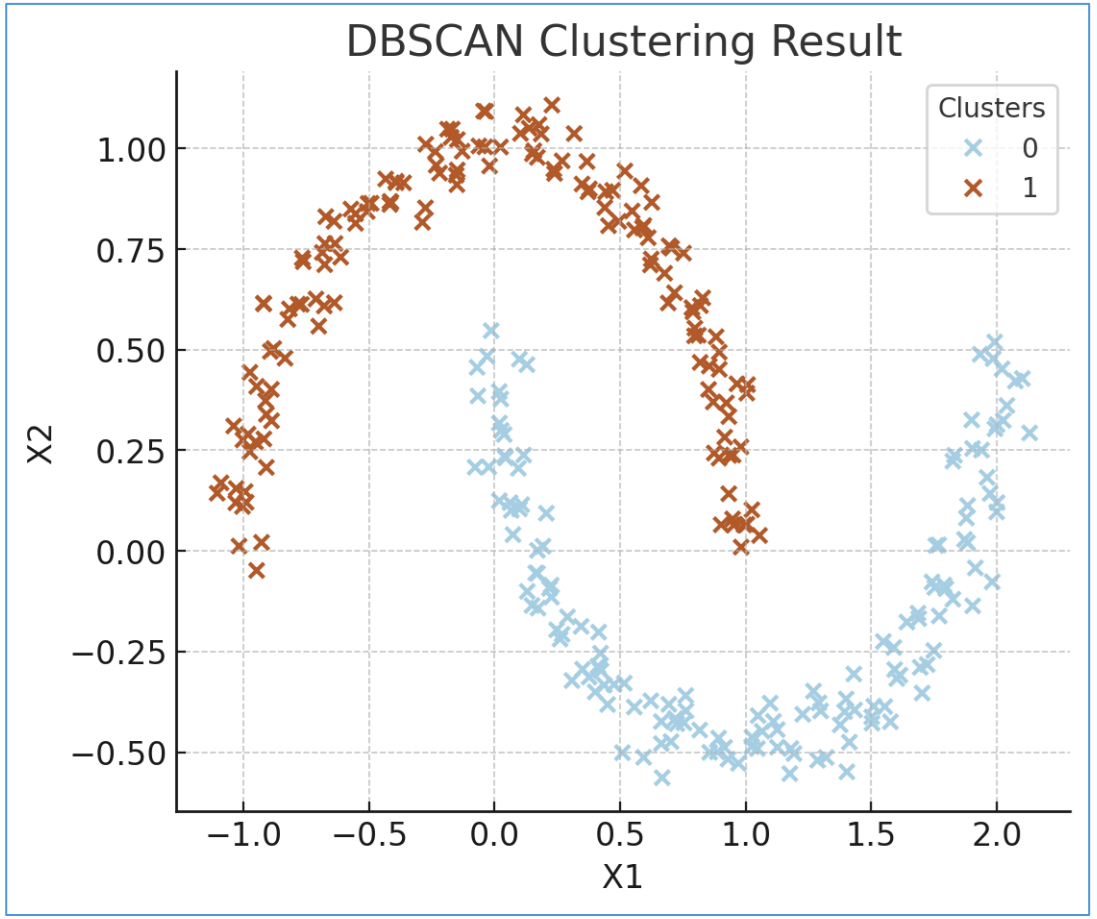
1. **统计方法:**
   1. Z-分数 (Z-score)
   2. 四分位距 (IQR)
   3. 马哈拉诺比斯距离 (Mahalanobis Distance)
2. **可视化方法:**
   1. 直方图(Histogram)
   2. 箱线图(**Box Plot** 或者 **Box-and-Whisker Plot**)
   3. 散点图 （Scatter Plot）
3. **基于机器学习的方法:**
   1. 基于密度的方法(Density-Based Methods)
   2. 基于聚类的方法(Clustering-Based Methods)
   3. 集成方法(Ensemble Methods)

SL13: 多变量异常值检测



2. DBSCAN算法:

- 基于密度的聚类
- 将低密度区域的数据点识别为异常值



Density-Based Spatial Clustering of Applications with Noise

SL13: 多变量异常值检测

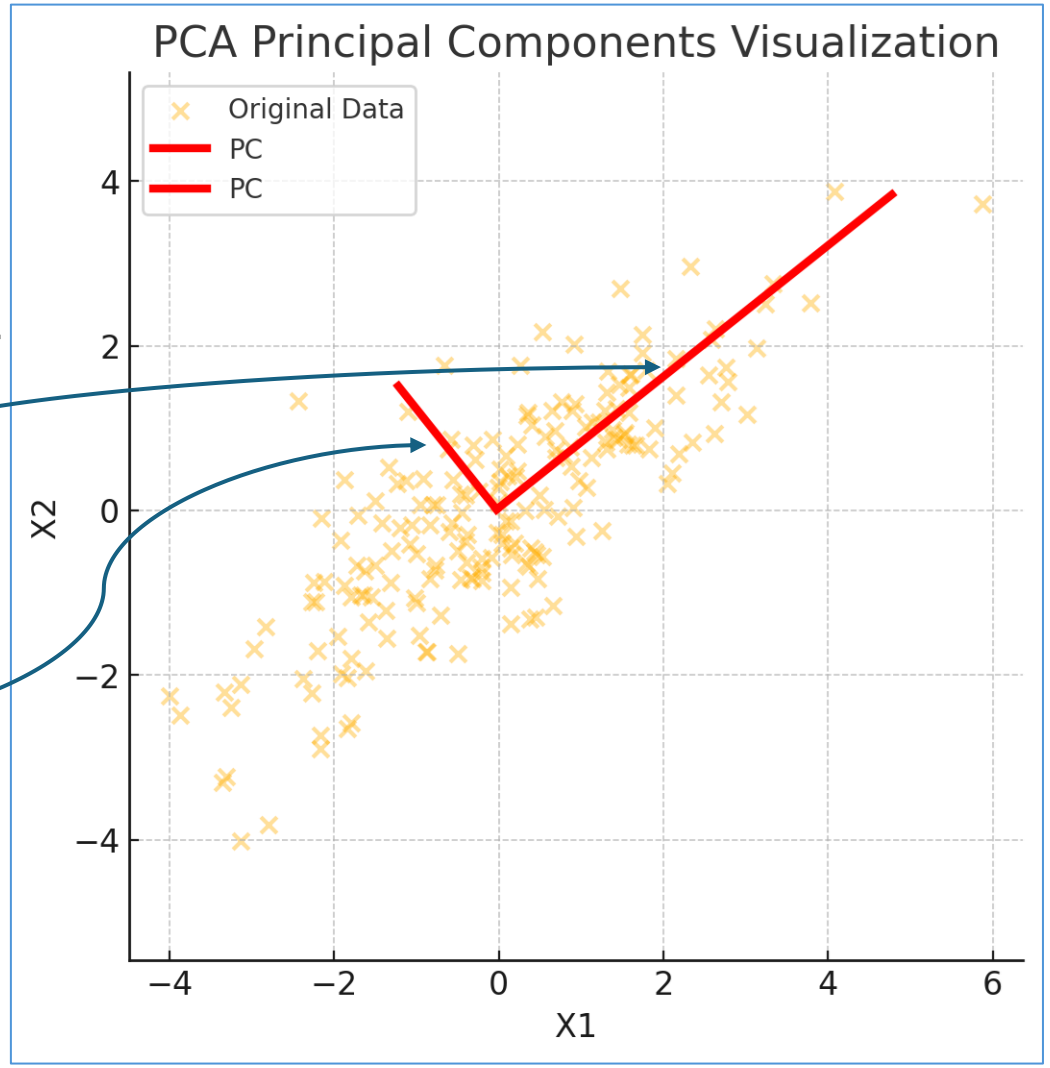


3. 主成分分析(PCA):

- 降维后利用重构误差
- 具有高重构误差的数据点为异常值

第一主成分（PC1）：数据分布**最广的方向**（方差最大）

第二主成分（PC2）：与第一主成分**垂直**，并且具有**第二大方差**的方向



SL14: 异常值处理方法



1. **删除(Deletion)**:
- 移除异常值
- 何时使用: 明确错误，数据量充足
2. **替换(Imputation)**:
- 用均值、中位数、众数等替换
- 基于预测模型的替换
3. **转换(Transformation)**:
- 对数转换，平方根转换
- 分位数转换(Quantile Transformation)
4. **单独处理(Separate Handling)**:
- 将异常值作为单独类别处理
- 分析异常值本身的模式

SL15:异常值处理注意事项



1. **利用领域知识**:
- 考虑行业/领域特性
- 参考专家意见
2. **理解异常值原因**:
- 区分是简单错误还是重要信号
- 根据原因选择适当处理方法
3. **记录数据处理过程**:
- 文档化异常值处理过程
- 比较处理前后的数据
4. **敏感度分析**:
- 确认处理异常值后结果的变化
- 比较不同处理方法

SL16: 异常值处理案例 - 金融数据



案例: 信用卡欺诈检测

数据特征:

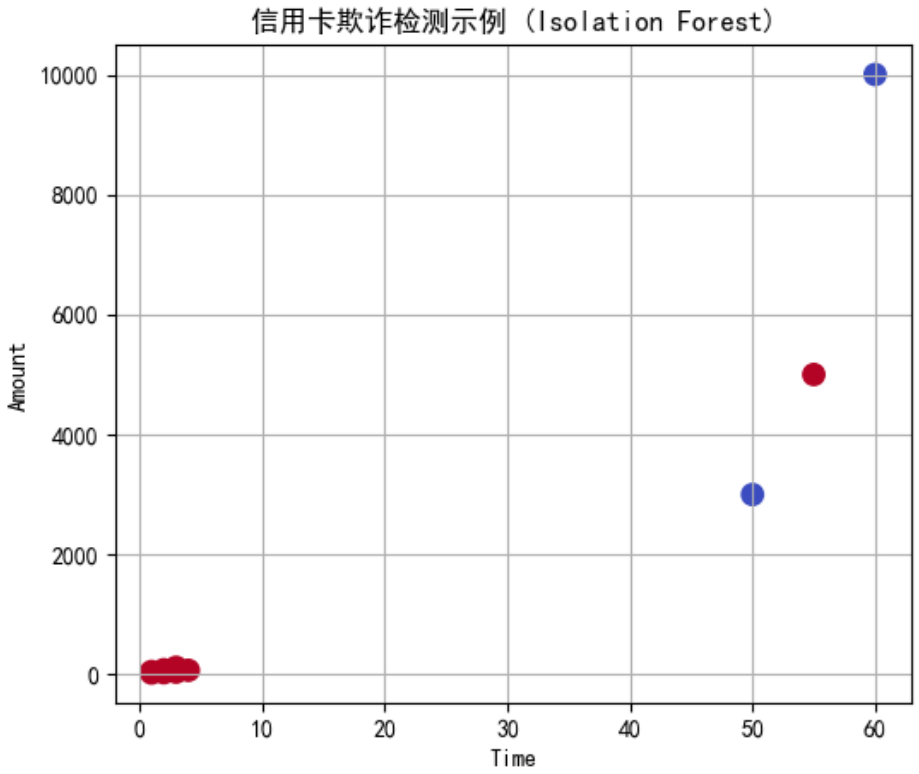
- 交易金额、位置、时间、频率等

异常值检测方法:

- 基于用户行为配置文件的异常检测
- 时间/空间模式分析

处理方法:

- 对异常交易进行标记
- 利用多重异常分数(Anomaly Score)



SL17: 异常值处理案例 - 传感器数据



案例: 工厂设备监控

数据特征:

- 温度、压力、振动、噪声等传感器数据
- 时间序列特性

异常值检测方法:

- 基于移动平均的波动性分析
- 季节性和趋势分解

处理方法:

- 应用平滑(Smoothing)技术
- 将异常值信息用于未来预测模型

时间(Time)	温度(°C)	振动(mm/s)	电流(A)
1	60	1.2	12
2	62	1.3	12.2
...	...	...	...
50	95	5.6	16.8 ← 이상치 가능성

SL18: 异常值处理案例 - 营销数据



案例: 客户行为分析

数据特征:

- 购买金额、频率、网站停留时间等

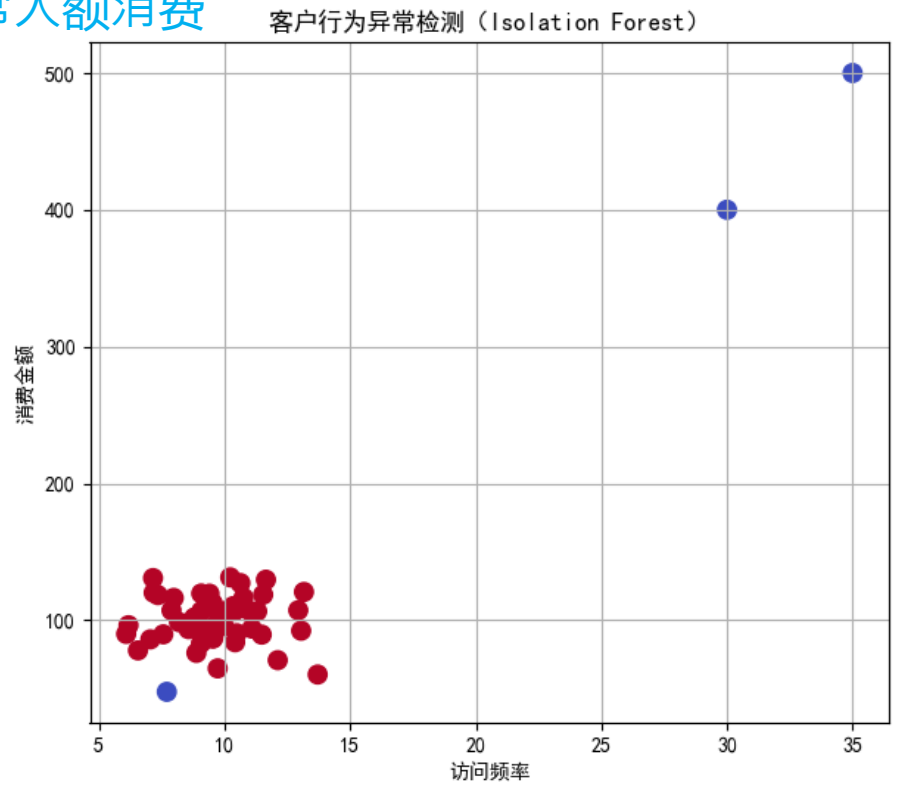
异常值检测方法:

- RFM分析中的异常值
- 基于聚类的客户细分分析

处理方法:

- 作为VIP客户单独管理理
- 用作特别营销活动的目标

- **红色点**（值为 1）：正常客户
- **蓝色点**（值为 -1）：异常客户，例如：
- 访问频率极高但消费金额异常
- 短时间内异常大额消费





数据预处理

理解度测验 · 20题 · 四选一

授课教师：李阳源 · 2026-03-02 · 第4章

题目

下列哪项不属于异常值（Outlier）的定义特征?

A 显著偏离整体数据模式的值

B 超出正常范围的观测值

C 与其他观测值特性明显不同的数据点

D 数据集中出现频率最高的值

题目

在Z-分数方法中，通常将哪个阈值作为判定异常值的标准?

A |Z| > 1

B |Z| > 2

C |Z| > 3

D |Z| > 5

题目

给定数据集 [10, 12, 15, 18, 20, 22, 25, 28, 30, 100]，其IQR（四分位距）值为多少?

A 13

B 15

C 18

D 21

题目

使用IQR方法，数据集 [10, 12, 15, 18, 20, 22, 25, 28, 30, 100] 的异常值判断上界为?

（ Q1=15, Q3=28, IQR=13 ）

A 43.0

B 47.5

C 50.0

D 55.5

题目

'冬季出现异常高温'属于哪种类型的异常值?

A 全局异常值（Global Outliers）

B 上下文异常值（Contextual Outliers）

C 集体异常值（Collective Outliers）

D 系统性异常值（Systematic Outliers）

题目

心电图中某段连续波形，单独看每个数据点都属于正常范围，但组合在一起却表现异常。这是哪种异常值类型?

A 全局异常值（Global Outliers ）

B 上下文异常值（Contextual Outliers ）

C 集体异常值（Collective Outliers ）

D 随机噪声（Random Noise ）

题目

关于Z-分数（Z-score）方法的局限性，以下哪项描述是正确的?

A 计算过于复杂，不易理解

B 无法处理正态分布数据

C 需要正态分布假设，且受多重异常值影响

D 只能用于单变量数据分析

题目

与Z-分数方法相比，IQR方法最主要的优点是什么？

A 计算结果更精确

B 不需要对数据分布做假设，具有鲁棒性

C 可以处理多维数据

D 适用于时间序列数据

题目

数据集 [10, 12, 15, 18, 20, 22, 25, 28, 30, 100] 中，值100的Z分数约为2.905，未超过3。根据Z-分数方法，以下判断正确的是?

（μ=28, σ=24.78）

A 100是异常值，因为它明显偏高

B 100不是异常值，因为Z分数未超过3

C 无法判断，需要更多数据

D 100是临界值，需要人工判断

题目

一家工厂的传感器监控系统中，某时刻同时出现温度、振动、电流均略高于正常值但未超过单项阈值的情况。从异常值类型角度，这最可能属于？

A 全局异常值

B 上下文异常值

C 集体异常值

D 测量误差

题目

在箱线图（Box Plot）中，须（Whisker）延伸至的位置表示什么?

A 数据的最大值和最小值

B 平均值±标准差的范围

C IQR方法中的上下限边界

D 数据的95%置信区间

题目

在哪种情况下，删除（Deletion）是处理异常值最合适的方法?

A 数据量较少且异常值原因不明

B 异常值明确是错误输入且数据量充足

C 异常值可能代表欺诈行为

D 数据集中异常值比例超过20%

题目

马哈拉**诺比斯距离（Mahalanobis Distance）**与普通**欧氏距离相比，最大的区别在于?**

A 计算速度更快

B 只能用于二维数据

C 考虑了变量间的相关性

D 不需要计算均值

题目

某研究员对一个样本量仅有15个的数据集使用IQR方法检测异常值，得到了5个异常值（占33%）。以下分析最合理的是？

A IQR方法完全可靠，结果有效

B 应考虑IQR在小样本中可靠性较低，结合领域知识验证

C 应该改用Z-分数方法重新计算

D 33%的异常值比例在正常范围内

题目

以下哪种方法不属于基于统计的异常值检测方法?

A Z-分数（Z-score）

B 四分位距（IQR）

C DBSCAN算法

D 马哈拉诺比斯距离（Mahalanobis Distance）

题目

在信用卡欺诈检测场景中，某用户平时在同一城市消费，某日突然出现海外高额消费记录。从异常值分析角度，最合适的描述是？

A 这是全局异常值，因为金额过高

B 这是上下文异常值，因为结合用户历史行为才判断为异常

C 这是集体异常值，因为多笔交易同时出现

D 这不是异常值，因为海外消费在技术上是合法的

题目

研究人员发现数据集明显不符合正态分布（严重右偏）。在检测异常值时，以下哪种方法最为合适?

A Z-分数方法，因为它最常用

B IQR方法，因为它不依赖分布假设

C 仅凭直方图目视判断即可

D 由于数据非正态，无法检测异常值

题目

使用主成分分析（PCA）检测异常值时，其核心判断依据是什么？

A 数据点离平均值的绝对距离

B 数据点在第一主成分方向上的投影值

C 降维后数据点的重构误差

D 数据点所属聚类的大小

题目

在营销数据分析中发现某VIP客户的购买金额是普通客户均值的50倍。以下处理方式中，最不恰当的是？

A 将其作为单独类别进行分析

B 直接删除该数据，避免影响模型

C 针对该类客户制定特别营销策略

D 分析该异常值背后的业务含义

题目

下列关于异常值处理注意事项的描述，哪项最能体现科学严谨的处理态度？

A 发现异常值后应立即删除，以保证数据整洁

B 应结合领域知识，理解异常值成因，记录处理过程并进行敏感度分析

C 异常值比例低于5%时可以忽略不处理

D 用中位数替换所有异常值是最安全的方法

数据预处理

异常值处理（Outlier Handling）

理解度测验 · 20题 · 四选一

授课教师：李阳源 · 2026-03-02 · 第4章

考试范围：SL01–SL18 | 总分：100分（每题5分）

题目

下列哪项不属于异常值（Outlier）的定义特征?

A 显著偏离整体数据模式的值

B 超出正常范围的观测值

C 与其他观测值特性明显不同的数据点

D 数据集中出现频率最高的值

题目

在Z-分数方法中，通常将哪个阈值作为判定异常值的标准?

A |Z| > 1

B |Z| > 2

C |Z| > 3

D |Z| > 5

题目

给定数据集 [10, 12, 15, 18, 20, 22, 25, 28, 30, 100]，其IQR（四分位距）值为多少?

A 13

B 15

C 18

D 21

题目

使用IQR方法，数据集 [10, 12, 15, 18, 20, 22, 25, 28, 30, 100] 的异常值判断上界为?

（ Q1=15, Q3=28, IQR=13 ）

A 43.0

B 47.5

C 50.0

D 55.5

题目

'冬季出现异常高温'属于哪种类型的异常值?

A 全局异常值（Global Outliers）

B 上下文异常值（Contextual Outliers）

C 集体异常值（Collective Outliers）

D 系统性异常值（Systematic Outliers）

题目

心电图中某段连续波形，单独看每个数据点都属于正常范围，但组合在一起却表现异常。这是哪种异常值类型?

A 全局异常值（Global Outliers ）

B 上下文异常值（Contextual Outliers ）

C 集体异常值（Collective Outliers ）

D 随机噪声（Random Noise ）

题目

关于Z-分数（Z-score）方法的局限性，以下哪项描述是正确的?

A 计算过于复杂，不易理解

B 无法处理正态分布数据

C 需要正态分布假设，且受多重异常值影响

D 只能用于单变量数据分析

题目

与Z-分数方法相比，IQR方法最主要的优点是什么？

A 计算结果更精确

B 不需要对数据分布做假设，具有鲁棒性

C 可以处理多维数据

D 适用于时间序列数据

题目

数据集 [10, 12, 15, 18, 20, 22, 25, 28, 30, 100] 中，值100的Z分数约为2.905，未超过3。根据Z-分数方法，以下判断正确的是?

（μ=28, σ=24.78）

A 100是异常值，因为它明显偏高

B 100不是异常值，因为Z分数未超过3

C 无法判断，需要更多数据

D 100是临界值，需要人工判断

题目

一家工厂的传感器监控系统中，某时刻同时出现温度、振动、电流均略高于正常值但未超过单项阈值的情况。从异常值类型角度，这最可能属于？

A 全局异常值

B 上下文异常值

C 集体异常值

D 测量误差

题目

在箱线图（Box Plot）中，须（Whisker）延伸至的位置表示什么?

A 数据的最大值和最小值

B 平均值±标准差的范围

C IQR方法中的上下限边界

D 数据的95%置信区间

题目

在哪种情况下，删除（Deletion）是处理异常值最合适的方法?

A 数据量较少且异常值原因不明

B 异常值明确是错误输入且数据量充足

C 异常值可能代表欺诈行为

D 数据集中异常值比例超过20%

题目

马哈拉**诺比斯距离（Mahalanobis Distance）**与普通**欧氏距离相比，最大的区别在于?**

A 计算速度更快

B 只能用于二维数据

C 考虑了变量间的相关性

D 不需要计算均值

题目

某研究员对一个样本量仅有15个的数据集使用IQR方法检测异常值，得到了5个异常值（占33%）。以下分析最合理的是？

A IQR方法完全可靠，结果有效

B 应考虑IQR在小样本中可靠性较低，结合领域知识验证

C 应该改用Z-分数方法重新计算

D 33%的异常值比例在正常范围内

题目

以下哪种方法不属于基于统计的异常值检测方法?

A Z-分数（Z-score）

B 四分位距（IQR）

C DBSCAN算法

D 马哈拉诺比斯距离（Mahalanobis Distance）

题目

在信用卡欺诈检测场景中，某用户平时在同一城市消费，某日突然出现海外高额消费记录。从异常值分析角度，最合适的描述是？

A 这是全局异常值，因为金额过高

B 这是上下文异常值，因为结合用户历史行为才判断为异常

C 这是集体异常值，因为多笔交易同时出现

D 这不是异常值，因为海外消费在技术上是合法的

题目

研究人**员发现数据集明显不符合正态分布（严重右偏）。在检测异常值时，以下哪种方法最为合适?**

A Z-分数方法，因为它最常用

B IQR方法，因为它不依赖分布假设

C 仅凭直方图目视判断即可

D 由于数据非正态，无法检测异常值

题目

使用主成分分析（PCA）检测异常值时，其核心判断依据是什么？

A 数据点离平均值的绝对距离

B 数据点在第一主成分方向上的投影值

C 降维后数据点的重构误差

D 数据点所属聚类的大小

题目

在营销数据分析中发现某VIP客户的购买金额是普通客户均值的50倍。以下处理方式中，最不恰当的是？

A 将其作为单独类别进行分析

B 直接删除该数据，避免影响模型

C 针对该类客户制定特别营销策略

D 分析该异常值背后的业务含义

题目

下列关于异常值处理注意事项的描述，哪项最能体现科学严谨的处理态度？

A 发现异常值后应立即删除，以保证数据整洁

B 应结合领域知识，理解异常值成因，记录处理过程并进行敏感度分析

C 异常值比例低于5%时可以忽略不处理

D 用中位数替换所有异常值是最安全的方法

SL20: 总结与下节课预告



总结:

- 异常值是偏离正常模式的数据点
- 主要检测方法: Z-分数、IQR、可视化方法、多变量方法
- 异常值处理: 删除、替换、转换、单独处理
- 根据领域知识和异常值原因选择适当处理方法

下节课预告:

- 主题: 数据归一化与标准化 (Data Normalization and Standardization)
- 内容: Min-Max归一化、Z-分数标准化、对数转换等

问答环节

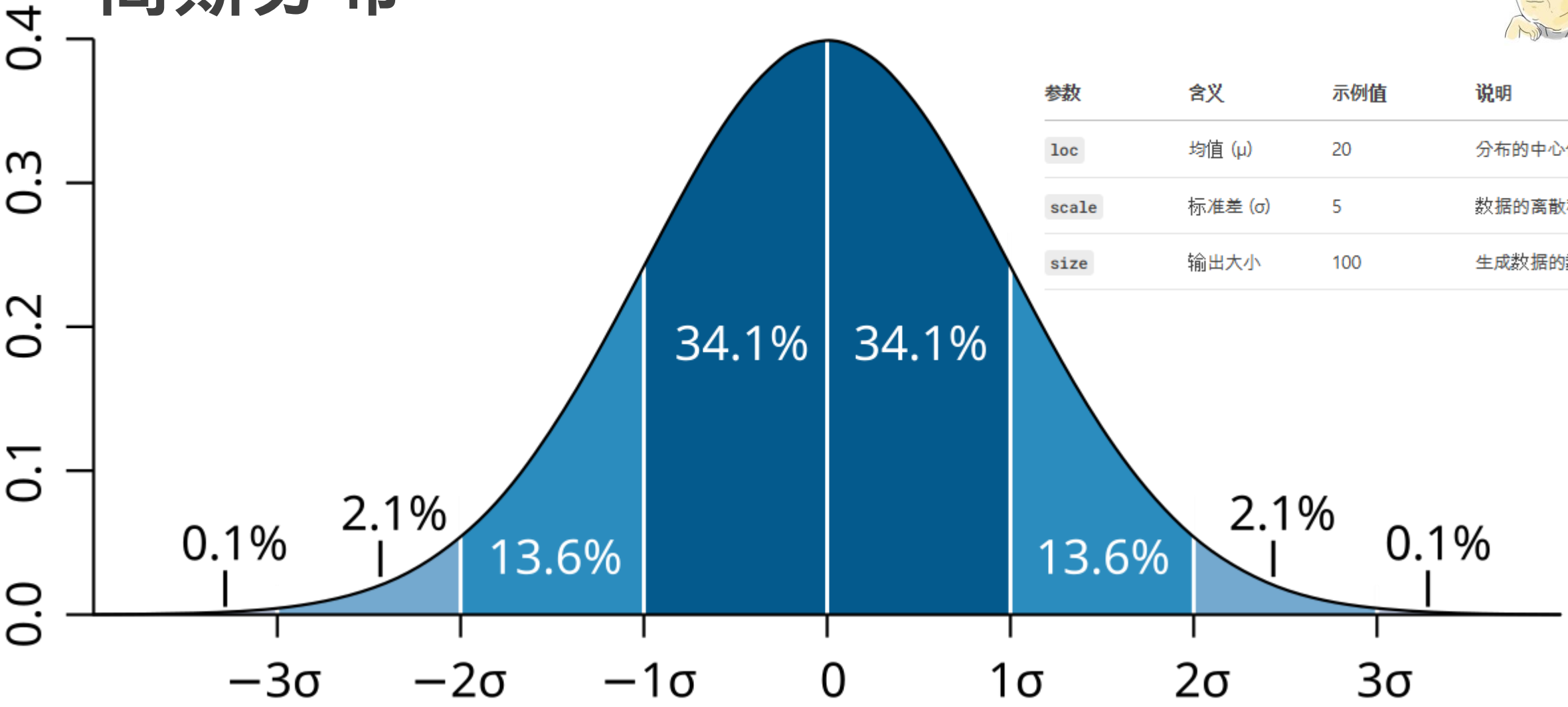
异常值处理实践



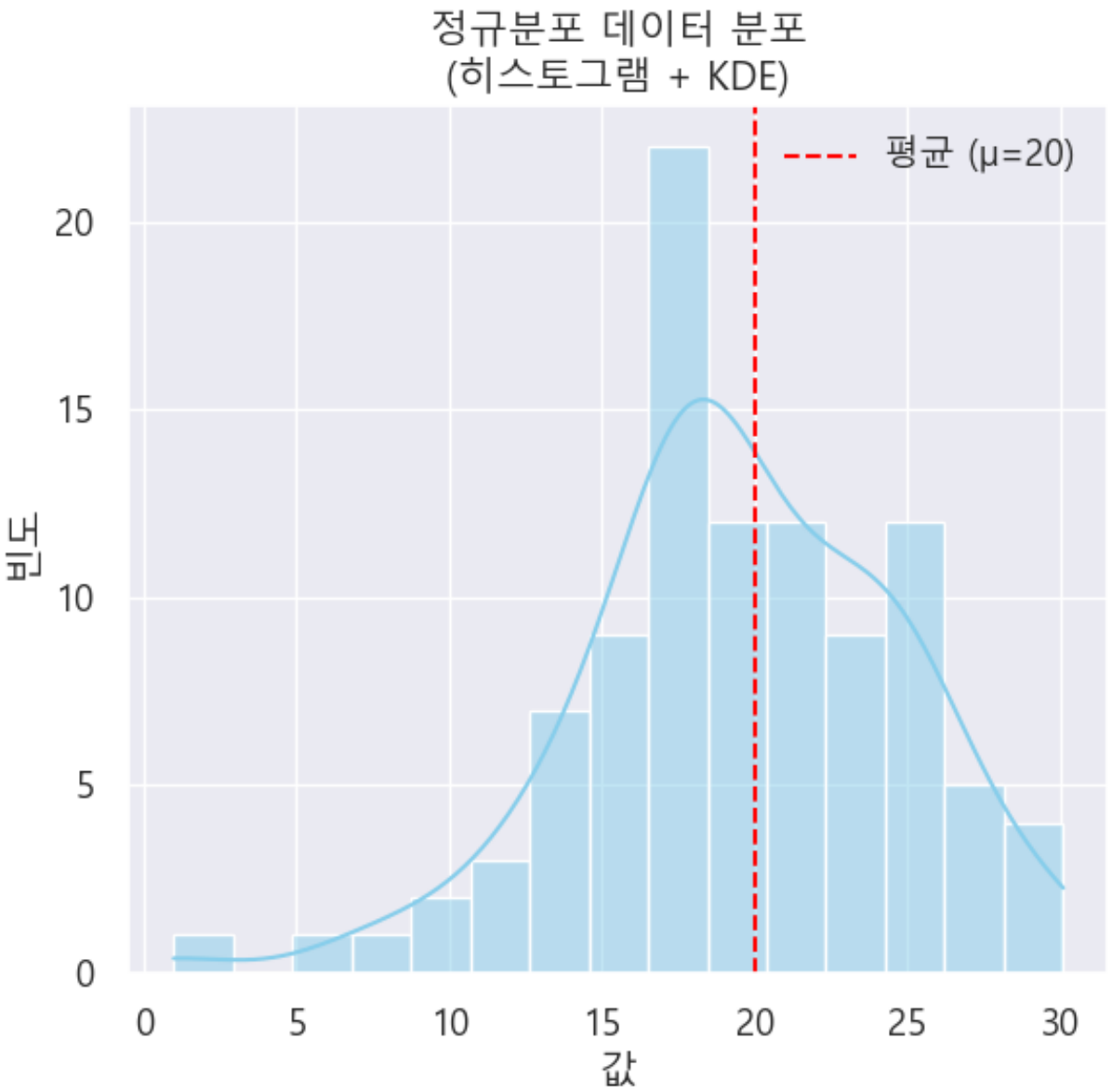
1. 数据集准备与探索
2. 使用Z-分数方法检测异常值
3. 使用IQR方法检测异常值
4. 通过可视化分析异常值
5. 应用异常值处理方法
6. 处理前后比较分析
7. 实践任务

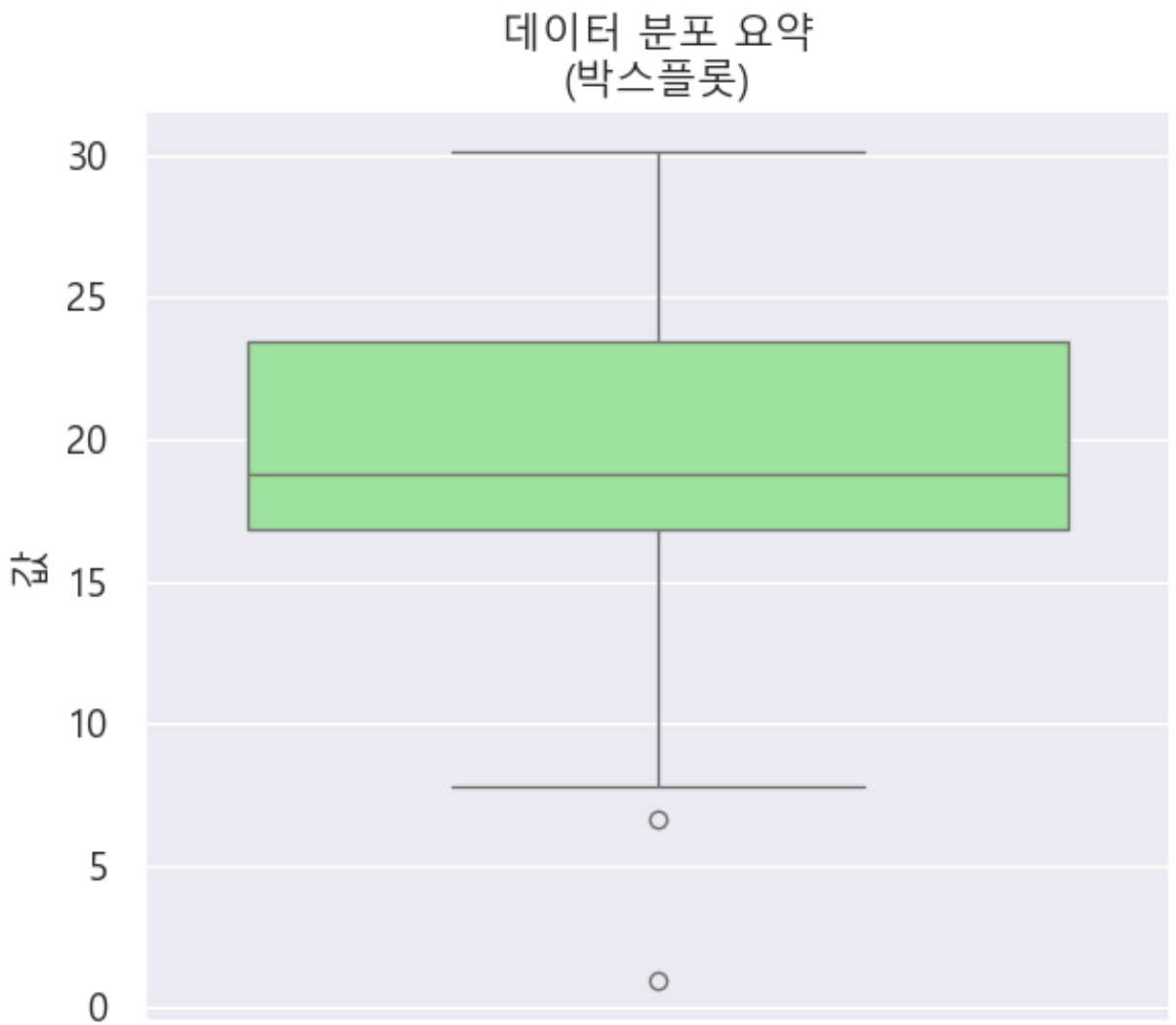
高斯分布

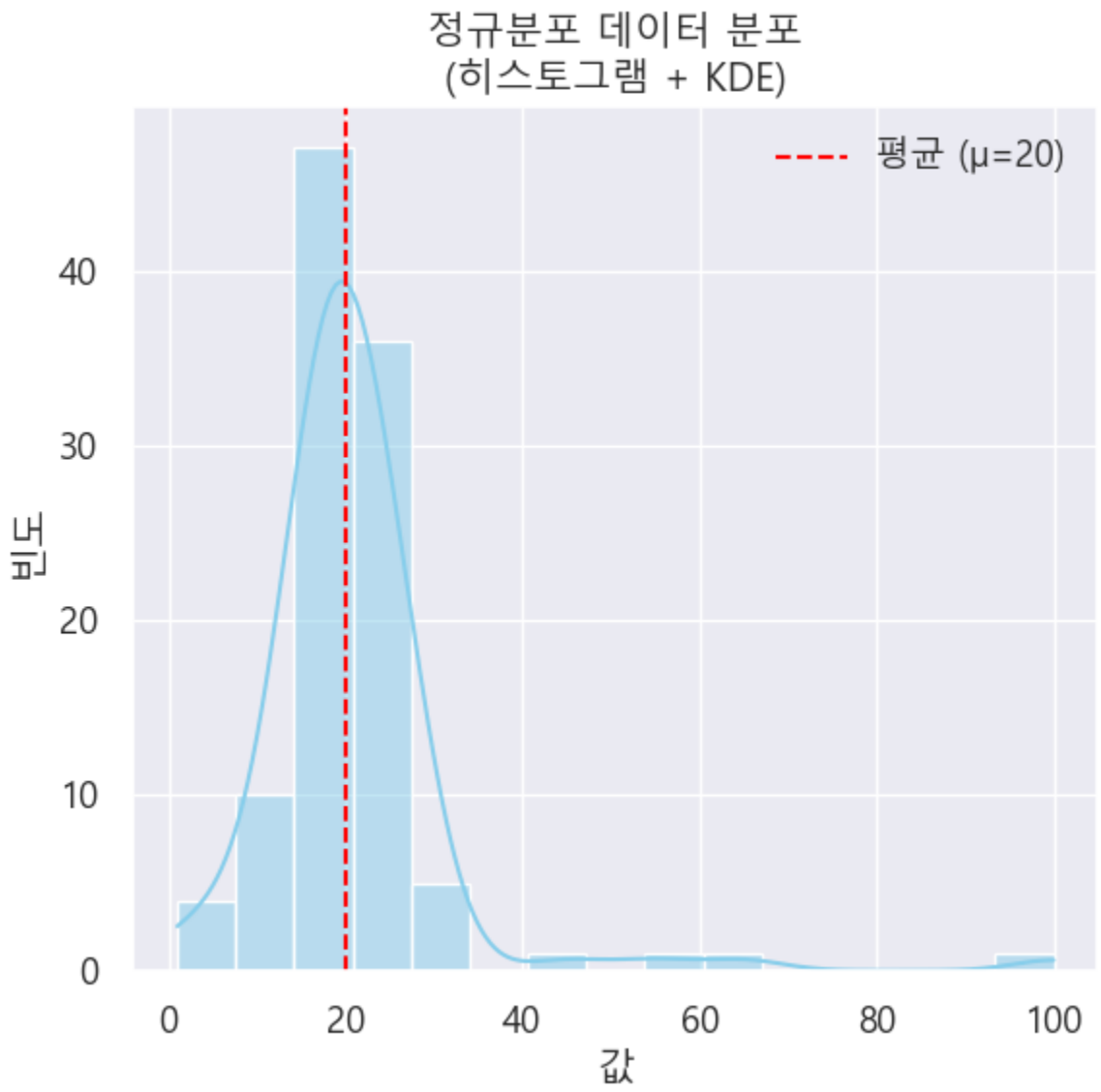


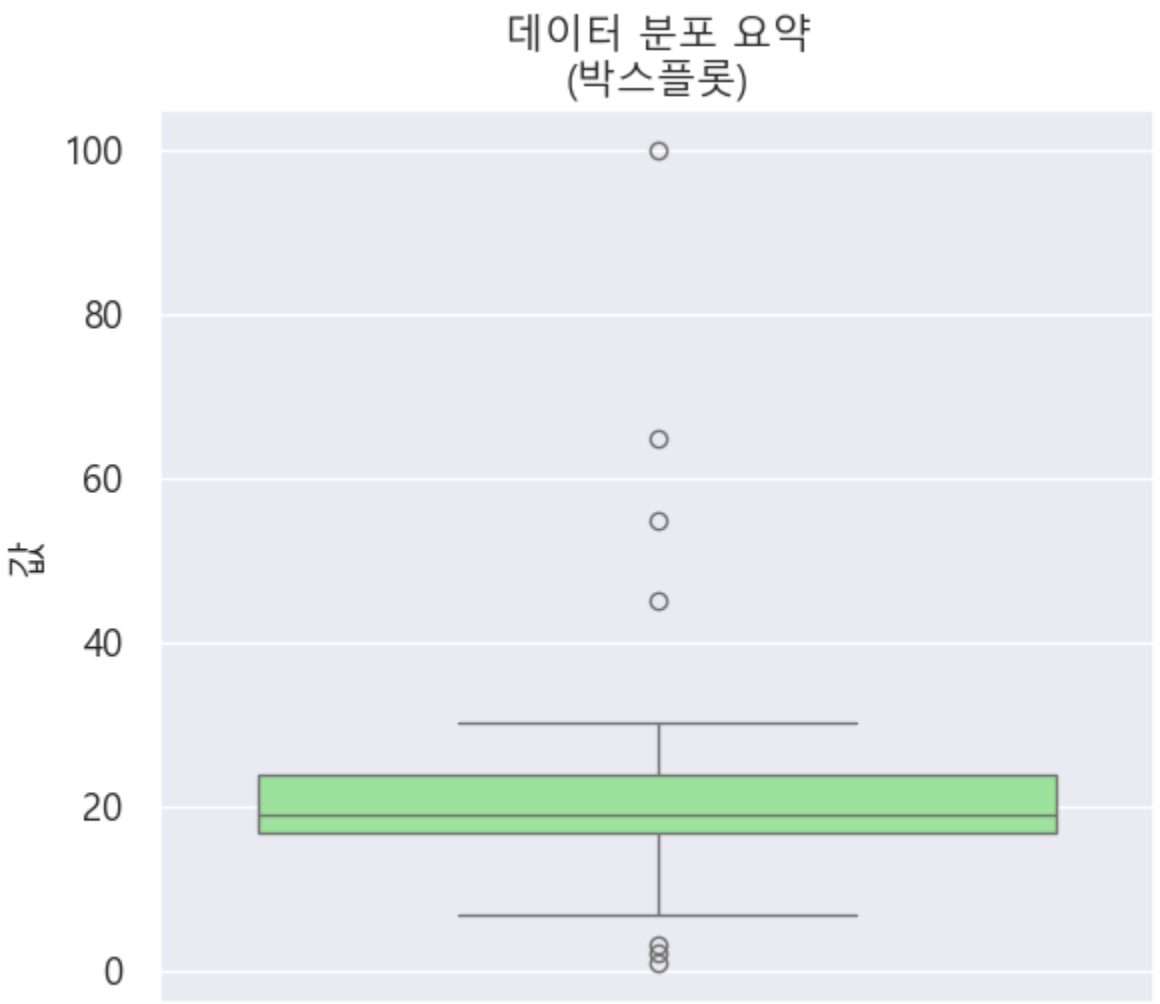


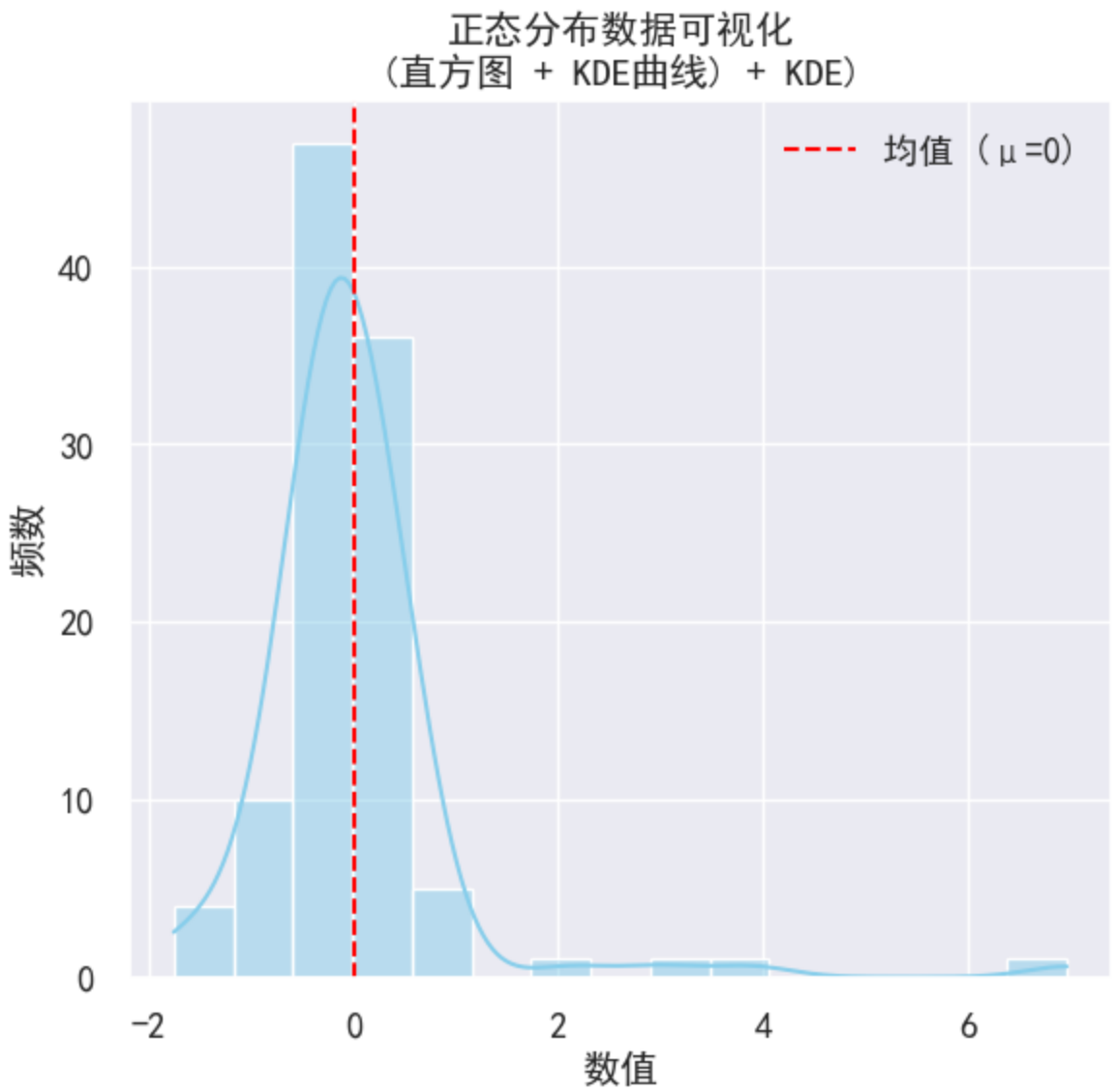
参数	含义	示例值	说明
`loc`	均值 (μ)	20	分布的中心位置
`scale`	标准差 (σ)	5	数据的离散程度
`size`	输出大小	100	生成数据的数量

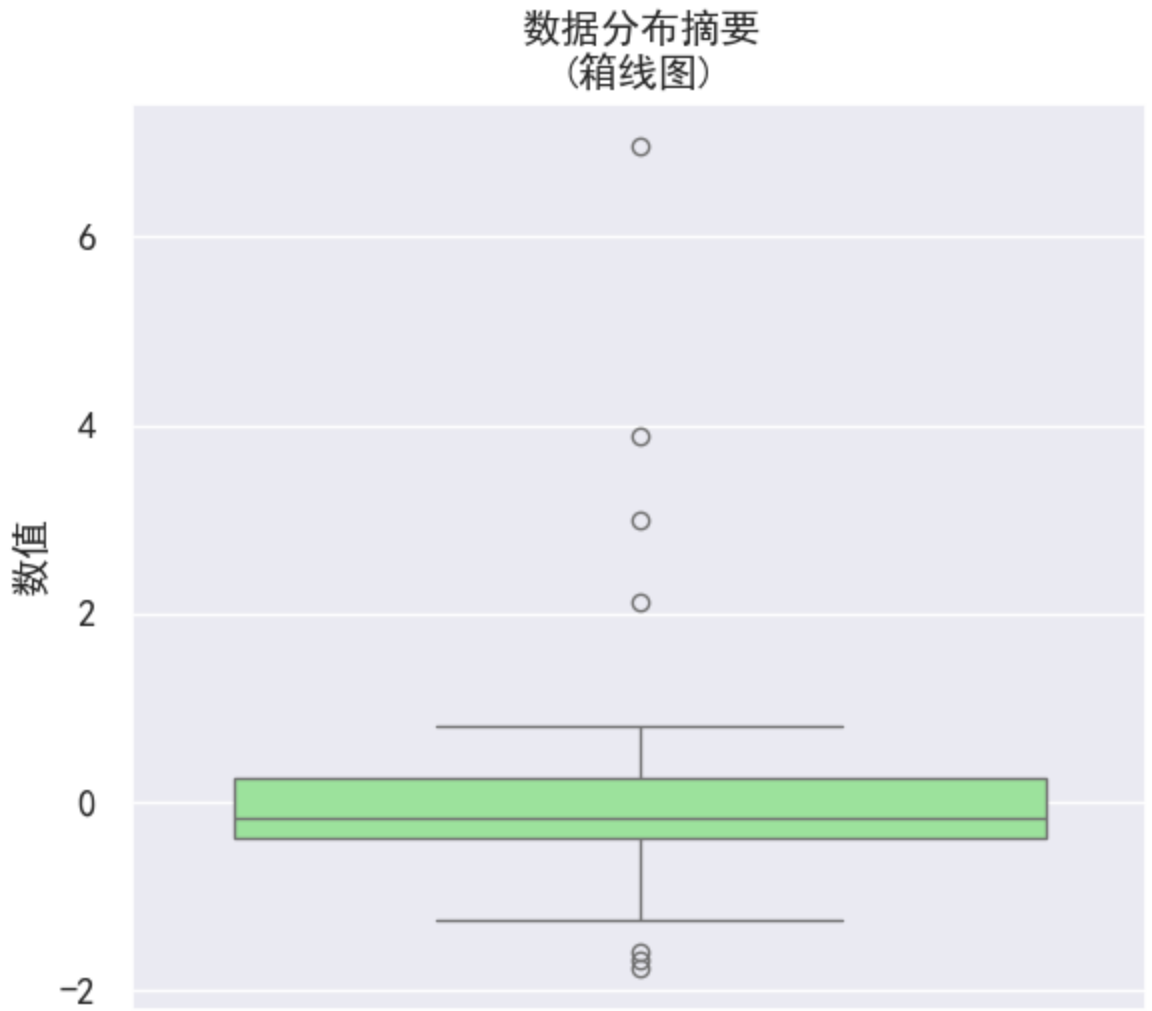

정규분포 데이터 분포
(히스토그램 + KDE)
평균 (μ=20)
빈도
값
0
5
10
15
20
25
30


데이터 분포 요약
(박스플롯)
값
0
5
10
15
20
25
30

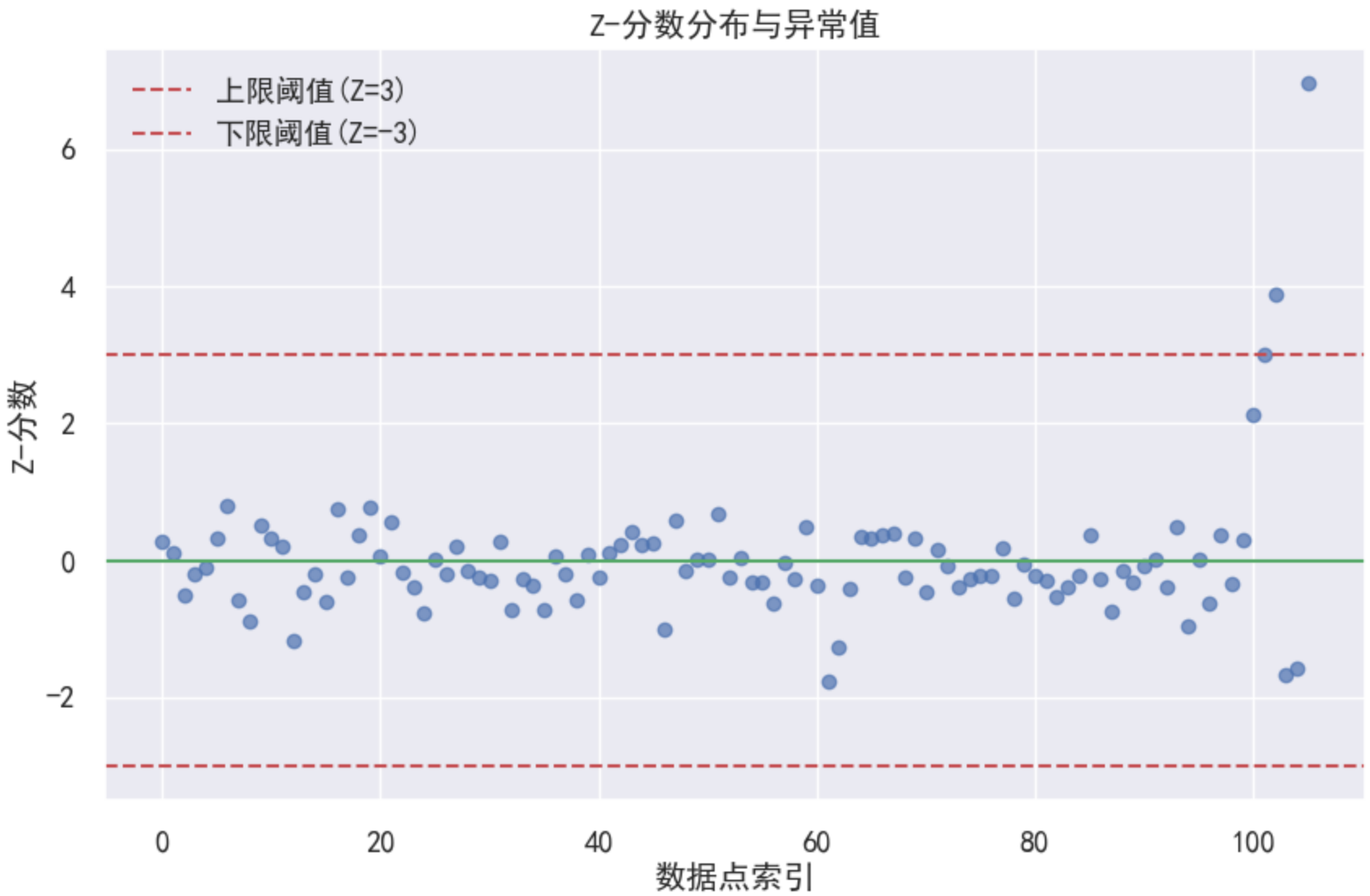

정규분포 데이터 분포
(히스토그램 + KDE)
평균 (μ=20)
빈도
값
0
10
20
30
40
0
20
40
60
80
100


데이터 분포 요약
(박스플롯)
값
0
20
40
60
80
100


正态分布数据可视化
（直方图 + KDE曲线） + KDE）
均值（μ=0）
频数
数值
0
10
20
30
40
−2
0
2
4
6


数据分布摘要
（箱线图）
数值
−2
0
2
4
6




Z-分数分布与异常值
上限阈值(Z=3)
下限阈值(Z=-3)
Z-分数
6
4
2
0
-2
0
20
40
60
80
100
数据点索引