

应用数理统计方法



第一章 总体特征描述及数据预处理

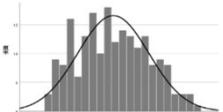
- 1.1 基本概念与采样方法
应用举例
- 1.2 总体特征及其表述
应用举例
- 1.3 数据预处理
应用举例

1.2.3

- 1.2.3 总体分布特征的统计表述
- 1.2.4 总体大小特征的统计表述
- 1.2.5 总体离散特征的统计表述
- 1.2.6 描述统计量的置信区间

随机变量的主要特征 复习 p.20

- **随机变量的特征**
 - 个体取值具随机性, 全部个体取值则有特定规律
 - 数理统计方法即研究这些规律的宏观特征
 - 许多学科的主要研究对象是随机变量 环境, 生态, 地学, 生物, 化学, 社会, 经济 ...
- **统计特征**
 - 分布 个体在不同取值区间出现的概率
 - 大小 分布中心取值 – 在数轴上的位置
 - 离散 个体对分布中心的偏离程度 – 聚集/分散



统计推断 复习 p.21

- **统计推断**
 - 经典数理统计的基本方法
 - 在概率论的基础上, 通过观察样本, 对总体特征做出判断
 - 包括参数估值和假设检验
- **参数估值**
 - 描述总体统计特征, 即根据样本统计量估计总体统计量
- **假设检验**
 - 据样本观测结果, 对针对总体的假设成立与否作出判断

参数估值 p.21

- **两类统计量**
 - 描述统计量 用于参数估值, 描述对象特征
如 算术均值, 标准差等
 - 检验统计量 用于假设检验, 计算判断的正确概率
如 异常值剔除中的 G, t-检验中的 t 等
- **参数估值**
 - 根据样本统计量估计总体统计量 如用 \bar{x} 估计 μ
 - 样本统计量 = 样本参数
 - 总体统计量 = 总体参数
- **参数方法与非参数方法**
 - 是否依赖总体参数 μ

参数估值方法 p.21

- **针对正态与非正态分布总体用不同方法**
- **正态分布总体**

| | |
|---------|------------|
| 大小与离散表征 | 算术均值, 标准差等 |
| 分布表征 | 偏度-峰态系数 |
- **非正态分布总体**

| | | |
|-------------|--------------------|--------|
| 可变换为正态分布的总体 | 几何均值等 | |
| 总体特征 | 中位数, 分位数, 半内四分范围 等 | |
| 特征 | 正态分布 | 从正态分布 |
| 大小 | 算术均值 | 中位数 |
| 离散 | 方差, 标准差 等 | 半内四分范围 |
| 分布 | 偏度系数, 峰态系数 | 分位数 |

随机变量的分布特征

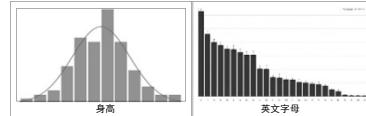
p.20

总体分布特征

个体在可能的取值范围内随机出现，且有特定概率。如黑白球、PM_{2.5}浓度。
分布特征：全部个体在可能取值范围内出现概率的分布。

样本分布特征

代表性样本表现出与总体类似的分布特征
通过对样本分布特征的描述判断总体分布特征
例 人群身高分布、文献中英文字母分布、鱼重量分布、鱼种类分布



分组统计

p.i-v

分组统计 - 构建数据的频数/频率分布图

将取值范围分为若干组，组距相等或不等。参见3.4.1 Kolmogorov检验。

分别统计各组频数 F 、样本数 或频率 f = 频数 / 总样本量。

结果表达为频数/频率分布图。横坐标为分组取值范围，纵坐标为频数或频率。

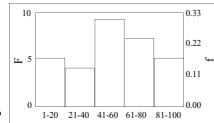
分组统计举例

数据 72, 6, 80, 7, 100, 77, 42, 77, 24, 10, 76, 57, 45, 81, 37, 28, 63, 48, 83, 98, 24, 41, 54, 59, 45, 41, 76, 18, 90, 12

分组 6, 7, 10, 12, 18, 24, 24, 28, 37, 41, 41, 42, 45, 45, 48, 54, 57, 59, 63, 72, 76, 76, 77, 77, 80, 81, 83, 90, 98, 100

统计 $F_i = 5, 4, 9, 7, 5 \quad \sum F_i = n = 30$

$f_i = 0.17, 0.13, 0.30, 0.23, 0.17 \quad \sum f_i = 1$



讨论

如果每个数据自成一组，会是什么样的分布？

不同属性变量的频率/频数分布

p.31

变量属性 - 按观测水平分类

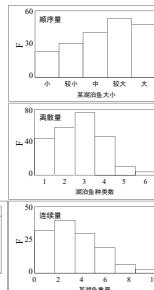
定量变量 连续量、离散量、顺序量 定性变量 类型量

定量变量的频数分布图

包含次序信息，按取值大小分组。

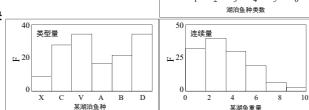
特定检验中可利用次序信息。

参见3.4.1 Kolmogorov检验。



定性变量的频数分布图

无次序关系，不同类型可任意互换。



连续变量的概率密度函数

p.31

从频率分布到概率密度函数

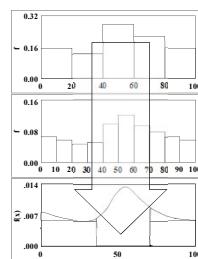
组距趋向零，组数趋向无穷，频率分布趋向概率密度函数。

概率密度函数取值范围内曲线下方总面积为1。

频率分布各组加和为1。

特定区间对应特定取值概率。

假设检验的判断依据。



理论分布与经验分布

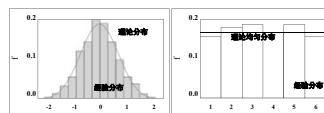
p.30

理论分布

符合特定理论的分布。

如：掷无穷多次完美骰子的理论均匀分布、测量误差的理论正态分布。

二项分布、对数正态分布、学生t-分布、F-分布等。



经验分布

实际观测数据的分布。

如：掷真实骰子的分布。

某测定仪器的误差分布。

特定人群的性别分布。

气溶胶的粒径分布等。

正态分布

p.33

正态分布 - 高斯分布

数理统计中最重要的连续变量理论分布。

由均值、分布中心 和标准差 分布形状 两个参数定义。

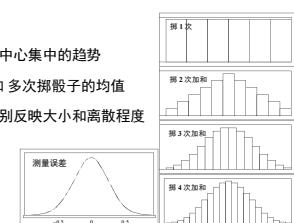
标准正态分布 均值为0，标准差为1的正态分布 参见 1.3.3 标准化。

特点与重要性

集中性与对称性：有对称地向分布中心集中的趋势。

任意分布的均值趋向正态分布，如多次掷骰子的均值。

参数方法的基础：均值和标准差分别反映大小和离散程度。



分位数

p.28-29

▪ 分位数

将所有个体分为若干个等份的分点 表现个体在不同取值范围的分布特点
个别数据的极端取值 如最大值或异常值 对多数分位数没有影响

▪ 常用分位数

二分位数 中位数

四分位数 将全部个体分为 4 个等样本量部分的 3 个分点

百分位数 将全部个体分为 100 个等样本量部分的 99 个分点

特点 反映大小、离散、偏斜等；不受异常影响 – 稳健



百分位数

p.28

▪ 百分位数

第 i 个百分位数记为 p_i

如 p_{25}, p_{50}, p_{75} 分别为第 25, 第 50 和第 75 个百分位数

▪ 计算

第 i 个百分位数

求 $p_i = i\%(n + 1)$, $p_i \leq 1$, $x_{pi} = x_1$

$p_i \geq n$, $x_{pi} = x_n$

$p_i > 1$ 且 $p_i < n$, 线性插值



百分位数的应用

p.28

▪ 百分位数计算

样本量不足 100 时同样可计算和使用百分位数 线性插值

▪ 百分位数选取

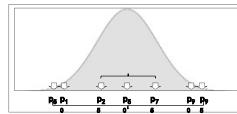
实际使用不需要 99 个百分位数 不是设计百分位数的目的

选取不同百分位数概括表现整体分布特征 针对向分布中心集中的趋势

分点描述 $p_5, p_{10}, p_{25}, p_{50}, p_{75}, p_{90}, p_{95}$

范围描述 50% $p_{25} - p_{75}$, 90% $p_5 - p_{95}$

用于不确定性表征



1.2.4

1.2.3 总体分布特征的统计表述

1.2.4 总体大小特征的统计表述

1.2.5 总体离散特征的统计表述

1.2.6 描述统计量的置信区间

总体大小表征

p.36-38

▪ 中心趋势

总体中的个体取值有向分布中心集中的趋势

▪ 大小表征

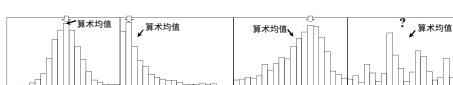
对分布中心位置的估计 与总体分布特征有关

正态分布 算术均值 $\bar{x} = \sum x_i / n$, $\mu = \sum x_i / N$ 样本与总体算术均值

对数正态分布 几何均值 $x_g = e^{(1/n)\sum(\ln x_i)}$, $\mu_g = e^{(1/N)\sum(\ln x_i)}$ 样本与总体几何均值

其他单峰分布 若可作正态变换, 类似几何均值; 否则用中位数

其他分布 中位数 $M = x_{0.5(n+1)}$, $M = (x_{0.5n} + x_{0.5n+1})/2$



1.2.5

1.2.3 总体分布特征的统计表述

1.2.4 总体大小特征的统计表述

1.2.5 总体离散特征的统计表述

1.2.6 描述统计量的置信区间

正态分布总体的离散表征

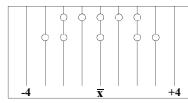
p.39-41

· 离散特征

个体距分布中心的远近

· 离散表征

正态分布 离差, 平方和, 方差, 标准差, 变异系数

**正态分布总体的离散表征**

p.39

· 离散特征

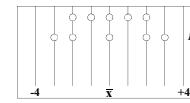
个体距分布中心的远近

· 离散表征

正态分布 离差, 平方和, 方差, 标准差, 变异系数

$$\sum |x_i - \bar{x}|$$

到均值距离的加和

**正态分布总体的离散表征**

p.39

· 离散特征

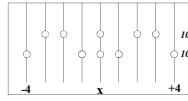
个体距分布中心的远近

· 离散表征

正态分布 离差, 平方和, 方差, 标准差, 变异系数

$$\sum |x_i - \bar{x}|$$

简单距离加和的问题: 近点与远点的差异

**正态分布总体的离散表征**

p.39-41

· 离散特征

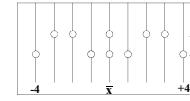
个体距分布中心的远近

· 离散表征

正态分布 离差, 平方和, 方差, 标准差, 变异系数

$$SS = \sum (x_i - \bar{x})^2$$

取平方强调远点, 类似空间插值中的距离平方反比

**正态分布总体的离散表征**

p.39-41

· 离散特征

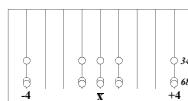
个体距分布中心的远近

· 离散表征

正态分布 离差, 平方和, 方差, 标准差, 变异系数

$$SS = \sum (x_i - \bar{x})^2$$

受样本量影响

**正态分布总体的离散表征**

p.39-41

· 离散特征

个体距分布中心的远近

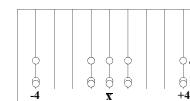
· 离散表征

正态分布 离差, 平方和, 方差, 标准差, 变异系数

$$s^2 = SS/(n-1)$$

除以自由度消除样本量影响

自由度 – 样本中独立或能自由变化的数据的个数 n-1

s² 和 σ² 分别代表样本和总体方差

正态分布总体的离散表征

p.39-41

▪ 离散特征

个体距分布中心的远近

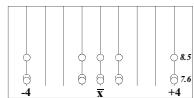
▪ 离散表征

正态分布 离差, 平方和, 方差, 标准差, 变异系数

$$s^2 = \frac{SS}{(n-1)}$$

取平方导致方差与算术均值尺度不同

不便同时表达大小与离散



正态分布总体的离散表征

p.39-41

▪ 离散特征

个体距分布中心的远近

▪ 离散表征

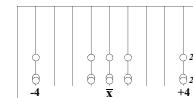
正态分布 离差, 平方和, 方差, 标准差, 变异系数

$$s = (s^2)^{0.5}$$

开方消除因平方导致的尺度变化

分别用 s 和 σ 代表样本和总体标准差

表达变量离散程度的常用参数



正态分布总体的离散表征

p.39-41

▪ 离散特征

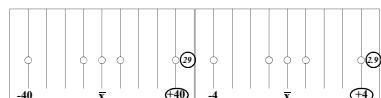
个体距分布中心的远近

▪ 离散表征

正态分布 离差, 平方和, 方差, 标准差, 变异系数

$$s = (s^2)^{0.5}$$

不同尺度变量的标准差不可直接对比



正态分布总体的离散表征

p.39-41

▪ 离散特征

个体距分布中心的远近

▪ 离散表征

正态分布 离差, 平方和, 方差, 标准差, 变异系数

$$V = 100 s/\bar{x}$$

除以均值以消除尺度差异影响

分别用 V 和 V_p 代表样本和总体变异系数

正态与非正态分布总体的离散表征

p.39-41

▪ 正态分布总体

正态分布 离差, 平方和, 方差, 标准差, 变异系数

对数正态分布 几何标准差 $S_g = e^{[1/(n-1) \sum (\ln x_i - \ln \bar{x})]^2/2}$

▪ 非正态分布总体

半内四分范围 $(P_{75} - P_{25})/2$

范围 $x_{\max} - x_{\min}$ 易受拖尾或异常值干扰, 不稳健

标准差与标准误差

p.39-41

▪ 标准差

表征个体对总体均值偏离的参数 表现离散程度

▪ 标准误差

表征抽样均值对总体均值偏离的参数 抽样均值的离散程度 – 估计的可靠性

标准误差 $SE = s/\sqrt{n}$

重复抽样样本均值的标准差

2 参数估值

6

1.2.6

- 1.2.3 总体分布特征的统计表述
- 1.2.4 总体大小特征的统计表述
- 1.2.5 总体离散特征的统计表述
- 1.2.6 描述统计量的置信区间

点估计与区间估计

p.42

▪ 参数估值

根据样本参数估计总体参数 计算样本统计量, 估计总体统计量

完整表达为: 点估计+区间估计

▪ 点估计

直接用样本参数估计对应的总体参数 特定条件下的最佳估计

如 用样本算术均值估计总体算术均值, 用样本相关系数估计总体相关系数

▪ 区间估计

特定概率条件下总体参数值的可能范围 估计的可靠性描述

如 总体均值的置信区间

如 回归方程的总体斜率, 截距, 预测值的置信区间



置信区间

p.43

▪ 置信区间的一般表达

一个区间及该区间包括待估总体参数的概率

$$P\{L_1 \leq \text{总体参数} \leq L_2\} = p$$

L_1, L_2 为该区间的上下界, p 为该区间覆盖总体参数的概率

▪ 例 算术均值的置信区间

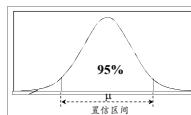
$$P\{L_1 \leq \mu \leq L_2\} = p \quad L_1 = \bar{x} - t_{(n-1)} s / n^{0.5}, \quad L_2 = \bar{x} + t_{(n-1)} s / n^{0.5}$$

如 $P\{5.5 \leq \mu \leq 6.7\} = 95\%$ 范围 $5.5-6.7$ 包含总体算术均值的概率为 95%

▪ 应用

点估计的可靠性表达 即区间估计

据预设可靠性估算样本量 见上式, 见下节举例



Box-Whisker 图

p.42

▪ 关键统计量的图示表达

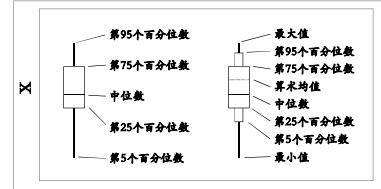
根据需要选择参数 均值, 中位数, 百分位数 等

直观表达多重特征 表现大小, 离散, 分布 等

直观比较不同总体的差异 比较小, 离散, 分布 等

▪ 讨论

如何读图?



应用数理统计方法

- #### 第一章 总体特征描述及数据预处理
- 1.1 基本概念与采样方法
应用举例
 - 1.2 总体特征及其表述
应用举例
 - 1.3 数据预处理
应用举例

应用实例 城市空气质量表征

<https://quotsoft.net/air/>

▪ 数据: 城市空气质量监测平台 PM2.5 浓度

21 个站点, 2015 年 1 月 17 日小时均值

▪ 基本统计量

列举最基本的样本统计量

包括: 最小值, 最大值, 均值, 标准差, 中位数, 百分位数 $p_5, p_{10}, p_{25}, p_{50}, p_{75}, p_{90}, p_{95}$

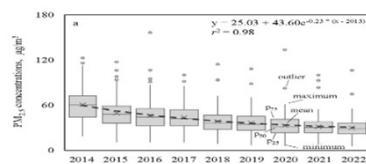
| 极值 | 0.01A | 10.0A | 150.0A | 174.0A | 228.0A | 223.7A | 229.8A | 251.7A | 255.2A | 258.7A | 263.2A | 263.6A | 265.4A | 267.5A | 268.4A | 268.7A | 269.3A | 269.8A | 270.7A | 271.1A |
|---------|-------|-------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 最小值 | 11.0 | 44.0 | 46.0 | 48.0 | 116.0 | 19.0 | 53.0 | 75.0 | 107.0 | 20.0 | 3.0 | 50.0 | 36.0 | 4.0 | 19.0 | 62.0 | 2.0 | 4.0 | 49.0 | 1.0 |
| 最大值 | 163.3 | 99.0 | 111.0 | 112.0 | 213.0 | 43.0 | 209.0 | 219.0 | 329.0 | 94.0 | 28.0 | 201.0 | 149.0 | 29.0 | 85.0 | 418.0 | 44.0 | 264.0 | 122.0 | 17.0 |
| 均值 | 68.1 | 66.0 | 75.1 | 80.8 | 162.3 | 30.1 | 111.7 | 117.7 | 208.7 | 72.4 | 9.8 | 101.5 | 62.0 | 14.0 | 40.5 | 178.3 | 22.0 | 110.8 | 83.0 | 6.1 |
| 标准差 | 55.1 | 15.7 | 17.1 | 20.3 | 103.5 | 38.7 | 103.5 | 103.5 | 219.9 | 7.6 | 41.9 | 21.1 | 7.9 | 14.3 | 100.7 | 11.0 | 61.9 | 23.2 | 4.2 | |
| 中位数 | 10.0 | 15.3 | 15.5 | 16.5 | 34.0 | 10.0 | 121.0 | 125.5 | 163.0 | 10.0 | 10.0 | 14.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 | |
| 第10百分位数 | 12.2 | 46.5 | 51.9 | 48.9 | 117.2 | 20.0 | 53.3 | 79.3 | 109.8 | 29.9 | 3.0 | 50.6 | 36.2 | 4.1 | 23.2 | 72.5 | 2.0 | 17.9 | 51.4 | 1.1 |
| 第25百分位数 | 13.0 | 50.5 | 57.3 | 54.3 | 124.6 | 20.0 | 58.6 | 81.3 | 126.6 | 36.5 | 3.0 | 54.9 | 37.0 | 5.2 | 24.3 | 82.8 | 4.4 | 59.7 | 59.3 | 1.1 |
| 第50百分位数 | 16.0 | 54.0 | 64.5 | 65.8 | 138.0 | 22.0 | 93.8 | 88.3 | 172.0 | 66.0 | 4.0 | 69.0 | 44.5 | 7.0 | 28.3 | 95.3 | 14.8 | 79.8 | 62.0 | 2.3 |
| 第75百分位数 | 19.0 | 59.0 | 70.5 | 84.5 | 160.5 | 34.0 | 111.3 | 121.0 | 202.5 | 79.0 | 6.5 | 99.0 | 52.0 | 14.0 | 40.0 | 138.0 | 20.0 | 96.0 | 87.5 | 5.5 |
| 第90百分位数 | 116.0 | 76.8 | 88.3 | 95.0 | 188.0 | 36.0 | 127.8 | 129.8 | 260.3 | 87.8 | 14.3 | 131.5 | 71.5 | 19.0 | 46.0 | 240.5 | 29.3 | 144.3 | 100.1 | 8.8 |
| 第95百分位数 | 135.7 | 88.1 | 98.6 | 109.2 | 207.0 | 38.0 | 143.4 | 143.4 | 279.2 | 92.0 | 22.6 | 152.4 | 92.0 | 26.6 | 52.8 | 309.3 | 37.3 | 208.3 | 111.0 | 12.0 |
| 第99百分位数 | 186.3 | 95.0 | 101.6 | 111.9 | 209.6 | 40.6 | 188.5 | 176.3 | 298.0 | 92.0 | 24.7 | 165.8 | 108.2 | 27.9 | 59.0 | 361.2 | 40.6 | 215.9 | 120.4 | 13.0 |

应用实例 城市空气质量表征<https://quotsoft.net/air/>**▪ 数据: 城市空气质量监测平台 PM2.5 浓度**

2014 – 2022年全国监测点日均浓度

▪ 基本统计量图示

直观表达与比较

包括: 均值, 中位数, 最小值, 最大值, 百分位数 $p_{10}, p_{25}, p_{50}, p_{75}, p_{90}$ **应用实例 城市空气质量表征**<https://quotsoft.net/air/>**▪ 数据: 城市空气质量监测平台 PM2.5 浓度**

1,494 个站点, 三天日均值, 原始数据与对数变换数据

▪ 频数分布图与 Box-Cox 图

原始数据右偏, 对数变换数据左偏

幂变换? 谨慎使用, 可比性差

未检出数据影响

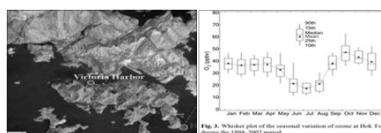
**应用实例 大气臭氧浓度的季节变化**

Wang et al., ACP 2009

▪ 香港鹤嘴大气臭氧浓度的季节变化

2007 – 2014 连续监测数据

研究季节变化趋势

▪ 动态变化的基本统计量图示Box-Cox图 算术均值, 百分位数 $p_{10}, p_{25}, p_{50}, p_{75}, p_{90}$
显著的季节变化 大小, 离散, 分布**应用实例 室内空气颗粒物浓度差异**

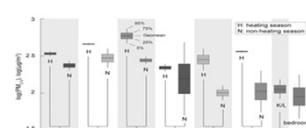
Chen et al., Environ. Int., 2018

▪ 农村室内空气 PM2.5 浓度统计数据

区分不同燃料 煤, 荚秆, 薪柴, 清洁燃料

区分不同季节 暖气期, 非暖气期

区分不同房间 卧室, 厨房, 客厅

▪ 基本统计量对数正态分布 几何均值和百分位数 $p_5, p_{25}, p_{75}, p_{95}$
纵标取对数 大小, 离散, 偏斜**应用实例 新冠潜伏期分布**

Qun et al., NEJM 2020

▪ 新冠潜伏期的统计分布

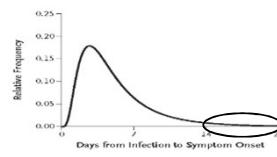
潜伏期: 从感染到出现症状天数

频率分布图 略尖峰, 明显拖尾

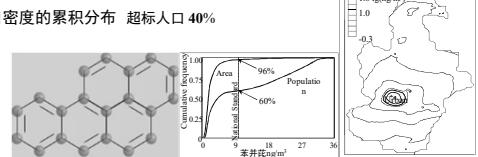
▪ 应用

传播模拟的重要参数

隔离期选择的参照依据 两周, 三周?

**应用实例 天津大气苯并芘超标评估**

Li et al., ES&T, 2006

▪ 天津市苯并芘呼吸暴露模拟近地面大气苯并芘浓度的空间分布 排放清单, 气象参数与大气传输模拟
近地面大气苯并芘浓度的累积频率分布 面积或人口**▪ 暴露风险分析**室外空气质量标准 $10/2.5 \text{ ng/m}^3$
按面积的累积分布 超标面积 4%
按人口密度的累积分布 超标人口 40%

应用实例 机动车污染物排放因子分布特征

· 机动车排放因子

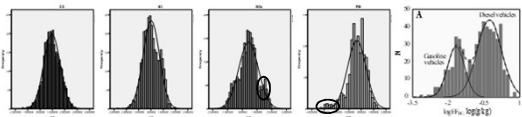
排放因子 单位质量燃料消耗导致的污染物排放量 CO, HC, NOx, PM_{2.5}, BC

· 观测排放因子的分布特征

前四种: 对数正态分布 局部偏倚

黑炭: 对数变换后双峰分布 柴油车/汽油车两总体,

PeakFit拆分两个对数正态分布总体



应用实例 不确定性表征

Tao et al., 2003 ES&T

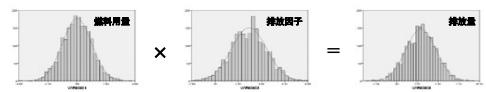
· 蒙特卡洛模拟

基于随机抽样的统计实验方法 环境研究中用于不确定性分析

据已知概率分布随机抽样, 估计量排放量的分布

· 例: 排放量及不确定性估算

排放量 = 燃料用量 × 排放因子



应用实例 变异与不确定性拆分

Tao et al., 2003 ES&T

· 变异与真实不确定性

总不确定性包含真实不确定性与随机变异 Overall / true uncertainties, variability

· 例: 天津表土多环芳烃浓度

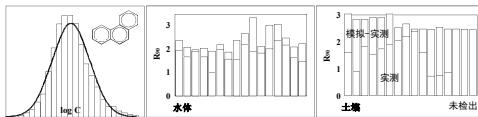
实测 采集188个表土样品, 实测16种多环芳烃浓度 – 对数正态分布

模拟 多介质模型计算16种多环芳烃介质浓度与界面通量 – 对数正态分布

模拟得到总不确定性, 观测提供空间变异 定义 $R_{90} = p_{90} - p_5$: 表现量级尺度差异

两者之差为真实不确定性

即量级尺度模型误差



应用实例 调查样本量预估

p.44

· 置信区间

置信区间计算依据: 标准差与样本量 $L_{1,2} = \bar{x} \pm t_{\alpha/2} s / n^{0.5}$

可根据置信区间和标准差反算样本量

· 样本量估算

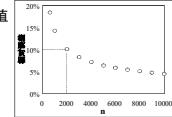
已知标准差 s , 主观设定最大误差 L_1, L_2 目标, 计算需要样本量

· 例: 农村燃料用量调查

通过预备性调查或文献资料估计均值与标准差

构建相对误差与样本量关系 相对误差=置信范围/均值

据采样成本确定样本量 如 10% 误差约需 2,000 样本



应用数理统计方法

第一章 总体特征描述及数据预处理

1.1 基本概念与采样方法

应用举例

1.2 总体特征及其表述

应用举例

1.3 数据预处理

应用举例



数据预处理的目的

p.48

· 检验和改善样本代表性

异常值剔除 排除异常值对统计分析结果的影响

独立性检验 检验个体采集过程是否受其它个体干扰

未检出数据 低于检出下限数据的处理

数据平滑 消除数据噪声

· 满足特定方法的要求

数据变换 如正态变换, 获得参数方法需要的正态分布数据

如求秩, 获得非参数方法需要的秩数据

1.3.2

1.3.2 异常值检验

1.3.3 数据变换

1.3.4 其它

异常值剔除

p.51

- 异常值

样本中出现概率很小的值 其存在可能影响统计分析结果

- 两类异常

统计异常 纯属偶然, 属于研究总体, 只是出现概率极低而已

非统计异常 采样失误, 不属于该总体

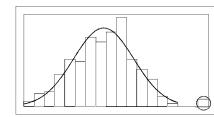
参数方法 两类异常都会干扰分析结果, 应剔除

非参数方法 不受异常值干扰

- 异常值剔除

正确剔除可以改善分析结果

主观随意剔除则可能导致非客观判断



异常值检验方法

p.52-54

- 简易方法

Panta, Chauvenet, 三倍标准差 无概率意义, 计算能力不足情况下使用

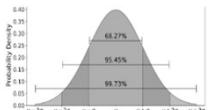
- 基于正态分布的方法

在特定概率基础上作出判断

Grubbs 检验 有确切错误率, 考虑样本量影响, 推荐方法

t-检验 比 Grubbs 法严格, 计算不包括可疑值

Dixon 检验 比 Grubbs 法严格, 需不同检验计算式



Grubbs 检验

p.52

- 方法

数据排序 从最可能是异常的最小和最大值, 即可疑值 X_i 开始

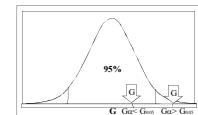
顺次检验 若剔除, 检验次大或次小值, 直至余下的最小和最大值都不是异常为止

- 计算

计算可疑值属于该总体的概率, 若小于特定概率如 5%, 则判定为异常值

或计算检验值 $G = |X_i - \bar{X}|/s$, 若大于临界值则判断为异常 临界值表 p.390

两种方法结果相同, 信息量不同 参见 2.1 假设检验



1.3.3

1.3.2 异常值检验

1.3.3 数据变换

1.3.4 其它

数据变换

p.57

- 目的

使数据满足特定统计方法的要求

- 类型变换

归类 将定量变量转换为类型变量 如 波动游程检验, 中位数检验

求秩 将连续量或离散量转换为顺序量 如 大多数非参数检验

- 算术变换

标准化 使均值为 0 标准差为 1 如 Grubbs 检验, 标准化多元回归

- 分布变换

正态化 使成为正态分布 如 对数变换, 角变换

类型变换

p.57

▪ 归类

将定量变量变换为类型变量 如波动游程检验中的正负号
由于放弃了定量变量中的取值信息, 据此进行的检验效率很低

▪ 求秩

对原始数据排序, 取各自序号 - 秩
同分 相等的观测值 取平均秩, 如
数据 32, 43, 23, 17, 22, 43, 18, 23, 43, 55, 31, 29, 50
求秩 8, 10, 4.5, 1, 3, 10, 2, 4.5, 10, 13, 7, 6, 12
在总体分布不服从正态分布的情况下, 用基于秩数据的非参数方法

标准化变换

p.58

▪ 标准化

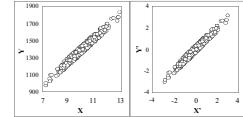
消除量纲和量级的影响 Min-Max 标准化, 标准差标准化等

▪ 标准差标准化

将正态分布数据转换为标准正态分布数据 使均值为 0, 标准差为 1
 $X'_i = (X_i - \bar{X}) / s$
 Grubbs 检验中的标准化过程
 可用于多元分析, 构建尺度一致的多维空间 标准化多元回归

▪ 例

两个变量的正态化 $X \sim 10 \pm 1.01$ $Y \sim 1400 \pm 145$
 统一量纲和量级, 不改变相互关系

**正态变换**

p.59

▪ 正态分布

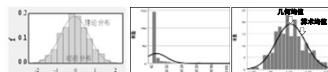
表征: 偏度系数和峰态系数
应用: 参数方法的前提 参数估值 算术均值与标准差, 假设检验 参数方法

▪ 正态检验

判断总体是否服从正态分布的假设检验 参见 3.3 正态分布检验

▪ 正态变换

将不服从正态分布的数据变换为服从正态分布 并非都能实现
变换前后检验 正确理解检验结果 参见 3.3 正态分布检验

**正态变换方法**

p.59

▪ 常用方法举例

特定分布数据有针对性方法 对数正态, 泊松, 二项, 左偏, 右偏等
对数变换最常用 取值低且不为负的变量可能服从对数正态分布

▪ 说明

多数情况下非正态分布总体无法做正态变换
正态变换与异常值剔除的顺序 基于经验的选择

| 正态变换方法 | 适用对象 | 变换式 |
|------------|------------------|-----------------------------------|
| 对数变换 | 对数正态分布数据 | $X'_i = \ln X_i$ |
| 平方根变换 | 泊松分布数据 (离散量) | $X'_i = (X_i + 0.5)^{0.5}$ |
| 角变换 | 二项分布数据 (百分数、比例) | $X'_i = \text{Arcsin}(X_i)^{0.5}$ |
| Box-Cox幂变换 | 任意分布数据 (左右偏、尖低峰) | 下页 |

Box-Cox 幂变换

p.60

▪ 幂变换

一类可以改变偏度和峰态的变换方法 Box-Cox 幂变化是其中之一

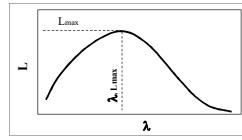
▪ 计算 λ 使以下最大似然函数 L 取极大值的 λ

$$L = -(v/2) \ln s^2 + (\lambda - 1)(v/n) \sum \ln X_i - s^2 \text{ 为变换后数据方差, 只能试算获得 } \lambda$$

▪ Box-Cox 变换式

$$X'_i = \ln X_i, \quad \lambda = 0 \quad \text{特例}$$

$$X'_i = (X_i^\lambda - 1) / \lambda, \quad \lambda \neq 0$$

 X'_i 为变换后数据**1.3.4****1.3.2 异常值检验****1.3.3 数据变换****1.3.4 其它**

数据独立性 p.49

- 数据独立性**
- 采集的个体不受其他采样个体影响
- 非独立性举例**
 - 随机布点采集土壤样品 空间自相关可能影响独立性
 - 系统动态布点测定室内颗粒物浓度 六小时周期可能错过峰或谷
- 独立性检验方法**
- 波动游程检验 参见下页

波动游程检验 p.49

- 中位数上下波动游程检验**
- 将数据变换为二分类类型数据 大于中位数记+，小于中位数记-
- 计算游程数 游程：连续出现的等值符号数据段
- 不独立性表现为相对极端的游程数 下例中的游程数 2 或 30 为最极端情况
- 方法评估**
- 数据信息利用率极低 二分类类型数据放弃了大部分信息
- 不能得出有价值的结论 无论是接受或拒绝原假设 参见 2.1 假设检验
- 数据独立性需要正确的采样方法来保证

| | |
|----|---|
| 30 | + - + - + - + - + - + - + - + - + - + - + - + - + - + - + - + - + - |
| 14 | + + + + - + + - + + - - + - - + - - + + - + - + - + - + + - + - |
| 2 | + + + + + + + + + + + + + + - |

未检出数据处理 非教材内容, 王斌等环境科学学报2009

- 未检出数据**
- 结果低于检出下限 噪声水平的 3-5 倍, 无法获得可靠的结果 记为 N.D.
- 常用处理方法**
 - 删除 样本量减少, 导致对大小的高估和对离散的低估, 影响分布
 - 取零 样本量不变, 导致对大小的低估和对离散的高估, 影响分布
 - 替代 2/3 或 1/2 检出下限, 常用, 样本量不变, 影响分布
- 基于分布的方法**
 - 思路 根据检出数据构建完整的分布, 包括未检出范围
 - 在未检出范围内对未检出数据赋值, 使服从预期分布
 - 方法 最大似然函数, 期望最大化, 顺序回归统计等

数据平滑 非教材内容

- 目的**
 - 消除系列数据噪声 避免过度平滑
 - 一维到多维
- 方法举例**
 - 滑动平均, 加权滑动平均 如五点三次平滑等
 - 样条函数平滑
 - 局部加权回归

应用数理统计方法

第一章 总体特征描述及数据预处理

- 1.1 基本概念与采样方法**
 - 应用举例
- 1.2 总体特征及其表述**
 - 应用举例
- 1.3 数据预处理**
 - 应用举例

应用实例 城市大气颗粒物浓度数据预处理 <https://quotsoft.net/air/>

- 数据: 全国大气 PM2.5 浓度**
1,451 个监测样点 2015.1.17 1500h, n = 1,451
- 对数变换**
显著右偏 对数变换后趋向正态, 略左偏 – 低浓度地区
- 异常值剔除**
原始数据 剔除高值端, 仍尖峰
变换数据 剔除低值端, 基本对称
讨论: 谨慎剔除
变换与剔除的顺序 下页

应用实例 异常值删除与分布检验的关系

▪ 异常值删除

基于正态分布,计算可疑值属于该分布的概率,判断是否剔除

▪ 分布检验

基于峰态和偏度,判断总体是否偏离正态分布 其它方法 参见 3.3 正态检验

▪ 次序

先剔除异常 若是正态分布,合理;若不是,则可能剔除拖尾数据,改变分布形态
先检验分布 若是偏态分布,合理;若是正态有正异常,则可能误判为右偏分布
据经验判断 文献积累信息



应用实例 非统计方法剔除异常

Tao et al. Sci. Total Environ. 2004

▪ 多环芳烃

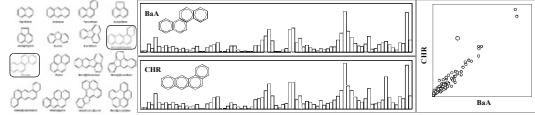
含两个或两个以上苯环的芳烃化合物 一类重要的有毒有机污染物

同分异构化合物行为相似、浓度显著相关 如 苯并蒽和䓛

▪ 例:天津土壤中 16 种多环芳烃含量的实测结果

苯并蒽和䓛浓度显著相关 谱图和散点图

个别样点例外,极可能是非统计异常,应剔除 并非统计意义异常
可表达为对比值的统计偏离



应用实例 口罩的颗粒物去除率数据预处理

Shen et al., EI 2021

▪ 数据: 口罩对空气中不同粒径颗粒物的去除率

六种粒径颗粒物的去除率 表达为均值和 $P_5, P_{10}, P_{25}, P_{50}, P_{75}, P_{90}, P_{95}$

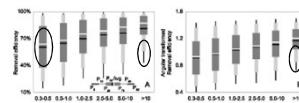
结果显著低峰 细颗粒 和左偏 粗颗粒

▪ 正态变换

百分数据的角变换 $y' = \text{Arcsin}(y^{0.5})$

各粒径数据均更接近正态分布 可做参数检验,粗颗粒仍有左偏倾向

$PM_{<10}$ 均值与标准差: 原始 $81 \pm 18\%$, 变换后 $85 \pm 6\%$



应用实例 遥感坡度数据转换

Csillik et al., Geomorphol. 2015

▪ 研究

数字高程模型中提取的坡度大多右偏 九个不同来源和分辨率数据集 A ... I

比较正态变换的不同方式 Box-Cox, 正弦, 正弦开方, 对数

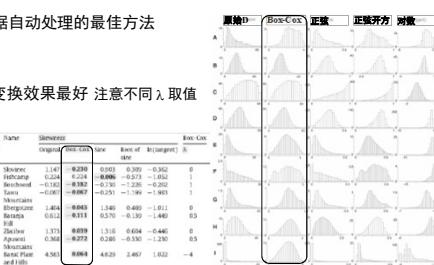
寻找用于数据自动处理的最佳方法

▪ 结果

Box-Cox 幂变换效果最好 注意不同 λ 取值

表 1

| Site Name | Slopeunit | Original | Box-Cox | New | Box-Cox | Box-Cox |
|-------------------------|-----------|----------|---------|--------|---------|---------|
| A Shewee | 2.14° | 0.239 | 0.063 | 0.399 | -0.362 | 0 |
| B Hargreaves | 2.14° | 0.239 | 0.063 | 0.399 | -0.362 | 0 |
| C Beechwood | -0.18° | -0.182 | -0.130 | -1.226 | -0.254 | 1 |
| D Tava | -0.06° | -0.067 | -0.125 | -1.190 | -1.043 | 1 |
| E Shrubland | 2.14° | 0.239 | 0.063 | 0.399 | -0.362 | 0 |
| F Shrubland | 2.14° | 0.239 | 0.063 | 0.399 | -0.362 | 0 |
| G Kibong | 1.37° | 0.059 | 1.115 | 0.604 | -0.446 | 0 |
| H Hill | 2.14° | 0.239 | 0.063 | 0.399 | -0.362 | 0 |
| I Kibong | 1.37° | 0.059 | 1.115 | 0.604 | -0.446 | 0 |
| J Apoatal | 0.30° | 0.272 | 0.205 | 0.350 | -1.230 | 0.5 |
| K Apoatal | 0.30° | 0.272 | 0.205 | 0.350 | -1.230 | 0.5 |
| L Rain Forest and Hills | 4.93° | 0.064 | 4.623 | 2.487 | 1.022 | -4 |



应用实例 网状细胞比数据预处理

Van Wyck et al., Int. J. Lab. Hematol., 2012

▪ 问题

网状细胞比可用于诊断透析效果

采集 30 名患者 12 次透析测定数据 $n = 360$

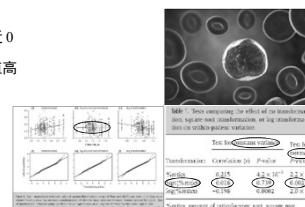
比较两种变换方式对同质性与正态性的影响 角变换, 对数变换

▪ 结果与讨论

方差同质性检验 相关系数接近 0

数据正态性检验 正态检验 p 值高

平方根变换最好 兼顾两者



要点

▪ 定义

描述统计量与检验统计量; 理论与经验分布, 累积分布

分组统计-频率分布-概率密度函数 正态分布, 对数正态分布, 其它分布

▪ 参数估值

分布 峰态-偏度系数 分位数

大小 算术均值 中位数

离散 标准差等 半内四分范围

点估计与区间估计

▪ 数据预处理

异常剔除, 独立性检验, 数据变换, 未检出数据

