

应用数理统计方法

2024.7

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

p.20

▪ 随机变量的特征

单一个体取值具 随机性, 总体中全部个体取值则有 特定规律

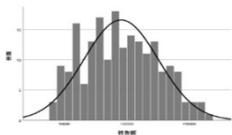
数理统计方法即研究这样的规律

▪ 主要特征

分布 个体在不同取值区间出现的概率

大小 分布中心在数轴上的位置

离散 个体对分布中心的偏离程度



参数估值

p.21

▪ 统计推断

以概率论为基础, 揭示随机变量的内在规律

参数估值 表征总体统计特征

假设检验 判断针对总体的假设是否成立, 如总体区间比较, 总体间关系

▪ 两类统计量

描述统计量 用于参数估值, 描述对象特征

检验统计量 用于假设检验, 计算特定判断的概率

▪ 参数估值

根据 样本 统计量估计 总体 统计量

样本统计量=样本参数, 总体统计量=总体参数

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 计算算术均值, 标准差等

分布表征 计算和检验偏度-峰态系数

▪ 非正态分布总体

总体特征 计算中位数、分位数, 半内四分范围 等

可变换为正态分布的总体 如几何均值等

特征	服从正态分布	不服从正态分布
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

单一个体: 取值有特定的出现概率 如黑白球, PM2.5浓度

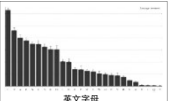
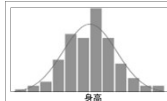
分布特征: 全部个体在可能取值范围内出现概率的分布

▪ 样本分布特征

代表性样本表现出与总体类似的分布特征 不了解总体分布

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 – 构建变量的频数/频率分布图

将取值范围分为若干组 等组距或不等组距 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/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

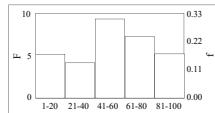
■ 分组统计举例

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 $F_i = 5, 4, 9, 7, 5$ $\Sigma F_i = n = 30$

$f_i = 0.17, 0.13, 0.30, 0.23, 0.17$ $\Sigma f_i = 1$



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

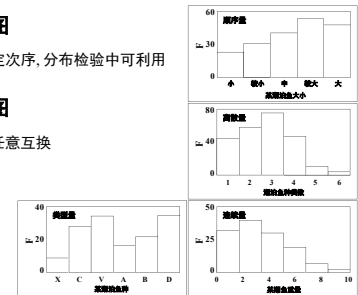
定量变量 连续量, 离散量, 顺序量 定性变量 类型量

■ 定量变量的频数分布图

包含次序信息 横坐标有固定次序, 分布检验中可利用

■ 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

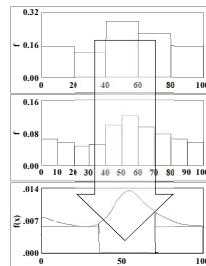
p.31

■ 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

取值范围内曲线下方总面积为 1 频率分布各组相加为 1

特定区间内有特定的取值概率



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

如 二项分布, 对数正态分布, 学生t-分布, F-分布 等

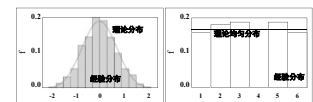
■ 经验分布

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

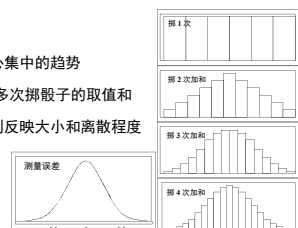
标准正态分布 均值为 0, 标准差为 1 的正态分布 参见 1.3.3 数据变换

■ 特点与重要性

集中性与对称性 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 多次掷骰子的取值和

参数方法的基础 均值和标准差分别反映大小和离散程度



其它常用理论分布举例

p.32-35

■ 对数正态分布

经对数变换后为正态分布 常见

取值小接近纵轴 且为正左侧截断的变量 如气溶胶粒径, 微量金属含量, 风速 ...

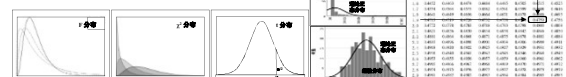
■ 其它

卡方分布, F-分布, 学生 t-分布 从标准正态分布中抽样

对称或非对称分布

取值与概率的对应关系 标准正态分布的例子

分布临界值表 在假设检验中应用



累积分布

p.32-35

■ 累积频率/频数分布

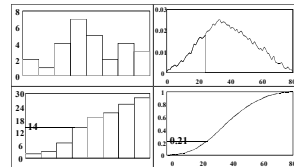
频数/频率分布 特定取值范围内频数/频率的累加

概率密度函数 特定取值范围内密度函数的积分

■ 应用

与一般频率/频率分布图信息量一样 表达方式不同

表现特定范围内的累积效应



区分频率分布曲线与动态变化曲线

■ 频率/频数分布曲线

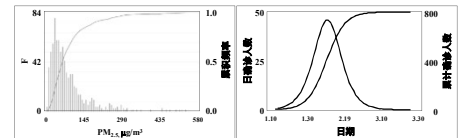
变量的分布特征 横标: 变量取值, 纵标: 频率/频数

例 某地区大气 $PM_{2.5}$ 观测数据的频数分布及累积频率分布

■ 动态变化曲线

变量随时间变化趋势 横标: 时间, 纵标: 变量取值

如 新冠确诊人数与累积确诊人数的日变化规律



总体分布表征

p.26-30

■ 制图

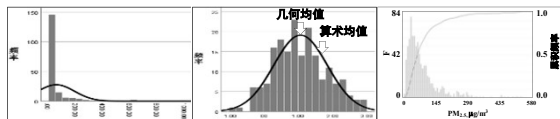
频数/频率分布图, 包括累积分布图

■ 偏度系数与峰态系数

表征对正态分布的偏离 可检验偏离是否显著, 正态检验方法 见3.3 正态检验

■ 分位数

从中位数到百分位数



偏度系数与峰态系数

p.26-27

■ 表征对正态分布偏离的参数

■ 偏度系数

描述两侧偏斜或拖尾 $g_1 = 1/ns^3 \sum (x_i - \bar{x})^3$, 正态 = 0, 左偏 < 0, 右偏 > 0

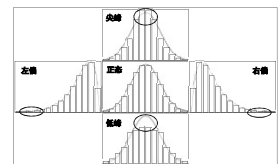
■ 峰态系数

描述中部集中趋势 $g_2 = 1/ns^4 \sum (x_i - \bar{x})^4 - 3$, 正态 = 0, 低峰 < 0, 尖峰 > 0

■ 关于检验

“=” 表示无显著差异 或 不显著

检验 参见 3.3 正态分布检验



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分布特点

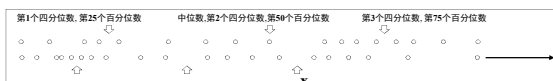
个别数据的剧烈变化 如最大值, 如异常值 对多数分位数无影响

■ 常用分位数

二分位数 中位数

四分位数 将全部个体分为 4 个等量部分的 3 个分点

百分位数 将全部个体分为 100 个等量部分的 99 个分点



分位数与等分

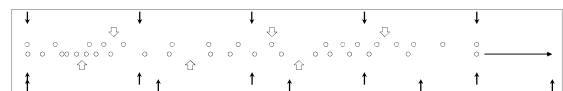
p.28-29

■ 分位数

根据个体出现概率等分, 体现分布及其它特征

■ 例

四分位数 蓝绿两组数据对比, 蓝组最大值右移不改变四分位数



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 略

样本量不足 100 时也可以计算和使用百分位数 线性插值

■ 百分位数选取

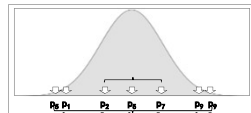
实际使用不需要 99 个百分位数 不是设计百分位数的目的

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% $p_{25} \sim p_{75}$, 90% $p_5 \sim p_{95}$

如用于不确定性表征



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

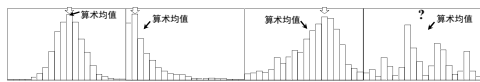
对分布中心位置的估计 与总体分布特征有关

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 $x_L = e^{(1/n)\sum(\ln x_i)}$, $\mu_L = e^{(1/N)\sum(\ln x_i)}$ 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, 兼用离散表征

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



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

总体离散表征

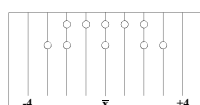
p.39-41

■ 离散特征

个体距分布中心的远近

■ 离散表征

正态分布 离散, 平方和, 方差, 标准差, 变异系数



总体离散表征

p.39

■ 离散特征

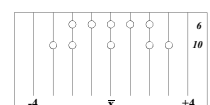
个体距分布中心的远近

■ 离散表征

正态分布 离散, 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



总体离散表征

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

总体离散表征

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

总体离散表征

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

总体离散表征

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

s^2 和 σ^2 分别为样本和总体方差

总体离散表征

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

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

取平方导致方差与算术均值尺度不同, 经常需要同时表达大小与离散

总体离散表征

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与均值的尺度差异

s 和 σ 分别为样本和总体标准差

表达变量离散程度的常用参数

总体离散表征

p.39-41

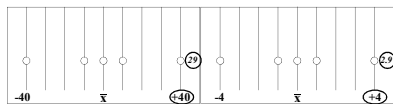
离散特征

个体距分布中心的远近

离散表征

正态分布 离散, 平方和, 方差, 标准差, 变异系数
尺度不同变量的标准差不可比

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



总体离散表征

p.39-41

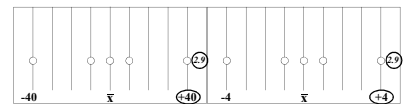
离散特征

个体距分布中心的远近

离散表征

正态分布 离散, 平方和, 方差, 标准差, 变异系数
除以均值消除尺度差异影响
 V 和 V_p 分别为样本和总体变异系数

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



非正态分布离散表征

p.39-41

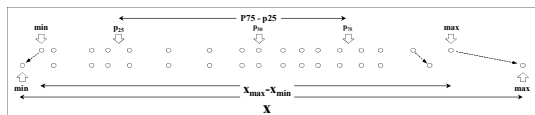
离散表征

正态分布 离散, 平方和, 方差, 标准差, 变异系数

对数正态分布 几何标准差 $s_L = e^{[1/(n-1) \sum (\ln x_i - \ln \bar{x})^2]^{0.5}}$

其他 半内四分范围 $p_{75} - p_{25}$

范围 $X_{\max} - X_{\min}$ 易受两侧拖尾, 异常值干扰, 不稳健



标准差与标准误差

p.39-41

标准差

个体对总体均值的偏离 表现个体离散程度

标准误差

抽样均值对总体均值的偏离 重复抽样样本均值的标准差

$$\text{标准误差 } SD = s/n^{0.5}$$

表现抽样均值的离散程度 - 估计的可靠性

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

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取值的可能范围 估计可靠性描述

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_2, L_1 为上下界, p 为该区间覆盖总体参数的概率

例 算术均值的置信区间

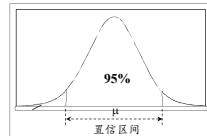
$P\{L_1 \leq \mu \leq L_2\} = p$ $L_1 = \bar{x} - t_{\alpha/2}(n-1)s / n^{0.5}$, $L_2 = \bar{x} + t_{\alpha/2}(n-1)s / n^{0.5}$

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

区间估计

根据预设可靠性估算样本量 估计的可靠性表达 见上例, 见举例



Box-Whisker 图

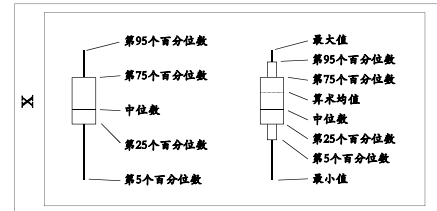
p.42

关键统计量的图示表达

根据需要进行选择参数 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

直观比较不同总体的差异 比较大小, 离散, 分布



应用数理统计方法 2024.7

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/>

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

最小值, 最大值, 均值, 标准差, 中位数, 百分位数 $P_5, P_{10}, P_{25}, P_{50}, P_{75}, P_{90}, P_{95}$

列举最基本的样本统计量

统计量	1001A	1009A	1040A	1041A	1042A	2071A	2072A	2073A	2074A	2075A	2076A	2077A	2078A	2079A	2080A	2081A	2082A	2083A	2084A	2085A	2086A	2087A	2088A	2089A	2090A	2091A	2092A	2093A	2094A	2095A	2096A	2097A	2098A	2099A	2100A							
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	20.0	165.0	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	100.0
最大值	163.0	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	100.0	68.1	66.0	75.1	80.8	162.3	30.8	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	6.0	51.1
中位数	59.0	70.5	84.5	160.5	34.0	111.5	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	46.0	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	1.0	21.2	
众数	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	23.2	16.8	54.0	64.5	65.8	130.0	22.0	93.8	88.3	172.0	66.0	4.0	69.0	44.5	7.0	28.8	95.3	14.8	79.8	62.0	2.3	33.3		
平均数	59.0	70.5	84.5	160.5	34.0	111.5	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	46.0	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	1.0	21.2	
标准差	16.8	54.0	64.5	65.8	130.0	22.0	93.8	88.3	172.0	66.0	4.0	69.0	44.5	7.0	28.8	95.3	14.8	79.8	62.0	2.3	33.3	59.0	59.0	70.5	84.5	160.5	34.0	111.5	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	46.0
方差	116.5	76.8	88.3	95.0	188.0	36.3	127.8	129.8	260.3	87.8	14.3	131.5	71.5	19.0	46.0	240.5	29.3	144.3	101.0	8.8	65.5	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	87.0
偏度	158.3	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	381.2	40.6	215.9	120.4	13.0	95.0	59.0	59.0	70.5	84.5	160.5	34.0	111.5	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	46.0

应用实例 城市空气质量

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

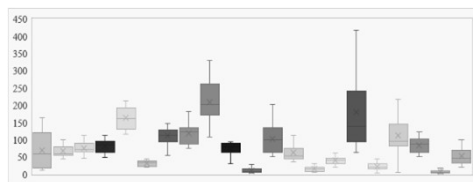
城市空气质量监测平台 PM2.5 浓度

21 个站点, 2015 年 1 月 17 日小时均值

基本统计量图示

均值, 百分位数 $P_{10}, P_{25}, P_{50}, P_{75}, P_{90}$

直观表达 表征与比较



应用实例 城市空气质量

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

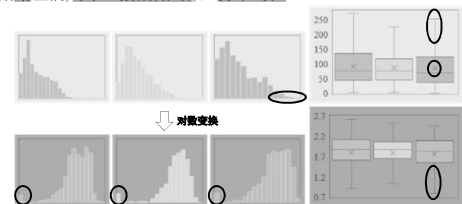
城市空气质量监测平台 PM2.5 浓度

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

对数变换数据略左偏, 未检出数据影响, 用算术均值?



应用实例 大气臭氧浓度

Wang et al., ACP 2009

香港鹤嘴大气臭氧浓度的季节变化

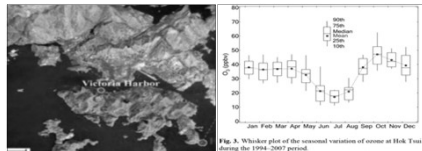
2007 – 2014 连续监测数据

表征季节变化趋势

动态变化的基本统计量图示

算术均值, 百分位数 P10, P25, P50, P75, P90

显著的季节变化 大小, 离散, 分布, 同高同低?



应用实例 室内颗粒物浓度

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, 秸秆, 薪柴, 清洁燃料

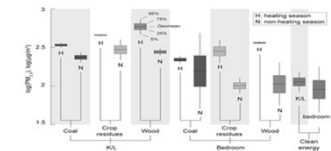
区分不同季节 取暖, 非取暖

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 P5, P25, P75, P95

纵标取对数 大小, 离散, 偏斜



应用实例 新冠潜伏期

Qun et al., NEJM 2020

新冠潜伏期的统计分布

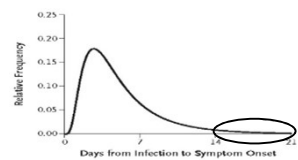
从感染到出现症状天数

频率分布图 略尖峰, 明显拖尾, 类似对数正态分布

应用

传播模拟的重要参数

隔离期选择的依据 两周隔离期?



应用实例 天津市苯并芘超标评估

Li et al., ES&T, 2006

天津市苯并芘呼吸暴露风险

近地面大气苯并芘浓度的空间分布 大气传输模拟

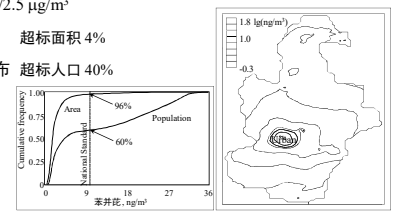
近地面大气苯并芘浓度的频率分布 面积与人口

暴露风险

室外空气质量标准 10/2.5 $\mu\text{g}/\text{m}^3$

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 华北清洁取暖健康效益评估

Meng et al., PNAS

26+2 煤改电的健康效益评估

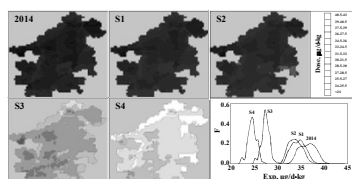
五情景模拟

2014 背景: S1 – S4 干预强度递增情景

人群暴露

五种情景暴露量的空间分布 类似空间分布, 整体下降

五种情景暴露量的频率分布 差异的另一种表达, 特定浓度范围的概率



应用实例 机动车污染物排放因子

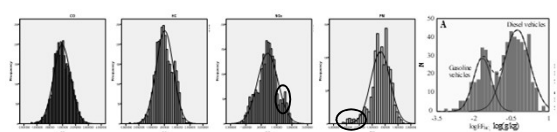
机动车排放因子

排放因子 单位燃料消耗导致的污染物排放量 CO, HC, NOx, PM2.5, BC

排放因子分布特征

典型对数正态分布 局部偏离

黑炭排放因子的双峰分布 柴油车与汽油车两个总体, PeakFit



应用实例 不确定性表征

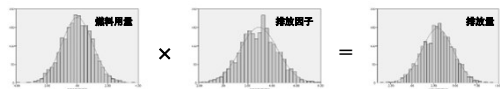
Tao et al., 2003 ES&T

▪ 蒙特卡洛模拟

随机抽样或统计实验方法 环境研究中常用于不确定性分析
据已知概率分布抽样, 建立估计量

▪ 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

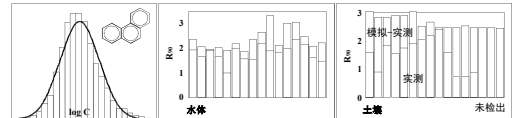
Tao et al., 2003 ES&T

▪ 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

▪ 变异与模型误差拆分

总不确定性包含真实不确定性和变异 Overall / true uncertainties, variability
模拟得到总不确定性, 观测得到空间变异 定义为 $R_{90} = p_{95} - p_5$
两者之差为真实不确定性 即模型误差



应用实例 调查样本量设计

p.44

▪ 置信区间

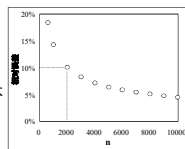
置信区间计算依据: 标准差与样本量 $L_{1,2} = \bar{x} \pm t_{\alpha/2, n-1} s / n^{0.5}$
可根据置信区间和标准差反算样本量

▪ 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, 可计算最低样本量

▪ 实例 农村燃料用量调查

通过预备性调查或文献资料估计均值和标准差
定义相对误差 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 2024.7

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

▪ 检验和保证样本代表性

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
未检出处理 低于测定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

两类异常都会干扰分析结果, 均应剔除

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, 计算能力不足条件下使用

基于正态分布的方法

基于特定理论分布的方法, 可在特定概率基础上作出判断

Grubbs 检验 基于正态分布总体, 有确切的错误率, 考虑样本量影响, 推荐方法

t-检验 基于正态分布总体, 比 Grubbs 法严格, 计算不包括可疑值

Dixon 检验 基于正态分布总体, 比 Grubbs 法严格, 需要不同检验计算式

Grubbs 检验

p.52

方法

数据从小到大排序

从最可能是异常的最小和最大值, 即可疑值 X_i 开始, 顺次判断

直至余下的最小和最大值都不是异常为止

计算

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

或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

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

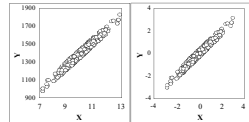
$$X_i' = (X_i - \bar{X}) / s \quad \text{如 Grubbs 检验}$$

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 $X: 10 \pm 1.01 \quad Y: 1400 \pm 145$

统一量纲和量级, 不改变相互关系



正态变换

p.59

■ 正态分布

特征: 偏度系数和峰态系数

应用: 部分数理统计方法的前提 参数估值 算术均值与标准差, 假设检验 参数方法

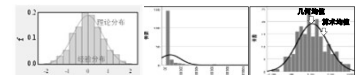
■ 正态检验

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 对检验结果的理解 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, 右偏等分布

对数变换最常用 取值低且不为负的变量常服从对数正态分布

■ 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 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

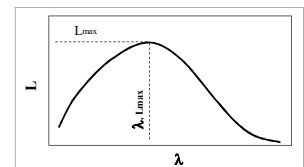
$$L = - (v/2) \ln s^2 + (\lambda - 1)(v/n) \sum \ln X_i \quad s^2 \text{ 为变换后数据方差, 试算获得 } \lambda$$

■ Box-Cox 变换式

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

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

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

所有个体不受其他个体影响

■ 非独立性举例

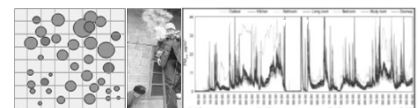
随机布点采集土壤样品 空间自相关影响独立性

随机布点测定排放特征 连续采集样品的相似性影响独立性

系统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 下页



p.49

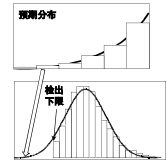
不独立性表现为极端游程数 实例中游程数 2 或 30 为最极端情况

数据独立性需要正确的采样方法来保证

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

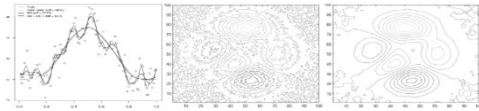
非教材内容 王斌等 环境科学学报 2009

方法 最大似然函数, 期望最大化, 顺序回归统计 等



非教材内容

• • • • •



应用举例



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

讨论: 变换与剔除的顺序 [下页](#)



据经验判断 文献积累信息



应用实例 非统计方法剔除异常

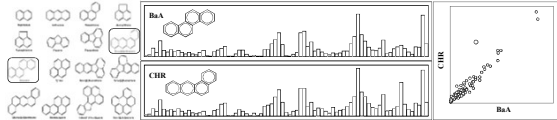
Tao et al. Sci. Total Environ. 2004

多环芳烃

含两个或两个以上苯环的芳烃化合物 一类重要的有机污染物
同分异构化合物行为相似,一般浓度显著相关 如 苯并蒽和蒽

天津土壤中 16 种多环芳烃含量的实测结果

苯并蒽和蒽浓度显著相关 谱图和散点图
个别样点例外,极可能是非统计异常,应剔除 注意该样品的其他化合物
可表达为对比值的偏离



应用实例 口罩的颗粒物去除效果

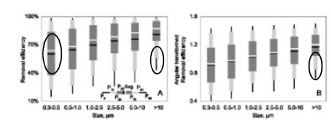
Shen et al., EI 2021

测定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 粗颗粒仍有左偏倾向,可做参数检验
 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

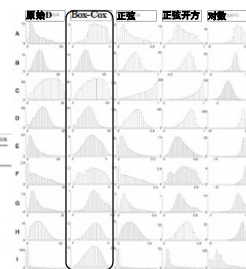
研究

数字高程模型中提取的坡度数据大多右偏 九个不同来源和分辨率数据集
比较不同方式做正态变换 Box-Cox, 正弦, 正弦开方, 对数
寻求合适的变换方法用于数据自动处理

结果

Box-Cox 幂变换结果最好 注意不同 λ 取值

Set	Source	Skewness	Original	Box-Cox	Power	Box-Cox	Box-Cox
A	Shuttle	1.447	0.230	0.303	0.303	0	0
B	FullRange	0.244	0.244	-0.006	-0.013	-1.052	1
C	WorldView	-0.183	0.000	-0.178	-1.258	-0.202	1
D	Topo	-0.007	0.000	-0.251	-1.188	-1.003	1
E	Changjiang	1.414	0.045	1.346	0.400	-1.011	0
F	Topo	0.013	0.013	0.570	-0.138	-1.480	0.5
G	Topo	1.373	0.000	1.310	0.004	-0.446	0
H	Topo	0.360	0.000	0.280	-0.330	-1.230	0.5
I	WorldView	4.563	0.000	4.819	2.407	1.022	-4



应用实例 网状细胞比

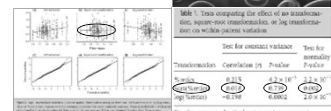
Van Wyck et al., Int. J. Lab. Hematol., 2012

问题

网状细胞比可用于诊断病人透析效果
采集 30 名患者 12 次透析测定数据 $n = 360$
比较不同变换方式对同质性与正态性影响 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
平方根变换效果最好



谢谢

