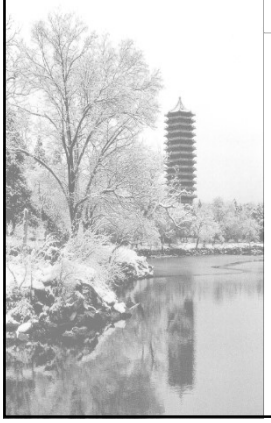


应用数理统计方法

前言

1

1. 课程安排
2. 随机变量与变量特征
3. 数理统计方法与统计推断
4. 统计方法应用举例

安排

- **课时安排**

  26.1.12 – 18　章节和进度大致划分

  08:00 – 12:00

课	日期	内容 编号同教科书章节号
1	01.12	序、1.1 采样、变量
2	01.13	1.2-3 特征,预处理
3	01.14	2.1 显著性检验
4	01.15	2.2-4 大小比较
6	01.16	3.3-4 离散与分布检验
7	01.17	4.1 方差分析
8	01.18	5.1-4 相关分析、回归分析

教材与参考书

- **教材**

  应用数理统计方法　环境科学出版社, 电子版, 勘误表

- **课件**

  Power Point – PDF版

- **主要参考书**

  Rohlf FJ, Sokal RR, **Biometry**, *Freeman, San Francisco*

课程重点

- **核心要求**

  正确使用　方法原理, 适用范围, 应用前提, 使用局限, 实例

- **基本概念**

  方法基础　如 随机变量, 分布概念, 假设检验, 变量类型, 不确定性 等

- **方法原理**

  参数方法　如 从两总体大小比较 到 方差分析

  非参数方法　思维训练, 如 U-检验

  课程内容　部分基于经验的个人理解

- **学习要求**

  各得其所　理解方法, 灵活运用, 正确使用, 考试通过

以应用为主线的章节安排

p.3-4

- **围绕三大特征**

  大小, 离散, 分布

- **从表征到判断**

  参数估值 – 描述特征　包括 大小, 离散, 分布

  假设检验 – 判断假设　如 总体差异, 分布形态, 影响因素, 变量共变, 预测模型

2

1. **课程安排**
2. **随机变量与变量特征**
3. **数理统计方法与统计推断**
4. **统计方法应用举例**

确定性现象与随机现象

p.1

- **确定性现象**

  一定条件下必然发生的现象

  太阳从西面下山, 水往低处流

  理想气体状态方程 – 压强/体积/温度

- **随机现象**

  相同条件下会得到不同结果的现象

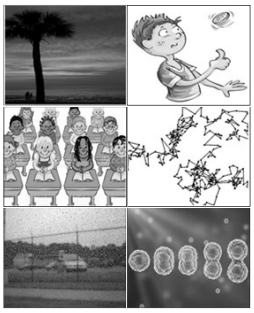
  摄　食　量: 与体重关系

  降水现象: 发生时间, 地点和降水量

  投抛硬币: 正面或反面

  溶液分子: 布朗运动

  细胞复制: 正确或错误



随机现象的内在规律

p.1, Tomasetti1 et al., Science, 2015

- **数理统计**

  揭示随机变量的内在规律

- **举例**

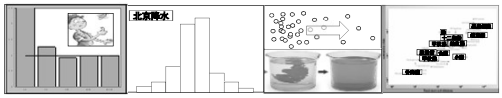
  体　重 与摄食量相关

  抛硬币 重复次数增加, 正反比趋向 1:1

  降　水 多年年均降水量, 降水季节分布, 极端天气出现概率

  稀溶液 沿浓度梯度递降方向的扩散现象 – 溶质分子趋向均匀分布

  干细胞 不同器官癌症发生率与干细胞复制速率正相关



随机变量的主要特征

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

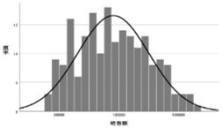
  许多学科的主要研究对象是随机变量　环境, 生态, 地学, 生物, 化学, 社会, 经济 …

- **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



3

1. **课程安排**
2. **随机变量与变量特征**
3. **数理统计方法与统计推断**
4. **统计方法应用举例**

数理统计方法

- **经典统计方法**

  以概率论为基础的统计推断　揭示随机变量的规律

  本课程内容　　只包括基本方法

- **其他统计方法举例**

  多元分析　如聚类分析, 主成分分析 等

  空间分析　如半变异函数分析, 克里格插值 等

  时间序列分析

  ……

统计推断 *p.2-3*

- **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
  如 根据样本均值估计总体均值, 根据样本分布判断总体分布特征
- **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
  如 两总体大小是否相同, 某因素是否有显著影响, 两变量是否相关

统计推断的其他分类举例 *p. 3*

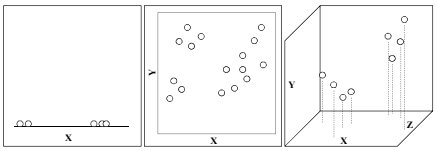
- **参数估值**
- **假设检验**
- **回归问题**
  包含参数估值与假设检验 如回归参数计算及斜率的显著性检验
- **多重决策**
  如多重比较 构建多总体大小关系
- **其他问题**
  如采样方法, 实验设计 等

课程安排及教材内容 *i - v*

- **前言**
  基本概念与采样方法 第一章
- **参数估值**
  随机变量表征 第一章
- **假设检验**
  假设检验 第二章
  大小比较 第二章
  离散比较, 分布比较 第三章
  方差分析 第四章
  相关分析 第五章
  回归分析 第五章

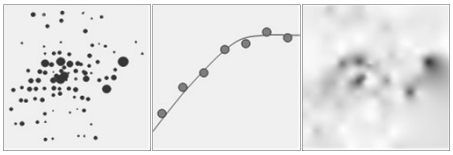
其他统计方法简介

- **多元分析**
  研究多变量相互关系
  基于多元数据在多维空间中的位置
  无概率意义
- **举例**
  聚类分析 欧氏距离, 夹角余弦



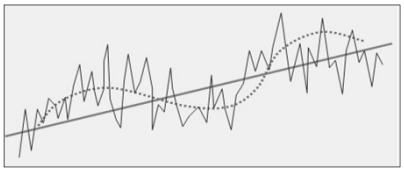
其他统计方法简介

- **空间分析**
  研究变量的空间分布规律
  可用于空间结构分析和空间插值
- **举例**
  半变异函数分析与克里格插值



其他统计方法简介

- **时间序列分析**
  研究变量的动态变化特征
  可用于动态变化驱动因素解析
- **举例**
  不同时间尺度影响因素的分解



4

1. **课程安排**
2. **随机变量与变量特征**
3. **数理统计方法与统计推断**
4. **统计方法应用举例**

应用实例 概率计算

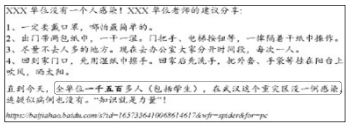
- **有时可直接计算概率作出判断**
- **新冠疫情期间的一条网上评论**

背景: 武汉 1,400 万人, 致2020年2月4日全市确诊人数达 6,384 人

事实: 武汉某单位 1,500 员工, 无一个人感染！

解释: 戴口罩, 纸巾开门, 不去人多的地方, 洗手 …

结论: 知识就是力量


XXX 单位没有一个人感染！XXX 单位老师的建议分享：
1、一定要戴口罩，哪怕戴简单的。
2、出门带两包纸巾，一干一湿。门把手、电梯按钮等，一律隔着干纸巾操作。
3、尽量不去人多的地方。现在去办公室大家分开时间段，每次一人。
4、回到家门口，先用湿纸巾擦手。回家后先洗手，把外套、手袋等挂在阳台上吹风，晒太阳。
直到今天，全单位一千五百多人（包括保洁），在武汉这个重灾区没一例感染，连疑似病例也没有。"知识就是力量"！
https://baijiahao.baidu.com/s?id=1653336410668614617&wfr=spider&for=pc

应用实例 概率计算

- **有时可直接计算概率作出判断**
- **新冠疫情期间的一条网上评论**

背景: 武汉 1,400 万人, 致2020年2月4日全市确诊人数达 6,384 人

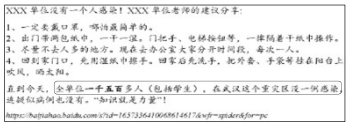
事实: 武汉某单位 1,500 员工, 无一个人感染！

解释: 戴口罩, 纸巾开门, 不去人多的地方, 洗手 …

结论: 知识就是力量

概率: 1500 人同期无一例感染的概率是 $(1\text{-}6384/14{,}000{,}000)^{1500} = 50.5\%$

讨论: 措施对, 证据不确切


XXX 单位没有一个人感染！XXX 单位老师的建议分享：
1、一定要戴口罩，哪怕戴简单的。
2、出门带两包纸巾，一干一湿。门把手、电梯按钮等，一律隔着干纸巾操作。
3、尽量不去人多的地方。现在去办公室大家分开时间段，每次一人。
4、回到家门口，先用湿纸巾擦手。回家后先洗手，把外套、手袋等挂在阳台上吹风，晒太阳。
直到今天，全单位一千五百多人（包括保洁），在武汉这个重灾区没一例感染，连疑似病例也没有。"知识就是力量"！
https://baijiahao.baidu.com/s?id=1653336410668614617&wfr=spider&for=pc

应用实例 大小表征

第一章

- **天津市土壤 DDT 污染水平**

问题: 区域土壤污染状况的统计表征　研究区域土壤污染状况

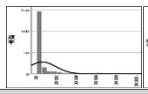
- **方法与结果**

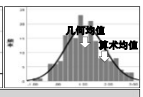
方法: 采集了 187 个表土样品, 测定 DDT 含量 – 数据表

结果: 数据特征: 典型的对数正态分布

算术均值: 56.3 ng/g ×

几何均值: 11.8 ng/g ✓





天津土壤 DDT含量 (ng/g)

2.4, 30.6, 16.6, 1.6, 4.2, 3.2, 7.8, 199.1, 1.6, 64.4, 1.1, 4.0, 4.7, 3.7, 8.6, 43., 5.2, 29.3, 152., 23.4, 41.1, 92., 103., 9, 36.4, 37.8, 37.4, 3.4, 7.1, 207.9, 8.0, 94.4, 102.0, 9.5, 84.4, 0.9, 6.0, 9.0, 19, 135.7, 870.0, 32.9, 16.5, 24.2, 3.8, 2.0, 8.1, 12.0, 167.4, 0.9, 3.9, 1.0, 6.5, 644.5, 3.1, 64.5, 30.7, 205.0, 0.2, 3.1, 20, 67, 34.537.6, 67.5, 19.2, 2.2, 1.7, 21.8, 0.6, 16.6, 218.5, 83.6, 36.7, 621.7, 22.8, 28.0, 2.1, 2.6, 16.2, 54.8, 379.9, 124.5, 972.2, 4.9, 0.7, 44.9, 3.0, 14.5, 304.7, 19.7, 0.9, 77.5, 510, 344.8, 3.0, 22., 2.7, 2370., 18.2, 81.1, 287., 10, 12.0, 1.5, 22, 36.4, 3.4, 56, 226.7, 73.5, 14.1, 17.9, 1.0, 135.7, 2089, 4.8, 72, 0.5, 34.7, 0.6, 10.6, 23, 208.8, 6.6, 10, 2.4, 17.4, 43.8, 7, 0.1, 0.5, 5.8, 194.0, 57, 259, 54, 3.7, 37.3, 0.9, 72, 3.3, 11.5, 19.5, 2.8, 8.4, 9.6, 2.1, 0.9, 15.9, 4.0, 126.4, 31.3, 0.6, 14.5, 8.4, 19.4, 65.4, 8.0, 0.6, 6.9, 2.2, 103.5, 54, 11.3, 27, 1.5, 115.3, 21.8, 95.6, 3.4, 6.4, 19.1, 24.8, 25.6, 37.7, 121.8, 33.4, 13.2, 15.3, 89.4, 446.3, 17, 8.8, 48.1, 6.8, 59.5, 2.8

应用实例 大小比较

Peng et al. J Nurs Sch. 2022,152-160, 第二章

- **吸烟口腔癌患者复吸研究**

问题: 高依赖性与中低依赖性患者间差异

- **方法与结果**

数据: 220 例患者, 确诊后中止吸烟

　　调查治愈后的复吸率

目的: 探讨两组差异, 影响因素

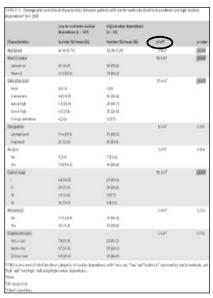
方法: 大小比较 t-检验和分布比较卡方检验

结果: 高依赖性患者复吸率高

　　年轻, 未婚, 教育和早期确诊依赖性强

　　职业, 宗教, 复发和抑郁状态无显著影响

讨论: 不同影响因素的交互影响？



应用实例 影响因素

Shen et al. Environ Int. 2021, 第二章

- **口罩对空气中颗粒物的去除效果**

问题: 比较不同品牌口罩对不同粒径颗粒物的去除效率

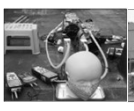
- **方法与结果**

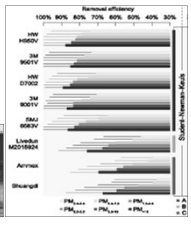
方法: 模拟和志愿者实验 测定口罩内外颗粒物, 计算和比较去除率

结果: 五种口罩没有显著差异, 另三类偏低

　　去除效率随粒径下降而下降

　　其他: 浓度, 性别, 污染 …





应用实例 变量相关

Dai et al. J Cent S Univ 2012,2634-2642, 第五章

- **沉积物重金属污染**

  问题: 探讨沉积物重金属污染浓度与粒径的关系

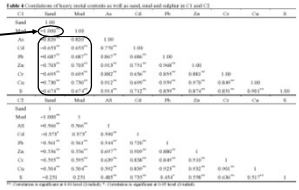
- **方法与结果**

  方法: 采集两套湘江沉积物柱获得分层样品, 测定重金属含量及泥, 沙含量

  统计: 金属, 泥, 沙含量间相关关系: 相关矩阵, 标注 0.05 和 0.01 两级显著性水平

  结果: 金属间相关,也与粒径相关

  讨论: 泥, 沙间为假相关



应用实例 线性预测模型

人教版初中化学, 第五章

- **分析方法的标准曲线**

  问题: 构建浓度 – 响应关系 吸收, 面积, 重量 …

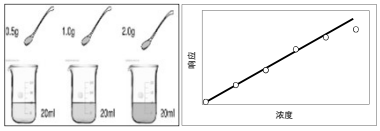
- **方法及应用**

  方法: 利用已知浓度标准样品, 获得特定方法的响应关系如比色法

  利用浓度 – 响应关系的近似线性段, 构建标准曲线方程

  应用: 利用标准曲线定量计算未知浓度样品的浓度

  讨论: 非线性, 非零截距, 拟合优劣判断



应用实例 多元预测模型

Lu et al., Environ. Int., 2020, 第五章

- **室内 PM2.5 浓度估计**

  问题: 基于高分辨在线监测数据构建估计室内$PM_{2.5}$浓度的统计模型

- **方法与结果**

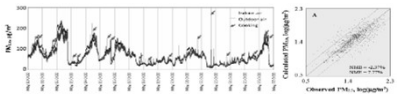
  方法: 同步监测 53 户居民室内外 $PM_{2.5}$ 浓度　　低成本传感器技术

  现象: 中高浓度时室内偏低, 低浓度时室内偏高 同步与滞后

  模型: $\log Ci = (1-0.01807P)(0.91695+0.00997W/H)\log Co$, $R^2=0.785$

  Ci, Co 室内外浓度, P, W, H 净化器使用状态, 开窗指数, 面积

  讨论: 炊事影响, 窄峰, 细颗粒



应用实例 非线性预测模型

第五章

- **预测新冠确诊人数的简单统计模型**

  问题: 基于已知数据及可能的定量关系

- **武汉疫情的早期预测**

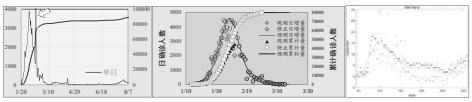
  方法: 基于至2月16日公布数据, 用Logistic 回归 $Y = k/(1+be^{at})$

  2.16 预测: 确诊人数降到两位数和一位数的日期, 最终累积确诊人数

  极端异常处理 基于漏检假设, 向前平滑

  预测: 降至两位和一位数的日期为 **2.29** 和 **3.10**, 实际为 **3.2** 84 和 **3.12** 9, 误差两天至 **3.10**, 累计确诊人数 **77,309**, 三月初实际 **80,991**, 低估 **4.5**%

  之后: …



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

1.1.1

1.1.1 基本概念

1.1.2 采样方法

个体, 总体与样本 *p.7*

- **个体**
  研究对象的基本单元 不一定有确切的大小, 如一份水样
- **总体**
  研究对象的全体 总体量 **N** = 总体中包含的个体数量
- **样本**
  从总体中抽取出来用于观测的个体 样本量 **n** = 样本中包含的个体数量
- **统计**
  从总体中抽取样本进行观测, 据此推断该总体的性质
  对比: 普查则穷尽全部个体

变量及变量属性 *p.15-19*

- **变量**
  关注的总体性质 如饮水量, 污染物浓度, 去除率, 感染率, … 湖泊/鱼参数
- **变量属性**

观测水平 { 连续变量, 离散变量, 顺序变量 } → 定量变量
观测水平 { 类型变量 } → 定性变量
取值性质 { 固定变量, 随机变量 }
获取方式 { 观测变量, 衍生变量 }

根据观测水平分类 *p.15-19*

- **连续量**
  理论上取值精度无限 如 身高, 体重, 浓度, 长度 …
- **离散量**
  正整数, 计数值 如 人口, 植株, 机动车保有量, 动物存栏数
- **顺序量**
  按大小排列的顺序值, 秩数据 如 高度次序, 大小次序, 质地次序, 优劣次序
- **类型量**
  只有属性意义 如 土壤类型, 健康状况, 季节, 性别

根据取值性质分类 *p.15-19*

- **随机变量**
  个体随机出现, 大量取值表现出宏观规律
  数理统计方法的直接研究对象或影响因素
- **固定变量**
  人为控制
  不是数理统计方法的直接研究对象, 可以是方法中的影响因素
- **区分随机变量与固定变量**
  是否可控
  是否可重复 多级分组方差分析中的次级变量特例

随机变量与固定变量举例 *p.15-19*

- **连续变量 – 温度对微生物生长的影响**
  固定 实验室内或装置内人为控制的温度
  随机 现场非控制条件下实际观测到的温度
- **属性变量 – 不同城市降水酸度的差别**
  固定 覆盖所有待研究城市
  随机 从所有研究城市中随机抽取部分代表性城市
- **验证**
  固定 可重复
  随机 不可重复

根据获取方式分类 *p.18*

- **观测变量**
  来自直接观测, 信息完整
- **衍生变量**
  根据观测变量计算得到, 精度低 如比例, 指数, 百分数, 比率 等

不同观测水平变量间的关系 *p.15*

- **信息量**

  沿以下方向递降 连续量 > 离散量 > 顺序量 > 类型量

- **属性转换**

  仅能向信息量递降方向, 如

  求秩 将连续量或离散量转换为顺序量, 如身高排序 1, 2, 3, …, n

  分类 将定量量转换为类型量, 如身高分类: 高, 较高, 中, 较矮, 矮

基本概念举例 *p.15*

- **总体, 样本与变量**

  研究者根据研究目定义

- **举例**

  中国儿童饮水量表征

  天津土壤滴滴涕含量表征

  两个口腔癌吸烟人群对比

  六种口罩颗粒物去除率对比

  沉积物重金属含量关系判断

  分析方法的标准曲线构建

  新冠确诊预测

  湖泊和鱼

1.1.2

1.1.1 基本概念

1.1.2 采样方法

采样方法 *p.10-11*

- **样本代表性**

  统计分析的必要前提 需要根据有限样本观测结果对总体作出判断

- **确保采样代表性的关键**

  正确的采样方法　与总体特征及研究要求有关

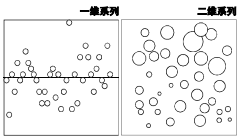
  足够大的样本量　影响检验的可靠性, 越大越好, 没有确切界限, 受限成本

- **基于总体中个体排序的抽样**

  基于特定规则排序 个体, 时间, 空间

  一维或多维

  抽样的基础



随机取样 *p.11-12*

- **从任意排序的总体中随机抽样**

  每个个体被抽取的机会相同

- **优缺点**

  优点 随机与客观

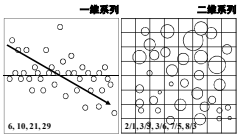
  缺点 分布不均匀, 易受空间自相关等因素影响

- **随机数获取**

  随机数表 教科书 p.376 表

  电子工作表格 Excel Rand() 函数

  其他 ….



系统取样 *p.12*

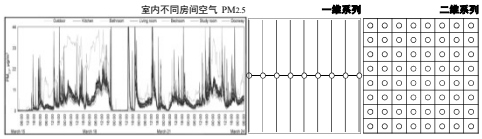
- **从随机排序的总体中按固定间隔抽样**

  获得等间隔的代表性样本

- **优缺点**

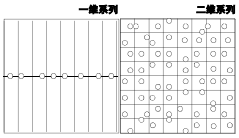
  优点 特定排序的均衡分布

  缺点 与周期性变化重叠的风险



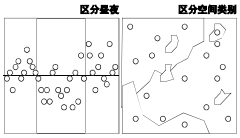
系统-随机取样 *p.12*

- **在特定间隔内随机抽样**
  结合系统取样与随机取样方法
  获得均衡分布的随机样本
- **优缺点**
  优点 综合随机取样和系统取样两者的优点
  缺点 操作难度大, 成本高



多层次取样 *p.13*

- **单层次取样**
  将总体视为一个整体对象
- **多层次取样**
  方法 将总体系统划分为有不同特征的多个层次 子系统
  各层次独立取样, 可设定不同层次的权重 时间, 面积, 个体数 等
  相当于将对象拆分为不同总体
  优点 代表性好, 可对各层次数据进行独立分析



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 采样方法

- **农村直接生活能源结构与用量调查**
  总体与样本 全部农户和抽取的农户
  变量 能源结构和用量, 其他辅助参数, 如人口, 收入等
- **分层取样**
  样本量: 按空气质量分 重点区 **0.43‰** 和非重点区 **0.22‰**
- **系统-随机取样**
  1 地级单元 样本量与户数成正比 不足 50 者合并 **276**/**334**
  2 县级单元 随机抽 **1**/**7**
  3 村级单元 每县随机 **2** 个
  4 户级单元 按设计样本量随机抽取

应用实例 采样方法

- **人群口罩佩戴率调查**
  目的 调查不同气温和污染条件下男, 女口罩佩戴率
  总体和样本 北大师生员工和抽取的人群, 在北大东门连续拍摄
  变量 口罩佩戴状态, 性别, 气温, 污染水平
- **系统-随机取样**
  抽样 系统随机抽样 每月随机抽取 **2** 天
  数据 区分性别和口罩佩戴状态
  记录气温与$PM_{2.5}$浓度
  缺点 进出东门人群的代表性

要点

- **随机变量及特征**
  随机现象, 随机变量与固定变量
  大小, 离散, 分布
- **定义**
  个体与属性, 总体与样本, 统计与普查
  变量属性 观测水平, 取值性质, 获取方式
- **统计推断**
  参数估值与假设检验
- **采样方法**
  随机, 系统, 系统-随机

谢谢