

应用数理统计方法
2024.7

前言

陶澍
北京大学 南方科技大学

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. 统计方法应用实例

课程安排

▪ 课时安排

共 7 天7.1 - 7.7

地点与时间三教 207, 9:00 - 12:00 左右, 中间休息一次

▪ 上课要求

无前续课程要求

不影响他人

▪ 联系方式

授课: 陶澍taos@sustech.edu.cn

练习: 朱雷, 沈慧中zhul3@sustech.edu.cn shenhz@sustech.edu.cn

课程内容

▪ 课程内容及次序

7.1 周一 前言 1.1 基本概念与采样方法

7.2 周二 1.2 总体特征及其表述 1.3 数据预处理

7.3 周三 2.1 假设检验与假设检验方法

7.4 周四 2.2-4 大小比较

7.5 周五 3.1-2 离散程度比较 3.3-4 拟合度检验

7.6 周六 4 方差分析

7.7 周日 5.1-2 回归分析 5.3-4 回归分析

▪ 课件与课下实习

https://zhu-group.github.io/ams 可用手机, 电脑, 平板打开

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

PDF 黑白每页六幅 教材页码或文献 课件右上角

▪ 主要参考书

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



教材内容

▪ 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

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

相关与回归 第五章



省略内容

▪ 省略小节

受课时限制,教科书中以下小节省略

第一章 1.2.1 1.3.1

第二章 2.2.2 2.3.2 2.3.3 2.3.5 2.4.2 2.4.4 2.5.2

第三章 3.1.1 3.1.2 3.1.3 3.2 3.4.3 3.4.4 3.5

第四章 4.3.3

第五章 5.1.3 5.1.5 5.1.6 5.2.4 5.3.4

课程重点

▪ 学习要求

正确使用 如 方法原理,适用范围,应用前提,使用局限,应用实例

▪ 基本概念

正确使用的基础 如 随机变量,分布概念,假设检验,变量类型,不确定性

▪ 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

▪ 学习要求

各得其所 理解方法,灵活运用,正确使用

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

假设检验 – 判断假设 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. 统计方法应用实例

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

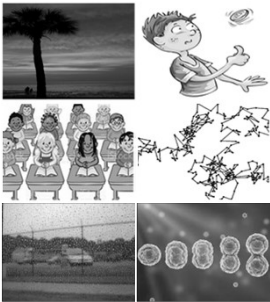
上课次数,与成绩的关系

降水现象,发生时间,地点和降水量

抛抛硬币,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

▪ 举例

抛硬币 重复次数增加,正反比趋向 1:1

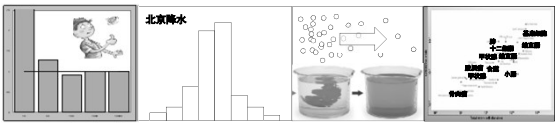
降水 多年年均降水量,降水季节分布,极端天气出现概率

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

▪ 数理统计

揭示随机变量的内在规律



随机变量的主要特征

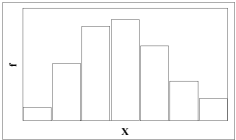
p.20

▪ 随机变量的特征

宏观规律的表现
数理统计方法的研究对象
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 ...

▪ 统计学特征

分布 个体在不同取值 (范围) 的 出现概率
大小 分布中心在数轴上的 位置
离散 个体的聚集/分散 程度



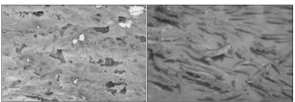
随机变量举例

▪ 湖泊与鱼

某区域有多个湖泊, 湖中有若干种鱼, 湖泊和鱼有多种属性 – 随机变量
湖 流域面积, 水面面积, 平均水深, 径流量, 蓄水量, 水温, 水质参数, 营养水平 ...
鱼 种类, 数量, 重量, 大小, 食性 ...

▪ 研究对象

研究者根据需求自行定义
全部或部分湖泊, 全部或部分鱼, 某个或某些特性



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. 统计方法应用实例

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
空间分析 如半变异函数分析, 克里格插值
时间序列分析
.....

统计推断

p.2-3

▪ 统计推断

在概率论的基础上, 根据对 样本 观测判断 总体 特征
包括 参数估值 和 假设检验

▪ 参数估值

描述总体统计特征, 即根据样本统计量对总体统计量进行估计
如 根据样本均值估计总体均值, 如湖深, 鱼重

▪ 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 ... 如湖深与鱼重

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 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

其他统计方法简介

多元分析

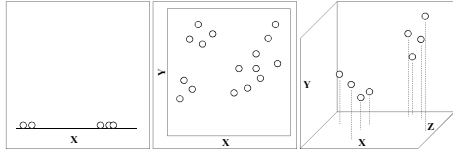
研究多变量相互关系

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

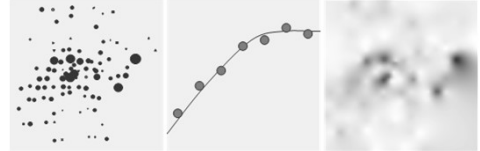
空间分析

研究变量的空间分布特征

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

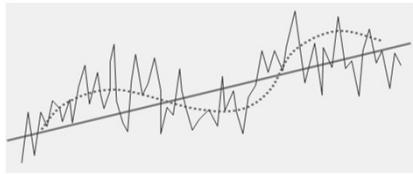
时间序列分析

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. 统计方法应用实例

应用实例 概率计算

有时只需要直接计算概率就可以作出判断

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 1500 员工, 无一个人感染, 连疑似的都没有!

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 $(1-6384/14,000,000)^{1500} = 50.5\%$

讨论: 措施正确, 表述不科学

XXX 单位没有一个人感染! XXX 单位老师的建议分步:
1. 一定要戴口罩, 哪怕最简单的。
2. 出门常带包纸巾, 一干一湿, 门把手, 电梯按钮等, 一擦随着手纸中操作。
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。回家前先洗手, 把外套、书包等挂在阳台上吹风, 晒太阳。
直到今天, 单位一千五百多人(包括学生), 在武汉这个重灾区无一例感染, 是挺厉害的, 也没有“知识就是力量”!

<https://weibo.com/16573564109068614617649-qqdkerdfw-pc>

应用实例 描述统计量

中国人群暴露参数

问题: 获取表征暴露特征的基本统计量

为风险分析提供基础参数

方法与结果

方法: 基于大规模入户调查, 获取相关参数, 如:

例子: 中国成年人日均间接日饮水量

区分: 区域, 城乡, 性别, 年龄

列举: 均值和百分位数

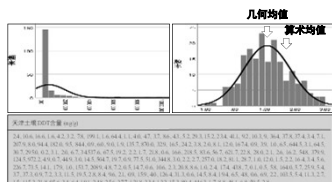
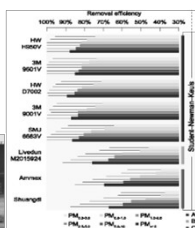


Table 2 Descriptive statistics of the variables used in the multivariate model. Descriptive statistics of the variables used in the multivariate model			
	Sample size (n = 100)	Age at onset (n = 100)	Age at diagnosis (n = 100)
Gender			
Male	50 (50%)	50 (50%)	50 (50%)
Female	50 (50%)	50 (50%)	50 (50%)
Marital status			
Married	35 (35%)	35 (35%)	35 (35%)
Single	65 (65%)	65 (65%)	65 (65%)
Work			
Employed	40 (40%)	40 (40%)	40 (40%)
Unemployed	60 (60%)	60 (60%)	60 (60%)
Family			
Family size	4.5 (4.5)	4.5 (4.5)	4.5 (4.5)
Family income	1000 (1000)	1000 (1000)	1000 (1000)
Family structure	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family type	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family size	4.5 (4.5)	4.5 (4.5)	4.5 (4.5)
Family income	1000 (1000)	1000 (1000)	1000 (1000)
Family structure	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family type	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family size	4.5 (4.5)	4.5 (4.5)	4.5 (4.5)
Family income	1000 (1000)	1000 (1000)	1000 (1000)
Family structure	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family type	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family size	4.5 (4.5)	4.5 (4.5)	4.5 (4.5)
Family income	1000 (1000)	1000 (1000)	1000 (1000)
Family structure	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family type	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family size	4.5 (4.5)	4.5 (4.5)	4.5 (4.5)
Family income	1000 (1000)	1000 (1000)	1000 (1000)
Family structure	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family type	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family size	4.5 (4.5)	4.5 (4.5)	4.5 (4.5)
Family income	1000 (1000)	1000 (1000)	1000 (1000)
Family structure	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family type	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family size	4.5 (4.5)	4.5 (4.5)	4.5 (4.5)
Family income	1000 (1000)	1000 (1000)	1000 (1000)
Family structure	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family type	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family size	4.5 (4.5)	4.5 (4.5)	4.5 (4.5)
Family income	1000 (1000)	1000 (1000)	1000 (1000)
Family structure	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family type	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family size	4.5 (4.5)	4.5 (4.5)	4.5 (4.5)
Family income	1000 (1000)	1000 (1000)	1000 (1000)
Family structure	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family type	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family size	4.5 (4.5)	4.5 (4.5)	4.5 (4.5)
Family income	1000 (1000)	1000 (1000)	1000 (1000)
Family structure	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family type	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family size	4.5 (4.5)	4.5 (4.5)	4.5 (4.5)
Family income	1000 (1000)	1000 (1000)	1000 (1000)
Family structure	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family type	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family size	4.5 (4.5)	4.5 (4.5)	4.5 (4.5)
Family income	1000 (1000)	1000 (1000)	1000 (1000)
Family structure	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family type	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family size	4.5 (4.5)	4.5 (4.5)	4.5 (4.5)
Family income	1000 (1000)	1000 (1000)	1000 (1000)
Family structure	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family type	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family size	4.5 (4.5)	4.5 (4.5)	4.5 (4.5)
Family income	1000 (1000)	1000 (1000)	1000 (1000)
Family structure	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family type	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family size	4.5 (4.5)	4.5 (4.5)	4.5 (4.5)
Family income	1000 (1000)	1000 (1000)	1000 (1000)
Family structure	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family type	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family size	4.5 (4.5)	4.5 (4.5)	4.5 (4.5)
Family income	1000 (1000)	1000 (1000)	1000 (1000)
Family structure	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)
Family type	1.5 (1.5)	1.5 (1.5)	1.5 (1.5)

[illegible]

应用实例 非线性预测模型

▪ 预测新冠确诊人数的简单统计模型

问题: 基于已知机理或基于已知数据

▪ 武汉疫情的早期预测

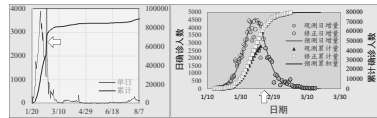
方法: 基于公布数据 Logistic 回归 2.16 预测日及累积确诊人数, $Y = k/(1+be^{at})$

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

累计至 3.10 确诊人数 77,309, 三月初实际 80,991, 低估 4.5%

之后:



应用实例 统计推断问题

▪ 参数估值问题

举例: 湖泊面积? 湖泊深度? 鱼重? 鱼龄?

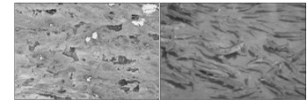
▪ 假设检验问题

问题: 不同湖泊中鱼的密度有差别吗? 两个最大湖泊中鱼的重量一样吗?

问题: 湖泊中鱼的数量与浮游生物数量有关吗?

问题: 两个湖泊中主要鱼种比例有差别吗?

问题: 能否根据某些湖泊参数估计湖泊中鱼的总量?



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

1.1.1

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察的个体 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: 普查需穷尽全部个体

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	定量变量
	离散变量	
	顺序变量	定性变量
取值性质	类型变量	
	固定变量	随机变量
获取方式	随机变量	
	观测变量	衍生变量
	衍生变量	

根据观测水平分类	p.15-19
<ul style="list-style-type: none"> 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
<ul style="list-style-type: none"> 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 <u>影响因素</u> 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 <u>影响因素</u> 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实例	p.15-19
<ul style="list-style-type: none"> 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 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城市 检验 固定 可重复 随机 不可重复 	

根据获取方式分类	p.18
<ul style="list-style-type: none"> 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关系	p.15
<ul style="list-style-type: none"> 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 向降低信息量方向, 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 	

基本概念实例	p.15
<ul style="list-style-type: none"> 总体, 样本与变量 研究者根据研究目的定义 举例 中国儿童饮水量 天津土壤滴滴涕含量 两个口腔癌吸烟人群的复吸率 六种口罩颗粒物去除率 沉积物重金属含量 	

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 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 与数据特征及研究要求有关

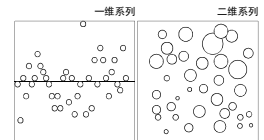
足够的样本量 影响检验的可靠性, 越大越好, 没有确切界限, 受限成本

■ 总体中个体排序

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取个体

每个个体被抽取的机会相同

■ 特点

优点 随机与客观

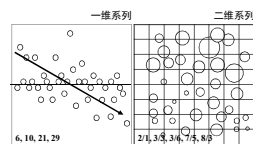
缺点 分布不均匀, 易受空间自相关影响

■ 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

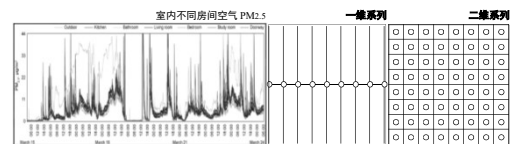
■ 从随机排序的总体中按固定间隔抽样

获得等间隔的规律样本

■ 特点

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

■ 在特定间隔内随机取样

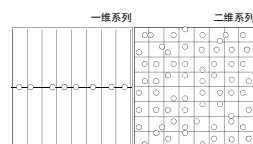
结合系统取样与随机取样方法

获得均衡分布的随机样本

■ 特点

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

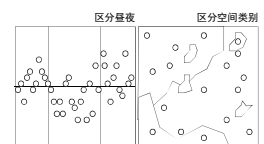
■ 多层次取样

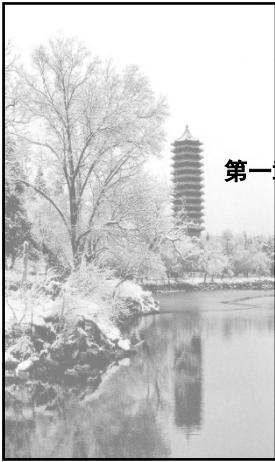
方法: 将总体系统划分为有特定特征的多个层次 子系统

各层次独立取样, 可设定不同层次的权重 时间, 面积, 个体数

相当于将对象拆分为不同 总体

优点: 代表性好, 各层次可独立分析





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 采样方法

问题

湖泊面积与 鱼重 表征

定义

总体: 研究区域的所有湖泊, 湖泊中所有鱼

样本: 抽样的湖泊和捕集到的鱼

变量: 水面面积 连续变量, 鱼的重量 连续变量

采样设计

系统随机抽取湖泊 经纬度网格—网格内按编号随机

从抽取湖泊中随机捕捞 随机布点



应用实例 采样方法

问题

水面面积大于1 km² 湖泊中鲮鱼平均 长度 是否与 蓄水量 有关

定义

总体 研究区域 水面面积 >1 km² 湖泊, 其中所有鲮鱼

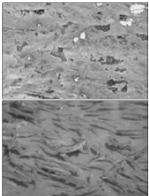
样本 抽样湖泊和捕集到的鱼

变量 湖泊蓄水量 连续变量, 鱼的长度 连续变量

采样设计

系统随机抽取湖泊 经纬度网格—网格内按编号随机

从抽样湖泊中随机捕捞 随机布点捕捞



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

方法: 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



调查点位



应用实例 采样方法

Shen et al. Environ Int. 2021

师生员工口罩佩戴率调查

方法: 北大东门, 连续拍摄 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


变量: 口罩佩戴状态、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

缺点: 进出东门人群的代表性



作业

作业

自愿完成, 课程结束时提交

内容

一个研究实例, 说明研究问题 自己的研究或文献例子

给出数据表 数据量大可用Excel表

提出统计假设, 选择统计方法

给出检验结果和结论

格式: Word文档 清晰简洁的文字描述

谢谢

