统 计 学 Statistics

陈 灯 塔

Econ, XJTUCC

CDT.WISE@G

2021 年 4 月 28 日

内容梗概

1. 数据收集
2. 数据整理
3. 软件操作

Genius

Genius: one percent inspiration and 99 percent perspiration

——Thomas Edison

None of my inventions came by accident. I **see a worthwhile need to be met** and I make trial after trial until it comes. What it boils down to is one per cent inspiration and ninety-nine percent perspiration. —— Thomas Edison

Without the one percent of inspiration, all the perspiration in the world is only a bucket of sweat. —— Cindi Myers

灵感来自思考。多复习，多思考，多对比，从而温故知新，想法就来了。

复习前一次课：概念

数据

数据是所收集、分析、汇总的，用以描述和解释事实的文本、数字与编码 (密文，图片、音频和视频等)

- 本课程的教科书只讨论简单的分类数据 (文本) 和数值。
- 编码类的，特别是图片 (如人口登记的相片)、音频和视频等没有讨论。
- 准确性, 及时性, 完整性, 系统性 (全面), 还要考虑成本

数据类型

量化尺度

分类数据, 顺序数据, 间距数据, 连续数据

1. 4.ppt.6–8
2. Wiki: Level of measurement 里的 Comparison 表格

常用类型

- 来源：原始数据与次级数据
- 量化尺度：品质型数据与数值型数据
- 时空维度：截面数据、时序数据和面板数据
  1. 使用面板数据的动机
  2. 政策分析: pooled cross-sections and two-period panel data analysis

数据获取

原始数据

对目标对象进行观察、交流与实验，直接取得所需要的数据。来自统计调查

- 调查方法
  - 观测 (如气温，客人流量)
  - 访问法 (邮件、电话、电邮调查法; 网络调查法)
  - 实验 (试点)
- 调查方式: 总结 4.ppt.27; 展开 4.ppt.19–27
  - 抽样调查
    - 概率型抽样: 组织形式有简单随机抽样、分层抽样、等距抽样、整群抽样、多阶段抽样
    - 非概率型抽样: 方便抽样, 判断抽样, 滚雪球抽样, ...
  - 普查：为某一特定目的而专门组织的一次性全面调查方式
  - 统计报表制度：自上而下统一布置，自下而上提供基本统计数据
  - 重点调查：选择少数重点单位进行调查
  - 典型调查：选择一个或几个有代表性的单位（典型单位）进行深入细致调查

次级数据

通过查阅或询问获得他人已经加工、整理的可使用的现存数据。来源有统计年鉴、有关期刊和相关网站

- 官方统计网站的次级数据 ppt.47
- 商业与会计数据提供的次级数据 ppt.54
- 中国部门统计与行政记录的次级数据
- 专为研究提供的次级数据
- 统计推算的次级数据

统计调查

调查方案

统计调查: 根据统计研究的目的和任务，在对统计工作内容和程序作出通盘考虑和安排的基础上，运用统计学的理论与方法，有计划、有组织地向各个体搜集资料的过程

1. 确定调查目的 [目的和任务有时很难区分]
2. 确定调查对象和调查单位
3. 确定调查项目 (内容)
4. 确定调查时间和调查期限
5. 明确调查方式和调查方法 [幻灯片: 原始数据]
6. 制定调查的组织实施计划

调查载体

体现能力，但不适合考试

1. 调查表：表头、表体、表外附加，填报说明
   - 单一表，一览表
2. 调查问卷：开头、甄别、主体和背景，调查过程记录
   - 设计：ppt.42–4
   - 提问技巧，冗余校验 (不诚实回答)，问题顺序
   - 五花八门：“五个要素、七大原则”

误差

- 抽样误差: 样本统计量与总体参数常常不相等。
  - 代表性误差的影响因素：样本量的大小，总体的变异大小，抽样的方法和组织形式 (我搞不清方式与形式的区别)
  - 系统性误差: 非随机性引起的
    - 抽样框误差: 涵盖不足, 过涵盖, 非一一对应, 内容偏差 (辅助信息), 老化
- 非抽样误差
  - 无回答误差: 遗漏, 失联, 放弃
  - 计量误差: 设计不周 (主观喜好、利益歪曲), 被调查者 (记忆、倾向), 调查者 (诱导), 工具 (仪器、环境)
  - 登记误差 (编码, 记录, 抄录, 汇总计算)

误差的控制: 尽量扩大样本量；调查全程的质量控制 (设计、培训、督导、评估、奖惩)

总结

1. 数据 (理解)
2. 数据的收集 (掌握)
3. 认识统计调查担负提供基础统计资料的任务 (理解)
4. 调查方案的制定 (掌握, 自己实践)
5. 二手资料的获取 (掌握)

Keyword: 数据, 分类数据, 顺序数据, 间距数据, 连续数据, 面板数据, 原始数据, 调查方法, 调查方式, 次级数据, 统计调查, 调查表, 调查问卷, 抽样误差, 非抽样误差

Homework: ($*$ by group)

- Time loss of smart phones, pads and computers in classroom
- 基本概念要牢记，牢记，牢记!
- $*$XBRL: 近期前 5 大股东及其持股比例, 近 5 年的基本每股收益 (6006x6, 6007x7 其中 x 为小组号)
- Read
  1. Textbook: S2.3–5 (S2.5 的数据)
  2. Reference book: corresponding chapters

2 数据整理

- 分组
- 统计指标
- 展示
- 总结

统计整理

统计调查得到数据，看不出整体的分布是什么样子，有什么特征。

根据统计研究的任务与要求，对统计调查所搜集到的原始资料进行分组、汇总，使信息条理化、特征明晰化，归纳出说明性概要，并系统地将数据的分布和变动等规律用图表直观展示的工作过程称统计数据的整理。步骤

1. 审核和订正：审核它们是否准确、及时、完整
2. 分组：筛选，排序；数据分组的恰当与否关系到能否显示现象的根本特征
3. 汇总和计算：计算各组的总数和合计总数，各组指标和综合指标等。
4. 展示：编制统计表与绘制统计图
5. 积累、保管和公布

分组

分组

分组是根据统计研究的需要，将统计总体按照一定的标志区分为若干组或部分的一种统计方法。数据分组是贯穿于整个统计工作过程的重要方法，它的主要作用有以下几个方面：

1. 划分现象的类型；
2. 研究现象的内部结构；
3. 分析现象之间的依存关系 (Lorenz 洛仑兹曲线: 人口收入比重与人口比重 →Gini 基尼系数)

统计整理中统计分组是关键：组内差异尽可能小, 组间差异尽可能大

- 分类数据：分组、频数和频率
- 顺序数据：+ 累积频数和累积频率
- 数值型数据：+ 组中值, 建议增加“组内均值和组内标准差”
  - 单值分组 (单项式): 把每一个具体的变量值作为一组 [变量值较少的离散变量]
  - 区间分组 (组距式): $K = 1 + 3.322 \log_{10} N$

组限和组中值

确定组限的细节问题

- 间断式组限：保证无数据限落在相邻两组上下限之间 (年龄 0-6 7-15 16-59 60 以上)
- 连续式组限：遵循“上限不在内”原则 (年龄 0-7 7-16 16-60 60 以上)
- 第一组的下限：应略小于或等于最小变量值
- 最后一组的上限：应略大于或等于最大变量值
- 开口组：第一组采取“xx 以下”；最后一组采取“xx 以上”

组中值= (本组下限值+本组上限值)/2

- “xx 以上”组：组中值=本组下限值+相邻组组距/2
- “xx 以下”组：组中值=本组上限值−相邻组组距/2
- 间断式组限：组中值= (本组下限值+后组下限值)/2. 也就是说先转换成连续式再计算组距，例如 “0-6 7-15 16-59 60 以上”转为 “0-7 7-16 16-60 60 以上”

次数分配

总体中的所有单位按一定标志分组整理，并将各组按一定顺序排列，形成总体中各个单位在各组间的分布。

数分配 (分配数列) 是进行统计分析的重要方法，是统计资料整理的一个重要结果。它可以表明总体的分布特征以及内部结构情况，并可据此研究总体某一标志的平均水平及其变动的规律性: 钟形, J 形和 U 形

统计指标

统计指标

主课本里没有集中讨论

- 总量指标：反映被研究总体的单位总数或它的某种标志的总量。计算相对指标和平均指标的基础
  - 时期指标与时点指标: 时点指标各时期数值不能累计数值大小与时期长短无关系
- 平均指标：是概括地反映现象的一般水平的综合指标 [$\frac{\text{变量值总和}}{\text{个体数目}}$, 以后专门讲]
- 相对指标：除了平均指标外，两个互有联系的指标对比所得的结果都是相对指标。
  - 分类：结构相对指标 (恩格尔系数 $=\frac{\text{食物支出金额}}{\text{总支出金额}}$), 比例相对指标 (5 至 9 岁男女性别比例 123.05:100), 比较相对指标, 计划完成程度相对指标, 强度相对指标 (人口出生率 12.9‰; 人口密度 141 人/平方公里), 动态相对指标
  - 构建、计算与运用相对指标时：正确选择对比的基数, 指标间对比要有可比性, 指标与基数结合使用, 多种指标结合使用

展示

统计表

表头, 行列标题, 数据区 ppt.50

统计图

一图知万言: 通过几何图形或具体事物的形象和符号来表现统计资料的方式

统计图的分类 (自己归类)

- 几何图: 利用点、线、面等几何图形来表示统计资料的统计图形
- 象形图: 是利用事物本身的形象来表明现象的数量特征的统计图形
- 统计地图: 是利用点、线、面或事物的形象在地图上显示现象的分布状况的统计图形

统计图的构成: 图号, 标题 (图题), 图注, 坐标系 (轴标题, 标度 $\rightarrow$ 图目, 刻度 $\rightarrow$ 图尺, 网格线), 图例, 图形 ppt.53 (有图例, 但没有标出)

种类与选择

各种统计图的用途 ppt.55–71

- 分类数据的显示：柱形图、条形图、饼图、圆环图
- 顺序数据的显示：柱形图、条形图、饼图、圆环图、累积频数分布图、累积频率分布图
- 数值数据的显示：直方图、折线图、曲线图
- 两个变量之间的关系：散点图
- 三个变量之间的关系：气泡图
- 多个变量之间的关系：雷达图

茎叶图：探索性数据分析。优点：完整数据；随时添加

总结

1. 数据的整理: 分组方法和汇总技术 (掌握)
2. 总量指标和相对指标的概念和特点 (掌握)
3. 统计表的编制并能熟练运用 (掌握)
4. 统计图的绘制并能熟练运用 (掌握)

Keyword: 数据分组, 组限, 组中值, 次数分配 (分配数列), 总量指标, 相对指标, 统计表, 统计图

Homework: ($*$ by group)

- Time loss of smart phones, pads and computers in classroom
- Y70-stu.xlsx + Y70-stu.docx, 熟练一下 Excel. 准备好 E2-3, E2-4 数据
- Read
  1. Textbook: review S2.3–5
  2. Reference book: corresponding chapters

3 软件操作

- Excel 应用
- 总结

统计软件

统计软件很多，如 SAS, SPSS, Stata, S-PLUS, Minitab; R, Python

- SAS 把数据存取、管理、分析和展现有机地融为一体; 药厂要用 (FDA 认证)
- SPSS 傻瓜直接，初学动动鼠标就可以
- Stata 异军突起, 在中国可能有盗版的功劳
- R 堪称是统计学家的计算机语言，但 R 入门难
- Python 是一门多功能的语言。数据挖掘

软件只是工具，要对基础理论有足够了解，才能正确使用软件。课程学习用 Excel, 生产工作时用专业软件 (需求是多样的, 选择合适的、有效率的, 不存在一劳永逸什么都最 NB 的)

- Excel 的数据分析插件 XLSTAT，也能进行数据统计分析
- 随着问题的深入，Excel 就不那么“傻瓜”：统计方法不全，比如 t 检定和方差分析 (新版本才提供), 必须在充分知道其原理和手法后才能实现

Excel 应用

Excel

Microsoft 公司推出的 Office 系列产品之一，是一个功能强大的电子表格软件。凡是有 Microsoft Office 的计算机，基本上都装有 Excel

Excel: 它严格说来并不是统计软件, 对表格的管理和统计图制作功能强大

- 数据透视功能: 一个数据透视表演变出 10 几种报表
- 统计分析: 常用的检验方式一键搞定
- 图表功能: 傻瓜化
- 高级筛选, 自动汇总功能: 简便灵活

如果 数据 菜单下面没有 数据分析 项，请搜索“加载项”并勾选“分析工具库”

例子

- Y70 数据透视
- E2-3
- E2-4

总结

1. Excel 数据透视 (掌握)
2. Excel 图表功能 (掌握)

Keyword: 数据透视表, 柱形图, 条形图, 饼图, 圆环图, 累积分布图, 直方图、折线图、曲线图, 散点图, 气泡图, 雷达图

Homework: ($*$ by group)

- Time loss of smart phones, pads and computers in classroom
- Read
  1. Textbook: S3.1
  2. Reference book: corresponding chapters
  3. 平均数的是与非;“平均数”掩盖了什么 (光明日报 2009-04-29; 2012-02-04 还有)