统计学原理(Statistic)

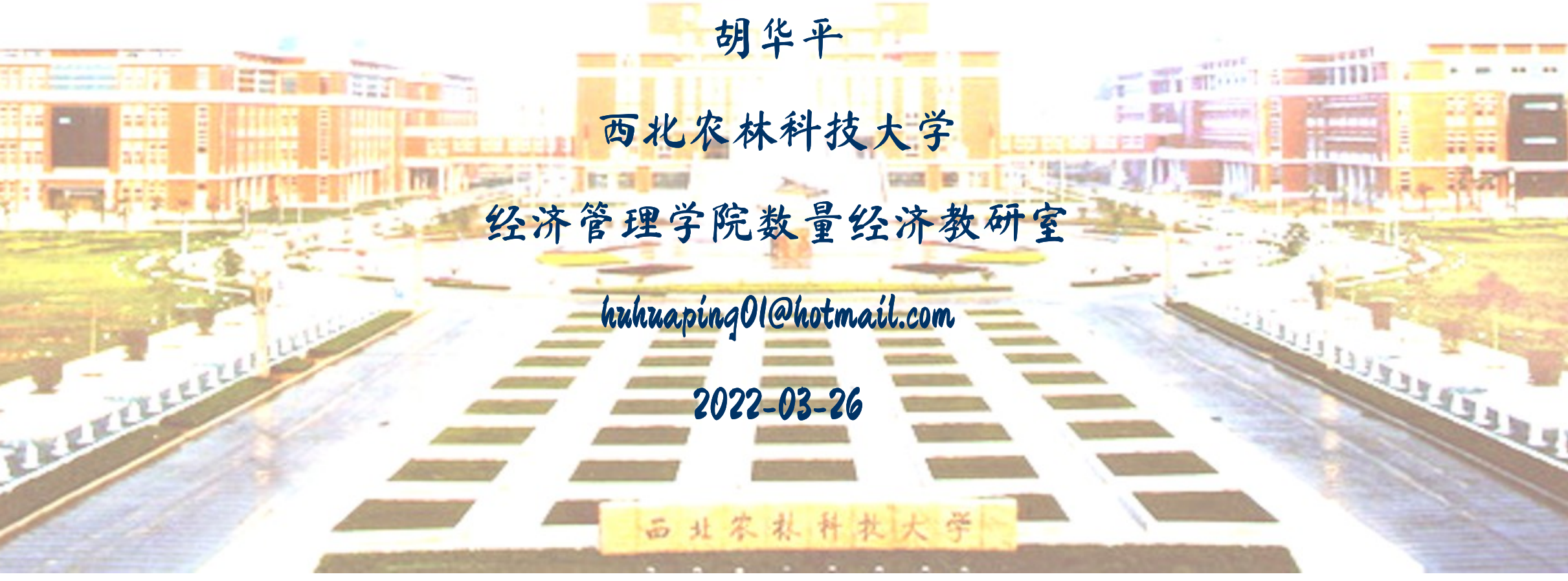
胡华平

西北农林科技大学

经济管理学院数量经济教研室

huhuaping01@hotmail.com

2022-03-26



第二章 数据收集、整理和清洗

2.1 数据目标

2.2 数据收集

2.3 资料整理和数据清洗

2.4 数据的数据库化

2.5 数据质量

2.6 抽样设计

2.7 抽样分布和抽样误差

2.8 问卷设计技术

2.5 数据质量

数据质量内涵与评估

影响数据质量的主要原因

数据质量（评判原则）

> 没有质量的数据，就是垃圾。“垃圾进，垃圾出”。
> 数据的质量会受哪些因素的影响？如何评估数据质量？

数据质量评估是一项专门的技术，对不同来源的数据，有不同的评估方法。判断数据质量的基本原则有三项：

1. **真实性**。真实性指的就是数据确实来源于调查，与数据产生有关的过程真实存在，调查对象真实存在；访问、观察真实存在；应答、场景、文献真实存在。
2. **准确性**。数据的调查人员准确按照研究设计在执行，准确地处理了调查对象和调查对象的反馈，或者是，准确地转录了原始数据。
3. **时效性**。 对于有时效要求的数据，还要考虑调查的 实施过程是不是符合规定的时间要求。 如果上述三项原则都能得到满足， 就可以进一步考察数据的基本质量，那就是符合性。

数据质量（评判维度）

对于数据质量的评估，总体上有两个维度：

1. **正向评估**，是与标准要求的距离到底有多远，也就是符合性问题。
2. **反向评估**，就是误差的大小。

数据质量（误差分类1）

事实上，数据收集、整理、清理的每一个环节都有可能产生误差。

1. 覆盖性误差。就是涉及到调查对象的备选机会而可能产生的误差。抽样问卷调查、访谈调查、观察调查、文献调查都有可能产生覆盖性的问题。
2. 测量性误差。就是调查数据中可能产生的误差。调查大都涉及到测量的信度和效度。只要信效度有问题，那么测量性误差就可能存在。
3. 应答性误差。观察调查、文献调查看起来没有应答类型的问题，实际上不是。只要是访问员提出的要求都存在应答类型的问题。只是不同类型的调查，应答性误差的表现形式、计算方法不同。因此，应答性误差也是调查数据中可能存在的误差。
4. 抽样性误差，仅出现在抽样问卷调查中的一类误差。

数据质量（误差分类2）

以上误差，如果依据在调查活动中的**可改进性**来看，又可以被归纳为两类误差：

1. 随机误差，就是在调查活动中随机产生的误差.
   - 比如访问员的不规范行为产生的误差。通过规范访问员行为就可以减少这一类型的误差。在表现形式上，这类误差会增大变量测量的方差。
2. 系统误差，是由设计因素影响所产生的误差。
   - 比如测量工具带来的误差，由于测量工具有问题，导致凡是采用这个测量工具的调查都会产生同一类、甚至同样程度的误差。 在表现形式上，这类误差会增大测量的偏移量，就是bias。

调查总误差：所有这些由数据收集、整理、清洗活动产生的误差的综合，被称之为**调查总误差**。通常用**均方误（MSE）**来表示。

覆盖性误差（概念）

覆盖性误差，又称为**抽样框误差**，指的就是**目标总体**与**抽样框总体**不一致所导致的调查对象错位所产生的误差。

覆盖性误差存在于所有通过调查方法获取数据的研究活动中。

- **目标总体**就是调查对象总体，有明确的调查对象所指。
- **抽样框总体**，简称框总体，是用于抽样的所有调查对象的集合。
- **样本总体**，是被抽中的，且被作为调查对象的集合。

> 文献调查中，已知需要查阅的涉及某件事的所有文献，在查阅之前，却打算把查阅文献的范围扩大或者缩小，这就产生了覆盖性误差。

覆盖性误差（来源）

1. 丢失或者重叠目标总体要素。
   - 框总体小于或者看起来大于目标总体，进而让部分要素失去或者获得了多次被抽中的机会，这里既有覆盖过度的现象，也有覆盖不足的现象。
   - 比如在“北京大学本科生入学机会地区不平等”的调查中，如果以已经入学的学生为总体，丢掉了某个院系， 或者既用院系、又用地区做抽样框，就会产生丢失，或者重叠问题。
2. 在抽样框总体中，包含着非目标总体要素。
   - 这会使得况总体看起来会大于目标总体，进而让目标总体的备选概率小于理论概率
   - 比如“北京大学本科生入学机会的地区不平等”研究，把北京大学的保安纳入到了抽样框，就会让目标总体学生的备选概率降低。
3. 不正确的辅助信息。

覆盖性误差（影响）

那么覆盖性误差对调查误差到底会有怎样的影响呢？

- 如果是**抽样问卷调查**，那么就会通过影响等概率，进而影响到代表性，影响了代表性，就影响到数据质量。
- 在**非抽样问卷调查**中，虽然不存在影响等概率的问题，但覆盖性问题依然存在，只是表现形式不同而已。如果覆盖过度，虽然不会对调查数据质量造成可计算的影响，却可能会干扰研究判断，比如冲淡了真正对象的变异性或者影响。如果覆盖不足呢，则有可能对研究判断造成致命的影响。
    - 在文献调查中，缺失了最关键的文献就有可能会认为没有这类文献，进而出现错误判断。
    - 在访谈调查中，如果没有访问到事件的当事人，就有可能出现关键信息不全甚至缺失，进而也导致错误的判断。
    - 在观察调查中，漏掉了关键的场景，比如研究庙会的，却没有去观察某个庙会，就无法对场景的现象做正确的判断。

测量性误差

测量性误差，指来源于测量工具的误差，和运用测量工具的误差。

> 在测量长度的时候你拿着尺子来量，尺子很准，很可靠，不过呢你的眼神不好，测量过程就有可能带来误差。
> 如果工具不好，即使你非常认真，也会产生误差。如果工具很好，没有用好，也不行，也会产生误差。

两个来源的误差都会反映在测量的质量参数上来，这就是**信度**和**效度**。

- **信度测量**：前后测信度、折半信度、复本信度、一致性信度。
- **效度测量**：表面效度、准则效度（校标效度）、建构效度、内在效度。

信度和效度的测量是针对**结构式测量**的。事实上**无结构式调查**中也同样存在信度和效度问题。只是因为没有结构，对信度和效度的测量比较困难而已。

测量质量的检验1（信度）：概念

信度（reliability）：是指测量工具的可靠性，也即使用同一个测量工具、重复测量同一个对象，得到相同结果的概率。

- 得到相同结果的概率越高，测量工具的信度也就越高。
- 信度对测量而言，就是测量工具的稳定性。
- “重测信度”，就是看前后之间有没有差异，前后之间的差异越小，信度就越高。

测量质量的检验1（信度）：实践类型

假设我们在做调查，用问卷在做调查，用访题在做测量。

- **垂直重复信度**：又叫**前-后测信度**，在实践上一前一后测试两次。适合变量**不随**时间变化的测量。
- **水平重复信度**：又称为**复本信度**，或**等值信度**，也是水平的重复测量。要求测量对象具有等价性。

假设我们有一组访题，一般是5-6道，或者6-7道访题。针对主观变量又如何检验测量的信度呢？

测量质量的检验1（信度）：计算办法4

折半信度法：如果访题的一致性很好，奇数题得分与偶数题得分之间的相关系数也应该很高，如果访题之间的一致性有问题，相关系数也不会高就说明访题的稳定性不高。

- 把访题编号，编成奇偶数。
- 对同一组对象，用奇数题和偶数题分别进行一次测量
- 计算奇数题偶数题得分的相关系数。
- 再用Spearman Brown公式计算信度。

测量质量的检验1（信度）：计算办法B

克隆巴赫系数法，一般记为为“Cronbach α”，主要运用了内部方差原理，也就是如果访题的内部方差越大，则测量的一致性也就越差。

一般表达式为：

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K}\sigma_{Y_i}^2}{\sigma_{X_i}^2}\right)$$

或者也可以表达为：

$$\alpha = \frac{K\bar{c}}{(\bar{v} + (K-1)\bar{c})}$$

其中，$\bar{v}$表示每个题项之间的平均方差，$\bar{c}$表示不同被测者之间在不同题项上的平均协方差。

- $\alpha \in [0.9, 1)$表明测量可靠性极高
- $\alpha \in [0.8, 0.9)$表明测量可靠性较好
- $\alpha \in [0.7, 0.8)$表明测量可靠性能够接受

测量质量的检验2（效度）：概念和类型

效度（validity）：指的是测量工具是否正确和有效。

- **预测效度**：一个试点的测量结果与另一个试点测量结果之间的相关程度，相关程度越高 预测效度也就越高。

> 一模、二模的成绩能在多大程度上预测高考成绩就是模考的预测效度。

- **同时效度**：指的是测量结果与既有的有效测量之间的相关程度，相关程度越高同时效度也就越高

> 笔试与实际能力之间的关系既涉及到预测效度，也涉及到同时效度

测量质量的检验2（效度）：概念和类型

- **结构效度**：指一组题在多大程度上可以测量到理论上期望的特征，或者说在多大程度上能测量到事物之间的关系模式。

> 一组题，能在多大程度上发现婚姻满意度与夫妻之间相互忠诚之间的关系模式。

- **内容效度**：直接测量变量的属性，是指测量在多大的意义上包含了概念的含义。

> 身高和体重，用什么测？

测量质量的检验2（效度）：示例

效度的检验不像信度的检验，总是需要用到统计检验，大多数的情况下都是主观判断。也有复杂的效度测量，难度超出了课程的要求。

北京大学本科生入学机会地区不平等的研究案例：

- 把**地区之间的差距**操作化为**地区之间的人均GDP**，虽然测量起来比较容易可是测量的并不是每个地区人们可以用于教育的资源。
- 测量人均GDP倒是很稳定，**信度**很好，却没有很准确地测量到我们希望测量的、可以用于教育的资源。
- 如果追求**效度**，测量每一个毕业生家庭可以用于教育的资源，虽然测量到了要测量的内容，可是测量起来却很困难。

应答性误差（概念）

应答性误差，是指访员发出了调查请求，调查对象却没有做出回应或者做出应答，由此带来的直接后果就是调查数据的缺失，从而引起数据误差。

在不同的调查中，应答性误差的表现形式并不一样，

数据缺失如果是样本层面的、对象群体层面的、场景层面的、文献类别层面的，那么应答性误差就可以被理解为**广义覆盖性误差**中的一种。

即使我们获得了等概率样本，或者必须调查的对象列表，在调查中，调查对象拒访、场景不可及、文献不可及等等情况总是会有的。

即使接受了访问，场景也可及，文献也找到了，可是某几道访题受访者不作答，或者不知道如何作答；或者没有遇到具体的场景。

> 希望看婚庆，但没有遇到有人结婚；或者文献中的某几页缺失了。

如此，就相当于覆盖不足，或者数据缺失。

应答性误差（概念）

无应答从类型上看主要有两种。

1. **对象无应答：**
   - 在抽样问卷调查中，常常被称之为**样本无应答**，或者**单元无应答**，英文是**unit nonresponse**；
   - 在非抽样问卷调查中，对象无应答被称之为**“失访”**，就是没有接触到、观察到或者访问到设计中需要调查的对象、文献、痕迹。
2. **某些议题没有得到应答：**
   - 在抽样问卷调查中，如果部分访题没有得到应答， 就会被称之为**选项无应答**，又被称之为 项目无应答，英文叫**item nonresponse**。
   - 在非抽样问卷调查中，指一个或者具体几个议题，没有“访到”，自己忘记了、遗漏了，或者缺失了。

应答性误差（应答率）

在抽样问卷调查中，**应答率**是评估数据质量的基本参数之一。应答率等于**应答样本数**除上**样本总数**，再乘上百分之一百。

从**分子**角度来看：

- 一种情形是完全应答，完成了所有应回答的访题。
- 另一种情形是如果只是部分地应答了，没有完成所有应该回答的访题呢，那么到底完成了多少算是应答了呢？通常会根据访题的数量算出一个百分数，也就是完成了百分之多少访题的应答率是多少。

从**分母**角度来看：

- 无效的样本，比如不符合样本约束条件的对象；
- 未接触到的样本，也不知道是不是符合样本的约束条件；
- 接触到了，却完全无应答的样本；
- 即使没有接触到，却被认为是有效的样本；

应答性误差（影响）

应答率对数据质量有什么影响呢？

假设应答率为 p，无应答率其实就是 $1-p$，由于无应答既可能是**随机现象**，也可能是**系统现象**。

- 随机现象，比如某个访题遗漏了，某个样本遗漏了；
- 系统现象，比如高收入的人群完全接触不到。

因此，无应答对样本估计值的影响主要来自于满足约束条件的样本的无应答，对代表性的影响。

- 高收入人群完全访问不到就会造成这一部分人群没有样本，进而影响到让样本满足等概率性。

抽样性误差（内涵）

在抽样调查中，覆盖性误差、测量性误差、 应答性误差，三类误差都是可计算的。

抽样调查中抽样性误差的来源

- 主要来自于制作抽样框时候形成的误差，比如对样本的覆盖性。换句话说，在抽样调查中，覆盖性误差其实是抽样性误差的一部分。
- 还有在抽样过程中形成的误差，比如分层、多阶段，尤其是在末端抽样中，采用的方法、抽样的人都有可能形成误差。

在文献调查中，因为使用二手文献、因为选择版本等所带来的误差；

在观察调查中，因为选择场景所带来的误差。

在访谈调查中，因为访谈对象变动所带来的误差。

抽样性误差（计算）

抽样误差的计算也是针对具体变量的。

> 抽样的目的是为了获得有代表性的样本；获得有代表性的样本是为了用样本推论总体，误差尽可能的小；而推论是针对具体变量的推论；可是任何一项调查，误差总是要体现在这个变量上的，没有变量，哪来的误差呢？

1. **均值的变异系数**。等于样本均值除以标准误，也即 $\frac{\bar{X}}{\sigma}$。如果是比例值，则为 $\frac{p}{\sqrt{p(1-p)}}$。经验上，如果一项调查样本均值的变异系数小于50%，就认为质量是可以接受的。
2. **样本均值的相对方差**。等于样本方差除上均值的平方，也即 $\frac{\bar{X}}{\sigma^2}$。如果是比例值，则为 $\frac{p}{p(1-p)}$。

本节结束