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

white[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

2.6 抽样设计

抽样的要素

抽样的逻辑

概率/非概率抽样

非概率抽样

抽样方案和实施

抽样误差

什么是抽样

假设我们希望通过自己调查来获得一手数据，就需要回答一系列问题：

1. 抽样的基本原理是什么？
2. 抽样的基本要素有哪些？
3. 抽样的逻辑是什么？
4. 什么条件该要采用概率抽样方法？怎么样做概率抽样设计？
5. 什么条件下要采用非概率抽样方法？又如何做非概率抽样？
6. 一个抽样方案应该包括哪些内容？
7. 怎么样去实施抽样工作？

抽样的要素（总体）

总体：是研究问题指涉对象的集合体，也就是研究问题涉及的全部对象。

- CFPS的总体是中国所有的家庭户
- CGSS的总体是中国所有的个体
- 入学机会的地区不平等研究的总体，就是某年所有的高中毕业生。

问题是：

- 什么叫中国所有家庭户，中国所有个体？
- 什么叫所有，台湾算不算？香港和澳门算不算？
- 住在中国的还是有中国户籍的？住在中国的外国人算不算？
- 长期出国却依然有着中国户籍的人算不算？
- 什么叫家庭户？没有生活在一起，户口在一起算不算？生活在一起，
- 户口不在一起的，算不算？怎么才算是某个地方的家庭户？
- 户口在甲地，却很少在甲地居住，算不算甲地的家庭户？
- 什么叫所有高中毕业生？没有参加高考的算不算？因非主观原因有参加高考的算不算？

抽样的要素（研究总体）

研究总体，是指可操作的研究对象，或称为**可及总体**。

CFPS把总体定义为中国的家庭户，是指有中国户籍的家庭户，指住在一起的，不管户籍是不是在一起的家庭户。

提问：

- CFPS中，家庭户指居住在二十五个省级单位内的家庭户吗？
- 户籍不在本地的算不算？
- 住又是什么意思呢？
- 住多长算是住？
- 一个人打工住在本地算是一户吗？

抽样的要素（抽样框和抽样单位）

抽样框：又叫抽样总体、框总体，是从研究总体中获得的用于抽取样本的研究对象的集合。

- CFPS的**总体**是中国所有的家庭户；
- CFPS的**研究总体**是二十五个省、市、自治区的常住户；
- CFPS的 **抽样框**是二十五个省、市、自治区在一个地方连续居住六个月或以上常住户。
- 从覆盖面和覆盖的对象数量出发，**总体** $\geq$ **研究总体** $\geq$ **抽样框**。

抽样的要素（抽样单位和样本）

抽样单位：是抽样指涉的基本单位，或包括基本单位的单位集合体。

- CFPS在抽到家庭户之前还要抽样本区县，样本村居。每一次抽样面对的基本单位 就是抽样单位。

样本：是从抽样框中运用抽样策略和抽样方法获取的样本单位的集合。

- CFPS的样本是，从25个省市自治区抽取的160个区县样本，从160区县样本中抽取的640村居样本，从640村居样本中抽取的16000个家庭户样本。

抽样的逻辑（样本代表性）

抽样的基本逻辑1：选择一定数量的样本，来拟合总体中个体变异性的分布，进而代表总体。

抽样的基本逻辑2：用尽量少的样本，在可接受的误差范围内，来代表总体的研究特征。

柯西的思想：用代表性的样本就可以估计总体的研究特征。

> 柯西曾去美国国会作证，他反对在美国实施人口普查，认为每十年一次的人口普查，耗费太多的资源，实在没有必要。

如果个体在总体的分布是随机的，根据随机性原则抽取的样本就能代表总体，就是**代表性样本**。

在研究实践中，样本与总体之间总是有差异的。即时在随机条件下，尽管每个抽样单位被抽中的概率是相等的。由**样本特征**与**总体特征**之间，总是有差距的。

抽样的逻辑（抽样误差）

抽样误差：是样本研究特征与总体研究特征之间的差异。误差的大小一般取决于样本的代表性。样本对总体的代表性越好，误差就越小，否则误差就会越大。

依据误差的来源环节，可分为：

1. **随机误差**：误差就是由抽样环节造成的误差。随机误差是希望**尽量**避免的误差。
2. **系统误差**：误差具有规律性，主要是由抽样设计造成的。系统误差是我们**最应当**避免的。因为一旦出现的系统误差，几乎就没有补救的余地，
   - 假设希望知道性别与成就之间的关系。严格按照抽样方案完成的抽样，抽到的样本却都是男性的，没有女性。

抽样的逻辑（抽样误差）

依据抽样活动涉及的对象，可以把误差来源分为：

1. **覆盖性误差**：是抽样活动没有正确的覆盖需要覆盖的总体，要么对总体覆盖过度，要么覆盖不住，过度和不足都会导致误差。
   - 假设界定的**总体**为参加高考的高中毕业生
   - 如果在抽样中把自愿或者是因为其他原因没有参加高考的毕业生都纳入到了抽样的范围，这就是覆盖过度。
   - 如果我们把复读并参加了高考的学生排除在了抽样的范围， 这就是覆盖不足。
2. **选择性偏差**：在设计与执行中，因偏好或者抽样活动而导致某个特定类型的样本的分布出现问题。
   - 某一类人群过多或者过少或者缺失。
   - 某个人群不在抽样框，被选机会就没了。

抽样的逻辑（样本分布）

抽样的逻辑3：利用重复多次抽样，提高抽样代表性，减小抽样误差。

抽样分布：又称统计量分布，指样本估计值的分布。抽样分布可以用来测量抽样方法的稳定性。

总体分布：是指总体特征值的分布。总体分布并不总是可得的，即使可得，也不满足经济性原则。

抽样方式（总览）

概率抽样：就是运用等概率原则进行抽样的总称。**等概率原则**，是指总体中每一个研究对象被抽中的概率是相等的。包括：简单随机抽样、系统抽样、整群抽样、与规模成比例的概率抽样、分层抽样以及隐含的分层抽样、多阶段混合抽样。

非概率抽样：抽取样本时不是依据随机原则，而是根据研究目的对数据的要求，采用某种方式从总体中抽出部分单位对其实施调查。包括：方便抽样、判断抽样、自愿样本、滚雪球抽样、配额抽样。

从抽样方式的具体运用，又可分为：

- **直接抽样**：一次抽样或独立抽样。简单随机抽样、系统抽样和整群抽样都是直接抽样
- **半截抽样**：通常不可以独立地用，要结合前直接抽样来使用。规模成比例的概率抽样、分层抽样以及隐含的分层抽样、多阶段混合抽样都属于半截抽样。

概率抽样1：简单随机抽样（步骤方法）

简单随机抽样(simple random sampling)：从总体N个单位中随机地抽取n个单位作为样本，每个单位入抽样本的概率是相等的。它是最基本的抽样方法，也是其它抽样方法的基础。

实施方法：

- 第一步，制备抽样框
- 第二步，对要素进行编码
- 第三步，根据抽样的要求抽取样本。
  - 直接抽选法
  - 抽签法
  - 随机数码表法（或kish table）
  - 软件抽取法

概率抽样1：简单随机抽样（优缺点）

优点：

- 简单、直观，在抽样框完整时，可直接从中抽取样本
- 用样本统计量对目标量进行估计比较方便

缺点：

- 当N很大时，不易构造抽样框
- 抽出的单位很分散，给实施调查增加困难
- 没有利用其它辅助信息以提高估计的效率
- 使用随机数表抽样的效率往往比较低下，即使用到，也会使用随机数表的一些变体如kish table

概率抽样1：简单随机抽样（kish table）

> Kish, L. (1949). A Procedure for Objective Respondent Selection Within the Household,Journal of the American Statistical Association, 380-387.

- 第一步，制备末端抽样框，将样本家户所有符合要素资格的成员，按照规则顺序编号，依据性别也好，年龄也好，逆序也好，顺序也好，怎么排都行，要求是不重，不漏。
- 第二步，拿出事先准备好的kish表，根据指引，抽取样本。抽样的约定是不管家里有几个要素，只抽取其中的一个要素作为样本。

概率抽样1：简单随机抽样（kish table）

Household	Kish Table
1	A
2	A
3	B1
4	B2
5	C
6	C
7	D
8	D
9	E1
10	E2
11	F
12	F
13	A
14	A
15	B1
16	B2
17	C
18	C
19	D
20	D
21	E1
22	E2
23	F
24	F
25	A
26	A
27	B1
Etc……..	

Sequentially work down the list

Selection table D

If the number of adults in household is:	Select adult numbered:
1	1
2	2
3	2
4	3
5	4
6 or more	4

Overall Selection Probabilities

	If the number of adults in household is:					
Adult numbered	1	2	3	4	5	6 or more
1	1	1/2	1/3	1/4	1/6	1/6
2		1/2	1/3	1/4	1/6	1/6
3			1/3	1/4	1/4	1/6
4				1/4	1/6	1/6
5					1/4	1/6
6						1/6
7 or more						0

概率抽样1：简单随机抽样（软件实现）

利用**统计软件**能快速实现简单随机抽样:

- SPSS
- excel
- R

简单随机抽样的两点忠告:

- 简单随机抽样是不得已的办法，不是最先选用的办法
- 只有在总体的信息所知甚少的情况下，才用它。

概率抽样1：简单随机抽样（R示例）

任务：从教学班上随机抽取6人。

- 第一步，确认当前的班级是样本班级，制作抽样框。
- 第二步，对班级的83位同学从1到83实行顺序编码。编码顺序可以按学号、按座位等，只要是有规则，并且保证每一位同学只有一个唯一的编号就行。
- 第三步，选择一个随机数表，大家可以找到很多的随机数表（教材附录）。在查阅随机数表之前，说出第一个样本的行列位置作为起点。
- 第四步，在随机数表上找到上面的起点，然后取一组随机数的固定位置，按照事先制定的规则，依次选中随机数字中的一位。

概率抽样1：简单随机抽样（R示例）

全体学生名单（按班级和学号排序）：共83人。

序号	学号	姓名	班级
1	2017014588	马丽	农管1801
2	2018011379	安舒心	农管1801
3	2018013874	刘照润青	农管1801
4	2018014553	李铮	农管1801
5	2018014556	张晓旭	农管1801

Showing 1 to 5 of 83 entries

Previous 1 2 3 4 5 … 17 Next

概率抽样1：简单随机抽样（R示例）

不放回-简单随机抽样：

从1-83中产生6个随机数：

```
choice <- base::sample(1:n, size = 6, replace = FALSE )
choice
```

```
[1]  5 23 82 17 80 36
```

概率抽样1：简单随机抽样（R示例）

不放回-简单随机抽样：

抽取到的6个学生：

序号	学号	姓名	班级
5	2018014556	张晓旭	农管1801
17	2018014636	胡亚宁	农管1801
23	2018014660	王赛	农管1801
36	2018014711	王佳鹏	农管1802
80	2018014790	孔雨欣	营销1801
82	2018014805	王楠	营销1801

概率抽样1：简单随机抽样（kish table）R示例1

下面是一张kish随机数表：

一份kish随机数表

adults	A	B1	B2	C	D	E1	E2	F
1	1	1	1	1	1	1	1	1
2	1	1	1	1	2	2	2	2
3	1	1	1	2	2	3	3	3
4	1	1	2	2	3	3	4	4
5	1	2	2	3	4	3	5	5
>=6	1	2	2	3	4	5	5	6

概率抽样1：简单随机抽样（kish table）R示例2

假设需要调查共30户家庭，并对每户的成年人进行了编号：

id	adults.num
1	7
2	7
3	3
4	6
5	3
6	2

Showing 1 to 6 of 30 entries

Previous 1 2 3 4 5 Next

概率抽样1：简单随机抽样（kish table）R示例3

按家庭成年人总数做分类，结合kish表可以得到：

id	adults.num	adults	A	B1	B2	C	D	E1	E2	F
1	7	>=6	1	2	2	3	4	5	5	6
2	7	>=6	1	2	2	3	4	5	5	6
3	3	3	1	1	1	2	2	3	3	3
4	6	>=6	1	2	2	3	4	5	5	6
5	3	3	1	1	1	2	2	3	3	3
6	2	2	1	1	1	1	2	2	2	2

Showing 1 to 6 of 30 entries

Previous 1 2 3 4 5 Next

概率抽样1：简单随机抽样（kish table）R示例4

进一步地，每户都可以在8张表（A-F）中做出随机选择：

id	adults.num	adults	table	select
1	7	>=6	A	1
1	7	>=6	B1	2
1	7	>=6	B2	2
1	7	>=6	C	3
1	7	>=6	D	4
1	7	>=6	E1	5

Showing 1 to 6 of 240 entries

Previous 1 2 3 4 5 … 40 Next

概率抽样1：简单随机抽样（kish table）R示例5

最后随机抽取kish表的结果如下

id	adults.num	adults	table	select
1	7	>=6	A	1
2	7	>=6	E2	5
3	3	3	B1	1
4	6	>=6	E1	5
5	3	3	B1	1
6	2	2	B1	1

Showing 1 to 6 of 30 entries

Previous 1 2 3 4 5 Next

概率抽样2：系统抽样（实施方法）

系统抽样(systematic sampling)：将总体中的所有单位按一定顺序排列（变量要素），在规定的范围内随机地抽取一个单位作为初始单位，然后按事先规定好的规则确定其它样本单位。**系统抽样**也称为等距抽样。

应用情景：

- 总体要素与抽样对象一致
- 总体通常规模也不大
- 变量异质性没有大到需要分层处理的程度
- 要素的特征在排列中没有**周期性**变化

实施方法是：

1. 把抽样框的要素按照规则进行**编码**。
2. 用要素总体数除以样本数，得到**抽样距**（不是整数怎么办）。
3. 选择任何一个**随机起点**，依照抽样距或者顺序抽样或者循环抽样。

> 假设一个班级有50个人，男生25位，女生25位，在排列时，每位男生的后面或者前面都是女生。这样男生跟女生之间的排列就是周期性的排列。万一要素的排列的周期，与抽样距吻合了，抽到的就只有一类样本，从而引起选择性偏差。

概率抽样2：系统抽样（示例）

任务：用**系统抽样**方法从16名学生中随机抽取3人：

- 第一步，把班内所有的的学生名单按照按照学号进行排列。
- 第二步，把排列好的学号，从1开始顺序编号。
- 第三步，假设我们要在16位学生中，抽出3个样本，抽样距为5，样本量为3。
- 第四步，假设我们把要素排列成一个循环圈，选择一个随机起点为编号8，编号8就是第一个样本。顺时针数第5个也就是编号11，就是第二个样本，以此类推。

> 在排列要素的时候，我们不仅可以排列成循环圈，也可以排列为直线。
> 数到16不够测量距了怎么办？回头接着数到编号2，就是第三个样本。

概率抽样2：系统抽样（优缺点）

优点：

- 操作简便，可提高估计的精度

缺点：

- 系统抽样的框不能太大。太大了就很费事，仅就要素编号 就比较费事，
- 要素的排列特征不能呈现周期性变化。

概率抽样3：整群抽样（应用情景）

整群抽样(cluster sampling)：将总体中若干个单位合并为组(群)，抽样时直接抽取群，然后对中选群中的所有单位全部实施调查。

应用情景：

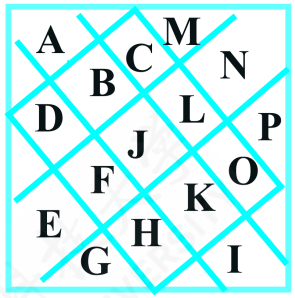
- 是群内具有异质性，不过异质性还没有还没有大到需要专门处理的程度。
- 群之间具有差异性，但也没有大到需要专门处理的程度。
- 通常不作为独立抽样的方法使用，而是用于多阶段、多层次抽样的末端。

概率抽样3：整群抽样（步骤方法）

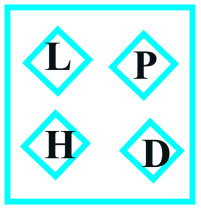
实施步骤：

1. 确定抽样框
2. 根据变量或辅助变量把总体分成若干子群
3. 确定样本容量和样本子群数
4. 依照简单随机抽样方法随机抽取子群

总体群数R=16



样本群数r=4



样本容量

$$n=n_d+n_p+n_l+n_h$$

概率抽样3：整群抽样（示例）

任务：为了分析教学法在16个班级上的效果，请按照**整群抽样法**，抽取90个学生。

注意：

- 采用整群抽样法，是假设了班与班之间特征的差异不大。每个班级同学的学习成绩有一个分布，在班与班之间，具有相似性。
- 整群抽样法实施中，分群过程非常重要。分群的基本原则是：
  - 在选择研究变量或者辅助变量时，让它在群间具有相似性，在群内具有异质性。
  - 如果群内同质群间非常异质，那就不适合用整群抽样了。
- 相似的可以用做分群标准的辅助变量，比如说行政区划、组织、行业、班级、年龄 、性别等等之类。

概率抽样3：整群抽样（优缺点）

优点：

- 抽样时只需群的抽样框，可简化工作量
- 调查的地点相对集中，节省调查费用，方便调查的实施

缺点：

- 估计的精度较差
- 在分群中有一点需要注意，群的规模不宜过大，否则就有可能出现内部同质性。影响抽样的效率，操作起来也很麻烦。

概率抽样4：成比例抽样（原理）

成比例的概率抽样(Probability Proportionate to Size Sampling)： 又称按规模大小成比例的概率抽样或PPS抽样。

原理：

- 如果总体的要素之间在研究变量上有异质性，不同规模要素群体之间异质性的分布不是随机的。在这样的条件下，就要考虑把规模因素纳入抽样的考量了。
- PPS抽样理论上运用了等概率原理，希望让每一个抽样单位被抽中的概率 与抽样单位的规模成比例。

概率抽样4：成比例抽样（实施方案）

实施方案：

- **PPS抽样**是一种按概率的比例抽样，在多阶段抽样中，尤其是二阶段抽样中，初级抽样单位被抽中的机率取决于其初级抽样单位的规模大小
- 初级抽样单位规模越大，被抽中机会就越大，初级抽样单位规模越小，被抽中机率就越小。
- PPS抽样也可以运用软件工具执行
  - Stata工具下ADO模块的gsample或者samplepps。

概率抽样4：成比例抽样（示例）

海淀区西北旺乡有100个社区，4万户。假设抽样要求是，要抽取10个社区，每个样本社区抽取20户，一共抽200户。假设针对海淀区西北旺乡A、B两个社区抽样，其中A社区有2000户，B社区只有500户。

在社区层次来看：

- A社区的总户数占西北旺乡的总户数的比例为 $2000/40000 = 0.05$。
- B社区被抽中的概率为 $500/40000 = 0.0125$。

从社区家户来看：

- A社区家户的被选概率就是 $20/2000 = 0.01$；
- B社区家户的被选概率为 $20/500 = 0.04$。

从整个西北旺乡来看：

- A社区家庭户在西北旺乡的被选概率就是 $0.05 * 0.01 = 0.0005$。
- B社区家庭户的被选概率是 $0.0125 * 0.04 = 0.0005$。

概率抽样4：成比例抽样（规模度量）

概率抽样基本的条件是具有抽样大小规模的辅助变量，又叫**规模度量**。

规模度量如何选择：

- 主要是是代表规模的，比如说社区的家庭户数。
- 规模量度的变量可以有多个，最常用的方法是依据研究变量相关程度来挑选。
- 选择规模度量的影响因素还有获取资料的难易程度、可靠程度等。
- 在两阶段/多阶段抽样中，每一阶段使用的规模度量一定要相同。

> 西北旺乡案例中：第一阶段是社区抽样，被选概率计算还是采用了家庭户数。 第二阶段是家庭户抽样，备选概率的计算也采用了家庭户数。

概率抽样4：成比例抽样（特点）

成比例抽样PPS的特点：

- 第一，PPS抽样常常会考虑抽样面对的现实，一般是进行多阶段抽样，不是抽一回。
- 第二，有些信息，抽样时并不知道，常常要步步为营，充分利用已经知道的信息。
- 第三，每一个阶段的抽样概率不一定相等。
- 第四，总的原则是总体要素的被选概率一定要相等。

概率抽样5：分层抽样（实施步骤）

分层抽样(stratified sampling)：将抽样单位按某种特征或某种规则划分为不同的层，然后从不同的层中独立、随机地抽取样本。

实施步骤：

1. 把研究总体按照研究特征变量进行分层。
2. 在每一层采用合适的方法来抽样
   - 简单随机抽样或者等距抽样、整群抽样
   - 等比例或者不等比例的抽样，甚至pps抽样都行。
3. 把每个层的样本合起来加总，计算得到对总体进行推断的样本容量。

概率抽样5：分层抽样（应用情景）

决定是否采用分层抽样，需要：

- 对研究总体同质性程度有一定了解，知道总体的同质性、异质性如何。
- 了解了总体的异质性的程度是不是大到了必须分层的程度。总体在研究变量上的同质性越高，对分层的要求就越低。
- 分层抽样通常不会独立使用，通常用来构造子抽样框、子总体，它不是独立抽样的方法，也不是末端抽样的方法。
- 对研究变量了解越充分，采用合适的分层方式，就越有利于降低抽样误差。

概率抽样5：分层抽样（示例）

任务：一项研究拟讨论教育模式对高校学生能力的影响，研究者打算采用分层抽样方法抽取n名学生。

- 从学校到院系，简单起见可以先分文和理两个大类的院系。
- 从院系到班，可以采用任何简单抽样的方法。
- 从班抽到学生呢，就可以采用整群抽样的办法。
- 把文和理两类样本加起来，就是一所学校的样本。
- 如果文理之间学生的数量相差的太大，也可以考虑按学生数量的比例分配样本。

概率抽样5：分层抽样（分层依据）

分层依据是分层抽样中关键的环节：

- 分层依据的变量通常与研究目标有关，与研究变量有关系。
- **分层**并不就是**分等级**，大多数情况下是**分类别**（提问）。
- 研究目的越复杂，分层变量越多，要区分的层数也就越多。
- 实践中一般希望尽可能地选取**主要的分层变量**，因为分层越多，看起来越精准。
- 在抽样实践中，有些分层明显，有一些分层则不太明显，可能实际上还携带着层变量的分层，称之为**内隐分层** 或者叫**隐含分层**。

概率抽样5：分层抽样（示例）

学生教育模式对学生能力的影响研究案例。

有的院系一个年级有多个班，如经济管理学院，有的学院只有一个班，如农学院。

如果有多个班的学院用平均能力对班进行排序，再抽取班级样本，则抽到的班样本不仅携带了院系信息，也携带了能力信息。

不仅按文理院系在分层，也在按照能力进行分层，只是按能力分层被隐含在了按文理院系分层之中。

> 大学里的院系，院系之间是平行的，不是层级关系。同一个院系的不同年级之间的分层，实际上是垂直的序列关系，但也叫分层。

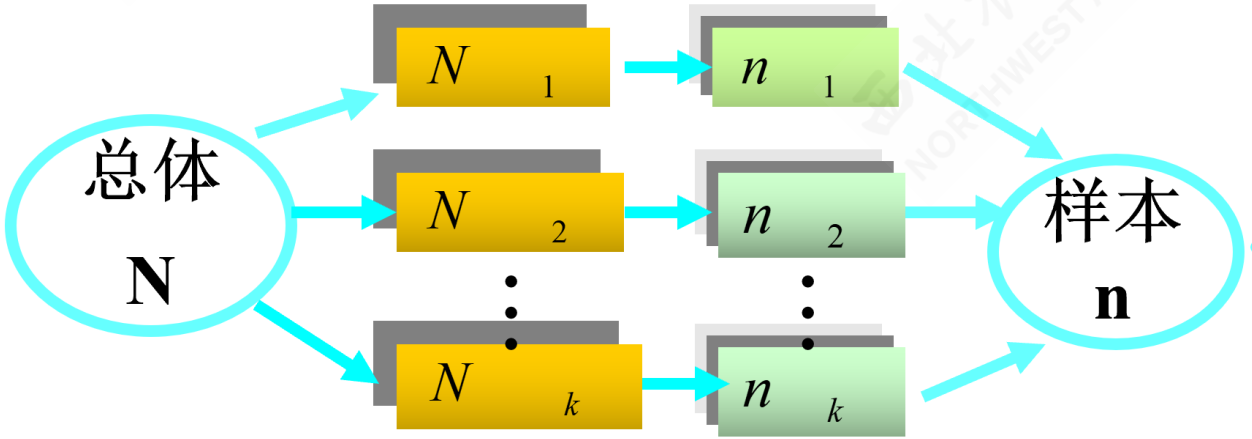
概率抽样5：分层抽样（优缺点）

- 保证样本的结构与总体的结构比较相近，从而提高估计的精度
- 组织实施调查方便
- 既可以对总体参数进行估计，也可以对各层的目标量进行估计

概率抽样5：分层抽样（样本分配）

在各层次中**样本量的分配**有两种基本的方法，

- 等比例分层抽样：各层的样本量与要素的规模成比例。
- 不等比例分层抽样：依据经验或者既有的研究结论减少或增加特定群体的样本量比例。



等额 $n_1 = n_2 = \cdots = n_k$，$n = \sum_{i=1}^{k} n_i$

等比例 $n_1 = \frac{N_i}{N} \cdot n$

概率抽样5：分层抽样（CFPS分层和样本分配）

CFPS的分层和分配示例：

- 第一个层：区分大省（5个）和小省（20个）。
- 第二个层1：大省（5个）
  - 子层1：4个省（辽宁、甘肃、河南、广东），各省为一个抽样框，但遵循相同的抽样策略。
  - 子层2：1个省（上海），为一个独立的抽样框
- 第二个层2：小省（20个）
  - 20个省级行政区，按照人均社会经济指标降序排列
  - 每一个省级行政区内，地级市按照人均GDP指标降序排列
  - 地级市内，分为区、县级市和县三个层。层内按人均GDP降序排列。

概率抽样5：分层抽样（CFPS分层和样本分配）

CFPS的分层和分配示例：

- 分配样本数。
  - 抽取样本县、区的初级抽样单位（PSU），分配各层样本数。
  - 实现既有发达的，也有不发达的，既有城市，也有县，人多的地区有样本，人少的地区也有样本。

概率抽样6：多阶段抽样（复习）

- 如果总体规模不大，要素在研究变量上的异质性分布具有**随机性**，则我们可以采用**简单随机抽样**、**系统抽样**。
- 如果不同群之间的异质性不大，群内的异质性对总体具有代表性，就可以采用**整群抽样**。
- 如果总体规模比较大，总体要素的**异质性也比较大**，且与不同特征群体的规模**无关**，研究变量在要素中呈现出某种非随机的分布，则需要采用**分层抽样**。
- 如果总体规模比较大，总体要素的**异质性也比较大**，且与不同特征群体的规模**有关**，那么至少要采用两个阶段的抽样，并且采用与群体规模**成比例的概率抽样**。
- 如果遇到搜集数据的范围非常大，要素的异质性分布也很复杂，那么采用上述任何一种方法都不足以解决抽样问题，而应该采用**多阶段抽样**。

概率抽样6：多阶段抽样（实施方案）

多阶段抽样(multi-stage sampling)：先抽取子群，但并不是调查群内的所有单位，而是再进行一步抽样，从选中的群中抽取出若干个单位进行调查。在多阶段抽样的每个阶段，采用的**抽样方法**也不一定相同。

实施方法：

- 先抽大单位（可用分层抽样）
- 再在大单位中抽小单位（可用成比例抽样）
- 小单位中再抽更小的单位（可用简单随机抽样）

概率抽样6：多阶段抽样（示例）

CGSS调查中基本要素是家庭中年满18岁或以上的个体。

假设研究者希望一次直接抽到个体，就需要编制一份有18岁或以上中国常住人口的抽样框。一个差不多有10亿人口的列表，这是不可能的

CGSS 2010年的抽样方案：

- 第一阶段，采用了分层抽样（覆盖全国区、县级市、县）。
- 第二阶段，抽到了村居，采用PPS抽样。
- 第三阶段，抽到了家户，采用了简单随机抽样。
- 末端抽样，抽到个体，采用了Kish表抽样。

概率抽样6：多阶段抽样（抽样单位）

- **初级抽样单位（PSU）**：初级阶段样本框的抽样单位。
  - CGSS的PSU就有160个区县
- **次级抽样单位（SSU）**：次级阶段样本框的抽样单位。
  - 对于上海，每个PSU只抽两个村居，也就是32乘2等于64，总的SSU的数量与其他大省一致。
- **末端抽样单位（USU）**：最末端阶段样本框的抽样单位。
  - CGSS的末端抽样单位是在样本户中抽到个人，是由调查员去抽取的

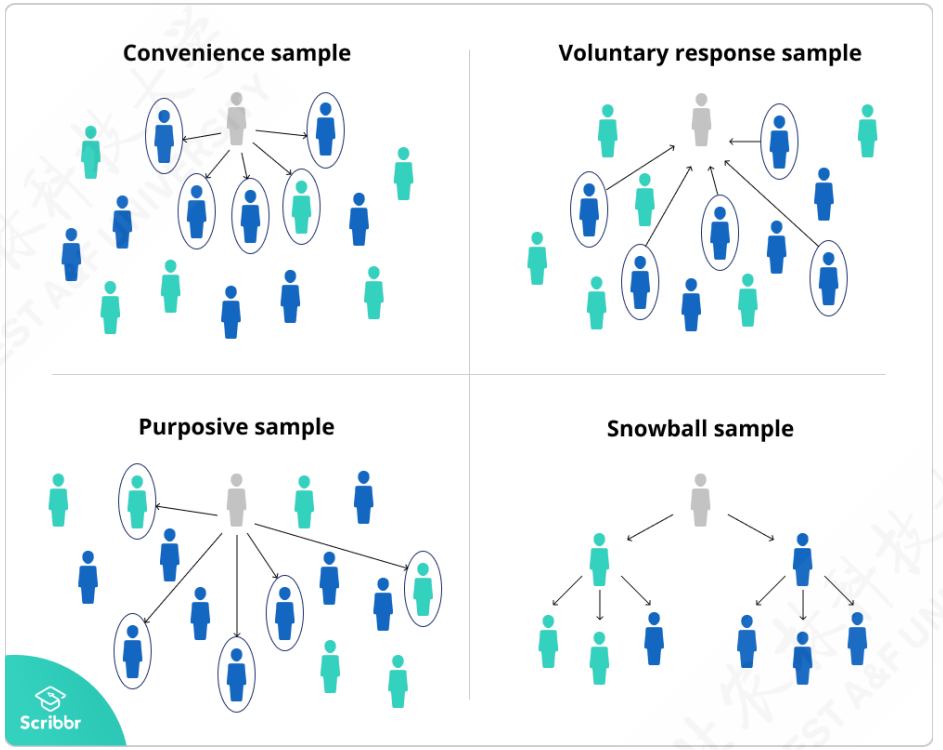
概率抽样6：多阶段抽样（优缺点）

- 具有整群抽样优点，保证样本相对集中，节约调查费用
- 需要包含所有低阶段抽样单位的抽样框；同时由于实行了再抽样，使调查单位在更广泛的范围内展开
- 在大规模的抽样调查中，是经常被采用的方法

非概率抽样

非概率抽样：抽取样本时不是依据随机原则，而是根据研究目的对数据的要求，采用某种方式从总体中抽出部分单位对其实施调查。

1. **方便抽样**（convenience sample）
2. **判断抽样**（purposive sample）
3. **自愿样本**（voluntary response sample）
4. **滚雪球抽样**（snowball sample）
5. **配额抽样**



非概率抽样1：方便抽样

方便抽样：调查过程中由调查员依据方便的原则，自行确定入抽样本的单位。

- 调查员在街头、公园、商店等公共场所进行拦截调查
- 厂家在出售产品柜台前对路过顾客进行的调查

优点：容易实施，调查的成本低

缺点：

- 样本单位的确定带有随意性
- 样本无法代表有明确定义的总体
- 调查结果不宜推断总体

非概率抽样2：判断抽样

判断抽样：研究人员根据经验、判断和对研究对象的了解，有目的选择一些单位作为样本。具体方式有：

- 重点抽样
- 典型抽样
- 代表抽样

缺点：

- 判断抽样是主观的，样本选择的好坏取决于调研者的判断、经验、专业程度和创造性
- 样本是人为确定的，没有依据随机的原则，调查结果不能用于推断总体

优点：抽样成本比较低，容易操作

非概率抽样3：判断抽样

自愿样本：被调查者自愿参加，成为样本中的一分子，向调查人员提供有关信息

- 参与报刊上和互联网上刊登的调查问卷活动
- 向某类节目拨打热线电话等

特点：

- 自愿样本与抽样的随机性无关
- 样本是有偏的
- 不能依据样本的信息推断总体

非概率抽样4：滚雪球抽样

滚雪球抽样：先选择一组调查单位，对其实施调查，再请他们提供另外一些属于研究总体的调查对象；调查人员根据所提供的线索，进行此后的调查。 持续这一过程，就会形成滚雪球效应。

特点：

- 适合于对稀少群体和特定群体研究
- 容易找到那些属于特定群体的被调查者
- 调查的成本也比较低

非概率抽样5：配额抽样

配额抽样：先将总体中的所有单位按一定的标志(变量)分为若干类，然后在每个类中采用方便抽样或判断抽样的方式选取样本单位。

特点：

- 操作简单，可以保证总体中不同类别的单位都能包括在所抽的样本之中，使得样本的结构和总体的结构类似
- 抽取具体样本单位时，不是依据随机原则，属于非概率抽样

抽样方案（抽样方法选择）

为了保证抽样过程的严谨，也需要一个文本，用来指导抽样活动，这就是**抽样方案**。

在抽样方案中，抽样方法的选择是核心内容。一般情况下，抽样方法的取舍取决于三个基本因素：要素的同质性、总体的规模、变量的多少。

- 第一，如果总体规模很大，异质性很强，研究变量很多，通常会采用多阶段、分层的PPS抽样。
- 第二，如果总体规模很大，异质性也很强，研究变量很少，通常采用多阶段抽样，末端通常采用整群抽样或者配额抽样。
- 第三，如果总体规模也很大，同质性也很强，这个时候，变量的多少没有太大关系一般的情况下会采用非概率抽样，比如说末端采用就近抽样、判别抽样。
- 第四，如果总体规模很小，异质性很强，变量多少都没关系，通常会采用滚雪球、RDS抽样或者是知情人抽样。

抽样方案（文本内容1）

抽样方案一般需要说明，采用什么方法，采用哪些步骤，获得用于收集数据的样本。一份抽样方案在内容上至少要有以下的内容，

- 第一，总体。不仅要把总体界定清楚，还要明确地界定研究总体、框总体，或者叫抽样总体。如果采用多阶段抽样、分层抽样的，还要说明每一个阶段、每一层的抽样框。
- 第二，研究对象。包括调查对象或者研究对象，就是收集数据的对象、受访者。比如说CGSS调查对象是家庭中的个人，CFPS调查对象是家庭中的所有成员。
- 第三，样本量。尤其是末端抽样单位的数量要做明确的说明。
- 第四，抽样方法。如果采用复杂设计，例如多阶段混合抽样，那么每一个阶段的抽样方法都要做说明。

抽样方案（文本内容2）

- 第五，如果采用多阶段混合抽样，或者多阶段抽样，每一个阶段的抽样单位、抽样框、抽样方法、样本量的配置以及末端抽样的方法也都需要写清楚。否则读者就无法知道每一个阶段的权重。
- 第六，如果不是采用大量熟悉的抽样框，自己在制备抽样框，还需要说明抽样框的制备方法。大型调查中的抽样框制备也是一项复杂的工程。
- 第七，还包括估计量的计算方法。比如说，权重到底怎么算，怎么配权重，如果是多阶段抽样，等概率又怎么保证。

抽样实施（工作安排）

即使有很好的抽样方案，如果不落到实处还是没有样本。抽样的实施一般来讲，根据抽样方案按照研究设计做就行了。听起来很简单，不过千万别大意。获得样本真的是一个非常艰难的过程。

- 第一，正确地理解方案，制定每一个环节的实施方案。抽样方案只是指引，指南，索引，在实践中在操作中还需要实施方案。
- 第二，组织资源。比如人力，社会关系，设备，后勤保障等等。稍稍大一点的调查就得请人，请学生，请朋友，怎么计酬，怎么支付这就是后勤问题。后勤对社会调查与研究也非常重要。
- 第三，培训抽样人员，督导人员，后勤人员。把实施中可能遇到的问题讲透彻，把合作与分工讲透彻，让每一个人明确的知道自己到底要干什么。
- 第四，逐步实施。一般来讲，前三步工作做完以后就一步一步地实施，先从制作抽样框开始，再抽样， 最后再做质量检验和误差估计。

抽样实施（经验建议）

在抽样的实施中，多问自己几个问题：

- 第一，总体到底有多大？到底多大范围的调查？
- 第二，研究总体在哪里？有哪些会影响到对调查对象的识别？
- 第三，有没有可用的抽样框？比如说，有没有可能让执行人提供一个抽样框？ 如果没有怎么制备抽样框？
- 第四，选择什么样的抽样方法可以减少误差？
- 第五，执行的难点到底是什么？怎么样去组织资源能够使得花最少的钱最有效地办事？
- 第六，最重要的一条经验就是多沟通。与相关各方尽可能就抽样设计，抽样实施的目标达成一致。

抽样误差（误差来源I）

抽样方法搜集数据，误差来源可能会出现在多个阶段：

- 第一，在**发起阶段**，由研究者带来的误差。理论假设不好，概念界定不清，对样本要求不明确。
- 第二，在**设计阶段**，由设计者带来的误差。比如测量工具选的不对，实施策略选的不对，抽样设计也有问题。
- 第三，在**抽样阶段**，由抽样员带来的误差。如果末端抽样框的界定不明确，抽样过程监管也不明确，就有可能产生随机性误差。
- 第四，在**访问阶段**，由访问员带来的误差。如果访员作弊、作假，轻易地接受拒访，诱导性提问，不规范的提问，也会造成随机误差、应答误差，甚至系统误差。

抽样误差（误差来源2）

- 第五，在**访问阶段**，由受访者带来的误差。
  - 如果受访者拒绝访问，或者没有能力作答，作假、作弊、随意作答、回忆误差，也会造成随机误差、应答误差。
- 第六，在**数据清理阶段**，由数据管理者带来的误差。
  - 如果数据的管理者编制的数据录入程序有问题，编码有问题，清理程序有问题，管理程序也有问题，也就有可能会前功尽弃，既可能产生随机误差，也可能产生系统误差。
- 第七，在**数据分析阶段**，由分析者带来的误差。
  - 如果分析者分析工具选择不当，模型建构不当，对数据有误读，也会造成研究误差。

抽样误差（误差类型）

涉及到调查活动的有三个阶段，也就是**设计阶段**、**抽样阶段**和**访问阶段**。这三个阶段涉及到的误差主要有：

- 第一，覆盖性误差，与抽样设计和抽样活动有关。
- 第二，抽样性误差，是抽样活动造成的误差。
- 第三，应答性误差，指访问阶段产生的误差。
- 第四，测量性误差，指测量、测量工具产生的误差。

抽样误差（4覆盖性误差）

覆盖性误差，主要指因抽样方制作不当带来的误差。它属于抽样设计和抽样活动有关一类误差。

- 如果抽样方与研究总体不一致，就会产生误差。
  - 假定CGSS使用电话号码作为抽样框，就会出现覆盖性误差。
  - 覆盖不足产生的误差：比如有些人没有电话，就会被抽样方忽略，太穷的、太富的都有可能没有电话，或者呢，有电话，却不在电话簿的列表中。
  - 覆盖过度产生的误差：比如很多人有多部电话，这些人就有可能被过度代表
- 任何抽样方法都不可避免地会带来误差。
  - 忽略样本特征而随意选择抽样方法，就会直接带来误差。
  - 即使让抽样方正确地反映了研究总体，抽样活动不可避免地也会带来误差。

抽样误差（3主要变量的抽样误差）

主要变量的抽样误差，由变量特征带来的。

- 每一个变量都有自己的抽样误差
- **主要变量的抽样误差**一般是指的**均值**的误差，用**均值的标准误** $\sigma_{\bar{x}}$ 来代表误差。
- **主要变量的抽样误差**也能用相对误差来表示，比如说**均值的变异系数** $V_{\bar{x}}$。

抽样误差（C应答性误差）

访问阶段的误差也会涉及到抽样误差，尤其是**应答性误差**。它属于访问阶段活动有关一类误差。

- **样本无应答**：又叫**单人无应答**，是指如果受访者对整个访问无论是问卷，还是访谈，都不回答的情形。简言之，就是无法从样本得到任何应答，比如说受访人拒访， 或者根本联系不上。
- **选项无应答**：又叫访题无应答，指受访者接受了访问，可能对某些访题不提供应答。

看起来这样的误差属于纯粹的访问误差，实际上不一定，也可以被认为是抽样误差的一种，比如，某些访题涉及到**稀有应答**，在抽样设计中，就需要予以考虑。

本节结束