抽样调查方法及应用

Part 1：概述

暨南大学 经济学院统计学系

陈光慧 教授

常用的抽样调查方法

非概率抽样调查

- 自愿样本
- 方便抽样
- 判断抽样
  (包括重点调查和典型调查)
- 配额抽样
- 滚雪球抽样
- ……

概率抽样调查（重点讲述）

- 简单随机抽样
- 分层抽样
- 系统抽样
- 不等概率抽样
- 整群抽样
- 多阶段抽样
- ……

（概率）抽样调查的基本步骤

- 第一、明确调查目的、目标总体、抽样单位等；
- 第二、确定抽样设计方法；
- 第三、确定必要的样本量；
- 第四、确定抽样估计方法；
- 第五、计算抽样误差。

实施抽样调查注意事项

- 一、两组容易混淆的概念
- 二、两个容易忽略的问题
- 三、两个容易陷入的误区

一、两组容易混淆的概念

- 目标总体与抽样总体
- 调查单位与抽样单元

1.目标总体与抽样总体

- **目标总体**
  - 由研究问题和调查目的确定的调查对象的全体，它是由研究对象中所有性质相同的个体所组成；
  - 组成总体的各个个体可称为总体单位或单元。
- **抽样总体：**从中实际抽取样本的总体，一般由所构建的抽样框决定。
- **难题：**由于抽样框不完整，导致抽样总体和目标总体有时不一致，引起抽样框误差。

2.调查单位与抽样单元

- 调查单位：就是调查项目的承担者，即我们想通过调查取得其观测值的单位，它通常是构成总体的最基本单位。但有时调查单位与基本单位并不相同。
- 抽样单元：就是用以抽取调查单位进入样本的中介单位，是抽样总体与调查单位之间的联接单位。在多阶抽样调查中，抽样单元需要分阶段确定，即初级抽样单元、次级抽样单元、……、最终抽样单元。

二、两个容易忽略的问题

- 抽样框的构建
- 辅助信息的利用

1.抽样框的构建

- **抽样框问题，是大家一直忽略的一个问题。实际上，任何概率抽样都需在抽样之前构建一个抽样框。**
- 抽样框是抽样总体的具体表现，通常是一份包含所有抽样单元的名单。
- 主要形式：**名录框**（目录、手册、电子数据库）、**地域框**（地图）、**时间框**等等。
- 抽样框的要求：

  （1）抽样框必须是有序的，即抽样单元必须编号，且根据某种顺序进行排列。

  （2）抽样框中包含的抽样单元务必要“不重不漏”，否则将出现抽样框误差。

例：企业名录抽样框

企业编号	企业名称	企业地址	联络人（姓名、	企业人数	备注
1	广州岭南佳国连锁	沿江东路405号	陈静/83856853	258	越秀区白云街
2	中华广场物业经营	中山三路33号	聂美萍/83739006	293	越秀区大东街
3	广州联新能源发展	东风中路268号广州交	区伟霞/83329640	372	越秀区东风街
4	广州市越秀区体育	应元路32号地下1层	廖玮菁/83560421	194	越秀区洪桥街
5	广东星煌文化传播	环市东路326号之一广	张惠建/62608117	149	越秀区建设街
6	广州唐会餐饮娱乐	起义路1号缤缤时装9	范燕妮/83280468	305	越秀区人民街
7	广州东亚大酒店有	长堤大马路320号	根洁/1380276267	92	越秀区人民街
8	广东新原地产代理	东风路410号健力宝	孙玮/83486398	139	越秀区广卫街
9	东浚物业管理有限	东风路836号4座21楼	琴/87605666（人事部	242	越秀区梅花街
10	广州钱柜餐饮娱乐	先烈中路汇华商贸大	小姐/37619222-7	250	越秀区黄花街

例：地域抽样框

广州主校区校园地
JINAN UNIVERSITY
参考比例尺 1:2000
中山大道
师大暨大站
师大暨大站 ①
师大暨大站 ②
岗顶站
人行天桥
往天河体育中心、广州火车东站
西门
临时停车场
暨南大学出版社
离退休工作处
北门
教育学院楼
附属幼儿园（吴炳昌楼）
附属中学（新址）（郭棣活楼）
暨大附中田径场
颜乐天纪念楼
旭日集团运动场（主场）
运动场（副场）
学术报告厅
学术会议厅
学生活动中心
校门诊部
印刷厂
艺术学院楼
学生宿舍1
学生宿舍2
学生宿舍6
附属中学教学楼
苏州苑41栋
篮球场
网球场
门球场
运动广场
体育路
真如路
花园路
苏州路
羊城路
苏州苑小花园
小花园
18栋之一
15 暨南大学建筑设计研究院
西侧门

失败调查案例解析
——不完备抽样框的危害，造成样本偏误

1936年，美国《文学文摘》杂志利用抽样调查的方法预测了美国1936年总统大选。调查方法是向1000万杂志订户邮寄了调查问卷，调查其选举意向，最终收回230万份，调查结果是兰登（London）与罗斯福（Roosevelt）两人得票比例为370：161。但是，选举结果却与调查结果正好相反，此项调查也因此成为抽样调查失败的经典案例。

与此同时，1935年由年轻调查员乔治·盖洛普创立的美国民意研究所却成功预测了罗斯福将大选获胜。盖洛普使用的方法是从全美随机抽取5万个选民进行调查，抽取的原则是保证全美所有的选民被抽中的概率相等。

此后，《文学文摘》杂志于次年破产，而盖洛普公司（www.gallup.com）却一举成名，直到现在仍然是世界上最著名的调查公司之一。

失败调查案例解析

——不完备抽样框的危害，造成样本偏误

2005年，北京某大学组织了一项调查，目的是想了解该校大学生学习生活情况。调查方法是在周末晚上去学生宿舍，通过随机抽取宿舍对宿舍内学生进行调查。

那么，这项抽样调查方法存在什么问题？

2.辅助信息的利用

- 如何充分利用辅助信息，也是大家一直忽略的问题。实际上，辅助信息可以渗透到抽样设计、估计等抽样调查的各个方面。
- 主要来源
  - ✓ 抽样框信息
  - ✓ 普查资料
  - ✓ 前期调查资料
  - ✓ 各级政府部门的行政记录数据
  - ✓ 企事业单位的生产经营数据

三、两个容易陷入的误区

- 随机与随意（或随便）的区别
- 概率抽样与等概率抽样的区别

1.随机与随意（或随便）的区别

- 两者的理论含义：随机有科学的含义，“随机”的结果可以用概率来描述，而随意则更多地带有人的主观性，“随意”的结果难以用概率来表示。
- 两者的根本区别：是否通过构建抽样框，确保总体中的每个单位都有事先可以计算和确定的、非零的概率被抽中。
- 举例：一个大笼子里养了很多兔子，
  - ✓ 随意抽样：饲养员进去（甚至闭上眼睛）随手抓几只兔子出来；
  - ✓ 随机抽样：将兔子编号构建抽样框，再产生随机数，确定被抽中编号的兔子。

2.概率抽样与等概率抽样的区别

- 两者的理论含义：概率抽样是指总体中的各单位都有非零的概率被抽中，各单位被抽中的概率可以相等，也可以不相等。如果各单位被抽中的概率相等，称为等概率抽样；如果各单位被抽中的概率不相等，则称为不等概率抽样。
- 两者的区别与联系：有些概率抽样是等概率抽样，有些则是不等概率抽样。这需要根据实际抽样调查的需要来确定，不可等同看待。

抽样调查方法及应用

Part 2：常见的抽样设计方法

暨南大学 经济学院统计学系

陈光慧 教授

常见的抽样设计方法

- 简单随机抽样
- 分层抽样
- 不等概率抽样
- 整群抽样
- 多阶抽样
- 系统抽样
- 二重抽样

一、简单随机抽样

(Simple Random Sampling)

- **理论定义：**从含有 N个单元的总体中抽取 n 个单元组成样本，如果抽样是不放回的，则所有可能的样本有C_N^n个，若每个样本被抽中的概率相同，都为$1/C_N^n$，这种抽样方法就是简单随机抽样。
- **操作步骤：**

✓ 第一步：编制涵盖N个总体单元的抽样框；

✓ 第二步：从1- N中产生n个随机数（可用抽签法、随机数表法、统计软件产生）；

✓ 第三步：n 个随机数对应的总体单元被抽出，便构成简单随机样本 。

一、简单随机抽样

(Simple Random Sampling)

- 优点：
  - 理论简单直观
  - 最基本的抽样设计方法
- 缺点：
  - N很大时难以构建抽样框（如广州市住户抽样，需构建全市所有住户的名单抽样框，几乎不可能）
  - 样本分散不易实施，调查费用高
- 应用：一般不单独使用；可结合其它方法使用
- 评价：精度中等（虽未充分利用辅助信息，但样本较为分散，对总体的代表性并不差），但有时难以构建抽样框，难以单独使用。

二、分层抽样(Stratified Sampling)

- **理论定义：**先将总体所有单元按某些重要标志进行分层(类)，然后在各层（类)中分别采用某种抽样方法抽取样本的一种抽样方式。
- **分层原则：**层内尽量相似、层间尽量差异。
- **操作步骤：**

✓ 第一步：选择分层标志，对总体单元进行分层；

✓ 第二步：分别编制涵盖各层单元的抽样框；

✓ 第三步：各层独立实施抽样（可使用简单随机抽样或其它抽样方式）。

二、分层抽样(Stratified Sampling)

- 应用：
  - 只要能构建各层抽样框便可使用
  - 可结合其它抽样方法使用
- 举例：
  - 某地企业抽样：通过该地企业名录库，可按照行业性质、规模大小等标志对企业进行分层，构建各层抽样框实施分层抽样；
  - 某高校学生抽样：通过该校学籍管理系统，可按照学科、年级对学生进行分层，构建各层抽样框实施分层抽样；若得不到学籍管理系统，则这种方法不可行；
  - 某地住户抽样：理论上可以按照住户的职业进行分层，但实际上却难以构建各个职业人群的抽样框，所以这种分层抽样实际上难以实施。

二、分层抽样(Stratified Sampling)

- 优点：
  - 样本代表性高，抽样估计精度高
- 缺点：
  - 需要构建各层的抽样框
  - 理论方法复杂（选择什么分层标志、分为几层、分层界限、各层样本量分配等问题）
- 评价：**精度非常高（层内越相似、层间越差异，精度便会越高），需要构建各层抽样框。**

三、不等概率抽样
(Sampling with Unequal Probabilities)

- **理论定义：**与抽样单元的规模大小成比例地确定被抽中进入样本的概率（即入样概率），然后针对不同的入样概率实施不等概率抽样抽出样本。
- **主要分类：**
  - ✓ PPS抽样（放回抽样，$z_i = x_i / \Sigma_{i=1}^{N} x_i$）
  - ✓ πPS抽样（不放回抽样，$\pi_i = n\, x_i / \Sigma_{i=1}^{N} x_i$）
- **操作步骤：**
  - ✓ 第一步：构建抽样框；
  - ✓ 第二步：获取与研究变量相关的辅助变量x_i信息；
  - ✓ 第三步：确定每个单元的抽取概率，实施不等概率抽样（可使用代码法）。

三、不等概率抽样
(Sampling with Unequal Probabilities)

- 应用：
  - 适用于抽样单元规模差异大的总体
  - 一般需结合其它抽样方法使用
- 举例：
  - 省抽县：按照与该省各县人口规模大小成比例，实施不等概率抽样抽取县；
  - 省抽高校：按照与该省各高校在校人数规模大小成比例，实施不等概率抽样抽取高校；
  - 某市抽零售企业：虽然理论上可以按照与该市各零售企业销售额大小成比例实施不等概率抽样，但考虑到企业数量大，构建抽样框和抽样过程复杂等因素，此时并不推荐单独使用不等概率抽样。

三、不等概率抽样
(Sampling with Unequal Probabilities)

- 优点：
  - 样本代表性高，抽样估计精度高
- 缺点：
  - 需要获得能够衡量规模大小的辅助变量信息
  - 抽样过程和理论方法较复杂
- 评价：精度非常高（辅助变量与研究变量相关性越高，精度便会越高），但需要辅助变量信息，实施过程复杂。

四、整群抽样(Cluster Sampling)

- **理论定义：**先将总体划分成很多不重合的子总体或群，然后以群为抽样单元，按某种随机方式从中抽取若干个群，形成一个“群”的随机样本，对抽中的群内所有单元全部进行调查。
- **分群原则：**群内尽量差异、群间尽量相似。
- **操作步骤：**

✓ 第一步：选择分群标志，将总体单元划分不同的群；

✓ 第二步：编制涵盖各群的抽样框；

✓ 第三步：实施某种抽样，抽出若干个群作为样本；

✓ 第四步：对群内所有单元进行调查。

四、整群抽样(Cluster Sampling)

- 应用：
  - 群规模相等时整群抽样（可结合简单随机抽样）
  - 群规模不等时整群抽样（可结合不等概率抽样）
- 举例：
  - 群规模相等：将包装好的整箱产品作为群，实施整群抽样，检验产品质量；
  - 群规模不等：将某县所有的行政村作为群，实施整群抽样，开展农户调查（村规模大小有差异）；
  - 群规模不等：将某高校所有的班级作为群，实施整群抽样，开展学生调查（班级规模大小有差异）。

四、整群抽样(Cluster Sampling)

- 优点：
  - 不需要构建涵盖所有总体单元的抽样框，仅需群的抽样框
  - 一般按照现成的分类（如行政区划）划分群，构建群的抽样框较容易
  - 由于群内单元全部接受调查，样本点集中，节约调查费用
- 缺点：一般群内单元难以遵守群划分的原则，相似的单元集中在群内，导致精度较低。
- 评价：**精度比较低，但非常节约调查费用。**

五、多阶抽样(Multi-stage Sampling)

- **理论定义：**将抽样过程分成几个不同阶段分别抽取各阶段的抽样单元，开展抽样调查。
- **操作步骤：**

✓ 第一步：将总体按某种原则划分成不同的初级抽样单元，实施随机抽样抽出若干个初级抽样单元作为第一阶样本；

✓ 第二步：针对被抽中的初级抽样单元，再抽取若干个次级抽样单元作为第二阶样本；

✓ 第三步：以此类推，直到抽出最终抽样单元构成最后一阶样本，再实施调查，即形成多阶段的抽样。

五、多阶抽样(Multi-stage Sampling)

- 应用：
  - 适用于大规模的抽样调查，通常都要采用多阶抽样的形式；
  - 各阶段抽样可灵活采用其它各种抽样方法。
- 举例：
  - 以省为总体的多阶抽样：省抽县、县抽乡、乡抽村、村抽住户；
  - 以某高校为总体的多阶抽样：高校抽学院、学院抽班级、班级抽学生。

五、多阶抽样(Multi-stage Sampling)

- 优点：
  - 不需要构建全部总体单元的抽样框，仅需分别构建各阶抽样单元的抽样框；
  - 一般按照现成的行政区划作为各阶抽样单元，构建抽样框较容易；
  - 最终调查样本较为集中，实施调查比较方便。
- 缺点：抽样估计较复杂。
- 评价：**精度较低（比简单随机抽样低，但一般比整群抽样高），非常节约调查费用。**

六、系统抽样(Systematic Sampling)

- **理论定义：**系统抽样是将N个总体单元按一定顺序排列，先随机抽取一个单元作为第一个样本单元，然后再按某种确定的规则抽取后续的样本单元。比如按照等距的方式间隔若干个单元便抽取一个作为样本单元，即常见的等距抽样。
- **应用举例：**
  - 按无关标志排队：按学号或姓氏笔画顺序抽学生、按学生宿舍编号顺序抽宿舍；
  - 按有关标志排队：某县所有的行政村按照平均收入高低排列，实施系统抽样（可考虑等距、分组对称、总体对称等形式）。

六、系统抽样(Systematic sampling)

- 优点：
  - 不需要构建全部总体单元的抽样框，仅需按照某种顺序排列即可；
- 缺点：抽样精度有时与排列顺序有关。
- 评价：**精度有时很高、有时又很低（关键看排列顺序），抽取样本方便，节约调查费用。**

七、二重抽样(Two-phase Sampling)

- **理论定义：**先从总体中抽取一个大的初始样本，从而获得总体的辅助信息，然后从初始样本中再抽一个子样本。
- **操作步骤：**

✓ 第一步：先从总体N中抽取一个样本量较大的样本，称为第一重样本

✓ 第二步：对其进行简单调查以获取总体的某些辅助信息，为下一步的抽样估计提供条件；

✓ 第三步：进行第二重抽样，第二重抽样的样本量相对较小；

✓ 第四步：针对第二重样本开展详细调查；

✓ 第五步：通过上述调查信息对总体进行推断估计。

七、二重抽样(Two-phase Sampling)

- 应用：
  - 适用于难以构建抽样框的情形。
- 举例：
  - 对某市患有高血压的人群开展抽样调查：难以构建抽样框，可实施二重抽样；
  - 对某高校内使用苹果手机的学生开展抽样调查：难以构建抽样框，可实施二重抽样；

七、二重抽样(Two-phase Sampling)

- 优点：
  - 解决了难以直接构建抽样框的难题；
  - 在第一重抽样调查时可搜集相关的辅助信息。
- 缺点：抽样估计较为复杂。
- 评价：**虽然精度不高，但在难以直接构建抽样框时可考虑使用。**

抽样调查方法及应用

Part 3：实际抽样设计应用案例

暨南大学 经济学院统计学系

陈光慧 教授

实际抽样设计的几点经验

- 一般应实施多阶抽样设计，各阶段抽样要结合其它抽样设计方法；
- 根据调查费用及可行性，合理构建各阶段抽样框和抽样设计方法；
- 各阶段要充分利用相关的辅助变量信息（普查、行政记录、各类大数据等），提高抽样估计精度。

应用案例1：中国劳动力抽样调查

此调查以全国为总体，各省为子总体独立实施抽样。可采用的抽样设计有：分层PPS两阶整群抽样、分层PPS三阶整群抽样、分层PPS二阶整群抽样。

- 1. 分层PPS四阶整群抽样

（1）对县级单元（县、县级市或市辖区）按照人口和劳动力等特征分层，在第一阶段采用PPS抽样在每一层抽取县级单位；

（2）对抽中的县级单位，采用PPS抽样抽取乡级单位（街道、镇或乡）；

（3）对抽中的乡级单位，采用PPS抽样抽取村级单位（居委会或村委会）；

（4）对抽中的村级单位，采用简单随机抽样或系统抽样抽取调查小区（由30个最相邻的住址组成）。抽中的小区调查其全部住户及家庭成员。

- 2. 分层PPS三阶整群抽样

  （1）对县级单元（县、县级市或市辖区）按照人口和劳动力等特征分层，在第一阶段采用PPS抽样在每一层抽取县级单位；

  （2）对抽中的县级单位，采用PPS抽样抽取村级单位；

  （3）对抽中的村级单位，采用简单随机抽样或系统抽样抽取调查小区。抽中的小区调查其全部住户及家庭成员。

- 3. 分层PPS二阶整群抽样

  （1）对村级单元按照城乡及人口和劳动力等特征分层，在第一阶段采用PPS抽样在每一层抽取村级单位；

  （2）对抽中的村级单位，采用简单随机抽样或系统抽样抽取调查小区。抽中的小区调查其全部住户及家庭成员。

应用案例2：某高校学生抽样调查

此调查以该校所有在籍学生为总体实施抽样。可采用的抽样设计有：分层PPS三阶抽样、分层PPS二阶抽样、分层PPS二阶整群抽样。

- 1. 分层PPS三阶抽样（按名录构建抽样框）

（1）对各学院按照学科差异分层，在第一阶段采用与学院学生数成比例的PPS抽样在每一层抽取若干个学院；

（2）对抽中的学院，采用与班级人数成比例的PPS抽样抽取若干个班级；

（3）对抽中的班级，采用简单随机抽样或系统抽样抽取一定数量的学生开展调查。

- 2. 分层PPS二阶抽样（按名录构建抽样框）

  （1）对校内所有班级按照所属学科差异进行分层，在第一阶段采用与班级人数成比例的PPS抽样在每一层抽取若干个班级；

  （2）对抽中的班级，采用简单随机抽样或系统抽样抽取一定数量的学生开展调查。

- 3. 分层PPS二阶整群抽样（按地域构建抽样框）

  （1）将该校所有的学生宿舍楼按照男女分层，在第一阶段采用与宿舍楼人数规模成比例的PPS抽样在每一层抽取宿舍楼；

  （2）对抽中的宿舍楼，采用简单随机抽样或系统抽样抽取宿舍，对于抽中宿舍的所有学生全部开展调查。

抽样调查方法及应用

Part 4：样本量的确定

暨南大学 经济学院统计学系

陈光慧 教授

关于样本量的几点认识

- 样本量不可过小，过小可能导致样本代表性不足，调查结果不可靠；
- 样本量也并不是越大越好，样本量过大将导致调查费用及工作量增加，引起非抽样误差增大；
- 合理的样本量应该根据调查费用、精度要求和抽样设计方法科学计算确定。

一、样本量过小的危害

- 美国霍普金斯大学原来是一所只招男生的学校，后来校方尝试也同时招收女学生。但是，有某些反对人士反对男女同校。为了证明男女同校的弊端，他们通过调查某一年该校招收的女生的毕业去向，结果得出结果是该校那一年招收的女学生中，毕业后有1/3嫁给了该校的老师。公布此结果后，社会一片哗然。最后，通过校方的仔细研究，才找出了问题的症结所在。
- **原因是：那年该校仅招收了3名女学生，其中有1名女学生嫁给了该校老师。**

二、样本量也不需要很大

教材案例：在简单随机抽样下，估计比例**P**，置信度**95%**，允许误差**5%**，在**P=0.5**条件下，

总体规模（N）	所需样本量（n）
50	44
100	80
500	222
1000	286
5000	370
10000	385
100000	398
1000000	400
10000000	400

经验结论：当总体规模成倍增大，样本量并不需要显著增加！

注：其它抽样方式下的样本量还需在此基础上进行调整。

主要国家人口调查样本量对比

国别	人口调查	总人口	样本量
加拿大	LFS：The Labor Force Survey	0.3亿	5万
美国	CPS：Current Population Survey	3亿	?
中国	人口住户类调查	14亿	?

主要国家人口调查样本量对比

国别	人口调查	总人口	样本量
加拿大	LFS：The Labor Force Survey	0.3亿	5万
美国	CPS：Current Population Survey	3亿	5万
中国	人口住户类调查	14亿	？

三、样本量的科学计算方法

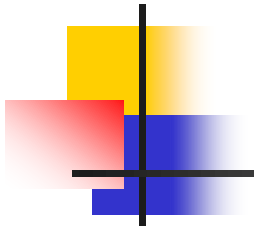
- **第1步：**假定以总体比例的抽样估计为例，确定委托机构所要求或认可的估计精度水平，包括绝对误差限d（或相对误差限r）、置信度1-α，且满足：$Pr(|p-P| \geq d)=\alpha$。
- **第2步：**根据上述确定的精度水平，对比例P进行预估计（若不确定，可以最保守的水平确定为0.5），在考虑某地常住人口规模N的情况下，以简单随机抽样时的初始样本量为

$$n_0 = \frac{t^2 P(1-P)/d^2}{1+\frac{1}{N}\left(\frac{t^2 P(1-p)}{d^2}-1\right)}$$

允许误差范围	在 90%置信度下	在 95%置信度下	在 99%置信度下
±0.01	6773	9614	16589
±0.02	1693	2403	4147
±0.03	753	1068	1843
±0.04	423	601	1037
±0.05	271	385	664
±0.06	188	267	461
±0.07	138	196	339
±0.08	106	150	259
±0.09	84	119	205
±0.1	68	96	166

三、样本量的科学计算方法

- **第3步：**根据本抽样方案实际使用的三阶PPS等距抽样设计，计算其抽样设计效应（$deff$），并对初始样本量进行调整：$n_1 = n_0 \times deff$。根据理论测算，多阶抽样的设计效应一般为1-3。
- **第4步：**通过过去各期调查的经验，估计有效回答率R，并对样本量进行再调整为：$n_2 = n_1/R$。
- **第5步：**在上述理论方法确定样本量的基础上，还需考虑并权衡调查经费、时间、调查机构可动用的各种资源等各方面的限制，综合确定最终的样本量。



谢谢！ 欢迎多联系交流！