普通高等学校“十一五”国家级规划教材
面 向 21 世 纪 课 程 教 材

市场营销调研,高等教育出版社，主编：景奉杰

第七章 样本设计

学习要求

- 掌握抽样的基本概念和步骤
- 熟悉非概率样本设计
- 熟悉概率样本设计
- 了解样本量的的确定原则

主要内容

第一节 样本设计概述

第二节 抽样技术

第三节 样本容量

第一节 样本设计概述

概念

- 总体 由市场研究目的目标明确规定的整个集合。
- 样本是总体的一个子集，应具有样本代表性
- 样本单位是样本的基本单位
- 普查 对整个总体的报告
- 抽样误差是因使用样本而发生的误差。来源：
  - 样本的选择方法
  - 样本容量

第一节 样本设计概述

- 抽样框是总体所有样本单位的完整列表
- 抽样框误差是抽样框不能解释总体的程度。
- 抽样框误差来源：
  - 总体的一部分成员不在抽样框；
  - 抽样框内一部分成员不属于目标总体。
- 抽样的原因
  - 成本、可行性

二、抽样的原因

- 对于即使是中等容量的总体的普查，其成本也非常昂贵，并且耗时很长
- 在某些情况下，普查不可行

三、抽样的程序

- 定义总体
- 识别抽样框
- 设计样本计划（方法、容量）
- 抽取样本收集数据
- 样本有效性检验
- 必要时再抽样

第二节 抽样技术

- 概率抽样（总体每个成员入选样本的概率已知、非零）

  又称随机抽样.概率抽样以概率理论为依据,通过随机化的机械操作程序取得样本,所以能避免抽样过程中的人为因素的影响,保证样本的客观性.虽然随机样本一般不会与总体完全一致,但它所依据的是大数定律,而且能计算和控制抽样误差,因此可以正确地说明样本的统计值在多大程度上适合于总体,根据样本调查的结果可以从数量上推断总体,也可在一定程度上说明总体的性质,特征.

- 非概率抽样（无法估计每个成员入选样本的概率）

  又称为不等概率抽样或非随机抽样,调查者根据自己的方便或主观判断抽取样本的方法.它不是严格按随机抽样原则来抽取样本,所以失去了大数定律的存在基础,也就无法确定抽样误差,无法正确地说明样本的统计值在多大程度上适合于总体.虽然根据样本调查的结果也可在一定程度上说明总体的性质,特征,但不能从数量上推断总体

一、概率抽样

- 简单随机抽样
- 系统抽样
- 整群抽样
- 分层抽样

随机抽样—简单随机抽样

- 若总体中每个个体被抽到的机会是均等的（即抽样的随机性），且在抽样取走一个个体之后总体内成分不变（即抽样的独立性），这种抽样方式称为简单随机抽样。。
- 抽签法。把总体中的每一个个体都编上号码，并做成签，充分混合后从中随机抽取一部分，这部分所对应的个体就组成一个样本。
- 查表法。查随机数表，确定从总体中所抽取个体的号码，则号码所对应的个体就进入样本。随机数表可随意从任何一区、任何一个数目开始，依次向各个方向顺序进行。
- 计算机造数法。用电子计算机编造随机数程序，把随机数作为总体中抽出个体进入样本的号码。
- 上述三种抽样方法是基本抽样方法，它虽然最符合随机原则，随时可用，但它十分费事，效率不高。仅适用于总体单位数较少，范围也很有限的情形，要进行大规模的抽样、编号、抽签或查随机数表都是很困难的。

系统抽样（等距抽样）

- 等间隔法的机械抽样。它把总体中所有个体按一定顺序编号，然后依固定间隔取样，间隔的大小视所需样本容量与总体中个体数目的比率而定，起始数字必须是随机决定的。这种方法与简单随机抽样相比，方便、易学、易做，当总体按一定顺序排定后，第一个样本一经确定，其他样本也随之确定。但是，这种抽样方法在名单排列中，如果存在周期性部分，则会造成偏差。因此，在等距抽样间距确定以后，选择起点时，应根据掌握的信息，尽量避开总体可能存在周期的点。
- 系统抽样具体方法如下：

（1）设总体共有N个单位，现需要从中抽出n个单位作为样本。先将总体的N个单位按与总体特征标志无关的标志进行排队。

（2）确定取样间隔，将N划分为n个单位相等的部分，每部分间隔为k

（3）决定起点，抽样起点的选定有多种方式，通常是在第一部分顺序为1，2，3，…，i…，K个单位中随机取一个单位i作为抽样的起点。对于总体单位N是奇数时，也可按R=（K+1）/2算出R值，就按某一部分的第R个单位作为抽样起点。对于总体单位N是偶数时，则按R=（K+2）/2算出起点位置。

（4）在第一部分中，随机以i为起点抽出第一个样本后，继续在第二部分中抽出第i+K单位为样本；如此类推，在第n部分则抽取第i+(n-1)K单位为样本。
这样一共抽出了n个单位组成样本，而且每个样本的间隔都是K，所以称这种抽样方法为等距抽样。

系统抽样例子

- 现有180名学生，要利用系统抽样法从中抽取15名学生作研究样本，其方法如下。
先将学生按与学生学习成绩无关的标志编号，假设按学生座位顺序把学生编为1－180号，然后按下述步骤抽取：
（1）确定间隔距离
（2）决定起点R=（K+2）/2=（12+2）/2=7，即决定从第一部分的第7号单位作为第一个样本。第二个样本为7+12=19号单位；如此类推，抽出的15个样本为：（7），（19），（31），（43），（55），（67），（79），（91），（103），（115），（127），（139），（151），（163），（175）。

整群抽样

- 整群抽样是先将各单位划分为若干群（子集合），每个子集合都可以代表整个总体。然后以群为单位从中随机抽取一些群，对抽中的群的所有单位进行调查。

  例如，某地要了解各校学生的学习情况，可在该校随机抽取几个班级，对抽中的班级的全部学生进行调查。但是，整群抽样在总体中不是抽取几个个体，而是随机抽取整群为单位进入样本。此种抽样在小范围内无实际意义，其抽样误差大，对总体的代表性差。由于总体中各个个体间存在差异，因此所得到的样本与总体间也有一定的差异，这个差异即抽样误差。

分层抽样（stratified sampling）

- 将总体的单位按某种特征分为若干次级总体（层），然后再从每一层内进行单纯随机抽样，组成一个样本。分层可以提高总体指标估计值的精确度，它可以将一个内部变异很大的总体分成一些内部变异较小的层（次总体）。每一层内个体变异越小越好，层间变异则越大越好。
- 分层抽样比单纯随机抽样所得到的结果准确性更高，组织管理更方便，而且它能保证总体中每一层都有个体被抽到。这样除了能估计总体的参数值，还可以分别估计各个层内的情况，因此分层抽样技术常被采用
- 例如，一个单位的职工有500人，其中不到35岁有125人，35岁至49岁的有280人，50岁以上的有95人.为了了解这个单位职工与身体状况有关的某项指标，要从中抽取一个容量为100的样本，由于职工年龄与这项指标有关，决定采用分层抽样方法进行抽取.因为样本容量与总体的个数的比为1：5，所以在各年龄段抽取的个数依次为125/5，280/5，95/5，即25，56，19。

二、非概率抽样—方便抽样（偶然抽样）

- 组成样本的元素以“偶然”的方式进入。
- 研究者按自己的意愿或可能，去抽取最接近、最有可能进行研究的对象为样本的抽样方法，它是一种非概率抽样的方法。

  这种抽样就具有随意性。它的缺点是，由于总体中每一个对象被抽取的概率是未知的，研究者一般不能说样本对于较大的总体具有何种程度的代表性，限制了把研究成果推广到样本范围之外的可能。而且，无法计算抽样误差。

二、非概率抽样—判断抽样

- 精心挑选样本元素希望它们能服务于研究目的。
- 通常运用于市场研究的早期阶段。例如，研究者可能选择一些观点差异悬殊的人，来检验问题的设计是否恰当。虽然这种调查结果不能代表任何有意义的总体，但会有效地发现问卷设计中的缺失。

二、非概率抽样—参考抽样、定额抽样

- 参考抽样（滚雪球样本）
- 配额抽样
  - 配额抽样使样本中拥有某种特征的元素比例与该类元素在总体中的比例一致，以此来代表总体。如此收集数据，从理论上讲应当能够代表总体。此种方法存在的问题是：配额的比例必须精确，但由于最新的关于总体性质变化的信息并不容易得到，而造成抽样中的偏差。

非概率抽样法评价

尽管非概率抽样存在许多不足，但非概率抽样在市场研究调查中的应用却并不少见。非概率抽样的最大优势在于它的简单易行与价廉。 当然，在实际研究中，既要尽可能按随机化原则抽样，又不能机械地套用上述抽样方法，而要求对课题的特点和客观条件作具体分析，将几种抽样方法结合起来使用。

第三节 样本容量

- 样本容量与样本代表性和精确度之间的关系
- 简单随机抽样下的样本容量
- 分层抽样下的样本容量

一、样本容量与样本代表性和精确度之间的关系

- 样本容量的决定与其说与总体容量有关，不如说与客户的预算、研究目标、数据用途和报告的时间期限更直接相关
- 样本容量与样本对总体的代表性无关
- 样本容量不决定代表性，然而影响结果的精确度

二、简单随机抽样下的样本容量

简单随机抽样时，总体均值的置信度为1-α的置信区间　自由度df=n-1

$$m = \bar{X} \pm t_{a/2} SE = \bar{X} \pm t_{a/2} \frac{S}{\sqrt{n}}$$

样本量足够大时，t分布可用相应的Z分布代替

最大允许绝对误差　$\Delta_{\bar{X}} = t_{a/2} SE$

最大允许相对误差　$r_{\bar{X}} = \frac{\Delta_{\bar{X}}}{\bar{X}}$

二、简单随机抽样下的样本容量

- 决定样本容量的程序
  - 规定可接受的容许水平h
  - 决定可靠性系数$Z_{\alpha/2}$
  - 取得在目标总体中测量的特征的标准差$\hat{s}_y$的估计值
  - 使用公式求得样本容量n*

$$n^* = \frac{(Z_{a/2})^2(\hat{s}_y)^2}{h^2}$$

二、简单随机抽样下的样本容量

- 在预期平均值固定百分比范围内决定样本容量的程序
  - 规定可接受的相对容许固定百分比r
  - 决定可靠性系数$Z_{\alpha/2}$
  - 取得在目标总体中测量的特征的变异系数 σ / μ 的估计值 $\hat{C}_Y = \hat{s}_Y / \overline{Y}$
  - 使用公式求得样本容量n*

$$n^* = \frac{(Z_{a/2})^2 (\hat{C}_Y)^2}{r^2}$$

二、简单随机抽样下的样本容量

- 用百分比决定样本容量的公式

$$n^* = \frac{(Z_{a/2})^2 p(1-p)}{r^2}$$

  - 对于以上求出的初始样本容量，要根据总体大小、设计效果和回答率分别进行调整
  - 当目标样本占目标总体的10%至20%或更多时（一般情况下为小总体），要使用修正系数fpc

$$fpc = \frac{N - n}{N - 1}$$

三、分层抽样下的样本容量

- 本部分只考虑对总体平均值的估计
- 设总体容量为N，分为H个互斥和穷举的层，每层的样本容量为n_h

$$N = \sum_{h=1}^{H} N_h \qquad n = \sum_{h=1}^{H} n_h$$

三、分层抽样下的样本容量

- 分层随机抽样的均值与误差估计

目标总体的平均值估计 $\bar{y}_{(ST)} = \sum_{h=1}^{H} W_h \bar{y}_h$

标准差 $S_{y(ST)} = \sqrt{\sum_{h=1}^{H} W_h^2 S_{yh}^2}$

标准误 $S_{\bar{y}(ST)} = \sqrt{\sum_{h=1}^{H} W_h^2 \frac{S_{yh}^2}{n_h}}$

小总体时的标准误 $S_{yh}^2 = \sqrt{\sum_{h=1}^{H} W_h^2 \left(\frac{S_{yn}^2}{n_h}\right)\left(\frac{N_h - n_h}{N_h - 1}\right)}$

- W_h：第h层的权，等于N_h/N
- $\bar{y}_h$：第h层的平均值
- S_{yh}^2：第h层的方差

第 h 层的方差计算 $S_{yh}^2 = \frac{\sum_{i=1}^{n_h}(y_{ih} - \bar{y}_h)^2}{n_h - 1}$

三、分层抽样下的样本容量

- 分层抽样中层间的差异性没有进入整个分层样本的标准误的计算，由于标准误代表抽样误差，所以分层误差比简单随机抽样的抽样误差更小，精确度更高

分层随机抽样的样本容量公式 $n^* \approx \dfrac{N\left(Z_{\frac{\alpha}{2}}\right)^2\left(\dfrac{\hat{\sigma}^2_{\omega y}}{\bar{Y}}\right)}{\left(Z_{\frac{\alpha}{2}}\right)^2\left(\dfrac{\hat{\sigma}^2_{\omega y}}{\bar{Y}}\right)+N_P^2}$

单个层内方差估计的加权平均数 $\hat{\sigma}^2_{\omega y}=\sum_{h=1}^{H}\dfrac{N_h}{N}\hat{\sigma}^2_{yh}$

三、分层抽样下的样本容量

- 分层抽样中的每层样本量的确定

比例分配时的每层样本容量确定 $n_h = \left(\frac{N_h}{N}\right) n$

最优分配时的每层样本容量确定 $n_h = (N_h \hat{\sigma}_{yh} / \sum_{h=1}^{H} N_h \hat{\sigma}_{yh}) n$

三、分层抽样下的样本容量

- 最优分配中变量分布标准差未知时的处理

- （1）基于以前调查了相似的抽样变量和使用了相似的分类变量的研究调查，通过平均或某种其他方法，取得每一层内抽样变量的分布的标准差的估计 $\hat{\sigma}_{yh}$，使用这些估计值计算从每层中抽取的元素的最优分配

- （2）从每层中抽取一个小的附属样本，在附属样本的基础上计算抽样变量的抽样分布的标准差 s^*_{yh}。使用 s^*_{yh} 计算从每一层中抽取的元素的最优分配

思考题

1. 解释下列概念：

总体 样本 抽样 普查 抽样框 抽样误差 概率抽样 非概率抽样 样本容量 简单随机抽样 系统抽样 整群抽样 分层抽样 方便抽样 判断抽样 参考抽样 限额抽样

2. 在市场研究中为什么要进行抽样?
3. 试述开发样本计划的过程。
4. 试比较4种概率抽样的优缺点。
5. 如何科学地确定样本容量?

案例：计算机品牌

某市场调查公司受一家计算机制造商委托对以下问题做出决定：哪些具体因素促使人们选择某个特定的品牌?家庭(个人)用户和企业用户在选择品牌时，有哪些不同的决定因素?为什么人们选择某一个品牌而不选择其他?去年有多少用户改变了他们的品牌?为什么他们要改变?人们对其品牌现在的满意度如何?使用者还想从其品牌得到哪些服务?

讨论

- 1．对这项研究你如何定义总体?
- 2．你将在这项研究中使用什么样的抽样框?
- 3．在选择的抽样框中，你将采取什么步骤进行简单随机抽样?
- 4．你将采取哪种概率抽样法，为什么?