法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

- 课程详情请咨询
  - 微信公众号：小象
  - 新浪微博：ChinaHadoop

数理统计

目录

目 录

统计调查的概念

统计调查是根据统计任务的要求，运用科学的调查方法，有计划、有组织地收集统计资料的过程。**统计调查的目的**是取得尽可能准确的综合数字和资料，用以说明调查对象的总体性质和规律。

统计调查是对代表个人、机构或实质物体单位所组成的存在总体的的科学研究。统计调查是试图通过对自然存在的总体进行观察以获得对它的了解，并对其综合的总体特征作出数量描述。

统计调查

统计调查可分为全面调查与非全面调查。

全面调查是对总体的全部单元，按某一时期逐个进行观察、记录，说明总体在某一方面或某些方面的性质和规律。

非全面调查（也称为意识抽样）是对调查对象中的一部分单元进行调查，又分为重点调查、典型调查和抽样调查。

重点调查

重点调查：抽取总体中在**数量或比重上有重要地位**的少数单位进行调查。例如，对一些大型油田的调查，掌握国家石油生产的情况；对几家大型商场的调查，掌握市场物价，购买动向等情况；对几重点大学的的调查，了解高等教育和大学生的状况等。

典型调查

典型调查(国外称为判定抽样或目的抽样)是调查者在对总体已有一定了解的基础上，根据自己的调查目的，有计划地从总体中抽取个别的典型单位，通过解剖麻雀、系统分析来观察趋势，指导一般。例如，毛泽东的《兴国调查》和《湖南农民运动考察报告》；费孝通的《江村调查》。

抽样调查

抽样调查是按照随机原则，即保证总体中每个单元都有一定概率被抽取的原则，抽取部分总体单元进行调查。所抽部分称为随机样本，简称样本。所以，随机抽样又称为概率抽样。精心设计的概率抽样可以提供一个能够反映总体的样本。

抽样调查的特点

- 按照**随机原则**抽取样本；
- 在数量上以样本推断总体；
- 可以在一定的概率保证下，计算抽样误差，并把它控制在允许的范围之内。

抽样调查的优越性

- **费用较低**：如果数据是从总体的一很小的部分取得的，那么它的费用就比普查小。当总体较大时，则从只占总体一小部分的样本中就能得到精确度足够有用的结果。
- **速度较快**：搜集和综合样本资料要比综合全面调查的资料快。在迫切需要有关的信息时，考虑这点极为重要。
- **应用范围广**：在某些种类的调查中，必需使用受过高度训练的人员或专用设备来获得有着的数据，而这种人员的设备在数量上又有很要进行普查实际上是办不到的。抽样调查目前广泛地应用于经济学，社会学，人类学心理学及政治学等学科。在公共卫生，生物统计，教育，社会工作，行政管理和工商管理等领域也广泛地应用抽样调查方法。

抽样调查的优越性

- **准确度较高**：在工作量减少后，由于能雇用质量较高的工作人员，并对他们进行深入的培训，还由于实地调查工作可以受到更仔细的检查监督，调查资料的处理能够完成，因此与可能进行的全面调查相比，抽样调查可能取得更准确的结果。
- 另外，抽样调查可对所抽出的样本做更详细的调查。在一次调查中，抽样调查的调查项目可设置较多。抽样调查特别适合商业调查、选举调查、社会、经济问题的研究。

统计调查的实施

1.2 统计调查的实施

1.2.1 统计调查计划

统计调查计划为具体施行调查作好准备。此阶段需要确定如何搜集数据，需要什么样的数据。

统计调查计划主要包括以下内容：

a. 确定调查任务与目的；

b. 确定调查对象、调查单位与报告单位；

c. 确定调查项目与设计调查表问卷与表格；

统计调查的实施

d. 确定调查时间、调查工作完成的期限及工作进度；

e. 确定调查方式与汇总的方法；

f. 必要的质量保证措施（包括：组织领导；培训调查人员；答卷的复核）。

1.2.2 抽样调查方案

抽样调查方案主要包括：调查的总体，调查的项目，调查的问卷，以及用什么样的方式进行抽样，具体抽取多少调查单位才能保证调查要求的精度和信度，提出相应的汇总资料和处理数据的方法，给出统计推断与预测的方法等。

统计调查的实施

1.2.3 问卷与表格的设计

抽样调查广泛地应用于人文科学方面。为了记录测量值以及指导测量过程，需要设计一份调查意问卷以及一些管理的表格。精心设计的标准问卷可以从不同的研究对象那里获得同样形式的数据。问卷应当能够使被调查者完全明确调查者的意图而乐意配合得出正确的回答，并能使调查机构便于对问卷进行计算机自理，得出正确的统计推断与预测。

1．问卷的结构与格式

a. 问卷首页应印有访员的自我介绍，表明代表哪个公司或机构作什么样调查；

统计调查的实施

b. 确定被访问的人员；

c. 主体问卷：主体问卷是指为了达到调查目的，设计的一系列问答项目。它安排的顺序是先易后难，注意逻辑关系；

d. 被调查者的背景资料：主要是指年龄、性别、文化程度、职业、收入等。一般安排在卷未。

2．主体问卷的设计

以减少非抽样误差为目标。由于人类认知与自身行为通常存在差异，在回答有关涉及到对某此问题的态度及个人隐私问题时会出现较大的偏差。例如，在问卷中提问：你喜欢学习吗？你有暴力倾向吗？你喜欢打架吗？都是

统计调查的实施

不妥的。这时需要通过婉转的提问方式。例如，在问卷中提问：你认为绝大部分时间都放在学习上好吗？等等。

也要避免问卷中出现诱导性的问题以至得出不正确的回答。例如，提问为：“你不同意……吗？”

另外，问卷不能太长。设计的问卷与被调查者可建立和谐的气氛。

1.2.4 调查与数据处理

1．访问过程中始终保持中立，不使用任何诱导性语言；

2．忠实被访问者的回答，严禁在不经询问或被访问

统计调查的实施

人未回答时，擅自代答或代填问卷；

3．除出现某种情况下规定的跳答题外，任何问题都要按规定询问；

4．在访问即将结束时，访问员应当场浏览遍问卷，把漏问、漏记和不明确的地方与被访问者核对，补充完整；

5．应尽可能地在问卷中详细记录被访问人的姓名、单位、地址及电话等，以便于复核；

6．问卷的所有答案都要编号、编码。编号为问题和答案的顺序；编码为答案结果的代码。

统计调查的实施

1.2.5 调查总结报告

任何一次调查都应该有一份总结报告。报告的主体部分是调查的结果必须完全依据调查数据及其分析来写。所有的推断与预测，都必须依数据而行。总结报告中对于调查的各阶段都要有详细的说明。总结报告是调查组织对对该项调查进行评价的主要依据之一。总结报告必须具有客观性。

总结报告一般包括以下几点：

1. 进行调查的背景与目标；
2. 用于指导调查活动的各项工作；
3. 包括对调查结果的介绍和讨论。

抽样误差与非抽样误差

1.3 抽样误差与非抽样误差

1.3.1 抽样误差

由抽样调查所抽取总体的一部分得出样本来估计总体的指标或参数而产生的误差称为抽样误差。抽样误差本身并不是错误的结果，尽管在抽样设计时，判断上的错误可能导致更大的不必要的错误。

如果记总体参数真值为 μ，用样本估计它的值为 $\tilde{\mu}$，抽样误差主要指 μ 与 $\tilde{\mu}$ 的差异。

1．实际误差：$e = \mu - \tilde{\mu}$. 从抽样的角度看，e 是一个随机变量，对于一个具体的样本，e 为一数字。

抽样误差与非抽样误差

2．平均抽样误差：

a. 平均绝对误差：$E|\mu - \tilde{\mu}|$；

b. 平均平方（误）差(mean square error)：$MSE(\rho) = E(\mu - \tilde{\mu})^2$.

c. 标准（误）差：$\sqrt{MSE(\rho)}$.

3．极限误差，又称误差上限，通常是指误差允许的最大上限，记为 Δ（或称公差）。例如：$MSE(\tilde{\mu}) \leq \Delta^2$. Δ 就是标准误差的上限，即标准误的极限误差。

4．相对误差：平均误差与真值的比

a. 相对绝对误差：$\frac{E|\mu-\tilde{\mu}|}{|\mu|}$.

b. 相对均方误差：$\frac{MSE(\tilde{\mu})}{\mu^2}$.

抽样误差与非抽样误差

c. 相对标准误差：$\sqrt{\frac{MSE(\tilde{\mu})}{\mu^2}}$.

1.3.2 非抽样误差

由于其它原因产生的误差叫做非抽样误差，非抽样误差往往使所得资料偏离调查对象的真实情况，致使预测、推断陷入困境，产生极大的偏差。人们通常认为非抽样误差的发生完全由于调查程序发展的执行中的错误和不足。

产生非抽样误差的的几个主要原因：

1．调查计划不周的误差；

2．划分调查对象范围的误差；

3．抽样方案不当产生的误差；

抽样调查的基本知识

2.1.1 总体

总体（ Population, 亦译为母体）表示从中抽取样本的那个集合体，也即研究对象的全体。总体元素是搜集信息的单位。它们是一些个体，是要进行推断的总体所构成的基本单位。它们也是分析单位性质由调查目的所确定。

总体和元素是共同定义的。总体是元素的集合，而元素是构成和确定总体的基本单位。总体的定义必须包括：(1) 内容； (2) 单位； (3) 范围； (4) 时间。

2.1.2 总体的特征数

总体特征数，也称总体指标，它是说明总体数量特征或规律性的数字。一个总体指标是整个总体中所有 N 个

抽样调查的基本知识

元素的某种特征的概括值的数量表现，它是特定总体变量分布的某些特征的综合计量。虽然抽样是为了各种目的，但通常的兴趣都来集中于总体的四项标志：均值 $\overline{X}$; 总值 X; 两个总值的比率或两个均值的比率 $R = \frac{Y}{X} = \frac{\overline{Y}}{\overline{X}}$; 属于某一特定类的单位所占的比例 (proportion).

1. 总体平均数与总体成数

设总体的容量为 N ,各单元标志值为 $X_1, X_2, \cdots, X_N$ 则

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

抽样调查的基本知识

称为该标志上的总体平均数，而

$$T = \sum_{i=1}^{N} X_i$$

称为在该标志上的总体总量。

又设 N 个总体单元中，具有某特征的单元数为 M，则

$$P = \frac{M}{N}$$

称为该特征的总体成数，也称为总体比例或总体频率。

2. 总体方差、修正总体方差与总体标准差

总体方差：$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2 = \frac{1}{N}\sum_{i=1}^{N} X_i^2 - \overline{X}^2$.

抽样调查的基本知识

修正总体方差： $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2.$

总体标准差： $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2}$.

一般，若 X_i 出现的频数为 $f_i, i = 1, 2, \cdots, k$ ，则

$$\overline{X} = \sum_{i=1}^{k} f_i X_i, \quad \sigma^2 = \frac{1}{N}\sum_{i=1}^{k} f_1 X_i^2 - \overline{X}^2.$$

2.1.3 样本与统计量

由总体按照随机原则（保证总体中每个单元都的一定概率被抽取的原则）抽取的部分单元称为随机样本，简称为样本 (Sample) ，被抽入的样本单元，称为样本单元，

抽样调查的基本知识

样本中所含的单元个数，称为样本容量或样本的大小，通常用 n 表示。

严格地说，样本所代表的只是调查总体的样本值（或统计量），是从样本的 n 个元素计算出的估计值。一个特定的估计值只是同一样本设计所能得到的许多可能的估计值中的一个。相反，总体值是由总体中全部 N 个值决定的。虽然总体值是未知的，但它是一个常数，不受各种各样变化的影响。

以 $x_1, x_2, \cdots, x_n$ 表示各样本单元所可能的取值，样本是 n 个随机变量；在样本取定后，它们又是具体的数字，仍用 $x_1, x_2, \cdots, x_n$ 表示，称为样本观察值。

抽样调查的基本知识

抽样调查的基本方法是以样本推断总体，需要针对总体特征数构造相应的特征数。

一般地，设 $\hat{\theta} = \hat{\theta}(x_1, x_2, \cdots, x_n)$ 为样本的连续函数，不包括未知参数，则称 $\hat{\theta}$ 为一个统计量。

1. 样本平均数与样本成数

样本平均数： $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$

样本总量： $T_s = \sum_{i=1}^{n} x_i.$

样本成数： $p = \frac{m}{n}$，其中 m 为具有某特征的单元数。

2. 样本方差、样本标准差与样本二阶中心矩

抽样调查的基本知识

样本方差：$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2 = \frac{1}{n-1}(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2)$.

样本标准差：$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}$.

样本二阶中心矩：$s_n = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$.

2.1.4 概率抽样与非概率抽样

概率抽样又称为随机抽样，具有以下特点：

1．能够确切地定义（或划分）不同的样本；

2．对于每一个被抽取的单位，均赋与一个被抽取有概率；

抽样调查的基本知识

3．抽取是通过与上述相一致的随机抽取方法进行的；

4．用样本估计总体时，要考虑上述被抽取的概率。

概率最佳方法的原因：

1．概率抽样避免调查者在抽样过程中有意无意的偏差；

2．具有深厚的理论基础，概率抽样可以对抽样误差做出估计；

3. 其可度量性可导致客观的统计推断；

4. 可以把误差的来源区分开来，并对戎作客观评价，从而使之逐步改善。

抽样调查的基本知识

非概率抽样的常见类型：

1．样本限于总体中易于取得的部分进行抽取；例如从卡车上取煤检验煤的质量。实际上，我们经常用任意的和不正规的样本作出有关总体的推断。如尝一个葡萄作出好坏的判断，随意地接待几件物品检验后变接收一批进货；

2．样本是随机选取的，但不是经过有意识、有计划设计的。没有赋于每一个个体被抽取的概率，对总体无知。例如在游行队伍中随意找两个人访问，从实验室的大笼子随意取一只兔子做实验等等；

3．对于一个小而内部判别比较大的总体，抽选一小部分“典型的”单位作为样本，即很接近调查者对总体

抽样调查的基本知识

平均数印象的那些单位；

4．调查过程对被套调查对象是不愉快的或麻烦的，调查中的样本是自愿者构成的样本。如药品试验，特别是与军事有关的人体试验；

2.1.5 抽样单位与样本框

1．抽样单位

将总体划分为互不重叠且有限多个部分，每个部分称为一个抽样单位。抽样单位又可以有大小之分。一个大的抽样单位（例如省）又可以划分为若干个小的单位（例如县）。前者称为初级单位或一级单位；后者称为次级单位

抽样调查的基本知识

或二级单位。如此下去，最小的抽样单位称为基本抽样单位。

2．抽样框

划分抽样框单位并将它们编号的进程就是构成抽样框的过程。抽样框是有关抽样单位的一本多册，一份地图或一份档案，其中每一个抽样单位都被编上适当的号码。

包括所有样本单位的抽样框，称为“完全抽样框”。完全抽样框包含总体的所有个体，且每个个体只出现一次。

编制抽样单位的名单（称为抽样框），常常是主要的实际问题之一。抽样单位的名单常常是不完全的，或有一部分是模糊不清、难以辨认的，或含有未知的重复部分。

几种常见的抽样方法

2.2 几种常用的抽样方法

抽样从方法上讲包括两方面：样本抽选和总体推估。

样本抽选是指样本的抽选方式是等概率还是不等概率，重复还是不重复；总体推估是指总体推估方式是简单推估还是比率推估或回归推估。

2.2.1 简单随机抽样

简单随机抽样包括简单重复抽样和简单不重复抽样。通常为其它抽样方法的基础。

几种常见的抽样方法

2.2.1 分层抽样

分层抽样又称类型抽样，它是将总体中的单位分成若干个组（层），设总体分为 K 层（组），将每层看作是一个子总体，对这些子总体分别进行简单随机抽样，称为分层抽样。

分层的原则是将彼此比较相似的单位归为一层（组）（例如在居民收入调查中，以职业分层，如工人、农民、公务员、教师等等分类），所以又称为类型抽样。分层标志无疑是与调查变量密切相关的那些标志，但这些标志既有品质标志，也有数量标志。而且，两者在抽样效率上也有差别。

几种常见的抽样方法

2.2.3 整群抽样

整群抽样又称为集团抽样，它是将总体分成若干个组（整群）。要求每个群尽可能的代表总体（如在全国范围内分省调查）。整群抽样分组的原则是使组内方差较大，而组间方差较小。

抽样方法：先从 N 个群中抽出 K 个群，在 K 个入样群中进行全面调查，所以称为整群抽样或集团抽样。

2.2.4 等距抽样

等距抽样又称系统抽样或机械抽样，是单阶段中花样最多的一种抽样方法。它简单明了，快速经济，操作方便，

几种常见的抽样方法

适用面广。

2.2.5 多阶抽样

多阶抽样又称多级抽样，它是将调查分成两个或两个以上的阶段进行抽样。多阶段抽样主要用于调查那些用单阶段调查不可行或花费太大的范围广大的总体，如农产品的抽样调查，居民家计抽样调查和人口变动的抽样调查等。

目 录

简单随机抽样

3.1.1 简单随机抽样的概念

从总体中抽取样本单元，就抽样方式而言，可以分为重复抽样和不重复抽样。重复抽样是从总体中每抽取一个单元观察、记录后放回去，然后再抽取下一个单元，直到抽取决所需的单元数为止；不重复抽样是从总体中每抽取一个单元观察、记录后不放回去，直到抽足所需要的单元数为止。

简单随机抽样

简单随机抽样：从容量为 N 的有限总体 X 中，随机抽取 n 个单元组成样本，每个这样的样本被抽取的概率相同。

对于重复抽样，简单随机抽样具有下述特性：

1．代表性：即 $x_1, x_2, \cdots, x_n$ 具有与总体 X 相同和概率分布；

2．独立性：即 $x_1, x_2, \cdots, x_n$ 相互独立。

对于有限不重复抽样，每个样本有相同的概率 $\frac{1}{C_N^n}$ 被抽中；当 N 很大，抽样比 $f = \frac{n}{N}(\text{Sampling fraction})$ 很小时，不重复抽样与重复抽样的估计结果很接近。

设总体为 $\Omega = \{ X_1, X_2, \cdots, X_N \}$. 设 ξ_i 为第 i 次抽

简单随机抽样

取的随机变量,则在重复抽样下抽取的样本为 $x_1, x_2, \cdots, x_n$ 的概率为

$$P\{\xi_1 = x_1, \xi_2 = x_2, \cdots, \xi_n = x_n\} = \prod_{i}^{n} P\{\xi_i = x_i\} = \frac{1}{N^n}.$$

在不重复抽取的样本为 $x_1, x_2, \cdots, x_n$ 的概率为

$$P\{\xi_1 = x_1, \xi_2 = x_2, \cdots, \xi_n = x_n\} = \frac{n}{N}\frac{n-1}{N-1}\cdots\frac{1}{N-(n-1)}.$$

简单随机抽样

注 1. 从名单上选出的号码必须代表惟一的 n 个不同元素。就名单本身而言，总体中的每个元素都确切的由一个号码来代表。

注 2. 当采用不重复抽样时，样本的容量 n 不能超过总体的容量 N.

注 3. 在实际的抽样调查中，我们很少采用简单随机抽样。

3.1.3 简单随机抽样的理论意义

1. 数学性质很简单；很多统计理论和假定元素是用简单随机抽样方法来选取的；

简单随机抽样

2. 所有的概率抽样都可以看成是对简单随机抽样的限制;

3. 相对简单的简单随机抽样公式经常用于得到复杂抽样中所需的相关数据;

4. 用于是“设计效果“的比较中。

3.1.2 简单随机抽样的方法

先将总体编号，即编制样本框。

1．另制 N 个号签，也同样分别标以 $1, 2, \cdots, N$ 等号码，将全部号签充分摇匀，根据需要按照重复方式或不重复方式，从中随机抽取 n 个号签。

简单随机抽样

2．随机数法

a. 随机数字表的构造：附表是中国科学院数学研究所概率统计室编印的《常用数理统计表》中的一种，该表分两页，每页分 50 行， 50 列，共有 50000 个数字。

b. 随机数字表的使用原则

1). 每次使用时，确定哪页哪行为起点，必需是随机的；

2). 设总体容量为 N ，若 N 为 r 位数，则要从 r 位数中抽取，遇到 1 至 N 的数可直接使用；其它的不能直接使用；

3). 当 $n \geq$ 时，可以从含有起点数字左边的 r 位数

简单随机抽样

开始，也可从右边的 r 位数开始；

4). 在重复抽样抽样时，遇到重复的数字应重复使用；在不重复抽样时，

遇到重复数字应舍去不用。

3．随机数字表的使用方法

1). 确定起点页码：用笔尖在表上随机指定一点，落点的位置为奇数，则起点为第一页，否则为第二页；

2). 确定起点的行数与列数：先在表上随机指定一点，由落点处的两位数确定起点的行数。当起点的两位数大于 50 时，取其减去 50 的差数为行数；当落点的处的数为 “00” 时，行数取为 50.

简单随机抽样

3). 确定抽取单位的号码：从上述确定的起点开始向下（或向右），每次取一个 r 位数，达到最后一行（列）时，向右（下）移 10 位，再从第一行（列）开始继续向下（右）抽取，直到取足所需的 n 个 r 位数为止。

例：设某县有 320 个村，今抽取 10 个村为样本，试利用随机数字表，按不重复抽样方式确定样本的号码。

3.1.4 估计量及其误差

在对总体参数进行抽样估计时，一般需要解决下述三个问题：

估计量的构造：针对待估计的总体参数，通过样本构

简单随机抽样

造一个合适的统计作为总体参数的估计量；

判别估计量的优良性：对所构造的估计量的优良性作出判断，并且在必要时，对估计量进行修（改）正；

估计量的误差范围：在给定的可靠程度下，求出抽样估计的误差范围。

1．估计量

定义：对于总体的未知参数 θ，由样本构造一个统计量 $\hat{\theta}(x_1, x_2, \cdot, x_n)$ 对 θ 作出估计，则称 $\hat{\theta}$ 为 θ 的估计量。

例：以样本平均数 $\bar{x}$ 作为 $\overline{X}$ 的估计量；

以 $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ 作为 σ^2 的估计量；

简单随机抽样

以 $\hat{T} = N\hat{\overline{X}} = N\overline{x}$ 作为 $T = N\overline{X}$ 的估计量。

2．估计量的优良性

a. 无偏性

定义：设 $\hat{\theta}$ 为总体未知参数 θ 的估计量，若 $E(\hat{\theta}) = \theta$，则称 $\hat{\theta}$ 为 θ 的无偏估计量。

例： $E(\overline{x}) = \overline{X}.$

因为有

$$E(\overline{x}) = E(\frac{1}{n}\sum_{i=1}^{n} x_i) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \overline{X}.$$

注. 设 ξ 为在容量为 N 总体中一次等概抽取元素的随机变量，则 ξ 满足如下分布

简单随机抽样

ξ	X_1	X_2	$\cdots$	X_N
P	$\frac{1}{N}$	$\frac{1}{N}$	$\cdots$	$\frac{1}{N}$

于是,

$$E(\xi) = \sum_{i=1}^{N} P\{\xi = X_i\} X_i = \frac{1}{N} \sum_{i=1}^{N} X_i = \overline{X},$$

$$E(\xi^2) = \sum_{i=1}^{N} P\{\xi = X_i\} X_i^2 = \frac{1}{N} \sum_{i=1}^{N} X_i^2,$$

$$E[\xi - E(\xi)]^2 = E(\xi^2) - [E(\xi)]^2$$

$$= \frac{1}{N} \sum_{i=1}^{N} X_i^2 - \overline{X}^2 = \sigma^2.$$

简单随机抽样

命题. 在简单重复抽样条件下， $E(s^2) = \sigma^2$.

证：

$$
\begin{aligned}
E(s^2) &= E[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2] \\
&= \frac{1}{n-1}E\{\sum_{i=1}^{n}[(x_i - \overline{X}) - (\overline{x} - \overline{X})]^2\} \\
&= \frac{1}{n-1}E\{\sum_{n=1}^{n}[(x_i - \overline{X})^2 - 2(x_i - \overline{X}) \\
&\quad (\overline{x} - \overline{X}) + (\overline{x} - \overline{X})^2]\} \\
&= \frac{1}{n-1}\sum_{i=1}^{n}E[(x_i - \overline{X})^2 - nE(\overline{x} - \overline{X})^2] \\
&= \frac{1}{n-1}(n\sigma^2 - n\frac{\sigma^2}{n}) = \sigma^2.
\end{aligned}
$$

简单随机抽样

由此可见,

$$E(s_n^2) = E(\frac{n-1}{n}s^2) = \frac{n-1}{n}\sigma^2 < \sigma^2,$$

样本的二阶中心矩是总体方差的有偏估计量。

注. 在证明中我们用到重复抽样下随机变量独立同分布的性质

$$E(\overline{x} - \overline{X})^2 = E[\overline{x} - E(\overline{x})]^2 = D(\overline{x})$$

$$= D[\frac{1}{n}\sum_{i=1}^{n} x_i] = \frac{1}{n^2}\sum_{i=1}^{n} D(x_i) = \frac{\sigma^2}{n}.$$

即 $\sigma^2(\overline{x}) = \frac{\sigma^2}{n}$.

b. 有效性

简单随机抽样

定义：设 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 为总体未知参数 θ 的无偏估计量，即 $E(\hat{\theta}_1) = E(\hat{\theta}_2) = \theta$，若有 $\sigma^2(\hat{\theta}_1) < \sigma^2(\hat{\theta}_2)$，则称 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 有效。

注. “最小方差无偏估计”是一种最优的估计量，可惜它有时并不存在，有时无偏性标准显得不合理。此标准表明：离散趋势越小，出现误差的可能性较小，即样本的估计在一定可信度下越精确。

c. 一致性

定义：对于任意给定的 $\epsilon > 0$，如果当 $n \to \infty$ 时，恒有

$$\lim_{n\to\infty} P\{|\hat{\theta}_n - \theta| < \epsilon\} = 1,$$

简单随机抽样

则称 θ_n 为 θ 的一致估计量。

例. $\lim\limits_{n\to\infty} P\{|\overline{x}_n - \overline{X}| < \epsilon\} = 1.$

注. 此标准表明：按照某种特定方法的一组与样本数目有关的统计量，当样本数目增大时，估计越精确。

定义：$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2.$

对于 $MES(\hat{\theta})$ 可作如下分解：

$$\begin{aligned} MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 &= E\{[\hat{\theta} - E(\hat{\theta})] + [E(\hat{\theta}) - \theta]\}^2 \\ &= E[\hat{\theta} - E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2 \\ &\quad + 2E\{[\hat{\theta} - E(\hat{\theta})][E(\hat{\theta}) - \theta]\} \\ &= E[\hat{\theta} - E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2 = \sigma^2(\hat{\theta}) + B^2(\hat{\theta}). \end{aligned}$$

简单随机抽样

其中， $B(\hat{\theta})$ 为估计量的偏差。于是有

$$\text{总误差} = \text{变量误差} + \text{偏差}$$

对于无偏估计量，有 $B(\hat{\theta}) = 0,\ MSE(\hat{\theta}) = \sigma^2(\hat{\theta}).$

定义：估计量的标准差

$$\sigma(\hat{\theta}) = \sqrt{E[\hat{\theta} - E(\hat{\theta})]^2}$$

称为标准误，简称标准差。

标准误度量一个无偏估计量抽样误差平均的大小，故又称抽样平均误差。例如，对于简单重复抽样，有

$$\sigma^2(\overline{x}) = \frac{1}{n^2}\sum_{i=1}^{n} D(x_i) = \frac{\sigma^2}{n}, \quad \sigma(\overline{x}) = \frac{\sigma}{\sqrt{n}}.$$

简单随机抽样

研究和计算估计量方差和标准差的作用：

1). 比较各种抽样技术的估计效率；

2). 在设计抽样方案时，估计最低需要抽取的样本单元数；

3). 计算抽样估计的误差范围。

3.1.5 区间估计

定义：对于总体未知参数 θ，根据样本构造两个统计量

$$\hat{\theta}_1 = \hat{\theta}_1(x_1, x_2, \cdots, x_n),\ \hat{\theta}_2 = \hat{\theta}_2(x_1, x_2, \cdots, x_n),$$

且 $\hat{\theta}_1 < \hat{\theta}_2$，而使随机区间 $(\hat{\theta}_1, \hat{\theta}_2)$ 包含 θ 的概率等于给

简单随机抽样

定值 $1-\alpha$，即 $P\{\hat{\theta}_1 < \theta < \hat{\theta}_2\} = 1-\alpha$ 成立，则 $1-\alpha$ 称置为信概率或置信度，$(\hat{\theta}_1, \hat{\theta}_2)$ 称为置信区间，$\hat{\theta}_1$ 和 $\hat{\theta}_2$ 分别称为置信下限和置信上限。

误差限：$\Delta(\hat{\theta})$，即估计 θ 落在 $\hat{\theta} - \Delta(\hat{\theta})$ 与 $\hat{\theta} + \Delta(\hat{\theta})$ 之间。

在抽样中，若能保证 $P\{|\hat{\theta} - \theta| < \Delta(\hat{\theta})\} = 1-\alpha$，则可作如下结论：总体参数 θ 的估计值为 $\hat{\theta}$，误差限为 $\Delta(\hat{\theta})$，可靠程度为 $1-\alpha$.

$\hat{\theta}$ 的相对误差：$\Delta'(\hat{\theta}) = \frac{\Delta(\hat{\theta})}{\theta}$.

估计的精度：$P_c = 1 - \Delta'(\hat{\theta}) = 1 - \frac{\Delta(\hat{\theta})}{\theta}$. 通常 θ 是未知的，可用估计量 $\hat{\theta}$ 作出估计。

简单随机抽样

注 1. 优良的置信区间应是区间长度，即

$$\hat{\theta}_2(x_1, x_2, \cdots, x_n) - \hat{\theta}_1(x_1, x_2, \cdots, x_n)$$

比较小。

注 2. 置信区间的好处在于它以一定把握保证 θ 在此区间中。

3.2 简单重复抽样

3.2.1 总体平均数的估计方法

1．估计量及无偏性

在简单重复抽样条件下，$\bar{x}$ 为 $\overline{X}$ 的无偏估计量。

简单随机抽样

2．估计量 $\overline{x}$ 的概率分布

根据中心极限定理，在大样本下，估计量 $\overline{x}$ 近似服从正态分布。

3．估计的方法

假定 $\overline{x}$ 服从正态分布，则由

$$E(\overline{x}) = \overline{X},\ D(\overline{x}) = \sigma(\overline{x}),$$

随机变量 $u = \frac{\overline{x}-\overline{X}}{\sigma(\overline{x})}$ 满足标准正态分布 $N(0,1)$. 于是有，

$$P\{|u| \geq u_\alpha\} = 1 - \frac{1}{\sqrt{2\pi}} \int_{-u_\alpha}^{u_\alpha} e^{-\frac{t^2}{2}} \mathrm{d}t = \alpha,$$

$$P\{|u| < \alpha\} = 1 - \alpha.$$

简单随机抽样

即 $P\{|\frac{\overline{x}-\overline{X}}{\sigma(\overline{x})}| < u_\alpha\} = 1-\alpha$ 等价地

$$P\{\overline{x} - u_\alpha\sigma(\overline{x}) < \overline{X} < \overline{x} + \sigma(\overline{x})\} = 1-\alpha.$$

于是估计量 $\overline{x}$ 的误差限为：

$$\Delta(\overline{x}) = u_\alpha\sigma(\overline{x}).$$

在简单重复抽样的条件下为：

$$\Delta(\overline{x}) = u_\alpha\frac{\sigma}{\sqrt{n}}.$$

总体平均数 $\overline{X}$ 的 $1-\alpha$ 置信区间为：

$$(\overline{x} - u_\alpha\sigma(\overline{x}),\ \overline{x} + u_\alpha\sigma(\overline{x})).$$

简单随机抽样

在简单重复抽样的条件下为：

$$(\overline{x} - u_\alpha \frac{\sigma}{\sqrt{n}},\ \overline{x} + u_\alpha \frac{\sigma}{\sqrt{n}}).$$

例 1．某市自来水公司以简单重复抽样方法随机抽取了 400 户居民，调查每户平均用水量为 4.51 吨。根据以往的资料可知，用水量服从正态分布，且标准差为 0.8 吨。试以 95% 的可靠程度估计该市居民每户每月的平均用水量。

解. 由 $n = 400,\ \overline{x} = 4.51,\ \sigma = 0.8$ 及 $1 - \alpha = 95\%,\ u_\alpha = 1.96$ 得

$$\Delta(\overline{x}) = u_\alpha \frac{\sigma}{\sqrt{n}} = 1.96 \times \frac{0.8}{\sqrt{400}} \approx 0.08.$$

简单随机抽样

结论：该市居民每月平均用水量为 4.51 吨，误差限为 0.08 吨，估计的可靠程度为 95% 。该市每户每月平均用水量可靠程度为 95% 的置信区间为 $(4.51-0.08, 4.51+0.08)$.

例 2. 某食品工厂为控制产品包装质量，从生产的大批罐头中随机抽取 30 瓶，测得每瓶净重数据如下（单位：克）

401	400	401	399	402	402	398	397	400	401
402	401	399	398	400	402	397	401	398	400
398	401	396	400	404	405	396	397	401	400

假定每瓶罐头的重量服从正态分布，试以 95% 的可靠程度估计该厂生产的罐头每瓶平均重量。

简单随机抽样

解. 已知 $n = 30, 1-\alpha = 95\%, u_\alpha = 1.96$. 由于总体方差 σ^2 未知，用样本方差 s^2 作为估计值。

经计算 $\overline{x} = 399.9$, $s = 2.2$. 所以， $\Delta(\overline{x}) \approx 1.96 \times \frac{2.2}{\sqrt{30}} = 0.8$(克).

结论：每瓶罐头平均重量为 399.9 克，误差限为 0.8 克，可靠程度为 95%.

3.2.2 总体总量的估计方法

1．估计量及无偏性

总体总量 T 的估计量为： $\widehat{T} = N\overline{x} = \frac{N}{n}\sum_{i=1}^{n} x_i$, $\frac{N}{n}$ 称为放大系数。

简单随机抽样

由

$$E(\widehat{T}) = NE(\overline{x}) = N\overline{X} = T,$$

$\widehat{T}$ 是 T 的无偏估计量。

2．估计方法

由 $\sigma^2(\widehat{T}) = \sigma^2(N\overline{x}) = N^2\sigma(\overline{x})$，估计量 $\widehat{T}$ 的标准误为：$\sigma(\widehat{T}) = N\sigma(\overline{x})$.

概率论中已证明，正态随机变量经线性变换后仍为正态随机变量。由于在正态总体或一般总体大样本条件下，$\overline{x}$ 近似服从正态分布，所以，$\widehat{T}$ 也服从或近似服从正态分布。

简单随机抽样

事实上，由 $\hat{T}=N\overline{x}, T=N\overline{X}, \frac{\hat{T}-T}{N\sigma(\overline{x})}\sim N(0,1)$. 而

$$P\left\{\left|\frac{\hat{T}-T}{N\sigma(\overline{x})}\right|<u_\alpha\right\}=1-\alpha$$

等价于

$$P\{\hat{T}-Nu_\alpha\sigma(\overline{x})<T<\hat{T}+Nu_\alpha\sigma(\overline{x})\}=1-\alpha.$$

于是，当给置信概率为 $1-\alpha$ 时(或概率度为 u_α 时), 估计量的误差限应为 $\Delta(\widehat{T})=u_\alpha(\widehat{T})=N\Delta(\overline{x})$.

由此，T 的估计值为 $\widehat{T}$，误差限为 $\Delta(\widehat{T})=u_\alpha(\widehat{T})=N\Delta(\overline{x})$, 可靠程度为 $1-\alpha$，$1-\alpha$ 置信区间为：

$$(N\overline{x}-N\Delta(\overline{x}), N\overline{x}+N\Delta(\overline{x})).$$

简单随机抽样

例 3. 在例 1 中，若该市有 100000 户居民，试以 95% 的可靠程度估计该市居民的每月用水总量。

解. 由 $N = 100000, \overline{x} = 4.51, \Delta(\overline{x}) = 0.08$ ，得

$$\widehat{T} = N\overline{x} = 451000(\text{吨}),\ \Delta(\widehat{T}) = N\Delta(\overline{x}) = 8000(\text{吨}).$$

可靠程度为 95%.

3.2.3 样本容量的确定

在简单重复抽样的条件下，如果误差限给定，则由 $\Delta(\overline{x}) = u_\alpha \frac{\sigma}{\sqrt{n}}$ 及 $\Delta(\widehat{T}) = N\Delta(\overline{x})$ 有

$$n_0 = \frac{u_\alpha^2 \sigma^2}{\Delta^2(\overline{x})} = N^2 \frac{u_\alpha^2 \sigma^2}{\Delta^2(\widehat{T})}.$$

简单随机抽样

例 4. 在例 2 中，将所抽样本作为试抽样本，以 $s^2 = (2.20)^2$ 作为 σ^2 的估计值，若要求误差不超过 0.5 克，可靠程度 95%，试求最少应检验多瓶罐头。

解. 已知 $\sigma^2 = s^2 = (2.20)^2$，$u_\alpha = 1.96$，$\Delta(\overline{x}) \leq 0.5$，则

$$n_0 = 1.96^2 \times \frac{(2.20)^2}{(0.5)^2} = 75,$$

即至少应检验 75 瓶，才能以 95% 的可靠程度保证最大误差不超过 0.5 克。

目 录

分层随机抽样

4.1.1 分层随机抽样的基本方法

分层抽样（ Stratified sampling ）也称类型抽样或分类抽样。先将包含 N 个单位的总体分成各包含 $N_1, N_2, \cdots, N_k$ 个单位的子总体。这些子总体互不重复，因此

$$N_1 + N_2 + \cdots + N_K = N.$$

这些子总体就称为层。要从分层获得充分的好处，必须知道 N_i 的值。层被确定后，就从每一层抽一个样本，抽样是在各层独立进行的。各层内的样本含量分别用 $n_1, n_2, \cdots, n_K$

分层随机抽样

表示。

如果从每层抽取一个简单随机样本，则整个方法就称为分层随机抽样 (Stratified random sampling).

大体上讲，分层抽样包括以下步骤：

1. 抽样单位的总体被分成互相区别的次级总体，称为层；

2. 在每一层中，从构成层的所有抽样单位中选出一个单独的样本；

3. 用从每一层所得的样本计算出一个单独的层平均数（或其它统计量）。层的平均数经过适当的加权就得到对总体一个合并估计量；

分层随机抽样

4. 方差也在各层分别计算，然后通过适当的加权合并并形成对总体的估计量。

在实际中，专业研究人员对本专业中研究与考察的总体一般都有某些了解，也对总体掌握了某些信息。

分层的主要理由：

1．需要有总体的某些分类的数据，且要具有规定的精度；

2．为了便于行政管理而要求分层。例如调查全面教育水平情况，大学、中学、小学分属不同的管理机构；

3．总体的各个不同部分的抽样问题可能显著不同。如在商业抽样调查中，我们可能掌握一份大商店的名单，

分层随机抽样

要将它们单独放在一个层内。对于较小的商店，则可能采用某种区域性的抽样方法。又如，为了调查的方便，在调查公司职员时，按白领和蓝领公层；

4．分层可能提高整个总体指标估计值的精确度。它可以将一个内部差异很大的总体分成一些内部比较相似的子总体。例如，男人和女人在购买行为上有较大的差异；

在总体名单的性质或标志值的大小明显呈现层次时，按其层次将总体划分为若干个组，每组称为一层。

分层时，使每层内部单元的差异较小，即齐性较好；使差异主要存在于各层之间。然后，在每一层都进行随机抽样，并针对总体的目标（即待估参数值）构造一统计量。

分层随机抽样

一个理想的例子是，分层使每层元素的统计值相同，些时层中的任一元素可作为整个层的代表。可见，分层增加样本的代表性，从而提高抽样的代表性。

分层时应遵守的原则是：尽可能使层内差异小，而使层间的差异大。

例：不同年龄段人群的消费习惯存在差异，如年龄儿童玩具选择；不同性别人群消费的习惯存在差异，如人们对化妆品挑选；不同文化程度的人群存在行为上的差异，如处事为人。

究竟选用哪个标志更好，无疑取决于我们的调查目的。如工业调查可按冶金、电力、机械等分层；农业经济

分层随机抽样

调查可按农、林、牧、副、渔业分层；农产量调查按平原、丘陵、山地分层。

调查学生的学习状况可按学生每天平均学习时间分层，也可按学习成绩的好坏分层；调查女性购买化妆品的情况，可按收入状况分层，也可按年龄分层。这些例子说明，同一总体可按不同的调查要求分层。

4.1.2 分层抽样中的一些名词和记号

假定总体容量为 N, 层数为 K, 第 i 层的单元数为 N_i. 设总体为 $\Omega = \{ X_1, X_2, \cdots, X_N \}$. 各次级总体，即各层为

$$\Omega_i = \{ X_{i1}, X_{i2}, \cdots, X_{iN_i} \},\ i = 1, 2, \cdots, K.$$

分层随机抽样

分层要求满足：

$$\Omega_i \cap \Omega_j = \Phi, \quad i \neq j, \quad i, j = 1, 2, \cdots, K,$$

$$N_1 + N + 2 \cdots + N_K = N.$$

记 $W_i = \frac{N_i}{N}, i = 1, 2, \cdot, K, \quad X_{ij}, i = 1, 2, \cdots, K; j = 1, 2, \cdots, N_i$ 为第 i 层第 j 个单元的标志值，则 $\overline{X}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} X_{ij}$ 为第 i 层总体平均数，总体平均数为

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{K} \sum_{j=1}^{N_i} X_{ij} = \frac{1}{N} \sum_{i=1}^{K} N_i \overline{X}_i = \sum_{i=1}^{K} W_i \overline{X}_i.$$

可见，总体平均数为各层平均数的加权平均数。

总体方差为：$\sigma^2 = \frac{1}{N} \sum_{i=1}^{K} \sum_{j=1}^{N_i} (X_{ij} - \overline{X})^2.$

分层随机抽样

总体的第 i 层方差、层内方差与层间方差分别定义为：

$$\sigma_i^2 = \frac{1}{N_i}\sum_{j=1}^{N_i}(X_{ij} - \overline{X}_i)^2,$$

$$\overline{\sigma_i^2} = \frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{N_i}(X_{ij} - \overline{X}_i)^2 = \frac{1}{N}\sum_{i=1}^{K} N_i\sigma_i^2 = \sum_{i=1}^{K} W_i\sigma_i^2,$$

$$\sigma_p^2 = \frac{1}{N}\sum_{i=1}^{K}(\overline{X}_i - \overline{X})^2.$$

可见，总体层内方差为总体各层方差的加权平均数；总体层间方差为各层平均数的方差。

命题 1：$\sigma^2 = \overline{\sigma_i^2} + \sigma_p^2$

分层随机抽样

证.

$$
\begin{aligned}
\sigma^2 &= \frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{N_i}(X_{ij}-\overline{X})^2 \\
&= \frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{N_i}[(X_{ij}-\overline{X}_i)+(\overline{X}_i-\overline{X})]^2 \\
&= \frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{N_i}[(X_{ij}-\overline{X}_i)^2 \\
&\quad + 2(X_{ij}-\overline{X}_i)(\overline{X}_i-\overline{X})+\overline{X}_i-\overline{X})^2] \\
&= \frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{N_i}(X_{ij}-\overline{X}_i)^2+\frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{N_i}(\overline{X}_i-\overline{X})^2 \\
&\quad + \frac{2}{N}\sum_{i=1}^{K}\sum_{j=1}^{N_i}(X_{ij}-\overline{X}_i)(\overline{X}_i-\overline{X})^2
\end{aligned}
$$

分层随机抽样

而

$$\sum_{i=1}^{K}\sum_{j=1}^{N_i}(X_{ij}-\overline{X}_i)(\overline{X}_i-\overline{X})=\sum_{i=1}^{K}[(\overline{X}_i-\overline{X})\sum_{j=1}^{N_i}(X_{ij}-\overline{X}_i)]$$

$$=\sum_{i=1}^{K}(\overline{X}_i-\overline{X})(N_i\overline{X}_i-N_i\overline{X}_i)$$

$$=0.$$

故 $\sigma^2=\overline{\sigma_i^2}+\sigma_p^2$.

4.1.3 分层抽样应满足的条件

1．任一总体单元属于且仅属于某一层；

2．N_i 是确定的，进而 $W_i=\frac{N_i}{N}$ 是确知的；

3．在各层所抽取的样本是相互独立的。

4.2 总体平均数的分层抽样估计

4.2.1 估计量及其无偏性

设抽取样本容量为 n, $n_i, i = 1, 2, \cdots, K$ 为第 i 层样本单元数 $(\sum_{i=1}^{K} n_i = n)$,

$$x_{ij}, i = 1, 2, \cdots, K, ; j = 1, 2, \cdots, n_i$$

第 i 层第 j 个样本单元的标志值，则第 i 层样本平均数为 $\overline{x}_i = \frac{1}{n_i} \sum_{i=1}^{n_i} x_{ij}, i = 1, 2, \cdots, K.$

第 i 个次级总体（层）为：

$$\Omega_i = \{ X_{i1}, X_{i2}, \cdots, X_{iN_i} \}.$$

从中以简单重复抽样或简单不重复抽样抽取 n_i 个样本，记为

$$x_{i1}, x_{i2}, \cdots, x_{in_i},$$

则第 i 个总体（层）的平均数为

$$\overline{x}_i = \frac{1}{n_i}\sum_{i=1}^{n_i} x_{ij}, i = 1, 2, \cdots, K$$

且 $\overline{x}_i$ 为 $\overline{X}_i$ 的无偏估计量。

以第 i 层的样本平均数，$\overline{x}_i$ 作为该层总体平均数 $\overline{X}_i$ 的估计量，则总体平均数 $\overline{X}$ 的估计量

$$\overline{x}_{st} = \frac{1}{N}\sum_{i=1}^{K} N_i\overline{x}_i = \sum_{i=1}^{K} W_i\overline{x}_i.$$

命题 2. $\overline{s}_{st}$ 为 $\overline{X}$ 的无偏估计量。

证. 由

$$E(\overline{x}_{st}) = E(\frac{1}{N}\sum_{i=1}^{K} N_i\overline{x}_i) = \frac{1}{N}\sum_{i=1}^{K} N_i E(\overline{x}_i)$$
$$= \frac{1}{N}\sum_{i=1}^{K} N_i\overline{X}_i = \overline{X}.$$

得证。

4.2.2 估计量的方差

由于在各层抽取的样本是相互独立的， $\overline{x}_1, \overline{x}_2, \cdots, \overline{x}_K$

也是相独立的。因此有

$$\sigma^2(\overline{x}_{st}) = \sigma^2(\frac{1}{N}\sum_{i=1}^{K} N_i\overline{x}_i) = \frac{1}{N^2}\sum_{i=1}^{K} N_i^2\sigma^2(\overline{x}_i) = \sum_{i=1}^{K} W_i^2\sigma^2(\overline{x}_i). \tag{*}$$

对于分层重复抽样，由 $\sigma^2(\overline{x}_i) = \frac{\sigma_i^2}{n}, i = 1, 2, \cdots, K$, 得

$$\sigma^2(\overline{x}_{st}) = \frac{1}{N^2}\sum_{i=1}^{K} N_i^2\frac{\sigma_i^2}{n_i}. \tag{4.1}$$

对于分层不重复抽样, 由 $\sigma^2(\overline{x}_i) = \frac{\sigma_i^2}{n_i}\frac{N_i - n_i}{N_i - 1}, i = 1, 2, \cdots, K$, 得

$$\begin{aligned}\sigma^2(\overline{x}_{st}) =& \frac{1}{N^2}\sum_{i=1}^{K} N_i^2 \frac{\sigma_i^2}{n_i}\frac{N_i - n_i}{N_i - 1} \\ \approx& \frac{1}{N^2}\sum_{i=1}^{K} N_i^2 \frac{\sigma_i^2}{n_i}(1 - \frac{n_i}{N_i}) \qquad (4.2) \\ =& \frac{1}{N^2}\sum_{i=1}^{K} N_i(N_i - n_i)\frac{\sigma_i^2}{n_i}.\end{aligned}$$

从 (4.1), (4.2) 可见，估计的方差与层内方差有关，而与层间方差无关。由命题 1，层间方差越大，层内方差越小，估计量的方差越小，估计越精确。

注 1. 第 i 个次级总体的方差为：

$$\sigma_i^2 = \frac{1}{N_i}\sum_{j=1}^{N_i}(X_{ij} - \overline{X}_i)^2.$$

注 2. 重复抽样时，估计量 $\overline{x}_i$ 的方差为：

$$\sigma^2(\overline{x}_i) = \frac{\sigma_i^2}{n_i};$$

不重复抽样时，估计量 $\overline{x}_i$ 的方差为：

$$\sigma^2(\overline{x}_i) = \frac{\sigma_i^2}{n_i}\frac{N_i - n_i}{N_i - 1}.$$

注 3. 比较 (4.1), (4.2) 知，在相同的分层下，分层不重复抽样的效率高于分层重复抽样。

4.2.3 估计量方差的估计量

在一般情况下，各层方差 σ_i^2 是未知的。对于分层重复抽样，在第 i 层以

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)^2$$

估计 σ_i^2, 则

$$s^2(\overline{x}_i) = \frac{s_i^2}{n_i}$$

为 σ_i^2 的无偏估计量. 令

$$s^2(\overline{x}_{st}) = \frac{1}{N^2} \sum_{i=1}^{K} N_i^2 \frac{s_i^2}{n_i}, \tag{4.3}$$

则由

$$Es^2(\overline{x}_{st}) = E(\frac{1}{N^2}\sum_{i=1}^{K}N_i^2\frac{s_i^2}{n_i}) = \frac{1}{N^2}\sum_{i=1}^{K}N_i^2E(\frac{s_i^2}{n_i})$$

$$= \frac{1}{N^2}\sum_{i=1}^{K}N_i^2\frac{\sigma_i^2}{n_i} = \sigma^2(\overline{x}_{st}).$$

$s^2(\overline{x}_{st})$ 为 $\sigma^2(\overline{x}_{st})$ 的无偏估计量。

对于分层不重复抽样，第 i 层方差 σ_i^2 的无偏估计量为 $\frac{N_i-1}{N_i}s_i^2$, 即有

$$E(\frac{N_i-1}{N_i}s_i^2) = \sigma_i^2.$$

又 $\sigma^2(\overline{x}_i) = \frac{\sigma_i^2}{n_i}\frac{N_i-n_i}{N_i-1}$, 上式两边同乘 $\frac{N_i-n_i}{n_i(N_i-1)}$ 得

$$E[\frac{s_i^2}{n_i}(1-\frac{n_i}{N_i})] = \sigma^2(\overline{x}_i).$$

令

$$s^2(\overline{x}_{st}) = \frac{1}{N^2}\sum_{i=1}^{K} N_i^2 \frac{s_i^2}{n_i}(1-\frac{n_i}{N_i})$$
$$= \frac{1}{N^2}\sum_{i=1}^{K} N_i(N_i-n_i)\frac{s_i^2}{n_i}, \qquad (4.4)$$

则由

$$Es^2(\overline{x}_{st}) = E(\frac{1}{N^2}\sum_{i=1}^{K} N_i^2 \frac{s_i^2}{n_i}(1-\frac{n_i}{N_i}))$$
$$= \frac{1}{N^2}\sum_{i=1}^{K} N_i^2 \frac{\sigma_i^2}{n_i} = \sigma^2(\overline{x}_{st}),$$

得 $s^2(\overline{x}_{st})$ 为 $\sigma^2(\overline{x}_{st})$ 的无偏估计量。

4.2.4 估计量误差限

若总体各层为正态分布（ x_{ij} 也服从正态分布），则各层的样本平均数 $\overline{x}_i$ 服从正态分布。于是，估计量

$$\overline{x}_{st} = \sum_{i=1}^{K} W_i \overline{x}_i$$

也服从正态分布。

若总体各层为非正态分布，则当 n_i 都充分大时， $\overline{x}_i$ 的分布近似正态分布，从而 $\overline{x}_{st}$ 的分布也近似正态分布。可见，在分层抽样中，只有当 N, N_i 及 n_i 都充分大时，才能讨论误差限的问题。因为在通常的情况下我们并不知道总体和各层的分布。

若给定了概率度 u_α，则误差限为

$$\Delta(\overline{x}_{st}) = u_\alpha \sigma(\overline{x}_{st}) \approx u_\alpha s(\overline{x}_{st}). \tag{4.5}$$

对于分层重复抽样 $\sigma(\overline{x}_{st}), s(\overline{x}_{st})$ 分别由 (4.1), (4.3) 给出；对于分层不重复抽样 $\sigma(\overline{x}_{st}), s(\overline{x}_{st})$ 分别由 (4.2), (4.4) 给出。

目 录

整群随机抽样

5.1.1 整群随机抽样的方法

一般地说, 如果将总体划分为若干组, 每组称为一个群或集团, 把每个群作为一个抽样单位, 整群地进行抽样; 然后, 在被抽中的群内进行全面调查。这就是整群抽样 (cluster sampling), 也称为分群抽样或集团抽样。

假定总体被划分为 N 群, 第 i 群有 $M_i, i = 1, 2, \cdots, N$ 个总体单元。记 $M_0 = \sum_{i=1}^{N} M_i$. 当 $M_i, i = 1, 2, \cdots, N$ 都相等时, 称为等群; 否则称为不等群。

整群随机抽样

在上述假设下，总体为

$$\Omega = \{X_{11}, \cdots, X_{1M_1}, X_{21}, \cdots, X_{2M_2}, X_{N1}, \cdots, X_{NM_N}\}.$$

定义

$$\Omega_i = \{X_{i1}, X_{i2}, \cdot, X_{iM_i}\}, i = 1, 2, \cdots, N$$

则 $\tilde{\Omega} = \{\Omega_1, \Omega_2, \cdots, \Omega_N\}$ 构成一个新的总体。满足：

$$\Omega_i \cap \Omega_j = \Phi,\ i \neq j,\ i, j = 1, 2, \cdots, N$$

且

$$\bigcup_{i=1}^{N} \Omega_i = \Omega.$$

在抽样方式上，可以重复抽样，也可以不重复抽样；可以是等概抽取，也可以是不等概抽取。

整群随机抽样

5.1.1 整群抽样与分层抽样的比较

与分层抽样类似，将全部总体单元分群时，要求满足：

1．总体中每个单元都必须属于且只能属于群中的某一群；

2．每群所含总体单元数是确知的。

分层的原则：尽量缩小层内差异，而扩大层间差异；

分群的原则：尽量扩大群内差异，而缩小群间差异。

整群随机抽样

表 5.1　　整群抽样与分层抽样比较

		分层（分类、类型）抽样	整群（分群、集团）抽样
分组（层或群）必须满足的条件		1. 既无重复又无遗漏； 2. 层单元数 N_i 确知； 3. 各层抽样相互独立。	1. 同左； 2. 群单元数 M_i 确知。
调查方式	组间	层间全面调查（在各层都抽取样本单元）	群间抽样调查（抽样单位是群，群是扩大了的总体单元）
	组内	层内抽样调查（层是缩小的总体，抽样单位总体单元）	群内全面调查
分组原则	组间	尽可能扩大层间差异（抽样误差与层间方差 σ_p^2 无关）	尽可能缩小群间差异（抽样误差只与群间方差 B^2 有关）
	组内	尽可能缩小层内差异（抽样误差只与层内方差 $\overline{\sigma_i^2}$ 有关）	尽可能扩大群内差异（抽样误差与群内方差 σ_W^2 无关）
总体方差分解		$\sigma^2 = \overline{\sigma_i^2} + \sigma_p^2$	$\sigma^2 = \sigma_W^2 + \sigma_B^2$

整群随机抽样

5.1.2 实际中采用整群抽样方法的原因

1. 当个别地抽选元素过于昂贵时整骆抽样来完成调查任务，即每个抽样单位包含若干元素。例如，例如公司职工可工作的群体抽选。学校的学生可按班抽选。城市可按各街区抽选等等。

2. 有些调查只适用整群抽样。例如整体包装的物品，整箱的烟、饮料、集装箱等。等容量的群通常是计划条件下 (如制造过程) 形成的结果；例如烟的制造。

3. 总体中的个体没有可靠的名单，而编造这样一份名单费用极高。

4. 在样本单元数相同的条件下，整群抽样与简单随机抽样相比，样本单元相对集中，虽然样本的代表性较差，但可大大节省人力、财力和时间。

我们在这里

- 问题答疑： http://www.xxwenda.com/
  - 可邀请老师或者其他人回答问题

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：小象
- 新浪微博：ChinaHadoop