第七章 统计抽样

本章系统介绍统计抽样基本概念以及简单随机抽样、分层抽样、整群抽样以及系统抽样相关理论。

7.1 统计抽样基本概念

在前面我们给出了总体和样本的定义，即

- 总体由研究对象的全体所组成。
- 样本是总体中的部分元素所组成的集合。

为说明这些概念，我们以四川长虹电子集团公司为例。为了制定企业下一步战略，该公司打算对其液晶电视机购买者进行一次调查。本次抽样调查的对象是长虹液晶电视机的购买者。总体由购买长虹液晶电视机的所有人组成，样本是长虹液晶电视机购买者的一个子集。

在抽样调查中，有必要区分目标总体与抽样总体。目标总体是我们要推断的总体，抽样总体是实际抽取样本的总体，这两个总体不总是一致的，明确这一点非常重要。在长虹电子集团公司的例子中，目标总体是购买长虹液晶电视机的所有人，抽样总体是将保修登记卡寄回长虹电子集团公司的所有液晶电视购买者。由于有一些长虹液晶电视购买者并没有寄回保修卡，其抽样总体和目标总体是不一致的。抽样调查获得的结果只适用于抽样总体，这些结果是否能扩展到目标总体需要依靠分析家的判断。

在抽样之前，应将总体划分为抽样单位。抽样单位既可以是一个简单的个体，也可以是一组个体。假设我们要调查持有律师资格证书的专业律师。如果可以利用所有专业律师名册，则抽样单位就是我们所要调查的专业律师。如果这样的名册不可利用，我们就必须依靠其他方法来找到调查对象。我们可以利用电话号码簿，查出所有律师事务所的名册，进而调查专业律师。这时，抽样单位是指每一家律师事务所。

在具体研究中，抽样单位的名册称为抽样框。在专业律师调查中，如果专业律师名册不可以利用，那么律师事务所名册就是抽样框。在实际抽样调查过程中，编制抽样框是一个既困难又重要的步骤。

7.2 抽样调查种类和抽样方法

最常用的三种调查是邮寄调查、电话调查和个人采访调查，而且每一种调查都需要设计和使用调查表。

在使用调查表的调查中，设计调查表是非常关键的问题。设计者必须要抵制想囊括所有要研究问题的诱惑，因为每增加一个问题都会增加调查表的长度。长的调查表不仅使回答者感到疲劳，也使采访者感到疲劳，尤其对邮寄和电话调查更是如此。但是，如果用个人采访调查，较长而且复杂的调查表是行得通的。

根据使用的抽样方法，抽样调查可分为概率抽样和非概率抽样。用概率抽样，可以计算出取得的每个可能样本的概率；用非概率抽样，则无法得知取得每个可能样本的概率。如果调查者想对估计的精度做出说明，应采用概率抽样。根据给定的允许误差，采用概率抽样方法可构造相应的置信区间。在后面几节中，我们将讨论四种概率抽样方法：简单随机抽样、分层简单随机抽样、整群抽样和系统抽样。

尽管统计学家喜欢用概率抽样方法，但非概率抽样方法常常是必要的。非概率抽样的优点是成本低且容易完成；缺点是不能对估计的精度做出准确的说明。两种最常用的非概率抽样方法是方便抽样和判断抽样。

方便抽样是根据调查者的方便性，以无目标、随意的方式进行的抽样调查活动。例如，一名教授在大学里进行一项调查研究，他可以邀请他的学生参加他的研究项目，仅仅是因为这些学生在他的班上。这时，学生样本称为方便样本。常见的街头拦访和随意的入户访问也是方便抽样的常见形式。

尽管方便抽样是选择样本和收集资料的一种相对简单的方法，但是对这样取得的样本统计量，无法评价由它们所估计出的总体参数的“优良性”。有时，研究人员将方便样本看成是一个随机样本，但是这样得出的结论会受到质疑。因此，用方便样本对总体参数进行推断时，必须非常小心。

在非概率抽样技术中，根据个人的主观意识来选择对总体有代表性的抽样单位的方法，称为判断抽样。尽管判断抽样常常是选择样本的一种相对容易的方法，但调查结果的使用者必须清楚地认识到，这些结果的质量依赖于个人在选择样本时的判断。因此，用判断样本对总体参数进行统计推断时也应非常小心。

7.3 调查误差

进行抽样调查可产生两类误差，一类是抽样误差，它是所得到的样本点估计值与总体参数之间的数量差异。换句话说，抽样误差是由于没有对总体的所有单位进行调查而产生的误差；另一类是非抽样误差，它包括进行一次抽样调查可能出现的所有其他类型的误差，如测量误差、采访者误差及数据处理误差等。抽样误差仅出现在抽样调查中，而非抽样误差则既可以出现在普查中，也可以出现在抽样调查中。

7.3.1 非抽样误差

我们不能准确地测量要研究的特征，这是最常见的非抽样误差形式之一，在调查中，调查人员必须十分仔细，保证测量工具（如调查表）非常准确，而且应对调查人员进行必要的培训。在多数情形下，注意细节是最好的防范措施。

由于没有回答所产生的误差是负责设计调查的统计人员和使用调查结果的管理人员都非常关心的问题。当不能得到或只能部分得到某些被调查单位的资料时，就会产生这类非抽样误差。出现偏差是很严重的问题。例如，对妇女外出工作的看法进行调查，若只在白天做家庭采访，就会出现明显的偏差。因为，外出工作的妇女没有包含在样本中。

另外两种类型的非抽样误差是选择误差和数据处理误差。当调查中包含不恰当的项目时，就会生产选择误差。假设设计一个抽样调查，来描述有胡须的男人外观。对“有胡须的男人”的理解，如果有些采访人员认为应包括有小胡子的男人，而其他采访人员则不这样认为，那么调查的结果将有缺陷。如果研究者没有将调查表中的资料正确输入计算机时，就会产生输入错误；而如果调查人员将被调查所选择信息填写错误时，便产生登记错误。当有登记错误或输入错误时，就会出现数据处理误差。

尽管在大多数的调查中，会出现一些非抽样误差，但通过周密的计划可使它们达到最小，这些计划诸如注意保证抽样总体与目标总体的一致、遵循良好调查表的设计原则、培训采访人员等。在调查结论中，研究者应对非抽样误差所可能产生的影响予以讨论。

7.3.2 抽样误差

回忆在 7.1 节介绍的长虹液晶电视调查问题中，假设长虹集团想估计购买长虹液晶电视的人的平均年龄。如果可以调查长虹液晶电视购买者的整个

总体（普查），则不存在抽样误差，同时我们也可以准确地计算他们的平均年龄。但如果不能调查长虹液晶电视拥有者的整个总体，调查结果将如何呢？这时，样本均值与总体均值之间可能存在差异，差异的绝对值即为抽样误差。

由于调查的只是一个样本，而不是整个总体，因此抽样误差必然存在。在实际调查中，由于总体均值是未知的，因此不可能知道抽样误差的大小，但可以对其进行概率说明。尽管抽样误差不可避免，但却是可以控制的。选择合适的抽样方法是控制这类误差的一个重要的方法。在下面几节中，我们将讨论四种概率抽样方法：简单随机抽样、分层简单随机抽样、整群抽样和系统抽样。

7.4 简单随机抽样

从一个容量为 N 的有限总体中抽取得到一个容量为 n 的简单随机样本，并且每一个容量为 n 的可能样本，都有相同的概率被抽中。

用简单随机抽样进行抽样调查，首先应建立一个抽样框，即抽样总体中所有个体的名册；然后根据随机数表进行抽样。使用随机数表，可以保证抽样总体中的每个个体都有相同的概率被抽中。在这一节中，我们将介绍用简单随机抽样对总体均值、总体总量及总体比率进行的估计。

7. 4.1 总体均值

在大多数的抽样调查中，总体概率分布的形式是未知的。例如，在长虹液晶电视抽样调查中，公司想估计购买长虹液晶电视的人的平均年龄 μ，显然，长虹集团不知道所有长虹液晶电视拥有者年龄的概率分布的形式。不知道总体概率分布形式通常不是问题，因为 μ 的点估计 $\bar{x}$ 的抽样分布性质，仅仅依赖于样本设计的选择。

如果选择大样本（n≥30），则中心极限定理可以保证 $\bar{x}$ 的抽样分布近似服从正态概率分布。当 $\bar{x}$ 的抽样分布近似服从正态概率分布时，μ 的区间估计为

$$\bar{x} \pm \mu_{\alpha/2}\sigma_{\bar{x}} \tag{7-1}$$

式中 $\sigma_{\bar{x}}$——均值的标准差。

1－α 称为置信度，$\mu_{\alpha/2}$ 为与之对应的临界值。例如，置信度为 95%时，$\mu_{0.025}=1.96$。

当从一个容量为 N 的有限总体中，抽取一个容量为 n 的简单随机样本时，均值的标准差的估计值为

$$s_{\bar{x}}=\sqrt{\frac{N-n}{N-1}}\left(\frac{s}{\sqrt{n}}\right) \tag{7-2}$$

这时，总体均值的区间估计为

$$\bar{x}\pm\mu_{\alpha/2}s_{\bar{x}} \tag{7-3}$$

在抽样调查中，当构造置信区间时，通常取 μ=2。因此，在使用简单随机样本时，总体均值的近似 95%的置信区间的表达式如下。

$$\bar{x}\pm 2s_{\bar{x}} \tag{7-4}$$

[例 7.1]《摄影》是一本推介摄影作品、报道摄影发展状况、介绍摄影器材的杂志，它目前拥有 8000 个订户。根据一个 484 个订户的简单随机样本，得出订户的年平均收入为 30500 元，标准差为 7040 元。因此，所有订户的年平均收入的无偏估计为 $\bar{x}=30500$ 元。

$$s_{\bar{x}}=\sqrt{\frac{8\,000-484}{8\,000-1}}\left(\frac{7\,040}{\sqrt{484}}\right)=310$$

因此，根据式（7－4），得到这本杂志订户的年平均收入的近似 95%的置信区间为 30500±2×310＝30500±620，即（29880，31120）。

上述过程也可用于对诸如总体总量或总体比率等其他总体参数的区间估计。对点估计的抽样分布近似服从正态概率分布的所有情形，其近似 95%的置信区间为：

点估计值±2×点估计量的标准误差的估计值

例如，在《摄影》杂志的抽样调查中，点估计量的标准误差的估计值为 310 元，允许误差为 2×310 元＝620 元。

7.4.2 总体比率

总体比率 p 是总体中具有某些兴趣特征的个体占总体的比重。

[例 7.2]在市场调查研究中，人们想了解喜欢某一品牌的消费者比重。样本比率 $\bar{p}$ 是总体比率 p 的无偏点估计。总体比率的标准差的估计值为

$$s_{\bar{p}} = \sqrt{\left(\frac{N-n}{N-1}\right)\left(\frac{\bar{p}(1-\bar{p})}{n}\right)} \tag{7-5}$$

因此，总体比率的近似 95%的置信区间的表达式如下：

$$\bar{p} \pm 2s_{\bar{p}} \tag{7-6}$$

例如，在大宇国际咨询公司的抽样调查中，还想估计所调查的 500 所学校中，使用天然气作为取暖燃料的学校比率。如果在抽出的 50 所学校中，有 35 所学校使用天然气作为取暖燃料，则总体 500 所学校中使用天然气比率的点估计值 $\bar{p} = 35/50 = 0.70$ 。根据式（7-5），比率的标准差的估计值为

$$s_{\bar{p}} = \sqrt{\frac{500-50}{500-1} \times \frac{0.7 \times (1-0.7)}{50}} = 0.0621$$

因此，根据式（7-6），总体比率的近似 95%置信区间为

$$0.7 \pm 2 \times 0.0621 = 0.7 \pm 0.1242$$

即（0.5758,0.8242 ）。

由此例可知，当估计总体比率时，置信区间的宽度可能很宽。一般地，需要大的样本容量，以保证获得总体比率的估计精度。

7.4.3 样本容量的确定

在抽样设计中，样本容量的选择是一个重要的问题，通常需要对经费和精度进行权衡。较大的样本可以提供较高的精度（允许误差较小），但费用较多。通常，研究的预算将决定样本容量的大小。在无预算限制的情况下，样本容量应该选取足够大来满足规定的精度水平。

通常，选择样本容量的方法是首先规定所需要的精度，然后确定满足精度的最小的样本容量。这里，精度涉及近似置信区间的大小，较小的置信区间可以提供较高的精度。因此，近似置信区间的大小依赖于允许误差 B，即选择精度水平相当于选择 B 的值。下面我们介绍估计总体均值时，选择所必需的样本容量的方法。

式（7-2）为均值的标准误差的估计公式，即

$$s_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{s}{\sqrt{n}}\right)$$

回忆前面提到的允许误差为“点估计的标准差估计值的 2 倍”，因此

$$B = 2\sqrt{\frac{N-n}{N-1}}\left(\frac{s}{\sqrt{n}}\right) \tag{7-7}$$

解式（7-7），则有

$$n = \frac{Ns^2}{(N-1)\left(\frac{B^2}{4}\right)+s^2} \tag{7-8}$$

可见，一旦给出了所需要的精度水平（通过选择 B 的值来实现）根据式（7-8），便可以得到满足所需要精度水平的 n 值。但是，对一个实际研究的问题，根据式（7－8）确定 n 值时，除了规定所需要的允许误差之外，还必须知道样本方差 s^2。但是只有实际得到样本时， s^2 才可以知道。

下面是三种估计 s^2 方法：

(1) 用两步抽样：由第一步抽取的部分单位，得到的 s^2 的估计值，将此值代入（7-8），确定出全部样本容量 n；然后根据第一步所确定的全部样本容量，再抽取剩余的单位数。

(2) 用试点调查或事先检验的结果估计 s^2。

(3) 根据以往的资料估计 s^2。

[例 7.3]某大学有 5000 名毕业生，我们想构造宽度在 1000 元之内的近似 95％的置信区间。对这样规定的置信区间，B＝500。在根据式（7-8）确定 n 之前，我们需估计 s^2。假设根据去年所做的同样研究，得知 s＝3000 元。我们可以用这个值来估计 s^2。根据 B=500、 s＝3000 及 N＝5000，根据式（7-8），则样本容量为

$$n=\frac{5000\times 3000^2}{(5000-1)\times\frac{500^2}{4}+3000^2}=140.0$$

综上所述，对规定宽度为 1000 美元的近似 95%的置信区间，所需要的样本容量为 140。但应当记住，这个计算结果依赖于 s 的最初估计值 s＝3000 元。如果在今年的抽样调查中，s 变得较大，则近似置信区间的宽度将大于 1000 元。因此，如果经费允许，调查设计者可选取样本容量为 150，以增加近似 95%的置信区间的宽度小于 1000 元的把握程度。

在估计总体比率时，选择样本容量的公式，与估计总体均值的公式类似。我们只需要将式（7-8）中 s^2 替换为 $\bar{p}(1-\bar{p})$，即

$$n=\frac{N\bar{p}(1-\bar{p})}{(N-1)\left(\frac{B^2}{4}\right)+\bar{p}(1-\bar{p})} \tag{7-9}$$

使用式（7-9）时，我们必须规定允许误差 B 和给出 $\bar{p}$ 的一个估计值。如果没有 $\bar{p}$ 的合适的估计值，我们可用 $\bar{p}=0.5$ 代替，这样将保证近似置信区间的允许误差比希望的要小的多。

7.5 分层简单随机抽样

在分层简单随机抽样中，首先将总体划分 H 组（称为层），然后从第 h 层中抽取一个容量为 n_h 的简单随机样本。由这 H 个简单随机样本的联合资料，可得出诸如总体均值、总体总量及总体比率等各种总体参数的估计。

如果各层内的差异比层间的差异小，则分层简单随机样本可得到更大的精度（总体参数的区间估计将更窄），各层的划分应依据样本设计者的判断。根据应用，总体可按部门、地区、年龄、产品类型、销售水平等分层。[例 7.4]某大学管理学院想对今年的毕业生进行一次调查，以便了解他们开始工作时的年薪。假设该学院有 5 个专业：会计、金融、信息系统、市场营销和经营管理。今年有 1500 名毕业生，其中会计专业 500 名，金融专业 350 名，信息系统专业 200 名，市场营销专业 300 名，经营管理专业 150 名。根据以往年薪资料的分析表明，开始工作时的年薪在专业间的差异比专业内大。因此，选择 180 名学生的一个分层简单随机样本，其中会计专业 45 名，

金融专业 40 名，信息系统专业 30 名，市场营销专业 35 名，经营管理专业 30 名。

7.5.1 总体均值

在分层抽样中，总体均值的无偏估计是各层样本均值的加权平均数，所用权数为总体在各层的比重。用 $\bar{x}_{st}$ —表示总体均值的点估计，其定义如下。

$$\bar{x}_{st}=\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)\bar{x}_h \tag{7-10}$$

式中 H ——层数； $\bar{x}_h$ ——第 h 层的样本均值；

N_h ——第 h 层的单位数； N ——总体单位数； $N=N_1+N_2+\cdots N_H$ 。

对分层简单随机样本，平均值的标准差的估计公式为

$$s_{\bar{x}_{st}}=\sqrt{\frac{1}{N^2}\sum_{h=1}^{H}N_h(N_h-n_h)\frac{s_h^2}{n_h}}$$

表 7-1 为某管理学院的 180 名毕业生的样本调查结果。

表 7-1 毕业生开始工作时年薪的抽样调查

专业(h)	$\bar{x}_h$ / 元	s_h	N_h	n_h
会计	30000	2000	500	45
金融	28500	1700	350	40
信息系统	31500	2300	200	30
市场营销	27000	1600	300	35
经营管理	31000	2250	150	30

各专业（层）的样本均值分别为：会计 30000 元、金融 28500 元、信息系统 31500 元、市场营销 27000 元，经营管理 31000 元。根据这些资料及式（7-10），总体均值的点估计为

$$\bar{x}_{st}=\frac{500}{1500}\times 30000+\frac{350}{1500}\times 28500+\frac{200}{1500}\times 31000$$
$$+\frac{300}{1500\times 27000}+\frac{150}{1500}\times 31000=29350\text{元}$$

表 7-2 给出了估计标准差所需要的部分计算结果

表 7-2　抽样调查中估计均值的标准差所需要的部分计算结果

专业	h	$N_h(N_h-n_h)\frac{s_h^2}{n_h}$
会计	1	20 222 222 222
金融	2	7 839 125 000
信息系统	3	5 995 333 333
市场营销	4	5 814 857 143
经营管理	5	3 037 500 000
合计		42 909 037 698

其中

$$\sum_{h=1}^{5} N_h(N_h-n_h)\frac{s_h^2}{n_h}=42909037698$$

因此

$$s_{\bar{x}_{st}}=\sqrt{\frac{1}{1500^2}\times 42909037698}=\sqrt{19070.68}=138$$

总体的近似 95%的置信区间为

$$29350\pm 2\times 138=29350\pm 276$$

即（29074，29626）。

7.5.2　总体比率

对分层简单随机抽样，总体比率 p 的无偏估计是各层比率的加权平均数，所用权数为总体在各层的比重。用 $\overline{p}_{st}$ 表示总体比率的点估计，其定义如下：

$$\overline{p}_{st}=\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)\overline{p}_h \qquad (7\text{-}11)$$

式中 H —层数； $\overline{p}_h$ —第 h 层的样本比率； N_h －第 h 层的单位数；

N －总体单位数； $N=N_1+N_2+\cdots N_H$ 。

$\overline{p}_{st}$ 的标准差的估计值为

$$s_{\bar{p}_{st}} = \sqrt{\frac{1}{N^2}\sum_{h=1}^{H} N_h(N_h - n_h)\left[\frac{\bar{p}_h(1-\bar{p}_h)}{n_h - 1}\right]} \tag{7-12}$$

因此，总体比率的近似 95%的置信区间的表达式如下

$$\bar{p}_{st} \pm 2s_{\bar{p}_{st}} \tag{7-13}$$

[例 7.5] 某大学的调查中，想了解毕业生开始工作时的年薪不低于 36 000 元的比率。180 名毕业生的抽样调查结果显示，有 20 名毕业生开始工作时的年薪不低于 36000 元，其中会计专业 4 名，金融专业 2 名，信息系统专业 7 名，市场营销专业 1 名，经营管理专业 6 名。

根据式（7-12），开始工作时的年薪不低于 36000 元的比率的点估计为：

$$\bar{p}_{st} = \frac{500}{1500}\times\frac{4}{45} + \frac{350}{1500}\times\frac{2}{40} + \frac{200}{1500}\times\frac{7}{30} + \frac{300}{1500}\times\frac{1}{35} + \frac{150}{1500}\times\frac{6}{30} = 0.0981$$

$$s_{\bar{p}_{st}} = \sqrt{\frac{1}{N^2}\sum_{h=1}^{H} N_h(N_h - n_h)\left[\frac{\bar{p}_h(1-\bar{p}_h)}{n_h - 1}\right]} = \sqrt{\frac{1}{1500^2}\times 924.8305} = 0.0203$$

根据式（7-13）毕业生开始时的年薪不低于 36000 元的比率近似 95%置信区间为（0.0575，0.1387）

7.5.3 样本容量的确定

对分层简单随机抽样，我们可用两阶段过程来选择样本容量。首先应确定总样本容量 n；然后决定各层应分配的样本单位数。或者，首先决定每层应选择的样本单位数，然后加总得到总样本容量。既然人们想估计各层的均值、总量及比率，这两种确定样本容量的方法都经常使用。确定总样本容量 n 及其分配，可对所有要研究的总体参数提供必要的精度。然而，对某些层，如果样本单位数没有达到满足层内估计量所需要的必要精度的数量要求，这些层的样本单位数应根据需要向上调整。本节我们将讨论恰当地分配样本容量的一些问题，给出一种选择总样本容量及分配的方法。

分配工作就是决定总样本被分配到各层的部分，即确定各层的简单随机样本的容量。进行分配时要考虑的重要因素是：

1．各层的单位数。

2．各层内的方差。

3．各层选择单位的费用。

一般地，单位数较多的层和方差较大的层应分配较多的样本数目。相反地，对于给定的费用，为了获得更多的信息，则抽样单位成本较大的层应分

配较少的样本数目。

在许多调查中，抽样单位成本在各层近似相等（例如邮寄和电话调查）。在这种情况下，进行分配时，可以忽略抽样成本。对这种情况，我们可以给出选择样本容量和进行分配的近似公式。抽样成本在层间差异显著情况下的公式，在有关抽样的高级教材中给出。本节我们给出的公式，满足对给定的精度水平使总的抽样成本达到最小的要求。这种方法，即为著名的 Neyman 分配法，将总样本容量 n 分配到各层的结果如下。

$$n_h = n\left(\frac{N_h s_h}{\sum_{h=1}^{H} N_h s_h}\right) \qquad (7\text{-}14)$$

式（7-14）表明分配到各层的单位数受各层容量和标准差的影响。在进行分配之前，我们必须先确定总样本容量 n。对于给定的精度水平 B，当估计总体均值和总体总量时，我们可使用下面的公式确定总样本容量。

估计总体均值时的样本容量

$$n = \frac{(\sum_{h=1}^{H} N_h s_h^2)^2}{N^2\left(\frac{B^2}{4}\right) + \sum_{h=1}^{H} N_h s_h^2} \qquad (7\text{-}15)$$

7.6 整群抽样

整群抽样需要将总体分为 N 组（也称作群），使总体中每个个体只属于一群。例如，我们想调查某省的登记选民。一种方法是建立包含该省所有登记选民的抽样框，然后根据抽样框，选择选民的一个简单随机样本；另一种方法是整群抽样，我们选择用该省各县的清单作抽样框。在这个方法中，每个县（或群）包含一组登记选民，而该省的每个登记选民只属于一群。

假设我们从 88 个县中选一个 n＝12 的简单随机样本。若收集 12 个中选群中所有登记选民的资料，这种方法称作单阶段整群抽样。若我们从 12 个中选群的每一群中，再选择登记选民的一个简单随机样本，这种方法称作二阶段整群抽样。这两种情形，都可用样本结果得到诸如总体均值、总体总量或总体比率等总体参数的点估计和区间估计。这一章我们只考虑单阶段整群抽样，二阶段整群抽样将在有关抽样的高级教材中予以论述。

分层抽样和整群抽样都将总体划分为组，因此这两种抽样过程感觉上是相似的。但是，选择整群抽样与分层抽样的原因是不同的。当群内的个体存

在差异时，整群抽样可提供较好的结果。理想情形是每一群都是整个总体的一个缩影。在这种情形下，抽取很少的群就可以获取关于整个总体特征的信息。

区域抽样是整群抽样的基本应用之一，在这里，群可以是县、区、城市街区或总体其他规定好的地理区域。因为只从整个地理区域（或群）的一个样本中搜集资料，而且群内的个体的代表性彼此相近，因此当资料搜集者或采访者去调查一个抽样单位时，可有效地节约时间和经费。因此，如果需要较大的总样本容量，整群抽样比简单随机抽样或分层简单随机抽样节省费用。另外，建立抽样框或被抽中的个体的清单时，整群抽样可使时间和经费达到最小。因为整群抽样不需要建立总体中每个个体的清单，只需要建立中选群中个体的清单。

[例 7.6]某省注册会计师协会打算对省内 12000 名执业注册会计师进行一项调查。作为调查的一部分，需要收集收入、性别以及与注册会计师生活方式有关的因素信息。因为用个人采访法去搜集所需要的信息，因此注册会计师协会采用整群抽样，以使总的差旅费和采访费用达到最小。抽样框中包含所有在该省登记注册的执业会计师事务所。

假设有 1000 个群，即在该省登记注册的从事会计活动的会计师事务所有 1000 个，选择 10 个会计师事务所为一个简单随机样本。

为了介绍在整群抽样中，构造总体均值、总体总量和总体比率的近似 95%置信区间需要的公式，我们使用如下的记号：

设 N —总体的群数；　　n—样本中选出的群数；

M_i—i 群的单位数；　　M —总体单位数；

$M = M_1 + M_2 + \cdots + M_N$；　$\overline{M} = M / N$ —每一群的平均单位数；

X_i—第 i 群所有观察值的总量；

a_i—第 i 群具有某特征的观察值的数量；

在注册会计师协会的抽样调查中，我们有如下资料。

$$N = 1000 \qquad n = 10$$

$$M = 12000 \qquad \overline{M} = 12000 / 1000 = 12$$

表 7-3 为每个中选群的 M_i 和 X_i 的值，以及中选事务所中女注册会计师

的数量（a_i）的资料。

表 7-3 注册会计师抽样调查的结果

事务所（i）	注册会计师数量（M_i）	第i个事务所年薪总额x_i/千元	女注册会计师数量（a_i）
1	8	320	2
2	25	1125	8
3	4	115	0
4	17	714	6
5	7	247	1
6	3	94	2
7	15	634	2
8	4	147	0
9	12	481	5
10	33	1567	9
合计	128	5444	35

7.6.1 总体均值

由整群抽样得到的总体均值的点估计的公式如下。

总体均值的点估计 $$\bar{x}_c = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} M_i} \tag{7-16}$$

该点估计量的标准差的估计为

$$s_{\bar{x}_c} = \sqrt{\left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{\sum_{i=1}^{n}(x_i - \bar{x}_c M_i)^2}{n-1}} \tag{7-17}$$

总体均值的近似 95%的置信区间 $\bar{x}_c \pm 2S_{\bar{x}_c}$

根据表 7-3 的资料，我们可以得到执业注册会计师平均年薪的点估计为

$$\bar{x}_c = \frac{5444}{128} = 42.531$$

由于表 7-3 中的年薪资料是以千元计量的，因此，执业注册会计师的平均年薪的估计值为 42531 元。

$$\sum_{i=1}^{n}(x_i-\bar{x}_cM_i)^2=39178.688$$

因此，

$$s_{\bar{x}_c}=\sqrt{\frac{1000-10}{1000\times10\times12^2}\times\frac{39178.688}{10-1}}=1.730$$

因此，标准差为 1.730。我们得到平均年薪的近似 95%置信区间

$$42.531\pm2\times1.730=42.531\pm3.460$$

即 （39.071，45.991）。

7.6.2 总体比率

整群抽样的总体比率的点估计公式如下：

总体比率的点估计 $\bar{p}_c=\dfrac{\sum_{i=1}^{n}a_i}{\sum_{i=1}^{n}M_i}$ （7-18）

式中a_i——第i群中具有某种感兴趣特征的个体的数量。

该点估计量的标准误差的估计为

$$s_{\bar{p}_c}=\sqrt{\left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{\sum_{i=1}^{n}(a_i-\bar{p}_cM_i)^2}{n-1}}$$ （7-19）

总体比率的近似 95%的置信区间 $\bar{p}_c\pm2s_{\bar{p}_c}$ （7-20）

对注册会计师抽样调查，可以得到女性执业注册会计师的比率的估计为

$$\bar{p}_c=\frac{2+8+\cdots+9}{8+25+\cdots+33}=\frac{35}{128}=0.2734$$

$$\sum_{i=1}^{n}(a_i-\bar{p}_cM_i)^2=15.2098$$

因此，$s_{\bar{p}_c}=\sqrt{\dfrac{1000-10}{1000\times10\times12^2}\times\dfrac{15.2098}{10-1}}=0.0341$

因此，女性职业注册会计师比率的近似 95%置信区间为

$$0.2734\pm2\times0.0341=0.2734\pm0.0682$$

即（0.2052,0.3416）。

7.6.3 样本容量的确定

一旦群形成，确定样本容量的基本问题就是选择群的数量。整群抽样的过程同其他抽样方法类似。首先通过选择β（即允许误差）的值，规定可接受的精度水平，然后是建立满足所需要的精度的 n 值的计算公式。每群平均个体的数量和群间方差是决定样本中包含群数多少的关键因素。如果各群相似，则群间方差小，因此中选群数就比较少。另外，如果每群平均个体数量较大，则中选群数也会比较少。准确确定样本容量的公式包含在有关抽样的高级教材中。

7.7 系统抽样

系统抽样常常用来代替简单随机抽样。对某些抽样情况，特别是大型总体，通过先确定随机数，然后根据抽样框寻找与随机数相对应的个体的方法来选择一个简单随机样本，这需要花费大量时间。在这种情况下，系统抽样可代替简单随机抽样。例如，需要从容量为 5000 的总体中抽取一个容量为 50 的样本，我们可以从总体中每 100（5 000 / 50）个个体中抽选一个个体。这种情况的系统样本，是从抽样框的前 100 个个体中随机选择一个；根据选中的第一个个体位置，然后在其后面的抽样框中，每隔 100 个个体选择一个，可得到样本中其余的个体。实际上，通过系统排列总体，及在随机抽取第一个个体后，每隔 100 个来选择一个个体，可以得到一个容量为 50 的样本。用这种方式选择容量为 50 的样本常常比用简单随机抽样容易。因为第一个个体的选择是随机的，因此系统样本常常假定具有简单随机样本的性质。当抽样框是由总体中的个体随机排列而形成时，这种假定通常是合适的。

习题

1、用简单随机抽样，从总体容量为 800 的总体中选择一个容量为 50 的样本，样本均值 $\bar{x} = 215$，样本标准差 s=20。

a.估计总体均值；

b.估计均值的标准误差；

c. 构造总体均值的近似 95％的置信区间。

2、选择一个样本来构造总体均值的 95％的置信区间。假设总体包含 450 个元素，试点研究的结果表明 s=70。如果想构造宽度为 30 的近似 95％的置信区间，应抽取多大容量的样本？

3、将一个总体划分为 3 层，其中N_1=300，N_2=600,和N_3=500。根据过去的调查，各层的标准差的估计如下：s_1=150,s_2=72,s_3=100。

a. 若总体均值估计的允许误差 B=20，则样本容量应为多少？每层应分配多少个单位？

b. 若总体均值估计的允许误差 B=10，则样本容量应为多少？每层应分配多少个单位？

4、一个会计事务所在银行业、保险业和经纪人行业有大量客户，其中有N_1=50 个银行，N_2=38 个保险公司和N_3=35 个经纪人公司。雇用一个市场调查公司来调查这 3 个行业中的该会计师事务所的客户，就客户的业务和他们对会计事务所提供服务的满意情况，询问了大量问题。假设要估计 123 个客户的平均雇员数量的近似 95%置信区间，允许误差为B=30

a.根据一个试点调查得到s_1=80,s_2=150 和s_3=45,选择总样本容量，并确定分配到 3 层中的样本单位数。

b.假设进行试点检验，得到在选择样本容量时，每层的标准差都等于 100，选择总样本容量，并确定每层的抽样单位数。

5、从包含 25 群 300 个元素的总体中选出 4 群。每群的 M_i 、 X_i 、 a_i 的值如下：

群（i）	M_i	x_i	a_i
1	7	95	1
2	18	325	6
3	15	190	6
4	10	140	2
合计	50	750	15

a.求总体均值、总体总量和总体比率的点估计值。

b.对（a）中的各个估计量，分别估计它们的标准误差

c.构造总体总量的近似 95%的置信区间

d.构造总体比率的近似 95%的置信区间.