第 1 章　问卷设计与数据收集

如果想了解某一具体经济领域或企业内外部的情况，以及某一具体目标市场的情况，获取准确的第二手数据是比较困难的，此时一般采用的方法是针对具体的目标进行市场调查，获得第一手数据，并基于此数据进行分析，得出结论。问卷调查是获得第一手数据的主要手段，其涉及的知识比较多，从问卷的设计与分析、调查方法的选择、抽样方案的确定、抽样实施、问卷回收、数据录入到数据分析并得出结论是一个整体。限于篇幅，本章不能对所有的流程进行详细说明，只针对问卷设计与分析、调查抽样方面做简要阐述。本章主要介绍在进行统计分析之前，如何通过问卷调查获得统计数据，以及在进行市场调查时，如何选择抽样方法、如何通过 SPSS 进行抽样和确定样本量，为进一步的统计分析做准备。

1.1　问卷设计与分析

1.1.1　问卷设计

对于问卷设计，主要是在清楚调查内容的基础上，通过对调查问题的合理设计和布局，让问题更好地反映调查内容，一项以第一手数据为基础的研究项目，其研究深度本质上由问卷的深度决定。对于问卷设计时没有考虑到的问题，在问卷调查完成后想研究就不太可能了，因为重新设计问卷、收集数据，无论是时间还是资金的损失，往往都是令人难以承受的。因此，按照所要研究的问题设计好问卷是科学研究中一项非常重要的工作。下面就问卷的构成、问卷的问题类型、问卷中量表的主要类型和问卷设计的注意事项几方面进行阐述。

1．问卷的构成

一份正式的问卷一般包括以下 4 个组成部分：标题、导语（前言）、正文和结束语。

（1）标题。

问卷的标题概括地说明调查主题，使被调查者对所要回答的问题有一个大致的了解。问卷标题要简明扼要，但又必须点明调查对象或调查主题。例如，“学生宿舍卫生间热水供应现状的调研”，而不要简单采用“热水问题调查问卷”这样的标题。这样无法使被调查者了解明确的主题内容，妨碍其回答问题的思路。

（2）导语。

导语主要是对调查目的、意义及填表要求等的说明。导语部分文字必须简明易懂，能激发被调查者的兴趣。导语的书写应注意以下 3 个问题。

- 与调查的内容和调查对象相吻合，突出本次调查的主要问题和现象。
- 要调动被调查者的积极性，体现被调查者完成本次调查的重要作用。
- 要对被调查者表示感谢，写一些祝愿的话语。

另外，卷首最好有说明（称呼、目的、被调查者受益情况、主办单位），如果涉及个人资料，则应该有隐私保护说明。

（3）正文。

将调查的若干问题及相应的选择项目有限度地排列成正文，要求被调查者回答。

（4）结束语。

结束语一般是一段短语，内容是向被调查者的合作再次表示感谢，以及关于不要漏填及复核的请求。结束语要简短明了，有的问卷也可以省略。

2．问卷的问题类型

（1）封闭型问题。

封闭型问题事先准备了答案，被调查者只能在事先准备的答案中选择。封闭型问题的数据转化工作量大为减少。例如，“我在英语课堂活动的参与程度上：A．十分活跃　B．较活跃　C．一般　D．不太活跃　E．根本不参与”，这是固定应答题，以指定答案的方式进行回答。

封闭型问题包括以下问题形式。

- 是否式。

是否式问题后列出两种相互对立的答案，从中选出一个，如“是”与“否”，或者“同意”与“不同意”。例如，“我觉得英语是一门十分有趣且很好学的学科：A．同意　B．不同意”。

- 选择式。

选择式问题是指根据自己的观点和实际情况，从列出的多个答案中挑选出一个或几个答案。例如，“我在课后对课堂所学知识的整理上：A．十分认真　B．较认真　C．一般　D．不太认真　E．一点儿也不认真”；“我觉得英语学习中最重要的是（最多选 3 项）：A．单词记忆　B．语法规则　C．阅读理解　D．口语表述　E．写作能力　F．听力　G．做题技巧　H．语言知识”。

- 评判式。

评判式也叫排列式、编序式，每个问题后列出很多选项，根据选项的重要性，用数字评定等级。例如，“英语学习涉及很多内容，请根据你的观点，按其重要程度由 1 到 5 顺序排列：（）听力技能　（）口语表达技能　（）阅读技能　（）写作技能　（）翻译技能”。

在多数情况下，要尽量用封闭型问题形成问卷。

（2）开放型问题。

开放型问题事先没有准备答案，允许被调查者用自己的话来回答问题，由于采取这种方式提问会得到各种不同的答案，不利于资料的统计，因此通常在问卷形成阶段使用，在最终问卷中要慎用。

3．问卷中量表的主要类型

量表是测量被调查者对某个问题（特别是复合型问题）的反应强度（或态度、看法）的工具。把单选问题的备选答案量化，就得到单问题量表，如表 1-1 所示。

表 1-1　单问题量表

金融危机对你的影响	非常大	大	一般	不大	无影响
	5	4	3	2	1

这就是一个简单的单项量表。“单项”是指该量表仅仅反映了被调查者对一个问题的态度。

（1）连续评分量表。

表 1-1 所示的量表的刻度仅从 1 到 5，如果采用从 0 到 100 的刻度，则称其为连续评分量表。

（2）分项评分量表。

表 1-1 所示的量表只涉及一个单选问题，如果量表涉及多个关联的单选问题，就称它为分项评分量表（Itemized Rating Scale）。分项评分量表中的多个单选问题必须是有关联的，是某个总项（上一层的变量）的一个分解。表 1-2 就是一个分项评分量表的例子。

表 1-2　高校辅导员压力问题的一个分项评分量表

项目	没有压力	有点儿压力	中度压力	压力较大	压力很大
辅导员外出进修和接受继续教育的机会	1	2	3	4	5
辅导员工作权利和义务不明确	1	2	3	4	5
辅导员事务性工作多，用于自身学习提高的时间少	1	2	3	4	5
辅导员工作见效周期长，成果无形化	1	2	3	4	5
辅导员评职称的条件及难度	1	2	3	4	5

这种分项评分量表又称 Likert 量表，是由美国社会心理学家 Rensis Likert 于 1932 年提出的。Likert 量表的度量级别通常为 5 级，但非一定为 5 级，在应用中，7 级、9 级均可，但通常不少于 5 级，不多于 9 级。

Likert 量表的关键特点是所有分项共同组成一个总项，分项的得分加总后就得到总项的得分，因此 Likert 量表又称为加总量表（或求和量表）。

（3）排序量表。

排序量表就是根据所研究的问题对几个因素排列出先后顺序。例如，根据重要性，对影响学生成绩的因素进行排序的量表如表 1-3 所示。

表 1-3　对影响学生成绩的因素进行排序的量表

比较因素	重要性等级
学生智商	5
家长监管	4
学习勤奋程度	3
教学质量	2
社会因素	1

4．问卷设计的注意事项

（1）目的明确。任何问卷调查都是有目的的，要么证实某个结论，要么证伪某个结论，只有目的明确，才能围绕目的设计题项。

（2）先易后难，先简后繁。问卷的前几个问题的设置必须谨慎，称呼语措辞要亲切、真诚。前几个问题要比较容易回答，不要使被调查者难以作答，给接下来的调查造成困难，因此，通常将被调查者熟悉的问题放在问卷的前面。

（3）提出的问题要具体，避免提一般性问题。一般性问题对实际调查工作并无指导意义。例如，“你认为食堂的饭菜供应怎么样？”这样的问题就很不具体，很难达到想了解被调查者对食堂饭菜供应状况总体印象的预期调查效果，应把这一类问题细化为具体询问关于价格、外观、卫生、服务质量等方面的印象。

（4）单选问题的备选答案应完整划分答案空间。单选问题的备选答案必须分布在同一个维度上，是同一个答案空间的完整分割，即备选答案之间不能有交集，也不能有遗漏。例如，“您的家庭月收入是：A．2000 元以下　B．2000～3999 元　C．4000～5999 元　D．6000～7999 元　E．8000 元及 8000 元以上”。这个问题的 5 个备选答案就是一个答案空间的完整划分。也就是说，收入的所有答案都可以在这 5 个备选答案中找到，备选答案之间既没有交集，又没有遗漏。

（5）多选题的备选答案必须分布在两个以上的维度上，并且至少有一部分不是互相排斥的。

（6）问题的陈述及备选答案不能有多重含义。

（7）问题设计的用语要含义明确，不能让被调查者产生不同的理解。

（8）在问题的陈述中，要对所询问行为的时间、方式、目的做必要的限定。

（9）对于得不到诚实回答而又必须了解的情况，可以通过变换问题的提法来获得相应的数据，或者通过了解相对数据来判断总体的情况。

（10）问卷不能太长，应答时间以 20～30min 为宜；对于商场拦截类的问卷，应答时间以 3～5min 为宜。

1.1.2 问卷分析

在问卷调查过程中，通过问卷得到的调查结果与真实情况之间不可避免地会存在误差，这些误差有可能是调查过程中的登记性误差，也有可能是由于问卷的结构质量不高造成的系统误差。因此，为了提高问卷的结构质量，在完成问卷设计后，还不能马上将其用于市场调查，还需要对问卷，特别是问卷中各量表的信度和效度进行评价。

问卷的信度（Reliability）是指进行重复测量后，一份问卷产生一致性结果的程度。系统误差对信度没有不利影响，因为它以不变的方式影响调查值，重复测量中的真实值不会改变，因而系统误差也不会改变；相反，登记性误差会影响信度，信度可以理解为调查值排除随机误差的程度，如果随机误差为 0，那么调查是完全可信的。通过确定问卷系统误差的比例来评价信度是通过确定用同一问卷得到的调查值之间的相关度来实现的，如果相关度高，则问卷可信，反之则问卷信度较低。对于调查问卷的信度，可以运用 SPSS 中的可靠性分析功能模块求解相关的系数，具体操作过程参见第 12 章“信度分析”的内容。

问卷的效度（Validity）可以定义为对象之间调查值的差异所能反映的对象之间真实值的差异程度。完美的效度要求没有测量误差，即系统误差和随机误差都为 0，SPSS 中有专门的信度分析模块，但没有效度分析模块，如果要进行效度分析，就需要利用相关分析等方法来实现。

1.2 调查抽样

如果只对被调查群体中的一些成员（部分个体）发放问卷，那么应该考虑选择这些成员的方法是否科学，以及所选成员的数量是否足够。因为这些决定了所选的成员能否有效代表他们所隶属的群体，也决定了所得到的结论的有效性。

1.2.1 样本具有代表性的条件

总体是指研究对象的全体，每个研究对象被称为个体，总体由同质的个体组成。抽取总体中部分个体组成集合，称为样本。因此，单从概念上来说，样本不一定能很好地反映其所隶属的总体。

样本对总体要具有较好的代表性，必须满足如下两个必要条件。

一是随机抽样，即从总体中抽取个体的方式不能是人为指定的，必须是随机的。只要能保证总体中每个个体都有相同的机会入选到样本中，就满足了抽样是随机的这项要求，称这种抽样方式为随机抽样。

二是样本量可使研究精度达到要求。样本对总体代表性的好坏除受抽样方式的影响外，还取决于样本量的大小。一般而言，大样本量的样本对总体的代表性优于小样本量的样本。样本量太小，样本便不能很好地代表其所隶属的总体；样本量越大，样本对总体的代表性越好。但这并不等于样本量越大越好，因为样本量越大，在研究中所需投入的人力、物力也会越多。

样本量的大小取决于如下因素。

（1）研究的类型和范围。由于定量问题比定性问题所需的样本量大，所以当问卷调查中涉及的问题既有定性的又有定量的时，应以各项所需的最大样本量作为问卷调查中所需的样本量。

（2）研究精度与允许的误差。研究精度越高，允许的误差越小，问卷调查中需要的样本量越大。

（3）总体容量。通常情况下，问卷调查中的总体容量越大，所需的样本量也越大。

（4）总体的变异性。总体的变异性越小，问卷调查中需要的样本量就越小；反之，样本量就要适当增大。

（5）研究经费。当研究经费宽松时，可适当增大样本量；反之，只需满足最低需求的样本量即可。

总之，为了获取有代表性的样本，在抽样实施之前，应事先了解研究课题涉及的背景资料、被调查者的分布情况、调查的复杂程度等，以此来确定抽样时要不要对总体进行分层、分群；再根据实际研究精度，在抽样阶段合理地选用下面提到的抽样方法，科学地做好抽样设计方案。

1.2.2　抽样方法

在抽样调查中，按抽样是否随机，可以将抽样方法分为随机抽样和非随机抽样两种。

随机抽样能使总体中的每个个体都有相同的机会入选到样本中，在用样本推断总体的过程中，以概率论中的大数法则和中心极限定理为理论依据，可以事先计算和控制抽样误差。因此，在问卷的抽样调查中，应以随机抽样方式来确定被调查者。随机抽样方法包括简单随机抽样、简单系统（等距）抽样、分层随机抽样、整群随机抽样和多阶随机抽样。

非随机抽样是在抽样时不遵从随机性原则，而由调查者根据调查目的和要求，用其主观设立的某个标准从总体中抽取样本的抽样方式。它主要包括方便抽样、判断抽样、配额抽样和滚雪球抽样等。例如，在做图书馆使用情况的满意度调查时，调查者只将问卷发放给在图书馆看书的人员，以此作为样本，这就是典型的方便抽样。非随机抽样抽取的样本有较大的偏差，不能代表总体的特征，因此，抽样设计中要予以避免。

在这些抽样方法中，分层随机抽样、整群随机抽样和多阶随机抽样是大规模复杂局面的抽样调查中用到的抽样方法。而其中最基本的抽样方法是简单随机抽样，其他抽样方法都是在它的基础上，在考虑到总体中各个个体不处在同等地位时，通过调整各个个体的抽样概率发展而来的。

限于篇幅，这里只介绍简单随机抽样和分层随机抽样。

1.2.3　简单随机抽样

当对研究总体的背景资料知之甚少，只有总体中各个个体的名录，没有总体中哪些个体更加重要的辅助信息，而只能将它们一视同仁时，即在总体中仅有一个简单的抽样框时，通常采用简单随机抽样。

抽样框是指对可以选为样本的总体单位列出的名册或序号，用来确定总体的抽样范围和结构。因此，抽样框还称为抽样框架或抽样结构。有了抽样框，便可采用抽签的方式或按照随机数表来抽取必要的单位数。一个好的抽样框应是完整且不重复的。街道派出所里的居民户籍

册、学校里的学生花名册、工商企业名录、意向购房人信息册等都是常见的抽样框。当没有现成的名册作为抽样框时，调查者要自己编制抽样框。即使在利用现有的名单作为抽样框时，调查者也要先对该名单进行核查，避免重复、遗漏，以改善所抽样本对总体的代表性。

简单随机抽样是在总体的 N 个有限个体（抽样单元）中抽取 n 个个体组成一个样本时，使每个个体出现的概率均等于 $1/C_N^n$ 的一种基本的抽样方法。前面提到，它是其他抽样方法的基础，其他抽样方法可以看成对它的修正。它的具体实施非常简单——从总体的 N 个个体中，无放回且机会均等地逐次抽取个体，直至选满 n 个个体。

1．按绝对精度确定样本量

如果问卷问题中涉及定量问题，并有估计的绝对误差要求，则可以使用绝对精度法公式来计算所需的样本量。

由正态分布区间估计理论推得的样本量 n 的计算公式为

$$n=\frac{(\mu_{(1-\frac{\alpha}{2})})^2S^2}{d^2+\frac{1}{N}(\mu_{(1-\frac{\alpha}{2})})^2S^2} \tag{1-1}$$

式中，d 为绝对精度；α 为显著性水平；$\mu_{1-\frac{\alpha}{2}}$ 为 $1-\frac{\alpha}{2}$ 处标准正态分布的位置值；S 为标准差。当总体容量 N 很大时，$n\approx\frac{(\mu_{(1-\frac{\alpha}{2})})^2S^2}{d^2}$。由此可知，当总体容量 N 很大时，样本量 n 与总体容量 N 的关系不大，而取决于总体方差。

【例 1-1】 要抽样调查某社区居民每月每户用于食物的消费支出，现已知该社区有 400 户，共 1500 人，那么在要求平均每月每户用于食物的消费支出的估计绝对误差不超过 30 元的前提下，应调查多少户?

已知条件为 d=30，N=400，但总体 S^2 未知，因此，首先需要获取 S^2 的一个粗略估计值，有以下几种常用方法。

（1）查阅资料法。如果所研究的总体以前被调查过，那么通过查阅有关资料，可以将以前获得的 S^2 的估计值作为其粗略估计值。

（2）预先调查法。在没有现成资料可查的情况下，可先从所要调查的总体中随机抽取一个样本量较小的样本，通过该样本得到方差，用它作为总体方差 S^2 的粗略估计值，再用该值确定所需的样本量。如果这个确定的样本量小于预先抽取的样本量，则预调查就成为正式调查；否则，补充调查不足部分的样本，使之达到所需的样本量。

（3）类推法。如果通过查阅有关资料能找到与所要研究的目标量 Y 高度关联的指标量 X 的信息，$\bar{X}$ 和 S_X^2 有估计值，且 X 与 Y 的变异系数接近，那么只需预先抽取一个样本量较小的样本就可得到 $\bar{Y}$ 的估计值，根据 $S_Y/\bar{Y}\approx S_X/\bar{X}$，便可推算得到 S^2 的粗略估计值。

本题的解题步骤如下。

☞ 第 1 步 预抽样。

在 SPSS 中，用简单随机抽样法，在 400 户中随机抽取 35 户（这个户数是预估数，50 户、30 户或其他较小的值均可），具体做法如下。

（1）在 SPSS 数据编辑窗口中，用数字编码的方式建立如图 1-1 所示的 400 户编号的抽样框数据文件 data1-1.sav。

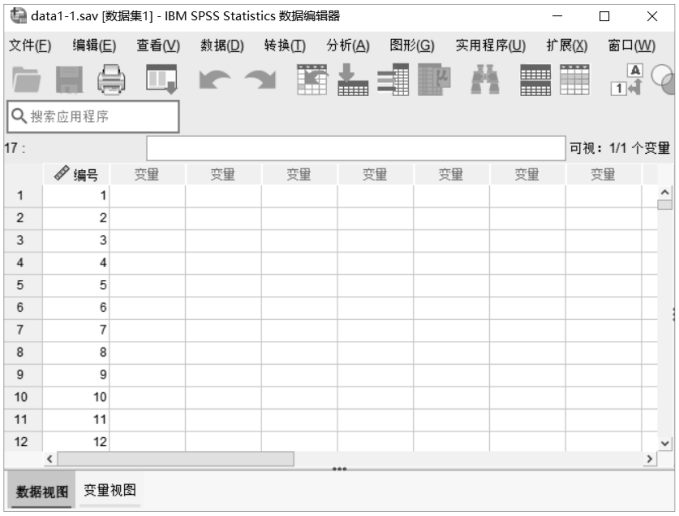

图 1-1　400 户编号的抽样框数据文件

（2）选择“分析”→“复杂采样”→“选择样本”选项，弹出“抽样向导”对话框，如图 1-2 所示。

图 1-2　“抽样向导”对话框

由于还没有抽样方案，所以在 SPSS 的复杂抽样过程中要从头做起，在“抽样向导”对话框中，选中“设计样本”单选按钮，将光标定位在其后的“文件”文本框中，通过直接输入“数据文件夹位置路径\data1-1.csplan”来定义抽样方案文件名；也可以通过单击“浏览”按钮，在弹出的对话框中逐级选择存放路径并定义文件名。

（3）单击“下一步”按钮，进入“阶段 1：设计变量”界面，如图 1-3 所示。

图 1-3 “阶段 1：设计变量”界面

在本例中，要做的是只有一个变量的简单随机抽样，因此，该变量就是抽样单元。在中间的“变量”列表框中选择“编号”变量，单击右侧第二个右移箭头，将其移入“聚类”列表框中，定义“编号”为第一阶群变量，由于后面不再有二阶抽样，所以它也是最终的抽样单元。至此，本阶段设定已经完成。

（4）单击“下一步”按钮，进入“阶段 1：抽样方法”界面，如图 1-4 所示。在此可以定义抽样方法和规模度量。

图 1-4 “阶段 1：抽样方法”界面

在“方法”选区的“类型”下拉列表中，共有 9 种抽样方法，如图 1-5 所示。由于系统默认的抽样方法为简单随机抽样，抽样方式为不放回抽样，故这里可不做任何选择，即保持默认设置。

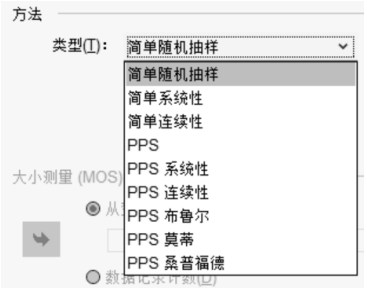

图 1-5　“方法”选区的“类型”下拉列表

（5）单击“下一步”按钮或在左侧导航栏中选择“样本大小”选项，可进入“阶段 1：样本大小”界面，如图 1-6 所示。在这里可以设定样本量。

图 1-6　“阶段 1：样本大小”界面

在“单元数”下拉列表中选择系统默认的“计数”选项。选中“值”单选按钮，并在数值框中输入 35，表示在本阶段选择 35 个样本单元。在做简单随机抽样时，无须考虑此界面中的其他选项。

（6）单击“下一步”按钮，进入“阶段 1：输出变量”界面，如图 1-7 所示。

在该界面中，4 个复选框全部选择，要求在数据文件中存取群体大小（总体容量）、样本大小（样本量）、样本比例和样本权重变量。

（7）单击“下一步”按钮，进入“阶段 1：计划摘要”界面，如图 1-8 所示，在这里可以查看抽样设计。

图 1-7 “阶段 1：输出变量”界面

图 1-8 “阶段 1：计划摘要”界面

单击“完成”按钮，完成简单随机抽样工作。在数据编辑窗口的原数据文件中出现本次抽样结果，如图 1-9 所示，7 个新增的带下画线的变量已存放到数据文件中。

新增变量说明如下。

InclusionProbability_ 1_表示第一阶段体现的抽样概率值，SampleWeightCumulative_ 1_表示第一阶段样本权重累积，PopulationSize_ 1_表示总体容量，SampleSize_ 1_表示样本量，SamplingRate_ 1_表示抽样比例，SampleWeight_ 1_表示样本权重，SampleWeight_ Final_表示最终的样本权重。

编号	InclusionProbability_1_	SampleWeightCumulative_1_	PopulationSize_1_	SampleSize_1_	SamplingRate_1_	SampleWeight_1_	SampleWeight_Final_
1	.	.	.	.	.	.	.
2	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.
4	.09	11.43	400	35	.09	11.43	11.43
5	.	.	.	.	.	.	.
6	.	.	.	.	.	.	.
7	.	.	.	.	.	.	.
8	.	.	.	.	.	.	.
9	.	.	.	.	.	.	.
10	.	.	.	.	.	.	.
11	.	.	.	.	.	.	.
12	.	.	.	.	.	.	.
13	.	.	.	.	.	.	.
14	.	.	.	.	.	.	.
15	.	.	.	.	.	.	.
16	.	.	.	.	.	.	.
17	.	.	.	.	.	.	.
18	.	.	.	.	.	.	.
19	.	.	.	.	.	.	.
20	.	.	.	.	.	.	.
21	.09	11.43	400	35	.09	11.43	11.43

图 1-9　抽样结果

在图 1-9 中，包含这 7 个新增变量值的样品被选入样本，而这 7 个新增变量值为系统缺失值的样品则不在样本之列。将包含抽样框的资料及抽取的 35 户的入选样本的概率、抽样权重、总体容量等相关资料的数据文件另存为 data1-1a.sav。

☞ 第 2 步　预调查。

获取 35 户原始信息，并在 SPSS 中建立相应的数据文件。

向所选择样本中该社区对应编号的居民发放问卷，收集包括户人数、人均月收入和户月食物支出的样本数据资料，以便获取每月每户用于食物的消费支出信息。将获取的资料存放在数据文件 data1-2.sav 中。

☞ 第 3 步　估计总体方差。

在 SPSS 中计算样本方差，用样本方差作为总体方差的粗略估计值。

选择“分析”→“描述统计”→“描述”选项，打开“描述”对话框，如图 1-10 所示，从左侧源变量列表框中选择“户月食物支出”变量，移入“变量”列表框，单击“选项”按钮，弹出如图 1-11 所示的“描述:选项”对话框。选择“均值”“标准差”“方差”作为输出项。

图 1-10　“描述”对话框

图 1-11　“描述:选项”对话框

单击“继续”按钮，返回“描述”对话框。单击“确定”按钮，在输出窗口得到描述统计结果，如表 1-4 所示。由此可得总体方差 S^2 的粗略估计值为 47419.328。

表 1-4 描述统计结果

	N	平均值	标准差	方差
户月食物支出	35	895.7143	217.75979	47419.328
有效个案数（成列）	35			

☞ 第 4 步 计算样本量。

在 SPSS 中计算所需绝对误差限下所需的样本量。

（1）选择“文件”→“新建”→“数据”选项，打开新的数据编辑窗口。建立数据文件 data1-3.sav，其中变量名为总体容量、绝对误差限和方差，其观测值分别为 400、30 和 47419.328。

（2）选择“转换”→“计算变量”选项，打开“计算变量”对话框，如图 1-12 所示。

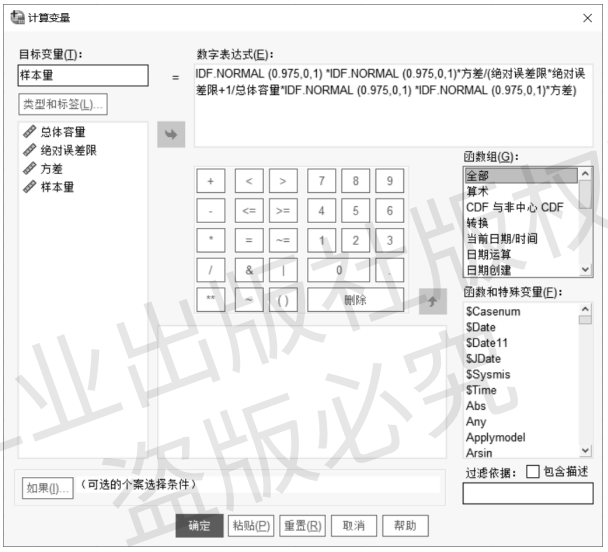

图 1-12 “计算变量”对话框

（3）在“目标变量”文本框中输入“样本量”，在“数字表达式”文本框中输入“IDF.NORMAL (0.975,0,1) *IDF.NORMAL (0.975,0,1)*方差/(绝对误差限*绝对误差限+1/总体容量*IDF.NORMAL (0.975,0,1) *IDF.NORMAL (0.975,0,1)*方差)”，第一个参数 0.975 由 1−0.05/2 计算得到，第二个和第三个参数 0、1 分别为标准正态分布的均值和方差。

（4）单击“确定”按钮，在数据文件中增加一个新变量——样本量及其计算结果值 134.40。

这说明，要满足绝对误差限的要求，至少要抽取 135 户进行调查。这是概算结果，在实际调查中，为确保研究精度，一般应在此基础上增加 10%左右的样本量，即本例应调查 150 户左右。在实际发放问卷的调查中，只需在 35 户的基础上增补调查 115 户即可。

在总体容量很大时，就可以使用公式 $n \approx \dfrac{(\mu_{(1-\frac{\alpha}{2})})^2 S^2}{d^2}$ 来近似估计简单随机抽样需要的样本量。

2．按相对精度确定样本量

当问卷调查的问题中涉及定量问题，并有估计的相对精度要求时，可以使用相对精度法公式来计算所需的样本量。

由正态分布区间估计理论推得的样本量 n 的计算公式为

$$n=\frac{(\mu_{(1-\frac{\alpha}{2})})^2S^2}{(\overline{Y}h)^2+\frac{1}{N}(\mu_{(1-\frac{\alpha}{2})})^2S^2}=\frac{(\mu_{(1-\frac{\alpha}{2})})^2C^2}{h^2+\frac{1}{N}(\mu_{(1-\frac{\alpha}{2})})^2C^2} \tag{1-2}$$

式中，$C=S/\overline{Y}$ 为变异系数。当总体容量 N 很大时，可取

$$n\approx\frac{(\mu_{(1-\frac{\alpha}{2})})^2C^2}{h^2} \tag{1-3}$$

【例 1-2】 前几年的抽样调查结果表明，北京市家庭汽车普及率为 25%～35%，为了解目前北京市家庭汽车普及情况，拟在北京市进行一次家庭问卷抽样调查，要求所估计的相对误差不超过 5%，则在简单随机抽样条件下，至少应抽取一个多大样本量的样本呢？

北京市是大都市，家庭户数达百万级，故可用 $n\approx\frac{(\mu_{(1-\frac{\alpha}{2})})^2C^2}{h^2}$ 来测算所需的样本量。已知估计的相对误差 h 为 5%，故只需知道 C^2 的一个粗略估计值即可计算 n。

假定北京市家庭汽车普及率为 P，则变异系数 $C=\sqrt{(1-P)/P}$。从变异系数的计算公式可见，当 P 从 0 到 1 增大时，1−P 的值减小；而当 P 取最小值时，C 取最大值。由于北京市早已采取了限制汽车牌照发放的办法，市民一般不会主动放弃已有的汽车牌照，所以可以肯定现在北京市的家庭汽车普及率不会低于 25%。故最大的 C^2=(1−0.25) /0.25 =0.75/0.25 = 3。

根据以上分析计算样本量，在 SPSS 中的操作步骤如下。

（1）选择“文件”→“新建”→“数据”选项，打开新的数据编辑窗口，建立数据文件 data1-4.sav，其中变量名为相对误差，其观测值为 0.05。

（2）选择“转换”→“计算变量”选项，打开“计算变量”对话框。

（3）在“目标变量”文本框中输入“样本量”，在“数字表达式”文本框中输入“IDF.NORMAL (0.975,0,1) *IDF.NORMAL (0.975,0,1)*0.75/0.25/(相对误差*相对误差)”。

（4）单击“确定”按钮，在数据编辑窗口的数据文件中增加一个新变量——样本量及其计算结果值 4609.75。

这说明，至少要调查 4610 户才可满足相对误差不超过 5%的要求。

1.2.4　分层随机抽样

分层随机抽样是指将总体分成若干不重叠的小总体，称每个小总体为一个层。分层随机抽样是指在不重叠的小总体或层中挑选独立样本。例如，层可以是种族、年龄组、工作类别等。使用分层随机抽样方法可以保证重要子群的样本量，改进全部估计的精度，并且在层与层之间可以使用不同的抽样方法。

1．分层样本量的确定

（1）当无法给定总样本量时，使用研究费用的最佳分配法来确定分层样本量。

分层随机抽样的费用一般可用以下公式来计算：

$$C=C_0+\sum_{i=1}^{k}n_iC_i \tag{1-4}$$

式中，C_0 为问卷调查中的基本费用；C_i（i=1,2,⋯,k）为第 i 层中调查一个样本单元所需的费用；k 为分层的层数。理论上可以证明，在分层随机抽样中，固定费用 C 使 $V(\overline{y}_{st})$（方差）最小，或者使 $V(\overline{y}_{st})$ 为固定值，而使费用 C 最小的各层样本量为

$$n_i \propto \frac{W_i S_i}{\sqrt{C_i}}（i=1,2,\cdots,k）\tag{1-5}$$

式中，$W_i=\frac{N_i}{N}$ 为各层单元数占总体容量的百分比，N 为总体容量，N_i 为第 i 层的单元数；S_i 为第 i 层样本的标准差。而此时的总样本量为

$$n=\frac{C-C_0}{\sum_{i=1}^{k}\sqrt{C_i}W_iS_i}\sum_{i=1}^{k}\frac{1}{\sqrt{C_i}}W_iS_i\tag{1-6}$$

总样本量可以由研究费用决定，但费用和精度是一对矛盾体，减少费用就会减小样本量，从而牺牲研究精度；要提高研究精度，就得多调查一些研究单元，增加一些研究费用。

（2）在给定总样本量 n 的前提下，分层样本量的确定方法。

在给定总样本量 n 的条件下，常用的分层随机抽样各层样本量的分配方法有以下 3 种。

① 等额样本：在每层都抽取相等的样本量，各层的样本量均为 $n_i=\frac{n}{k}$。此时的分层随机抽样称为不成比例（不等概率）分层随机抽样。

② 按比例分配：样本量按总体中各层的单元数量所占的比例来分配，$n_i=n\frac{N_i}{N}$。当各层的单元数量 N_{i} 已知，而其他信息很少时，通常采用这种分配方法。此时的分层随机抽样称为等比例（等概率）分层随机抽样。

③ 奈曼（Neyman）最优分配：在分层随机抽样中，$n=\sum_{i=1}^{k}n_i$ 固定，使 $V(\bar{y}_{\mathrm{st}})=\sum_{i=1}^{k}W_i^2\left(\frac{1}{n_i}-\frac{1}{N}\right)S_i^2$ 达到最小的样本量分配方法为

$$n_i=n\frac{W_iS_i}{\sum_{j=1}^{k}W_iS_i}=n\frac{N_iS_i}{\sum_{j=1}^{k}N_iS_i}（i=1,2,\cdots,k）\tag{1-7}$$

因此，在给定总样本量时，建议采用等概率分层随机抽样。

【例 1-3】 要了解某地区 904 家出口企业的经营情况，由于各出口企业的生产规模相差较大，现要求分层随机抽取 200 家进行调查，那么各层应如何分配抽样单元数量呢？可以参考的以往各出口企业的经营情况如表 1-5 所示。

表 1-5 以往各出口企业的经营情况

类别	以往出口额（万美元）	企业数量	平均出口额（万美元）	方差
1	2600～5000	40	3503.40	356554.26
2	1400～2600	74	1895.78	119535.60
3	700～1400	115	1036.38	34141.69
4	250～700	226	419.31	16565.56
5	50～250	449	123.26	3110.05

各层的抽样单元数量取决于选用的分配方法，这里只介绍按比例分配法，因为这种分配方法常常可以获得精度很高的估计结果。

首先在 SPSS 数据编辑窗口中将表 1-5 中的数据建成 SPSS 数据文件（data1-5.sav），并使其处于打开状态，然后按以下步骤进行操作。

① 选择“转换”→“计算变量”选项，打开“计算变量”对话框。

② 在“目标变量”文本框中输入“按比例样本量”，在“数字表达式”文本框中输入“rnd(200*企业数/904)”。rnd( )函数为四舍五入取整函数。

③ 单击“确定”按钮，在数据编辑窗口的数据文件中增加了一个新变量——按比例样本量，从上到下各层的值分别为 9、16、25、50、99，总数为 199，这里的总数比 200 小的原因是进行了四舍五入取整进位操作。

2．在 SPSS 中实现分层随机抽样

接上面各分层样本量的计算结果，下面介绍在 SPSS 分层随机抽样过程中实现按比例分配的具体做法。

（1）在 SPSS 数据编辑窗口中先建立“出口金额分类”变量和用来对应 904 家出口企业编号的“编号”变量的抽样框数据文件（data1-6.sav）。注意：分层变量必须定义值标签，否则会在步骤（5）中出错。

（2）选择“分析”→“复杂采样”→“选择样本”选项，打开“抽样向导”对话框。在“抽样向导”对话框中，选中“设计样本”单选按钮，在“文件”文本框中输入“数据文件夹位置路径\data1-6.csplan”，定义抽样方案文件名。

（3）单击“下一步”按钮，进入“阶段 1：设计变量”界面。在“变量”列表框中选择“出口金额分类”变量，将其移入“分层依据”列表框中，定义“出口金额分类”为第一阶层变量；选择“编号”变量并将其移入“聚类”列表框中，定义“编号”为第一阶群变量。

（4）单击“下一步”按钮，进入“阶段 1：抽样方法”界面。在此界面中，不进行任何选择，采用系统默认的简单随机抽样方法。

（5）单击“下一步”按钮，打开“阶段 1：样本大小”界面，定义样本量。在“单元数”下拉列表中选择系统默认的“计数”选项，定义抽样单位为计数值，而非比例值。选中“各层的不等值”单选按钮，表示各层选择不同的样品数量，单击“定义”按钮，展开“定义不等大小”对话框，由上到下分别单击各层后面的单元格，分别输入 9、16、25、50、99，完成对各层所要抽取的样品量的录入，如图 1-13 所示。

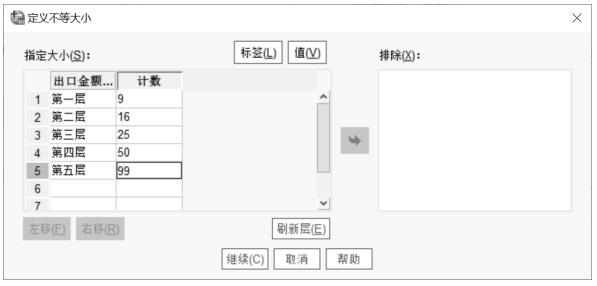

图 1-13　“定义不等大小”对话框

（6）单击“继续”按钮，返回“阶段 1：样本大小”界面。

（7）单击“完成”按钮，在数据编辑窗口的数据文件中出现如图 1-14 所示的分层随机抽样结果，新增变量中有数值的行所指的单元被抽中。

出口金额分类	编号	InclusionProbability_1_	SampleWeightCumulative_1_	SampleWeight_Final_
1	1	.	.	.
1	2	.	.	.
1	3	.	.	.
1	4	.	.	.
1	5	.	.	.
1	6	.23	4.44	4.44
1	7	.	.	.
1	8	.23	4.44	4.44
1	9	.	.	.
1	10	.	.	.
1	11	.	.	.
1	12	.	.	.
1	13	.	.	.
1	14	.	.	.
1	15	.	.	.
1	16	.	.	.
1	17	.	.	.
1	18	.23	4.44	4.44
1	19	.	.	.
1	20	.	.	.
1	21	.23	4.44	4.44
1	22	.23	4.44	4.44

图 1-14　分层随机抽样结果

在输出窗口中，得到第一阶段摘要，如表 1-6 所示。其中，第一大列所列为各层的名称，第二大列所列为各层抽样单元的数量（要求抽取的数量和实际抽取数量相同，分别为 9、16、25、50 和 99），第三大列所列为各层抽样单元的比例，各层比例基本围绕 22%上下波动。

表 1-6　第一阶段摘要

		抽样单元数		抽样单元比例	
		请求	实际	请求	实际
出口金额分类=	1	9	9	22.5%	22.5%
	2	16	16	21.6%	21.6%
	3	25	25	21.7%	21.7%
	4	50	50	22.1%	22.1%
	5	99	99	22.0%	22.0%

计划文件：数据文件夹位置路径

在数据编辑窗口中选择“文件”→“另存为”选项，弹出“将数据保存为”对话框，在此对话框中，可以选择将抽样结果数据文件保存下来。注意：文件名不能与现有的文件相同。

总之，在一般的问卷调查中，为确定随机抽样的总样本量，需要对有精度要求的每个具体问题测算满足精度要求的样本量，实际调查的样本量要大于或等于测算样本量中的最大值。

1.3　思考与练习

1．设计一份合格的问卷需要注意哪些问题？

2．试设计一份关于大学本科生手机需求情况的问卷。要求格式正确，题目中要包含开放型问题和封闭型问题。

3．要对总体具有较好的代表性，样本应满足哪些条件？

4．随机抽样方法主要有哪几种？

5．简述简单随机抽样中样本量的确定方法。

6．简述在给定总样本量的条件下确定分层样本量的方法。

7．简述在 SPSS 中如何实现分层随机抽样。