中央广播电视大学教材

Social Statistics

社会统计学

◎ 陈卫 编著

中央广播电视大学出版社

Social Statistics

社会统计学

中央广播电视大学教材

Social Statistics

社会统计学

◎陈卫 编著

中央广播电视大学出版社

原书缺页

原书缺页

原书缺页

原书缺页

原书缺页

到社会研究中数据分析是关键的一步，数据的分析是以统计学为工具的，社会研究离不开统计学，统计学是社会研究的重要工具。

第二节　数据收集与分析

中国人口究竟有多少？有多少生活在农村，又有多少农村人口流动到城市打工？这些问题都是关系社会发展的重要问题，可是，如何得到这些问题的答案？又如，人们如何看待国家的计划生育政策？妇女的生育意愿如何？这些问题的答案又是如何得到的？为了回答这些问题，我们就需要收集相关的数据。我们在第一节中提到，数据的收集是社会研究的关键步骤，数据的收集是社会调查的目的。在这一节中，我们主要讲解几种社会调查方法，对以上问题加以讨论。

一、变量类型

在统计中，我们将说明事物某种特征的概念称为变量（variable），如性别、受教育程度、智商、年龄等都是变量。变量的具体取值称为变量值。例如，性别可以是男、女；年龄可以是1岁、1.5岁、2岁等。根据不同的分类标准，变量可以分为不同的类型。

（一）按测度水平分类

不同情况下，对变量的描述也不同：对于人的性别的计量，我们用男、女来描述；对人们受教育程度的计量，我们用文盲、小学、初中、高中、大专等来描述；而对于人们的智商、年龄等的计量，我们用0，1，2等数字来表示。由于测度水平不同，变量可以分为定类变量、定序变量、定距变量和定比变量四种。

1. 定类变量

当变量值的含义仅表示个体的不同类别，而不能说明个体的大小、程度等其他特征时，这种变量称为定类变量（nominal variable）。定类变量是最低层次的变量。在社会调查中，定类变量有很多，如性别、婚姻状态、民族、地区、职业等。性别可以分为男、女两类，民族可以分为汉族、回族、壮族等。

定类变量只能区分个体之间的类别差异，而且各类别之间是平等的并列关系。其变量值只能用来表示类别，但为了便于统计处理和利用计算机进行统计分析，我们通常用数字来表示。但是数字1，2等只表示类别，1并不代表更高或更低的级别。例如，用“1”表示男性人口，“2”表示女性人口；用“1”表示未婚人口，“2”表示初婚人口，“3”表示再婚人口，“4”表示离婚人口，“5”表示丧偶人口等。当然，我们也可以用“1”表示女性人口，“2”表示男性人口。定类变量的变量值，只能用于比较“=”或“≠”，而不能用于其他运算。

定类变量中，对个体进行分类要符合穷尽和互斥的要求。类别穷尽是指每个个体必须能够归属于一个类别，不能有遗漏；类别互斥是指每个个体只能归属于一个类别，不能重复。例如，一个人，要么是男性人口，要么是女性人口，总会有他（她）的归属，不可能同时为男性和女性。

定类变量中，变量可以有两种或更多种分类。如性别只有男、女两类；民族可以分为汉族、少数民族两类，也可以分为汉族、回族、壮族等56类。当变量只分为两个类别时，我们称之为**二分变量（binary or dichotomous variable）**。

2. 定序变量

当变量值的含义不仅表示个体的不同类别，还可以区分个体之间大小、程度等序次差异时，这种变量称为定序变量（ordinal variable）。在社会调查中，受教育程度是定序变量，可以分为文盲、小学、初中、高中、大专及以上等；人们对某种制度的态度可以分为非常同意、同意、中立、不同意、非常不同意等。

同样，为了便于统计分析和计算机运算，定序变量也用数字表示不同类别。例如，用“1”表示文盲，“2”表示小学，“3”表示初中，“4”表示高中，“5”表示大专及以上。此外，数字1，2等不仅仅表示文化程度的分类，还表示文化程度的高低，1代表最低的文化程度，2表示的文化程度要高于1，5表示最高的文化程度。

与定类变量相比，定序变量更精确，它所包含的信息量大于定类变量，除可以进行“=”或“≠”比较外，还可以进行“>”或“<”运算。

3. 定距变量

当变量值不仅可以将个体区分为不同类别并进行排序，而且可以确定不同类别之间的数量差别和间隔差距时，这样的变量称为定距变量（interval variable）。

定距变量具有测量单位，这些测量单位具有不变的相等区间的标准，使个体之间的比较更具客观性。例如，智商、温度等都是定距变量。

定距变量的变量值是用数值表示的，我们可以准确地计算个体之间的差值。例如，学生甲的智商为120，学生乙的智商为110，则甲的智商比乙的智商高10。又如，地区甲的温度是0℃，地区乙的温度是8℃，则地区甲的温度比地区乙的温度低8℃。在定距变量中，0的选取只是为了方便或习惯，0只表示一个数值，而不表示该现象不存在。如地区乙的温度为0℃，并不是说地区乙没有温度；一个人的智商为0，并不是说这个人没有智商。

与定序变量相比，定距变量更精确，除可以进行“=”或“≠”比较、“>”或“<”运算之外，还可以进行“+”或“-”运算。

4. 定比变量

当变量除了具有上述三种变量的全部特征外，还可以计算两个变量值之间的比值时，这样的变量称为定比变量（ratio variable）。定比变量是最高层次的变量。在社会调查中，很多情况下我们使用的都是定比变量，如身高、年龄、收入、一个地区的人口数、某产品的生产量等。

定比变量的变量值也是用数值表示的，但是与定距变量相比，两者的唯一区别是定比变量有绝对零点，即定比变量中的“0”是有实际意义的数值。例如，一个人的身高是0米，则表示这个人不存在；一个人的收入是0元，则表示这个人没有收入。同样，由于定比变量中有绝对零点，除可以进行“=”或“≠”比较、“>”或“<”运算外，它还可以进行“+”，“-”，“×”，“÷”运算。例如，职工甲每月收入是7 500元，职工乙每月收入是2 500元，我们可以说职工甲比职工乙每月多收入5 000元，也可以说职工甲每月收入是职工乙每月收入的3倍。

上述四种变量的层次是由低级到高级、由粗略到精细逐步递进的。高层次的变量具有低层次变量的特征，但是反过来不成立。显然，我们可以将高层次变量转化为低层次变量，但转化过程中会丧失一些信息。例如，在定比变量中，受教育程度的变量值为0年、1年、2年等，我们可以将不同阶段受教育程度转化为文盲、小学、初中、高中、大专及以上。我们可以清楚地看到，转化后只能进行序次比较，而不能进行加减乘除运算。

在统计中，定类变量和定序变量说明事物的品质特征，通常用文字表述；定距变量和定比变量则说明事物的数量特征，通常用数值表示。因而，**我们经常将定类变量和定序变量称为定性变量（qualitative variable），而将定距变量**

和定比变量称为定量变量（quantitative variable）。

（二）其他分类

采用不同的分类标准，变量的分类就不同。上文中，根据测度水平不同我们把变量分为定类变量、定序变量、定距变量和定比变量四类；根据变量值是否连续，我们可以将变量分为离散变量和连续变量两类；根据变量之间的相互关系，我们可以将变量分为自变量和因变量。

1. 离散变量和连续变量

在社会调查中，我们发现，有些变量值是不连续的，如某妇女生育的子女数；有些变量值是连续的，如年龄。根据变量值是否连续，我们将变量分为离散变量（discrete variable）和连续变量（continuous variable）。

如果一个变量的变量值是间断的，可以一一列举的，这种变量称为离散变量。例如，某人兄弟姐妹数、结婚次数、工厂生产产品的数量等，其变量值的取值是0，1，2，3等。离散变量的取值是有限的，而且其取值都是以整数位断开的，是有最小计量单位的。例如，某人的兄弟姐妹数，只能是1个、2个、3个等，而不能是1.3个、2.7个等。

如果一个变量的变量值是连续不断的，即可以取无限多个数值，这种变量称为连续变量。例如，年龄、温度、灯泡的寿命等，它们的取值是连续不断的。连续变量可以取无限多个值，其取值是连续不断的，不可以一一列举的，而且，它们没有最小计量单位。例如，年龄可以是1岁整，也可以是1.2岁、1.45岁、2.544岁等。

2. 自变量和因变量

在社会调查中，我们发现，有的变量发生在前，有的变量发生在后，有的变量的变化必须以其他变量的变化为前提。例如，收入增加在前，生育率减少在后。根据变量之间的相互关系，我们将变量分为自变量（independent variable）和因变量（dependent variable）。

我们将引起其他变量变化的变量称为自变量，而将由于其他变量的变化而导致自身发生变化的变量称为因变量。自变量与因变量之间的关系不仅仅是先后关系，如我们常说的“小树长高，我也长高”。我们不能说“我长高”是因变量，“小树长高”是自变量，两者只是有时间上的先后关系，而没有必然联系。自变量与因变量之间还必须存在以下关系：因变量的变化以自变量的变化为前提。例如，受教育程度提高，收入增加，这两者就是因果关系，受教育程

度是自变量，收入是因变量。

在社会调查中，我们经常将性别、年龄、民族、居住地等属性变量作为自变量，而将生育水平、生育态度等行为或者态度变量作为因变量。同时，我们还会遇到这样的情况：有些变量在某种关系中是自变量，而在另一种关系中则是因变量。例如，我们经常提到“收入增加，生育率降低”、“受教育程度提高，收入增加”，在前一关系中收入是自变量，而在后一关系中收入则是因变量。

二、数据收集

（一）相关概念

在介绍调查时，我们先介绍几个相关概念：

总体（population）是构成它的所有个体的集合，个体（element）则是构成总体的最基本的单位。例如，我们对某学校学生的就业观进行观察，那么该学校的所有学生就是研究总体，而其中每一个学生就是个体。

样本（sample）就是从总体中按照一定方式抽取的一部分个体的集合。例如，我们要对某省360万育龄妇女进行调查，按一定方式抽取1 000人进行调查，那么这1 000名育龄妇女就是该总体的一个样本。当然，从总体中我们可以抽出若干不同的样本。

抽样单位（sampling unit）就是一次直接的抽样所使用的基本单位。抽样单位有时与构成总体的个体是相同的，有时是不同的。例如上面提到的对育龄妇女的调查，当我们直接抽取育龄妇女时，两者是相同的；当我们从总体中一次直接抽取户，以抽中的户中的育龄妇女作为样本时，抽样单位（户）与个体（育龄妇女）就不相同了。

抽样框（sampling frame）是指一次直接抽样时样本中所有抽样单位的名单。例如，从某校中抽取200名学生进行就业观的调查，那么这所学校的所有学生的名单就是这次抽样的抽样框。但是，当我们先抽取班级，以抽中班级中的所有学生作为样本时，这所学校所有班级的名单就是这次抽样的抽样框。

抽样框的形式有很多种，只要是包括所有总体单位的名单即可，如学生名单、住户门牌号等。抽样框不仅可以提供备选单位的名单，而且是计算各个单位入选样本概率的依据。

（二）调查类型

1. 普查

普查（census）是一种专门的调查，它是为了某种特定的目的而对总体中所有的个体进行的一次全面调查。例如，我们定期举行的人口普查、工业普查、农业普查、第三产业普查、经济普查、统计基本单位普查等都是普查。

作为一项特殊的调查，普查的优点是：（1）由于普查是对所有个体进行的调查，因而，普查具有信息全面、完整的特点，可以为其他抽样调查提供依据。（2）普查通常是一次性或周期性的。例如，我国目前在每逢末尾数字为“0”的年份进行人口普查，每逢“3”的年份进行第三产业普查，每逢“5”的年份进行工业普查，每逢“7”的年份进行农业普查，这都是周期性的。（3）为了保证调查数据的准确性，避免重复或遗漏，普查一般都有统一的标准调查时点。例如，我国2010年第六次人口普查的标准时点为11月1日0时。

普查的缺点有：（1）由于普查涉及范围广，调查单位多，因而，普查比较耗时、费力，成本比较高。（2）普查的适用范围比较窄，调查内容不深入，只适合调查一些基本的、一般的社会现象。

当然，尽管我们平时说到普查就会想到全国范围的人口普查等，但并不是说必须在全国范围内进行的调查才可以称为普查。就范围而言，只要是对所要研究的总体中所有个体都进行调查即可称为普查。例如，在一个省内、一个工厂内、一条街道上等都可以进行普查。

2. 抽样调查

前文提到，普查是一项耗时、费力的调查，而且适用范围较窄，在实际中应用也相对较少。而抽样调查则是实际中应用最广泛的一种调查方式。**抽样调查（sampling survey）是从总体中选取部分个体组成样本进行调查的一种方式，其目的在于根据样本的调查结果推断总体特征。**

与普查相比，抽样调查具有以下特征：（1）经济性。抽样调查涉及的调查单位较少，工作量较少，比较省时、省力，节约成本。（2）时效性强。抽样调查可以迅速、及时地获得所需信息，也可以较频繁地进行。（3）适用范围广。抽样调查可以用于各个领域、各种问题的调查，同时，涉及的调查问题可以更深入。（4）准确性较高。抽样调查的工作量较小，涉及人员少，工作中误差较小，准确性较高。

根据抽取样本的方法不同，抽样调查可以分为概率抽样和非概率抽样。

概率抽样（probability sampling），就是按照随机原则进行的抽样，总体中每个个体都有一定的、非零的概率入选样本，并且入选样本的概率都是已知的或可以计算的。当总体中每个个体入选样本的概率相等时，概率抽样称为等概率抽样；当入选概率不相等时，概率抽样称为不等概率抽样。

概率抽样主要的方式有以下几种：

简单随机抽样（sample random sampling），就是从包括总体 N 个单位的抽样框中随机地、一个一个地抽取 n（$n<N$）个单位作为样本，每个单位入选样本的概率是相等的。

简单随机抽样常用的方法类似于抓阄，就是将每个个体都编号并将号码写在小纸条上，然后放入一个盒子中，搅拌均匀后从中任意抽取，直到抽够样本数目为止。例如，从某专业 200 名学生中抽取 50 人进行调查。我们先找到这 200 名学生的名单，并将 200 个学生从 1 到 200 编号，用 200 张小纸条分别写上 001，002，003，…，200，然后将这些小纸条放入一个空盒子中，搅拌后，随意抽出 50 张小纸条，然后按照号码找到对应的学生即可。

对于总体数量很大的情形，我们一般采用随机数表来抽样。随机数表中的数码和排列都是随机形成的，没有任何的规律性，每一个数字都有 0～9 位，随机产生，这些数字可以随意地组成两位数、三位数等。例如，我们从包含 1 000个个体的总体中抽取 50 个个体作为样本。首先将个体从 1 到 1 000 编号，然后在随机数表中任意选取四行或四列组成一个四位数字，从这个数开始向下取数，当数字没有超过 1 000 时就选择，超过时就放弃，一直到选出 50 个为止。

简单随机抽样是最基本的概率抽样，也是其他抽样方法的基础。它操作简单，而且保证每个个体入选样本的概率是一样的。但是，简单随机抽样必须有一个完整的抽样框，因而，当样本量很大时，往往给实际操作带来很大的麻烦。

系统抽样（systematic sampling），也称为等距抽样，就是将总体中所有单位按照某一标志排序后，在规定的范围内随机抽取一个单位作为初始单位，然后按照一定的相等距离抽取调查单位。

系统抽样的具体方法是：首先，将每一个个体按照顺序排列并编号，然后计算出抽样间距，即 k（抽样间距）$=\dfrac{N\text{（总体规模）}}{n\text{（样本规模）}}$；其次，在最前面的 k 个

个体中，采用简单随机抽样的方法抽取一个个体并记下其编号（假设为 A），称为随机起点；最后，从 A 开始，每隔 k 个个体就抽取一个个体，这样抽中的个体的编号分别为：A，$A+k$，$A+2k$，…，$A+(n-1)k$。这些个体就构成了样本。

系统抽样的操作比简单随机抽样更为简单，但是同样需要抽样框。此外，系统抽样中有一个排序的问题，我们需要注意两种情况：一是个体的排列具有某种序次上的先后、高低之分；二是个体的排序有与抽样间隔相对应的周期性分布的情况。

分层抽样（stratified sampling），也称类型抽样，就是先将总体中的所有单位按某种特征或标志（如年龄、性别、职业等）划分成若干类型或层次，然后再在各个类型或层次中采用简单随机抽样或系统抽样的方法抽取一个子样本，最后将这些子样本合起来构成总体样本。

在分层时，我们通常采用以下原则：（1）以所要研究的主要变量或相关变量作为分层的标准。（2）以那些已有明显层次区分的变量作为分层变量。例如，我们以性别、年龄、受教育程度等作为分层标准。（3）以保证各层内部同质性强、各层之间异质性强、突出总体内在结构的变量作为分层变量。例如，在大学中，我们可以将学生根据所学专业分层，或者根据来源地分层。

分层抽样的两个重要的优点是：（1）在不增加样本规模的情况下降低抽样误差，提高抽样的精度。因为分层的目的就是把异质性较强的总体分成一个个同质性较强的子总体，从各子总体中抽取样本，从而保证总样本中包含各种类型的个体。（2）便于了解总体中不同层次的情况，以及对总体中不同的层次进行单独研究，或者进行比较。

整群抽样（cluster sampling），就是先将总体按照某种标志或特征划分为一些子群体，然后从总体中随机抽取一些子群体，再将这些抽出的若干小群体内的所有元素构成总体样本。

整群抽样不仅可以简化抽样的过程，同时可以降低收集资料的费用，能相应地扩大抽样范围。但是，它具有样本分布不广、样本对总体代表性相对较弱的缺点。

多阶段抽样（multi-stage sampling），也称多级抽样或分段抽样，就是根据抽样元素的隶属关系或层次关系，将抽样过程分为几个阶段进行。其具体做法是先在总体中抽取若干大群，然后从抽中的群中再抽取几个小群，这样一层层抽下来，一直抽到最基本元素为止。

多阶段抽样适用于总体范围特别广、对象层次特别多的情况。我们在确定每一阶段的抽样单位数目时，不仅要考虑各阶段子总体的同质性程度，还要考虑所拥有的人力和经费情况。当同质性较强时，我们可以使所抽的规模相对小一点；反之，则应大一点。当经费和人力允许时，在其他条件不变的情况下，样本所覆盖的范围越广，样本的代表性也越高。

前面我们已经提到概率抽样是按照随机原则进行的，实际中，我们也会**根据主观意愿、实际情况等进行抽样，而不依据随机原则进行抽样，这些不符合概率抽样要求的抽样都称为非概率抽样（non-probability sampling）**。非概率抽样中样本的代表性较低、误差也较大，而且个体入选样本的概率是未知的，不能估计误差，因而其结果不能用于推断总体情况。但是非概率抽样实施方便、简单，因而在实际中应用比较广。

非概率抽样主要有以下几种：

判断抽样（judgmental sampling），就是研究者根据自己的研究目的和主观判断从总体中选择有代表性的个体。例如，研究者根据判断选择一些有经验的计划生育工作人员调查有关计划生育工作的实行情况。

判断调查可以充分发挥研究人员的主观判断能力，但是对研究者的判断能力、实际经验等要求较高。当研究者对总体情况很了解时，样本的代表性就较高，但是它仍属于非概率抽样，其结果仍不能用于推断总体。

偶遇抽样（accidental sampling），也称方便抽样，就是根据实际情况以自己方便的方式选择样本，或者选择那些离得最近的、最容易找到的作为样本。例如，研究者在商场门口拦截行人进行有关某种商品的调查；在影院门口拦截观众进行有关电影方面的调查。

偶遇抽样实施很是方便，尤其对于没有经验的研究者来说更是实用。偶遇抽样尽管排除了主观因素，但是仍不是随机抽样。因为它只是从最容易看到的、最方便找到的人员中选取样本，而不是从所有个体中选择样本。所以研究者还是不能将结果用于推断总体状况。

滚雪球抽样（snowball sampling），就是先从总体中少数人员入手调查，并向他们询问其他符合条件的人，再去调查这些人并继续询问其他人，像滚雪球一样。该方法主要用于人数较少或者比较难找到的总体抽样。例如，对同性恋的调查，可以采用滚雪球的方法进行抽样，先找到几个人员进行调查，然后通过他们提供的名单再去寻找其他同性恋者，这样通过滚雪球的方法来找到符合样本的数量。