

抽样调查基本知识



查瑞传 著



中国统计出版社

抽样调查基本知识

查瑞传 著

中国统计出版社

抽样调查基本知识

查瑞传 著

中国统计出版社出版发行

中国建筑工业出版社印刷厂排版 河北省新城印刷厂印刷

787×1092毫米 32开本 5.75印张 120,000字

1983年7月第1版 1983年7月新城第1次印刷

印数：1—35,000

统一书号：4006·019 定价：0.70元

前 言

抽样调查是人们在实践中长期以来一直使用的一种调查方法。人们经常不自觉地运用它，而不知自己是在做抽样调查，也讲不清这样做的道理。有一段时期，在我们的统计工作中贬低抽样调查，给以种种不正确的指责。而另外一些时期，当人们提倡抽样调查时，又常常夸大了它的作用，把它神秘化。抽样调查究竟有没有用？为什么有用？道理在哪里？好处在哪里？有什么缺点和局限？它的误差由何而来？能否测定？能否避免？怎样缩小？只有弄清这些问题，才能清醒地而不是盲目地、科学地而不是迷信地充分运用抽样调查方法。

本书就是为了给广大统计工作者和学习统计学的人提供一本通俗读物，使读者能通过实例正确理解和掌握抽样调查的基本道理。抽样调查的中心问题是根据部分单位调查结果推论全体。这种推论之所以可能，是由于部分与全体之间存在着某种必然联系。而这种联系是一种“可能性”的联系。抽样调查自始至终包含着一个“可能性”的问题。本书的主要目的之一就是使读者对此有一个正确的理解。

出于这种考虑，全书分成四大部分。第一部分讲现象的实际分布，即频率分布。这是对总体的反映，是对抽样调查进行分析的基础。第二部分讲或然现象的分布，即概率分布。这里阐明频率分布与概率分布之异同，并介绍几种主要分布律。第三部分通过简单实例说明总体和样本的关系，即

抽样调查的原理。这一部分所讲的内容，扼要地涉及到抽样调查的所有基本问题。第四部分介绍几种基本的具体抽样方式，说明它们的特点及适用条件。

全书由浅入深，主要依靠逻辑推理，尽量避免单纯公式推导。重点在于使读者准确地掌握抽样调查的概念和道理，而不在于提供若干可以套用的计算公式。每部分之后附有若干习题，以便使所学原理得到进一步消化和掌握，并使学者体验原理的应用。

本书原是为培养统计干部而编写的教材，曾在中国人民大学用于统计干部进修班的教学。编写过程中，作者与陈余年教授共同承担数理统计学和抽样调查课程，在所有问题上都得到陈教授的全面指导和帮助，特此志谢。

目 录

一、现象的分布.....	(1)
(一) 总体和标识.....	(1)
1. 总体.....	(1)
2. 标识.....	(2)
(二) 单位在总体内的分布.....	(4)
1. 分布的概念.....	(4)
2. 分配数列——分布的一种表现形式.....	(6)
3. 分布的图示.....	(10)
(三) 表现分布情况的两类综合指标.....	(23)
1. 表示分布中心的指标.....	(23)
2. 表示全部单位离散程度的指标.....	(29)
二、对样本的观察与可能性问题.....	(38)
(一) 样本的概念.....	(38)
(二) 可能性的度量——概率.....	(39)
1. 概率的概念.....	(39)
2. 概率的测定.....	(43)
(三) 概率的基本运算规则.....	(46)
基本事件和复合事件.....	(46)
(四) 随机变数及其分布律.....	(52)
1. 随机变数的分布.....	(52)
2. 表现概率分布的综合指标.....	(55)
3. 几个随机变数之和的平均值与方差.....	(58)

(五) 几种重要的概率分布.....	(65)
1. 二项式分布.....	(65)
2. 超几何分布.....	(70)
3. 正态分布.....	(75)
三、总体与样本的关系.....	(84)
(一) 抽样调查的概念.....	(84)
1. 非全面调查的作用.....	(84)
2. 抽样调查的概念与特点.....	(86)
(二) 总体和样本.....	(89)
1. 一个简例.....	(90)
2. 从12块田地抽出 2 块的结果.....	(94)
3. 每个单位被抽入样本的概率.....	(97)
4. 样本平均数与总体平均值.....	(93)
5. 样本平均数的代表性与样本单位数的关系.....	(99)
(三) 抽样误差	(104)
1. 抽样误差的概念	(104)
2. 抽样误差的计算	(107)
3. 抽样误差与样本的代表性	(115)
4. 总体平均值所在范围的估计	(122)
5. 必要抽样数目的确定	(126)
6. 影响抽样误差大小的因素	(129)
四、几种基本抽样方式	(133)
(一) 抽样调查的几种基本组织方式	(134)
1. 简单随机抽样 (纯随机抽样)	(134)
2. 分类随机抽样 (分层抽样、类型抽样)	(139)
3. 等距抽样 (机械抽样)	(145)
4. 整群抽样	(153)

(二) 多级抽样(多阶段抽样)	(158)
五、抽样调查中的估计推算	(162)
(一) 直接估计法	(162)
1. 总体平均数 $\bar{X}$ 的估计	(162)
2. 总体标识总值 X 的估计	(165)
3. 总体成数(比重) P 的估计	(166)
4. 总体中某类单位数目 N_i 的估计	(167)
5. 总体单位总数 N 的估计	(168)
(二) 比例估计法	(168)
1. 用比例法估计总体标识总值 X	(168)
2. 修正系数法	(170)

一、现象的分布

(一) 总体和标识

1. 总体

从统计上研究社会现象时，首先要观察这个现象所包括的许许多多个别对象，登记它们的特征，并根据观察结果，对于全部对象加以综合说明。我们把所要研究的全部对象叫做“总体”，把组成这个总体的个别对象叫做“总体单位”。在统计研究中，最终要说明的是总体，但研究过程要从个别单位开始。

例如，研究一个城市的职工家庭生活水平时，每户职工家庭是一个单位，全市的全部职工家庭便是总体。研究一个造纸厂纸的质量时，一天所生产的全部纸张是调查对象的总体，每一令纸则是一个单位。

进行统计研究，必须首先确定总体范围，这包括两个方面：一方面要确定所研究的对象包括什么内容，也就是说，所统计的那一事物要有明确的定义。例如，统计工业总产值时，首先要确定哪些生产属于工业。这方面的问题由统计学及有关的社会科学解决。另一方面要确定总体的空间范围，例如统计工业总产值时，还要确定是哪个地区范围以内的。这个问题根据实际需要而定。

确定了总体，便确定了统计的对象。在总体之内还要确定总体单位，以便进行观察。总体单位大致可以分为两类。一类是各单位之间可以清楚地分得开的。例如，居民、机床、

家畜、生产队、工厂、家庭等。其中有些是天然分得清的，如人、机床、家畜；有些是需要人为地明确规定界限的，如工厂、家庭。另一类总体的单位是连在一起、没有天然界限的。例如，土地（耕地）和时间。这类总体由于没有划分单位的天然界限，因而可以采取任意大小的一块或一段作为一个单位。例如，研究粮食产量时，可以用一亩土地作为一个单位，也可以用一平方公尺土地作为一个单位；研究纺织机器停车次数时，可以用一小时作为一个单位，也可以用一分钟作为一个单位。当然，还可以用更小的土地面积或更短的时间间隔作为单位。这也就是说，这类总体的单位是可以任意划分的。同一个总体，由于划分的粗细不同，总体单位的数目便不一样，每个单位的数值也随之而不同。例如，假定一个生产队有147亩田地，若以一亩地为一个总体单位，则有147个单位，其单位产量是每亩地的产量；若以一分地为一个总体单位，则有1470个单位，其单位产量则是每分地的产量了。因此，进行调查时，总体单位的大小必须事先明确规定。由于这类总体单位没有明确的天然界限，即使规定了单位的大小，一个单位的划分究竟从哪里开始到哪里截止，仍然有许多种可能。把一片田地划分成一亩一亩的小地块，可以这样划分，也可以那样划分。用不同办法划分出来的一亩亩田地，产量便随之而不同。处理这类总体时，应该特别注意这些问题。

2. 标识

对总体单位进行调查时，除计点单位的数目外，就是观察各个单位的某些特征的数值。例如，观察各个炼钢厂的钢产量，各个生产队的粮食产量，各个工人的工种和工资收入，各个新学生的性别、年龄和身高，各趟公共汽车的乘客人数以

及在某站上下车的人数等等。每个单位的特征叫做总体单位的标识。标识有两类：一类是品质标识，表示每个单位的属性，如工人的工种、学生的性别等。另一类是数量标识，如产量、工资收入、年龄、身高、乘客人数等。

统计工作就是收集、整理和分析各种标识的资料，通过标识的数量表现，说明总体的各种问题。以下专就数量标识进行考察。

一个标识的具体数值，就总体中各个单位来说并不完全相同。实际统计工作中观察的正是各个单位的这些不同的数字，即标识数值。例如，调查某工厂263个工人的家庭人口数，得到数字如下（表1-1）：

某工厂全体263个工人家庭人口数目登记 表 1-1

3	5	6	4	7	5	3	2	6	3	5	4	2	5	6	5	3	5	4	3
8	4	1	3	4	3	6	5	3	3	6	3	9	2	4	3	0	3	7	1
4	6	2	5	2	10	4	4	7	6	4	6	5	4	2	6	5	6	3	8
2	3	4	7	5	1	2	6	5	4	1	3	3	7	5	5	3	4	6	4
1	5	9	5	3	2	5	3	2	2	8	1	4	5	6	4	6	3	2	4
6	3	3	4	6	4	7	5	4	5	4	5	6	3	3	8	3	9	4	6
4	1	5	6	5	4	5	6	2	7	3	10	4	3	5	5	5	2	6	5
7	6	7	3	4	6	2	4	9	5	6	5	4	6	1	2	7	4	1	3
2	5	3	5	2	7	4	5	4	4	4	2	7	1	3	6	5	3	4	2
5	2	6	2	1	5	6	4	1	2	7	5	3	3	5	3	1	7	3	6
7	5	5	5	8	3	1	7	6	3	5	6	5	3	7	5	4	4	4	7
3	4	2	6	4	3	5	3	5	1	2	3	1	2	4	1	6	4	8	4
4	6	3	1	3	4	3	8	3	6	2	7	6	4	2	4	4	1	1	3
2	4	3																	

全部263个工人家庭是我们研究的总体，每个工人家庭是一个单位，每一家庭的人数则是一个具体的标识值。

又如，把某一造纸厂某日生产的250令纸的重量登记下

来，得到如下数字（表1-2）：

某造纸厂生产的250令纸的重量(单位：公斤) 表 1-2

14.75	14.55	14.95	15.10	14.70	15.15	14.65	14.50	15.00	14.55
14.20	14.70	14.85	14.55	14.90	14.80	15.00	14.85	14.95	14.70
14.85	15.15	15.30	14.70	15.10	14.75	14.55	14.90	14.50	15.15
14.95	14.50	14.75	14.90	14.75	14.50	14.90	14.25	14.85	15.15
14.90	14.25	14.90	15.40	14.50	15.35	14.75	14.70	14.95	14.90
14.55	14.60	14.50	14.45	14.85	14.70	14.50	14.55	15.55	14.75
15.00	14.75	14.95	14.60	15.30	14.55	14.85	15.00	14.70	14.50
14.45	14.90	14.30	14.75	14.90	14.90	14.70	14.45	14.95	14.95
14.90	15.50	14.90	15.15	14.70	14.30	15.35	14.90	15.15	14.90
15.30	14.45	14.50	14.70	14.45	14.45	14.60	14.95	14.40	15.35
14.65	14.70	14.75	14.95	15.50	14.95	14.90	15.30	14.55	14.80
14.70	14.85	14.95	15.35	14.75	14.90	14.80	15.15	14.70	14.90
15.15	14.95	14.40	14.50	14.90	15.15	15.15	14.95	14.65	14.55
14.50	15.00	15.20	14.85	14.65	14.40	14.50	14.75	15.30	14.30
14.85	14.65	14.45	14.90	15.00	14.85	15.55	15.05	15.15	14.65
14.70	14.90	14.70	14.55	14.75	14.90	14.70	14.80	14.60	14.70
15.25	14.95	15.15	14.70	14.90	15.25	14.75	14.55	14.85	14.95
14.50	14.70	14.90	14.35	14.95	14.70	14.90	14.70	15.10	14.90
14.75	15.35	14.65	14.90	15.15	14.65	14.50	15.35	14.45	15.15
14.90	14.50	15.00	14.75	14.70	14.50	14.35	14.90	14.70	15.20
15.15	14.75	14.85	14.90	14.85	14.75	11.90	14.65	14.95	14.85
14.50	14.95	14.90	15.15	14.35	14.90	14.95	15.10	14.50	14.50
14.95	15.10	14.55	14.65	14.60	14.95	15.15	14.75	14.95	14.75
14.90	14.55	14.70	14.55	14.55	15.10	14.70	14.50	14.35	14.90
15.10	14.90	15.15	15.05	14.95	14.75	14.85	14.95	14.80	15.00

两张表中所列的是对总体各单位观察所得的原始数据。
以后的整理和分析都从这些数字开始。

(二) 单位在总体内的分布

1. 分布的概念

在统计工作中，除了统计总体单位数目（如某地的人口

数)或标识总量(如全国钢厂的钢产量)外,常常需要知道具有不同标识数值的各类单位在总体中各有多少,以及各占多大比重。例如,一个县的农业领导部门不仅要知道全县总的粮食生产水平,而且要知道各公社和大队的粮食生产水平,以及生产水平不同的公社和大队各占多大比重。一个造纸厂不仅要掌握一天生产的某种纸的数量和平均每令纸的重量,而且要考察同一类纸中厚薄不同的纸所占的比重,从而判断产品质量是否均匀。标识数值不同的各类单位在总体中所占的比例,称为总体的分布。在经济管理和统计工作中,必须掌握和研究现象的分布。通过现象的分布,往往能反映出规律性的东西。

以上两例讲的是一段时期内或一定时点上某一现象内部的分布状况。此外,还可研究一个现象在时间上的分布状况。例如,观察某百货公司某柜台每分钟的顾客人数,然后,加以整理,列出在柜台前同时出现多少顾客的情况各有多少分钟。又如,观察某区间某路公共汽车在一天内的各段不同时间各有多少乘客,找出乘客人数的增减规律。这是两种不同情形。前者将每段时间的顾客人数,作为各段时间单位的标识数值,而不管各段时间的先后顺序,因此这里的每一分钟仅被看作一个时间单位,其性质同上述的一个社队、一令纸等是一样的。后者不但要考察每段时间内的乘客人数,还要根据各时间单位的排列次序来考察乘客增减变化的趋势,因此也称为动态分布。

统计工作不仅要掌握现象的数量表现,而且要透过数量掌握现象的质量和规律性,而总体分布则是现象规律性的重要表现形式之一。例如,上述公共汽车乘客人数在一天内的分布,一天内生产出的同类纸张在每令重量上的分布,就反映

了乘客人数、纸张质量的规律性。抽样调查的理论和实践与现象的分布关系密切，掌握抽样调查的原理，必须从讨论现象的分布入手。

2. 分配数列——分布的一种表现形式

表1-1与表1-2中所列数字是杂乱无章的，很难从中直接得出结论。为使现象的分布清楚地显现出来，需把总体单位按标识数值归并成若干组，求得各组包含的单位数。根据现象的不同情况，可有两种不同处理方法。

(1) 当标识的数值不多而且简单整齐时，可以把原始调查数据按标识的数值分开，数一数每种标识数值的单位数，并按标识数值大小排列出来。例如，把表1-1中的某工厂全体工人家庭按家庭人口数分组排列起来得到如下的分组表(表1-3)：

某工厂全体工人家庭按人口数目分组表 表 1-3

家 庭 人 数 家 口	家 庭 数	频 率 (%)	家 庭 人 数 家 口	家 庭 数	频 率 (%)
1	20	7.6	7	19	7.2
2	29	11.0	8	8	3.0
3	48	18.3	9	4	1.5
4	50	19.0	10	3	1.2
5	46	17.5	总 计	263	100.0
6	36	13.7			

(2) 当标识的数值很多，无法一一列出时，最好先把标识数值分成若干段，把属于每一段的单位合并在一起(可以把每个单位的标识值写在卡片或纸条上，按每一组的上下限归组)，并按标识数值大小排列出来。例如，把表1-2中的纸重分成组距为0.2公斤的7个组，写出每一组的令数，如

表1-4所示：

某工厂生产的250令纸按重量分组

表 1-4

每令纸重 (公斤)	令数	频 率 (%)	密度	每令纸重 (公斤)	令数	频 率 (%)	密度
14.2~14.4	10	4	0.2	15.0~15.2	35	14	0.7
14.4~14.6	45	18	0.9	15.2~15.4	15	6	0.3
14.6~14.8	60	24	1.2	15.4~15.6	5	2	0.1
14.8~15.0	80	32	1.6	合 计	250	100	—

注：相邻两组共有的纸重数字归入数值较大的一组。

表1-3和表1-4中都是把总体中的全部单位按其标识数值大小分组排列，并加总出每一组的单位数目，这样得出的数列（如各种人数不同的家庭数，各种重量不同的纸的令数），称为分配数列。分配数列是表现总体单位分布情况的一种形式。在分配数列中，分组所依据的标识数值叫做标识值或变量。每个组的单位数目叫做次数或频数。通过分配数列便可比较清楚地看出总体中全部单位按所考察的标识数值分配的情况，有助于进一步了解和掌握现象的规律性。例如，从表1-3不仅可以知道某厂工人家庭人口数从1人到10人不等，而且可以准确地知道人数不同的每种家庭各有多少。如每户4个人的工人家庭共有50户，是最多的。次数虽能说明每一组的单位有多少个，但还不能清楚表明各个组在全部单位中所占比重，不便于在不同总体之间进行对比。例如，在另一个工厂511户工人家庭当中，一家4口的有95户。与前一工厂相比是多还是少，就很难看清。为了更鲜明地表现出现象的分布，最好用总体单位总数除各组的次数，计算各组

在总体中所占百分比，称为“频率”。如表1-3中第一组的频率为7.6%，第二组的频率为11.0%等等。第二个工厂里一家4口的工人家庭虽然有95户，比第一个工厂多得多，但其频率只有18.6% ($95/511=0.186$)，低于第一个工厂。次数和频率各有自己的用途。次数可以表明各个组与总体的规模，并且也在一定程度上反映了总体的分布，但不很清楚。频率可以清晰地表明总体的分布，但却不能反映总体和总体的规模。统计工作应按实际需要，采用次数或频率来说明总体分布，也可象表1-3和表1-4那样将两者同时列出。

从表1-3和表1-4可以看到，把总体单位分组时有两种情形。一种是把标识的每个数值列出，作为一个组，例如工人家庭按家中人数分组。当标识的数值是离散的，一个一个分得清楚，且数值不多时，便可采取这种形式，称为单值数列。另一种情形是把一段段的标识数值列出，作为各个组的上下限，例如表1-4中250令纸按重量分组和表1-5中某市350个商店按售货员人数分组。当标识的数值呈连续变化时（如每令纸的重量），便需要采取这种形式，称为组距数列。有时，标识数值虽然可以一一分开，但数值较多，不宜一一列出时，也可把若干标识数值合并起来，写成组距数列形式，如表1-5；

某市350个商店按售货员人数分组表 表 1-5

售货员人数	商店数	售货员人数	商店数
1~2	25	11~20	118
3~5	45	21~30	49
6~10	113	合 计	350

在组距数列中，每个组内的最高和最低标识值之间的间隔称为“组距”。当标识值为连续变动的数量时，各个组的组距等于组中标识的最大值（上限）减最小值（下限）。如表1-4中纸的重量，第一组的组距为 $14.4-14.2=0.2$ 公斤，第二组的组距为 $14.6-14.4=0.2$ 公斤，等等。当标识值本身是离散的（即不连续的）而写成组距数列形式时，则组距为各个组内所包括的标识值的个数。例如，表5中商店售货员人数，第一组包括1人和2人两个数值，组距为2；第二组包括3人、4人和5人等3个数值，组距为3；第三组组距为5人，第四组组距为10人，等等。

根据标识数值分组时，首先要确定组距。选择组距的基本要求是能分出类别，看清总体的分布，能够说明问题。各组的组距可以是相等的（如表1-4），也可以是不等的（如表1-5）。分组的原理是统计学阐述的内容，在这里我们只从研究现象分布的角度来考虑。为了清楚地显示现象的分布情况，组距最好相等，同时，组距大小也要适宜。例如，表1-4中的250令纸按0.2公斤的组距分组，分布情况看得比较清楚。如以0.4公斤、0.5公斤、0.6公斤为组距，便会得到不同的分配数列，如表1-6所示：

250令纸按重量分布情况

表 1-6

每令重量 (公斤)	令 数	每令重量 (公斤)	令 数	每令重量 (公斤)	令 数
14.2~14.6	55	14.2~14.7	80	14.2~14.8	115
14.6~15.0	140	14.7~15.2	150	14.8~15.4	130
15.0~15.4	50	15.2~15.7	20	15.4~16.0	5
15.4~15.8	5				
合 计	250	合 计	250	合 计	250