第八章　资料的整理与简化

资料的整理是调查研究活动中工作量极大的一个环节。在结束了资料收集阶段后，面对大量的资料，首先要解决的就是如何整理这些资料的问题。调查资料量多、无规范且很难查找，只有通过整理、加工使它们条理化、系统化、规范化，才会便于检索、查阅，才有可能把调查研究转向分析阶段。这一阶段之所以对调查研究工作具有重要意义，就在于通过整理、加工后的资料，能够保证调查结论具有确切的根据和可靠性。在这一阶段发现的一些新问题和工作失误，也可以得到及时补救和解决。所以，必须掌握一定的整理资料的技术，保证整个研究的科学性，以获得调查应有的效益。

第一节　资料的整理

整理资料的基本要求是：真实、具体、简明、扼要。真实是宗旨，即在整理资料过程中不应渗及个人的意向，不能从个人的主观愿望出发来取舍资料；具体，就是要防止把资料整理得空洞抽象，要有事实、有数量依据，带结论性的概括应在分析资料阶段进行，应尽可能地把收集到的事实整理出来；简明扼要，即要求整理资料时应留取事实，去掉空洞无物的内容，要抓住主要特点、主要方面，不拖泥带水。

整理资料工作的内容很多，按照调查资料的不同类型，有不同的整理方式。

社会调查资料主要有四种：（1）问卷调查资料；（2）有关统计资料；（3）文献调查资料；（4）运用观察、访谈等手段收集的各种初步处理了的记录资料。这些资料因其收集的方式不同，资料在其形态、内容、特征上也有所区别，在使用时也有不同的价值。

一般说来，问卷资料通过初级的编码和登录后主要借助电子计算机进行整理分析；统计资料则较灵活，可以人工处理，也可计算机处理；文献资料及观察、访谈的记录资料，就只能靠人工处理。

整理资料的工作主要有：

一、资料的检查和校订

这是指对收集到的所有原始资料按照完整性、一致性和可靠性的要求，进行全面的检查、鉴别、校订。它是提高资料的有效性，确保研究结论科学性的一个必要步骤。

完整性的检查就是要检查事先确定的调查对象、研究项目、问卷内容等是否完备无缺。如发现有缺访者或项目有遗漏的，应尽可能重访或用其他方式补足。对于因调查对象出于各种原因不接受调查或未回答的空缺资料，应注明原因和情况，以免影响对总体的推断。

一致性检查是指检查资料在以下三个方面的一致：第一，计量标准是否一致，如计算人均收入时，有的把人均年收入误作为月收入等。第二，填答、记录的方法和方式是否一致。第三，补充、修改答案时所采用方式是否一致。以上三个方面的差错在调查中是很容易出现的。如访问调查因访问员在理解上的差异产生不一致；观察过程因用于记录的符号太多而产生不一致；问卷调查因被询问者对问题的理解不同也会产生不一致。这些不一致的资料会使资料之间失去可比性，因而也就无法保证调查的真实

性。对于以上三方面，凡未按事先规定的要求填写的，都应进行纠正。

可靠性检查是指对资料的真伪的鉴别以及资料的准确程度的检查。可靠性检查通常有几种方法：

（1）计算检查：即通过计算的方法来检查数据中各项指标及计算结果的准确性。

（2）核对检查：依据可靠的或权威性的相关信息资料来对照、比较调查资料，以便发现和纠正调查资料中的某些差错或失真。

（3）抽查：即对调查资料所反映的情况再次到实地调查几个子样，以检查资料的可靠性。

（4）逻辑检查：对资料进行逻辑的分析，找出差错。如资料中出现的一些逻辑矛盾：实际年龄20岁，而工龄已经10年等。可以参照其参加工作的时间加以纠正。如无法改正，可改为“不知道”。

在校订资料过程中，还应注意那些容易偏差的项目。如家庭收入、妇女人工流产指标，往往低于事实，而日常生活支出、家长对子女的评价等，又往往高于实际。除了有必要抽取部分样本再次调查核实以估计偏差范围的大小外，一般在分析时不应作准确度较高的推论。

此外，对一些未填写主要项目，或错误太多的资料，以及无法校订的资料，应作为无效卷处理。有些可以作必要的技术上的处理。如对两者择一的问题，如果调查对象同时选择了两个答案，或填写方式不合要求，可以用丢硬币的方式进行随机处理选取一种答案。

二、摘要

作摘要卡片，是处理定性资料的方法之一。在收集资料时，

有用的资料越多越丰富越好。但在整理时，就需要进行一番筛选，把重要的、主要的资料保留下来。作摘要就是有系统地记录那些内容丰富、生动具体的，具有内在联系的原始资料。它尤其适用于整理通过观察法、文献法、座谈会等方法获得的各种记录资料。摘要必须注意抓住反映现象本质的资料，也就是说要抓住问题的要点及特点来进行整理、摘录。马克思、列宁在研究资本主义社会时都收集了大量的文献资料，但并不全部一一照抄，而是通过思考摘取其中最能表现社会现象特点的资料。摘要的内容除了围绕主题外，还要按照研究计划，系统地摘录具有内在联系的事实。

做摘要往往是使用卡片或活页纸，要注明调查时间、对象、内容等具体项目，以便查找、分类及核实。

在整理、摘录资料时，应按研究的需要，把资料进行分类。可以按专题、调查对象、调查时间、资料性质等不同标准来分门别类。关于分类标准问题，将在下一节进行介绍。

三、资料的编码与登录

对资料进行编码、登录，是整理定量资料的方法，主要用于整理问卷的资料。经过编码和登录后的资料，便于进行统计分析。在社会调查中因大多采用了数量较多的样本，所以，对于收集的大量的定量资料，如采用手工处理的方法来整理，不仅工作量大和速度慢，而且难以进行完整的统计分析，同时又极易发生人为的错误。如果要进行复查和纠正，则更加困难。因此在有条件的地方，应尽量使用电子计算机，这是社会调查的必然发展趋势。对资料进行编码和登录，不仅是使用电子计算机计算的必要步骤，也可以运用于手工计算。资料通过编码和登录以后，可以大大提高工作效率和计算的准确性。

1．编码

它是指把问题的答案分成若干有意义的类别，使原来详细的信息变成几个较简单的类别，然后赋以一定的符号或码值，以便对这些资料进行描述和分析。编码的目的就是为了方便描述和分析。

编码可以分为先编码和后编码两种。

先编码是计量的一个组成部分，属于赋予数字或数目的阶段，它是在问卷尚未正式使用之前，就先把问题的有关答案编好码值。这种编码办法只适用于问卷中的封闭性问题。这类问题的答案都是预先给定的，所以可以对它们进行编码、赋值。而开放性问题，因预先不知道有多少种答案，就无法给予码值。

后编码是在问卷收回后进行的。它适用于整理开放性的问题，研究者根据回答的情况把答案进行分类然后赋予一定的码值。对于通过观察、文献、访问等方法得到的记录资料，只要分类得当，也同样可以使用后编码的方法进行整理。

先编码与后编码比较，其优点是：

第一，可以节省大量的劳动，被调查者在回答问题时，就已经选出了代表自己答案的码值，所以整理资料时可以直接进行登录。

第二，节省了编码簿。编码簿（也称译码簿）是对编码的说明和索引，用来说明每个符号或数值代表的资料的意义以及标明各项资料在“调查资料汇总表”上的位置。使用先编码就可以利用问卷作为现成的编码簿。研究者可以直接从问卷上了解到某一问题的码值，而不必要再编制有关的说明。

后编码则较麻烦。它需要花费较长的时间和较多的经费，而且容易出现人为的差错。如需把大量定性资料进行分类，再赋码值，还要编制编码簿等。但它有一个优点。即对事先无法估计答案数量的问题，可以待资料收集后再编码，避免了先编码所可能

造成的某些遗漏。

2．调查资料汇总表与资料的登录

调查资料汇总表，是用来汇集经过数量化后的资料，它是我们进行统计分析的依据。使用这张表可以囊括问卷上所有的码值，它比把问卷资料直接输入计算机要方便，并便于核查。

我们在设计问卷时，无论是先编码还是后编码，都应注意两个问题：

第一，在决定每一答案的码值时，必须确定它在“汇总表”上的位置，这个位置通常用“栏号”来表示。一个问题在“汇总表”上可以占有两个以上的栏，也可以只有一栏，这需要根据答案的数量来定，例如：

45．您认为自己现在的工作是

1．非常理想_____　　栏号

2．较理想　_____　　077_____

3．一般　　_____

4．较不理想_____

5．很不理想_____

如果调查对象认为自己的工作是“较理想”的，则他在这一问题上的码值就是2．在整理资料时，必须在“栏号”：077后的空白处填上“2”。077则代表了第45个问题在“汇总表”上所处的位置（即栏数）。登录的工作就是把问卷上的这些码值登记、转录到“汇总表”上。我们便可以从该表的第77栏中找到每一个调查对象对第45个问题的所有答案了。

第二，对所有问题可能出现的答案，都应当给予不同的码值。如果答案过多，则可以只列出几个主要的答案，然后用“其他”来包括剩余的答案。如：

（1）您对当前的改革最关心的问题是：　　027＿＿＿＿

1．物价问题＿＿＿＿＿＿＿＿　　028＿＿＿＿

2．党政机关为政清廉的问题＿＿＿＿＿＿

3．住房改革问题＿＿＿＿＿＿

4．其他问题（请具体说明）＿＿＿＿＿＿＿

在栏数的安排上要充分照顾到可能出现的答案，既要尽可能节省栏数，以免“汇总表”过于庞大，又必须留足栏数。例如上表：实际是先编码后与后编码的结合点。因为人们最关心的问题可能全集中在第一、二个答案，但又考虑到各种复杂因素，把第4个答案作为开放式的答案，这样就无法预计可能出现的答案数量，为了保险起见。这个问题可以留2个栏数，这样就足以对99个答案进行编码了。

第二节　简化资料的基本统计技术

在整理资料阶段，对于收集到的大量的统计资料以及各种标准化询问表中的数据，运用统计的方法加以分门别类的整理，使原来分散的、无系统的资料，变成集中的、系统的资料，使原来只说明个别事物的资料，变成说明整个总体的资料，这一过程，我们称之为统计整理。它是整理、简化定量资料的重要方法。

例如我们进行生活方式的调查，获得大量调查表，每张表上都记载着每个调查对象的收入、开支、消费结构、闲暇时间利用等各种资料。经过统计整理加工，可以从各种统计表中获得整个调查对象在以上各方面的基本状况的说明。列宁曾指出；“全部问题在于如何整理这些出色的按户调查资料，如果没有全面的、编制得合理的分组和复合表，极丰富的按户调查简直毫无用处，

这就是现代统计工作的最大危险。”[①]所以，统计资料的整理和简化，对整个统计分析具有十分重大的意义。

统计整理的最主要内容，就是通过统计分组，把资料简化成各种图、表的形式，从而对社会现象的规模、构成等作出初步的说明。本节将介绍有关的统计分组方法，统计图表的制作等简化资料的基本技术。

一、统计分组法

1．统计分组的意义和分类标准

统计分组法实际上就是分类法，它是根据调查的任务和社会现象的性质，按照一定的标志将被调查的事物划分为不同的组或类。统计分组是统计整理中的一个重要环节，也是进行统计分析的基础。我们在前几章的介绍中，已多次涉及到分类问题，如随机抽样时对调查对象的分类，对不同类型现象的不同计量层次，对定性记录资料的分类等。统计分组同以上各种分类都具有共同的作用。首先，它们都可以区分社会现象的不同类型。如对我国现有多种经济成份的所有制形式的分类：国营、集体、个体、各种形式的横向联合企业股份制，以及“三资”企业等的划分。利用分组的方法区分不同的社会类型，就可以深入认识每一类型的特殊本质和类型之间的关系。其次，这些分类有助于研究事物的内部结构，从而认识事物的内在联系和发展趋势。就统计分组来讲，还可以研究事物之间在数量上的依存关系。

进行分类工作，最关键的在于选择适当的分类标志，即对事物进行分类时所依据的标准。它是统计分组的首要问题。同一个资料，因分类方法不同，会得出完全相反的结论来。正确地选择分类标志，就能得出正确的结论。否则，只能导致结论的错误。

---

① 《列宁全集》，人民出版社，第20卷，第71页。

如何选择分类（组）标志呢？

第一，要根据事物内部矛盾的分析，找出反映事物本质的标志。如研究人口状况，就要按性别、年龄、文化程度等进行分类。研究社会保险，就要考虑年龄、工龄等标准。研究生活方式，就要按收入、年龄、婚姻状况等标志来分类。

第二，根据研究的不同目的来选择划分标志。对于任何事物，都可以从不同的角度进行调查研究。研究目的不同，分类的方法和标志也不相同，社会学的调查研究要从社会学的角度来选择有社会学理论意义的标志。

第三，结合研究对象所处的具体历史条件或社会发展条件来选择标志。反映现象本质的重要标志具有条件性、地区性和历史性。标志要依时间、地点、条件为转移，这样的分组才有现实意义。例如，在经济体制改革以前，我们主要是以生产资料所有制公有化的程度来划分经济类型为全民所有制经济和集体所有制经济。改革以后，多种经济成份的企业层出不穷，如“三资”企业、各种横向联合企业、私营经济、股份制等，再以公有化程度来划分标准显然就很不适应了。

可见，选择分组标志不仅仅是一个简单的技术问题，而更重要的是理论问题和政策问题。

2．统计分组的主要形式

1）单项式分组。即在进行分组时只采用一个分组标志进行，如把调查对象按性别分类，或按年龄分类。在这种分组基础上可形成分组表。

2）复合式分组。即在分类时同时使用两个或两个以上的标志。例如，把职工按职业进行分类，然后再按性别的标志把每一种职业的职工分为男、女两类。在这种分组基础上形成复合表或交叉表。

按照分组标志，把样本单位加以分配，就构成分配数列，在

此基础上可以形成次数分配表。经过不同计量层次测量并量化了的社会现象均可以作次数的分配。

二、次数分配

1．什么是次数分配

也称次数分布，它是按照某种标志把总体（或样本）的单位加以分配。以显示总体（或样本）单位在某一类别中出现的次数（或频率），次数通常用符号 f 来表示。次数分配以及在此基础上形成的次数分配表，是整理、简化资料的一个重要内容。例如，我们抽550个样本进行文化程度调查，这样我们可以得到550个数据。为了使资料的进一步分析成为可能，就必须把这些零散的数据进行组织、整理，通过次数分配把各种变量的值分成若干组，看原始资料在每组中出现的次数(见表8．1)，文化程度划分为五级（文盲、小学、初中、高中、大专以上）这是定序变量的次数分配表。定类、定距变量的次数分配表（见表8．2、表8．3）。

表8．1 定序变量的次数分配表

文化程度 （X）	次数 （f）
1	20
2	30
3	335
4	120
5	45
合 计	550

表8．2　定类变量的次数分配表

性　　别	次数（f）
男	250
女	300
合　　计	550

表8．3　定距变量的次数分配表

年龄分组（岁）	次　　数
21—30	305
31—40	135
41—50	90
51—60	20
合　　计	550

2．组距、组中值

对于定距变量的统计分组一般都有一定的组距。组距，即每一组区间的距离；在组距两端的数值称为组限，每组的起点数值称为下限，其终点数值称为上限。所以，组距就是上限与下限之差：组距＝上限－下限。

上限和下限之间的中点数值称为组中值，通常的计算方式是：

$$\text{组中值} = \frac{\text{上限} + \text{下限}}{2}$$

在表8—3中，第一组年龄的下限为21岁，上限为30岁，组距 = 30 − 21 = 9（岁）。其组中值 $= \frac{30+21}{2} = 25.5$（岁）。

3．累积次数

定序以上变量的资料，除了可以使用上述次数分布外，还可以使用累积次数，即将各组的次数逐级相加。累积次数分为向上累积（cf↑，即向上限累积次数）和向下累积（cf↓，即向下限累积次数）两种。其作用在于可以使我们了解到在某一数值以下（或以上）的次数（表8·4）。

表8·4　累积次数表

文化程度(X)	次数（f）	cf↑	cf↓
1	20	20	550
2	30	50	530
3	335	385	500
4	120	505	165
5	45	550	45
合　计	550	—	—

同理，累积也可以对成数p和百分数进行计算，可分别用符号 cp，c%来表示。

三、比例法与对比法

1．比例法

进行次数分配可以帮助我们简化资料，但不能比较资料。可以通过比例法来解决这一问题。比例法是用于表示各类成分在总体上所占的比重大小，用以比较构成比重的情况，也叫成数，用符号p表示。计算公式为：

$$P = \frac{\text{次数}}{\text{总数}} = \frac{f}{n}$$

例：甲、乙两单位中知识分子的成数为：

$$P_{甲} = \frac{152}{555} = 0.274 \qquad P_{乙} = \frac{135}{480} = 0.281$$

如果只看两单位知识分子的绝对数，乙＜甲，由于基数不同，所以不能进行比较，而使用了比例法后，则可以比较，实际上乙单位的知识分子所占比重要比甲单位多。

使用比例法通常因所得数字太小而不易给人以明确的概念，可以通过比率法把数字变为百分率、千分率、万分率。百分率的计算方法为：$p \times 100\%$。

用比率法时，要注意根据不同的精确度要求保留小数点后一定位数的数字。一般小数点后保留两位数。现代统计学中，在决定小数点后的最后一个数值时一般不采用“四舍五入”的方法，而使用“前单5入”，即前面是单数时就进位，否则舍去（零作双数）。①

① 参见李沛良《社会研究中的统计分析》，湖北人民出版社，第35页。

2．对比法

即将两类次数进行比较，如a与b两类资料之比为$\frac{n_a}{n_b}$

例：某校有教师409人，学生10447人。则该校师生之比为409∶10447＝1∶25.5。即每一个教师要教25.5个学生。最常见的对比有性比例，一般以女性为100。

四、次数分布的图示法

次数分布不仅可以用统计表的形式来表现，而且还可以用适当的图形加以表示。统计图是将抽象的数字，用点、线、面、体等几何图形、实物形象、地图以及各种色彩等绘制的图形来表现。它具有直观、形象、生动的特点。

1．制图的基本规则

第一，图形必须根据研究目的准确显示资料；

第二，图形应力求简单、形象生动。一般不使用面积图和立体图。要在准确反映资料的基础上富于艺术性。

第三，图形的排列顺序一般为自左到右、从上到下，图形的高、宽比例以3∶4为宜；

第四，制图时所依据的坐标轴、基准点和线，应慎重设计，需要时应注明单位和尺度；

第五，图形应标有完整、明确的标题，图形所依据的资料必须列出。

2．统计图

用图示法制作的统计图多种多样，如条形图、圆瓣图、直方图、折线图、曲线图、面积图等等。在社会调查的整理资料阶段，最常采用的是条形图、圆瓣图、折线图和曲线图。

1）条形图。它是以相同宽度的条形的长短或高低来表示资

料的次数或百分率。如对500个学生父亲职业的调查结果，用条形图显示出来（图8·1）。

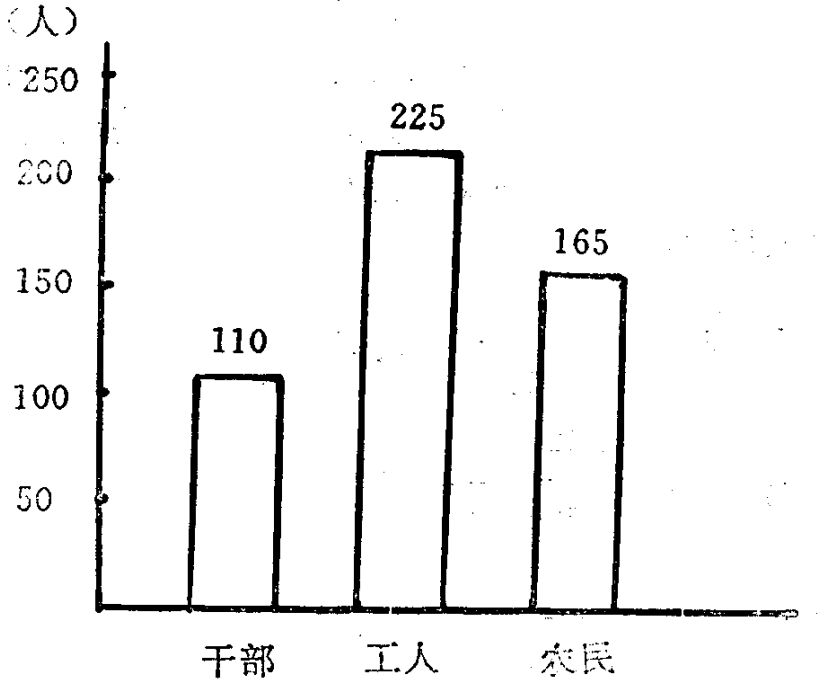

图8·1　条形图

制图时，横轴表示分组的标志。纵轴表示次数或比率，按数据标出实际的高度。对于连续变量的组距分组。在条形之间不留间隔，构成直方图（见图8·2）。

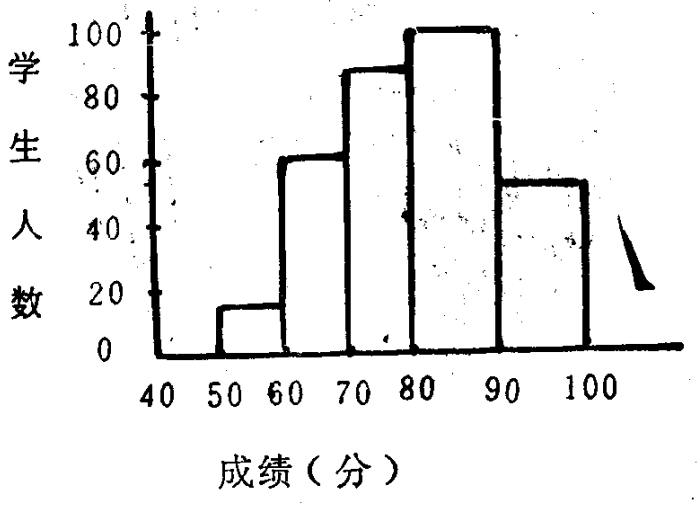

图8·2　某校学生英语成绩直方图

（学生总数305人，其中：90—100分45人，80—90分100人，
70—80分90人，60—70分60人，
50—60分10人。）

2）圆瓣图。它可以把相对数的资料展示在一个圆平面上，以图形中各个扇形面积来代表各部分的比重，使人看上去一目了然（图8·3）。根据图8·1的资料，学生父亲各类职业的比重及扇形面角度为：

$$干部 = 360° \times \frac{110}{500} = 79.2°$$

$$工人 = 360° \times \frac{225}{500} = 162°$$

$$农民 = 360° \times \frac{165}{500} = 118.8°$$

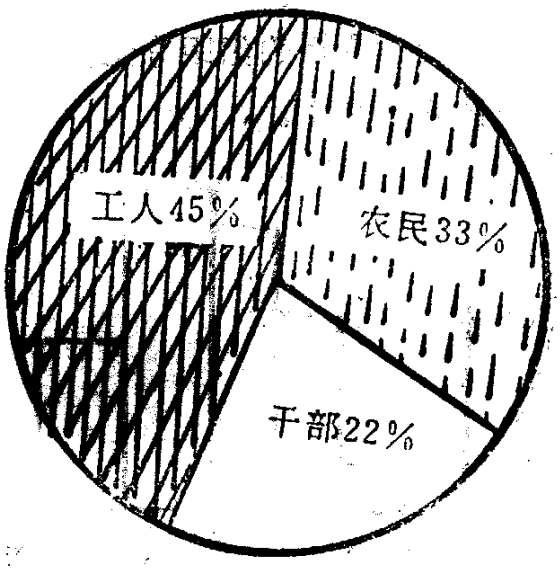

图8·3 圆瓣图

3）折线图。它是在直方图的基础上，取各组方形上端的中点，并连接各点形成一条折线。折线图不仅可以表示次数的分布，还可以绘制向上、向下累积的两种次数分布图。我们将图8·2的资料绘制出以下两张图。图8·4表示次数折线。图8·5表示两种累积方向的次数分布。

在绘制连续变量的组距分组资料时，应注意：

第一，次数分布折线图的开端要从资料中最小一组低一组的组中点开始，到此资料最大一组高一组中点为止(见图8·4)。

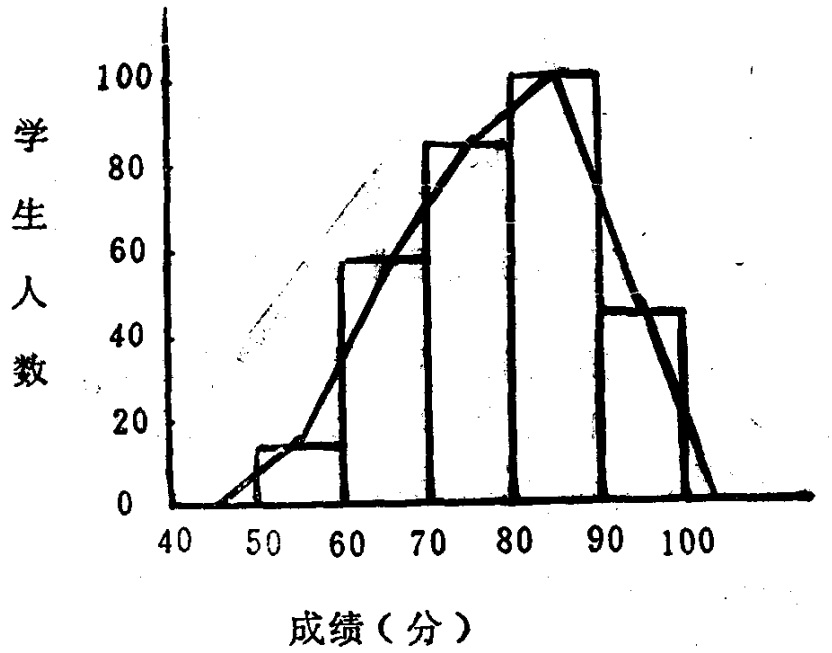

图8·4　某校学生英语成绩次数分布折线图

第二，累积次数分布折线，向上累积线的起点在第一组（最小组）的下限，连续各累积组的上限的纵座标，终点在最后一组（最高组）的上限；向下累积线则相反，其起点在最后组的上限，连接各累积组的下限的纵座标，终点在第一组的下限（图8·5）。

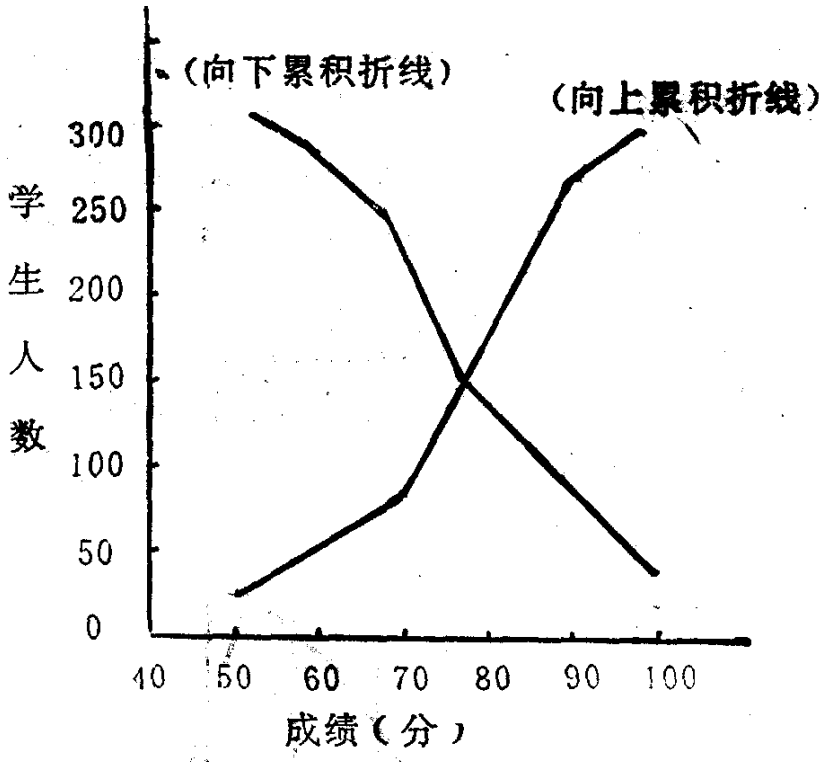

图8·5　某校学生英语成绩累积次数分布折线图

4）曲线图。它是变量值的数量趋于无限多时，折线图的极限描绘，是一种理论曲线，它实质上是对应于连续变量的次数分布的函数关系图，它是一条平滑的曲线，如正态分布曲线等。

五、统计表的制作

对调查表中的数据进行汇总的结果，便可以得到一系列被简化了的数字，把这些统计数字用一定形式的表格表现出来，就是统计表。统计表可以对调查的对象及有关现象作出集中、系统的说明，便于进行比较分析。

统计表是由标题、标目、横行、纵栏和统计数字所构成的：

标题。即统计表的名称，位于表的上端中央处；

标目。指对象的分组名称及各种具体项目；

横行、纵栏构成表格，决定统计表的尺寸，并区分各个统计数字。

表8—5　×企业职工文化程度统计表（标题）

（××××年）

文化程度	人数
（标目）	
1．文盲	20
2．小学	30
3．初中	335
4．高中	120
5．大专以上	45
合计	550

在制作统计表时要遵守以下规则：

第一，在内容安排上，标题在表上方。标目的排列要合乎逻辑。

第二，表的标题要清楚，能概括地反映表的内容。

第三，统计表要注明其内容所属的时间和地域。要注明指标的计量单位。

第四，统计表左右两侧不划纵线，一般采用“开口”表式。

第五，表中的同一栏数字，应对准位数，要有统一的精确度，即有效数字的位数一致。缺少的数字，用“……”表示；当数字为0时，用“——”来表示。

第六，统计表的资料来源及需附加的说明，要在表的下面注明，以便查考。

统计表根据使用分组标志的数量，主要分为分组表和复合表（交叉表），分组表使用一个标志进行分组，复合表则使用两个以上的标志分组，这里不再赘述。

总之，调查后的资料，通过检查、校订、摘录、分类、编码、登录以及各种简化的技术和环节的整理、加工，就初步形成了集中的、系统的、规范的、可供分析的次级资料。