第3章 统计整理

技能目标：

1. 能分析统计整理的实际意义；
2. 能懂得分布数列的编制要点；
3. 能进行科学的统计分组，形成系统的资料。

知识目标：

1. 掌握统计整理的过程；
2. 明确统计分组方法；
3. 了解统计图表的具体构成。

案例导入：

通过统计调查可以取得大量直接反映现象数量特征的具体的数据资料，这些数据资料是否能直接用于对现象特征和规律的归纳及进行统计分析呢？

3.1 统计整理的意义和步骤

3.1.1 统计整理的意义

统计整理是根据统计研究的目的与任务，对统计调查所获取的大量原始资料进行科学加工，以得出能够系统反映现象总体数量特征的综合数字资料的工作过程。统计整理还包括对次级资料（含历史资料）的再加工。

统计调查取得的第一手资料，是各单位具体情况的汇总，是分散、零碎、表面的。要反映总体的数量特征，还必须对这些资料进行加工整理。

统计整理也不是简单的数据汇总，而是对资料进行科学的分类加工。它是现象由个体量的观察到总体量认识的连接点，所以统计整理是统计工作和研究过程的中间环节，它是对客观现象从感性材料上升到理性认识的过渡阶段。统计整理既是统计调查的继续，又是统计分析的前提。只有通过科学的整理，才能形成反映事物整体全貌的数量特征。另外，统计整理也是对历史资料进行再加工的必要手段。对历史资料的甄选，以及按现有的口径对其重新整理，都必须通过统计整理工作来完成。

知识链接 3-1　数据的种类

- 横截面数据：在同一时刻或几乎同一时点所收集到的数据。
- 时间序列数据：在若干个连续时点内收集到的数据。

3.1.2　统计整理的步骤

统计整理通常包括统计资料的审核、分组、汇总、编表或制图四个环节。

1. 统计资料的审核

对调查第一手资料的审核是统计整理的基础环节，只有通过对收集到的资料进行全方位的审核，才能保证统计资料符合统计研究的目的和要求，确保资料的质量。统计资料的审核，主要包括以下两个方面。

(1) 审核资料的完整性和及时性。完整性审核主要是看调查单位或填报单位是否齐全、调查项目是否完整；及时性审核主要是指填报单位是否符合规定的时间要求，以及是否按时报送资料，有无不报、漏报或迟报等现象。

(2) 正确性审核。资料的正确性审核是审核的关键，主要检查所填报的资料是否准确可信。常用的审核方式有逻辑检查和计算检查。逻辑检查是指从理论上或常识上检查统计资料是否有违常理，有无不切实际或不合逻辑的地方。如人口普查登记中的某人年龄 7 岁，职业工人，就有违常理。计算检查是指检查有关指标的计算范围、计算方法、计量单位是否符合要求，指标之间是否相互衔接。

2. 统计资料的分组

根据统计研究的目的和要求，对经审核合格的资料进行科学的分类和分组，以保证得到科学的综合指标。

3. 统计资料的汇总

在统计分组的基础上，将各组资料进行综合汇总，得出反映各组和总体数量特征的指标。

4. 编表或制图

经过汇总整理的资料，我们还必须以系统简明的方式加以表现，以方便统计分析。这种方式一般就是编制统计表或绘制统计图。

3.2　统计分组

3.2.1　统计分组的概念和作用

所谓统计分组，就是根据统计研究的目的和要求及现象的特点，按照特定的标志和原则，把总体或总体单位进行科学的分类，便于系统地整理资料的方法。

要理解这一概念，通常可以从以下两个方面来认识：一方面，是划分总体，即把统计总体按个性特征分组标志划分为若干个统计组；另一方面，是合并个体，即把总体单位按共性特征(分组标志)合并为若干个统计组。

统计分组是统计整理的基础，也是统计分析的基本方法。没有科学的统计分组，也就没有科学的统计。其作用主要体现在以下几个方面。

1. 区分现象质的差别

社会经济现象千差万别，要了解各种现象的属性、特征及相互关系，首先必须对其进行科学的区分。统计分组的根本作用就在于划分现象的不同类型，以便揭示不同社会经济现象质的差异。例如，国企经济按产业结构划分为第一产业、第二产业、第三产业；企业按所有制属性划分为国有及国有控股企业、集体企业、私营企业、股份制企业、外资企业等类型。通过这种分类和分组，可以反映现象的地位及作用，也为进一步的统计分析提供便利条件。

2. 体现总体的内部结构

通过统计分组反映总体内部结构，可以分析和研究总体内部各组成部分的差别及相互联系，从而认识事物的本质特征，掌握其发展变化的过程及规律。例如，我国历年出口商品构成的变化，反映经济发展水平和结构的变化；而产业结构中从业人员构成的变化，则反映我国第一产业从业人员比重呈下降趋势，第三产业从业人员比重逐年呈上升趋势。

3. 分析现象之间的相互依存关系

社会经济现象之间存在着相互联系、相互依存、相互制约的关系。例如，农作物的产量与施肥量之间、投入与产出之间、市场商品供求与价格之间等。这些依存关系可以通过统计分组查明影响原因与得出结论之间的变动规律。例如，通过分析施肥量与粮食单位面积产量之间的依存关系，得出以怎样合适的施肥量才能达到粮食作物的最高单位面积产量。

3.2.2 统计分组的种类

统计分组一般包括如下两种。

1. 按标志的属性一般可分为品质分组和数量分组

品质分组是按品质标志进行的分组，即按事物的内在本质或属性为分组标志，如企业按所有制属性分组、人口按性别分组、大学生按专业分组等。品质分组可以反映总体各单位属性上的差异，从而体现不同属性的部分在总体中的地位和作用。在实际工作中，常常需要对研究现象进行复杂的品质分组，这种分组不仅涉及复杂的分组技术，而且涉及国家的政策和有关的科学理论。因此，国家为保证分类的统一性和完整性，专门制定了统一的分类目录和标准。

数量分组是按数量标志进行的分组，即选择反映事物数量差异的数量标志为分组标志。如企业按产值分组、人口按年龄分组、学生按成绩分组等。数量分组可以通过总体各单位数量上的差异分析，体现事物在性质上的差异。按标志的属性分组是统计分组中最重要的一种分组方式，统计分组方法主要是围绕品质分组和数量分组开展分析研究工作的。

2. 按分组标志的多少和形式可分为简单分组与复杂分组

简单分组就是按一个标志对总体进行的分组。如企业按生产规模分为大型企业、中

型企业和小型企业三组，货运量按运输方式分为铁路运输、公路运输、水路运输、航空运输和管道运输五组等。

复杂分组就是按两个或两个以上的标志对总体进行的分组。复杂分组又可分为重叠式分组和平行式分组两种。

重叠式分组是在按某一标志分组的基础上，再按另一标志进一步分组。例如，企业在按产业结构划分为第一产业、第二产业和第三产业的基础上，再按规模划分为大型企业、中型企业和小型企业。这样分组的结果，形成了几层重叠的组别，把我国产业结构的构成分析得更透彻、更详细。重叠式分组，可以对以下现象总体的各个层面进行更深入的分析。但是，重叠式分组会导致总体单位标志分类的复杂化，处理不当会很烦琐，不利于分析问题。因此在实际统计工作中，不可滥用，尤其不宜采用过多的标志进行重叠式分组。

平行式分组就是同时用两个以上的标志分别从不同的角度进行并列分组。例如，对学校教师总体，分别按性别、年龄段、职称进行的分组。平行式分组所形成的统计组别相互独立而不重叠，既可以从不同的方面反映总体的多重结构，又不致使分组过于烦琐，故在实际统计工作中被广泛采用。

3.2.3 分组标志的选择

分组标志是统计分组的依据或标准。统计分组的关键在于如何正确选择分组标志。正确选择分组标志，需要注意以下几个方面。

(1) 必须根据统计研究的目的和要求选择分组标志。统计总体有很多标志，每一个标志反映总体单位的某一方面的特征。而统计研究的目的不同，所选择的分组标志也应有所不同。例如，某高校在校大学生总体中，学生有性别、年龄、专业、学习成绩等许多标志；如果要研究学生的年龄结构状况，就必须以年龄为分组标志；如果要研究学生的学习成绩状况，就必须以每门课程的平均成绩为分组标志。所以，对于特定的研究目的，必须选择对应的分组标志分组。

(2) 必须选择最能反映事物本质特征的核心标志为分组标志。在统计总体的众多标志中，有些标志是反映现象本质的主要标志，而有的则是一般性的次要标志。例如，考核企业经济效益的好坏，可供选择的标志诸如总产值、净产值、增加值、销售收入、劳动生产率、单位产品成本、利税总额、资金占用额等。但其中最能综合反映企业经济效益的只能是利税总额。因此，我们只有选择最能反映事物本质特征的核心标志，才能抓住主要矛盾，得出科学的结论。

(3) 必须考虑现象所处的具体历史条件和经济条件。在不同的历史条件下，社会经济现象的特征呈现不同的形态，随着客观历史条件的变化，最能反映现象本质特征的标志也会发生变化。例如，在社会生产力水平较低下的条件下，人们把吃摆在首位，而把穿、住、用等放在次要位置；随着社会生产力的发展，产品丰富了，人们不愁吃了，逐渐把吃摆在次要位置，相反的为强调生活质量的提高，把穿、用、住等摆在了主要位置。再比如，同是划分企业规模，对于劳动密集型企业，是以职工人数为分组标志的；而技术密集型企业则是选择固定资产价值或产值为分组标志的。

知识链接 3-2　土地状况

我国的土地状况见表 3-1。

表 3-1　我国的土地状况

项　　目		面积/万平方千米	占总面积的比例/%
总面积		960	100.00
按地形分	山地	320	33.33
	高原	250	26.04
	盆地	180	18.75
	平原	115	11.98
	丘陵	95	9.90
按地高分	500m 以下	241.7	25.18
	500～1 000m	162.5	16.93
	1 000～2 000m	239.9	24.99
	2 000～3 000m	67.6	7.04
	3 000m 以上	248.3	25.86

3.3　分布数列

3.3.1　分布数列的概念

分布数列也称为分配数列或次数分布，是指在统计分组的前提下，把各组汇总的标志值按顺序排列成的一系列数据资料。其中分布在各组的总体单位数，我们称之为次数或频数，各组次数占总体的比重我们称之为比率或频率。例如，人口按性别分组形成的男性和女性人数的比重；生产工人按产量分组，所统计出的各组工人数。

分布数列包括两个组成要素：①组的名称，即按什么标志分的组；②各组的次数，即各组包含了多少个单位。分布数列直观地表明了总体单位的分布特征和结构状况，它是统计整理的一种重要形式。

3.3.2　分布数列的种类

反映复杂社会经济现象数量特征的分配数列，有多种形式。一般从以下三个方面进行分类。

(1) 根据分组标志的特征，分布数列可分为品质数列和变量数列。品质数列是按品质标志分组形成的分配数列(见表 3-2)；变量数列是按数量标志分组形成的分配数列(见表 3-3 和表 3-4)。

表 3-2　某企业产品质量检测分布

按产品是否合格	数量/件	比重/%
合格品	9 500	95
不合格品	500	5
合　计	10 000	100

表 3-3　某企业工人日产量情况分布

按日产量分组/件	工人数/人	比重/%
30	8	4.76
31	12	7.14
32	30	17.86
33	60	35.71
34	40	23.81
35	18	10.72
合　计	168	100.00

表 3-4　某企业职工工资构成分布

按月工资分组/元	工人数/人	比重/%
＜1 000	15	6.82
1 000～2 000	45	20.45
2 000～3 000	125	56.82
＞3 000	35	15.91
合　计	220	100.00

（2）根据数量标志的表现形式不同，分布数列可分为单值数列和组距数列。单值数列是指各组的分组标志为一个具体的变量值，由这些具体变量值表示的数列，例如表 3-3；组距数列是指各分组标志为一个变量区间，由这一系列变量区间表示的数列，例如表 3-4。

（3）按次数分布的状态不同，分布数列可分为钟形分布数列、U 形分布数列和 J 形分布数列。

钟形分布数列是指数列中变量值的次数分布呈两端次数少、中间次数多的形状。如车间工人日产量数列、班级学生学习成绩数列等。钟形分布数列又可分为正态分布数列和偏态分布数列两种。其中正态分布也就是以变量值中点为对称轴呈对称分布；偏态分布即变量值向某一方向偏斜，向左偏斜称左偏分布（见图 3-1），向右偏斜称右偏分布（见图 3-2）。

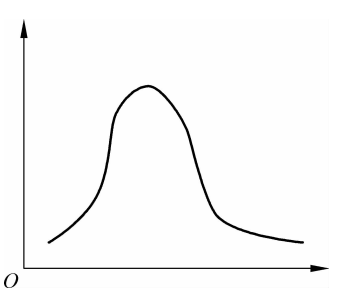

图　3-1

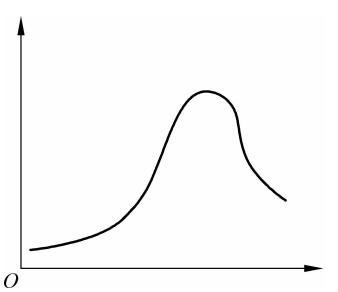

图　3-2

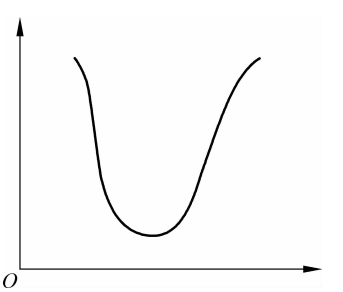

图　3-3

U形分布数列正好与钟形分布数列相反，即次数分布呈“两端次数多，中间次数少”的形状，也称为死亡率曲线、产品故障率曲线、浴盆曲线等（见图3-3）。例如，按年龄结构划分的人口死亡率，一天中天空云量百分比的分布等。

J形分布数列，即次数分布呈“一端次数多，另一端次数少”的形状。它包括两种情况：①次数随变量值的增大而增加，如供给曲线随价格（横轴）的增加，供应量（纵轴）相应增加（见图3-4）；②次数随变量值的增大而减少，如需求曲线随价格（横轴）的增加，需求量（纵轴）相应减少（见图3-5）。

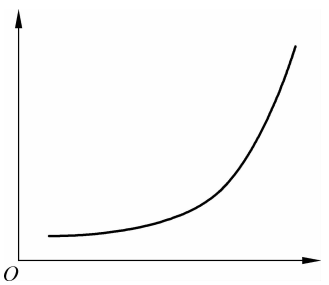

图 3-4

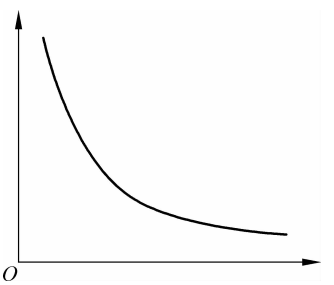

图 3-5

3.3.3 分布数列的编制

就编制品质数列而言，只要选择的分组标志正确，就比较简单，只需要按分组标志对总体单位中的相同标志进行直接加总即可。但对变量数列来说，由于它有单位值数列和组距数列之分，要按不同的变量值形式来对总体资料进行分组加工，所以相对较为复杂，并且组距数列还需要合理确定组数、组距、组限等问题，所以是最复杂的。以下以组距数列的编制为例来介绍变量数列的编制原理。

1. 基本概念

全距：总体单位标志中最大标志值与最小标志值之差。它反映总体单位标志值的最大差距状况，一般用 R 表示。

$$R = \text{最大标志值} - \text{最小标志值} \tag{3-1}$$

如果是分组数列，则是最大组上限与最小组下限之差。

组距：各变量组中最大标志值（上限）与最小标志值（下限）之差，表示为

$$\text{组距} = \text{上限} - \text{下限} \tag{3-2}$$

组限：统计组变量值两端的界限。最大变量值称上限，最小变量值称下限。

组限的形式与变量的特点紧密关联。如果分组标志是连续变量，组限一般采用重合式，即相邻两组中前一组上限与后一组下限相同；如果分组标志是离散变量，组限一般采用间断式，即相邻两组中前一组上限与后一组下限紧密相连而不重复。但为了方便统计工作，我们一般统一采用重合式，只是遵循“上限在下一组”或“只含下限不含上限”的原则。

组数：统计组的个数。在等距分组的条件下，组数等于全距除以组距。

闭口组和开口组：闭口组是既有下限，又有上限的统计组；开口组是只有上限或者只有下限的统计组，其中只有上限的称下开口组（最小组），只有下限的称上开口组（最大组）。

等距组与异距组：等距组即各统计组的组距相等，异距组是各统计组的组距不等。

组中值：是各组内标志值的一般水平。它是以组内标志值呈正态分布为假设前提的，在统计分析中应用非常广泛。其计算方法因组限形式不同而不同。

当统计组为重合式组距时

$$\text{组中值} = \frac{\text{下限} + \text{上限}}{2} \tag{3-3}$$

当统计组为间断组距时

$$\text{组中值} = \text{下限} + \frac{\text{组距}}{2} \tag{3-4}$$

当统计组为开口组时

$$\text{组中值} = \text{下限} + \frac{\text{相邻组组距}}{2} \quad \text{（上开口组）} \tag{3-5}$$

或

$$\text{组中值} = \text{上限} - \frac{\text{相邻组组距}}{2} \quad \text{（下开口组）} \tag{3-6}$$

2. 组距数列的编制原理

编制组距数列，一般包括以下几个步骤：①原始资料的初步整理，即把零散无章的数据资料按从小到大的顺序排列；②合理确定组数与组距，以次数分布能体现各组的分布规律为前提；③编制次数分布表，以显现总体各单位的数量分布情况；④绘制次数分布图，以检验编制的分布数列是否反映了现象数量变化的规律性。

对某城市居民家庭生活支出水平进行抽样调查，现取得72个家庭人均月生活支出资料。

1 260	1 060	1 670	970	1 480	2 000	1 870	1 620	1 440
1 250	1 000	1 840	2 200	1 360	820	1 610	1 480	1 140
950	1 320	1 570	1 300	1 210	1 980	1 450	1 290	1 180
1 020	740	1 550	1 020	1 440	1 400	1 770	1 630	1 230
710	980	1 490	1 360	1 250	1 780	1 530	1 390	1 100
1 760	890	1 990	1 650	1 470	1 200	1 740	770	1 960
810	1 000	1 200	1 050	1 880	1 430	1 180	1 050	1 340
1 140	1 250	1 480	2 180	1 650	980	1 580	2 130	1 290

首先，对上述无法看出其数量特征的数据进行排序整理。

710	740	770	810	820	890	950	970	980
980	1 000	1 000	1 020	1 020	1 050	1 050	1 060	1 100
1 140	1 140	1 140	1 180	1 180	1 200	1 200	1 210	1 230
1 250	1 250	1 250	1 260	1 290	1 290	1 300	1 320	1 340
1 360	1 360	1 390	1 400	1 430	1 440	1 440	1 450	1 470
1 480	1 480	1 480	1 490	1 530	1 550	1 570	1 580	1 610
1 620	1 630	1 650	1 650	1 670	1 740	1 760	1 770	1 780
1 840	1 870	1 880	1 960	1 980	1 990	2 000	2 130	2 180

其次，确定组距和组数。经过整理，我们发现该市居民家庭人均月生活费支出大多数在1 100～1 800元。数据中全距为1 490元，我们选择组距为200，则组数为8组(1 490/200＝7.45)。

再次，编制变量数列（见表 3-5）。

表 3-5　某市家庭人均月生活费支出情况

月生活费支出额/元	家庭数/户	比率/%
700～900	6	8.33
900～1 100	11	15.28
1 100～1 300	15	20.83
1 300～1 500	16	22.22
1 500～1 700	10	13.89
1 700～1 900	7	9.72
1 900～2 100	4	5.56
2 100～2 300	3	4.17
合　计	72	100.00

最后，绘制次数分布图（见图 3-6），检验所编制的变量数列是否反映现象发展变化的规律。

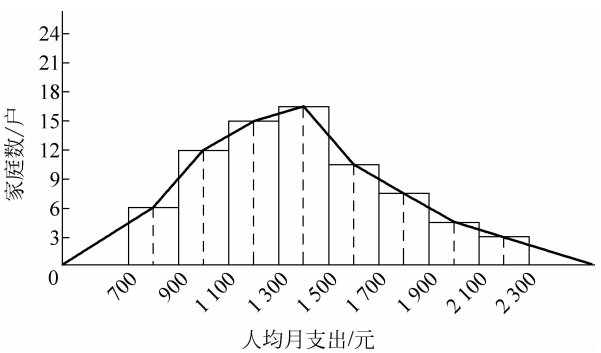

图 3-6　次数分布图

3. 累计次数分布的编制

要全面而深入地分析分布数列，除了简单统计各变量组出现的次数，体现在整个数列中次数分布的基本规律外，通常还要知道事物发展的进程等情况，即我们还要编制截至某一变量组总共的分布次数是多少。这是通过编制累计次数分布来实现的。

编制累计次数分布要计算累计次数和累计频率，包括两种计算方法：①向上累计（或称以下累计），是指从变量值低的组向变量值高的组进行的累计，它表明该组以下（如果是组距分组，则是该组上限以下）的累计次数和累计频率；②向下累计（或称以上累计），是指从变量值高的组向变量值低的组进行的累计，它表明该组以上（如果是组距分组，则是该组下限以上）的累计次数和累计频率。具体例子如表 3-6 和表 3-7 所示。

表 3-6　某班级学生按年龄分组

年龄/岁	学生人数/人	向上累计次数/次	向下累计次数/次
17	5	5	50
18	8	13	45
19	26	39	37
20	9	48	11
21	2	50	2
合　计	50	—	—

表 3-7　某班级学生按成绩分组

按成绩分组/分	学生人数/人	向上累计次数/次	向下累计次数/次
50～60	20	20	200
60～70	40	60	180
70～80	80	140	140
80～90	50	190	60
90～100	10	200	10
合　计	200	—	—

3.4　统计资料的汇总与显示

在统计分组的基础上，将统计资料分配到各组中，并计算各组和总体的累计数称为统计资料的汇总。而在汇总过程中，只有采用科学的汇总技术，才能节约人、财、物、时，保证汇总资料的准确和迅速，为统计分析打下良好的基础。

3.4.1　统计汇总的组织形式和方法

统计汇总有逐级汇总和集中汇总两种组织形式。

逐级汇总即从最基层的统计组织开始逐级向上汇总上报统计资料。逐级汇总便于就地审核和订正统计调查资料，并满足各级单位对统计资料的需要；但其经过的中间环节很多，需要的时间较长，出现误差的可能性较大。集中汇总即是把全部的统计资料集中到组织调查的最高机构进行的汇总。这种汇总形式不经过中间环节，可节省时间，减少误差；但原始资料如有差错就很难更正，并且很难满足各级部门对统计资料的需要。

在实行统计工作中，通常把上述两种组织形式结合起来运用，称之为综合汇总。即对于各级部门都需要的基本资料实行逐级汇总，对于只需要掌握整体情况的资料采用集中汇总的形式。如我国第五次全国人口普查，就是对于各地均需要的总人数、总户数及按性别、民族、文化程度分组的资料采用逐级汇总，而对其他资料则由省、市和中央进行集中汇总。

统计汇总是一项技术性很强的工作，汇总方法通常包括手工汇总和计算机汇总。

对于一些小规模、小范围的统计资料的汇总，一般采用手工汇总的形式。通过画记法、过录法、折叠法、卡片法等方式汇总相关资料。

(1) 画记法，即通过点线符号来汇总各组和总体单位数。画记法简便易行，但不能汇总标志值。如大学生新生入学后进行的班团干部民主选举。

(2) 过录法，即将统计资料过录到整理表上，然后再按整理表汇总资料。过录法既可汇总总体单位，也能汇总标志值，但工作量大，还会产生转录误差。

(3) 折叠法，即将调查表中需要整理的项目和数值折叠在边上汇总每一张表中的同一项目和数值。此法简单，但必须认真仔细，若出错则很难查找。

(4) 卡片法，即把总体各单位资料摘录到不同卡片上再按卡片进行整理。

对于大量的统计资料和复杂的统计工作，往往借助现代计算机技术进行计算机汇总。通过编制统计资料汇总的相关程序，录入原始资料，由计算机进行逻辑检查并完成