第二章　　统计资料的整理

调查收集到的统计资料常常是大量的。按照一般情形，如有上百个原始的数字和事项，人们就很难比较并发现它们的相互关系或规律。所以统计资料必须经过整理。统计资料的整理是统计方法中的关键环节，是以后进行统计分析的基础。

2．1　统计表

调查得到的原始统计资料必需加工整理，如分类归并汇总，按时间前后或按数值大小重新排列等等，才容易发现数据的规律性并便于做进一步的统计分析。表 2-1 为原始的统计资料。

表 2-1　某校 200 个学生高等数学考试成绩

98	54	79	73	99	78	55	65	65	72
52	65	96	72	70	84	77	62	63	88
83	60	78	88	76	67	94	86	85	72
65	73	80	72	100	60	58	63	76	63
76	66	83	63	60	69	68	56	85	87
84	62	75	87	86	70	82	85	65	72
69	81	100	72	71	85	70	75	72	55
92	65	56	55	60	45	75	65	76	70
65	82	85	70	62	75	70	62	75	57
72	82	65	57	71	80	88	91	81	60
64	75	78	60	58	76	60	65	72	78
65	53	62	85	75	71	65	84	76	80
75	60	61	86	86	78	81	76	62	83
72	73	67	61	85	91	86	92	86	75
75	86	65	86	65	94	86	65	91	100
85	70	61	94	86	85	90	86	82	56
76	73	85	65	88	74	76	95	73	85
85	61	52	78	85	86	71	95	85	65
61	62	85	65	85	96	75	64	62	78
65	85	75	76	62	86	96	65	98	62

序列表 将变量所取值按时间顺序或按地域排列的表，并分别称为时间序列表和地域序列表。见表 2-2。

表 2-2 我国全部职工平均工资指数（以 1952 年为 100）

年份	指数		年份	指数	
	货币工资	实际工资		货币工资	实际工资
1978	100	100	1995	1158.5	261.7
1980	130	121.7	1996	1287.1	262.5
1985	221.2	174.9	1997	1385.6	278.9
1990	415	189.8	1998	1609.4	324.1
1991	448.7	190.5	1999	1787.7	365.1
1992	545.8	213.2	2000	2007.6	410.1
1993	704.7	231.9			
1994	969.6	254.6			

分类表 可以按性质分类（常称为定性分布），也可以按数值分类（常称为频数分布）。

定性分布 先建立一个关于元素的类别系统，各类要互相排斥，而且是完备的，使被观测的各元素能既不重复又无遗漏地分到各类中去。记录分到同类中的元素个数，或将同类中各元素对所研究的变量的观测值加以归并，这样得到定性分布。见表 2-3。

表 2-3 全国高等学校情况（2000）

类别	学校数	学校百分比	在校学生数	学生百分比
综合大学	83	9.3	1108166	25.27
理工院校	293	32.85	1786372	40.73
师范院校	221	24.78	266778	6.08
医药院校	100	11.21	312440	7.12
农林院校	50	5.61	316778	7.22
财经院校	68	7.62	337099	7.69
其他院校	77	8.63	258270	5.89
总计	892	100.00	4385903	100.00

频数分布 按变量所取的值进行分类，分类的原则与定性分布相同，于是资料中每个观测值都分到相应类中去。记录各类中观测值出现的次数，制成表格形式，就是频数分布表。

在作频数分布表时，如果变量所能取值的数目很小，就按取值大小顺序排列，每个值为一类。如果变量所能取值的数目很大，特别当变量是连续的情形，就将变量所取的值分组，记录观测值落在各组中的资料（称为频数），作为表格形式，常称为频数分布表。

做分组频数分布表时，要先确定分几组（组数），每组变量取值范围的大小（组距）和取值范围的上、下限（组限）。

分组的目的是要简明扼要地了解大量数据的数值分布情况，所以组数不能太多。但分组后，落在同一组内的数据不再加以区分（都以该组的中点值代替），因此损失了不少原始数据的信息，所以组数也不能太少。对于观测值在 100 个或更多的资料，一般以分 10-15 组为宜。

各组的组距可以是相同的，称为等组距分布表。在等组距的情形下，当找出全部数据中的最大值 $x_{\max}$ 和最小值 $x_{\min}$ 后，组距 h 可由二者之差被组数 k 除得的商 $\frac{x_{\max}-x_{\min}}{k}$ 来表示。

当然，组距要取整数或便于计算的数。但是分组并不一定非“等组距”不可。有些数据按“等组距”分组后可能频数分布很不正常，例如有些组的频数太小甚至为 0，就不如用不等的组距为好。还有一种通常称作“开口组”的，即最小组只有上限没有下限，最大组只有下限没有上限，写成：“×××以下”“×××以上”，常用在事先不能确定组限就陆续收集数据，无法预计全体数据的最大值和最小值的情形。在确定了组数和组距后，就应写出各组的上下限，然后将各观测值一一归入相应的组即可做出分布表。现通过例题说明。数据采用表 2-1。

研究表 2-1 中 200 个学生的成绩分布，知最低分 $x_{\min}=45$，最高分 $x_{\max}=100$。若组距定为 10 分，对 200 名学生的成绩进行分组，在组数、组距和各组上、下限都确定了以后，就可对观测值逐一检查它们所属的组，在所属组的记录栏做一记号，按照我国习惯，用写“正”字方法，将各观测值检查、记录完毕，就可计算出各组的频数，此时，频数分布表就完成了。本例的频数分布表如表 2-4 所示：

表 2-4　某校 200 个学生高等数学考试成绩的频数分布表

分数	计　　数	人数（f）
40—49	一	1
50—59	正正下	14
60—69	正正正正正正正正正正	55
70—79	正正正正正正正正正正下	58
80—89	正正正正正正正正正丁	52
90—99	正正正丁	17
100—109	下	3
总数		200

在此数表中，我们可以看出，资料的许多细节已经失去。表中 58 人的分数在 70 分至 79 分之间,但却不知 58 人的分数在此 10 分全距中所呈分布的细节。若将组距缩小，组数增多，则细节的损失，就可减少；但组数过多，频数表则不易一目了然。而且当组数适当，频数分布较有规则性（Regularity），即两极端之组所含频数较小，渐近中央的组频数逐渐增大时，若组数增多，这种规则性会变得不明显。观察表 2-4，就可明白 200 个同学高数成绩的频数分布，也具有这种对称的规则性。假使组数增多，便渐渐地失去这种规则性。现在将组距改为 6 分及 2 分，则频数表分别如 2-5 和 2-6 所示：

表 2-5　某校 200 个学生高等数学考试成绩的频数分布表

分数	人数（f）	分数	人数（f）
40—45	1	76—81	25
46—51	0	82—87	42
52—57	12	88—93	10
58—63	29	94—99	11
64—69	28	100—105	3
70—75	39	总　数	200

表 2-6 某校 200 个学生高等数学考试成绩的频数分布表

分数	人数（f）	分数	人数（f）
45—46	1	75—76	22
47—48	0	77—78	8
49—50	0	79—80	4
51—52	2	81—82	7
53—54	2	83—84	6
55—56	6	85—86	30
57—58	4	87—88	6
59—60	8	89—90	1
61—62	15	91—92	5
63—64	6	93—94	3
65—66	21	95—96	5
67—68	3	97—98	2
69—70	9	99—100	4
71—72	14	总　数	200
73—74	6		

表 2-5 中组距等于 6，频数分布的规则性，仍然可以维持，同时细节的损失也可减轻；而表 2-6 中的组距等于 2，各组频数分布就变得很不规则了。由此可见，组数的确定应适当，亦不宜太多。

累积频数（Cumulative Frequency）：

由第一组起至第 i 组止各频数之和称为第 i 组的累积频数，记为 F_i，即

$$F_i = \sum_{k=1}^{i} f_k = F_{i-1} + f_i \qquad (i > 1) \tag{2-1}$$

频率（Percent Frequency）和累积频率：

频率是频数被总数 n 除：f_i/n，经常以百分数表示，见表 2-7 第 5 列。各组频率之和为 1。累积频率是频率的累加，可以和累积频数比照，见表 2-7 第 6 列。频率分布和累积频率分布由于不受总数 n 的影响，所以便于不同资料的比较。

表 2-7　某校 200 个学生高等数学考试成绩的频数表

组数	中点值	频数f_i	累积频数	频率F_i	累积频率
40—45	42.5	1	1	0.005	0.005
46—51	48.5	0	1	0.000	0.005
52—57	54.5	12	13	0.060	0.065
58—63	60.5	29	42	0.145	0.200
64—69	66.5	28	70	0.140	0.340
70—75	72.5	39	109	0.195	0.535
76—81	78.5	25	134	0.125	0.660
82—87	86.5	42	176	0.210	0.870
88—93	92.5	10	186	0.050	0.920
94—99	97.5	11	197	0.055	0.975
100—105	103.5	3	200	0.015	1.000
合计		200		1.000	

2. 2　统计图

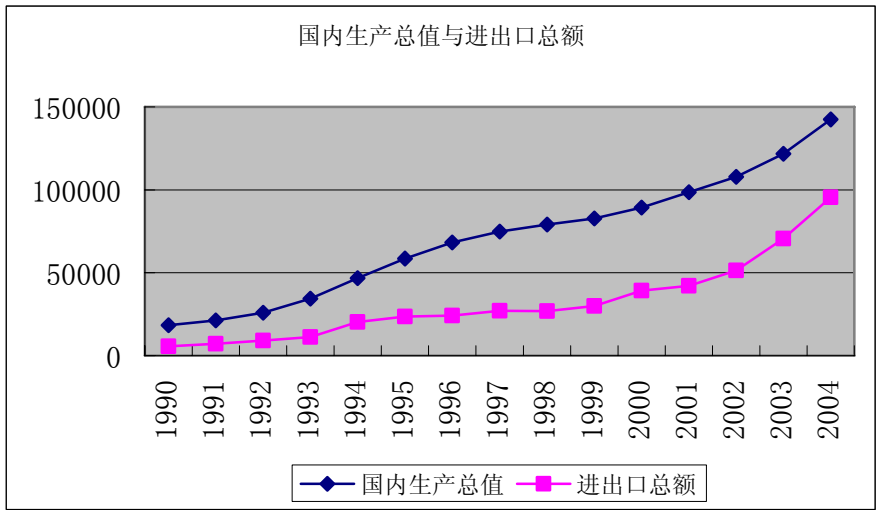
统计资料整理成统计表后，便于清晰地展示变量的变化规律。为了使这种规律更有直观性，也常用图形表示，称为统计图。

1. 线图（Line graph）

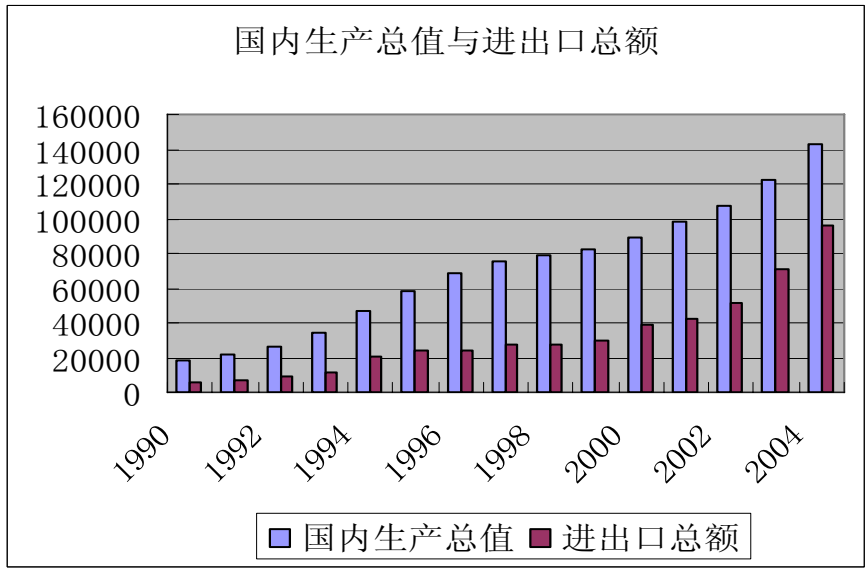
表 2-8　国内生产总值与进出口总额

年份	支出法国内生产总值（亿元）	进出口总额（亿元）
1990	18319.5	5560.1
1991	21280.4	7225.8
1992	25863.7	9119.6
1993	34500.7	11271
1994	46690.7	20381.9
1995	58510.5	23499.9
1996	68330.4	24133.8
1997	74894.2	26967.2
1998	79003.3	26857.7
1999	82673.1	29896.3

2000	89340.9	39274.2
2001	98592.9	42183.6
2002	107897.6	51378.2
2003	121730.3	70483.5
2004	142394.2	95539.1



2. **条形图（Bar chart）数据**(采用表 2-8)

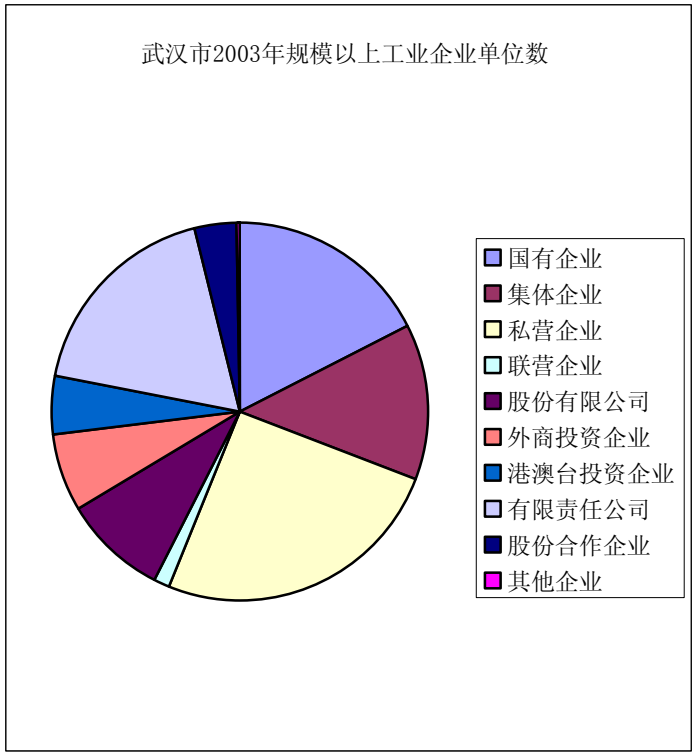


3. 圆饼图 (Pie chart)

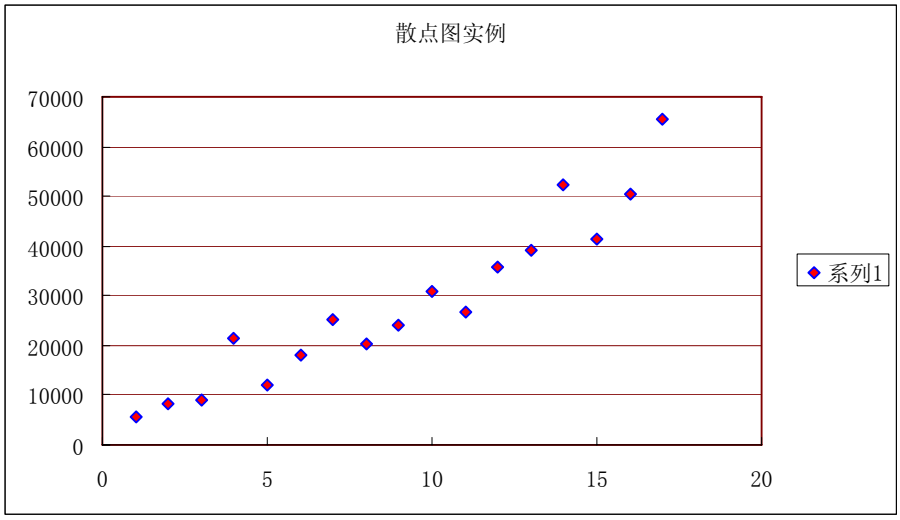
表 2-9 武汉市 2003 年规模以上工业企业单位数

类别	企业单位数
国有企业	252
集体企业	181
私营企业	337
联营企业	17
股份有限公司	121
外商投资企业	87
港澳台投资企业	67
有限责任公司	240
股份合作企业	48
其他企业	5
合计	1335

注：规模以上指销售收入 500 万以上



4. **散点图 (Scatter diagram)**



2.3 双变量的二元分布

对每一元素观测两个特征，记录观测结果，就是双变量的统计资料。双变量常用（X，Y）形式表示，以区别两个单变量 X 和 Y。整理双变量的统计资料时，将两变量分别分类（或按数值分组）：

$$X: x_1, x_2, \cdots, x_n$$
$$Y: y_1, y_2, \cdots, y_m$$

检查每一元素的两个特征应属于的类别，记录属于同类（x_i, y_i）的元素的数，即频数 f_{ij}，就得到二元分布。二元分布用矩形表表示，称为二元分布表，或称交叉表（Cross Table）。元素的总数:

$$n = \sum_{i=1}^{l} \sum_{j=1}^{m} f_{ij} \tag{2-2}$$

[例 2.1]在飞行模拟训练时，用计算机测定并打印出飞行动作的错误，从两方面进行测定：

（1）错误发生时的飞行状态，分起飞（T），巡航（C）和着陆(L)三种。

（2）错误发生的原因，分规范理解错误(R)，仪表读数错误(M)和其它原因(O)三种。

测定 45 次的打印记录如下：

TM	TO	LM	LO	CO	LM	TR	CM	TM
LO	TM	CO	LR	CM	TR	LO	TR	LO
CO	LO	LM	TM	TO	CM	TO	LM	TO
CR	CM	TM	TR	LR	TM	LR	TR	TM
LM	TR	TR	LO	CR	TR	LO	LM	TM

（3）根据该记录整理的二元分布表如下：

		错 误 原 因			合 计
		R	M	O	
飞行状态	T	8	8	4	20
	C	2	4	3	9
	L	3	6	7	16
合 计		13	18	14	45

从表中看出，在起飞（T）时容易发生规范理解错误（R）和仪表读数错误（M），而着陆（L）时不太容易发生规范理解错误。

在上述分布表中变量也可以是定量的。

[例 2.2]某旅行社 322 个旅游团的旅游天数和支出费用

费用（元）＼天数	3	4	5	6	7	8	合 计
0—299	53	43	9	—	—	—	105
300—599	20	45	37	12	26	1	141
600—899	4	9	11	6	17	12	59
900—1200	—	—	1	1	5	10	17
合 计	77	97	58	19	48	23	322

从表中看出，旅游费用是直接与天数有关的，天数为 3 天且费用不到 300 元的很多。

在二元分布表最下行（合计行）和最右列（合计列）分别是 X 和 Y 的单变量分布，称为边际分布。

一个双变量的二元分布绝不同于两个单变量的一元分布，它不仅说明两

变量各自的分布情况，而且说明两变量之间（飞行状态与错误原因之间，旅游天数与旅游费用之间）的相互关联情况。而这种关联情况（即是否存在关联以及关联的性态和程度等）正是研究二元分布的主要任务。

对于三变量（X，Y，Z）的统计资料，整理成分布表的形式是困难的，常用的方法是对于 X 的每一特定值 x_i，研究(Y, Z)的二元分布。更多变量的情形也类似。

习 题

1．最近的 20 天中一名员工生产的产品数量数据列示如下。

160	170	181	156	176
148	198	179	162	150
162	156	179	179	151
157	154	179	148	156

通过构筑：

a．频数分布

b．累积频数分布

汇总数据

2．某财经类杂志为了解订阅者进行了一次调查。调查问题之一是询问订阅者的投资证券组合（股票、公司债券、互助基金、存款）的价值。下列频数分布是调查的结果。

投资价值（元）	频数	投资价值/元	频数
25 000 以下	17	250 000-499 999	13
25 000- 49 999	9	500 000-999 999	13
50 000- 99 999	12	1 000 000 及以上	16
100 000- 249 999	20	合计	100

a．订阅者中投资少于 100 000 元的百分比有多大？

b．订阅者中投资在 100 000 ~499 999 元间的频率有多大？

c．订阅者中投资在 500 000 元及以上的频率有多大？

3．一项关于家用电脑使用时间的调查，显示了我国年龄在 12 岁及以上的电脑使用者使用电脑时间的情况。下列数据是一周时间内 50 人的样本使用个人计算机的小时数。

4.1	1.5	10.4	5.9	3.4	5.7	1.6	6.1	3.0	3.7
3.1	4.8	2.0	14.8	5.4	4.2	3.9	4.1	11.1	3.5
4.1	4.1	8.8	5.6	4.3	3.3	7.1	10.3	6.2	7.6
10.8	2.8	9.5	12.9	12.1	0.7	4.0	9.2	4.4	5.7
7.2	6.1	5.7	5.9	4.7	3.9	3.7	3.1	6.1	3.1

汇总数据，构建

a．频数分布（用 3 小时作组宽）。

b．频率分布。

c．直方图。

d．上述数据显示了在家使用个人计算机的何种情况。