第三章
图表描述

内容提要

通过调查所获得的原始数据，往往都是观测值众多且结构零散的，不加以系统整理，很难进行有效的整体性观察。大量、零散的原始数据有其内在的条理化结构，图表描述便是揭示原始数据内在结构条理的有效手段。

学习目标

知识目标

- 掌握品质型数据的图表描述方法
- 掌握数值型数据的图表描述方法
- 熟悉双变量关系的图表描述方法

能力目标

- 能用图表正确描述各种数据

引导案例

260 名毕业生就业意向的调查

某财经大学学生就业指导处每年都要吸引许多企业来校园参加招聘工作洽谈会，为此，就业指导处专门在应届毕业生中做了一项调查，询问每名学生的就业意向，并获得如表 3-1 所示的原始数据。

表 3-1　260 名毕业生就业意向的调查数据

	1 列	2 列	3 列	4 列	5 列	6 列	7 列	8 列	9 列	10 列
1 行	金融	会计	管理	会计	金融	会计	会计	管理	会计	营销
2 行	会计	营销	会计	其他	金融	会计	营销	金融	营销	营销
3 行	其他	管理	金融	管理	营销	管理	会计	会计	营销	其他
4 行	金融	营销	营销	金融	金融	营销	金融	营销	管理	会计
5 行	营销	金融	营销	营销	金融	会计	营销	营销	会计	营销
6 行	营销	营销	管理	金融	金融	会计	营销	营销	营销	营销
7 行	金融	会计	会计	金融	管理	金融	其他	会计	其他	金融
8 行	营销	会计	营销	会计	其他	管理	会计	金融	金融	金融
9 行	金融	其他	其他	营销	营销	会计	金融	营销	金融	营销
10 行	金融	会计	其他	会计	营销	金融	会计	金融	其他	其他
11 行	营销	其他	营销	金融	会计	会计	其他	其他	会计	金融
12 行	管理	管理	会计	管理	会计	金融	管理	会计	营销	金融
13 行	其他	管理	其他	会计	会计	其他	会计	会计	会计	金融
14 行	营销	会计	会计	会计	金融	会计	会计	管理	管理	会计
15 行	会计	会计	管理	会计	其他	会计	会计	会计	会计	营销
16 行	管理	金融	会计	管理	管理	会计	会计	营销	会计	其他
17 行	其他	营销	营销	金融	营销	金融	金融	其他	其他	其他
18 行	营销	金融	营销	金融	其他	管理	会计	营销	营销	会计
19 行	会计	会计	其他	其他	营销	金融	管理	管理	金融	金融
20 行	会计	金融	金融	营销	营销	会计	营销	会计	营销	会计
21 行	金融	金融	会计	会计	管理	营销	会计	其他	会计	金融
22 行	管理	营销	会计	营销	管理	营销	管理	管理	金融	管理
23 行	营销	营销	管理	营销	会计	会计	金融	营销	管理	营销
24 行	金融	营销	会计	营销	会计	营销	金融	会计	其他	金融
25 行	金融	会计	其他	会计	管理	管理	会计	营销	营销	会计
26 行	营销	会计	营销	营销	营销	金融	营销	营销	会计	金融

毫无疑问，上述数据中包含毕业生就业意向的有用信息，但仅从表面观察，很难从中获得有用的观察结果，更无法进行深入分析和计算。而如果用图表进行描述，就能很清楚地看出其内在联系。

图表描述的目的是以图和表为工具，对最初大量、零散、杂乱无章的样本数据加以分组和汇总，从而显示出数据本身的内在结构条理，以便进一步的分析和计算。数据整理过程包括两个基本步骤：一是分组，二是汇总。分组是指将样本数据中的观测值，按照变量的不同取值区分为不同组；汇总则是在分组的基础上统计出不同组的频数。

数据整理的直接结果是频数分布表。频数分布表显示了样本数据本身所固有的结构条理——频数分布状况。为使数据的频数分布状况表现得更为生动而直观，可以在频数分布表的基础上绘制频数分布图。

第一节　品质型数据的图表描述

不同类型的变量，其样本数据的图表描述方法略有不同。对于品质型变量的样本数据，常用的频数分布表是单项式频数分布表，常用的频数分布图有条形图和饼形图。

一、单项式频数分布表

表 3-1 中的原始数据涉及一个变量，即“就业意向”，这是一个定类变量，有 5 个变量值，即会计、金融、管理、营销、其他。尽管 260 个观测值之间存在差异，但这种差异并不是漫无边际的，它们分别归属 5 个不同变量值中的一个。如果先将 260 个观测值按所属变量值划分为 5 组，再汇总得出各组观测值的个数，原本大量、零散的原始数据就会在不损失任何原有细节的前提下得以简化，并显示出一种条理化的结构，如表 3-2 所示。

表 3-2　260 名毕业生就业意向频数分布表

就业意向	频数 / 人	频率 /%
会计	76	29.2
金融	54	20.8
管理	33	12.7
营销	68	26.2
其他	29	11.1
合计	260	100

观察表 3-2 中的数据，可以得出以下结论：在 260 名应届毕业生中，就业意向倾向于会计的人数为 76，占总人数的 29.2%；倾向于金融的人数为 54，占总人数的 20.8%；倾

向于管理的人数为 33，占总人数的 12.7%；倾向于营销的人数为 68，占总人数的 26.2%；其他为 29 人，占总人数的 11.1%。

表 3-2 称为单项式频数分布表，其编制过程包括两个步骤：第一步，按照变量值对原始数据进行分组；第二步，汇总得出各组观测值的个数，即该组的频数。频数分布表显示了全体观测值在不同组之间的分布状况，可以根据频数分布来把握数据的整体特征。

表 3-2 中的频率是对频数进一步加工计算所得出的派生结果，它是各组频数与全体观测值个数的比值。频率是对频数分布状况的进一步概括和抽象，它与频数所反映的内容在本质上是一致的。

二、条形图与饼形图

绘制条形图时，通常以横轴表示变量及其分组，以纵轴表示频数。每个条形的长短代表该组频数的多少；条形的宽窄及各条形之间的间隔没有实际含义，考虑到图形美观和避免引起歧义，通常取相等的宽窄和间隔。条形图的纵轴也可以表示频率，采用频率所绘制的条形图与采用频数所绘制的条形图的整体形状没有差别。如图 3-1 所示为 260 名毕业生就业意向的频数分布条形图。

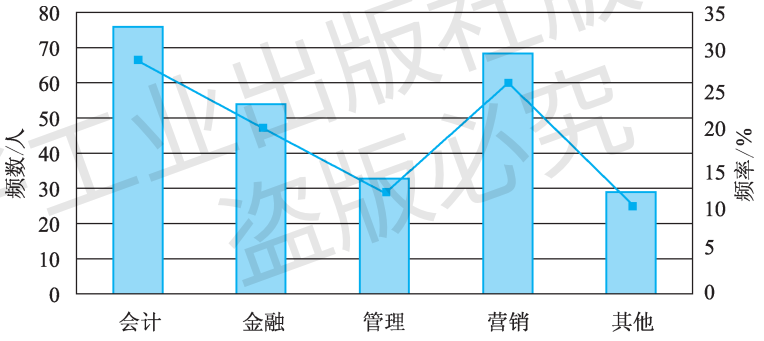

图 3-1　260 名毕业生就业意向的频数分布条形图

在饼形图中，整个圆的面积代表频数的 100%，各个扇形的面积代表各组的频率。饼形图的扇形面积也可以表示频数，但在实际工作中，人们一般习惯于在条形图中采用频数，在饼形图中采用频率。如图 3-2 所示为 260 名毕业生就业意向的频数分布饼形图。

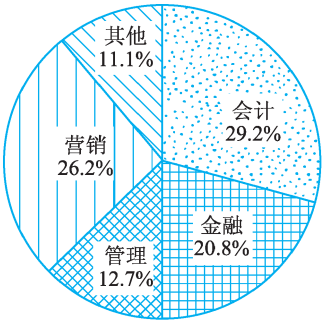

图 3-2　260 名毕业生就业意向的频数分布饼形图

小提示 IPS

饼形图与条形图所反映的内容是完全相同的，但从图示效果上看，饼形图可以更醒目地表现出频数分布的整体结构状态，条形图则更便于进行不同组之间频数差异的比较。

实际工作中，如果对各组频数高低的顺序感兴趣，还可以在条形图中重新排列各个条形的位置，如图 3-3 所示。图 3-3 称为帕累托图，它是按照各组频数高低排序绘制的条形图，由此图可以清楚地看到频数高低变化的整体情况。

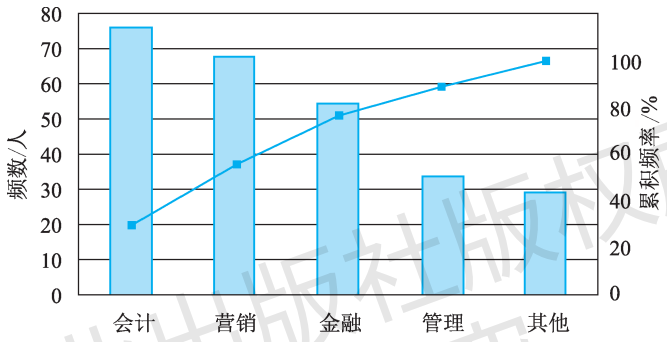

图 3-3　260 名毕业生就业意向频数分布帕累托图

以上是围绕定类数据所介绍的图表描述方法，这些方法同样适用于定序数据。例如，为评价某城市的空气质量状况，研究人员在该城市中测定了 300 个采样点，并获得表 3-3 所示的测定结果。

表 3-3　300 个采样点空气质量评价数据

空气质量等级	采样点个数	频　　率
优	193	64.4%
良	67	22.3%
轻度污染	28	9.3%
中度污染	7	2.3%
重度污染	5	1.7%
合计	300	100%

表 3-3 中的数据是定序变量数据，根据此数据所绘制的频数分布条形图及饼形图如图 3-4 和图 3-5 所示。

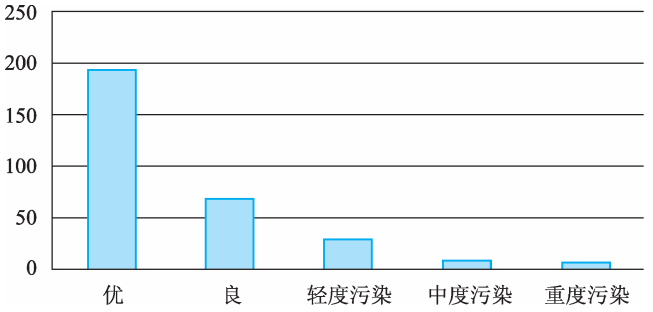

图 3-4　300 个采样点空气质量等级频数分布条形图

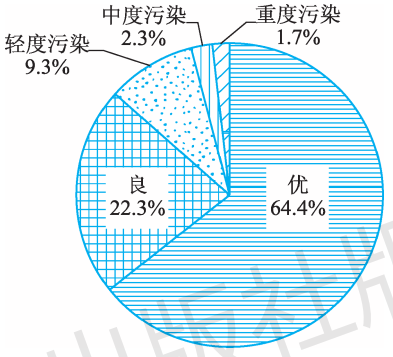

图 3-5　300 个采样点空气质量等级频数分布饼形图

由于定序数据的取值是具有顺序性的，所以其频数分布表及分布图中的分组数据不可随意调换位置，分组次序的混乱意味着原始数据中的信息损失。对于定序数据，如果需要直接从表中读出某一变量以上或以下累积频数的多少，则需要在频数分布表的基础上进一步绘制累积频数分布表。累积频数的计算分为向上累积与向下累积两种情况，向上累积回答某一变量值以下的累积频数是多少，向下累积回答某一变量值以上的累积频数是多少。此外，还可以根据累积频数计算出累计频率，如表 3-4 所示。

表 3-4　300 个采样点空气质量评价数据

空气质量等级	采样点个数	向上累积		向下累积	
		累积频数	累积频率 /%	累积频数	累积频率 /%
优	193	193	64.4	300	100
良	67	260	86.7	107	35.6
轻度污染	28	288	96	40	13.6
中度污染	7	295	98.3	12	4
重度污染	5	300	100	5	1.7
合计	300	—	—	—	—

第二节　数值型数据的图表描述

对于数值型变量的样本数据，常用的频数分布表是组距式频数分布表，常用的频数分布图主要包括直方图、盒形图和茎叶图。这些图表各有其优缺点，实际应用时应注意合理选择。

一、组距式频数分布表

（一）组距式频数分布表概述

如果数据中的变量值个数不是很多，可以参照与品质型数据相同的方法，以单个变量值作为分组标准来编制频数分布表。但在日常数据处理活动中所遇到的数值型数据，其变量值与观测值的个数往往很多，如果仍以单个变量值作为分组标准，最终得出的频数分布表就会由于组数太多而拖得很长，这样反倒不便于对频数分布状态进行整体性观察。

例如，为科学考核教学效果，任课教师每学期期末都要对自己所担任课程的学生考试成绩进行统计分析。如表 3-5 所示为某班级 60 名学生“统计学”课程的期末考试成绩数据。

表 3-5　60 名学生“统计学”课程的期末考试成绩数据　　单位：分

65	81	72	59	71	51	85	66	66	70
72	71	79	76	77	68	65	73	64	72
56	73	77	75	80	85	89	74	86	83
77	77	67	80	78	69	64	67	79	60
62	78	59	99	74	68	63	69	67	67
84	69	72	62	74	73	68	83	74	65

表 3-5 中共有 60 个观测值，其变量值个数多达 31 个。若以单个变量值来确定组别，就会有 31 组，组数太多，已经失去了用来进行整体性观察的意义。

实践中，规模比较大、变量值个数比较多的数值型数据采用组距式频数分布表，能够更好地概括显示频数分布状态。如表 3-6 所示是根据表 3-5 中的原始数据所编制的组距式频数分布表，该表概括地描述了 60 名学生“统计学”课程期末考试成绩的频数分布状态。一般来讲，学生期末考试成绩的频数分布呈现出“两头小，中间大”的特征是合理的。

表 3-6　60 名学生“统计学”课程期末考试成绩组距式频数分布表

成绩分组 / 分	频数 / 人
50 ～ 60	4
60 ～ 70	21
70 ～ 80	24
80 ～ 90	10
90 ～ 100	1
合计	60

组距式频数分布表不是以单个变量值来确定组别的，而是以表示一定取值范围的两个变量值来确定组别的，并以此为标准进行各组频数的汇总。在组距式频数分布表中，每组较小的那个变量值称作该组的下限，较大的那个变量值称作该组的上限，下限与上限之间的距离称作该组的组距，下限与上限之间的中点距离称作该组的组中值。例如，在“70 ～ 80”这一组中，下限为 70，上限为 80，组距为 10，组中值为 75。组距式频数分布表通常都是等距的，这主要是为了保证各组之间分布频数的可比性。

组距式频数分布表具有很强的概括性，无论数据规模多大，都可以通过组距的延伸加以分组和汇总。但它也有一个严重的缺陷，即组距越大，原始数据中的细节损失就越多。例如，“70 ～ 80”这一组所对应的频数为 24，但无从知晓这 24 名学生的考试成绩具体是多少。

（二）组距式频数分布表的编制步骤

由原始数据编制组距式频数分布表的具体步骤如下。

1．确定组数

确定组数时应以能够充分显示频数分布的整体特征为原则。组距过长，组数过少，会损失原始数据中的大量细节；组距过短，组数过多，又不便于对数据的频数分布特征进行整体性观察。

在实际工作中，很难找到一个确定组数的可操作的客观标准，数据分析人员往往根据自身的经验来确定组数。表 3-6 中将数据分为 5 组，从整理的结果上看，还是比较直观地显示了频数分布的整体特征。

小提示 IPS

有人曾提出确定组数的一个公式：组数 $=1+3.3\log(n)$。根据该公式，表 3-5 中数据的组数 $=1+3.3\log 60=6.87$，四舍五入，60 名学生期末考试成绩的组数应确定为 7 组。但这只是一个经验公式，不是绝对的标准。

2．确定组距

组距＝（最大观测值－最小观测值）/ 组数

在表 3-5 的原始数据中，最大观测值为 99，最小观测值为 51，如果组数确定为 5，则各组的组距＝(99－51) / 5＝9.6，四舍五入，组距可确定为 10。

3．确定组限

有了组距之后，只要确定了最小组的下限，则其余各组的组限也将随之确定。确定最小组的下限也带有一定的主观性，但必须遵循一个重要原则：最小组的下限必须包含数据中的最小观测值。考虑到表 3-5 数据中的最小观测值为 51，组距为 10，可将最小组的下限确定为 50。于是，各组的组限依次为“50 ～ 60”“60 ～ 70”“70 ～ 80”“80 ～ 90”“90 ～ 100”。

实践中所遇到的数值型变量多为连续型的，对于连续型变量数据来说，任何两个变量之间都存在着无数个可能的观测值，为避免频数汇总过程中的遗漏，相邻两组之间，较小组的上限应当与较大组的下限重合。例如，在“60 ～ 70”与“70 ～ 80”两组之间，较小组的上限“70”与较大组的下限“70”是重合的。对于离散型数据，则没有这种硬性规定。

4．频数汇总

频数汇总过程中要遵循“不重不漏”的原则，其中，“不重”是指同一个观测值在频数汇总过程中不能重复统计，“不漏”是指原始数据中的全部观测值必须包含在最小组的下限与最大组的上限所界定的范围之内。例如，如果某一观测值为 80，则不得在“70 ～ 80”与“80 ～ 90”两组中同时统计频数。实际工作中一般遵循“上限不计入本组内”的原则，即取值为 80 的观测值要计入以 80 为下限的“80 ～ 90”这一组的频数之内，而不应计入以 80 为上限的“70 ～ 80”这一组的频数之内。

此外，也可在数值型变量数据频数分布表的基础上，编制累积频数分布表，如表 3-7 所示。

表 3-7　60 名学生“统计学”课程期末考试成绩组距式累积频数分布表

成绩分组	频数 / 人	向上累积		向下累积	
		累积频数	累积频率 /%	累积频数	累积频率 /%
50 ～ 60	4	4	7	60	100
60 ～ 70	21	25	42	56	93
70 ～ 80	24	49	82	35	58
80 ～ 90	10	59	98	11	18
90 ～ 100	1	60	100	1	2
合计	60	—	—	—	—

二、直方图

组距式频数分布表所描述的频数分布状态可以通过直方图更为直观地显示出来。如图 3-6 所示即为根据表 3-6 绘制的 60 名学生“统计学”课程期末考试成绩频数分布直方图。

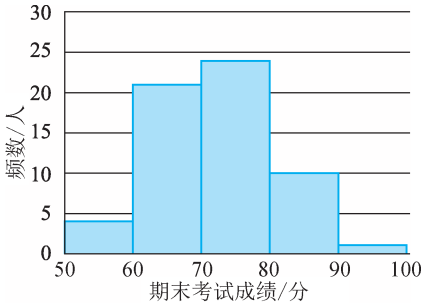

图 3-6　60 名学生“统计学”课程期末考试成绩频数分布直方图

直方图是直接根据组距式频数分布表绘制出来的，通常以横轴表示变量分组，纵轴表示各组的频数。直方图与条形图的形状类似，但两者之间有着本质区别。条形图的宽窄是没有含义的，直方图的宽窄则表示各组的组距；制作条形图时，通常要使各个条形之间保持一定的间隔，在直方图中各组之间则是没有间隔的。直方图是以面积来显示数据的，当某一组的频数为零时，代表该组数据的条形高度为零。相应地，条形面积也为零。

为强调频数分布的整体特征，还可以在直方图的基础上进一步加工制作出频数分布折线图或曲线图。折线图是将直方图中各个条形上端的中点用直线连接起来所形成的图形，它可以通过折线与横轴所围成的面积来显示数据。如图 3-7 所示是根据图 3-6 的直方图所绘制的折线图。

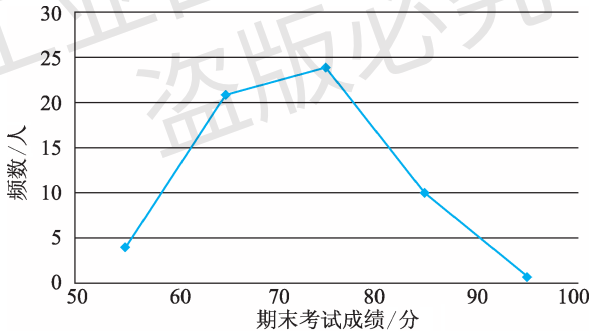

图 3-7　60 名学生“统计学”课程期末考试成绩频数分布折线图

假定数据规模无限扩大，同时组距无限缩小，而组数又无限增多，那么折线图就将趋近于一条平滑的曲线，从而形成频数分布曲线图。如图 3-8 所示是根据图 3-6 的直方图所绘制的曲线图。

直方图及其所派生的折线图与曲线图具有很强的概括性，适用于规模较大的数值型数据的频数分布显示，而且比组距式频数分布表更为生动、直观，因而成为实际数据处理工作中最为常用的一种频数分布图形。但它有一个明显的缺陷，即在确定组数和组距时带有一定的主观性。

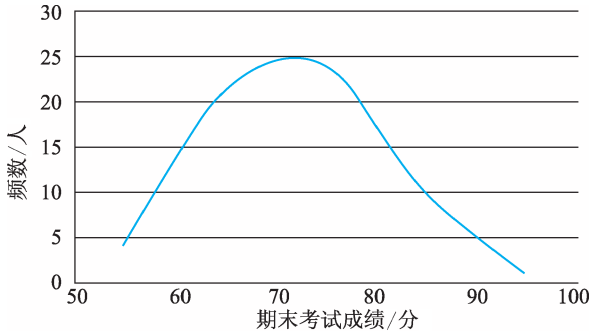

图 3-8　60 名学生“统计学”课程期末考试成绩频数分布曲线图

三、盒形图

盒形图也称箱线图，是利用数据中的最小观测值、下四分位数、中位数、上四分位数和最大观测值五个统计量来描述数据的方法。将数据中的全部观测值按照从小到大的顺序排成一列，处于第一位置上的观测值即为该数据的最小观测值，处于第 1/4 位置上的观测值即为下四分位数，处于第 1/2 位置上的观测值即为中位数，处于第 3/4 位置上的观测值即为上四分位数，处于最后位置上的观测值即为该数据的最大观测值。依此定义，可得如表 3-5 所示的五个统计量分别为 51、67、72.5、79、99，据此可绘制出 60 名学生“统计学”课程期末考试成绩频数分布盒形图，如图 3-9 所示。

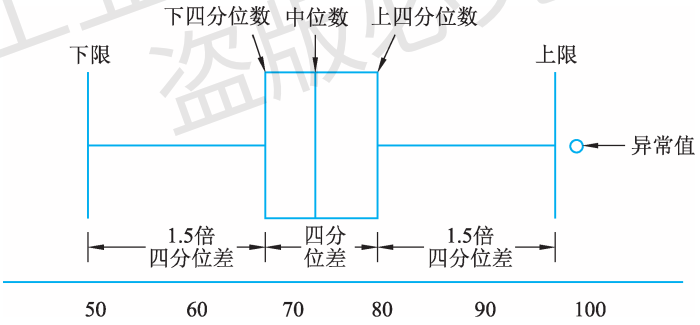

图 3-9　60 名学生“统计学”课程期末考试成绩频数分布盒形图

图 3-9 中，方盒的左侧边界对应下四分位数，右侧边界对应上四分位数。从下四分位数到上四分位数之间的距离称为四分位差，也就是说，方盒的宽窄代表四分位差的大小。方盒内的竖线对应中位数。

由下四分位数和上四分位数的定义可知，数据中大于下四分位数、小于上四分位数的观测值的个数占全部观测值个数的 50%，也就是说，方盒中包含了数据中 50% 的处于中间位置的观测值。由中位数的定义可知，数据中小于或大于中位数的观测值的个数各占全部观测值个数的 50%，也就是说，中位数处于全部数据的中间位置。

于是，通过观察方盒的宽窄以及中位数在方盒中的相对位置，就可以大体判断出频数分布的离散程度及其对称性。方盒越宽，表明频数分布的离散程度越高；方盒越窄，

表明离散程度越低；中位数在方盒中的相对位置趋近于中间，则表明频数分布具有较强的对称性。

方盒两侧延伸出来的线段是一个人为给定的变动范围，其两侧限制线分别在比下四分位数低 1.5 倍四分位差和比上四分位数高 1.5 倍四分位差的位置上。在数据处理活动中，如果某个观测值的大小超出了这个范围，则将被数据处理人员识别为异常值。图 3-9 中右侧的圆点就是一个偏大的异常值。

小提示 IPS

这里主要介绍盒形图的结构，有关上四分位数和下四分位数的计算方法，将在第四章中进行详细介绍。

四、茎叶图

茎叶图也称枝叶图，其基本思路是将数组中的数按位数进行比较，将数的大小基本不变或变化不大的位作为一个主干（茎），将变化大的位作为分枝（叶），列在主干的后面，这样就可以清楚地看到每个主干后面有几个数，每个数具体是多少。如图 3-10 所示是根据表 3-5 中的数据所绘制的频数分布茎叶图。

频数	茎	叶
1	5	1
2	6	99
6	6	022344
15	6	555667777888999
14	7	01122223334444
9	7	567778899
7	8	0012334
5	8	55679
1	9	9

图 3-10　60 名学生“统计学”课程期末考试成绩频数分布茎叶图

茎叶图包括“茎”与“叶”两个要素。图 3-10 中竖向排列的茎，显示各个观测值的十位数；对应每节茎向右横向排列的叶，显示各个观测值的个位数。图形右侧由全部观测值的个位数堆积形成的外部轮廓线，正好显示了频数分布的整体特征。与此同时，原始数据的全部细节并没有任何损失。

此外，图形左侧还给出了对应每节茎所确定的各个分组的频数，这使得整个茎叶图具有多功能的性质，既是频数分布图，又是频数分布表。

五、频数分布的类型

频数分布所显示出来的特征是由变量本身的性质决定的，而不是由描述方法决定的。不同性质的变量，其频数分布会表现出不同的特征。实践中，常见的频数分布可分为钟形分布、U 形分布和 J 形分布三种基本类型。

（一）钟形分布

这是最常见的频数分布类型，其频数分布的图形轮廓好像一座倒扣过来的钟，呈现出“两头小，中间大”的特征。变量的取值越靠近中间，频数就越高；越靠近两边，频数就越低。钟形分布又可分为正态分布、左偏分布和右偏分布三种情况，如图 3-11 所示。

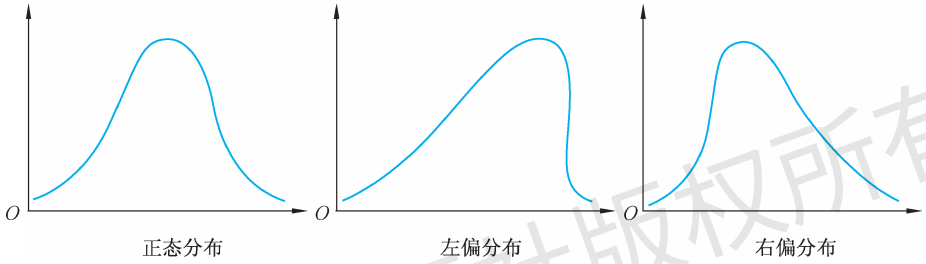

图 3-11　钟形分布

（二）U 形分布

U 形分布的特征与钟形分布刚好相反，变量的取值越靠近中间，频数越低；越靠近两边，频数越高。频数分布的图形轮廓好像字母“U”，如图 3-12 所示。

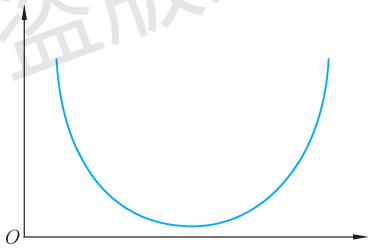

图 3-12　U 形分布

（三）J 形分布

J 形分布的特征是：频数随着变量取值的增大或减小逐渐增高，频数分布的图形轮廓像字母“J”。J 形分布又可分为正 J 形分布和反 J 形分布两种情况，如图 3-13 所示。

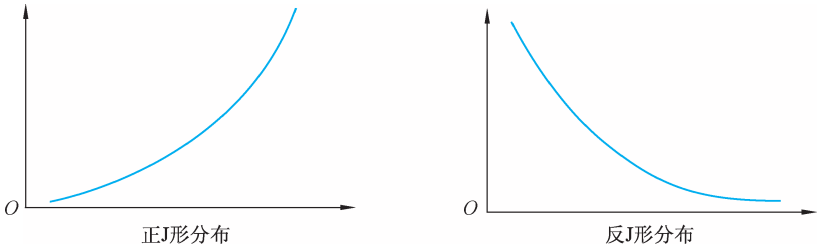

图 3-13　J 形分布

第三节 双变量关系的图表描述

前面介绍的都是单变量数据的图表描述方法，实践中经常会遇到双变量数据，如果要通过图形来观察变量之间的关系，就要用到双变量关系的图表描述。本节主要介绍交叉频数分布图和散点图，交叉频数分布图适用于两个品质型变量关系的描述，散点图适用于两个数值型变量关系的描述。

一、交叉频数分布图

若要通过图形来观察两个品质型变量之间的关系，可以根据样本数据绘制交叉频数分布图。例如，为开展诺基亚、摩托罗拉、爱立信、三星四种品牌手机的广告设计活动，广告公司经理需要知道四种品牌手机在学生群体中的市场占有情况。为此，专门针对不同身份的学生进行了一次关于手机使用情况的调查。被调查的学生需要回答他们愿意购置哪种品牌的手机，并说明他们是初中生、高中生还是大学生。调查获得的数据如表 3-8 所示。

表 3-8 四种品牌手机在学生群体中的市场占有情况

组　别	初 中 生	高 中 生	大 学 生	合　计
诺基亚	27	29	33	89
摩托罗拉	18	43	51	112
爱立信	38	15	24	77
三星	37	21	18	76
合计	120	108	126	354

表 3-8 是由原始数据加工整理所获得的一个交叉频数分布表，涉及手机品牌与使用者身份两个品质型变量。观察表中的数字，可以得出有关手机市场占有情况的初步判断。若要进一步观察两个变量之间的内在关联情况，可以绘制交叉频数分布图，如图 3-14 所示。

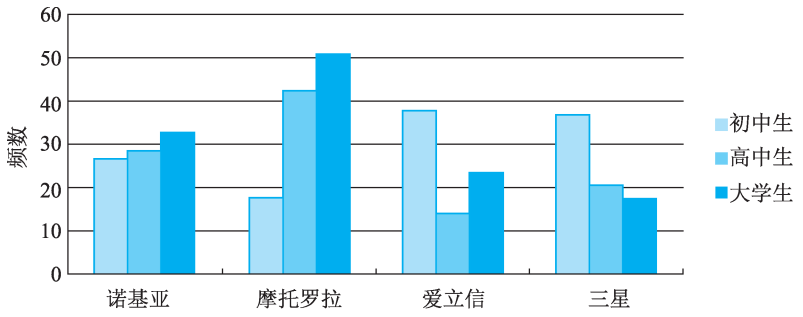

图 3-14 手机品牌与使用者身份交叉频数分布图

从图 3-14 中可以清楚地看到，初中生、高中生和大学生手机使用者在不同品牌手机中的分布是有所差异的。交叉频数分布图实际上是对单变量条形图的一种组合应用，借助此图可以直观地观察两个品质型数据的交叉频数分布情况。

二、散点图

某家具销售商认为家具销售与住宅面积密切相关，为证实这一想法，专门收集了其所在地区近 10 年来的相关统计数据，如表 3-9 所示。

表 3-9　某地区近 10 年新增住宅面积与家具销售额

年　　份	新增住宅面积 / 万平方米	家具销售额 / 万元
1	65	9
2	62	12
3	69	15
4	76	19
5	89	22
6	105	32
7	116	43
8	134	55
9	164	67
10	170	76
合计	1 050	350

表 3-9 中的数据涉及新增住宅面积与家具销售额两个数值型变量，仅观察数字，难以对两个变量之间的关联性做出明确判断。此时，可以用散点图加以描述。散点图是通过样本数据判断两个变量之间关联性的图形工具，适用于描述两个数值型变量之间的关系。如图 3-15 所示是根据表 3-9 中的数据所绘制的散点图。

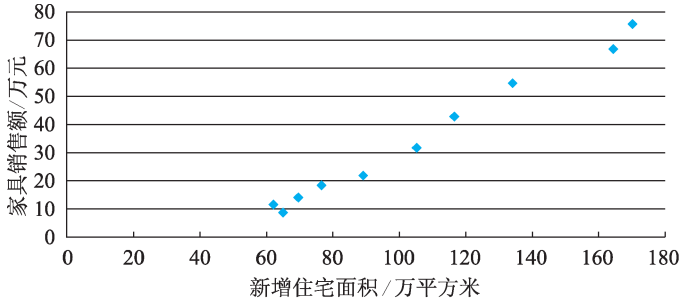

图 3-15　新增住宅面积与家具销售额散点图

图 3-15 中的横轴表示新增住宅面积的取值，纵轴表示家具销售额的取值，这两个变量在样本数据中的每对取值决定了图中的一个点。观察这些点的分布状况，可以帮助判

断和把握两个变量之间的关系类型及其相互关联的密切程度。从图 3-15 中不难看出，新增住宅面积与家具销售额之间具有一种正向的线性关联，各点整体上沿着一条向上的直线上下波动。

散点图具有一个明显的优点，就是在图形的绘制过程中没有损失原始数据的任何细节。取每个点在横轴和纵轴上的投影，即可重新完整获得原始数据。

第四节　运用 SPSS 进行图表描述

一、运用 SPSS 制作单项式频数分布表

根据 260 名毕业生就业意向调查数据制作单项式频数分布表的主要操作步骤如下。

（1）打开表 3-1 对应的 SPSS 数据集“data3.1”。在 SPSS 菜单栏中选择【Analyze】→【Descriptive Statistics】→【Frequencies】菜单命令，弹出如图 3-16 所示的“Frequencies”对话框。

（2）选择变量“就业意向 [jyyx]”，单击按钮，将其移到“Variable(s)”列表框中。选中“Display frequency tables”复选框，单击【OK】按钮。系统输出 260 名毕业生就业意向单项频数分布表，如图 3-17 所示。

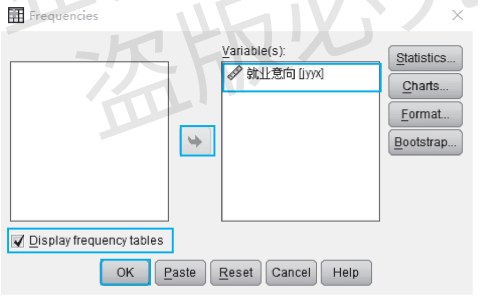

图 3-16　“Frequencies”对话框

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	会计	76	29.2	29.2	29.2
	金融	59	22.7	22.7	51.9
	管理	33	12.7	12.7	64.6
	营销	63	24.2	24.2	88.8
	其他	29	11.2	11.2	100.0
	Total	260	100.0	100.0	—

图 3-17　260 名毕业生就业意向单项频数分布表

二、运用 SPSS 制作条形图

根据 260 名毕业生就业意向调查数据制作频数分布条形图的主要操作步骤如下。

（1）打开表 3-1 对应的 SPSS 数据集“data3.1”。在 SPSS 菜单栏中选择【Graphs】→【Legacy Dialogs】→【Bar...】菜单命令，系统弹出如图 3-18 所示的“Bar Charts”对话框。

（2）单击【Define】按钮，系统弹出如图 3-19 所示的“Define Simple Bar: Summaries for Groups of Cases”对话框。

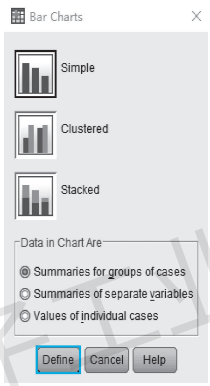

图 3-18　“Bar Charts”对话框

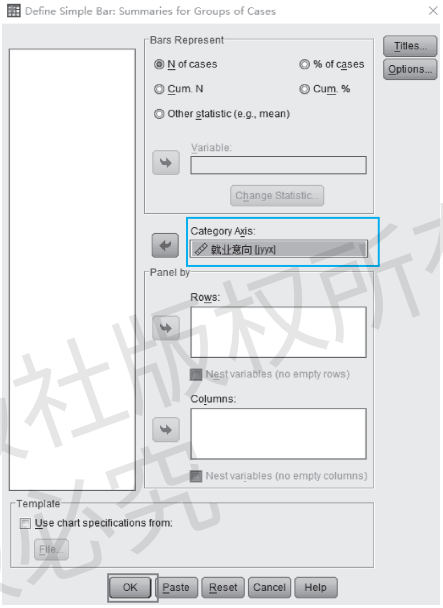

图 3-19　“Define Simple Bar: Summaries for Groups of Cases”对话框

（3）选择变量“就业意向 [jyyx]”，单击第二个按钮，将其移到“Category Axis：”列表框中，单击【OK】按钮。系统输出 260 名毕业生就业意向频数分布条形图，如图 3-20 所示。

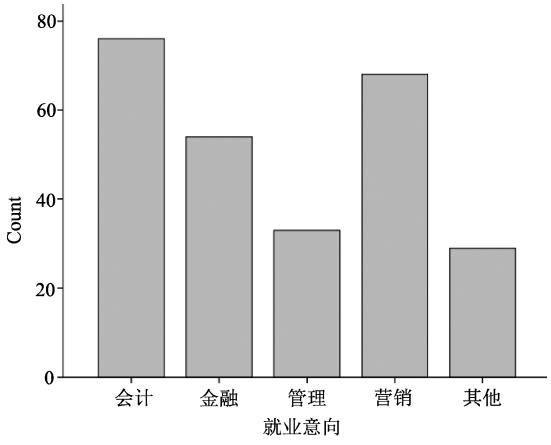

图 3-20　260 名毕业生就业意向频数分布条形图

三、运用 SPSS 制作饼形图

根据 260 名毕业就业意向调查数据制作频数分布饼形图的主要操作步骤如下。

（1）打开表 3-1 对应的 SPSS 数据集“data3.1”。在 SPSS 菜单栏中选择【Graphs】→【Legacy Dialogs】→【Pie...】菜单命令，系统弹出如图 3-21 所示的“Pie Charts”对话框。

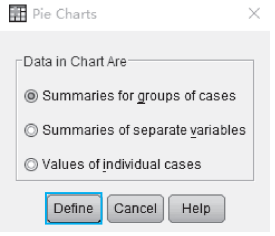

图 3-21　“Pie Charts”对话框

（2）单击【Define】按钮，系统弹出如图 3-22 所示的“Define Pie:Summaries for Groups of Cases”对话框。

（3）选择变量“就业意向 [jyyx]”，单击第二个按钮，将其移到“Define Slices by:”列表框中，单击【OK】按钮。系统输出 260 名毕业生就业意向频数分布饼形图，如图 3-23 所示。

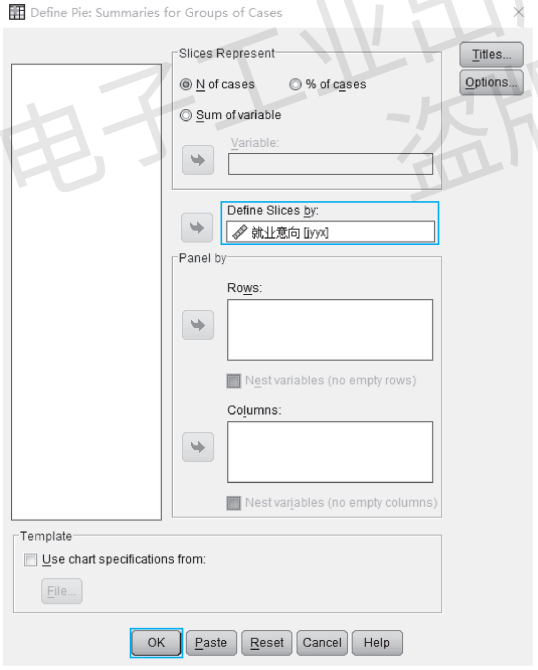

图 3-22　“Define Pie：Summaries for Groups of Case”对话框

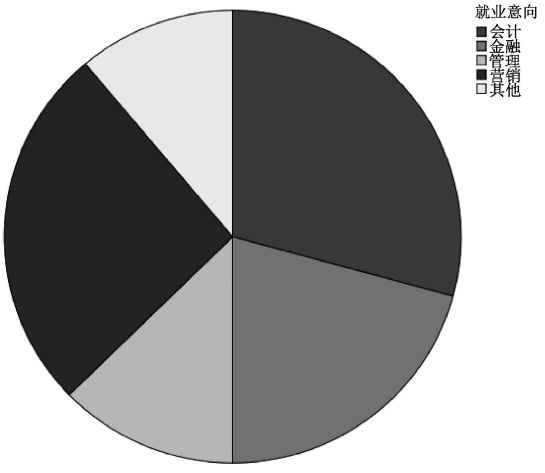

图 3-23　260 名毕业生就业意向频数分布饼形图

四、运用 SPSS 制作直方图

根据 60 名学生“统计学”课程的期末考试成绩数据制作直方图的主要操作步骤如下。

（1）打开表 3-5 对应的 SPSS 数据集“data3.5”。在 SPSS 菜单栏中选择【Graphs】→【Legacy Dialogs】→【histogram...】菜单命令，系统弹出如图 3-24 所示的“Histogram”对话框。

（2）选择变量“期末成绩 [qmcj]”，单击第一个按钮，将其移到“Variable:”列表框中，单击【OK】按钮。系统输出 60 名学生“统计学”课程的期末考试成绩频数分布直方图，如图 3-25 所示。需要指出的是，可以根据实际需要对 SPSS 输出结果的坐标轴刻度、背景色、分组数进行相应的调整。

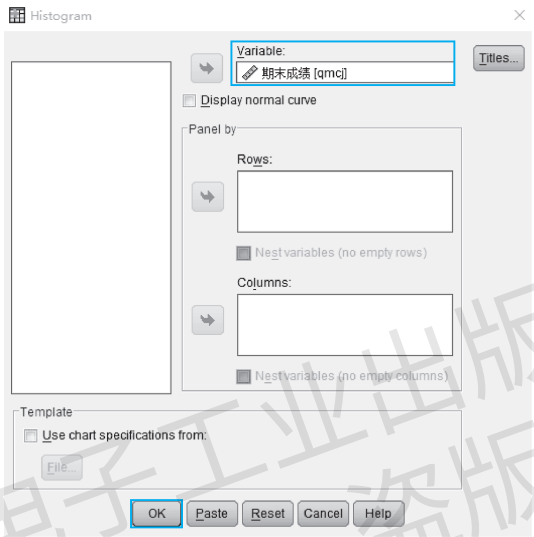

图 3-24　“Histogram”对话框

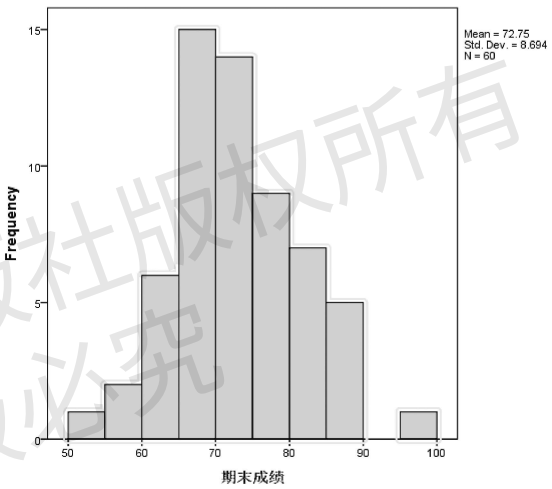

图 3-25　60 名学生“统计学”课程的期末考试成绩频数分布直方图

五、运用 SPSS 制作盒形图

根据 60 名学生“统计学”课程的期末考试成绩数据制作盒形图的主要操作如下。

（1）打开表 3-5 对应的 SPSS 数据集“data3.5”。在 SPSS 菜单栏中选择【Graphs】→【Legacy Dialogs】→【Boxplot...】菜单命令，系统弹出如图 3-26 所示的“Boxplot”对话框。

（2）在“Data in Chart Are”栏内选中“Summaries of separate variables”单选按钮，单击【Define】按钮，系统弹出如图 3-27 所示的“Define Simple Boxplot: Summaries of Separate Variables”对话框。

（3）选择变量“期末成绩 [qmcj]”，单击第一个按钮，

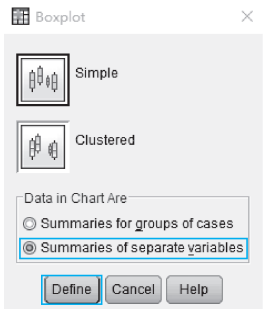

图 3-26　“Boxplot”对话框

将其移到“Boxes Represent:”列表框中，单击【OK】按钮。系统输出 60 名学生“统计学”课程的期末考试成绩频数分布盒形图，如图 3-28 所示。

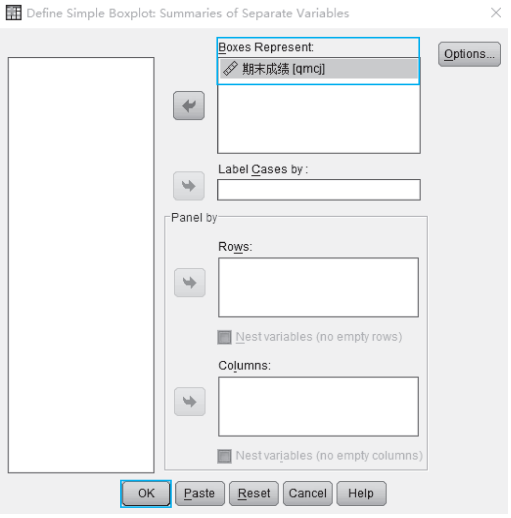

图 3-27 “Define simple boxplot: summaries of separate variables”对话框

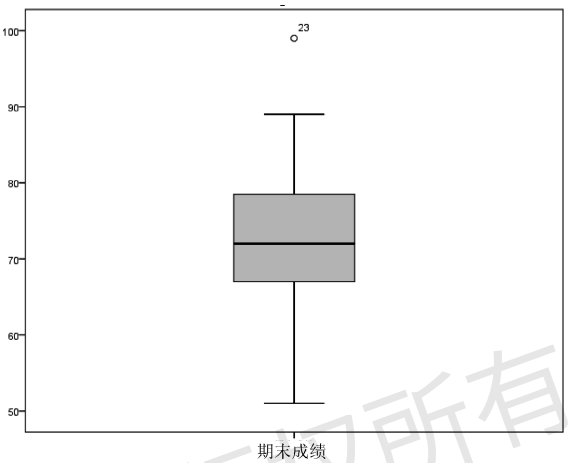

图 3-28 60 名学生“统计学”课程的期末考试成绩频数分布盒形图

六、运用 SPSS 制作茎叶图

根据 60 名学生“统计学”课程的期末考试成绩数据制作茎叶图的主要操作步骤如下。

（1）打开表 3-5 对应的 SPSS 数据集“data3.5”。在 SPSS 菜单栏中选择【Analyze】→【Descriptive Statistics】→【Explore...】菜单命令，系统弹出如图 3-29 所示的“Explore”对话框。

（2）选择变量“期末成绩 [qmcj]”，单击第一个按钮，将其移到“Dependent List:”列表框中。在“Display”栏内选中“Plots”单选按钮，单击【Plots...】按钮，系统弹出如图 3-30 所示的“Explore: Plots”对话框。

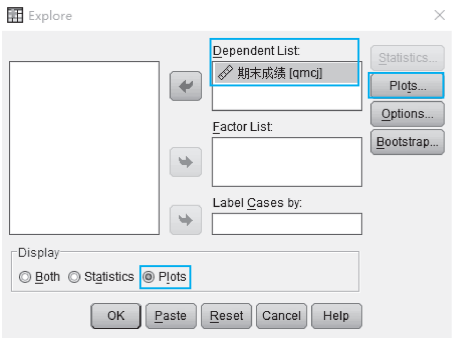

图 3-29 “Explore”对话框

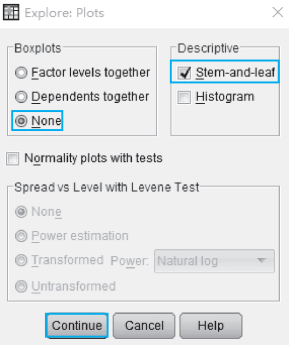

图 3-30 “Explore：Plots”对话框

（3）在“Boxplots”栏内选中“None”单选按钮，在“Descriptive”栏内选中“Stem-and-leaf”复选框，依次单击【Continue】和【OK】按钮。系统输出 60 名学生“统计学”课程的期末考试成绩频数分布茎叶图，如图 3-31 所示。

```
期末成绩  Stem-and-Leaf  Plot

 Frequency       Stem  &    Leaf

       1.00        5  .     1
       2.00        5  .     99
       6.00        6  .     022344
      15.00        6  .     555667777888999
      14.00        7  .     01122223334444
       9.00        7  .     567778899
       7.00        8  .     0012334
       5.00        8  .     55679
       1.00  Extremes       (>=99)

 Stem  width:              10
 Each  leaf:          1  case(s)
```

图 3-31　60 名学生“统计学”课程的期末考试成绩频数分布茎叶图

七、运用 SPSS 制作交叉频数分布表和交叉频数分布图

针对表 3-8 中四种品牌手机在学生手机市场占有情况的调查数据制作交叉频数分布表和交叉频数分布图的主要操作步骤如下。

（1）打开表 3-8 对应的 SPSS 数据集“datd3.8”。在 SPSS 菜单栏中选择【Data】→【Weight Cases】菜单命令，系统弹出如图 3-32 所示的“Weight Cases”对话框。在此对话中选中“Weight cases by”单选按钮，并选择变量“交叉频数 [f]”，单击按钮，将其移到“Frequency Variable”列表框中，然后单击【OK】按钮。此项操作是一个赋权过程，适用于已分组数据集。如果是未分组的数据集，则无须此项操作。

（2）在 SPSS 菜单栏中选择【Analyze】→【Descriptive Statisties】→【Crosstabs...】菜单命令，系统弹出如图 3-33 所示的“Crosstabs”对话框。

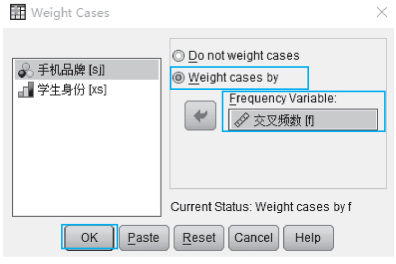

图 3-32　“Weight Cases”对话框

图 3-33　“Crosstabs”对话框

（3）选择变量“手机品牌 [sj]”，单击第一个按钮，将其移到“Row(s)：”列表框中；再选择变量“学生身份 [xs]”，单击第二个按钮，将其移到“Column(s)：”列表框中。系统输出如图 3-34 所示的交叉频数分布表和如图 3-35 所示的交叉频数分布图。

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
手机品牌 * 学生身份	354	100.0%	0	.0%	354	100.0%

手机品牌 * 学生身份 Crosstabulation

Count

		学生身份			Total
		初中生	高中生	大学生	
手机品牌	诺基亚	27	29	33	89
	摩托罗拉	18	43	51	112
	爱立信	38	15	24	77
	三星	37	21	18	76
Total		120	108	126	354

图 3-34　交叉频数分布表

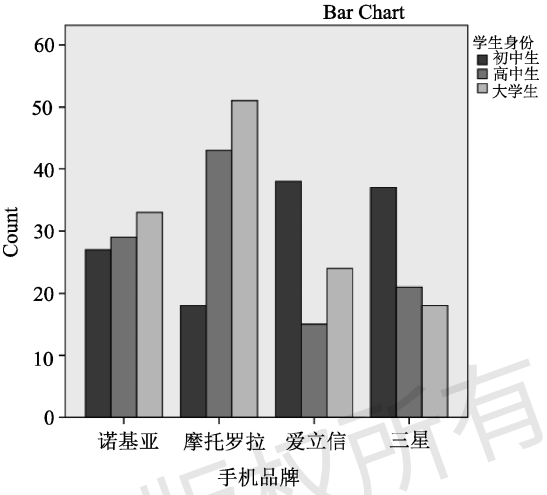

图 3-35　交叉频数分布图

八、运用 SPSS 制作散点图

根据表 3-9 中的某地区近 10 年新增住宅面积与家具销售额数据制作散点图的主要操作步骤如下。

（1）打开表 3-9 对应的 SPSS 数据集“data3.9”。在 SPSS 菜单栏中选择【Graphs】→【Legacy Dialogs】→【Scatter/Dot...】菜单命令，系统弹出如图 3-36 所示的“Scatter/Dot”对话框。

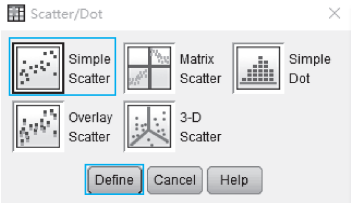

图 3-36　“Scatter/Dot”对话框

（2）选中“Simple Scatter”选项，并单击【Define】按钮，系统弹出如图 3-37 所示的“Simple Scatterplot”对话框。

（3）选择变量“新增住宅面积 [x]”，单击第二个按钮，将其移到“X Axis：”列表框中，再选择变量“家具销售额 [y]”，单击第一个按钮，将其移到“Y Axis：”列表框中，单击【OK】按钮，系统输出新增住宅面积与家具销售额散点图，如图 3-38 所示。

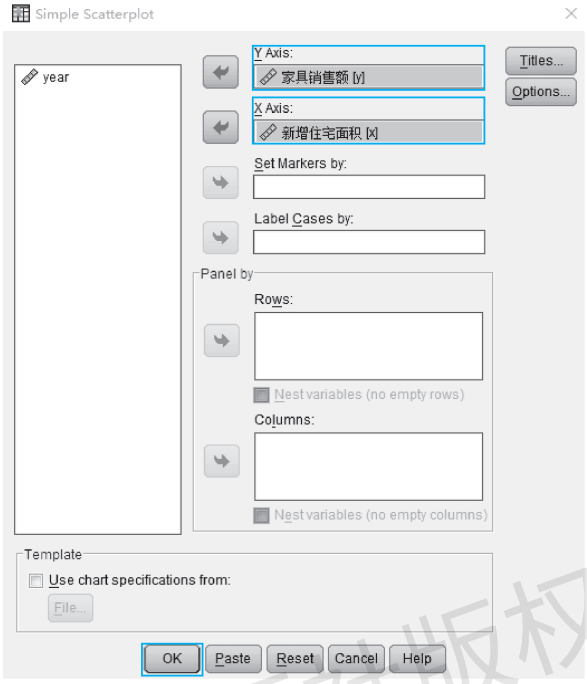

图 3-37　“Simple Scatterplot”对话框

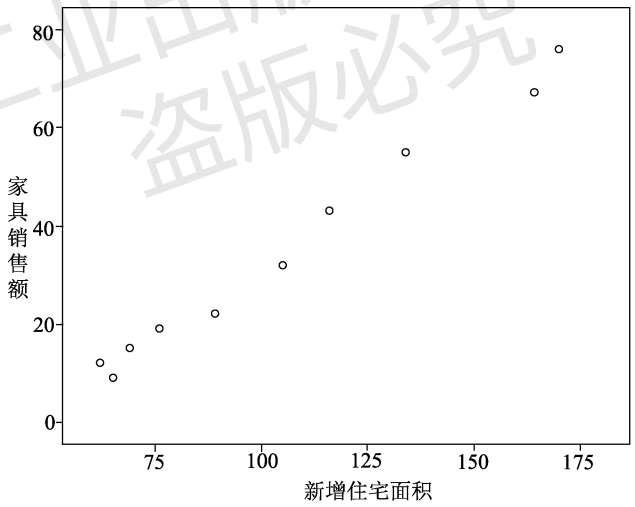

图 3-38　新增住宅面积与家具销售额散点图

本章小结

对于品质型变量的样本数据，常用的频数分布表是单项式频数分布表，常用的频数分布图主要有条形图和饼形图。

对于数值型变量的样本数据，常用的频数分布表是组距式频数分布表，常用的频数分布图主要包括直方图、盒形图和茎叶图。

实践中，常见的频数分布可分为钟形分布、U 形分布和 J 形分布三种基本类型。

交叉频数分布图适用于两个品质型变量关系的描述，散点图适用于两个数值型变量关系的描述。

思考与练习

一、单项选择题

1．以下类型的频数分布图中，（　　）可用于描述品质型数据。

A．直方图　　B．条形图

C．盒形图　　D．茎叶图

2．在组距式频数分布表中，下限与上限之间的中点称作该组的（　　）。

A．组距　　B．频数

C．组限　　D．组中值

3．直方图的纵轴通常表示各组的（　　）。

A．组距　　B．频数

C．组限　　D．组数

4．在盒形图中，方盒的右侧边界对应（　　）。

A．最小观测值　　B．下四分位数

C．中位数　　D．上四分位数

5．（　　）适用于描述两个数值型变量之间的关系。

A．条形图　　B．盒形图

C．散点图　　D．茎叶图

二、简答题

1．在组距式频数分布表中，如何确定组限？

2．频数分布有哪些主要类型？各种类型的特征是什么？

三、综合题

1．为评价家电行业售后服务质量，随机抽取了一个由 100 个家庭所构成的样本进行调查。服务质量的等级分别表示为：A．好；B．较好；C．一般；D．较差；E．差。调查结果如表 3-10 所示。

表 3-10　家电行业售后服务质量评价调查表

B	E	C	C	A	D	C	B	A	E
D	A	C	B	C	D	E	C	E	E
A	D	B	C	C	A	E	D	C	B
B	A	C	D	E	A	B	D	D	C
C	B	C	E	D	B	C	C	B	C
D	A	C	B	C	D	E	C	E	B
B	E	C	C	A	D	C	B	A	E
B	A	C	D	E	A	B	D	D	C
A	D	B	C	C	A	E	D	C	B
C	B	C	E	D	B	C	C	B	C

要求：

（1）编制服务质量等级频数分布表。

（2）绘制服务质量等级频数分布条形图和饼形图。

2．为确定一批灯泡的使用寿命，从中随机抽取了 100 只进行测试，测试结果如表 3-11 所示。

表 3-11　100 只灯泡使用寿命测试结果　　单位：小时

700	716	728	719	685	709	691	684	705	718
706	715	712	722	691	708	690	692	707	701
708	729	694	681	695	685	706	661	735	665
668	710	693	697	674	658	698	666	696	698
706	692	691	747	699	682	698	700	710	722
694	690	736	689	696	651	673	749	708	727
688	689	683	685	702	741	698	713	676	702
701	671	718	707	683	717	733	712	683	692
693	697	664	681	721	720	677	679	695	691
713	699	725	726	704	729	703	696	717	688

要求：

（1）编制 100 只灯泡使用寿命组距式频数分布表。

（2）绘制 100 只灯泡使用寿命频数分布直方图、盒形图与茎叶图。

3．A、B 两班共 80 名学生的数学考试成绩如表 3-12 所示。

表 3-12　A、B 两班学生的数学考试成绩（满分为 150 分）

A 班	141	150	141	149	118	110	120	145	148	103
	141	149	123	128	135	143	150	120	128	106
	150	131	132	146	128	132	127	144	112	130
	119	133	110	137	132	126	105	142	129	119
B 班	149	148	140	146	134	138	136	142	141	125
	123	126	119	119	137	120	128	117	116	71
	141	130	122	121	131	120	132	121	112	74
	137	128	128	136	118	136	129	113	108	82

要求：

（1）绘制 80 名学生的数学考试成绩频数分布盒形图。

（2）绘制 A、B 两班各 40 名学生的数学考试成绩频数分布盒形图。

4．睡觉打鼾与心脏病是否有关联？一次调查所获得的数据如表 3-13 所示。

表 3-13　打鼾与心脏病关联调查表

	从不打鼾	每晚打鼾	总计
有心脏病	24	30	54
没有心脏病	1 355	224	1 579
总计	1 379	254	1 633

要求：

（1）根据上述数据绘制交叉频数分布图。

（2）分析打鼾与心脏病之间的关联性。

5．在某大学，“微积分”是“统计学”课程的先修科目。抽取 15 名学生作为样本，并记录每个学生的“微积分”和“统计学”成绩，其结果如表 3-14 所示。

表 3-14　15 名学生的“微积分”和“统计学”成绩

微积分	65	58	93	68	74	81	58	85	88	75	63	79	80	54	72
统计学	74	72	84	71	68	85	63	73	79	65	62	71	74	68	73

要求：

（1）根据上述数据绘制散点图。

（2）通过散点图分析“微积分”与“统计学”成绩之间的关系。