CHAPTER 3 第3章 描述数据的图表方法

在我国加入WTO以后，关税逐渐降低，非关税壁垒也在逐渐消除，我国经济正在快速地融入世界经济之中。在这个过程中，我国企业在境外屡遭倾销指控，甚至发生多起我国某一类产品被迫退出某一市场的事件。而在国内，一些外国企业对我国实行大肆倾销以打击我国企业。于是产生一个问题，如何判断一种商品在某地以某一价格出售是否属于倾销。

当我国某一类企业投诉外国企业倾销后，商务部应收集这些相关企业的成本数据，并设立参照企业，对比判断该外国企业是否属于倾销。对这些数据进行整理，并用图形方法来阐明其意义，这正是本章要研究的主题。

3.1 数据描述的意义及分类

我们收集数据的意义在于对其进行处理以得到其中的含义，使之符合我们的需要。对数据的处理有多种方法，本章我们将论述如何绘制各种图表，以表达数据的性质。在对数据进行处理时，首先要弄清数据的类型，对于不同种类的数据的处理方法是不同的。

我们所处理的数据分为两类：定量数据和定性数据，用数值来表现观察值，称为定量数据；只能归入某一类而不能用数值进行测度的数据，称为定性数据。由定性数据反映的变量有定类变量和定序变量；由定量数据反映的变量有数字变量，相关内容在第2章中已经讲述。

3.2 定量数据的图形描述

本章首先介绍如何对数字变量的定量数据值进行描述。收集统计数据之后，首先要对获取的数据进行系统化、条理化地整理，然后进行恰当的图形描述，以提取有用的信息。

3.2.1 定量数据整理

对定量数据进行统计分组是数据整理中的主要内容。根据统计研究的目的和客观现象的内在特点，按某个标志（或几个标志）把被研究的总体划分为若干个不同性质的组，称

为统计分组。统计分组的对象是总体。从分组的性质来看，分组兼有分合双重含义。对于总体而言，是“分”，即把总体分为性质相异的若干部分；而对于单位而言，又是“合”，即把性质相同的许多单位结合为一组。例如收集到了某班所有同学的英语考试成绩，为了研究需要划分高、中、低三个成绩段，每个成绩段的范围分别是 85～100，70～85，0～70，然后将每个成绩归入相应的组中。

频数分布表反映以上数据整理的结果信息。将数据按其分组标志进行分组的过程，就是频数分布或频率分布形成的过程。表示各组的单位的次数称为频数；各组次数与总次数之比为频率；频数分布则是观察值按其分组标志分配在各组内的次数，由分组标志序列和各组对应的分布次数两个要素构成。在对这些定量数据进行分组时，需要建立频数分布表，以便更有效地显示数据的特征和分布。

对于某一变量，其定量数据为离散值且数据个数较少时，可以把每一个定量数据值作为一组，整理出每一组（即每一个定量数据值）出现的频数，从而得到频数分布表。在实际情况下，所分析的定量数据很多都是连续型的，即使是离散型的数据往往也是数据个数较多，在这些情况下就无法像上面那样简单地得到频数分布表了。因此，下面具体介绍在定量数据为连续值或定量数据为离散值但个数较多情况下，编制频数分布表的过程。

（1）选择组数。所处理的数据集合分多少组合适？这取决于数据本身的特点和数据量的大小。由于分组的目的之一是观察数据分布的特征，因此组数的多少应适中。若组数太少，数据的分布会过于集中，而组数太多则数据的分布就会过于分散。所以组数的确定以能够显示数据的分布特征和规律为目的。实际分组时可以参考 Sturges 的经验公式 $K=1+\frac{\lg n}{\lg 2}$，式中 n 为数据的个数，并根据数据量的大小、数据的特点及分析的要求来决定 K。

（2）确定各组的宽度。一个组的最小值称为下限（lower limit），最大值称为上限（upper limit）。组的宽度即为其上限和下限之差，可根据全部数据的最大值和最小值及所分的组数来确定，即

$$\text{组宽度}=(\text{最大值}-\text{最小值})/\text{区间数}$$

（3）根据确定好的组宽度，计算出组界，即上限和下限。

（4）计算组中值，组中值反映了该组数据的一个代表值，它体现各组数据的一般水平，组中值＝（该组下限值－该组上限值）/2。

（5）根据分组整理成频数分布表。计算出每个组的频数，即算出落入每个组中的观察值数，同时得到与组频数对应的组相对频数，组相对频数＝类频数/观察值总数。为了统计分析的需要，有时要观察某一数值以上或某一数值以下频数或频率之和，这就需要在频数分布表基本分组的基础上绘出累积频数或累积频率。由表的上方向表的下方的频数

或频率相加就称为“向下累积”，反之称为“向上累积”。

例 3.1　某电商平台 2015 年 4 月份至 7 月份的销售收入数据如下(单位：万元)。

492	494	355	463	434	406	392	512	478	420	453	377
409	444	386	382	407	432	436	489	376	494	355	382
470	399	379	433	447	398	372	469	494	444	386	416
446	439	388	413	366	405	451	354	424	399	381	512
394	463	427	339	377	375	446	422	325	440	388	399
481	388	373	445	382	507	347	409	436	463	427	473
372	443	388	527	416	469	376	390	364	388	373	406
407	421	510	471	512	424	428	454	417	443	388	432
398	375	493	421	399	371	366	407	392	421	510	398
409	453	379	466	473	416	389	409	325	375	493	405

描述这些销售收入数据的频数分布，编制出包含累积频数分布的完整的频数分布表，如表 3-1 所示。

表 3-1　销售收入频数分布表

按销售收入分组/万元	组中值/万元	组距/万元	频数	百分数/%	向下累计频数	向下累计百分数/%
290～320	305	30	0	0	0	0
320～350	335	30	4	3.3	4	3.3
350～380	365	30	20	16.7	24	20.0
380～410	395	30	35	29.2	59	49.2
410～440	425	30	23	19.2	82	68.4
440～470	455	30	19	15.8	101	84.2
470～500	485	30	12	10.0	113	94.2
500～530	515	30	7	5.8	120	100.0

3.2.2　单变量定量数据的图形描述

将定量数据整理成频数分布形式后，已经可以初步看出数据的一些规律了。例如，从表 3-1 就可以大致看出该电商平台 120 天中每天的销售收入大多在 350 万元～470 万元，其中以 380 万元～410 万元居多，低于该营业额的销售天数比例占 20.0%，高于该营业额的销售天数占 50.8%，因而该电商平台在 2015 年 4—7 月的销售收入所呈现的是一种非对称的分布，如果通过一些图形来表示这个分布会更形象直观。下面介绍最常用的图形表示方法：直方图、折线图、累积折线图、茎叶图、箱线图。

1. 直方图、折线图和累积折线图

直方图是用来描述定量数据集最普及的图形方法，它将频数分布表的信息以图形的

方式表达出来。直方图是用矩形的高度和宽度来表示频数分布的图形。在直角坐标系中以横轴表示所分的组，纵轴表示频数或频率，因此直方图可分为频数直方图和相对频数直方图。在得到频数分布表的基础上，绘制直方图的过程很简单，即在平面直角坐标系上，将分组标志作为横轴并将各组频数（或频率）作为纵轴，给出各组的长方形图，即得到直方图。

折线图也称频数多边形图，其作用与直方图相似。以直方图中各组标志值中点位置作为该组标志的代表值，然后用折线将各组频数连接起来，再把原来的直方图去掉，就形成了折线图。

当组距很小并且组数很多时，所绘出的折线图就会越来越光滑，逐渐形成一条光滑的曲线，这种曲线即频数分布曲线，它反映了数据的分布规律。统计曲线在统计学中很重要，是描绘各种分布规律的有效方法。常见的频数分布曲线有正态分布曲线、偏态分布曲线、J 形分布曲线和 U 形分布曲线等。

当编制频数分布表的时候，常会根据实际需要计算每组数据的累积频数或频率，累积折线图正是用来描述累积频数信息的。

根据表 3-1，绘制出该电商平台 2015 年 4—7 月 4 个月销售收入的直方图、折线图和累积折线图，如图 3-1～图 3-3 所示。

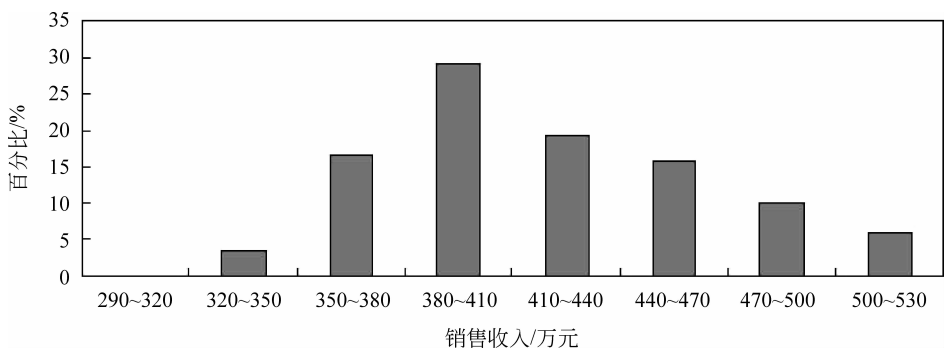

图 3-1　销售收入直方图

直方图与折线图表示了相同的销售收入分布规律。图 3-1 中显示的数据分布有点向左边倾斜的趋势，因此这里的数据分布是有些右偏的。

2. 茎叶图与箱线图

前面介绍的对数据进行图形描述一般是先将数据分组，得到频数分布表，然后将分布表中的信息画成直方图或折线图，以观察数据的分布规律。这种传统数据整理方法的局限性表现为整理后就损失了原始数据的信息。因此国外在 20 世纪 70 年代末出现了探索

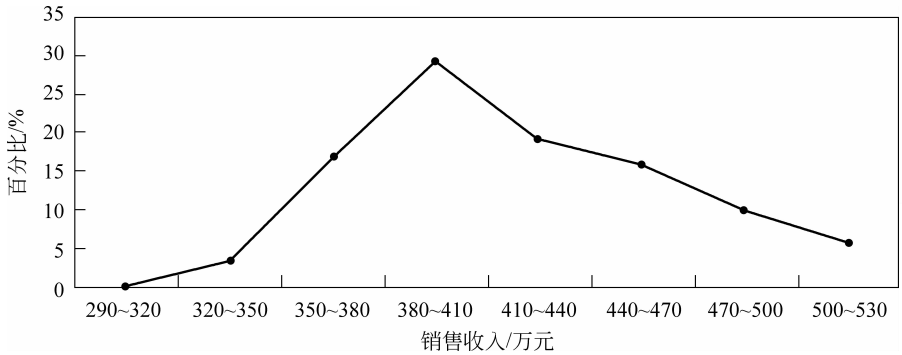

图 3-2　销售收入折线图

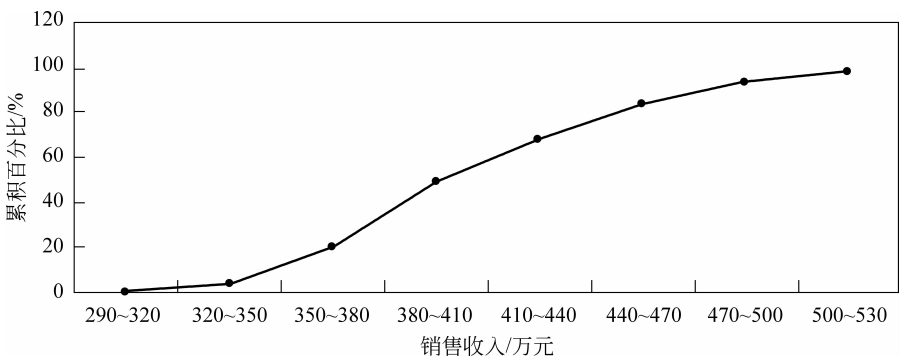

图 3-3　销售收入累积折线图

性数据分析统计，即直接描述和分析未分组的原始数据，直观地描述原始数据的分布特征，并能根据数据的特点选择适当的分析工具探索数据的内在数量规律，这样有助于用户思考对数据进一步分析的方案。茎叶图和箱线图是探索性数据分析中最简单的图形，SPSS 中有专门用于探索性数据分析统计的功能，因此在 SPSS 中绘制茎叶图和箱线图很方便。

茎叶图将传统的统计分组与画直方图两步工作一次完成，既保留了数据的原始信息，又为准确计算均值等提供了方便和可能。通过茎叶图可以看出数据的分布形状及数据的离散状况。比如分布是否对称，数据是否集中，是否有极端值等。在茎叶图画好后，不仅可以一目了然地看出频数分布的形状，而且茎叶图中还保留了原始数据的信息。利用茎叶图进行分组还有一个好处，就是在连续数据的分组中，不会出现重复分组的可能性。

我们还可以用箱线图对未分组的原始数据描述其分布特征。当只有一组数据时，可以绘制单个箱线图来进行描述。当有多组数据需要处理时，可绘制组箱线图（组箱线图将

在下一节中介绍)。从箱线图我们不仅可看出一组数据的分布特征,还可以进行多组数据分布特征之间的比较。

箱线图由一个长方形“箱子”和两段线段组成,其中长方形中部某处被一线段隔开。因此,要绘制一个箱线图,需要确定五个点,从左向右依次为这一组数据的最小值、下四分位数、中位数、上四分位数、最大值。首先我们将这一组数据按大小进行排序,其中排序后处在中间位置的变量值称为中位数,如果数据有 $2n+1$ 个,则中位数恰好是第 $n+1$ 个数据;如果数据有 $2n$ 个,则中位数为第 n 个数和第 $n+1$ 个数的均值。同理可得下四分位数和上四分位数。下四分位数是处在排序数据 25%位置的值,上四分位数是处在排序数据 75%位置的值。连接两个四分位数画出长方形“箱子”,再将两个极值点与箱子相连接。一般形式如图 3-4 所示。

图 3-4 单个箱线图

根据例 3.1 的原始数据,在 SPSS 中得到其茎叶图和箱线图的过程如下。

在 SPSS 中进行如下操作:Analysis→Descriptive Statistics→Explore,进入“Explore”定义框,如图 3-5 所示,将“销售收入”放入“Dependent List”中,表示要探索这个变量中的数据,然后单击“OK”。最后出现茎叶图和箱线图,分别如图 3-6 和图 3-7 所示。

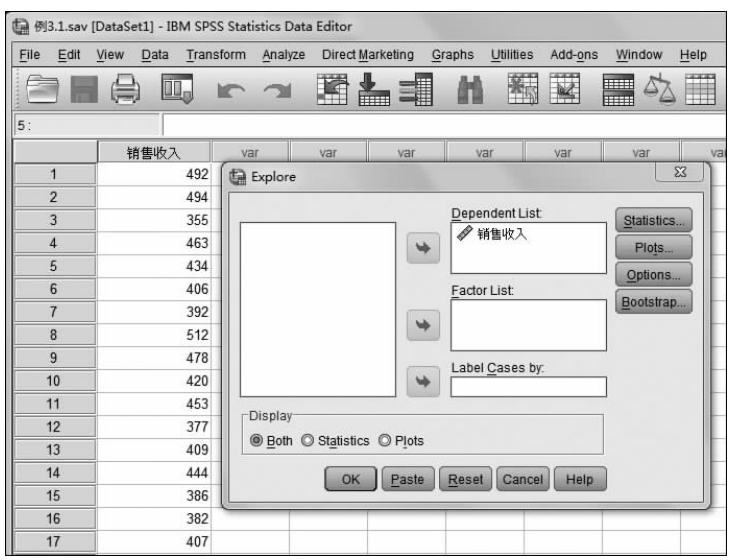

图 3-5 Explore 定义框

销售收入　Stem-and-Leaf Plot

Frequency	Stem &	Leaf
3.00	3 .	223
4.00	3 .	4555
18.00	3 .	666777777777777777
23.00	3 .	88888888888999999999
16.00	4 .	0000000000011111
15.00	4 .	222222222233333
15.00	4 .	444444444445555
11.00	4 .	66666677777
8.00	4 .	88999999
6.00	5 .	011111
1.00	5 .	2

Stem width: 100
Each leaf: 1 case(s)

图 3-6　销售收入的茎叶图

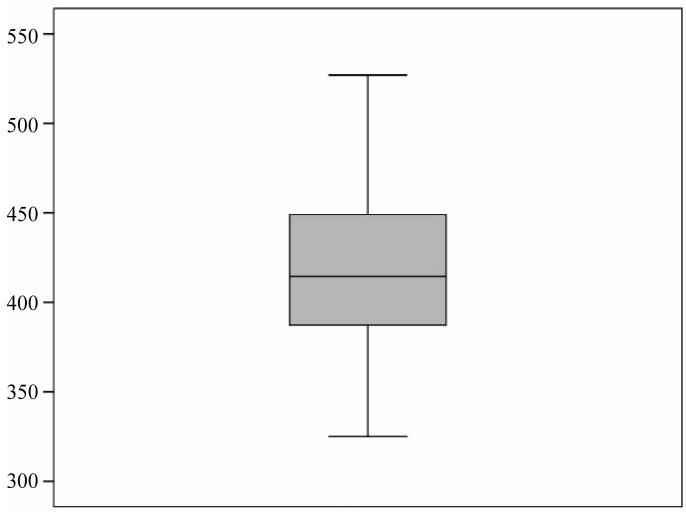

图 3-7　销售收入的箱线图

从图 3-6 可知该茎叶图中的茎宽度为 100。因为这里所描述的变量值——销售收入都是三位数的，所以 Stem(树茎)表示销售收入值得百位数，而 Leaf(树叶)则表示销售收入值的十位数。如茎叶图的第一行内容显示 Frequency 等于 3，Stem 为 3，Leaf 分别为 2，2，3，这就表示销售收入在 320 多万元的记录有 2 条，销售收入在 330 多万元的记录有 1 条，可以通过查看原始数据来验证这个结果，根据原始数据可以知道销售收入在 320 多万元和 330 多万元的具体值分别为 325，325，339。将 120 天的销售收入全部分成树茎和树

叶两部分，按照每一数字百位数和十位数的数值分别选定树茎并写上树叶，就得到了完整的茎叶图。

在画图时，要注意树叶竖行要对齐，这样，树叶的个数是各组的频数。当我们将图画好后，不难看出这就是一个放倒了的直方图，各树茎上树叶的个数就是各组的频数。从图中可以大致了解销售收入的分布是有些右偏的。

图 3-7 是销售收入的箱线图，图中“箱子”中那条黑色的线即为销售收入的中位数（值为 414.5），最外的上下两条线分别代表了销售收入的最大值（值为 527）和最小值（值为 325），箱子的上下两条线分别为下四分位数（值为 386.5）和上四分位数（值为 450）。从图中可见，中位数 414.5 离下四分位数 386.5 的距离更近，因此销售收入反映的数据分布有点右偏，这和由茎叶图中得到的结果相一致。

3.2.3 多变量定量数据的图形描述

以上我们对如何以图表描述单变量的定量数据进行了讨论，而实际上往往只对一个变量进行数据分析是不能满足研究目的的，通常把多个变量放在一起来描述，并进行分析比较。本章主要介绍四种比较常见的多变量定量数据的图形表示方法：散点图、线图、组箱线图和雷达图。

1. 散点图

在我们的生活和工作中，有许多现象和原因之间呈规则性或不规则性的关联。因此我们往往需要同时处理多个变量的定量数据，以揭示它们之间的关系。

在讨论两个变量的关系时，首先可以对其定义分类。当一个变量可以视为另一个变量的函数时，称为相关变量，通常也称为反应变量。当一个变量对另一个变量有影响时，称为独立变量或解释变量，通常它是可控的。

散点图是描述两个数字变量之间关系的图形方法。在绘制散点图时，独立变量或解释变量应放置在 X 轴上，相关变量或反应变量应放置在 Y 轴上。

例 3.2 北京市某高校 12 名女大学生的体重和肺活量的数据如表 3-2 所示。

表 3-2 北京市某高校 12 名女大学生体重和肺活量

女大学生	体重/kg	肺活量/L	女大学生	体重/kg	肺活量/L
1	42	2.20	7	50	3.10
2	42	2.55	8	50	3.40
3	46	2.40	9	52	2.90
4	46	2.75	10	52	3.40
5	46	2.85	11	58	3.00
6	50	2.80	12	58	3.50

根据表 3-2 中的数据在 SPSS 中绘制散点图。

在 SPSS 菜单项中执行如下步骤：Graphs→Legacy Dialog→Scatter/Dot，进入“Scatter/Dot”对话框，如图 3-8 所示，依次选择“Simple Scatter”（简单分布）和“Define”，进入简单散点图定义框。

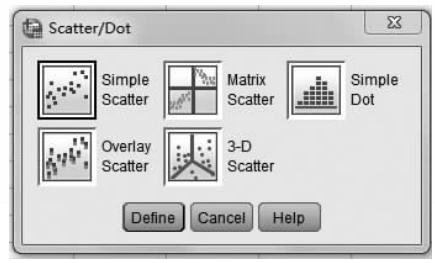

图 3-8　进入散点图

在定义框中分别选择“体重/kg”和“肺活量/L”作为 X Axis 和 Y Axis 上的值，如图 3-9所示，单击“OK”，即可得到散点图，如图 3-10 所示。

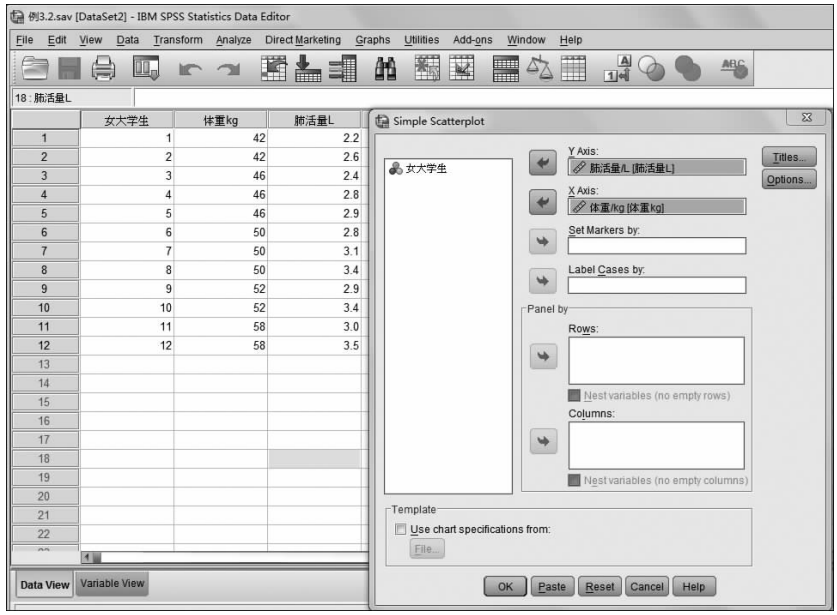

图 3-9　定义散点图

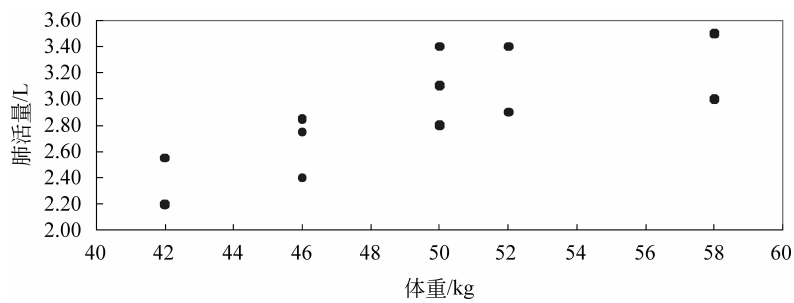

图 3-10　体重与肺活量之间的关系

2. 线图

如果数据是在不同时点取得的，称为时间序列数据，这时还可绘制线图和面积图。

线图是在平面坐标系中用折线表示数量变化特征和规律的统计图，主要用于描述时间序列数据，以反映事物发展变化的趋势。

例 3.3 某智能手机制造公司为了进入某市市场，对该市居民 2005—2014 年的智能手机消费状况进行了市场调研，其中智能手机的平均消费支出一项的数据如表 3-3 所示。

表 3-3 2005—2014 年某市居民的智能手机消费支出

年份	智能手机消费支出/(元/人·年)	年份	智能手机消费支出/(元/人·年)
2005	5.3	2010	30.5
2006	8.1	2011	49.5
2007	12.5	2012	72.5
2008	15.1	2013	100.4
2009	17.4	2014	122.6

根据表 3-3 中的数据在 SPSS 中绘制线图。执行如下步骤：Graphs→Legacy Dialog→Line，弹出“Line Charts”折线图对话窗口。在折线图对话窗口，如图 3-11 所示，依次点击“Simple”，“Summaries for group of cases”(个案组摘要)，“Define”，弹出定义简单线图窗口，如图 3-12 所示。在“Line Represents”栏中选择“Other statistic(e.g.,mean)”，然后选择“智能手机消费支出”放入“Variable”，“年份”放入“Category Axis”，然后点击“OK”，得到简单线图，如图 3-13 所示。

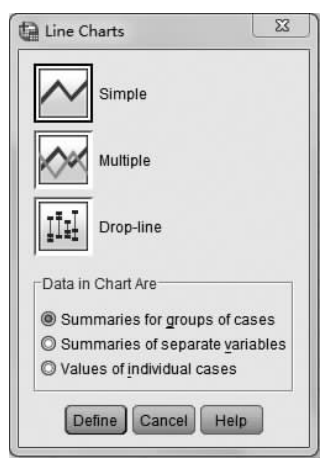

图 3-11 进入线图

从图 3-13 中可以看出，某市居民的智能手机消费支出逐年提高，尤其是 2010—2014 年的提高速度日益增长。

3. 组箱线图

对于多组数据，我们可以依据同样的方法来绘制箱线图，然后将各组数据的箱线图并列起来，以比较其分布特征。这里多组数据可以出自同一总体的不同组样本数据，或来自不同总体的不同组样本数据。

例 3.4 根据例 3.1 的数据，将该电商平台 4 月份到 7 月份的销售收入数据经排序后结果如表 3-4 所示。