第二讲：一般数据分析

教学目的：能应用 SPSS 软件进行：描述分析、频数分析、数据探索、交叉表分析、图形分析等

教学内容：1）描述分析

2）频数分析

3）数据探索

4）交叉表分析

教学重点：描述分析、频数分析、交叉表

教学难点：数据探索、交叉表分析

教学时间：1 学时

描述性统计分析 Descriptive Statistics

描述性统计分析是统计分析的第一步，做好这第一步是下面进行正确统计推断的先决条件。SPSS 的许多模块均可完成描述性分析，但专门为该目的而设计的几个模块则集中在 Descriptive Statistics 菜单中，最常用的是列在最前面的四个过程：

●Frequencies 过程的特色是产生频数表；

●Descriptives 过程则进行一般性的统计描述；

●Explore 过程用于对数据概况不清时的探索性分析；

●Crosstabs 过程则完成计数资料和等级资料的统计描述和一般的统计检验，常用的 X2 检验也在其中完成。

1.1 Frequencies 过程

频数分布表是描述性统计中最常用的方法之一，Frequencies 过程就是专门为产生频数表而设计的。它不仅可以产生详细的频数表，还可以按要求给出某百分位点的数值，以及常用的条图、饼图等统计图。和国内常用的频数表不同，几乎所有统计软件给出的都是详细频数表，即并不按某种要求确定组段数和组距，而是按照数值精确列表。如果想用 Frequencies 过程得到熟悉的频数表，请先用第 3 章学过的 Recode 过程产生一个新变量来代表所需的各

组段。

1.1.1　界面说明

Frequencies 对话框的界面如图 1.1a 所示。选取 Analyze→Descriptive Statistics →Frequencies，系统就会弹出该对话框，其各部分的功能如下：

1. Variable（s）框：左侧的变量可全部选入右侧的 Variable（s）框内，一次性完成所有变量的频数分析；也可逐一选入右侧，进行分析 n 次分析（这样就太累了）。

2. Display frequency tables 复选框：确定是否在结果中输出频数表。

图 1.1a　Frequencies 对话框

3. Statistics：单击后弹出 Statistics 对话框如图 1.1b，用于定义需要计算的其他描述统计量。其中：

●Percentile Values 复选框组：定义需要输出的百分位数，可计算四分位数(Quartiles)、每隔指定百分位输出当前百分位数(Cut points for　equal groups)、或直接指定某个百分位数(Percentiles)，如直接指定输出 P2.5（即累计百分数为 2.5%处的变量值）和 P97.5（即累计达到 97.5%处的变量值）。

●Central tendency 复选框组：用于定义描述集中趋势的一组指标：均值(Mean)、中位数(Median)、众数(Mode)、总和(Sum)。

●Dispersion 复选框组：用于定义描述离散趋势的一组指标：标准差(Std.deviation)、方差(Variance)、全距 (Range)、最小值(Minimum)、最大值(Maximum)、标准误(S.E.mean)。

●Distribution 复选框组：用于定义描述分布特征的两个指标：偏度系数（Skewness）和峰度系数(Kurtosis)。

●Values are group midpoints 复选框：当输出的数据是分组频数数据，并且具体数值是组中值时，选中该复选框，以通知 SPSS，免得它犯错误。

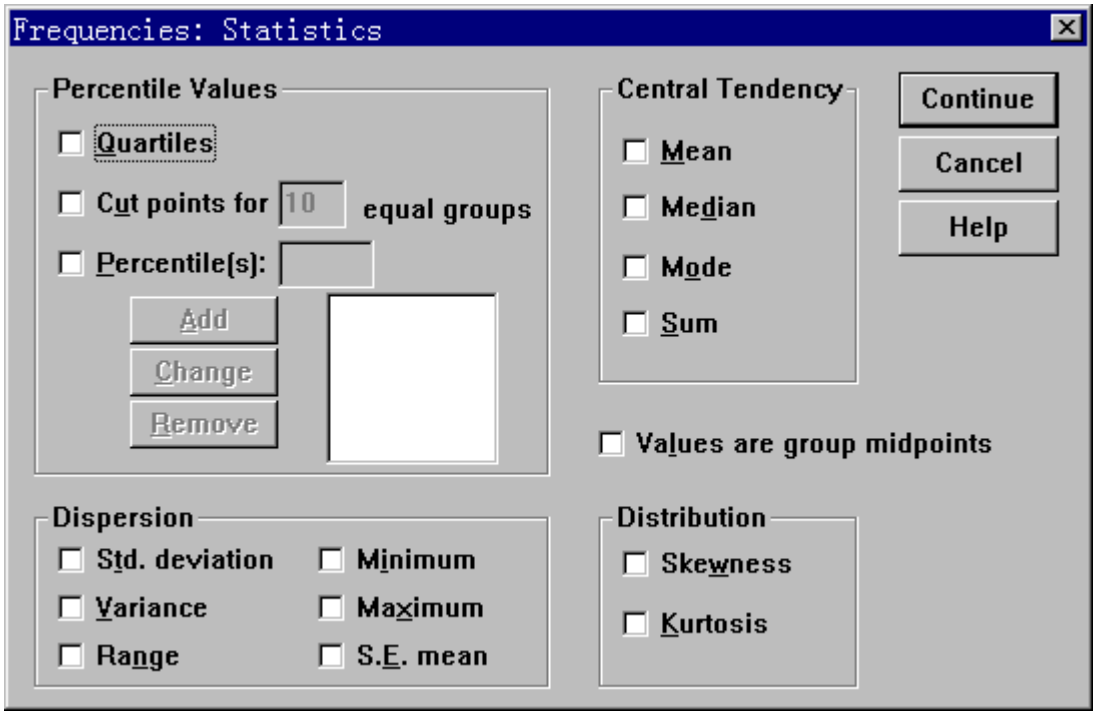


图 1.1b　Frequencies 的 Statistics 对话框

4. Charts：弹出 Charts 对话框，用于设定所做的统计图。

●Chart type 单选钮组 定义统计图类型，有四种选择：无、条图（Bar chart）、圆图(Pie chart)、直方图 Histogram)，其中直方图还可以选择是否加上正态曲线（With normal curve)。

●Chart Values 单选钮组 定义是按照频数还是按百分比做图（即影响纵坐标刻度)。

5. Format：弹出 Format 对话框，用于定义输出频数表的格式，不过用处不大，一般不管。

●Order by 单选钮组 定义频数表的排列次序，有四个选项：Ascending values 为根据数值大小按升序从小到大作频数分布；Descending values 为根据数值大小按降序从大到小作频数分布；Ascending counts 为根据频数多少按升序从少到多作频数分布；Descending counts 为根据频数多少按降序从多到少作频数分布。

●Multiple Variables 单选钮组 如果选择了两个以上变量做频数表，则 Compare variables 可以将他们的结果在同一个频数表过程输出结果中显示，便于互相比较，Organize output by variables 则将结果在不同的频数表过程输出结果中显示。

●Suppress Tables more than...复选框 当频数表的分组数大于下面设定数值时禁止它在结果中输出，这样可以避免产生巨型表格。

1.1.2 实例分析

例 1.1 利用 111.sav 文件中 q9（即被访问者最近一次参加促销活动的消费）的调查数据，绘制频数表、直方图，计算平均值、标准差、变异系数 CV、中位数 Mode、p2.5 和 p97.5。

●求解

上述要求中，除 CV 需用手工计算外，其他问题都可通过 Frequencie 解决。其主要操作如下：

1. 从程序中打开 SPSS，选择 File→open→data，打开 111.sav；
2. Analyze→Descriptive Statistics→Frequencies，弹出 Frequencies 对话框；
3. Variables 框：选入 q9
4. 单击 Statistics
5. 选中 Mean、Std.deviation、Median 复选框
6. 单击 Percentiles：输入 2.5：单击 Add：输入 97.5：单击 Add：
7. 单击 Continue
8. 单击 Charts：
9. 选中 Bar charts
10. 单击 Continue
11. 单击 OK，系统即在 SPSS Viewer 中显示所有结果，详见结果解释。

而 CV 可用得到的 Std. deviation 与 Mean 相除求得。

●问题与处理

图 1.2 是 q9 的次数分布直方图，它表明：由于 q9 的取值点较多 ，使得按变量取值分组进行的 Frequencies 分析表很长，绘出的直方图也因分组太多而显得不清爽，需要进一步处理。可先对 q9 分组，可通过重新赋值于新变量来实现，再作直方图。

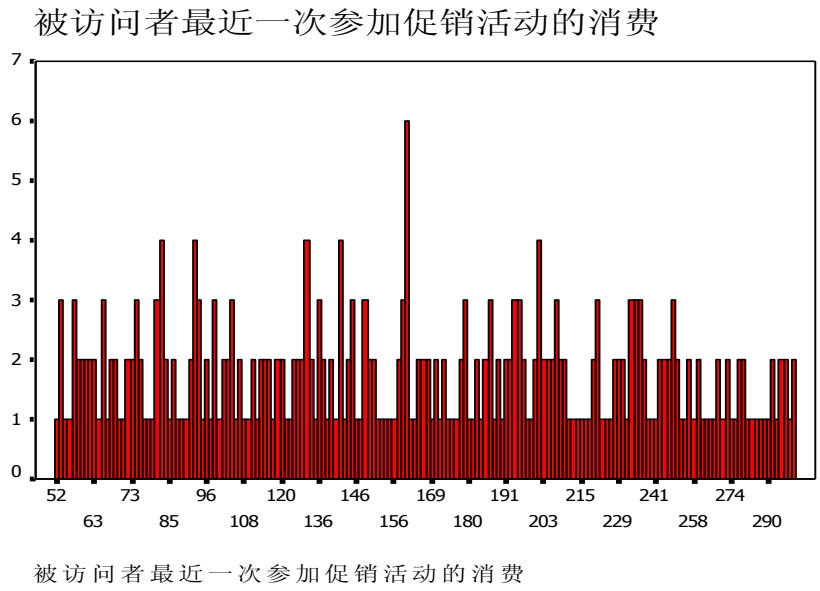


图 1.2　q9 的次数分布直方图

1.1.3　结果解释

●Statistics Table

Statistics

被访问者最近一次参加促销活动的消费

N	Valid	312
	Missing	0
Mean		114.03
Mode		113
Std. Deviation		18.158
Percentiles	2.5	58.00
	97.5	293.70

表的最上方是表名，接下来是变量 q9 的标签——被访问者最近一次参加促销活动的消费；表的左侧是统计变量名称，右侧是统计结果。表中数据显示：样本量 N 为 312 个，缺失值 0 个，平均值 Mean=114.03，中位数 Median=113，标准差 STD=18.158，P2.5=58，P97.5=293.7。

●Frequencies Table

被访问者最近一次参加促销活动的消费

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	52	1	.3	.3	.3
	53	3	.9	1.0	1.3
	54	1	.3	.3	1.6
	55	1	.3	.3	1.9
	58	3	.9	1.0	2.9
	59	2	.6	.6	3.5
	60	2	.6	.6	4.2
	61	2	.6	.6	4.8
	62	2	.6	.6	5.4

上表是系统对变量 q9 作的频数分布表(此处只列出了开头部分),Vaild 右侧为原始值，Frequency 为频数，Percent 为各组频数占总例数的百分比（包括缺失记录在内），Valid percent 为各组频数占总例数的有效百分比，Cum Percent 为各组频数占总例数的累积百分比。

1.2 Descriptives 过程

Descriptives 过程是连续资料统计描述应用最多的一个过程，他可对变量进行描述性统计分析，计算并列出一系列相应的统计指标。这和其他过程相比并无不同。但该过程还有个特殊功能就是可将原始数据转换成标准正态评分值并以变量的形式存入数据库供以后分析。

1.2.1 界面说明

Descriptives 对话框的界面如图 1.3a 所示。选取 Analyze→Descriptive Statistics →Descriptives，系统就会弹出该对话框，其各部分的功能如下：

●Save standardized values as variables 复选框：确定是否将原始数据的标准正态评分存为新变量。

●Options：Options 对话框（见图 1.3b）中的大部分内容均在前面 Frequences 过程的 Statistics 对话框中见过，只有最下方的 Display Order 单选钮组是新的，可以选择为变量列表顺序、字母顺序、均值升序或均值降序。

1.2.2 结果解释

利用 111.sav 文件中的 q9 数据，选择 Analyze→Descriptive Statistics→ Descriptives，在弹出的 Descriptives 对话框中选 q9 到 Variable（s）框中，点击 ok，即可得到如下一个典型的 Descriptives 过程结果统计表：

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
被访问者最近一次参加促销活动的消费	312	52	300	114.03	18.158
Valid N (listwise)	312				

表中各统计项在前面都有解释，这里就不再啰嗦了。

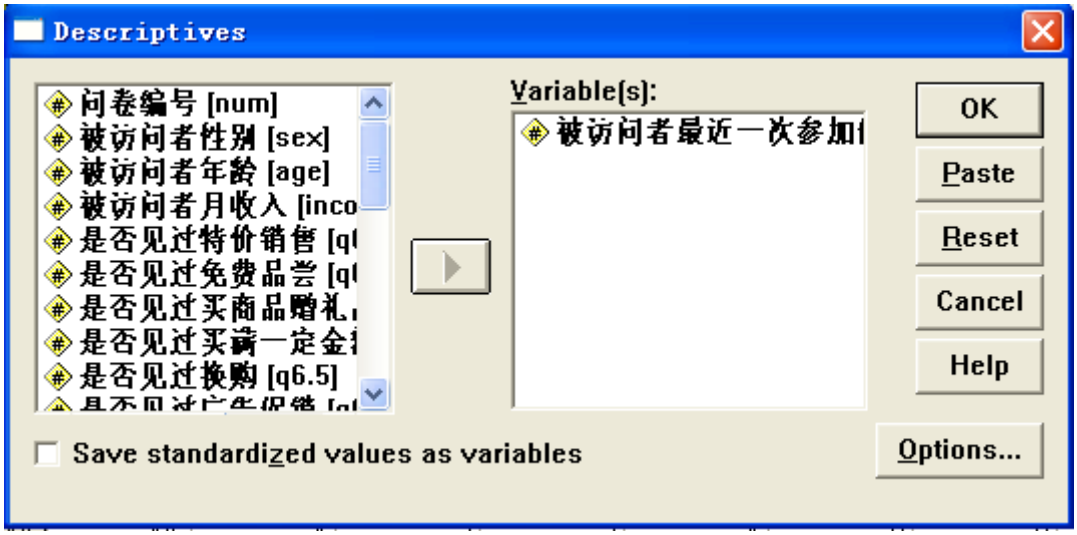


图 1.3a　Descriptives 对话框

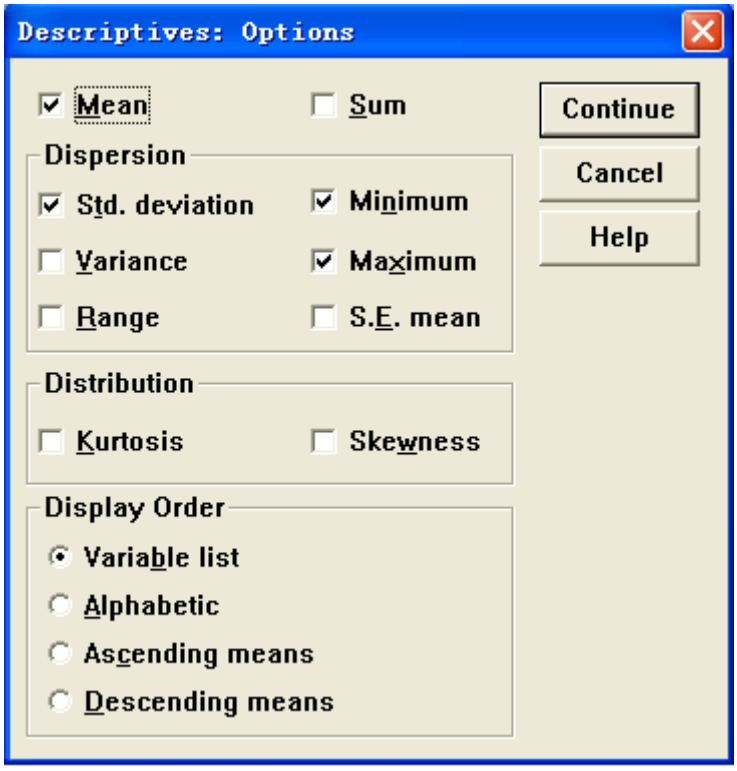


图 1.3b　Descriptives 中的 Options 对话框

1.3 Explore 过程

Explore 过程可对变量进行更为深入详尽的描述性统计分析，主要用于对资料的性质、分布特点等完全不清楚时，故又称之为探索性分析。它在一般描述性统计指标的基础上，增加有关数据其他特征的文字与图形描述，如茎叶图、箱图等，显得更加详细、全面，有助于用户制定继续分析的方案。

1.3.1　界面说明

Explore 对话框的界面如图 1.4a 所示。选取 Analyze→Descriptive Statistics→Explore，系统就会弹出该对话框，其各部分的功能如下：

图 1.4a　Explore 对话框

●Display 单选钮组：用于选择输出结果中是否包含统计描述、统计图或两者均包括。

●Dependent List 框：用于选入需要分析的变量。

●Factor List 框：如果想让所分析的变量按某种因素取值分组分析，则在这里选入分组变量。

●Label cases by 框：选择一个变量，他的取值将作为每条记录的标签。最典型的情况是使用记录 ID 号的变量。

●Statistics：弹出 Statistics 对话框（见图 1.4b），用于选择所需要的描述统计量。有如下选项：

Descriptives 复选框：输出平均值、中位数、众数、5%修正平均值、标准误、方差、标准差、最小值、最大值、全距、四分位全距、峰度系数、峰度系数的标准误、偏度系数、偏度系数的标准误及指定的均值可信区间。

M-estimators 复选框：作中心趋势的粗略最大似然确定，输出四个不同权重的最大似然确定数。

Outliers 复选框：输出五个最大值与五个最小值。

Percentiles 复选框：输出第 5%、10%、25%、50%、75%、90%、95%位数。

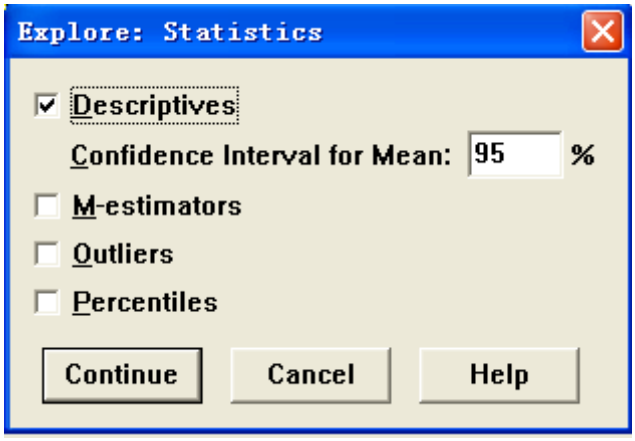


图 1.4b Explore 中的 Statistics 对话框

●Plot：弹出 Plot 对话框（见图 1.4c），用于选择所需要的统计图。有如下选项：

Boxplots 单选框组：确定箱式图的绘制方式，可以是按组别分组绘制(Factor levels together)，也可以不分组一起绘制(Depentends together)，或者不绘制(None)。

Descriptive 复选框组：可以选择绘制茎叶图(Stem-and-leaf)和直方图(Histogram)。

Normality plots with test 复选框：绘制正态分布图并进行变量是否符合正态分布的检验。

Spread vs. Level with Levene Test 单选框组：当选择了分组变量时，绘制 spread-versus-level 图，设置绘图时变量的转换方式，并进行组间方差齐性检验。

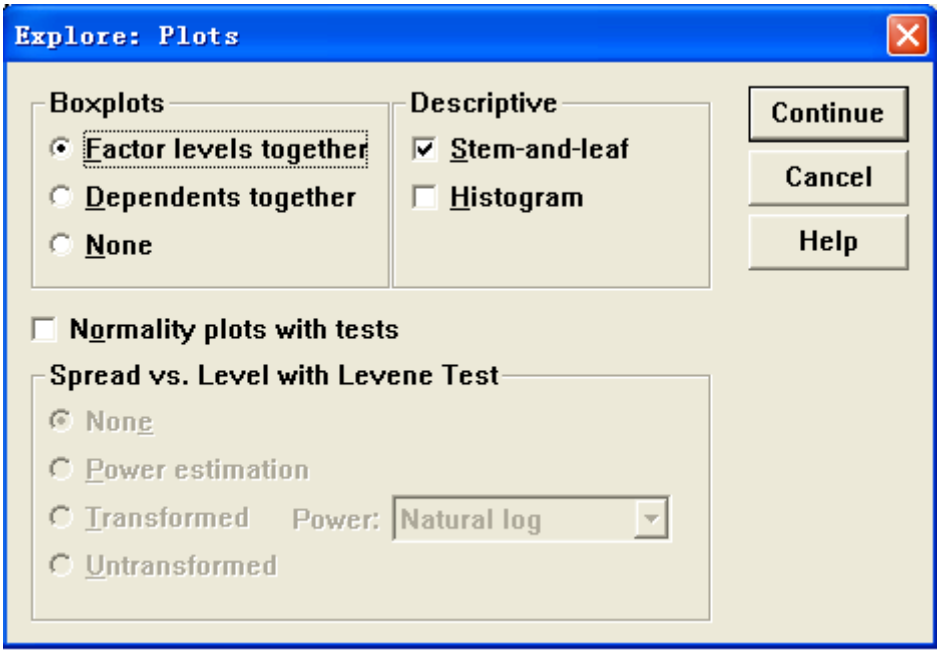


图 1.4c　Explore 中的 Plots 对话框

●Options：用于选择对缺失值的处理方式，可以是不分析有任一缺失值的记录、不分析计算某统计量时有缺失值的记录，或报告缺失值，如图 1.4d 所示。

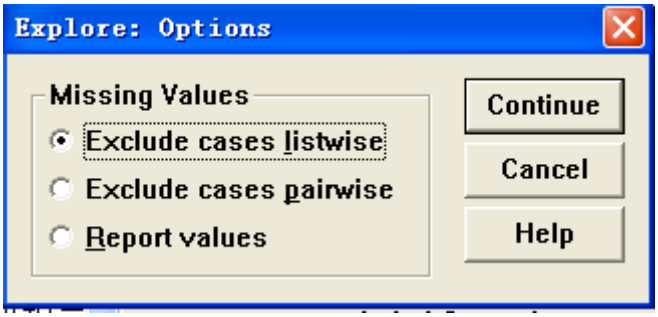


图 1.4d　Explore 中的 Options 对话框

1.3.2　结果解释

以例 1.1 的数据为例，按默认方式下的选择，Explore 过程的输出如下：

●首先是例行的处理记录缺失值情况报告，可见 312 例均为有效值。

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
被访问者最近一次参加促销活动的消费	312	100.0%	0	.0%	312	100.0%

●其次是描述统计结果，包括平均数 Mean 及其 95%的置信区间、中位数 Median、方差 Variance、标准差 Std. Deviation、偏度 Skewness、峰度 Kurtosis 等，几乎常见的描述统计量都出现了，比较全面。

Descriptives

			Statistic	Std. Error
被访问者最近一次参加促销活动的消费	Mean		114.03	3.887
	95% Confidence Interval for Mean	Lower Bound	151.38	
		Upper Bound	171.18	
	5% Trimmed Mean		112.82	
	Median		113.00	
	Variance		4713.891	
	Std. Deviation		18.158	
	Minimum		52	
	Maximum		300	
	Range		248	

	Interquartile Range	114.50	
	Skewness	.113	.138
	Kurtosis	-1.027	.275

●然后是茎叶图，整数位为茎，小数位为叶。这样可以非常直观的看出数据的分布范围及形态，在国外非常流行。

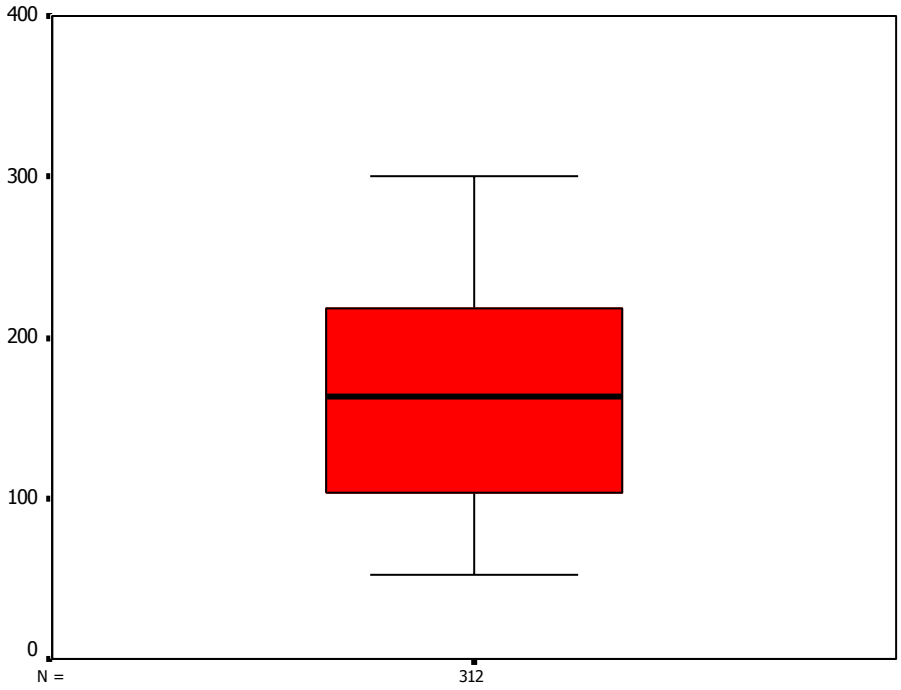
被访问者最近一次参加促销活动的消费 Stem-and-Leaf Plot

```
 Frequency    Stem &  Leaf

    11.00        0 .  55555555555
    30.00        0 .  111111111111111117777777777777
    32.00        0 .  88888888888888899999999999999999
    25.00        1 .  0000000000000111111111111
    25.00        1 .  22222222333333333333333333
    27.00        1 .  44444444444444445555555555
    31.00        1 .  111111111111111117777777777777
    30.00        1 .  88888888888889999999999999999
    25.00        2 .  00000000000000000011111111
    25.00        2 .  22222222223333333333333333
    18.00        2 .  444444444444455555
    17.00        2 .  11111111177777777
    14.00        2 .  88888999999999
     2.00        3 .  00

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

●最后还有箱式图，中间的黑粗线为均值，红框为四分位间距的范围，上下两个细线为最大、最小值。



1.4 Crosstabs 过程

Crosstabs 过程用于对计数资料和有序分类资料进行统计描述和简单的统计推断。在分析时可以产生二维至 n 维列联表，并计算相应的百分数指标。统计推断则包括了常用的 X2 检验、Kappa 值，分层 X2（X2M-H)。如果安装了相应模块，还可计算 n 维列联表的确切概率（Fisher's Exact Test）值。这里只介绍一些常用的。

1.4.1 界面说明

Crosstabs 对话框的界面如图 1.5a 所示。选取 Analyze→Descriptive Statistics→Crosstabs，系统就会弹出该对话框，其各部分的功能如下：

●Rows 框：用于选择行*列表中的行变量。

●Columns 框：用于选择行*列表中的列变量。

●Layer 框：Layer 指的是层，对话框中的许多设置都可以分层设定，在同一层中的变量使用相同的设置，而不同层中的变量分别使用各自层的设置。如果要让不同的变量做不同的分析，则将其选入 Layer 框，并用 Previous 和 Next 钮设为不同层。Layer 在这里用的比较少，在多元回归中将进行详细的解释。

●Display clustered bar charts 复选框：显示重叠条图。

●Suppress table 复选框：禁止在结果中输出行*列表。

●Exact:针对 2*2 以上的行*列表设定计算确切概率的方法,可以是不计算(Asymptotic only)、蒙特卡罗模拟（Monte Carlo）或确切计算（Exact)。蒙特卡罗模拟默认进行 10000 次模拟，给出 99%可信区间；确切计算默认计算时间限制在 5 分钟内。这些默认值均可更改。

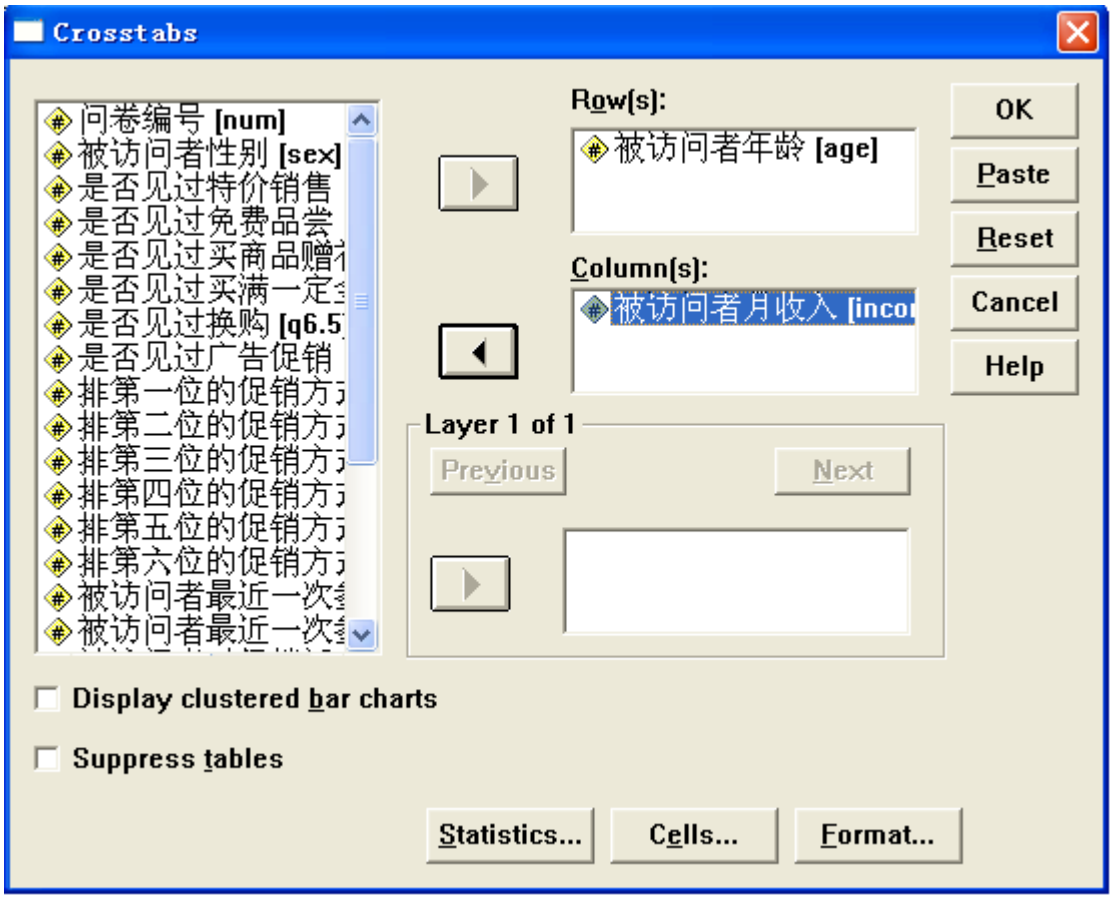


图 1.5a Crosstabs 对话框

●Statistics：弹出 Statistics 对话框，用于定义所需计算的统计量,见图 1.5b。

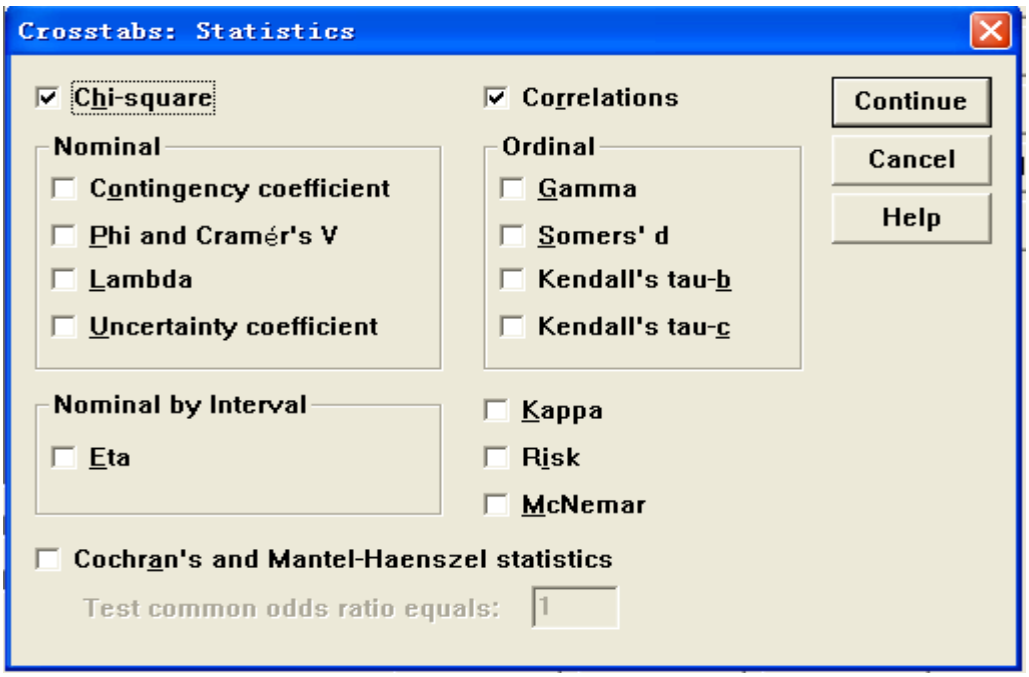


图 1.5b Crosstabs 中的 Statistics 对话框

- Chi-square 复选框：计算 X^2 值。
- Correlations 复选框：计算行、列两变量的 Pearson 相关系数和 Spearman 等级相关系数。
- Norminal 复选框组：选择是否输出反映分类资料相关性的指标，很少使用。

  a. Contingency coefficient 复选框：即列联系数，其值界于 0～1 之间；
  b. Phi and Cramer's V 复选框：这两者也是基于 X^2 值的，Phi 在四格表 X^2 检验中界于-1～1 之间，在 R*C 表 X^2 检验中界于 0～1 之间；Cramer's V 则界于 0～1 之间；
  c. Lambda 复选框：在自变量预测中用于反映比例缩减误差，其值为 1 时表明自变量预测因变量好，为 0 时表明自变量预测因变量差；
  d. Uncertainty coefficient 复选框：不确定系数，以熵为标准的比例缩减误差，其值接近 1 时表明后一变量的信息很大程度来自前一变量，其值接近 0 时表明后一变量的信息与前一变量无关。

- Ordinal 复选框组：选择是否输出反映有序分类资料相关性的指标，很少使用。

  a. Gamma 复选框：界于 0～1 之间，所有观察实际数集中于左上角和右下角时，其值为 1；
  b. Somers'd 复选框：为独立变量上不存在同分的偶对中，同序对子数超过异序对子数的比例；
  c. Kendall's tau-b 复选框：界于-1～1 之间；
  d. Kendall's tau-c 复选框：界于-1～1 之间；

- Eta 复选框：计算 Eta 值，其平方值可认为是因变量受不同因素影响所致方差的比例；
- Kappa 复选框：计算 Kappa 值，即内部一致性系数；
- Risk 复选框：计算比数比 OR 值；
- McNemanr 复选框：进行 McNemanr 检验（一种非参检验）；
- Cochran's and Mantel-Haenszel statistics 复选框：计算 X^2_{M-H} 统计量（分层 X^2，也有写为 X^2_{CMH} 的），可在下方输出 H_0 假设的 OR 值，默认为 1。

●Cells：弹出 Cells 对话框（见图 1.5c），用于定义列联表单元格中需要计算的指标：

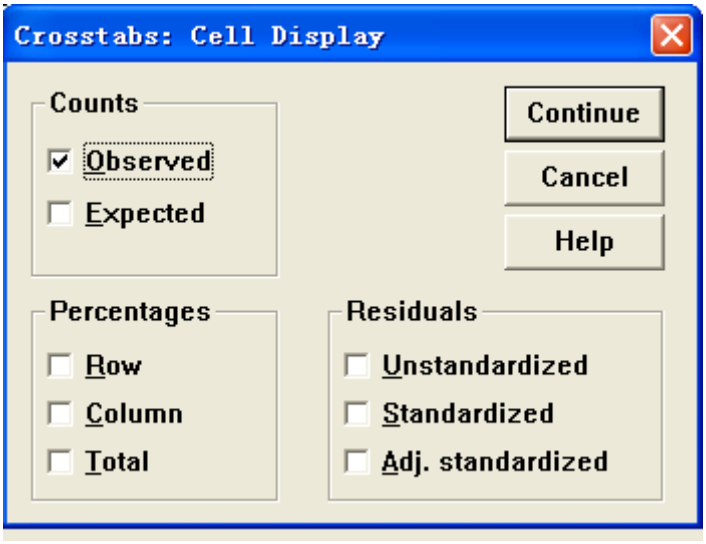


图 1.5c Crosstabs 中的 Cell Display 对话框

- Counts 复选框组：是否输出实际观察数(Observed)和理论数（Expected）；
- Percentages 复选框组：是否输出行百分数（Row）、列百分数（Column）以及合计百分数（Total）；
- Residuals 复选框组：选择残差的显示方式，可以是实际数与理论数的差值（Unstandardized）、标化后的差值（Standardized，实际数与理论数的差值除理论数），或者由标准误确立的单元格残差（Adj. Standardized）；

●Format：用于选择行变量是升序还是降序排列。

1.4.2 分析实例

例 1.2 利用 111.sav 文件中调查数据，做年龄 age 与月收入 income 的交叉分析表，并分析在“性别 sex”变量控制下的年龄与收入的关系。

这两个问题都可以通过 Crosstabs 来完成，在默认 111.sav 文件已打开时，第一个问题的操作步骤如下：

1. Analyze→Descriptive Statistics→Crosstabs
2. Rows 框：选入 age
3. Columns 框：选入 income
4. 单击 Cells：选中 Observed 下的 Counts，和 Percentage 下的 Row，单击 Continue
5. 单击 OK

第二个问题的操作步骤如下：

1. Analyze→Descriptive Statistics→Crosstabs
2. Rows 框：选入 age
3. Columns 框：选入 income
4. Layer 框：选入 sex
5. 单击 Statistics：选中 Chi-square 和 Correlation，单击 Continue

6. 单击 Cells：选中 Observed 下的 Counts，和 Percentage 下的 Row，单击 Continue
7. 单击 OK

1.4.3 结果解释

第一题的结果如下：

●首先是处理记录缺失值情况报告，可见 312 个 cases 均为有效值。

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
被访问者年龄 * 被访问者月收入	312	100.0%	0	.0%	312	100.0%

●被访问者年龄 age 与月收入 income 的交叉分析表，行是年龄分组，列是与收入分组，中间的数据是各组人数和在各年龄组中月收入的人数比重。分析表结果显示： 25 岁以下年龄组中，被访问者月收入在 1000 元以下的占 14.2%；而 25-35 岁年龄组的占 51.1%，35-45 岁组占 24.1%，45 岁以上组占 42.9%。但将收入级别调高后，35-45 岁组的人数比重都为最高，两边年龄组的则逐渐下降。这表明：随年龄变化，月收入既有先减后增的趋势，也有先增后减的趋势，而且 35-45 岁的中坚人群高月收入者明显较多，可以认为年龄越与月收入之间的关系密切。

被访问者年龄 * 被访问者月收入 Crosstabulation

			被访问者月收入				Total
			1000元以下	1000-1500元	1500-2000元	2000元以上	
被访问者年龄	25岁以下	Count	34	14	4	1	53
		% within 被访问者年龄	14.2%	21.4%	7.5%	1.9%	100.0%
	25-35岁	Count	97	58	21	7	188
		% within 被访问者年龄	51.1%	30.9%	13.8%	3.7%	100.0%
	35-45岁	Count	14	22	14	7	57
		% within 被访问者年龄	24.1%	38.1%	24.1%	12.3%	100.0%
	45岁以上	Count	1	4	3	1	14
		% within 被访问者年龄	42.9%	28.1%	21.4%	7.1%	100.0%
Total		Count	151	98	47	11	312
		% within 被访问者年龄	48.4%	31.4%	15.1%	5.1%	100.0%

第二个问题的结果如下：

●首先仍然是处理记录缺失值情况报告，312 个 cases 也都有效值。

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
被访问者年龄 * 被访问者月收入 * 被访问者性别	312	100.0%	0	.0%	312	100.0%

●在“性别 sex”变量控制下的被访问者年龄 age 与月收入 income 的交叉分析表。与第一题有所不同，这张表多了一个性别分层，可分别研究男性和女性被访问者的年龄与月收入的关系。分析表结果显示：无论是女性组还是男性组，年龄与收入的关系表现都与前面一致；但是，各性别组还是有些区别，如女性组 25 岁以下、月收入 1000 元以下的人数比重要高于男性，而 1000-2000 元的却与男性组相差无几。据此是否能判断年龄与月收入关系在女性与男性之间是有差别的？

被访问者年龄 * 被访问者月收入 * 被访问者性别 Crosstabulation

被访问者性别			被访问者月收入				Total
			1000元以下	1000-1500元	1500-2000元	2000元以上	
女	被访问者年龄	25岁以下	Count	20	8	2	0	30
			% within 被访问者年龄	11.7%	21.7%	1.7%	.0%	100.0%
		25-35岁	Count	51	31	8	0	90
			% within 被访问者年龄	51.7%	34.4%	8.9%	.0%	100.0%
		35-45岁	Count	9	13	1	3	31
			% within 被访问者年龄	29.0%	41.9%	19.4%	9.7%	100.0%
		45岁以上	Count	3	2	2	1	8
			% within 被访问者年龄	37.5%	25.0%	25.0%	12.5%	100.0%
Total			Count	83	54	18	4	159
			% within 被访问者年龄	52.2%	34.0%	11.3%	2.5%	100.0%
男	被访问者年龄	25岁以下	Count	14	1	2	1	23
			% within 被访问者年龄	10.9%	21.1%	8.7%	4.3%	100.0%

	25-35岁	Count	41	27	18	7	98
		% within 被访问者年龄	41.9%	27.1%	18.4%	7.1%	100.0%
	35-45岁	Count	5	9	8	4	21
		% within 被访问者年龄	19.2%	34.1%	30.8%	15.4%	100.0%
	45岁以上	Count	3	2	1	0	1
		% within 被访问者年龄	50.0%	33.3%	11.7%	.0%	100.0%
Total		Count	18	44	29	12	153
		% within 被访问者年龄	44.4%	28.8%	19.0%	7.8%	100.0%

●Chi-square Tests 表和 Symmetric Measures 表。在交叉分析表中发现的问题，可以通过 Symmetric Measures 表和 Chi-square 检验结果来说明。从 Symmetric Measures 表中结果可以看出：女性组的年龄与月收入的相关系数为 0.31，略呈显著正相关关系；而男性组的相关系数为 0.182，也有正相关关系，但明显弱于女性组。但是，从 Chi-Square Tests 表中可推断：女性组 Pearson Chi-Square=22.954，Asymp. Sig.=0.001<0.05；而男性组 Pearson Chi-Square=11.751，Asymp. Sig.=0.227>0.05。因此，可以拒绝系统默认的原假设——年龄与月收入没有关系，说明总体上这两者是有关系的；但是，还不能确定女性组年龄与月收入一定有显著关系，须进一步用其他方法再作检验。

Symmetric Measures

被访问者性别			Value	Asymp. Std. Error(a)	Approx. T(b)	Approx. Sig.
女	Interval by Interval	Pearson's R	.310	.079	4.081	.000(c)
	Ordinal by Ordinal	Spearman Correlation	.284	.077	3.709	.000(c)
	N of Valid Cases		159			
男	Interval by Interval	Pearson's R	.182	.071	2.278	.024(c)
	Ordinal by Ordinal	Spearman Correlation	.224	.075	2.819	.005(c)
	N of Valid Cases		153			

a Not assuming the null hypothesis.

b Using the asymptotic standard error assuming the null hypothesis.

c Based on normal approximation.

Chi-Square Tests

被访问者性别		Value	df	Asymp. Sig. (2-sided)

女	Pearson Chi-Square	22.954(a)	9	.001
	Likelihood Ratio	21.780	9	.010
	Linear-by-Linear Association	15.189	1	.000
	N of Valid Cases	159		
男	Pearson Chi-Square	11.751(b)	9	.227
	Likelihood Ratio	12.747	9	.174
	Linear-by-Linear Association	5.051	1	.025
	N of Valid Cases	153		

a 9 cells (51.3%) have expected count less than 5. The minimum expected count is .20.

b 8 cells (50.0%) have expected count less than 5. The minimum expected count is .47.

SPSS 统计绘图功能详解

2.1 常用统计图

在 SPSS 10.0 版中，除了生存分析所用的生存曲线图被整合到 ANALYZE 菜单中外，其他的统计绘图功能均放置在 graph 菜单中。该菜单具体分为以下几部分：

- Gallery：相当于一个自学向导，将统计绘图功能做了简单的介绍，初学者可以通过它对 SPSS 的绘图能力有一个大致的了解。
- Interactive：交互式统计图，这是 SPSS 9.0 版新增的内容。
- Map：统计地图，这是 SPSS 10.0 以上版本新增的内容。
- 下方的其他菜单项是最为常用的普通统计图，具体来说有：

其中后面几种图形用于时间序列分析。这里只讲解一些常规统计图，对交互式统计图和统计地图只举例介绍一下。下面以 SPSS 自带的 anxiety.sav 和 car 两个数据文件为基础，学习常规统计图的做法。

2.1.1　操作界面介绍（条形图）

●条形图的通用界面

由于不同图形的绘图对话框有相当强的共性，下面通过一个简单的例子来看看绘图菜单的大致界面是怎么样的，通过这个例子大家可以举一反三。

例 2.1：在数据集 anxiety.sav 中分不同的 subject 对变量 score 值（之和）绘制条图。

选择 graphs→bar 后，系统首先会弹出一个简单的导航对话框，如图 2.1a 所示：

绘制简单条图（单式条图）

绘制复式条图

绘制堆积条图（分段条图）

定义统计图中数据的表达类型：

条图反映了同一变量若干条记录的分组汇总

条图反映了不同变量的汇总

条图反映了个体观察值

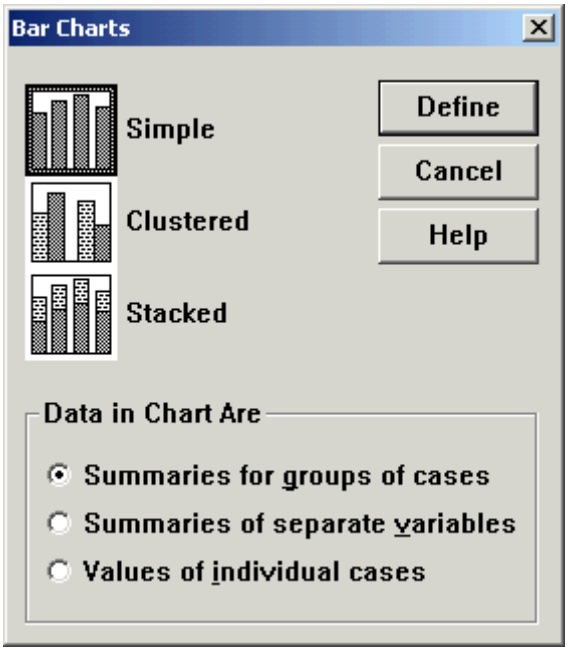


图 2.1a　bar 导航图及各部分含义

在该对话框中，SPSS 将条图进行了大致的分类，对话框的上半部分用于选择条图类型，下半部分的 Data in Chart are 单选框组用于定义条图中数据的表达类型。这里根据所需绘制条图的类型，应该选择简单条图，在表达类型中则应选择"Summaries for groups of cases"。选好后单击 DEFINE，系统开启正式的条图定义对话框如图 2.1b。

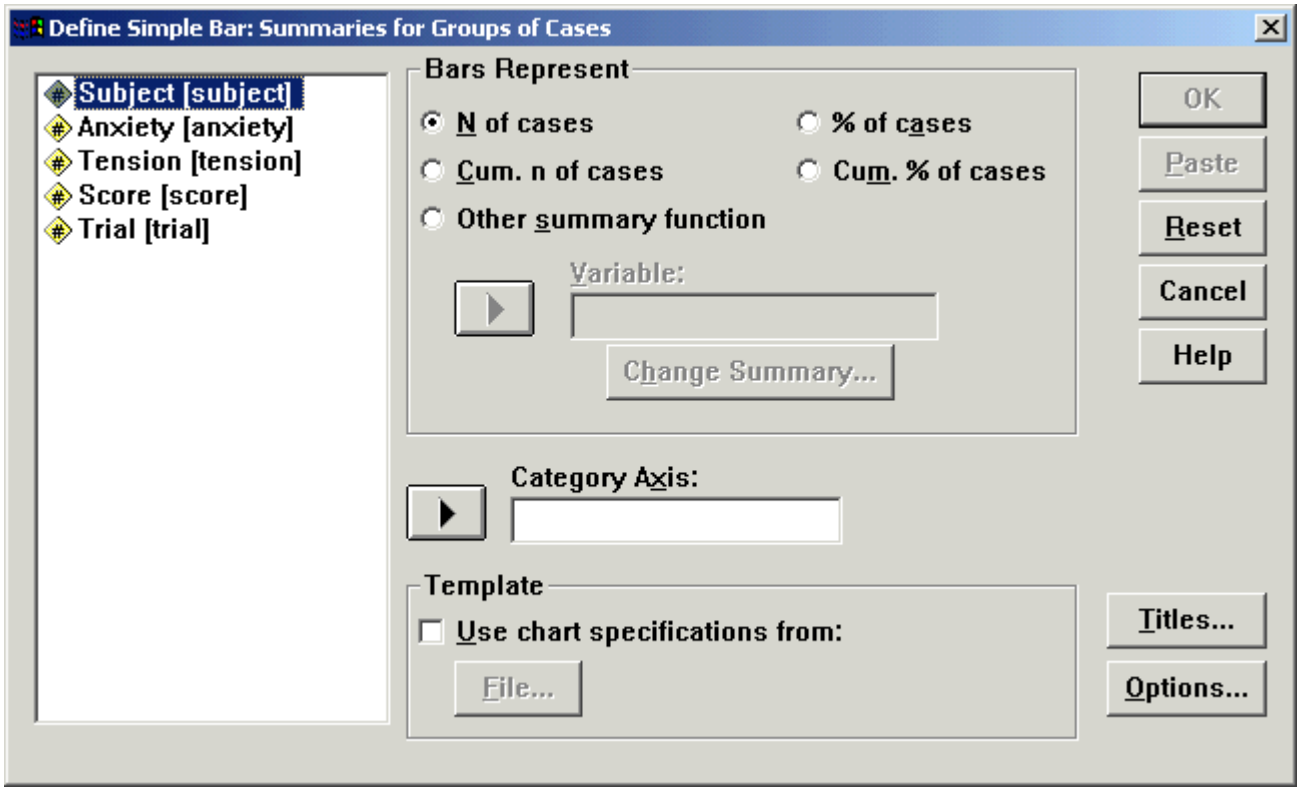


图 2.1b Define Simple Bar 对话框

对话框左侧为通用的侯选变量列表框，右侧的对话框元素依次解释如下：

1、Bars Represent 单选框组

用于定义条图中直条所代表的含义，可以是样本例数、样本数所占的百分比、累计样本例数、累计样本数所占的百分比或其余汇总函数。例 2.1 要求对变量 score 之和绘图，因此选择最后一项"Other summary function"，系统开启 summary function 对话框，见图 2.1c。

该对话框中列出了更多的统计汇总函数，可以满足绝大多数情况的需要。具体有：

上部：包括大多数常用统计汇总函数，如均值、标准差、中位数、方差、众数、最大、最小值、样本例数、变量值之和、累计变量值。

中部：可对各记录按大小进行筛选，如上侧百分之多少，或者只选择小于某个数值的记录。具体的数值在 value 框中输入。

下部：可按数值大小值选择取值在某个范围内的记录，具体的范围在 low 和 high 框中输入。

对话框最下侧还有一个 Values are grouped midpoints 复选框。当选中 median of values 或 percentile 单选框时，该框变为可选；选中则表明数据为频数表格式，所输入的数值为组中值。

根据例 2 的绘图要求，这里应该选择 sum of values 单选框。然后单击 continue 回到上图 2.1b。

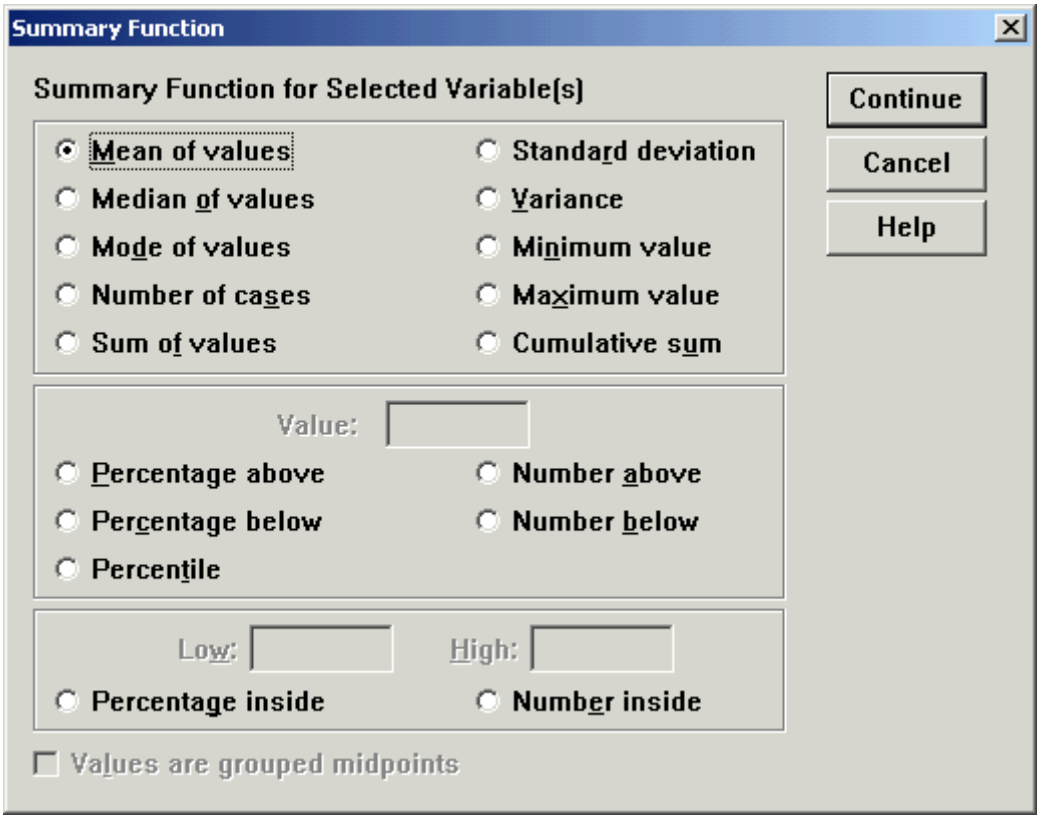


图 2.1c summary function 对话框

2、Category Axis 框：用于选择所需的分类变量，此处必选。这里根据要求，将 subject 选入，可以见到此时 OK 已经变黑可用了。

3、Template 框：用于选择绘制条图的模板，一般较少用。

4、Titles 钮：用于输入统计图的标题和脚注，最多可以输入两行主标题，一行副标题，两行脚注。

5、Options 钮：弹出 Options 对话框，用于定义相关的选项，包括 Displays a report of missing values 和 Missing Values 选项。

其中，Missing Values 单选框组用于定义分析中对缺失值的处理方法，可以是具体分析用到的变量有缺失值才去除该记录（Excludes cases analysis by analysis），或只要相关变量有缺失值，则在所有分析中均将该记录去除(Excludes cases listwise)。默认为前者，以充分利用数据。

至此，例 2.1 要求绘制的条形图操作已实施，只需单击 OK，系统就能绘出统计图如图 2.1d 所示。

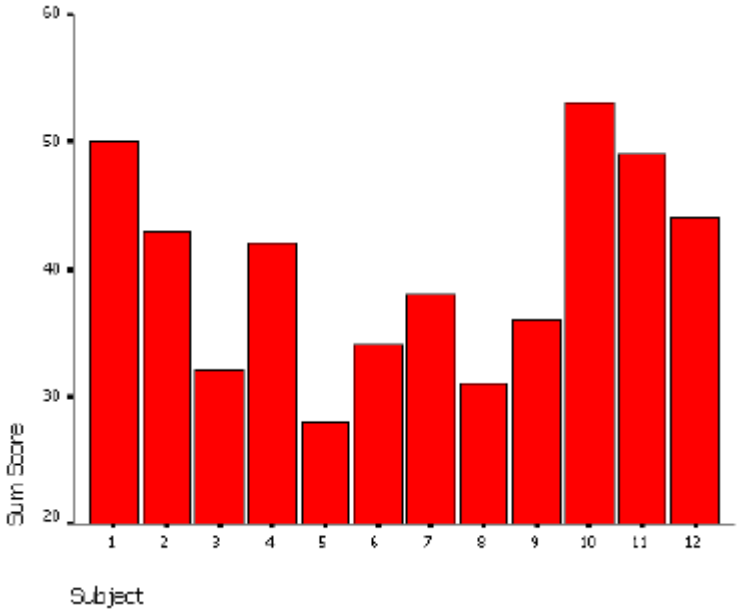


图 2.1d 按 Subject 分组的 scores 之和分布条形图

至于 Data in Chart Are 中的另两种情况 Summaries of separate variables 和 Values of individual cases，其对话框界面极为简单，可以说是一目了然，这里不再多讲，只指出以下几点：

- 在 Summaries of separate variables 的对话框中，可以用 Change summary 钮更改汇总函数。
- 在 Values of individual cases 的对话框中，下方 category labels 的选择并不影响做出直条的多少，只会影响 X 轴表示的内容，默认是记录号。

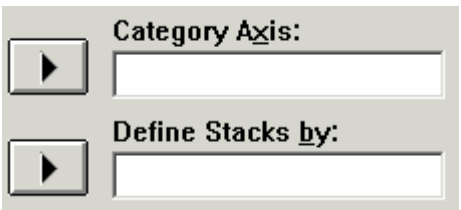
●复式条图与分段条图的界面

复式条图与分段条图的界面并非全新的东西，只是在前面的简单界面上增加了一些元素，下面看看一个例子。

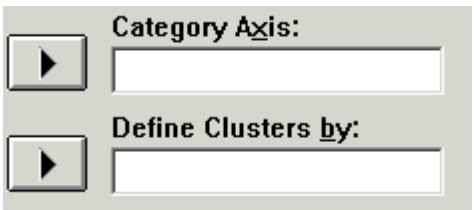
例 2.2 在 anxiety.sav 中分不同的 subject 对变量 score 值（之和）绘制条图，并且按变量 trial 的不同取值堆积（分段）。

由于要按变量 trial 的不同取值分段，因此在导航对话框中就不能选 simple，而应根据目的选择 stacked，单击 define 后系统开启的条图定义对话框和前面所用的略有不同。

具体讲，在 Category Axis 框附近不同，现在 Category Axis 框下面多了些东西如下所示：



选择 stacks 时的情况



选择 clusters 时的情况

显然，当需要做复式条图时，将所需的分类变量选入 stacks 框中即可，做分段条图的情况也与此类似。

以例 2.2 为例，其操作步骤如下：

1. Graphs→bar
2. Clustered：选中
3. Summarizes for groups of cases 单选框：选中
4. 单击 Define
5. Bars repesent 框：选入。
6. Other summary function 单选框：选中
7. Variable 框：选入 score
8. Change summary：单击
9. Sum of values 单选框：单击
10. 单击 continue
11. Category Axis 框：选入 subject
12. Define stacks by 框：选入 trial
13. 单击 OK

绘出的条图如图 2.2 所示。

但是，在 Summarizes for groups of cases 和 Values of individual cases 的对话框中情况有些不同，原先 Bars repersent 框只能选入一个变量，做复式条图和分段条图时该框中可以选入多个变量了，其他的内容不变。

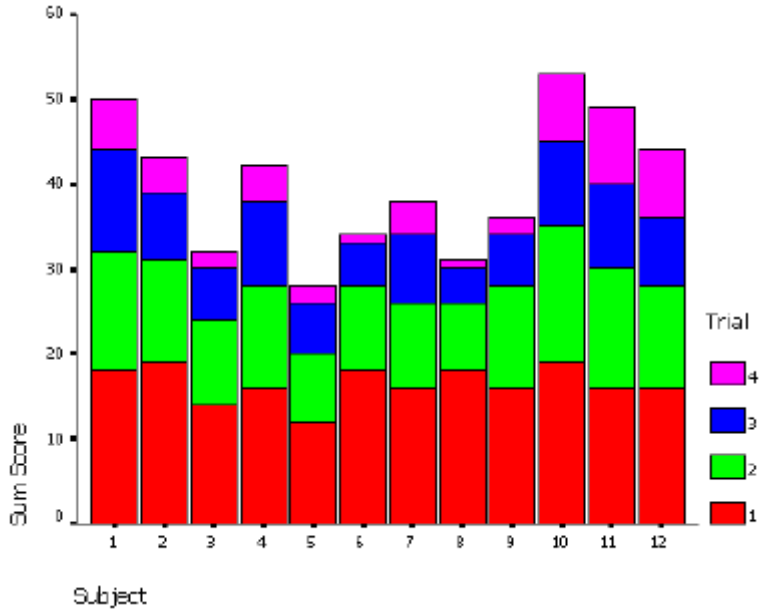


图 2.2 满足例 2.2 分组（或分段）要求的 Scores 之和的分布条形图

2.1.2　其他常用统计图

●散点图

散点图是各种统计图中比较简单的一种，共分为 simple、matrix（以矩阵的形式显示多个变量间两两的散点图）、overlay（将多个变量间两两的散点图同时做在一张图上）和 3D（将 X、Y、Z 三个变量间的相关散点图做在一个立体空间中）四种。

选择 graphs→scatter，系统弹出 Scatterplot 对话框，显示四种类型散点图，见图 2.3a。选择一种类型，单击 Define 就进入相应类型的 plot 对话框，如选 Simple 后就进入 Simple Scatterplot 对话框。

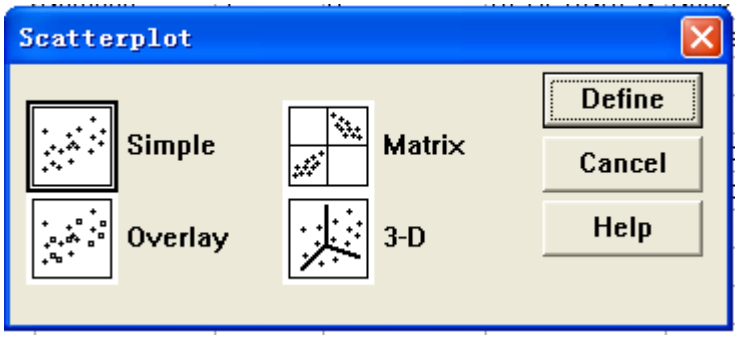
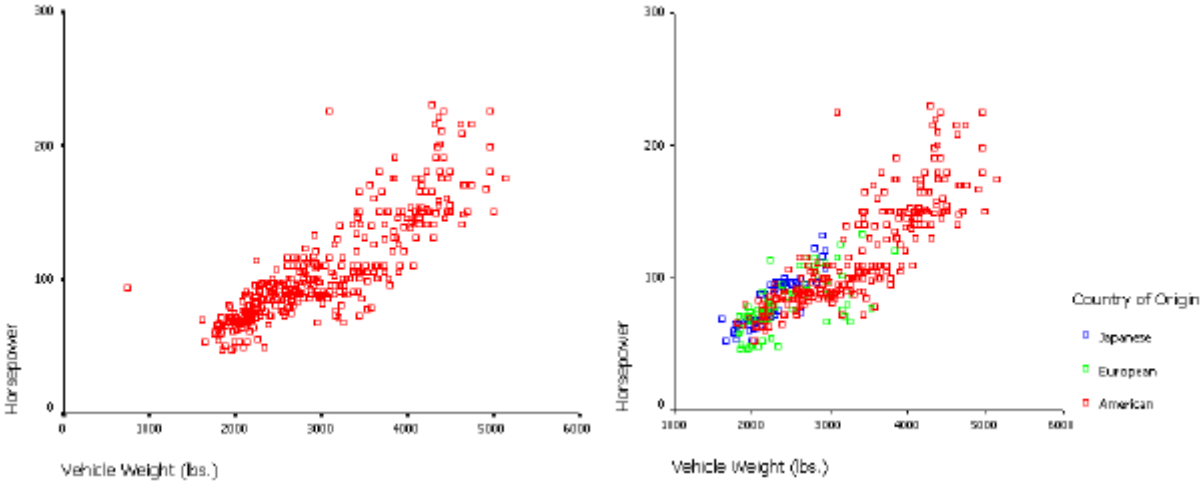


图 2.3a Scatterplot 的定义对话框

各类 Scatterplot 对话框都大同小异，而且容易理解，但其中需要特别解释一下的有：

- Set marks by 框：选入一个标记变量到该框内，系统将根据该变量取值的不同，而对同一个散点图中的各点标以不同的颜色（或形状）。例如，在数据文件 cars 中，以 horse 和 weight 作 Simple Scatter 图，没有引入作 Marks 变量时，条形图如图 2.3b 左边所示；而用 orgion 的大小来作 marks 变量时，条形图如图 2.3b 右边所示。
- Label cases 框：当编辑图形在图形选项中选择显示 labels 时，图形默认显示记录号，如果在这里选择了 label 变量，则显示该变量的取值。
- 做出的 3D 图形可以在编辑时进行三维旋转，从多个角度进行观察。



没有 mark 变量时的情况　　用 orgion 做 mark 变量时的情况

图 2.3b 有没有 mark 变量下的 horse 和 weight 的 Simple Scatterplot

●线图 Line

线图实际上和条图是一回事，可以认为它就是条图的变形，条图是用直条的高低表示多少，而线图是用点的高低来表示，然后又用直线将各点连接而成。

●饼图 Pie

饼图的做法也比较简单，这里就不再累赘了。

●面积图 Area

面积图的做法是和线图、饼图类似的，比如堆积面积图是将各个指标值相加而成，和分段式条图非常类似。

●直方图

直方图用于观察某个变量的分布情况，如果选择了 display normal curve 复选框，则会同时做出一条当前变量理想状况的正态分布曲线来和该曲线相比，这样就可以知道变量的实际分布究竟差了多远。图 2.4 是利用 Car.sav 文件中的 Vehicle Weight 数据作出的直方图，明显呈右偏分布。

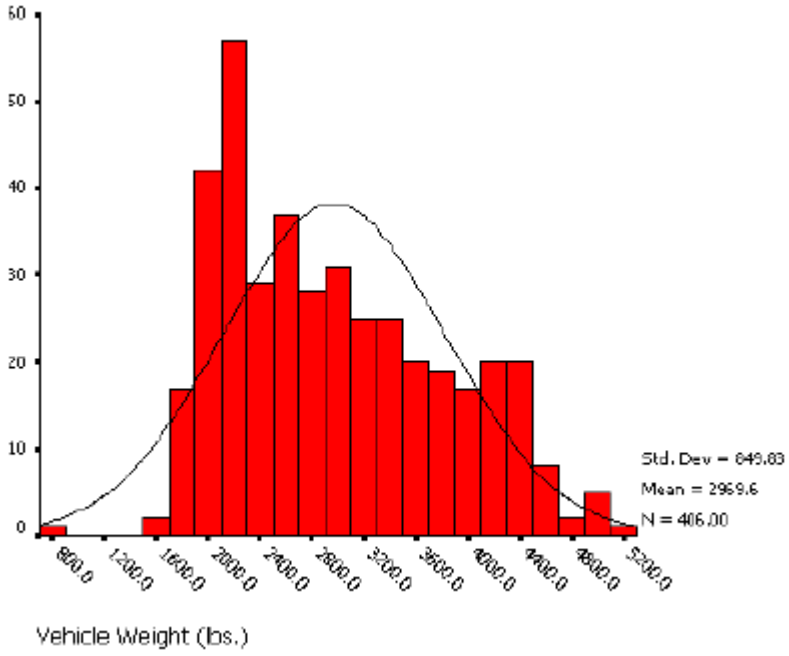


图 2.4　Vehicle Weight 的直方图

●其他

P-P 图和 Q-Q 图都是用来观察变量是否服从正态分布的；质量控制图则用来观察个体值是否有超过正常值范围的情况出现；箱式图的作用和它类似，只是换了一种表达方式；其余的几种图几乎都是用与时间序列模型的。

2.2 交互式统计图

交互式统计图是 SPSS 8.0 版新增的绘图类型，包括了交互式条图、线图、面积图、饼图、散点图、箱式图、误差限图和和直方图共七种类型。交互式统计图和普通的统计图相比有优越性，主要体现在“交互式”这三个字上：

- 对话框的交互。它的对话框全部采用拖方式操作，并且每一个元素的可操作性都大大强于普通对话框，以前需要两至三层对话框才能完成的工作，现在在一

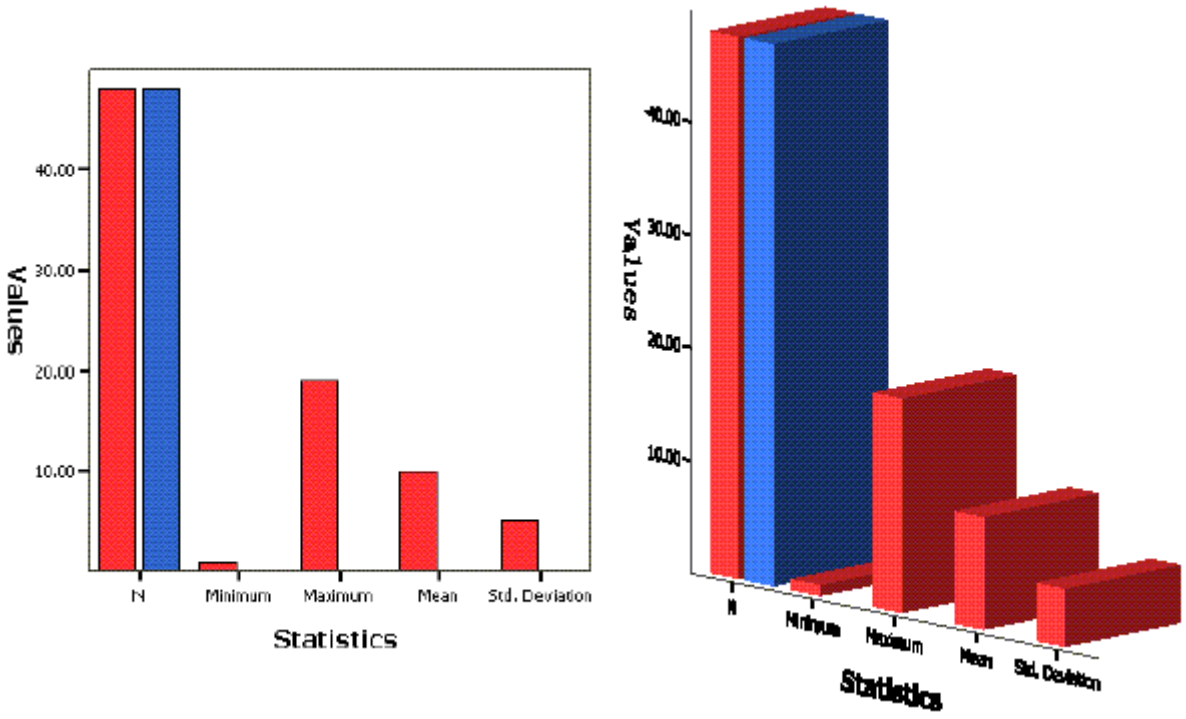
层对话框中就可以完成了。

- 图形内容的交互。在技术上，普通统计图存储的是图形元素，因此编辑时只能就图形元素的特征，如颜色、线型等加以修改；而现在的交互式统计图完全不同，它存储的是原始数据或者绘图用的中间结果（如均值、标准差等），因此当图形绘制完毕后仍能对图形进行彻底更改，如加入新的变量（在散点图中加入标示变量，甚至二维变三维）、删除某一部分数据、甚至改变所绘图形的基本类型，如将条图改绘为线图等，只要所需信息相同，随你如何转换！不但如此，由于这个存储特点，现在还可以绘出以前无法直接得到的图形，如将一个数据透视表的内容用图形来表示！

图 2.5 中的 Descriptive Statistics 表是对 anxiety.sav 文件中 Score 数据作的简单描述统计分析结果，包括观察单位数 N、最大最小值（maximum/minimum）、平均值 mean 和标准差 Std. Deviation；下方两个图是分别用二维或三维做出的 Score 统计分析结果条形图。

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Score	48	1	19	10.00	5.17
Valid N (listwise)	48				

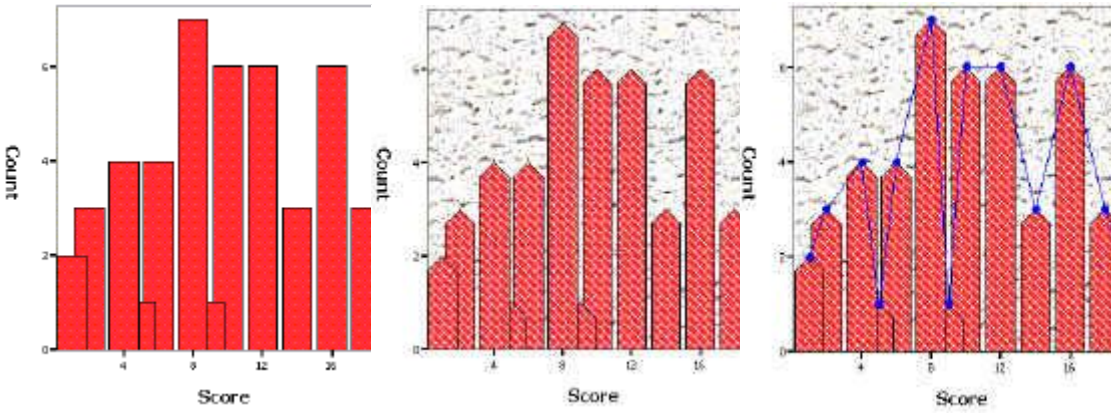


二维交互式统计图　　三维交互式统计图

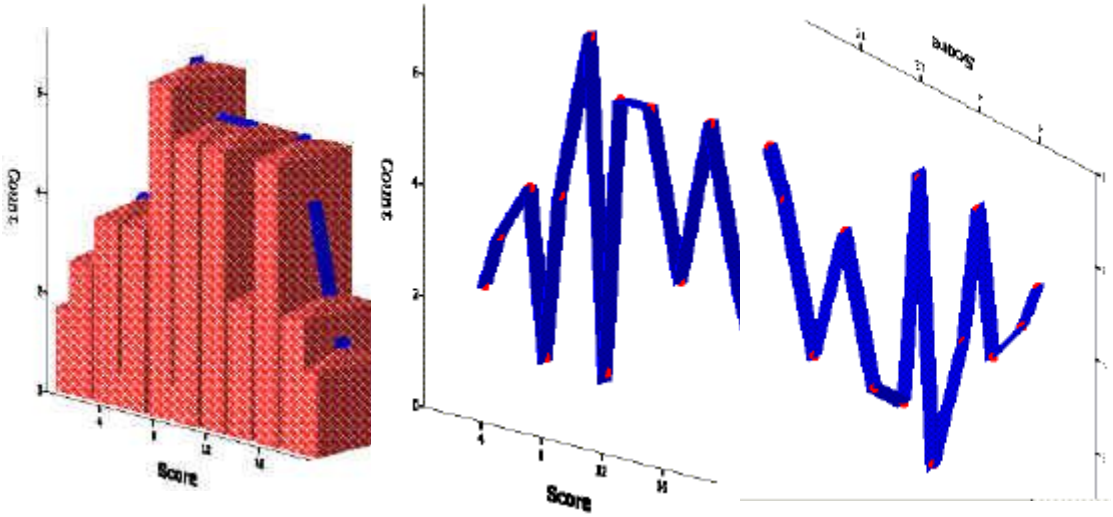
图 2.5 Score 描述统计结果的交互式统计图

- 增强的图形编辑能力。同样由于它的存储特点，现在交互式统计图的图形编辑

能力非常强，几乎任何东西都可以拿来改，也可以往里添加许多辅助线，就如图 2.1 所示的各种按 score 分组的次数分布图一样。



朴素的美　　削尖脑袋的模样　　墙上长满了爬山虎



太阳照在红墙内外　　反客为主的蓝飘带　　放倒？

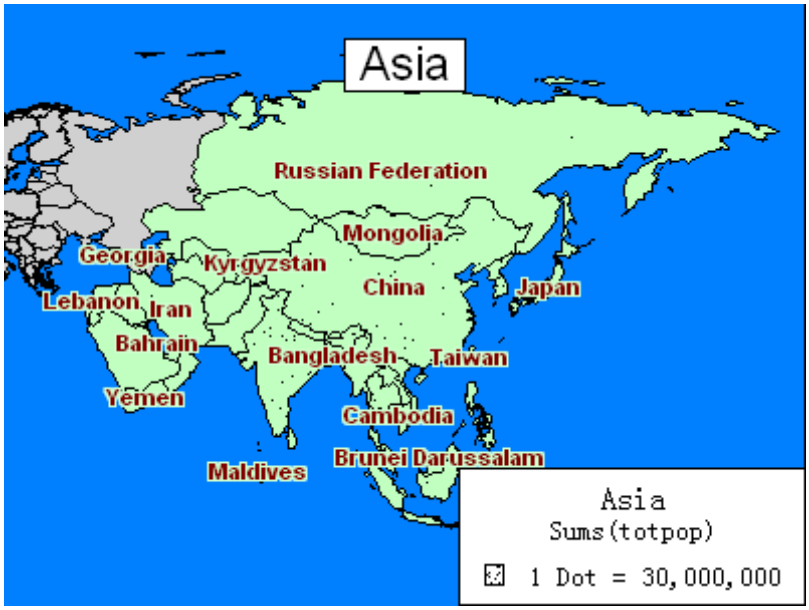
图 2.1 各种按 score 分组的次数分布图

请注意,最后一幅图是在三维实时旋转时截取的（三维实时旋转也是交互式统计图新增的功能之一）。

2.3 统计地图

统计地图是 SPSS 10.0 的新增功能，该功能可以将收集到的数据和地图相联系，从而绘出统计地图来。该功能共分为区域值统计地图、渐近符号统计地图、点密度统计地图、个体值统计地图、分类计数条图统计地图、饼图统计地图和多主题统计地图七种。但是，该地图集关于中国的部分简直就是一塌糊涂，所以对国内用户来说它更多的使用来玩，而不是工作。

统计地图在操作上和交互式统计图完全一致,实际上,它就是一类特殊的交互式统计图。它所用的数据集应该和所选的地图相对应，否则会给出错误信息，并停止做图。这是用 SPSS 附带的亚州数据集做出的亚洲国家人口点密度图：



SPSS 在根目录下的 MapData 目录中放有许多绘制统计地图用的数据集，有兴趣的朋友可以自己做几个图试试。

附：实验项目 1：SPSS 数据文件建立与数据预处理操作

实验项目	能应用 SPSS 软件进行：描述分析、频数分析、探索分析、交叉表和图形分析
实验日期	
实验环境	SPSS for WINDOWS
实验内容	依据上个实验的数据文件或选择 SPSS 数据库中的文件，进行： 1、描述分析 2、频数分析 3、探索分析 4、自拟研究目的的交叉表分析 5、自拟研究目的的图形分析
实验步骤	根据实验自己认真填写.
实验结论（或实验体会）	1. 写出求解问题的主要结果。 2. 谈谈实验体会。
实验批改	