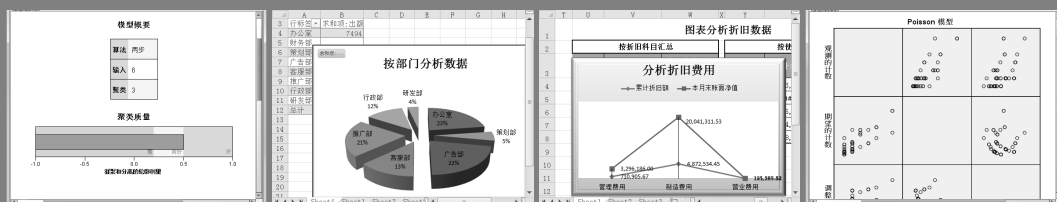


第 3 章

描述统计分析



统计分析的目的是研究数据的总体特征，描述统计分析内容是统计分析学中的基础内容，是对分析数据进行正确统计推断的先决条件。通过描述统计分析，可以使杂乱无章的数据呈现出其规律性，以帮助用户把握数据的整体分布形态，为数据建模提供数据依据。在本章中，将详细介绍描述统计分析的基础知识与分析方法，为用户学习高深的统计分析技术奠定基础。

本章学习目标：

- 频数分析
- 描述分析
- 交叉列联表分析
- 探索性统计分析
- 比率分析

3.1 频数分析

频数分析是基本统计分析中的一种重要形式,是把握数据分布特征的统计分析方法之一。通过频数分析,可以清晰地分析数据中各变量的取值状况。

3.1.1 频数分析概述

频数也称为次数,是指同一观测值在一组数据中出现的次数。而频率则是每个小组的频数与总数值的比值。另外,在变量分配数列中,频数表明对应组标志值的作用程度,其值越大表明该组标志值对于总体水平所起的作用越大,反之亦然。用户在使用频数分析数据之前,还需要先了解一下频数分析中的统计量、参数、频率分析图表等一些频数分析的基础内容。

1. 频数统计量

在 SPSS 中,频数分析是运用统计量和图形来描述多种类型的变量,并可以在其分析结果报告中,运用升序和降序方法来排列不同的变量。另外,当变量具有多个值时,可以通过提取频率报告,或者使用默认值或百分比标记图表的方法来分析。在使用频数统计分析数据之前,还需要先了解一下频数的统计量。

- **百分位值** 用于描述数值在一组数据中的相对位置,包括百分位数、Z 分数等。
- **集中趋势** 是描述分布位置的统计量,包括均值、中位数、众数等。
- **离散程度** 是测量数据中变异和展开的统计量,包括标准方差、方差、最小值、最大值等。
- **分布指标** 是描述分布形状和对称性的统计量,包括偏度系数、峰度系数等,这些统计量与其标准误一起显示。

2. 频数分布表中的参数

在使用 SPSS 编制频数分布表时,还需要了解下列术语及参数:

- **频数** 是指变量值在某个区间内出现的次数。
- **百分比** 是指各频数值与总样本数的比率。
- **有效百分比** 是指各频数与总体有效样本数的比率。其中,有效样本数是总样本数量减去缺失样本数量的差值。有效百分比适用于分析变量中存在缺失值的数据。
- **累积百分比** 是指各百分比的累计值。

3. 频数分析中常用图表

在频数分析中,经常会使用图表功能,形象且直观地显示变量之间的取值情况。其中,最常用的统计图表为下列 3 种:

- **条形图** 条形图是使用条形直观地显示频数分布变化的图形,包括单式和复式两种条形图,该图形适用于分析定序和定类变量。
- **饼图** 饼图是使用圆形或扇形来显示频数百分比变化的一种图形,主要用于显示

各部分对整体的贡献情况,该图形适用于数据的结构组成分析。

- **直方图** 直方图是使用矩形的面积来显示频数分布变化的一种图形,用于显示分布的形状、中心和分布。该图形适用于分析定距型变量。另外,还可以通过为直方图添加正态分布曲线的方法,来比较正态分布结果。

3.1.2 设置频率统计量

在 SPSS 软件中,执行【分析】|【描述统计】|【频率】命令,弹出【频率】对话框,如图 3-1 所示。

在【频率】对话框中,左侧的列表框中显示当前数据表中的变量,而在【变量】列表框中将显示用户需要进行频率分析所添加的变量。例如,在左侧的列表框中选择【办公软件】选项,单击【添加】按钮,如图 3-2 所示。

在【频率】对话框中,单击【统计量】按钮,打开【频率:统计量】对话框,如图 3-3 所示。在该对话框中,设置完各选项之后,单击【继续】按钮可以返回到【频率】对话框中。



图 3-1 【频率】对话框

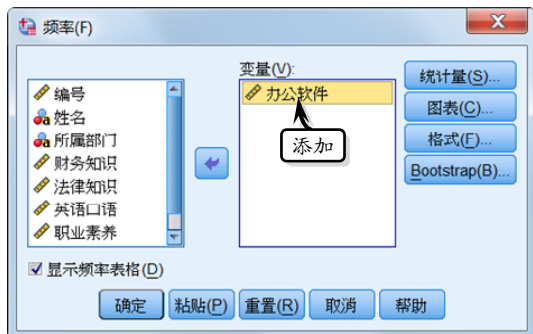


图 3-2 添加变量

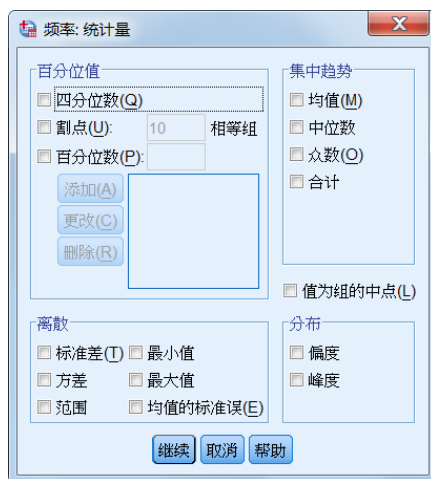


图 3-3 【频率:统计量】对话框

在【频率:统计量】对话框中,主要包括如表 3-1 所示中的选项。

表 3-1 【频率:统计量】对话框选项

选项组	选 项	作 用
百分位数	四分位数	表示将观察值分为 4 个大小相等的组,例如 25%、50%、75%个百分位数
	割点...相等组	表示将观察值分为指定的 n 个大小相等的组

续表

选项组	选 项	作 用
百分位值	百分位数	表示由用户随意指定单个百分位值，例如指定 95% 个百分位数，表示将有 95% 的观察值大于该值。在该选项中，指定百分位值之后，可通过单击【添加】按钮，添加百分位值；单击【更改】按钮，更改百分位值；同样通过单击【删除】按钮，删除百分位值
集中趋势	均值	表示算数平均值，是总和除以个案的结果值，该方法是一种集中趋势的测量
	中位数	又称为中数，表示第 50 个百分位的数值。当个案数为偶数时，则中位数是个案在升序或降序排列的情况下最中间两个个案的平均。另外，中位数是集中趋势的测量，对于远离中心的值并不敏感
	众数	众数是一组数组中最频繁出现的值，当数组中出现多个频繁出现的值时，则每一个数值都是一个众数，但频率分析过程中只会显示多个众数中最小的数值
	合计	表示所有带有非缺失值的个案值的合计值
离散	标准差	表示对围绕均值的离差的测量，其值越大表示数据的离散程度越大
	方差	该值等于与均值的差的平方和除以个案数量减去 1，其度量方差的单位是变量本身的单位的平方
	范围	范围又称为全距，表示数值变量的最大值与最小值之间的差，即最大值减去最小值，是描述数据离散情况最简单的一种分析方法
	最小值	表示数值变量的最小值
	最大值	表示数值变量的最大值
	均值的标准误	表示取自同一分布的样本与样本之间均值差的测量
分布	偏度	表示分布的不对称性度量，当偏度值为 0 时表示正态分布，当偏度值超过标准误差的两倍时，表示不具有对称性
	峰度	用于观测值聚焦在中点周围的程度的一种测量，当峰度值为 0 时表示正态分布，其分布的峰度比较高狭
值为组的中点		表示当数据中的值为组中点时，应启用该复选框以用来估计原始未分组数据的中位数和百分位数

3.1.3 设置频率格式

在使用频率方法分析数据时，还需要根据分析要求设置频率分析方法的格式，其格式主要包括数据的排列方式与变量的显示方式。

执行【分析】|【描述统计】|【频率】命令，在【频率】对话框中，单击【格式】按钮，弹出【频率:格式】对话框，如图 3-4 所示。在该对话框中，设置完各选项之后，单击【继续】按钮可以返回到【频率】对话框中。

在【频率:格式】对话框中，主要包括如表 3-2 所示中的选项。

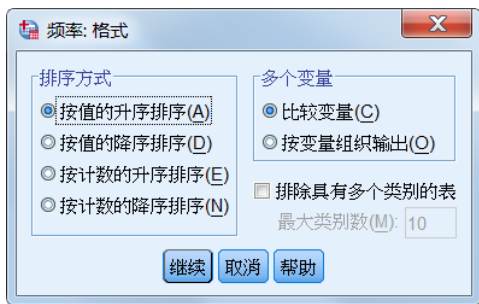


图 3-4 【频率:格式】对话框

表 3-2 【频率：是】对话框选项

选项组	选 项	作 用
排序方式	按值的升序排列	表示对频数分布按照数值的大小进行升序排列
	按值的降序排列	表示对频数分布按照数值的大小进行降序排列
	按计数的升序排列	表示对频数分布按照频数的大小进行升序排列
	按计数的降序排列	表示对频数分布按照频数的大小进行降序排列
多个变量	比较变量	表示在单个表中显示所有的变量
	按变量组织输出	表示显示每个变量的独立统计量表
排除具有多个类别的表		表示显示用于指定数量值的表

3.1.4 设置频率图表

SPSS 软件为用户提供了显示频数分析结果的条形图、饼图与直方图图表。只需执行【分析】|【描述统计】|【频率】命令，在【频率】对话框中，单击【图表】按钮，即可弹出【频率：图表】对话框，如图 3-5 所示。在该对话框中，设置完各选项之后，单击【继续】按钮可以返回到【频率】对话框中。

在【频率：图表】对话框中，主要包括“图表类型”与“图表值”两个选项组。其中，在“图表类型”选项组中主要包括下列选项：

- ☐ 无 选中该选项，表示在结果中将不显示图表。
- ☐ 条形图 选中该选项，表示在结果中将以条形图的样式显示分析数据。
- ☐ 饼图 选中该选项，表示在结果中将以饼图的样式显示分析数据。
- ☐ 直方图 选中该选项，表示在结果中将以直方图的样式显示分析数据。
- ☐ 在直方图上显示正态曲线 该复选框只有在选中【直方图】选项后才可用。启用该复选框，表示在直方图中显示正态分布曲线，用以判断分析结果数据是否接近于正态分布。

另外，在“图表类型”选项组中，选中【条形图】与【饼图】选项时，“图表值”选项组中才变成可用状态。而在“图表值”选项组中，主要包括【频率】和【百分比】选项。其中，【频率】选项表示所选择的图表将以频数为单位进行显示；而【百分比】选项表示所选择的图表将以百分比为单位进行显示。

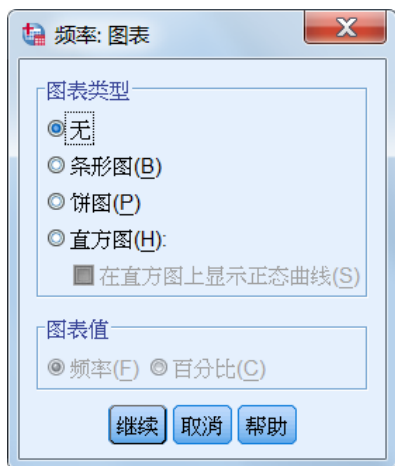


图 3-5 【频率：图表】对话框

3.1.5 Bootstrap 分析方法

Bootstrap 是 SPSS 新增功能，是一种非参数统计方法，也是一种从样本计算得到的

估计值来做出有关这些总体参数的推论。使用 Bootstrap 方法可以导出稳健的标准误差值,并能为诸如均值、中位数、比例、相关系数或回归系数等估计值导出置信区间。另外,Bootstrap 方法还可以用于构建假设检验。

执行【分析】|【统计分析】|【频率】命令,在【频率】对话框中,单击【Bootstrap】按钮,即可弹出【Bootstrap】对话框,如图 3-6 所示。在该对话框中,设置完各选项之后,单击【继续】按钮可以返回到【频率】对话框中。

在【Bootstrap】对话框中,启用【执行 bootstrap】选项,使 Bootstrap 各选项处于可用状态。各选项的具体功能如下所述:

- ❑ **样本数** 可通过在文本框中指定一个正整数的方法,来设置 Bootstrap 执行时所需要的样本个数。而当用户需要生成的百分位数和 BCa 区间时,至少需要 1000 个 Bootstrap 样本。其取值范围介于 0 ~ 2147483647 之间的整数。
- ❑ **设置 Mersenne Twister 种子** 启用该复选框,可以允许用户复制分析,另外所设置种子会保留随机数生成器的当前状态并分析完成后恢复该状态。其取值范围介于 1 ~ 2000000000 之间。

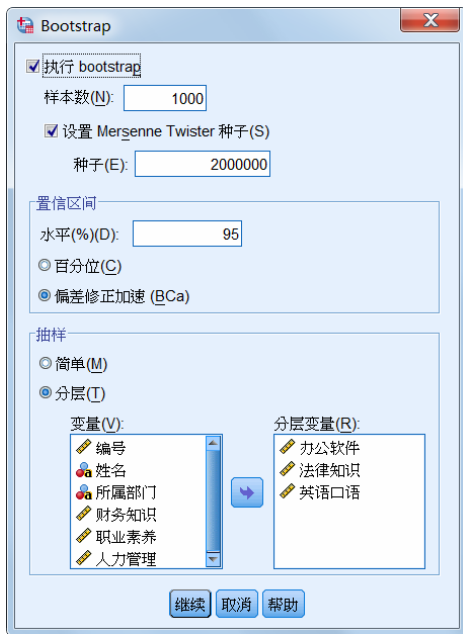


图 3-6 【Bootstrap】对话框

- ❑ **置信区间** 指定一个大于 50 且小于 100 的置信水平。其中,“百分位”选项表示简单地使用对应于所需置信区间百分位数的有序 Bootstrap 值。而“偏差修正加速 (BCa)”选项表示该区间为调整区间,分析比较准确,但运算时间比较长。
- ❑ **抽样** 在该选项组中主要包括“简单”与“分层”两种选项,其中,“简单”选项表示为通过放回方式从原始数据集进行个案重新取样;而“分层”选项表示为通过放回方式从原始数据集进行个案重新取样,但在层次变量的交叉分类定义的层内。如果层中的单元相对均一,且不同层间的单元相差较大,则分层 Bootstrap 抽样非常有用。

另外,用户在使用 Bootstrap 方法进行频数分析时,还需要注意下列几点:

- ❑ Bootstrap 不能用于多重插补数据集。如果在数据集中存在 Imputation_ 变量,Bootstrap 对话框将被禁用。
- ❑ Bootstrap 使用列表删除来确定个案基础,即在任何分析变量上具有取缺失值的个案将从分析中删除,因此当 Bootstrap 生效时,不管分析过程中是否制定了其他处理缺失值的方法,该列表删除照样处于生效状态。
- ❑ 频率分析结果中的统计表支持均值、标准差、方差、中位数、偏度、峰度和百分

位数的 Bootstrap 估计。

- ☐ 统计分析结果中的频率表支出百分比的 Bootstrap 估计。

3.1.6 显示分析结果

设置完所有参数之后，在【频率】对话框中，单击【确定】按钮。此时，系统会自动弹出【输出】窗口，并显示分析结果，如图 3-7 所示。

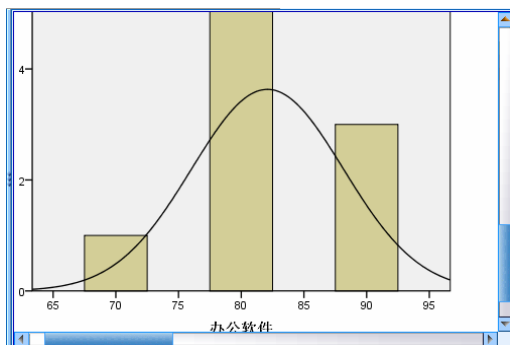


图 3-7 显示分析结果

3.2 描述分析

描述分析像频率分析那样，属于 SPSS 数据分析中“描述分析”的一部分。它是将研究中的数据加以整理、归类、简化或绘制成图表，以此分析数据的观测个数、中心趋势，以及到中心值的变异或离散程度的一种过程。通过描述分析，可以清晰、准确地分析数据的分布特点。

3.2.1 描述分析概述

描述统计的过程为单个表中若干变量显示单变量摘要的统计量，并以此计算标准化值。其中，描述统计主要涉及数据的集中趋势、离散程度和分布形态，最常用的指标有平均数、标准差和方差等。

1. 集中趋势

集中趋势是指一组数据向某一中心值靠拢的程度，反映了该组数据中心点的位置。集中趋势统计主要是寻找数据水平的代表值或中心值，其度量包括均值、中位数、众数和中列数。

☐ 均值

均值又称为算术平均数，表示一组数据或统计总体的平均特征值，是最常见的代表值或中心值，主要反映了某个变量在该组观测数据中的集中趋势和平均水平。均值是计算平均指标最常用的方法和形式，其计算公式表现为：

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

式中： n 代表总体样本数， x_i 代表各样本值。通过该公式，用户可以发现均值的大小比较容易受到数据中极端值的影响。

☐ 众数

众数是指一组数据中出现最多的数值，也是明显集中趋势的数值。在统计分析数据中，鉴于数据分组区别于单项式和组距不同类型的分组，所以众数的方法也各不相同。

其中,由单项式分组确定众数的方法比较简单,即表示出现次数最多的数值,该方法也是最常用的方法之一。

另外,由组距分组确定的众数需要先确定众数组,然后再根据计算公式计算出众数的近似值。而众数值是依据众数组的次数与众数组相邻的两组次数的关系近似值,其计算公式分为上限与下限公式,表示为:

$$\text{上限公式: } M_o = L_{M_o} + \frac{f_{M_o} - f_{M_{o-1}}}{(f_{M_o} - f_{M_{o-1}}) + (f_{M_o} - f_{M_{o+1}})} \times d_{M_o}$$

$$\text{下限公式: } M_o = U_{M_o} - \frac{f_{M_o} - f_{M_{o+1}}}{(f_{M_o} - f_{M_{o-1}}) + (f_{M_o} - f_{M_{o+1}})} \times d_{M_o}$$

式中: M_o 代表众数; L_{M_o} 代表众数组的下限; U_{M_o} 代表众数的上限; f_{M_o} 代表众数组的次数; $f_{M_{o-1}}$ 代表众数组前一次的次数; $f_{M_{o+1}}$ 代表众数组后一组的次数; d_{M_o} 代表众数组的组距。

除了上述介绍的均值与众数之外,还包括集中趋势描述统计中的中位数与中列数等,用户应根据集中趋势度量的不同计算特点与实际数据的分析类型,合理选择相应的集中趋势度量类型,以便准确地分析不同类型的数据。

2. 离散程度

离散程度是观测各个取值之间的差异程度,也可以理解为一组数据远离其中心值的程度。通过对数据离散程度的分析,不仅可以反映各个数值之间的差异大小,而且还可以反映分布中心的数值对各个数值代表性的高低。观测度量包括样本方差、标准差与全距。

□ 样本方差

样本方差是各个数据与平均数之差的平方的平均数,主要用于研究随机变量和均值之间的偏离程度。样本方差的值越大,表示变量之间的差异性越大。其计算公式表现为:

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

□ 样本标准差

样本标准差也称为样本均方差,是随机变量偏离平均数的距离的平均数,是方差的平方根,是反映了随机变量分布离散程度的一种指标。标准差跟方差一样,其数值越大表明变量值之间的差异越大。在实际应用中,需要注意平均数相同的,标准差未必相同。标准差的计算公式表现为:

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$$

□ 全距

全距又称为极差,是观测变量的最大值与最小值之间的绝对差,也可以理解为是观测变量的最大值与最小值之间的区间跨度。全距的计算公式表现为:

$$R = \max(x_i) - \min(x_i)$$

在使用全距统计分析数据时，在相同数值的情况下其值越大，表明数据的分散程度越大，反之则表明数据的集中性越好。

3. 分布形态

分布形态主要用于分析数据的具体分布情况，例如分析数据的分布是否对称、数据的偏斜度，以及分析数据的分布陡缓分布情况等。一般情况下，可以使用偏度与峰度两个统计量进行分析。

□ 偏度

偏度是描述统计量取值分布对称或偏斜程度的一种指标，其计算公式表现为：

$$\alpha = \frac{\sum_{i=1}^n (x_i - \bar{x})^k f_i}{\sigma^3 \sum_{i=1}^n f_i}$$

当数据分布为正态时，偏度值等于 0；当数据分布为正偏斜时，偏度值大于 0；当数据分布为负偏斜时，偏度值小于 0；而偏度的绝对值越大，表明数据分布的偏斜度越大。

□ 峰度

峰度是描述统计量取值分布是否陡缓程度的统计量，也是衡量取值分配的集中程度，其计算公式表现为：

$$\beta = \frac{\sum (x - \bar{x})^4 f}{\sigma^4 \sum f} - 3$$

当峰值结果值等于 0 时，表示数据分布与正态分布的陡缓程度相同；当峰值结果值大于 0 时，表示数据分布比正态分布更加集中，其平均数的代表性更大；而当峰值结果值小于 0 时，表示数据分布比正态分布更加分散，其平均数的代表性更小。

3.2.2 SPSS 描述统计分析方法

SPSS 为用户提供了简单的描述统计分析方法，无需用户根据计算公式一步一步地进行计算，只需执行【描述】命令，即可快速且准确地对数据进行描述分析。

1. 添加变量

在 SPSS 软件中，执行【分析】|【描述统计】|【描述】命令，弹出【描述性】对话框，在【描述性】对话框中，左侧的列表框中显示当前数据表中的变量，而在【变量】列表框中将显示用户需要进行频率分析所添加的变量。例如，在左侧的列表框中选择【总成绩】选项，单击【添加】按钮，如图 3-8 所示。

在【描述性】对话框中，启用【将标

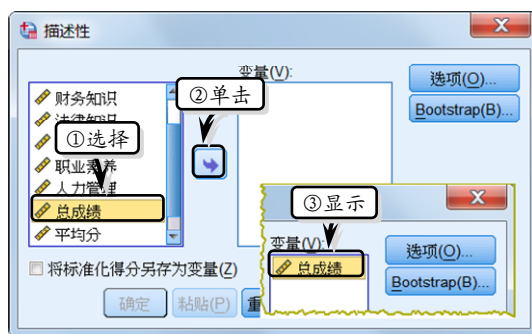


图 3-8 添加变量

【均值(M)】复选框，系统将会对所选择的变量进行标准化处理，并产生相应的 Z 分值作为新变量保存在数据窗口中。Z 分值的表达式为：

$$Z_i = \frac{x_i - \bar{x}}{s}$$

2. 设置统计选项

在【描述性】对话框中，单击【选项】按钮，弹出【描述:选项】对话框，如图 3-9 所示。

在【描述:选项】对话框中，各选项的具体功能如表 3-3 所示。

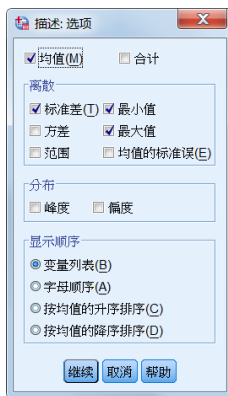


图 3-9 【描述:选项】对话框

表 3-3 【描述:选项】对话框选项的具体功能

选项组	选项	功能
均值	均值	启用该复选框，可以计算默认情况下变量的平均值
合计	合计	启用该复选框，可以计算默认情况下变量的总和
离散	标准差	启用该复选框，可以按照标准差的方式测量数据的分布情况
	方差	启用该复选框，可以按照方差的方式测量数据的分布情况
	范围	启用该复选框，可以计算数值变量最大值和最小值之间的差
	最小值	启用该复选框，可以计算数值变量中的最小值
	最大值	启用该复选框，可以计算数值变量中的最大值
	均值的标准误	启用该复选框，可以计算取值同一分布的样本与样本之间的均值之差
分布	峰度	启用该复选框，可以测量观察值聚集在中点周围的程度
	偏度	启用该复选框，可以测量数据分布的对称性
显示顺序	变量列表	选中该选项，表示将按变量的出现顺序进行显示
	字母顺序	选中该选项，表示将按字母的排列顺序进行显示
	按均值的升序排序	选中该选项，表示将按均值的升序方式进行显示
	按均值的降序排序	选中该选项，表示将按均值的降序方式进行显示

3. 显示分析结果

设置完所有参数之后，在【描述性】对话框中，单击【确定】按钮。此时，系统会自动弹出【输出】窗口，并显示分析结果，如图 3-10 所示。

提示

在【描述性】对话框中，也存在一个【Bootstrap】选项，由于该选项与“频数分析”中的选项一样，所以在此不做详细介绍。

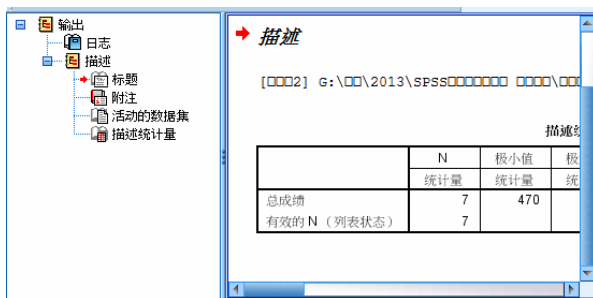


图 3-10 显示分析结果

3.3 交叉表分析

交叉表分析是描述性统计分析中的一种分析方法，是对两个或两个以上分类变量的描述与推断的统计分析。在本小节中，将详细介绍交叉表分析数据的基础知识，以及在 SPSS 中利用交叉表分析功能分析数据的操作方法。

3.3.1 交叉表分析概述

交叉表分析是同时将两个或两个以上具有一定联系的变量及变量值，按照一定的顺序交叉排列在同一张统计表中，使各变量值可以成为不同变量的结点，以便可以掌握多变量的联合分布特征，进而便于分析变量之间的相互性。

1. 交叉表分析步骤

在使用交叉表分析数据时，首先需要根据样本数据制作一个二维或多维交叉列联表。然后，在交叉列联表的基础上，分析变量之间的相关性。其中，交叉列联表是两个或两个以上变量分组后形成的频率分布表。

2. 交叉表分析的方法

在进行交叉表分析时，经常使用的分析方法为卡方检验法。卡方检验法属于非参数检验方法，是一种用途广泛的计数资料的假设检验方法，主要是比较两个或两个以上样本率以及两个或两个以上分类变量的关联性分析。

卡方检验方法与一般假设检验方法一样，首先需要建立行变量和列变量相互独立的零假设；然后，选择和计算检验统计量；最后，得出分析结果并做出决策。

在选择和计算检验统计量时，其分析中的检验统计量为 Pearson 卡方统计量，表现公式为：

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(f_{ij}^o - f_{ij}^e)^2}{f_{ij}^e}$$

式中： r 表示行数； c 表示列数； f^o 表示实际观测频数； f^e 为期望观测频数。通过公式可以发现实际观测频数和期望频数的总差值决定了卡方统计量取值的大小，卡方取值越大，表明行列变量之间的相关性越高；反之卡方取值越小，则表明行列变量之间独立性就越高。

3. 交叉表分析中的注意事项

在使用交叉表分析方法时，还需要注意下列注意事项：

- **数据** 需要定义每个表变量的类别，应该使用数值或字符串变量的值。
- **假设** 对于基于卡方的统计量，数据应为来自多项分布的随机样本。
- **有序变量** 有序变量可以为代表类别的数值代码，也可以为字符串值。但是，字符串值的字母顺序将假定反映了类别的真实顺序。鉴于字符串值的特殊性，使用数值代码代表有序数据更为可靠。

3.3.2 添加交叉表变量

交叉表变量与其他分析表中的变量不一样,它需要设置行、列与层3个变量。在SPSS软件中,执行【分析】|【描述统计】|【交叉表】命令,弹出【交叉表】对话框,如图3-11所示。

在【交叉表】对话框中,可以将左侧列表框中的变量分别添加到【行】、【列】和【层1的1】中,其中:

- **行** 表示用于显示交叉表中的行变量,适用于二维列联表分析。在左侧列表框中选择变量,单击【添加】按钮即可添加变量至该列表框中。
- **列** 表示用于显示交叉表中的列变量,适用于二维列联表分析。在左侧列表框中选择变量,单击【添加】按钮即可添加变量至该列表框中。
- **层** 该列表框主要用于放置分层的变量,适用于三维以上列联表分析。在同一层中的变量使用相同的

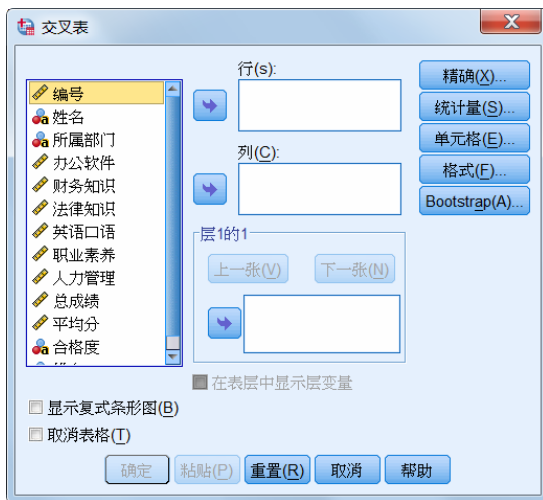


图 3-11 【交叉表】对话框

设置;但对多个层变量来讲,系统将对每个层变量的每个类别产生单独的交叉制表。

在【交叉表】对话框中,为【层1的1】列表框中添加变量之后,可以通过单击【下一张】按钮,继续添加下一个变量,依此类推。同样,单击【上一张】按钮可以返回到所设置的上一层变量中,单击【移除】按钮可以将其移除。

另外,【交叉表】对话框中,还包括下列两个选项:

- **显示复式条形图** 启用该复选框,在结果中将输出每个变量不同取值情况下的复式条形图,以汇总个案组的数据。其【行】列表框中的变量将作为聚类条形图,而【列】列表框中的变量则用于定义每个聚类内的条形图的变量。
- **取消表格** 启用该复选框,表示在结果中将不输出交叉列联表的表格。

3.3.3 交叉表中的精确检验

交叉表分析中的精确检验,主要用于设定计算的确切概率,适用于两行及以上和两列及以上的列联表。

在【交叉表】对话框中,单击【精确】按钮,将弹出【精确检验】对话框,如图3-12所示。在该对话框中,根据分析需求设置相应的选项,并单击【继续】按钮,返回到【交叉表】对话框中。

在【精确检验】对话框中,还包括下列3种类别:

- **仅渐进法** 该选项为默认选项,适用于大样本中具有渐进分布的数据,该方法只能计算出近似的显著性水平。
- **Monte Carlo** 该选项用于蒙特卡洛计算确切的显著性水平。该方法默认样本数

为 10000, 取值范围介于 1 ~ 1000000000 之间; 置信水平为 99%, 取值范围介于 0.001 ~ 99.9 之间。

- **精确** 该选项表示确切概率法, 用户可以根据实际需求需要定义检验的时间限制, 系统默认为 5 分组, 其取值范围介于 1 ~ 9999999 之间。

3.3.4 交叉表中的统计量

交叉表中的统计量主要用于设置获取二阶表或子表的关联检验和度量。在【交叉表】对话框中, 单击【统计量】按钮, 弹出【交叉表:统计量】对话框, 如图 3-13 所示。

在【交叉表: 统计量】对话框中, 用户需要根据分析目的, 有选择性地设置下列选项:

- **卡方** 启用该复选框, 若表示对于两行两列的表, 则选择卡方来计算 Pearson 卡方、似然比卡方、Fisher 的精确检验和 Yates 修正卡方的计算。对于 2×2 的表, 当表不是从具有期望频率小于 5 的单元的较大表中的缺失行或列中得来, 则计算 Fisher 的精确检验; 对具有任意行列的表, 则计算 Pearson 卡方和似然比卡方。而当两个变量都是定量变量时, 卡方将产生线性关联检验。

- **相关性** 启用该复选框, 表示对于行和列都包含排序值的表, 将生产 Spearman 相关系数 ρ (仅数值数据)。当两个表变量都是定量变量时, 产生 Pearson 相关系数 r , 即变量之间的线性关联的定量。
- **相依系数** 选择该复选框, 可以进行基于卡方的关联性测量。其值的取值范围介于 0 ~ 1 之间, 其中 0 表示行变量和列变量之间不相关, 而越接近 1 的数值, 则表示变量之间的相关度越高。
- **Phi 和 Cramer 变量** 是基于卡方统计量的关联性测量, 其绝对值越大表示变量的相关性越高。
- **Lambda** 启用该复选框, 可以反映自变量相对于因变量的预测效果, 其值为 1 时, 表示自变量可以完全预测因变量, 而值为 0 时, 则表示自变量无法预测因变量。
- **不定性系数** 为一种相关性测量, 表示当一个变量的值用来预测其它变量的值时, 误差成比例下降的情况。启用该复选框, 在结果输出时将输出对称时、行变

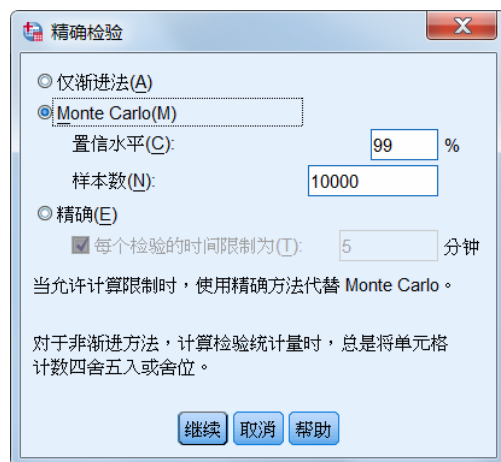


图 3-12 【精确检验】对话框



图 3-13 【交叉表:统计量】对话框

量为因变量时和列变量为因变量时的结果。

- ❑ **Gamma** 启用该复选框，可以显示两个有序变量之间的相关性的对称度量，其范围介于-1~1之间。其绝对值越接近1的值，表示两个变量之间存在的关系越紧密；而绝对值越接近0的值，则表示两个变量的关系渐弱或没有关系。该选项对于二阶表显示零阶 gamma，而对于三阶以上的表则显示条件 gamma。
- ❑ **Somers'd** 启用该复选框，可以测量两个有序变量之间的相关性，其取值范围介于-1~1之间。其中，绝对值越接近1的值，表示两个变量之间存在的关系越紧密；而绝对值越接近0的值，则表示两个变量的关系渐弱或没有关系。Somers的d是gamma的不对称扩展。
- ❑ **Kendall的tau-b** 启用该复选框，表示将结考虑在内的有序变量和排序变量的非参数相关性测量。系数的符号指示了关系的方向，而绝对值则指示强度，其绝对值越大，表明强度越高，取值范围介于-1~1之间，但-1或1值只能从正方表中取得。
- ❑ **Kendall's tau-c** 启用该复选框，可以对忽略结的有序变量的非参数关联性测量。系数的符号指示了关系的方向，而绝对值则指示强度，其绝对值越大，表明强度越高。取值范围介于-1~1之间，但-1或1值只能从正方表中取得。
- ❑ **Eta** 该选项用于范围在0~1之间的相关性测量，其中0值表示行变量和列变量之间无相关性，接近1的值则表示行变量和列变量之间高度相关。Eta适用于在区间尺度上测量的因变量以及具有有限类别的自变量。
- ❑ **Kappa** 该选项用于多方对同一估计对象的一致性。当其值为1时表示完全一致，当其值为0时表示完全不一致。如果用于估计的两个变量的数据存储类型不同，则Kappa将不进行计算。另外，对于字符串变量，变量之间的定义长度必须一致。
- ❑ **风险** 该选项用于测量因子的存在与事件的发生之间的关联性强度。当该统计量的置信区间包含值1，则不能假设因子与事件相关。
- ❑ **McNemar** 该选项用于两个相关二分变量的非参数检验，即使用卡方分布检验响应其改变。对于较大的正方表，会报告对称性的 McNemar-Bowker 检验。
- ❑ **Cochran's and Mantel-Haenszel 统计量** 该选项用于检验二分因子变量和二分类响应变量之间的条件独立性，条件是给定一个或多个分层（控制）变量定义的协变量模式。

3.3.5 交叉表中的单元显示

交叉表中的单元显示主要是获取观察值和期望值、百分比值和残差。在【交叉表】对话框中，单击【单元格】按钮，弹出【交叉表:单元显示】对话框，如图3-14所示。

在【交叉表:单元显示】对话框中的各选项的作用如表3-4所示。

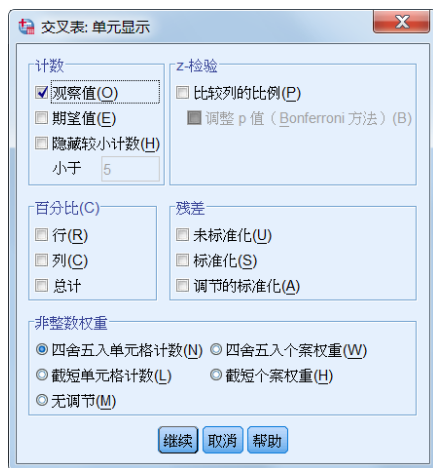


图 3-14 【交叉表:单元显示】对话框

表 3-4 【交叉表:单元显示】对话框选项的作用

选项组	选 项	作 用
Z-检验	比较列的比例	该选项将计算列属性的成对比较, 并指出给定行中明显不同的列。该选项使用下标字母以 APA 样式格式在交叉表中表示显著性差异, 并以 0.05 显著性水平对齐进行计算
计数	观察值	该选项表示计算数据中的观察值个数
	期望值	该选项表示计算数据中的期望值个数
	隐藏较小计数	该选项表示可以隐藏低于指定整数的计数, 其指定整数必须大于或等于 2, 当指定值为 0 时表示不隐藏计数
百分比	行	该选项表示统计行变量的百分比
	列	该选项表示统计列变量的百分比
	总计	该选项表示输出行、列变量的百分比与合计百分比
残差	未标准化	该选项表示观察值与期望值之间的差。当两个变量之间不存在联系时, 则期望值是期望在单元格中出现的个案数; 当行变量和列变量独立时, 则正的残差表示单元中的实际个案数多于期望的个案数
	标准化	该选项表示残差除以其标准差的估计。其标准化残差也成为 Pearson 残差, 它的均值为 0, 标准差为 1
	调节的标准化	该选项表示单元格的残差除以其标准误的估计值, 其生成的标准化残差表示为均值上下的标准差单位
非整数权重	四舍五入单元格计数	该选项表示, 在计算任何统计之前, 个案权重不变, 单元中的累计权重进行四舍五入
	四舍五入个案权重	该选项表示, 在计算任何统计之前, 对个案权重进行四舍五入
	截短单元格计数	该选项表示, 在计算任何统计之前, 个案权重不变, 单截短单元中的累计权重
	截短个案权重	该选项表示, 在计算任何统计之前, 对个案权重进行截短
	无调节	该选项表示个案权重按原样使用且使用小数单元计算

3.3.6 显示交叉表分析结果

在【交叉表】中设置完所有的选项参数之后, 便可以输出交叉表分析结果了。此时, 在输出结果之前, 还需要单击【格式】按钮, 在弹出的【交叉表:表格格式】对话框中, 设置类别的显示顺序, 如图 3-15 所示。

最后, 在【交叉表】对话框中, 单击【确定】按钮, 系统将自动弹出输出结果的“查卡器”窗口, 如图 3-16 所示。

提 示

在【描述性】对话框中, 也存在一个【Bootstrap】选项, 由于该选项与“交叉表分析”中的选项一样, 所以在此不做详细介绍。

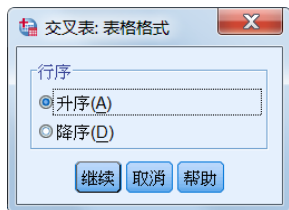


图 3-15 设置显示顺序

案例处理摘要

	案例					
	有效的		缺失		合计	
	N	百分比	N	百分比	N	百分比
姓名 * 英语口语	14	100.0%	0	0.0%	14	100.0%

姓名* 英语口语 交叉制表

计数

		英语口语						合计
		76	82	87	89	90	95	
姓名	陈鑫	0	0	0	0	1	0	1
	陈旭	0	0	0	0	1	0	1
	杜红	0	0	1	0	0	0	1
	金山	0	0	0	0	0	1	1
	金鑫	0	1	0	0	0	0	1
	刘能	0	0	0	1	0	0	1
	刘洋	0	0	0	1	0	0	1
	柳红	0	0	0	1	0	0	1

图 3-16 输出结果

3.4 探索性统计分析

探索性统计分析是在对统计数据进行了基本描述性统计之后，通过图形与数据筛选等方法对数据进行更深入的描述分析，从而帮助用户根据分析结果选择进一步更详细的分析方法。

3.4.1 探索性统计分析概述

“探索”过程既可以为所有个案也可以分别为个案组生成摘要统计量和图形显示。当用户需要进行筛选数据、离群值识别、描述、假设检验以及描述子总体（个案组）之间差异的特征时，可以使用探索性分析方法。例如，通过数据筛选可以得知数据中具有异常值、极值、数据中的缺口或其他特性。

在使用探索性统计分析之前，还需要先了解一下该分析方法中的一些常用图形与指标。例如，茎叶图、箱图、正态分布的检验等内容。

1. 茎叶图

茎叶图又称为“枝叶图”，是用来描述连续变量次数分布的一种统计方法。该方法是由统计学家托奇设计，其思路是将数组中的数位按位数进行比较，将数的大小基本不变或变化不大的位作为茎（主干），将变化大的位的数作为叶（分支），而列（频率）位于茎的后方，便于查看具体数据。茎叶图与直方图相类似，不仅可以帮助用户统计数据变量的次数，而且还在保留原数据的基础上计算各数据段的频率或百分比。

2. 箱图

箱图又称为盒须图、盒式图或箱线图，是一种用作显示一组数据分散情况资料的统计图，适用于提供有关数据的位置和分散的参考。通过箱图，不仅可以表现一组数据的四分位数、中位数、分不断全距与形态，而且还可以检测数据的异常性。

箱图中最大值和最小值显示在箱图的顶部和底端，而四分位数显示在中间箱子的底部的线段处、中位数显示在箱子的中间线段处、第三个分位数则显示在中间箱子的顶端线段处。

当中位数不在显示在箱子的中间线段处，而偏于箱子底部时，则表示呈正偏态分布；而当中位数偏于箱子顶部时，则表示呈负偏态分布。

3. 正态分布检验

正态分布检验是判断一样本所代表总体与理论正态分布是否存在显著差异的检验，是最为广泛的检验方法，是参数统计分布的前提。在探索性统计分析中，正态分布检验主要用于检验一组数据是否符合正态分布。

4. 方差齐性检验

方差齐性检验是在对两组以上的数据进行比较时，既检验了数据的正态分布性，又检验了数据的方差相等性。方差齐性检验与正态分布检验一样，也是进行探索统计分析过程的前提条件。

3.4.2 设置统计变量

交叉表中的变量与其他分析表中的变量不一样，它包括因变量列表、因子列表和标注个案 3 个选项。在 SPSS 软件中，执行【分析】|【描述统计】|【探索】命令，弹出【探索】对话框，并将“英语口语”变量添加到【因变量列表】列表框中，如图 3-17 所示。

在【探索】对话框中的 3 个选项的含义分别为：

- ☐ **因变量列表** 用于显示要对其进行探索分析的变量，为必选变量。该变量应具有合理数量的不相同的值（类别），这些值可以为短字符串或数值。
- ☐ **因子列表** 用于显示需要对【因变量列表】列表中的变量，在不同类别上进行探索性统计分析的必要变量。
- ☐ **标注个案** 该选项主要用来显示用来区分观测值的变量。



图 3-17 添加变量

提示

在【探索】对话框中，还可以通过选中【两者都】选项要求系统同时输出统计量表与图。另外，还可以选中【统计量】选项要求系统只输出统计量表而不输出图。

3.4.3 设置探索统计量

设置探索统计量，主要用于设置探索分析中所需要进行的统计量分析。在【探索】

对话框中,单击【统计量】按钮,弹出【探索:统计量】对话框,如图 3-18 所示。设置完毕之后,单击【继续】按钮,返回到【探索】对话框中。

在【探索:统计量】对话框中,主要包括下列 4 种统计量类型:

☐ **描述性** 该选项用于在默认情况下显示集中趋势度量和离差测量。其中,集中趋势度量表示分布的位置,包括中位数、均值、5%切尾均值;而离差测量主要显示值的不相似性,包括标准误、方差、标准差、最小值、最大值、范围、内距等。

☐ **M-估计量** 该选项表示样本均值和中位数的稳健替代值,用于估计位置。其计算出的估计被应用到个案的权重不同,并显示 Huber 的 M 估计、Andrews 波估计、Hampel 的重新下降 M 估计和 Tukey 的双权估计。

☐ **界外值** 该选项主要显示 5 个最大和 5 个最小带个案标签的值。

☐ **百分位数** 该选项主要用于显示第 5 个、第 10 个、第 25 个、第 50 个、第 75 个、第 90 个和第 95 个百分比值。

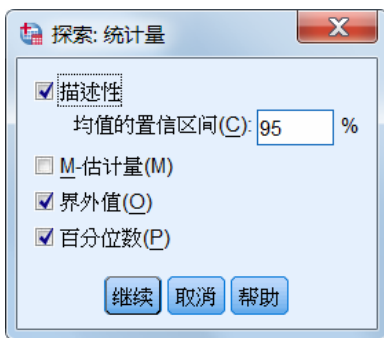


图 3-18 【探索:统计量】对话框

3.4.4 设置探索图

设置探索图主要用于设置在输出结果中所显示的分析图表类型。在【探索】对话框中,单击【绘制】按钮,弹出【探索:图】对话框,如图 3-19 所示。设置完毕之后,单击【继续】按钮,返回到【探索】对话框中。

在【探索:图】对话框中,主要包括箱图、描述性图与带检验的正态图 3 种图表类型,其具体功能与含义如表 3-5 所示。

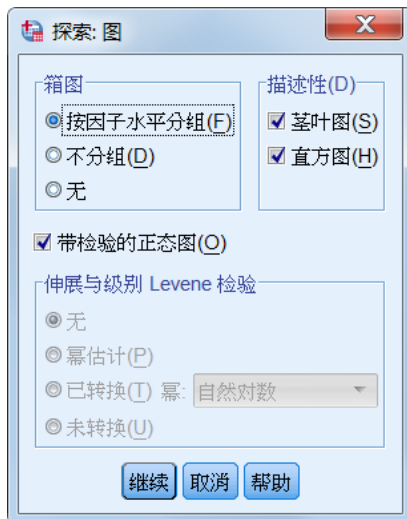


图 3-19 【探索:图】对话框

表 3-5 【探索:图】对话框选项的具体功能与含义

选项组	选 项	作 用
箱图	按因子水平分组	选择该选项,可以为每个因变量生成单独的显示。在每个显示中,将为因子变量定义的每个组显示箱图
	不分组	选择该选项,可以为因子变量定义的每个组生成单独的显示。在每个显示中,为每个因变量并排显示箱图
	无	选择该选项,表示在输出结果中将不显示箱图
描述性	茎叶图	选择该选项,表示在输出结果中显示茎叶图
	直方图	选择该选项,表示在输出结果中显示直方图
带检验的正态图		选择该选项,可以在输出结果中显示正态概论和反趋势正态概论图

续表

选项组	选 项	作 用
伸展与 级别 Levene 检验	无	选择该选项, 表示不进行 Levene 检验, 也不生成无分布-水平图
	幂估计	选择该选项, 表示可以针对所有单元的中位数的自然对数以及幂转换的估计值生成内距的自然对数图, 以在各单元中得到相等的方差。其中, 分布-水平图协助确定稳定组之间方差所需转换的幂
	已转换 (T) 幂	选择该选项, 可以通过选择幂替代值, 并生成转换的数据图, 绘制转换数据的内距和中位数
	未转换	选择该选项, 可以生成原始数据的图, 相当于幂等于 1 的转换

另外, 对于【伸展与级别 Levene 检验】选项组中的【已转换幂】选项中的幂转换, 包括下列转换类别:

- ☐ 自然对数 表示自然对数转换, 为默认值。
- ☐ 1/平方根 表示对于每个数据值, 来计算平方根的倒数。
- ☐ 倒数 表示计算每个数据值的倒数。
- ☐ 平方根 表示计算每个数据值的平方根。
- ☐ 平方 表示计算每个数据值的平方。
- ☐ 立方 表示计算每个数据值的立方。

3.4.5 显示探索分析结果

在显示探索性分析结果之前, 还需要单击【探索】对话框中的【选项】按钮, 在弹出的【探索:选项】对话框中, 设置缺失值的显示方式, 以控制对缺失值的处理, 如图 3-20 所示。

在【探索:选项】对话框中, 主要包括下列 3 个选项:

- ☐ 按列表排除个案 选择该选项, 表示从所有分析中排除任何因变量或因子变量具有缺失值的个案, 为默认值。
- ☐ 按对排除个案 选择该选项, 除了排除带有缺失值的个案之外, 还排除与其有成对关系的数值, 即表示在该组的分析中包含组 (单元) 中变量不具有缺失的个案。
- ☐ 报告值 选择该选项, 表示因子变量的缺失值被作为单独的类别进行输出。其中, 频率表包含缺失值的类别。而因子变量的缺失值也被包含在内, 但会被标记为缺失。

在【探索:选项】对话框中, 单击【继续】按钮返回到【探索】对话框中。此时, 已完成所有基础选项的设置, 单击【确定】按钮, 即可生成输出结果。结果中将包含描述、百分位数、极限、正态检验等表格, 如图 3-21 所示。

另外, 在输出结果中, 还将包含【带检验的正态图】选项中的“标准 Q-Q 图”与“趋势标准 Q-Q 图”, 如图 3-22 所示。

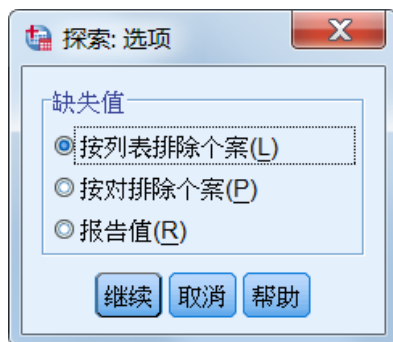


图 3-20 【探索:选项】对话框

极值			
英语口语	最高	案例号	值
	1	4	95
	2	3	90
	3	5	90
	4	7	90
	5	11	90 ^a
最低	1	12	76
	2	10	82
	3	13	87
	4	6	87
	5	9	89 ^b

a. 上限值表中仅显示一部分具有值 90 的案例。
b. 下限值表中仅显示一部分具有值 89 的案例。

图 3-21 极限输出表格

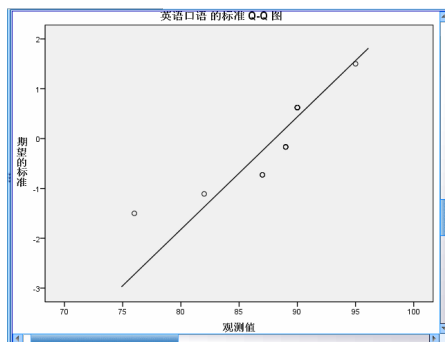


图 3-22 标准 Q-Q 图

提示

在【描述性】对话框中，也存在一个【Bootstrap】选项，由于该选项与“探索性分析”中的选项一样，所以在此不做详细介绍。

3.5 比率分析

比率分析主要用于对两个定距变量间变量值比率变化的描述分析，并生成比率变量以及比率变量的一些基本描述统计量，例如中位数、均值、标准差、最大值等。另外，比率分析还可以显示比率变量的集中趋势和离散程度，以帮助用户更详细且清晰地分析相关数据。

3.5.1 设置比率变量

在 SPSS 中，执行【分析】|【描述统计】|【比率】命令，弹出【比值统计量】对话框。在该对话框中，需要为【分子】、【分母】与【组变量】选项添加变量，如图 3-23 所示。其中，【分子】与【分布】选项中需要添加希望分析的两个变量，而【组变量】选项中主要用于添加想进行比较不同组的组变量，并且还可以为组变量设置排序方式。

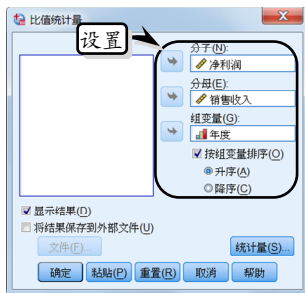


图 3-23 添加变量

3.5.2 设置比率统计量

设置比率统计量主要是设置统计分析中的集中趋势、离散、集中指数等统计量。在【比值统计表】对话框中，单击【统计量】按钮，弹出【比率统计量:统计量】对话框，如图 3-24 所示。

在【比率统计表:统计量】对话框中，主要包括



图 3-24 设置比率统计量

如下选项。

表 3-6 【比率统计量:统计量】对话框选项

选项组	选 项	作 用
集中趋势	中位数	表示小于该值的比率数与大于该值的比率数相等
	均值	表示比率的总和除以比率的总数所得到的结果
	权重均值	表示分子的均值除以分母的均值所得的结果, 该值也是比率按分母加权之后的均值
	置信区间	表示显示均值、中位数和加权权重的置信区间, 其置信区间的取值范围介于 0~100 之间
离散	AAD	表示平均绝对偏差是中位数比率的绝对离差求和并用结果除以比率总数所得的结果
	COD	表示离差系数是将平均绝对偏差表示为中位数的百分比结果
	PRT	也称为回归指数, 表示价格相关微分, 是均值除以加权均值所得到的结果
	中位数居中 COV	表示中位数居中的变异系数是将于中位数偏差的均方根表示为中位数百分比的结果
	均值居中 COV	表示均值居中的变异系数是将标准差表示为均值百分比的结果
	标准差	表示比率与均值间偏差的平方之和, 再除以比率总数减 1, 取正的平方根所得到的结果
	范围	表示最大的比率减去最小比率所得到的结果
	最小值	表示最小的比率
	最大值	表示最大的比率
集中指数	介于比例	用于指定集中指数的区间, 其区间是指区间的最小值和最大值而显示定义的。在文本框中输入最小值和最大值, 单击【添加】按钮即可获得区间
	中位数百分比之内	用于指定集中指数的中位数百分比值, 取值范围介于 0~100 之间。区间的下界等于 $(1-0.01 \times \text{值}) \times \text{中位数}$, 上界等于 $(1+0.01 \times \text{值}) \times \text{中位数}$

3.5.3 保存/显示分析结果

在比率分析中, SPSS 系统为用户提供了保存输出结果的功能。另外, 用户还可以以统计表的样式显示比率分析结果。

1. 保存分析结果

在【比值统计量】对话框中, 启用【将结果保存到外部文件】复选框, 并单击【文件】按钮, 如图 3-25 所示。

在弹出的【比率统计量:保存到文件】对话框中, 设置保存位置与名称, 单击【保存】按钮, 即可将结果保存到指定位置, 如图 3-26 所示。

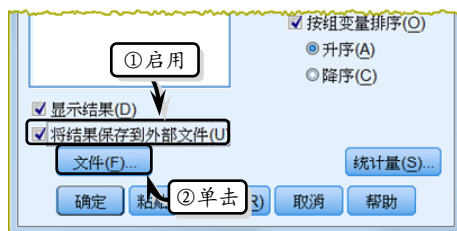


图 3-25 启用选项

系统会自动返回到【比值统计量】对话框中,此时,在【文件】按钮后面将显示所设置的文件位置,如图 3-27 所示。

2. 显示分析结果

在【比值统计量】对话框中,单击【确定】按钮,系统会自动弹出输出窗口,并显示分析统计表,如图 3-28 所示。

3.6 P-P 图和 Q-Q 图

SPSS 还提供了用于概率分析的 P-P 和 Q-Q 图,用于确定某个变量的分布是否与给定的分布相匹配。通过概率图可以详细、形象地显示数据,例如当选定变量与检验分布匹配,其点则聚集在某条直线周围,便于对数据进行观察。

3.6.1 P-P 图

P-P 图是对照一些检验分布的累计比例来绘制某个变量的累计比例图。在 P-P 图中可用的检验分布包括卡方、指数、半正态、排列等。另外,使用 P-P 图还可以获取转换值的概率图,包括自然对数、标准化值、差方和季节性差分等转换选项。除此之外,用户还可以指定计时期望分布,以及求解“连结”的方法。

在 SPSS 中,执行【分析】|【描述统计】|【P-P】命令,弹出【P-P 图】对话框,如图 3-29 所示。

在【P-P 图】对话框中,最左侧的列表框中显示当前数据文件中的变量,用户只需选择需要添加的变量名称,单击【添加】按钮即可将其添加到【变量】列表框中。该对话框中的其它选项如表 3-7 所示。

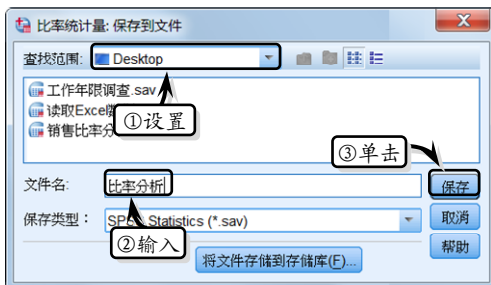


图 3-26 保存结果

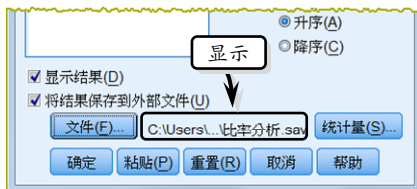


图 3-27 显示保存位置

案例处理摘要			
	年度	计数	百分比
	2011年	8	53.3%
	2012年	7	46.7%
总数		15	100.0%
排除的		0	
总计		15	

净利润/销售收入的比率统计量					
组	均值	中值	价格相关得分	离散系数	中值居中
2011年	.423	.406	1.004	.098	14.4%
2012年	.450	.391	1.040	.220	34.4%
总数	.436	.405	1.021	.154	23.0%

图 3-28 显示统计结果

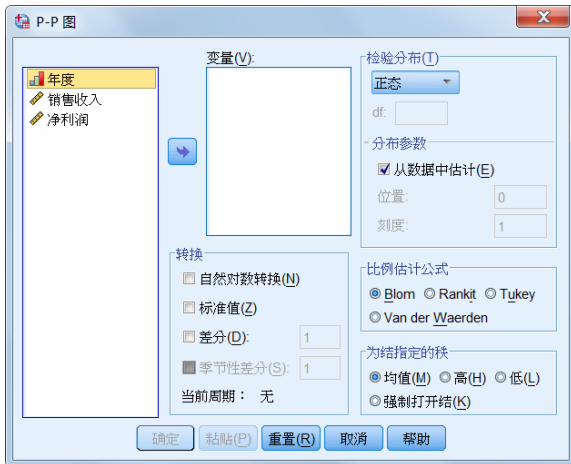


图 3-29 【P-P 图】对话框

表 3-7 【P-P 图】对话框选项

选项组	选 项	作 用
转换	自然对数转换	表示将当前变量转换为以自然数 e 为底的对数变量
	标准值	表示将当前变量转换为标准值, 转换后变量的均值为 0, 方差为 1
	差分	表示将当前变量转换为变量中连续数据之间的差值, 该选项的默认值为 1, 其值表示转换的差分度, 取值范围介于 1~99 之间
	季节性差分	表示将当前变量转换为计算时间序列的季节差分, 该选项需要对当前变量的数据序列定义了周期后才可用
	当前周期	用于显示当前变量中的周期性
检验分布		用于设置检验的分布类型, 包括卡方、正态、均匀等 13 种类型。其中, 分布类型下方的 <i>df</i> 表示分布的自由度
分布参数	从数据中估计	用于设置分布类型的参数, 其参数输入的窗口是根据分布参数类型的改变而改变
比例估计公式	Blom	其计算公式表示为 $(r-3/8)/(n+1/4)$ 。其中, n 为观测量的数据; r 为 1 从 n 的秩次
	Rankit	其计算公式表示为 $(r-1/2)/n$ 。其中, n 为观测量的数据; r 为 1 从 n 的秩次
	Tukey	其计算公式表示为 $(r-1/3)/(n+1/3)$ 。其中, n 为观测量的数据; r 为 1 从 n 的秩次
	Vander Waerden	其计算公式表示为 $r/(n+1)$ 。其中, n 为观测量的数据; r 为 1 从 n 的秩次
为结指定的秩	均值	表示用连接值的平均秩指定顺序
	高	表示用连接值的最大秩指定顺序
	低	表示用连接值的最小秩指定顺序
	强制打开结	表示忽略观测量权重的影响

在【P-P 图】对话框中设置相应选项之后, 单击【确定】按钮, 系统将自动在输出窗口中显示 P-P 图, 如图 3-30 所示。

3.6.2 Q-Q 图

Q-Q 图是对照一些检验分布的分位数, 绘制某个变量分布的分位数图。在 Q-Q 图中可用的检验分布包括卡方、指数、半正态、排列等。另外, 使用 Q-Q 图还可以获取转换值的概率图, 包括自然对数、标准化值、差方和季节性差分等转换选项。除此之外, 用户还可以指定计时期望分布, 以及求解“连结”的方法。

在 SPSS 中, 执行【分析】|【描述统计】|【Q-Q】命令, 弹出【Q-Q 图】对话框, 设置相应的选项, 如图 3-31

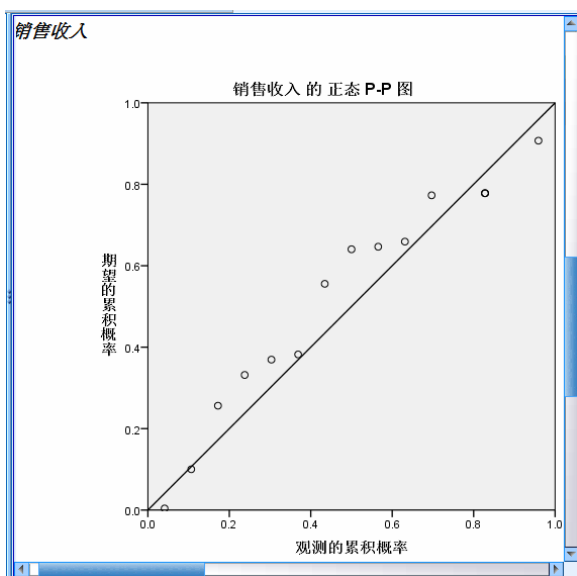


图 3-30 显示 P-P 图

所示。

用户会发现【Q-Q 图】对话框中的选项与【P-P 图】对话框中的选项完全一致，单击【确定】按钮，系统会自动在输出窗口中显示 Q-Q 图，如图 3-32 所示。



图 3-31 【Q-Q 图】对话框

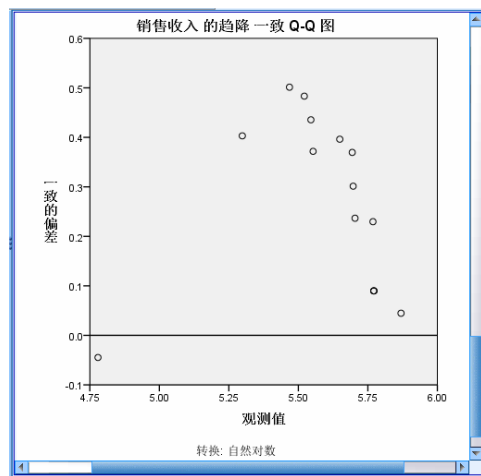


图 3-32 显示输出结果

3.7 课堂练习：交叉表分析工作年限

交叉表分析是描述性统计分析中的一种分析方法，在本练习中将运用 SPSS 中的交叉表分析工具分析已知数据中不同年龄和不同工作年限交叉分组下的频率分布情况，如图 3-33 所示。

年龄* 工作年限 交叉制表									
计数		工作年限							合计
		4	5	6	8	13	15	20	
年龄	26	0	1	0	0	0	0	0	1
	27	0	0	1	0	0	0	0	1
	28	1	0	0	0	0	0	0	1
	29	0	1	0	0	0	0	0	1
	32	0	0	0	2	0	0	0	2
	36	0	0	0	0	1	0	0	1
	37	0	0	0	0	1	0	0	1
	39	0	0	0	0	1	0	0	1
	45	0	0	0	0	0	1	0	1
	46	0	0	0	0	0	0	2	2
	52	0	0	0	0	0	0	1	1
合计		1	2	1	2	3	1	3	13

图 3-33 工作年限分析结果表

操作步骤

- 1 根据调查数据获得某公司 13 名员工的姓名、性别、年龄与工作年限。根据获取数据的类别，设置数据变量，如图 3-34 所示。

- 2 在“数据视图”窗口中，根据变量类别分布输入员工的姓名、性别、年龄与工作年限值，如图 3-35 所示。

	名称	类型	宽度	小数	标
1	姓名	字符串	8	0	
2	性别	字符串	8	0	
3	年龄	数值(N)	8	0	
4	工作年限	数值(N)	8	0	
5					
6					
7					
8					
9					

图 3-34 设置数据变量

	姓名	性别	年龄	工作年限	变量
1	刘能	男	36	13	
2	王二	男	37	13	
3	赵然	男	45	15	
4	冉欣	男	46	20	
5	金鑫	女	39	13	
6	柳红	男	26	5	
7	李四	女	27	6	
8	赵毅	男	32	8	

图 3-35 输入分析数据

- 3 执行【分析】|【描述统计】|【交叉表】命令，将“年龄”变量添加到【行】列表框中，将“工作年限”变量添加到【列】列表框中，如图 3-36 所示。

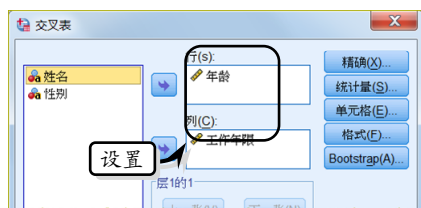


图 3-36 添加变量

- 4 单击【单元格】按钮，在弹出的【交叉表：单元显示】对话框中，启用【观察值】复选框，如图 3-37 所示。

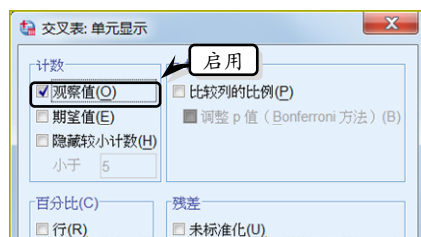


图 3-37 设置单元类型

- 5 在【交叉表】对话框中，禁用【显示复式条形图】复选框，并单击【确定】按钮，如图 3-38 所示。



图 3-38 设置输出显示

- 6 此时，系统会在弹出的输出窗口中显示交叉分析列表，如图 3-39 所示。

	工作年限					
	4	5	6	8	13	15
26	0	1	0	0	0	0
27	0	0	1	0	0	0
28	1	0	0	0	0	0
29	0	1	0	0	0	0
32	0	0	0	0	2	0
36	0	0	0	0	1	0
37	0	0	0	0	1	0
39	0	0	0	0	1	0
45	0	0	0	0	0	1
46	0	0	0	0	0	0
52	0	0	0	0	0	0
	1	2	1	2	3	1

图 3-39 显示交叉分析列表

分析结果：从图 3-39 中的交叉列表中可以看出，该统计表中的行变量为年龄，列变量为工作年限，而两组变量中的变量值分布为标题与列标题。位于统计表中间的为观测值，表示统计后每个工作年限值下的人数。例如，分析结果显示工作年限为 4、6 和 15 的人数为 1 人，而工作年限为 13 和 20 的人数为 3 人，工作年限为 5 和 8 的人数为 2 人。

3.8 课堂练习：频数分析考试成绩

某公司为提高员工的工作技能，特别组织了一系列的课程培训。培训结束后，人力资源部需要对考试结果进行分析与评价，为员工晋升提供数据依据。在本练习中，将运用 SPSS 中的频数分析工具，分析总分的频率分布情况，如图 3-40 所示。

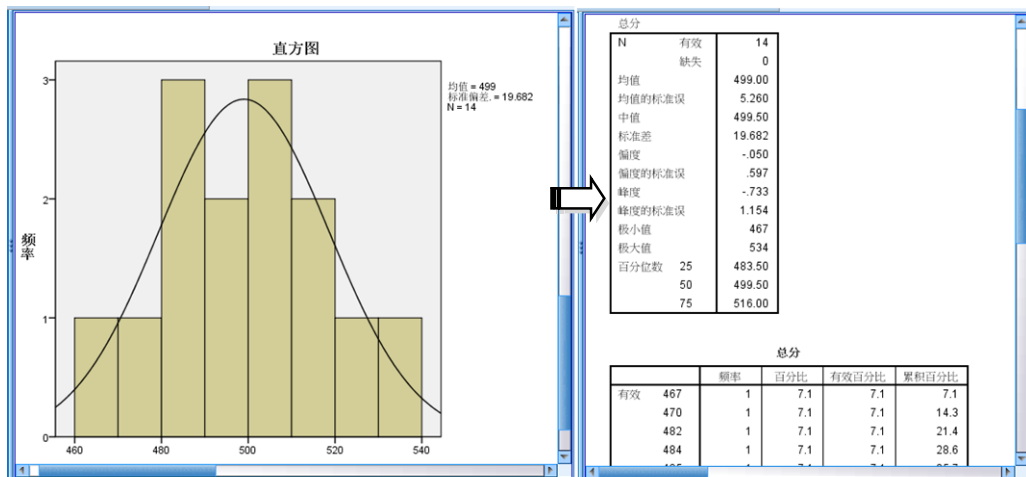


图 3-40 考试成绩频率统计表

操作步骤

- 1 根据调查数据获得某公司 13 名员工的姓名、性别、年龄与工作年限。根据获取数据的类别，设置数据变量，如图 3-41 所示。

	名称	类型	宽度	小数	标
1	编号	数值(N)	12	0	
2	姓名	字符串	6	0	
3	所属部门	字符串	9	0	
4	办公软件	数值(N)	12	0	
5	财务知识	数值(N)	12	0	
6	法律知识	数值(N)	12	0	
7	英语口语	数值(N)	12	0	
8	职业素养	数值(N)	12	0	
9	人力管理	数值(N)	12	0	
10	总分	数值(N)	8	0	

图 3-41 设置变量

- 2 在“数据视图”窗口中，根据变量类别分布输入员工考试数据，如图 3-42 所示。
- 3 执行【分析】|【描述统计】|【频率】命令，将“总分”变量添加到【变量】列表框中，如图 3-43 所示。

	编号	姓名	所属部门	办公软件	财务知识
1	1008	刘洋	财务部	83	89
2	2021	刘能	市场部	90	77
3	2011	陈鑫	市场部	82	61
4	2012	金山	市场部	70	73
5	2013	陈旭	市场部	92	80
6	1018	欣欣	行政部	80	84
7	3011	赵四	研发部	88	60
8	3012	然然	财务部	83	82
9	3013	樽红	研发部	90	70

图 3-42 输入分析数据



图 3-43 添加变量

- 4 单击【统计量】按钮，在弹出的【频率:统

【频率:统计量】对话框中,设置统计量选项,如图 3-44 所示。



图 3-44 设置统计量

- 5 在【频率】对话框中,单击【图表】按钮,在弹出的【频率:图表】对话框中,选择图表类型,如图 3-45 所示。



图 3-45 设置图表类型

- 6 在【频率】对话框中,单击【确定】按钮,在输出窗口中将显示“统计量”分析结果,如图 3-46 所示。

分析结果:通过该表可以发现数据的均值为 499,而标准差为 19.682,表示该数据的离散程度较高。

- 7 此时,分析结果中还将显示“总分”的频率分析表,如图 3-47 所示。在该表中可以看出某个分数的人数以及某分数的人数占总人

数的百分比情况。

统计量		
总分		
N	有效	14
	缺失	0
均值		499.00
均值的标准误		5.260
中值		499.50
标准差		19.682
偏度		-.050
偏度的标准误		.597
峰度		-.733
峰度的标准误		1.154
极小值		467
极大值		534
百分位数	25	483.50
	50	499.50
	75	516.00

图 3-46 统计量分析结果

总分				
	频率	百分比	有效百分比	累积百分比
有效	467	1	7.1	7.1
	470	1	7.1	14.3
	482	1	7.1	21.4
	484	1	7.1	28.6
	485	1	7.1	35.7
	495	1	7.1	42.9
	499	1	7.1	50.0
	500	1	7.1	57.1
	509	2	14.3	71.4
	516	2	14.3	85.7
	520	1	7.1	92.9
	534	1	7.1	100.0
合计	14	100.0	100.0	

图 3-47 频率分析表

- 8 最后,在分析结果中,系统还用直方图的方式直接显示了数据的正态分布情况,如图 3-48 所示。

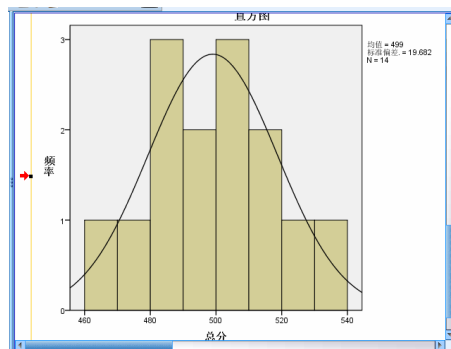


图 3-48 直方图

3.9 课堂练习：销售比率分析

已知某公司 2011 年与 2012 年销售统计数据,为保证下一年销售计划的制定精确度,还需要分析两年内的销售数据。在本练习中将运用 SPSS 中的比率分析方法,分析两年内公司销售净利润的差异性,如图 3-49 所示。

案例处理摘要			
		计数	百分比
年度	2011年	8	53.3%
	2012年	7	46.7%
总数		15	100.0%
排除的		0	
总计		15	

净利润/销售收入的比率统计量										
组	均值	中值	极小值	极大值	标准差	范围	价格相关得分	离散系数	方差系数	中值居中
2011年	.423	.406	.349	.505	.055	.156	1.004	.098	14.4%	
2012年	.450	.391	.342	.672	.118	.330	1.040	.220	34.4%	
总数	.436	.405	.342	.672	.088	.330	1.021	.154	23.0%	

图 3-49 比率分析销售净利润

操作步骤

- 1 根据销售数据,在“变量视图”窗口中设置数据变量,如图 3-50 所示。

	名称	类型	宽度	小数	标
1	年度	数值(N)	1	0	
2	销售收入	数值(N)	8	2	
3	净利润	数值(N)	8	2	
4					
5					
6					
7					
8					
9					
10					

图 3-50 设置变量

- 2 在“数据视图”窗口中,依次输入销售数据,如图 3-51 所示。

	年度	销售收入	净利润	变量	变
1	2011年	300.00	120.00		
2	2011年	200.00	100.00		
3	2011年	250.00	110.00		
4	2011年	320.00	130.00		
5	2012年	119.00	80.00		
6	2012年	321.00	160.00		
7	2011年	258.00	90.00		
8	2011年	297.00	150.00		
9	2012年	354.00	180.00		

图 3-51 输入分析数据

- 3 执行【分析】|【描述分析】|【比率】命令,在弹出的【比值统计量】对话框中,分别添加相应的变量,如图 3-52 所示。

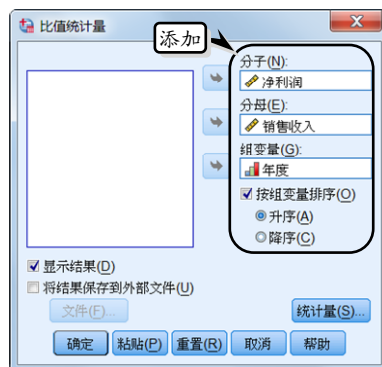


图 3-52 添加变量

- 4 单击【统计量】按钮,在弹出的【比率统计量:统计量】对话框中,设置统计量选项,如图 3-53 所示。
- 5 最后,在【比值统计量】对话框中单击【确定】按钮,将显示比率统计表,如图 3-54 所示。在该表中可以看出,2012 年的均值、标准差与离散系数都比 2011 年要高,表明 2012 年净利润的离散成本较高。

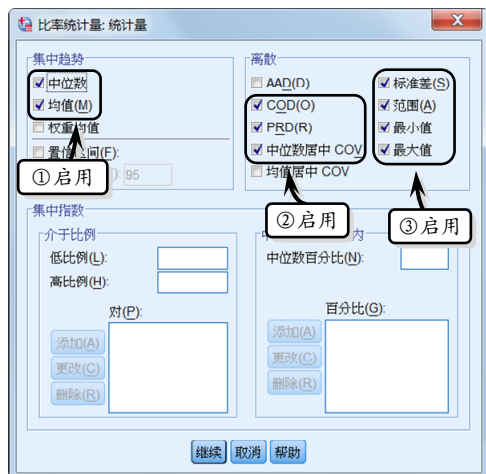


图 3-53 设置统计量

案例处理摘要					
	计数	百分比			
年度	2011年	8	53.3%		
	2012年	7	46.7%		
总数		15	100.0%		
排除的		0			
总计		15			

净利润 / 销售收入的比率					
组	均值	中值	极小值	极大值	标准差
2011年	.423	.406	.349	.505	.055
2012年	.450	.391	.342	.672	.118
总数	.436	.405	.342	.672	.088

图 3-54 显示分析结果

3.10 思考与练习

一、填空题

1. 统计分析的目的是研究数据的总体特征，主要包括频数分析、_____、_____、比率分析、描述分析 5 种分析方法。

2. 在变量分配数列中，频数表明_____的作用程度，其值_____表明该组标志值对于总体水平所起的作用越大，反之亦然。

3. 描述分析是将研究中的数据加以整理、归类、简化或绘制成图表，以此分析数据的_____、_____，以及到中心值的_____或_____程度的一种过程。

4. 交叉表分析是描述性统计分析中的一种分析方法，是对_____或_____以上分类变量的描述与推断的统计分析。

5. 探索性统计分析是在对统计数据进行基本描述性统计之后，通过_____与_____等方法对数据进行更深入的描述分析。

6. 比率分析主要用于对两个_____间变量值_____的描述分析，并生成比率变量以及比率变量的一些基本描述统计量。

二、选择题

1. 频数分析是运用统计量和图形来描述多种类型的变量，下列_____选项不属于频率分析中的统计量。

- A. 百分位值 B. 分布指标
C. 集中趋势 D. 卡方

2. 描述统计中包括集中趋势、离散程度与分布形状，下列描述说法错误的为_____。

- A. 集中趋势是指一组数据向某一中心值靠拢的程度，反映了该组数据中心点位置
B. 离散程度是观测各个数值之间的差异程度，反映了各个数值之间的差异大小
C. 分布形态主要用于分析数据的集中趋势与离散程度，是描述分析的综合运用
D. 集中趋势统计主要是用于寻找数据水平的代表值或中心值

3. P-P 图和 Q-Q 图是一种用于概论分布的图表，用于确定某个变量的分布是否与给定的分布相匹配的，下列选项中描述错误的一项为_____。

- A. P-P 图是对照一些检验分布的累计比例，来绘制某个变量的累计比例图
B. 在 P-P 图中可以转换值的概率图，而在 Q-Q 图中则不可以
C. P-P 图中的检验分布方法与 Q-Q 图中的检验分布方法一致
D. Q-Q 图是对照一些检验分布的分位数，来绘制某个变量分布的分位数图

4. 在比率分析中，统计量中的【集中指数】选项组包括【中位数百分比之内】选项，该选项

区间的下界等于_____。

- A. $(1-0.01 \times \text{值}) \times \text{中位数}$
- B. $(1+0.01 \times \text{值}) \times \text{中位数}$
- C. $(1 \times 0.01 \times \text{值}) \times \text{中位数}$
- D. $(1/0.01 \times \text{值}) \times \text{中位数}$

5. 在 SPSS 中除了单独使用 P-P 图与 Q-Q 图来分析数据之外, _____分析方法中还将运用该图来显示分析结果。

- A. 频数
- B. 比率
- C. 探索
- D. 描述

三、问答题

1. 频数分析中都具有哪些统计量和参数?
2. 描述分析方法中的集中趋势都包括哪些度量?

3. 如何使用 Q-Q 图来分析数据?

4. 什么是茎叶图与箱图?

5. 如何使用交叉分析方法对不同数据层中的数据进行分析?

四、上机练习

1. 描述分析销售数据

本练习运用 SPSS 中的描述分析方法, 分析一年内销售数据的中心趋势, 以及到中心值的变异或离散程度, 如图 3-55 所示。首先, 在 SPSS 中制作数据变量, 并依据变量类型依次输入销售数据。然后, 执行【分析】|【描述统计】|【描述】命令, 添加变量并设置统计选项为均值、标准差、最小值、最大值、峰度与偏度。最后, 单击【确定】按钮, 在输出窗口中显示分析结果。

描述统计量									
	N	极小值	极大值	均值	标准差	偏度		峰度	
	统计量	统计量	统计量	统计量	统计量	统计量	标准误	统计量	标准误
销售收入	12	260	400	316.33	39.320	1.074	.637	1.086	1.232
有效的 N (列表状态)	12								

图 3-55 销售数据分析结果

2. 探索性分析考试成绩

本练习将运用描述统计分析中的探索性分析方法, 分析考试成绩中的总成绩, 如图 3-56 所示。首先, 制作分析数据表。然后, 执行【分析】|【描述统计】|【探索】命令, 将“总分”变量添加到【因变量列表】列表框中, 将“所属部门”变量添加到【因子列表】列表框中。同时, 单击【统计量】按钮, 启用【描述性】、【界外值】与【百分位数】复选框。最后, 单击【绘制】按钮, 启用【带检验的正态图】复选框, 并选择【按因子水平分组】选项。

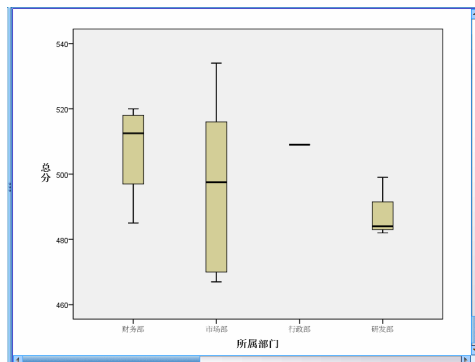


图 3-56 考试数据探索性分析结果