第四章 连续变量的统计描述与参数估计

统计分析的目的是研究总体特征。但是，由于各种各样的原因，研究者能够得到的往往只能是从总体中随机抽取的一部分观察对象，它们构成了样本。只有通过对样本的研究，才能对总体的实际情况做出可能的推断。因此，在数据收集、整理完毕后，进行深入分析之前，首要的工作就是去了解这个数据的整体情况，通过数据来掌握一定的行业背景，随后才能考虑作深入的推断。

用少量数字（即描述指标）概括大量原始数字，对数据进行描述的统计方法即为描述性统计分析。所谓描述性统计分析，是针对统计学的另一大类——推断性统计分析而言的，后者指从样本信息来回推总体特征。在第二章中我们讲到，变量按其测量类型可以分为：Nominal变量（即名义型）、Ordinal变量（即定序型）和Scale变量（即定距型）。针对不同测量类型的变量（属性、字段），有不同的描述指标体系和统计图形与之对应。本章将讲述Scale变量，或者说连续变量的统计描述，而下一章将讲述Nominal变量和Ordinal变量，以及多选题的统计描述。

4.1 连续变量的统计描述概述

当数据量较少时，如只有5个人的身高，或者7个人的性别资料时，研究者可以通过直接观察原始数据来了解几乎所有的信息。但是，实际工作中所接触到的数据量往往要远大于人脑可以直接处理、记忆的容量。这是就必须要借助于各种统计指标来辅助完成对数据的描述工作了。而为了方便统计指标的应用，又以此为基础衍生出了各种描述用工具，最终再使用各种统计软件来加以实现，而SPSS就是最常用的一种。

4.1.1 统计描述中可用的工具

首先，在统计描述中最基本的工具就是列表进行原始数据的频数描述，特别是对于分类数据而言，频数表现在仍然是最常用的描述工具。但是，当数据量较大时，原始频数表显得过于冗长，如果希望深入发掘数据中蕴含的信息，则需要对数据加以浓缩汇总。

1．各种初步汇总描述方法：最直接的汇总描述方法就是将原始数据按照其大小进行分组汇总，计算各组段的频数大小，最终汇总成相应的分组频数表，或相应的分组直方图，汇总频数表可以反应出数据的大致趋势。除分段汇总以外，百分位数也能够对数据的分布特征进行刻画，多个百分位数组合起来，也能够反应出数据的分布特征来。但是分组汇总和百分位数对信息的利用仍然比较粗糙，均只能反映比较基础的信息，如果希望对数据的分布特征描述的更为简练，还需要更进一步。

2．各种统计描述指标：这实际上是更复杂的各种描述工具的基础，是针对数据的某种特征进行精确的数字呈现的一系列指标。对于样本而言，这些统计描述指标也可被称

www.MedStatStar.com

为统计量。大家常用的统计描述指标在连续变量中有均数、标准差、四分位间距等，而在分类变量中则有比、率等。

3．统计表：当数据比较复杂，所计算的统计指标较多时，直接观察所计算出的数值比较麻烦，为此人们又会按照一定的排列方式将统计指标组织为一张表格，以方便使用，这就是所谓的统计表。在一张统计表中可以同时组织呈现多种统计指标，并进行复杂的样本分组、合并计算，是统计描述中常用的工具之一。

4．统计图：统计表虽然能做到非常精确、详细的对统计指标进行陈列，但是不够直观，如果希望结果更为直观一些，则可以按照统计指标的大小将其绘制为一张图形，这就是所谓的统计图。例如对于连续变量数据，我们常用直方图、箱图等工具加以展示。而对于分类变量，则常用条图、饼图等加以展示。

显然，统计表和统计图都是建立在各种统计描述指标的基础上的，因此本章和下一章将对统计描述指标体系做详细的讲解，而第六、七两章将进一步讲解如何利用统计指标制作统计表，第八、九两章则会讲解统计图的绘制方法。对于在本章和下一章中可能会提前涉及到的统计图形，本章将仅作简单解释，不详细讨论，请大家参阅随后各章的相应内容。

4.1.2 连续变量的统计描述指标体系

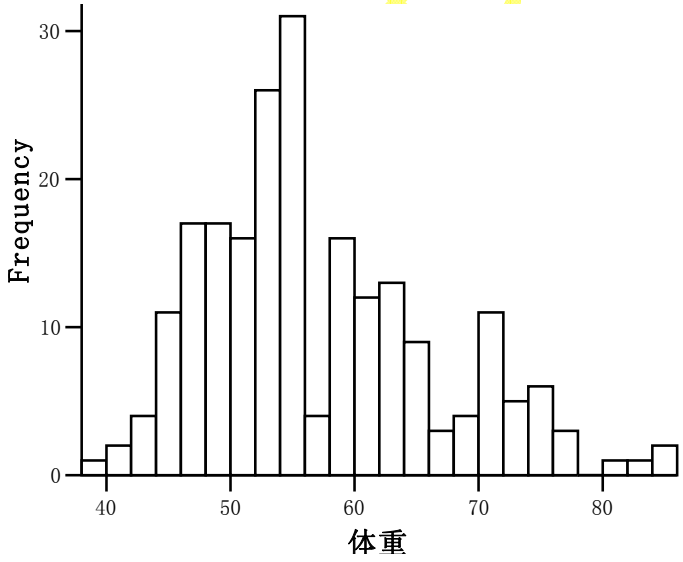


图4.1　体重的直方图

上图是对某人群体重分布情况绘制的直方图，这种图形是描述连续性变量最常用的工具，它实际上就是按照数据的大小将数值分成若干个组段，然后计算每个组段内的频数，最终用直条的高低反映出来，它可以直观的反应数据的分布状况。通过对这张图形的观察，我们可以发现如果要使用统计指标对该数据加以描述，则主要是表现以下几个趋势：

1．集中趋势(Central Tendency)：这个人群的平均体重是多少？这可能是人们希望了解的最基本的汇总信息，我们常说美国人比中国人高，这并不是说美国人比中国人都

高，比如姚明就要高于绝大多数美国人，这种说法实际上省略了“平均起来”这个定语。或者说，它实际上是关于数据的“中心位置”的某种表述。在统计学中，相应的用于描述集中趋势，或者说数据分布的中心位置的统计量就被称为位置统计量(location statistic)。常用的位置统计量有均数、中位数等，其中均数适用于正态分布和对称分布资料，中位数则适用于所有分布类型的资料，详后。

2．离散趋势(Dispersion Tendency)：显然，仅仅反映数据的集中趋势是远远不够的，上图中还反映出体重在该人群中的分散状况，最轻的不到40公斤，而最终的大约在90公斤上下。应当有某种指标可以反映数据波动范围的大小，这被称为数据的离散趋势。比如我们常说的某国的贫富分化严重，或者某国卫生资源分配的公平性很差，偏远地区还缺医少药的时候，大城市的CT等大型医疗设备却大量闲置，占用了大量资源。这些实际上都是在讨论数据的离散趋势，而描述该趋势的统计量就被称为尺度统计量（scale statistic）。常用的尺度统计量有标准差、方差、四分位数间距等，其中标准差、方差只适用于正态分布资料，而四分位间距则适用于各种分布类型的资料，详后。

3．分布特征(Distribution Tendency)：除以上两大基本趋势外，随着对数据特征了解的逐渐深入，研究者常常会提出假设，认为该数据所在的总体应当是服从某种分布的。那么，针对每一种分布类型，都可以由一系列的指标来描述数据偏离分布的程度。例如对于正态分布而言，偏度系数、峰度系数就可以用来反映当前数据偏离正态分布的严重程度。当然，相对而言，这些分布指标使用得较少。

4．其他趋势：统计描述中还会用于许多其它指标，如可同时反映集中趋势和离散趋势的百分位数指标(Percentile)，描述数据是呈单峰还是双峰分布，数据的分布是对称的还是偏态的，专门针对存在异常值的数据进行描述的M统计量（M-estimators）、极端值(Outlier)列表等，详后。

4.1.3 SPSS中的相应功能

SPSS的许多模块均可完成统计描述的任务，除各种用于统计推断的过程会附带进行相关的统计描述外，SPSS还专门提供了几个用于连续变量统计描述的过程，他们均集中在Descriptive Statistics子菜单中：

1．Frequencies过程：其特色是产生原始数据的频数表，并能计算各种百分位数。从图4.2可见，它所提供的统计描述功能非常全面，且对话框布置很有规律，基本上按照数据的集中趋势、离散趋势、百分位数和分布指标四大块将格描述指标进行了归类。有了上面的基础，读者使用它应当不存在任何的困难。

www.MedStatStar.com

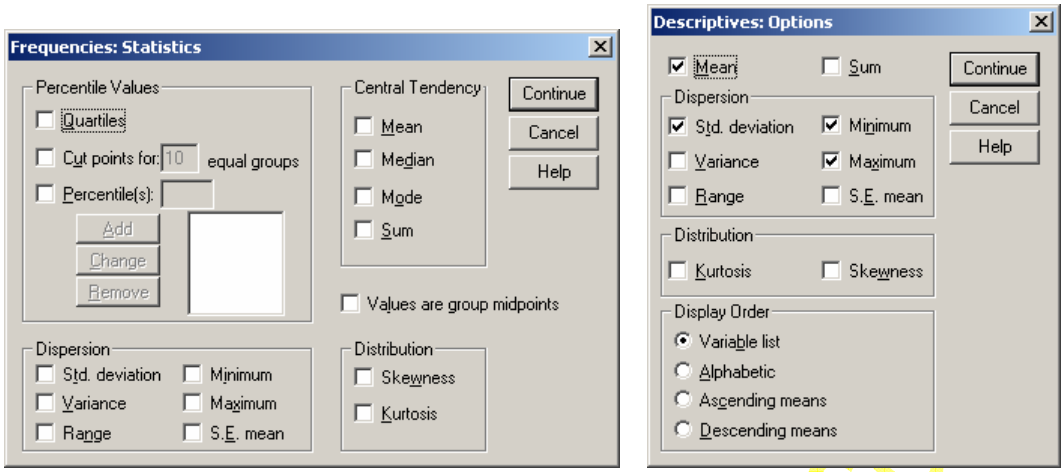


图4.2　Frequencies过程和Descriptive过程的统计选项子对话框

除统计指标外，Frequencies过程还可以为数据直接绘制相应的统计图，如用于连续型变量的直方图，用于分类变量的饼图和条图等。

2．Descriptive过程：该过程用于进行一般性的统计描述，相对于Frequencies过程而言，它不能绘制统计图，所能计算的统计量也较少，但使用频率却是最高的。实际上从统计选项可以看出，该过程适用于对服从正态分布的连续性变量进行描述。

3．Explore过程：顾名思义，该过程用于对连续性资料分布状况不清时的探索性分析，它可以计算许多描述统计量，给出各种统计图，并进行简单的参数估计。本章最后的分析实例就将以该过程为主加以讲解。

4．Ratio过程：功能比较特殊，用于对两个连续性变量计算相对比指标，它可以计算出一系列非常专业的相对比描述指标，相对而言使用面比较窄，因此本书将不对它做过多介绍，对此感兴趣的朋友请参见笔者前作《SPSS11统计分析教程》（基础篇）。

4.2 集中趋势的描述指标

怎样将一个变量的所有个体的值汇总为一个数字，使这个数字代表原数据的中心趋势或平均水平？统计学家提供了多种统计量来代表原始数据的中心趋势，如平均值、中位数和众数等。

4.2.1 算术均数

平均数用于反映一组数值的平均水平，包括算术均数、几何均数、调和均数等，但是以算术均数最为常用，往往也直接将算术均数，简称均数。

算术均数（Arithmetic Mean）是最常用的描述数据分布的集中趋势的统计量。总体均数(Population Mean)用希腊字母μ表示，样本均数常用 $\overline{X}$ 表示。

www.MedStatStar.com

算术平均数的定义和性质

实际上，大家从小学起就已经学习了相关的知识，对一组数据 $X_1,...,X_n$ 而言，其均数的算法为各数据直接相加，再除以总例数n，即：

$$\overline{X}=\frac{X_1+...+X_n}{n}=\frac{\sum X}{n}$$

显然，有各个变量值与均数离差之和等于零。即：

$$\sum(x-\bar{x})=0$$

算术平均数的这条数学性质说明，均数的实质是把总体各单位的差异全部抽象化，采用取长补短的方法把变量值小于平均数的负离差全部用大于平均数的正离差抵消补齐。

除上面的性质外，各个变量值与平均数离差平方之和为最小值。即：

$$\sum(x-\bar{x})^2=\min$$

算术平均数的这条数学性质说明，以任意不为平均数的数值为中心计算的离差平方和大于以平均数为中心的离差平方和，因此，算术平均数是误差最小的总体代表值。

均数的意义

任何一个平均数首先是同类现象的平均数，这是平均数的同质性。任何一个平均数总是一个平衡点。在这个平衡点的两边有多有少、有大有小、有高有低、有胖有瘦。而且总是多少相等，大小相同，高低适中，胖瘦相抵。这就是说，用平均数作为观测数据的代表在整体上是没有误差的，而且数学上可以证明，平均数的误差平方和也比其它任何一个数都小。统计学中著名的“最小二乘法”就是根据这个结论建立起来的。但是，由于平均数只是一个平衡点，如果两边加上或去掉相同的砝码，而不管砝码是多少这杆天平总能保持平衡。

平均数的最灿烂之处就在于它高度浓缩了数据的精华，使大量的观测数据转变为一个代表性数值。用平均数作为变量的集中值不仅考虑到变量值的频次、次序，而且还考虑到它的大小。数据资料中任何频次、次序和数值大小的变化，都会引起平均数的改变。因此它是灵敏的，也是对资料所提供信息运用得最为充分的。

但平均数在高度概括观测数据从而使问题简单话的同时，却丢失了某些有用的信息，一方面它把各个观测数据之间的差异性掩盖了起来，另一方面由于平均数对个别极端值反应比较灵敏，因而平均数在某些情况下可能具有一定的欺骗性，这时它就有可能走样。

均数的适用范围

虽然平均数对资料的信息利用最充分，但对严重偏态的分布，会失去它应有的代表性。例如，一个国家会因某些富翁的存在，使平均收入变得很高。假设某单位有6个人，5个员工，1个经理。员工的月收入分别是：360元、380元、400元、420元、440元，经理的月收入为40000元，他们的平均月收入为7000元。显然这时用平均数就不能很贴切的反

映他们收入的一般水平。所以，平均数的一个主要缺点是容易受极端值的影响。因此，对于偏态的分布，应使用中位数作为集中趋势的统计量。只有单峰和基本对称的分布资料，使用平均数作为集中趋势描述的统计量才是合理的。由于在统计技术中，发展更多的是平均数，而不是中位数或众数等。因此，我们应该更多地设法用上平均数，必要时可以考虑对数据进行变量变换，以达到对称分布的要求。

严格地讲平均数只适用于定距变量。但有时对于定序变量，如果求平均等级也可以使用平均数。对于定类变量，如果人为地把每一类赋予一个数值，例如用1代表男，2代表女。那么男性在总体中所占的比例，实际就是这种特殊的平均数。

4.2.2 中位数

中位数（Median）是将全体数据按大小顺序排列，在整个数列中处于中间位置的那个值。它把全部数值分成两部分，比它小和比它大的数值个数正好相等。

中位数的定义

对于未分组的原始资料，首先必须将数据按大小排序。设排序的结果为：

$X_1 \leqslant X_2 \leqslant X_3 \leqslant \ldots\ldots X_n$

则中位数就可以按下面的方式确定：

当n为奇数时，$M = X_{(n+1)/2}$　当n为偶数时，$M = \frac{1}{2}(X_{n/2} + X_{n/2+1})$

中位数作为分布数列中处于中等水平的代表值，能够将全部总体单位按数值的大小等分为两个部分。所以中位数又称为二分位数。

对于按照频数方式分组录入的资料，其中位数的确定方式相对复杂一些（主要涉及到了数值的内插），感兴趣的读者可以参看专业统计书籍。好在今天有计算机和统计软件的帮助，使我们摆脱了繁琐的计算过程。

中位数的适用范围

中位数是位置平均数，因此它不受极端值的影响，在具有个别极大或极小值的分布数列中，中位数比算术平均数更具有代表性。例如上面员工收入的例子，其中位数就是410元，显然要比均数更能够代表数据的集中趋势。

中位数适用于任意分布类型的资料，不过，由于中位数只考虑居中位置，其它变量值比中位数大多少或小多少，它是无法反映出来的。所以，用中位数来描述连续变量会损失很多信息。当样本量较小时，中位数会不太稳定，并不是一个好的选择。因此，对于对称分布的资料，分析者往往优先考虑使用均数，仅仅是对均数不能使用的情况才是用中位数加以描述。

中位数对于定序变量、连续变量都可以使用。对于定序变量来说，虽然有众数和中位数两种统计量可供选择，但是，由于众数不考虑变量的次序关系，在描述定序变量时会损失很多信息。因此，对于定序变量，应采用中位数来反映更多、更准确的信息。

4.2.3 其他集中趋势描述指标

www.MedStatStar.com

除上述最常用的两种指标外，在SPSS中还可以计算一些更为复杂和专业的统计描述指标，这里简介如下：

截尾均数

由于均数较易受极端值的影响，因此可以考虑将数据进行排序后，按照一定比例去掉最两端的数据，包括中部的数据来求均数。如果截尾均数和原均数相差不大，则说明数据不存在极端值，或者两侧极端值的影响正好抵消；反之，则说明数据中有极端值，截尾均数能更好的反映数据的集中趋势。

常用的截尾均数有5%截尾均数，即两端各去掉5%的数据。在SPSS中Explore过程可以自动计算5%截尾均数。

几何均数

几何均数适用于原始数据分布不对称，但经对数转换后呈对称分布的资料。例如医学中的血清滴度资料就常用几何均数描述其分布的集中趋势。样本几何均数常用G表示，其计算公式是：

$$G = \sqrt[n]{X_1 X_2 \cdots X_n}$$

利用对数的性质，上述公式可表达为：

$$G = \lg^{-1}\left(\frac{\sum \lg X}{n}\right)$$

可以发现，几何均数实际上就是对数转换后的数据lg X的算术均数的反对数。

在SPSS中，几何均数可以在Report子菜单中的四个报表过程中计算输出。

众数（Mode）

众数指的是样本数据中出现频次最大的那个数字，众数容易理解，也不受极端值影响，但不易确定，且没有太明确的统计特性。

众数适用于任何层次的变量，特别适用于单峰对称的情况。是比较两个分布是否相近首先要考虑的参数。但是，由于众数仅使用了资料中最大频次这一信息，所以它对资料的使用是不完全的，提供的信息是有限的。用它来反映连续变量会损失很多信息。对于多峰的图形分布，一般也不用它来描述。因此，我们在这里不做详细介绍，把它放在第五章：分类变量的统计描述里面。

在SPSS中，众数可以在Report子菜单和Tables子菜单的全部报表过程和制表过程中计算输出。

调和均数

调和均数用符号H表示，现在已经很少使用，它实际上是观察值X倒数之均数的倒数，常用于完成的工作量相等而所用时间不同的情况，主要用来求平均速度。实际上，中学物理中学习过的并联电路的总电阻就是各分电路电阻的调和均数，各原始数据的大小相差约悬殊，该均数的“调和”作用就越明显。

在SPSS中，调和均数可以在Report子菜单中的四个报表过程中计算输出。

www.MedStatStar.com

4.3 离散趋势的描述指标

和集中趋势一样，离散趋势也有一系列的描述指标，我们将就一些常用的指标一一加以讲解。

4.3.1 全距

全距（Range）又称为极差，是一组数据中最大值（Maximum）与最小值（Minimum）之差。它是最简单的变异指标：

$$R = x_{\max} - x_{\min}$$

极差反映的是变量分布的变异范围或离散幅度，在总体中，任何两个标志值之差都不可能超过极差。极差计算简单，含义直观，运用方便。但存在两点不足：一是它仅仅取决于两个极端值的水平，不能反映其间的变量分布情况，提供的信息太少；二是它容易受个别极端值的影响，不符合稳健性和耐抗性的要求。

一般情况下，全距只用于预备性检查，目的是大体上了解数据的分布范围，以便确定随后分析的方法。

4.3.2 方差和标准差

方差（Variance）和标准差（Standard Deviation）的定义

相对而言，方差和标准差的计算比较复杂，因此这里将从其计算原理开始谈起。首先，对于每个数据而言，其离散程度的大小就是和均数的差值，简称离均差，它可以用来描述个体的变异大小。那么，离均差值和能否表示整个样本的离散程度大小呢？答案是否定的，因为根据均数的性质，所有数据的离均差值和应当正好为0，这是由于大于均数和小于均数的离均差正好能够完全抵消。为此，可以考虑先将离均差取绝对值，然后再求和，这样就不会出现政府抵消的情况了。显然，离均差绝对值之和可以表示数据离散程度的大小。

但是，使用离均差绝对值之和来表示离散程度仍有不便之处，大家都知道绝对值符号在数学推导中是非常难处理的，该指标很难用来进行后续的统计推断，因此人们又改用将各离均差先平方再求和，这样仍然可以解决符号的问题，但同时又可以进行后续的数学推导，该指标被称为离均差平方和（sum of squares of deviations from mean, SS）。

离均差平方和在使用上比绝对值要方便一些，但是，它的大小显然是和样本量有关的，观察单位越多，该指标就会越大，因此如果要客观的反映变异程度的大小，就应当去除样本量的影响。为此将离均差平方和除以观察例数n所得，这就是方差：

$$\sigma^2 = \frac{\sum (x - \bar{x})^2}{N}$$

方差相当于平均了每个数据的离均差的平方值，从而克服了离均差平方和受样本含量影响的缺点。故方差可用于不同含量样本数据分布离散程度的比较。方差越大，数据分布离散程度越大。

www.MedStatStar.com

上面给出的实际上是总体方差的计算公式，对于更常见的样本数据而言，方差的计算公式有所不同：

$$S^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

其中的n-1被称为自由度(degree of freedom)，它描述了当 $\bar{x}$ 选定时n个X中能自由变动的X(变量值)的个数。由于公式中需要使用均数，这是一个限制条件，因此样本量为n的样本实际上能自由取值的x只有n-1个，最后一个数值可以通过均数反推出来。自由度在统计学中也是一个非常重要的概念，我们在后面还会反复遇到。

最后，方差在使用上还有一点小小的不便，就是量纲不合常理，以身高为例，原始数据的量纲为米，则方差的量纲就是其平方，即平方米，这显然很别扭。为此又将方差开平方，这就是所谓的标准差，总体和样本的标准差分别用σ和s来表示。标准差度量了偏离平均数的大小，相当于平均偏差，可以直接地、概括地、平均的描述数据变异的大小。对于同性质的数据来说，标准差越小，表明数据的变异程度越小，即数据越整齐，数据的分布范围越集中；标准差越大，表明数据的变异程度越大，即数据越参差不齐，分布越分散。

方差和标准差的适用范围

由于标准差和方差的计算涉及到每一个变量值，所以它们反映的信息在离散指标中是最全的，是最理想、最可靠的变异描述指标。方差还具有可加性，能够参与进一步的统计运算。不过，也正是由于标准差和方差的计算涉及到每一个变量值，所以，他们也会受到极端值的影响，当数据中有较明显的极端值时不宜使用。另外，他们在计算中实际上都使用了均数，因此实际上只有均数能反映集中趋势时才能使用方差和标准差来反映离散趋势。因此，实际上方差和标准差的适用范围应当是正态分布。

4.3.3 百分位数、四分位数与四分位间距

全距的数据最不可靠，因为全距只由数据中的两个极端数据来决定，其余数据均不起作用。为了尽量减少全距缺点，人们又使用了分位差。分位差是对极差指标的一种改进，是从变量数列中剔除了一部分极端值之后重新计算的类似于极差的指标。常用的分位差有四分位差、十分位差以及百分位差。这里以四分位差为例加以说明。

百分位数、四分位数与四分位间距的定义

百分位数(percentile)是一种位置指标，用P_x表示。一个百分位数P_x将一组观察值分为两部分，理论上有x%的观察值比它小，有(100-x)%的观察值比它大。前面所学习过的中位数实际上就是一个特定的百分位数，即P50。

除中位数外，常用的百分位数还有四分位数，它实际上是三个数值的总称，分别是P25、P50和P75分位数。这三个分位数正好是能够将全部总体单位按标志值的大小等分为四部分的三个数值，符号分别记为Q_1、Q_2和Q_3。在许多统计书籍中，也将第一个四分位数P25称为“下四分位数”；第三个四分位数P75称为“上四分位数”，分别用符号Q_L和Q_U表示。上、下四分位数的差值被称为四分位间距：

www.MedStatStar.com

$$Q.R = Q_3 - Q_1$$

显然，P25和P75这两个分位数间包括了中间50%的观察值，因此四分位间距既排除了两侧极端值的影响，又能够反映较多数据的离散程度，是当方差、标准差不适用时较好的离散程度描述指标。

同样的道理，我们还可以计算十分位差、百分位差等。它们的作用都是排除少数极端值对分布变异范围的异常影响。分位的程度越高，分位差所排除的极端值的比例就越小，保留的信息就越多。分位的程度越低，分位差所排除的极端值的比例就越大，保留的信息就越少。实际分析时，需要根据具体情况和要求选择使用。

四分位数与四分位数间距的适用范围

计算四分位差的直接目的是排除部分极端值对变异指标的影响，其计算可以看成是首先从总体分布中剔除最大和最小各四分之一的单位，再对剩下的总体半数单位计算“全距”。因此四分位间距可以适用于任意分布类型的资料，它与全距（极差）的区别在于计算范围较窄，反映的是处于分布中间半数单位的变异幅度。

百分位数并非由全部观察值总和计算而来，因此它不如均数和标准差精确，然而中间部分的百分位数因不受极端数据的影响，具有较好的稳定性。但是，靠近两端的百分位数只有在样本含量足够大的时候才比较稳定，例如当样本量为100例时，比P95大的数值只有5个，换言之，这5个数字就决定了P95的大小。显然，此时P95是很不稳定的。因此，当样本量较小时，不宜取太接近两端的百分位数。而当样本含量很少时，“百分”位数已名不符实，就更加不用考虑了。

最后需要指出的是，严格的讲百分位数并不应当被仅限于描述离散程度，显然，他也可以对数据的集中趋势等其它特征进行描述，而多个百分位数联合起来，实际上就可以完整的反映整个数据的分布规律。这一点在本章第一节已有所提及，这里再次强调一下。

4.3.4 变异系数

当需要比较两组数据离散程度大小的时候，往往直接使用标准差来进行比较并不合适。这可以被分为两种情况：

1．测量尺度相差太大：例如希望比较蚂蚁和大象的体重变异，蚂蚁的标准差单位是克，而大象的单位可能是吨，如果直接比较，显然永远都是大小的体重变异更大，但这显然是不合理的，因为体重相差一公斤对大象的体重而言根本就算不了什么，而蚂蚁则永远也做不到。

2．数据量纲不同：例如希望比较身高和体重的变异程度，两者的量纲分别是米和公斤，那么，究竟是1米大，还是2公斤大？根本就没法比较，完全是一笔糊涂账。

在以上情形中，就应当消除测量尺度和量纲的影响，而变异系数（coefficient of variation）就可以做到这一点，它是标准差与其平均数的比率。样本变异系数计算公式为：

$$CV = S / \overline{X}$$

计算出的CV没有量纲，同时又按照其均数大小进行了标化，这样就可以进行客观的比较。

4.4 连续变量统计描述实例

在系统学习了连续变量的统计描述指标体系后，下面我们将用一个具体的分析实例来看一下各种描述指标在SPSS中的实现方法。

4.4.1 数据背景介绍

本例是一次实际调查的部分问卷数据，调查对象为上海部分大专院校的大学生，文件名为student.sav。主要调查内容和封闭型题目的选项代码如下：性别（1 男，2 女）；出生年、月、日（具体数字）；身高（cm）；体重（kg）；血型（A、AB、B、O）；血型代码（1 A、2 AB、3 B、4 O）；教育背景（1 重点大学本科、2 普通大学本科、3 大专、4 中专/职校）；学科（1 文史、2 理工、3 其它）；男、女身高级别（1 低、2 中等、3 高，但两者的划分标准不一样）；男、女体重级别（1 轻、2 中等、3 重，两者的划分标准不一样）；季度（具体数字）。

需要说明的是，后面的5个变量：男生身高级别（hm）、女生身高级别（hf）、男生体重级别（wm）、女生体重级别（wf）和季度(quarter)，是我们通过SPSS的Recode过程，从前面的相应变量中，经过Into Different Variables…变换而来。

4.4.2 使用Explore过程进行分析

分析操作

这里以student.sav数据为例，对男性和女性身高数据分别进行描述，具体步骤如下：

Analyze➔Descriptive Statistics➔Explore
Dependent Variables框：height
Factor List框：sex
Plots…：
　Descriptive：☑ Histogram
　Continue
OK

相应的对话框如图所示，Dependent List框用于选入需要分析的变量，下方的Factor List框用于选入分组变量，从而将希望描述的变量按该因素的取值分组分析，本例中为性别。

www.MedStatStar.com

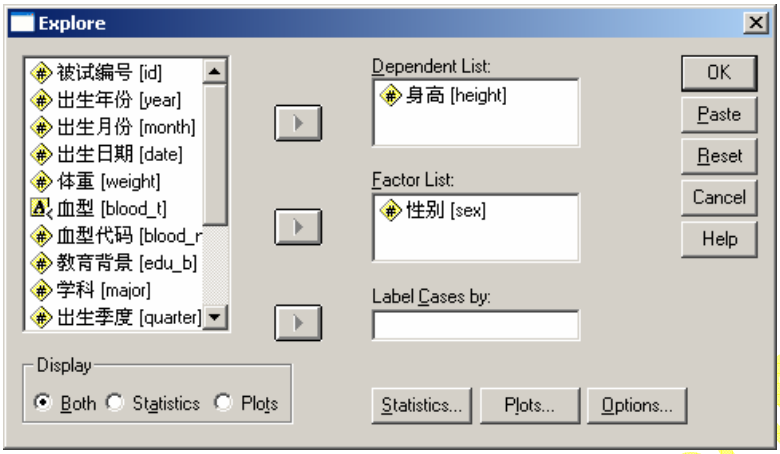


图4.3 对连续变量进行描述性分析的Explore过程主对话框

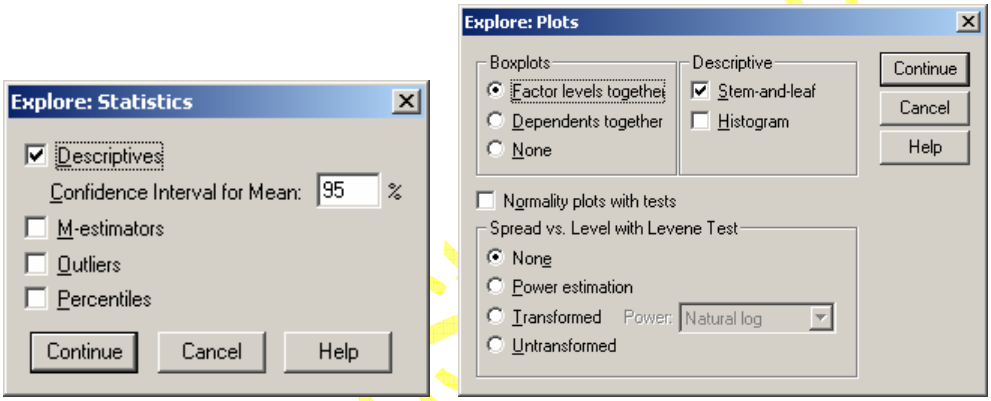


图4.4 Explore过程的Statistics和Plots子对话框

基本的分析结果

分析结果中首先会给出标题“Explore”，表明随后的输出都属于Explore过程。

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	性别	N	Percent	N	Percent	N	Percent
身高	男	69	95.8%	3	4.2%	72	100.0%
	女	146	99.3%	1	.7%	147	100.0%

首先是例行的处理记录缺失值情况报告，可见对于身高而言，男性、女性两组均存在缺失值，其中男性3例，女性1例，最终进入分析的各为69和146例有效值。

Descriptives

	性别			Statistic	Std. Error
身高	男	Mean		174.71	.671
		95% Confidence Interval for Mean	Lower Bound	173.37	
			Upper Bound	176.05	
		5% Trimmed Mean		174.70	
		Median		175.00	
		Variance		31.062	
		Std. Deviation		5.573	
		Minimum		159	
		Maximum		188	
		Range		29	
		Interquartile Range		8	
		Skewness		-.034	.289
		Kurtosis		.138	.570

记录汇总报告之后给出的就是身高的统计描述表格，因本例中的结果输出较长，为便于解释，这里仅给出表格上半部男性的分析结果。可见Explore过程的输出结果较多，这里依次解释如下：

1．集中趋势指标：首先可以看到69名男性学生的平均身高为174.71cm（mean），去掉两侧各5%的极端值后，截尾均数为174.70cm（5% Trimmed Mean），中位数为175cm（median）。对于对称分布，且不存在极端值的数据而言，均数、截尾均数和中位数应当基本相同，显然本例符合这种情况，因此从上述指标及可推测出数据应当是对称分布的。

2．离散趋势指标：身高的方差为31.062cm（Variance），其平方根即标准差，大小为5.573cm（Std. Deviation）。全部男生中最矮的为159cm（minimum），最高的为188cm（maximum）。两者之差即为全距29cm（range），中间一半的男生的身高差即为四分位间距8cm（interquartile range）。

3．分布特征指标：表格最下方还会给出表示数据偏离正态分布程度的偏度系数和峰度系数，及其各自的标准误，详后。

4．参数估计：以上结果实际上还会给出总体均数的参数估计结果，可见均数的标准误为0.671cm，相应的总体均数95%可信区间为173.37~176.05cm，详后。

女生身高情况请大家自己分析，这里不再详述。

在统计描述表格之后，Explore过程还会给出身高分性别的茎叶图和箱图，从图形分布上可以看出，分性别的升高基本上呈对称的分布状态。对这两种图形的介绍请读者参见第八、九两章，这里不再详述。

www.MedStatStar.com

输出百分位数和极端值列表

除默认的统计量输出外，Explore过程中还可以计算一些更深入的描述统计指标，如选中Statistic子对话框的Outliers复选框后，即可输出极端值列表如下：

Extreme Values

	性别			Case Number	Value
身高	男	Highest	1	180	188
			2	5	186
			3	154	186
			4	149	184
			5	150	183
		Lowest	1	20	159
			2	18	165
			3	11	165
			4	7	165
			5	183	167[a]

a. Only a partial list of cases with the value 167 are shown in the table of lower extremes.

这里同样只给出了男性的情况，表格中会输出五个最大值与五个最小值，以及这些数值所对应的记录号，从两侧极值的大小可见，在最大、最小两个方向上并没有特别明显的异常值，该结果同样支持前面得出的数据分布基本对称的结论。

如果选择Percentiles复选框，则会输出百分位数表如下：

Percentiles

			Percentiles						
		性别	5	10	25	50	75	90	95
Weighted Average(Definition 1)	身高	男	165.00	168.00	170.00	175.00	178.00	182.00	185.00
		女	155.00	156.70	159.00	163.00	166.00	170.00	172.00
Tukey's Hinges	身高	男			170.00	175.00	178.00		
		女			159.00	163.00	166.00		

上表会输出第5%、10%、25%、50%、75%、90%、95%分位数，并分别采用了两种算法，当数据量较大，且基本无重复值时，两法的结果相同，反之，则加权平均法会对数据进行内插，两种方法的结果会略有区别。

4.4.3 使用其它过程进行分析

上面我们使用Explore过程对数据进行了分析，下面来演示一下另外两个过程的分析结果。但是，由于另两个过程不能直接对身高进行分组描述，因此这里我们仅给出部分性别的分析结果，希望给出分组描述的读者可以先采用第三章介绍过的Select Cases过程进行数据拆分。

Descriptive过程的结果

该过程的操作非常简单，只需要将希望描述的变量选入即可，本例中身高的分析结果如下：

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
身高	215	151	188	166.67	7.668
Valid N (listwise)	215				

一望可知，这里的大部分内容都在上一节见过，因此就不再多解释了。

Frequencies过程的结果

Frequencies过程默认值给出原始频数表，如果希望得到各种统计量，则需要分析者自行加以指定。例如在上述的分析中，我们已经得到了描述集中趋势的均值、中位数等，以及描述离散趋势的方差、标准差、极差等统计量。如果还希望知道身高的具体四分位数及P5、P95百分位数是多少，则可以利用Frequencies过程来得到。具体步骤如下：

Analyze➔Descriptive Statistics➔Frequencies
Variables框：height
Statistics：
Percentile Value：☑ Quartiles
Percentile Value：☑ Percentiles: 5：Add | ☑ Percentiles: 95：Add
Continue
OK

Statistics

		身高
N	Valid	215
	Missing	4
Percentiles	5	155.80
	25	160.00
	50	165.00
	75	172.00
	95	180.00

从上表中我们可以看到，所有学生身高的四分位数为160、165和172cm。意味着，有1/4的学生身高矮于160cm，1/2的学生身高较165cm矮，1/4的学生身高高于172cm。另外，90%的学生的身高在155.8-180cm之间。

4.5 连续变量的参数估计

4.5.1 正态分布

正态分布的定义

正态分布的特征

标准正态分布

偏度和峰度

4.5.2 参数的点估计

矩法

极大似然法

稳健估计值

4.5.3 参数的区间估计

标准误

区间估计的计算

思考与练习

1．请就student.sav数据，分析学生的体重分布情况，尝试分性别和合并描述。

2．使用Descriptive过程，对student.sav中的身高和年龄变量进行标准正态变换，对变换后的变量进行统计描述。

参考文献

1. 吴喜之主编. 统计学基本概念和方法. 北京：高等教育出版社. 2003
2. 杨树勤主编. 中国医学百科全书 • 医学统计学分册. 上海：上海科学技术出版社. 1982
3. 杨树勤主编. 卫生统计学. 第三版. 北京：人民卫生出版社. 1995
4. 方积乾主编. 卫生统计学. 第五版. 北京：人民卫生出版社. 2003
5. 陈希孺. 数理统计学简史. 长沙：湖南教育出版社. 2002
6. 张文彤主编. SPSS11统计分析教程（基础篇）. 北京：北京希望电子出版社. 2002

www.MedStatStar.com