附录五　SPSS 在统计分析中的应用

§1　SPSS 软件基础

1.1　SPSS 概论

进行统计分析时，设计到的变量和样本数据很多，计算量很大。靠手工方法进行统计计算是不现实的，不借助于计算机难以实现，只有计算机才能快速得到精确的结果。在微机上使用的统计软件有许多种，在实际工作中应用比较普遍的主要有 SPSS、SAS、TSP、EViews、BMDP、TPL、CENTS、DET、SP、SARP、Excel、Lotus 1-2-3、Matlab、S-plus、Minitab 等。

SPSS 是英文 Statistical Package for the Social Science（社会科学统计软件包）的缩写。20 世纪 60 年代，美国斯坦福大学的三位研究生研制开发了最早的统计分析软件 SPSS，同时成立了 SPSS 公司，并于 1975 年在芝加哥组建了 SPSS 总部。20 世纪 80 年代以前，SPSS 统计软件主要应用于企事业单位。1984 年 SPSS 总部首先推出了世界第一个统计分析软件微机版本 SPSS/PC+，开创了 SPSS 微机系列产品的开发方向，极大地扩充了它的应用范围，并使其能很快地应用于自然科学、技术科学、社会科学的各个领域。SPSS 名为社会科学统计软件包，这是为了强调其在社会科学应用的一面（因为社会科学研究中的许多现象都是随机的，要使用统计学来进行研究），而实际上广泛应用于经济学、社会学、生物学、教育学、心理学、医学以及体育、工业、农业、林业、商业和金融等各个领域。

SPSS 现已推广到各种操作系统的计算机上，它和 SAS、BMDP 并称为国际上最有影响的三大统计软件。在国际学术界有条不成文的规定，即在国际学术交流中，凡是用 SPSS 软件完成的计算和统计分析，可以不必说明算法，由此可见其影响之大和信誉之高。

SPSS 的基本功能包括数据管理、统计分析、图表分析、输出管理等等。SPSS 统计分析过程包括描述性统计、均值比较、一般线性模型、相关分析、回归分析、对数线性模型、聚类分析、数据简化、生存分析、时间序列分析、多重响应等几大类，每类中又分好几个统计过程，比如回归分析中又分线性回归分析、曲线估计、Logistic 回归、Probit 回归、加权估计、二阶段最小二乘法、非线性回归等多个统计过程，而且每个过程中又允许用户选择不同的方法及参数。SPSS 也有专门的绘图系统，可以根据数据绘制各种图形。

SPSS 运行方式灵活，主要有四种方式：

（1）批处理方式

（2）完全窗口菜单运行方式

这种方式通过选择窗口菜单和对话框完成各种操作。用户无须学会编程，简单易用。

（3）程序运行方式

这种方式是在语句（Syntax）窗口中直接运行编写好的程序或者在脚本（Script）窗口中运行程序的一种运行方式。这种方式要求掌握 SPSS 的语句或脚本语言。

（4）混合运行方式

混合运行方式指以上各种方法的结合方式。

1.2　SPSS 基本操作

使用 SPSS 进行统计分析时，首先要录入数据或者打开一个已经存在的数据文件，根据需要进行数据转换；然后选择合适的统计分析过程，选择统计分析所采用的方法和

参数；最后分析 SPSS 输出的结果，并保存结果。

1.2.1. 数据管理（Data Management）

启动 SPSS 后，出现的界面是数据编辑器窗口，它的底部有两个标签：Data View（数据视图）和 Variable View（变量视图），它们提供了一种类似于电子表格的方法，用以产生和编辑 SPSS 数据文件。Data View 用于查看、录入和修改数据，Variable View 定义和修改变量的定义。如果使用过电子表格如 Microsoft Excel 等，那么数据编辑窗口的许多功能应该已经熟悉。但是，还有一些明显区别：（1）列是变量，即每一列代表一个变量（Variable）或一个被观测量的特征。（2）行是观测，即每一行代表一个个体、一个观测、一个样品，在 SPSS 中称为事件（Case）。（3）单元包含值，即每个单元包括一个观测中的单个变量值。单元（Cell）是观测和变量的交叉。与电子表格不同，单元只包括数据值而不能含公式。（4）数据文件是一张长方形的二维表。数据文件的范围是由观测和变量的数目决定的。可以在任一单元中输入数据。如果在定义好的数据文件边界以外键入数据，SPSS 将数据长方形延长到包括那个单元和文件边界之间的任何行和列。

如果要分析的数据还没有录入，可用数据编辑器来键入数据并保存为一个 SPSS 数据文件（其默认扩展名为 sav）。

1.定义变量

输入数据前首先要定义变量。定义变量即要定义变量名、变量类型、变量长度（小数位数）、变量标签（或值标签）和变量的格式，步骤如下：单击数据编辑窗口中的 Variable View 标签或双击列的题头（Var），显示变量定义视图，在出现的变量视图中定义变量。每一行存放一个变量的定义信息，包括 Name、Type、Width、Decimal、Label、Value、Missing、Columns、Align、Measure 等。

（1）Name：定义变量名

变量名必须以字母或字符@开头，其它字符可以是任何字母、数字或_、@、#、$等符号。变量名总长度不能超过 8 个字符（即 4 个汉字）。

（2）Type：定义变量类型

SPSS 的主要变量类型有：Numeric（标准数值型）、Comma（带逗号的数值型）、Dot（圆点作小数点的数值型）、Scientific Notation(科学记数法)、Date(日期型)、Dollar（带美元符号的数值型）、Custom Currency（自定义型）、String（字符型）。单击 Type 相应单元中的按钮，选择合适的变量类型并单击 OK。

（3）Width：变量长度

设置数值变量的长度，当变量为日期型时无效。

（4）Decimal：变量小数点位数

设置数值变量的小数点位数，当变量为日期型时无效。

（5）Label：变量标签

变量标签是对变量名的进一步描述，变量只能由不超过 8 个字符组成，8 个字符经常不足以表示变量的含义。而变量标签可长达 120 个字符，变量标签对大小写敏感，显示时与输入值完全一样，需要时可用变量标签对变量名的含义加以解释。

（6）Value：变量值标签

值标签是对变量的每一个可能取值的进一步描述。

（7）Missing：缺失值的定义方式

SPSS 有两类缺失值：系统缺失值和用户缺失值。在数据长方形中任何空的数字单元都被认为系统缺失值，用点号（●）表示。SPSS 可以指定那些由于特殊原因造成的信息缺失值，然后将它们标为用户缺失值，统计过程识别这种标识，带有缺失值的观测被特殊处理。默认值为 None。单击 Value 相应单元中的按钮，可改变缺失值定义方式。

（8）Column：变量的显示宽度

输入变量的显示宽度，默认为 8。

（9）Align：变量显示的对齐方式

选择变量值显示时的对齐方式：Left（左对齐）、Right（右对齐）、Center（居中对齐）。

（10）Scale：变量的测量尺度

根据变量测量精度不同，可把变量由低到高分为四种尺度：定类变量、定序变量、定距变量和定比变量。

1）定类变量

定类变量由称为名义（nominal）变量。这是一种测量精度最低、最粗略的基于“质”因素的变量，它的取值只代表观测对象的不同类别，例如“性别”变量、“职业”变量等都是定类变量。定类变量的取值称为定类数据或名义数据。定类数据的共同特点是用不多的名称来加以表达，并由被研究变量每一组出现的次数及其总计数所组成，这种数据是枚举性的，即由计数一一而得。唯一适合于定类数据的数学关系是“等价关系”。因而，在定类数据中，同一组内各单位是等价的，同时若更换各不同组的符号并不会改变数据原有的基本信息。因此，最常用来综合定类数据的统计量是频数、比率或百分比等。

2）定序变量

定序变量由称为有序（ordinal）变量、顺序变量，它的取值大小能够表示观测对象的某种顺序关系（等级、方位或大小等），也是基于“质”因素的变量。例如：“最高学历”变量的取值是：1－小学及以下、2－初中、3－高中、中专、技校、4－大学专科、5－大学本科、6－研究生以上。由小到大的取值能够代表学历由低到高。定序变量的取值称为定序数据或有序数据。适合于定序数据的数学关系是“大于（>）”和“小于（<）”关系。在定序数据中，同一组内各单位是等价的，相邻组之间的单位是不等价的，它们存在“大于”或“小于”的关系。而且进行保序变换（或称单调变换），不改变数据原有的基本信息即等级顺序。最适合用于综合定序数据取值的集中趋势的统计量是中位数。

3）定距变量

定距变量又称为间隔（interval）变量，它的取值之间可以比较大小，可以用加减法计算出差异的大小。例如，“年龄”变量，其取值 60 与 20 相比，表示 60 岁比 20 岁大，并且可以计算出大 40 岁（60-20）。定距变量的取值称为定距数据或间隔数据。定距数据是一些真实的数值，具有公共的、不变的测定单位，可以进行加减乘除运算。定距数据的基本特点是两个相同间隔的数值的差异相等。对于定距数据，不仅可以规定“等价关系”以及“大于关系”和“小于关系”，而且也可以规定任意两个相同间隔的比值或差值。如果将每个数值分别乘以一个正的常数再加上一个常数，即进行正线性变换，并不影响定距数据原有的基本信息。因此，常用的统计量如均值、标准差、相关系数等都可直接用于定距数据。

4）定比变量

定比变量又称为比率（ratio）变量，它与定距变量意义相近，细微差别在于定距变量中的“0”值只表示某一取值，不表示“没有”。例如，人的身高就是一个定比变量，如果身高值为“0”米，则表示这个人不存在。定比变量的“0”值表示“没有”。而在测定温度的摄氏表中，$0^{\circ}C$ 并不表示没有温度，因为还有在零度以下的温度。定比变量的取值称为定比数据或比率数据。定比数据也同样可进行算术运算和线性变换等。通常对定距变量和定比变量不需要加以区别，两者统称为定距变量或间隔变量。

一般地，定类变量和定序变量用于描述定性数据，属于定性变量；而定距变量和定比变量用于描述定量数据，属于定量变量。

同其它分类标准一样，一个变量在不同分析中可当作不同尺度的变量。例如，“年龄”在某些分析中（如回归分析）当作定距变量，而在另外一些分析中（如方差分析）可通过分组作为定类变量处理。

如果变量为定距变量或定比变量，则在 Scale 相应单元的下拉列表中选择 Scale；如果变量为定序变量，则选择 Ordinal；如果变量为定类变量，则选择 Nominal。

如果有许多个变量的类型相同，可以先定义一个变量，然后把该变量的定义信息复制给新变量。具体操作为：先定义一个变量，在该变量的行号上单击右键，弹出快捷菜单，选择 Copy；然后用鼠标右键选择多行，弹出快捷菜单，选择 Paste；再把自动产生的新变量名称（如 Var00001、Var00002、…）改为所要的变量名称。

定义了所有变量后，单击 Data View 即可在数据视图中输入数据。

2．数据的输入与编辑

定义了变量后就可以输入数据了。由于各种原因，已经输入的数据往往会有错误，这就需要进行编辑。用 Windows 的基本操作方式可实现对数据的编辑。如果数据文件较大且知道要修改的数据单元的行号，可通过选择 Data＝>Go to Case 打开对话框，在对话框中 Case Number 的右框输入行号来查找特定观测（行）。如果要查找某变量中的特定值或值标签，选择该变量，再选择 Edit＝>Find 或者按 Ctrl+F 打开对话框，在 Search for 右框中输入要查找的数值或标签。

3．数据转换

在理想情况下，输入的原始数据完全适合要执行的统计分析模型，遗憾的是，这种情况很罕见，经常需要通过数据转换来提示变量之间的真实关系。利用 SPSS 可进行从简单到复杂的数据转换。

（1）根据已存在的变量建立新变量

选择 Transform=>Compute，打开 Compute Variable（计算变量）对话框。在对话框中的 Target Variable（目标变量）下框中输入符合变量命名规则的变量名，目标变量可以是现存变量或新变量。对话框中 Numeric Expression（数值表达式）下的文本框用于输入计算目标变量值的表达式。表达式能够使用左下框列出的现存变量名、计算器板列出的算术运算符和常数和 Functions（函数）列表框显示的各种函数等。可以在文本框中直接输入和编辑表达式，也可以使用变量列表、计算器板和函数列表将元素粘贴到文本框中。

计算器板包括数字、算术运算符、关系运算符和逻辑运算符，可以象使用计算器一样使用它们。

函数表有 70 多个函数，包括算术函数、统计函数、分布函数、逻辑函数、日期和时间汇总与提取函数、缺失值函数、字符串函数、随机变量函数等等，例如对数函数 LN（）、绝对值函数 abs（）、求和函数 sum（）等。

计算器板下面有一个 IF 按钮，单击该按钮打开条件表达式对话框。在条件表达式对话框中指定一个逻辑表达式，一个逻辑表达式对每一个观测（case）返回真、假或缺失值。如果一个逻辑表达式的结果是真，就把转换应用于那个观测；如果结果是假或缺失值，就不对那个观测应用转换。

（2）对观测（case）记录进行排序

在数据文件中，可根据一个或多个排序变量的值重排观测的顺序。选择 Data=>Sort Cases，打开 Sort Cases 对话框，对选定变量的数据按升序或降序进行排列。

（3）观测或变量转置

SPSS 中将行作为观测，列作为变量。对那些观测和变量的行列关系与此相反的数

据文件，可以选择 Data=>Transpose 将行列互换。

（4）文件合并

可以将两个或更多个数据文件合并在一起,即可将具有相同变量但观测不同的文件合并，也可将观测相同变量不同的文件相合并。选择 Data=>Merge Files=>Add cases 从第二个文件即外部 SPSS 数据文件相当前工作数据文件追加观测。选择 Data=>Merge Files=>Add Variables 合并含有相同观测但不同变量的两个 SPSS 外部文件。

（5）选取观测子集

可以选择 Data=>Select Cases 根据包含变量和复杂的表达式的准则把统计分析限于某一特定观测子集，也可选取一个随机观测样本。这样就可以同时对不同的观测子集做不同的统计分析。

（6）其它转换

数据汇总，Data=>Aggregate；

数据加权，Data=>Weight Cases；

数值编码，Transform=>Recode；

数据求秩，Transform=>Rank Cases；

产生时间序列，Tranform=>Create Time Series；等等。

4．保存数据文件

在数据文件中所做的任何变化都仅在这个 SPSS 过程期间保留，除非明确地保存它们。要保存对前面建立的数据文件进行的任何改变，选择 File=>Save 或按 Ctrl+S 快捷键即可。如果要把数据文件保存为一个新文件或将数据以不同格式保存，可选择 File=>Save As，打开保存对话框，可以保存成 SPSS 默认格式，Microsoft Excel 格式或其它数据库格式等等。

5．打开已经存在的数据文件

选择 File=>Open 或按快捷键 Ctrl+O，显示 Open File（打开文件）对话框。选择要打开文件的文件类型和文件名，单击“打开”。

1.2.2 统计分析（Statistical Analysis）

在 SPSS 中建立了数据文件或打开一个数据文件之后，选择正确的统计分析方法，是得到正确分析结果的关键步骤。统计分析过程在主菜单 Analyze（分析）中的下拉菜单中。

1.2.3 图形分析（Graphical Analysis）

统计图是用点的位置、线段的升降、直条的长短或面积的大小等方法来表达统计数据的一种形式，它可以把资料所反映的变化趋势、数量多少、分布状态和相互关系等形象直观地表现出来，以便于读者的阅读、比较和分析。统计图具有简明生动、形象具体和通俗易懂的特点。SPSS 的图形分析功能很强，许多高精度的统计图形可从 Analyze 菜单的各种统计分析过程产生，也可以直接从 Graph 菜单中所包含的各个选项完成。图形分析的一般过程为：建立或打开数据文件，若数据文件结果不符合分析需要，则必须转换数据文件结果；生成图形；修饰生成的图形，保存结果。常用的统计图形有条形图、线图、面积图、圆饼图、散点图、直方图、箱线图等等。其中统计图形有两种形式，一种为一般图形，另一种为交互式图形，交互式图形提供了更多的选项，可绘制出更强大的图形。

1.2.4 输出窗口（Output Management）

不管是统计分析还是图形分析，其结果都输出到新的窗口－Viewer 窗口或 Draft Viewer 窗口，SPSS 默认输出窗口为 Viewer 窗口。Viewer 窗口的左边是输出大纲视图，可以单击统计过程名称左边的“+”和“-”展开或收缩输出大纲，也可以拖动输出内容项目改变改变项目的位置。Viewer 窗口的右边显示具体的输出内容，一般通过文字、

表格、图形显示统计计算结果。许多输出结果以数据透视表（Pivot Table）的表格形式显示，数据透视表功能强大，便于用户自行定义所需格式。如果要查看数据透视表中某个统计术语的含义，双击该数据透视表，右击术语，在弹出的快捷菜单中选择 What's This，就可获得该术语的简单定义。用户可通过与操作 Windows 应用程序一致的方法使用 Viewer 窗口。

§2　统计数据的收集、整理与描述

2.1　统计数据的收集

统计数据的收集就是统计调查，它按研究的目的和要求，有组织地向调查对象收集相关的各种资料。为了保证统计数据资料的完整性、准确性和及时性，必须熟悉各种收集方法及各自的特点。

1．问卷调查

问卷是调查者向被调查者了解情况或征询意见时所运用的同一设计的调查表。绝大多数旨在收集定量数据的调查都要采用某种形式的问卷，才会使调查得以顺利完成，并获得令人满意的数据。

2．普查法

普查，是按照一定标准时间对普查对象的全部单位无一例外地逐个进行的调查。普查按门类划分，可分为人口普查、工业普查、商业普查、农业普查、第三产业普查等。普查按区域划分，有宏观、中观和微观之分。一般而言，我们经常提起的普查为宏观普查。

3．抽样调查

普查的覆盖面宽，但其耗费的人力、物力、财力太大，在统计调查中抽样调查更为常用。抽样调查是从调查对象的总体中，按照一定的抽样原则抽取一部分单位作为样本，并以对样本进行调查的结果来推断总体的方法。

根据抽样方法是否随机，可将抽样调查分为随机抽样和非随机抽样两大类。

4．典型调查

典型调查是从调查对象的总体中选取一个或几个有代表性的单位进行全面、深入的调查。调查单位可依不同调查目的选取企业、学校、个人、家庭等。

典型调查的目的就是通过对某个典型的深入分析来概括和反映全面。因此，典型调查要求典型对总体推断有一定的代表性，这也是典型调查的关键。典型的代表性可以从动态、静态两个方面来衡量。从动态上来讲，是指事物的发展趋势；从静态上来讲，是指事物的共同属性与差异。

5．观察法

观察法是观察者深入现场或进入一定环境，观察调查对象，获取第一手资料的方法。调查人员直接到调查现场，耳闻目睹顾客对市场的反映和公开言行，或者利用照相机、监视器等现代化器械间接地进行观察来收集资料等，都属于观察法。

观察法的特点就是从侧面观察被观察者的言行和反映，一般不直接向被调查人提出问题，所以，被调查者往往是在不知情的状况下被调查的。

6．实验法

实验法是研究者根据一定的研究目的，控制某种市场条件，或在人工环境中使一定的现象产生，通过观察、记录收集资料，以揭示其发生原因或规律的方法，是一种复杂、高级调查方法。

7．集体访谈法

集体访谈法是访问调查法的延伸和扩展，是调查者邀请若干被调查者，通过集体访

谈的方式了解有关情况或研究实用统计学有关问题的方法。

2.2　统计数据的整理

收集统计数据之后，要对获取的数据进行系统化、条理化地整理，以提取有用的信息。

1．统计分组

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，分组具有分和双重含义。

2．频数分布与频率分布

将数据按其分组标志进行分组的过程，就是频数分布和频率分布形成的过程。表示各组的次数称为频数，各组次数与总次数之比称为频率。频数分布就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相对应的分布次数两个要素构成。由分组标志序列和各组相应的频率构成频率分布。

在平面直角坐标系上，将分组标志作为横轴并将各组频数（频率）作为纵轴，给出各组的长方形图即直方图。与直方图相似作用的图示是折线图，它以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。

当所观察的次数很多，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，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 型分布曲线和U 型分布曲线。

3．累计频数分布与频数分布

为了统计分析的需要，有时为了观察某一数值以上或某一数值以下频数或频率之和，这就需要在基本分组的基础上绘出累计频数或累计频率。由表的上方向表的下方的频数或频率相加就称为“向下累计”，反之称为“向上累计”。

累计频率（或频率）分布曲线，可用以研究财富、土地和工资收入的分配是否公平。这种累计分布曲线图最早由美国洛伦茨博士（Dr. M. O. Lorenz）提出的，故又称洛伦茨曲线图。

例 1　某车间 30 名工人安每天加工某种零件数如表 1 所示。

表 1　某车间工人每天加工某种零件件数

工人编号	加工零件数	工人标号	加工零件数
1	106	16	97
2	84	17	103
3	110	18	106
4	91	19	95
5	109	20	106
6	91	21	85
7	111	22	106
8	107	23	101
9	121	24	105
10	105	25	96
11	99	26	105
12	94	27	107
13	119	28	128
14	88	29	111
15	118	30	101

在 SPSS 中进行频数（率）分析的步骤为：

1）定义工人编号和加工零件数的变量名分布为 NO 和 X，然后输入变量 NO 和 X 的原始数据。

2）选择 Analyze=>Descriptive Statistics=>Frequencies...，弹出 Frequencies 主对话框。现欲对 X 进行频数分析，在对话框左侧的变量列表中选 X，单击按钮使之进入 Variable(s)列表框，并选择 Display FrequencyTables 显示频数分布表。

3）可单击 Format...按钮弹出 Frequencies：Format 子对话框，在 Order by 栏中有四个选项：

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。

这里选 Ascending values 项后点击 Continue 钮返回 Frequencies 主对话框。

4）可单击 Statistics...按钮，弹出 Frequencies：Statistics 子对话框，并单击相应项目，在作频数表分析的基础上，附带作各种统计指标的描述，特别是可进行任何水平的百分位数计算。这里不选。

5）可单击 Charts...钮，弹出 Frequencies：Charts 子对话框，用户可选三种图形:直条图(Bar Charts)、饼图(Pie Charts)和直方图(Histogram)。这里选择 Histogram 项,并选择 With Normal Curve 要求绘制正态曲线。单击 Continue 按钮返回 Frequencies 主对话框，再单击 OK 钮即可得到（累计）频数（频率）分布表和直方图。

应该注意的是，SPSS 在未特别指定的情形下，直方图或频数分布表是按照原始数值逐一作频数分布的，这与日常需要的等距分组、且组数保持在一定数目的要求不符。因此，在调用 Frequencies 统计过程命令之前，可先对原始数据进行预处理：已知最小值为 84，最大值为 128，故可要求分成 5 组，起点为 80，组距为 10。选择 Transform=>Recode=>Into Different Variable...，在弹出的 Recode Into Different Variable 对话框中选定 X，单击按钮使之进入 Numeric Variable→Output Variable 列表框，在 Output Variable 栏的 Name 文本框中输入 X1，单击 Change 按钮表示生成新生成的变量名为 X1。单击 Old and New Values 按钮弹出 Record Into Different Variable：Old and New Values 子对话框，在 Old Value 选项中单击 Range 项，输入第一个分组的数值范围：80～89，在 New Value 栏内输入新值：80，单击 Add 按钮，依此将各组的范围及对应的新值逐一输入，最后单击 Continue 按钮返回，再单击 OK 按钮即完成。系统在原数据库中生成一新变量为 X1，这时再调用 Frequencies 统计过程将输出等距分组且组数为 5 的频数分布表。

2.3 统计数据的描述

将数据整理成频率（频数）分布后，数据的数量规律性就可以大致地呈现在分布的类型和特点上。但频数分布给予我们的是一个大致的分布形状，还缺少代表性的数量特征值精确地描述出不同的统计数据分布。作为统计数据的代表值，一个是分布的中心，反映分布的集中趋势，另一个是分布的形状，反映分布的离散程序。

2.3.1 分布的中心

定义分布的中心有许多不同的方式。这里介绍三种最常用的，即众数、中位数和平均数。

1．众数（mode）

众数表示流行、时兴之意，有众多的意思。因而一个分布的众数就定义为频数出现最多的变量值。在正态分布和一般的偏态分布中，分布曲线最高点所对应的数值即是众

数。如果没有明显的最高点，众数可以不存在。当然，如果有两个最高点，也可以有两个众数。众数很容易求得，一般只要看一眼即可。它特别使用于描述定类变量和定序变量的数据。定距变量的数据分组后也可近似地用某个组的组中值来表示众数的大小。但众数并不是一个描述中心的很好的代表值，它常常依赖于数据的分组情况，即分组数改变的话众数可能就会有较大的变化。而且众数也可能不唯一。

2．中位数（median）与分位数

中位数是数据排序后，位置在最中间的数值。显然，中位数将数据分成两半，一半数据比中位数大，一半数据比中位数小。用中位数来代表总体标志值的一般水平，可以避免代表值受数列中极端值的影响，稳定性比较好，有时更有代表性。

与中位数相似的还有四分位数（quartiles）、十分位数（decile）和百分位数（percentile）。中位数是将统计分布从中间分成相等的两部分，而四分位数就是将数据分布四等分的三个数值，其中中间的四分位数就是中位数。十分位数和百分位数分别是将数据分布十等分和一百等分的数值。

3．平均值（均值）（mean）

平均数是数据集中趋势的最主要测度值。

2.3.2　分布的形状

只从均值来看待数据是片面的，我们还必须考虑数据的分布形状。用于描述数据分布形状即分布关于其中心的波动程度的代表值有：极差、内距、方差和标准差等，它们描述了分布的离散程度和差异程度。

1．极差（range）

极差也称为全距，是最大值与最小值之间的距离，它是数据离散或差异程度的最简单测度值。

2．内距（Inter-Quartile Range，IQR）

内距又称为四分位差，是两个四分位数之差，即内距 IQR=高四分位数-低四分位数。与极差类似，内距也是由两个值之差决定的，也是不全面的。但由于这两个值之差代表了中间 50％部分的长度，所以比极差能更好地描述分布的特征。例如，若内距比较小，则说明数据比较集中在中位数附近；反之则比较分散。内距常和中位数一起用来描述一个定距特别是定序测量数据的分布。

3．方差（variance）和标准差（standard deviation）

2.3.3　偏度与峰度

前面讨论了分布的集中趋势和离散趋势。要全面了解分布的特点，仅了解分布的集中趋势和离散程度是不够的，还需要了解分布是否对称和集中趋势高低等特征。偏度和峰度就是对分布的进一步描述。

1．偏度

所谓偏度是指反映频数分布偏态方向和程度的测度。从方向上看，偏度分左偏和右偏两种。

2．峰度

所谓峰度，是指频数分布曲线高峰的形态，即反映分布曲线的尖峭程度的测度。在频数分布中，有的频数分布曲线与正态曲线相比是尖顶，有的则是平顶，峰度就是用来衡量频数分布曲线的高耸程度的一个数字特征。当峰度大于 3 时，表示分布曲线的高峰是尖顶高峰；当峰度小于 3 时，表示分布曲线的高峰是平顶高峰。

2.3.4　SPSS 操作

在 SPSS 中计算例 1 各种指标的步骤为：

1）定义加工零件数的变量名为 X，并输入原始数据。

2）选择 Analyze=>Descriptives Statistics=>Descriptives...，打开

Descriptives 主对话框。在主对话框左边列表中选定变量 X，单击按钮使之进入 Variable(s)列表框。

3)单击 Options...按钮，打开 Descriptives:Options 子对话框。选择均值(Mean)、总和（Sum)、标准差（Std. Deviation)、方差（Variance)、极差（Range)、最小值（Minimum)、最大值（Maximum)、偏度（Skewness）和峰度（Kurtosis)，选好后单击 Continue 按钮返回 Descriptives 主对话框，再单击 OK 按钮即可得到各种统计量的计算结果。

§3 由样本推断总体

统计推断（Statistical inference）就是根据样本的实际数据，对总体的数量特征作出具有一定可靠程度的估计和判断。统计推断的基本内容有参数估计和假设检验两方面。概括地说，研究一个随机变量，推断它具有什么样的数量特征，按什么样的模式来变动，这属于估计理论的内容，而推断这些随机变量的数量特征和变动模式是否符合我们事先所作的假设，这属于检验理论的内容。

下面给出 SPSS 中假设检验的实现方法。

SPSS 提供了计算指定变量的综合描述统计量的过程和对均值进行比较检验的过程。

（1）用于计算变量的综合统计量的 Means 过程

Analyze=>Compare Means=>Means

（2）用于单独样本的 t 检验过程

Analyze=>Compare Means=>One-Sample T Test

（3）用于独立样本的 t 检验过程

Analyze=>Compare Means=>Independent-Sample T Test

用于检验是否两个不相关的样本来自具有相同均值的总体。

（4）用于配对样本的 t 检验过程

Analyze=>Compare Means=>Paired-Sample T Test

用于检验两个相关的样本是否来自具有相同均值的总体。

例 2　分别测得 14 例老年性慢性支气管炎病人及 11 例健康人的尿中 17 酮类固醇排出量（mg/dl）如下，试比较两组均值有无显著性差别（$\alpha = 0.05$）。

表 2　类固醇排出量数据

病人	2.90	5.41	5.48	4.60	4.03	5.10	4.97	4.24	4.36	2. 72	2.37
	2.09	7.10	5.92								
健康人	5.18	8.79	3.14	6.46	3.72	6.64	5.60	4.57	7.71	4.99	4.01

（1）定义变量：把实际观察值定义为 X，再定义一个变量 G 来区分病人和健康人。输入原始数据，在变量 G 中，病人输入 1，健康人输入 2。

（2）选择 Analyze=>Compare Means=>Independent-Samples T Test，打开 Independent-Samples T Test 主对话框。从主对话框左侧的变量列表中选 X，单击按钮使之进入 Test Variable(s)列表框，选 G 单击按钮使之进入 Grouping Variable 框，单击 Define Groups 按钮弹出 Define Groups 定义框，在 Group 1 中输入 1，在 Group 2 中输入 2，单击 Continue 按钮，返回 Independent-Samples T Test 主对话框，单击 OK 按钮即完成。

检验结果如下，经 Levene 方差齐性检验：$F = 0.440$，p 值$= 0.514$，$p > \alpha$，两总体方差无显著性差异。第三行表示方差齐性情况下的 t 检验的结果，第四行表示方

差不齐情况下的t检验的结果。依次显示t值（t-value）、自由度（df）、双侧检验p值（Sig 2-Tail）等。因本例属方差齐性，故采用第三行（即 Equal variances assumed）结果：$t=-1.807$，$p=0.084<0.1$，差异显著，即老年性慢性支气管炎病人的尿中 17 酮类固醇排出量低于健康人。

§4 方差分析

4.1 单因素方差分析

方差分析是检验两个总体或多个总体的均值间差异是否具有统计意义的一种方法。方差分析与回归分析之间存在一定的关系。对于方差分析，所有的自变量都被视为定类变量；而回归分析种，自变量可以是各种测度的变量（包括定类变量、定序变量、定距变量和定比变量）。事实上，经常把方差分析看作回归分析的一种特例，几乎所有方差分析模型可以由回归模型来表示，可以用回归分析的一般方法估计出相应的参数并进行推断。

为使方差分析更加有效，一般要假定所比较的总体具有相同的方差和正态分布。不过，方差分析方法在更宽的条件下也还是近似有效的。

我们通过一个具体的例子来说明方差分析的 SPSS 操作。

例 3　在 1990 年秋对“亚运会期间收看电视的时间”调查结果如表 3 所示。

表 3　三组居民的样本的态度得分

第一组	第二组	第三组
42	39	43
41	40	44
42	40	43
42	41	45
43	40	45

在 SPSS 中进行方差分析的步骤如下：

（1）定义“居民对亚运会的总态度得分”变量为 x（数值型），定义组类变量为 g（数值型），g=1，2，3 表示第一组、第二组、第三组。然后录入相应数据，如图 1 所示。

	g	x
1	1	42.00
2	1	41.00
3	1	42.00
4	1	42.00
5	1	43.00
6	2	39.00
7	2	40.00
8	2	40.00
9	2	41.00
10	2	40.00
11	3	43.00
12	3	44.00
13	3	43.00
14	3	45.00
15	3	45.00

图 1　方差分析数据格式

（2）选择 Analyze=>Compare Means=>One-Way ANOVA…，打开 One-Way ANOVA 主对话框。从主对话框左侧的变量列表中选定 x，单击按钮使之进入 Dependent List 框，再选定变量 g，单击按钮使之进入 Factor 框。单击 OK 按钮完成。

4.2 多因素方差分析

例 4 从由五名操作者操作的三台机器每小时产量中分别各抽取 1 个不同时段的产量,观测到的产量如表 4 所示。试进行产量是否依赖于机器类型及操作者的方差分析。

表 4 三台机器五名操作者的产量数据

	机器 1	机器 2	机器 3
操作者 1	53	61	55
操作者 2	47	55	51
操作者 3	46	52	49
操作者 4	50	58	54
操作者 5	49	54	51

SSPS 的操作步骤为：

（1）定义“操作者的产量”变量为 x（数值型），定义机器因素变量为 g1（数值型）、操作者因素变量为 g2（数值型），g1=1、2、3 分别表示第一、二、三台机器，g2=1、2、3、4、5 分别表示第 1、2、3、4、5 位操作者。录入相应数据，如图 2 所示。

	g1	g2	x
1	1	1	53.00
2	1	2	47.00
3	1	3	46.00
4	1	4	50.00
5	1	5	49.00
6	2	1	61.00
7	2	2	55.00
8	2	3	52.00
9	2	4	58.00
10	2	5	54.00
11	3	1	55.00
12	3	2	51.00
13	3	3	49.00
14	3	4	54.00
15	3	5	51.00

图 2 双因素方差分析数据格式

（2）选择 Analyze=>General Linear Model=>Univariate…，打开 Univariate 主对话框。从主对话框左侧的变量列表中选定 x，单击按钮使之进入 Dependent List 框，再选定变量 g1 和 g2，单击按钮使之进入 Fixed Factor(s)框。单击 OK 按钮就可以得到方差分析的结果，认为机器类型和操作者的影响均是显著的。

§5 相关分析

5.1　简单相关系数

1．简单相关系数的定义

简单相关分析是对两个变量之间的相关程度进行分析。单相关分析所用的指标称为单相关系数，又称为 Pearson（皮尔森）相关系数或相关系数。通常用 ρ 表示总体的相关系数，以 r 表示样本的相关系数。

前面我们已经给出总体相关系数的定义式为

$$\rho=\frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

样本相关系数的定义式是

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

2．简单相关系数的检验

在实际的客观分析研究中，相关系数一般都是利用样本数据计算的，因而带有一定的随机性，样本容量越小其可信程度就越差。因此也需要进行检验，即对总体相关系数 ρ 是否等于 0 进行检验。

数学上可以证明，在 X 与 Y 都服从于正态分布，并且又有 $\rho=0$ 的条件下，可以采用 t 检验来确定 r 的显著性。其步骤如下：

首先，计算相关系数 r 的 t 值：

$$t=\frac{r}{\sqrt{(1-r^2)/(n-2)}}$$

其次，根据给定的显著性水平和自由度 $(n-2)$，查找 t 分布表中相应的临界值 $t_{\alpha/2}$（或 p 值）。若 $|t|>t_{\alpha/2}$（或 $p<\alpha$）表明 r 在统计上是显著的。若 $|t|\le t_{\alpha/2}$（或 $p\ge\alpha$），表明 r 在统计上是不显著的。

例 5　某地区统计了机电行业的销售额 Y 和汽车产量 X（如表 5 所示），请使用 SPSS 计算 Y 与 X 的相关系数并进行显著性检验。

表 5　某地区机电行业销售额等数据

年份	销售额 Y（万元）	汽车 X（万辆）
1983	280.00	9.43
1984	10.36	10. 36
1985	337.4	14.50
1986	404.20	15.75
1987	402.10	16.78
1988	452.00	17.44
1989	431.70	19.77
1990	582.30	23.76
1991	596.60	31.61
1992	620.80	32.17
1993	513.60	35.09
1994	606.90	36.42
1995	629.00	36.58
1996	602.70	37.14

1997	656.70	41.3
1998	778.50	45.62
1999	877.60	47.38

解：（1）根据表 5 的数据创建 SPSS 数据文件。

（2）选择 Analyze=>Correlate=>Bivariate，在显示的对话框中，选择变量 Y 和 X 进入 Variables 框。采用默认设置，直接单击 OK 进行分析。

（3）计算结果如表6所示。

表6　相关系数结果输出

Correlations

		Y	X
Y	Pearson Correlation	1	.901**
	Sig. (2-tailed)		.000
	N	17	17
X	Pearson Correlation	.901**	1
	Sig. (2-tailed)	.000	
	N	17	17

**. Correlation is significant at the 0.01 level

从结果可以看出，Y 与 X 的相关系数 $r=0.901$，p 值＝0.000，在 $\alpha=0.01$ 水平下线性关系显著。

5.2 偏相关分析

在多变量的情况下，变量之间的相关关系是很复杂的。因此，多元相关分析除了要利用简单相关系数外，还要计算偏相关系数和复相关系数。这里仅讨论偏相关系数。

在对其它变量的影响进行控制的条件下，衡量多个变量中某两个变量之间的线性相关程度的指标称为偏相关系数。偏相关系数不同于简单相关系数。在计算简单相关系数时，只需要掌握两个变量的观测数据，并不考虑其它变量对这两个变量可能产生的影响。而在计算偏相关系数时，需要掌握多个变量的数据，一方面考虑多个变量相互之间可能产生的影响，一方面又采用一定的方法控制其它变量，专门考察两个特定变量的净相关关系。

在 SPSS 中计算偏相关系数的步骤是依次选择 Analyze=>Correlate=>Partial，再进行相关的操作即可。

§6　回归分析

6.1　一元线性回归分析

例 6　考虑家庭月可支配收入如何影响消费支出。进行了 10 观测，观测值如表 7 所示。

表 7　消费支出与可支配收入的观测值

消费支出 Y（千元）	可支配收入 X（千元）
1.6	2.0

2.0	2.5
2.3	3.0
2.4	3.5
3.0	4.0
3.2	4.5
3.1	5.0
3.5	5.5
3.6	6.0
4.4	6.5

在 SPSS 中进行一元线性回归方程估计的操作步骤为：

（1）建立数据文件，定义“消费支出”变量为 Y，定义“可支配收入”变量为 X，并录入相应数据。

（2）选择主菜单 Analyze=>Regression=>Linear，打开 Linear Regression 主对话框。在左边列表框中选定变量 Y，单击按钮，使之进入 Dependent 框，选定变量 X，单击按钮使之进入 Independent(s)框。

（3）单击 OK 按钮，得到如表 8 所示结果。

表 8　一元线性回归结果输出

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.607	.189		3.206	.013
	X	.542	.042	.977	12.832	.000

a. Dependent Variable: Y

可以看出，对于模型 $y = b_0 + b_1 x$，系数估计值分别为 $\hat{b}_0 = 0.607$，$\hat{b}_1 = 0.542$。

6.2　一元线性回归模型的检验

根据变量 X 和 Y 的样本观测值，应用最小二乘法求得了样本回归直线，作为总体回归的近似，这种近似是否合理，必须对其进行检验。如果通过检验发现模型有缺陷，则必须回到重新设定模型或估计参数。一元线性回归模型的检验包括经济意义检验、统计检验和计量检验。

1．经济意义检验

经济意义检验主要涉及参数估计值的符号和取值范围，如果它们与经济理论以及人们的实践经验不相符，就说明模型不能很好地解释现实的经济现象。例如，例 6 的家庭消费支出与可支配收入中，b_1 的取值范围应在 0 和 1 之间，如果估计出来的 b_1 小于 0 或大于 1，则不能通过经济意义检验。在对实际的经济现象进行回归分析时，常常会遇到经济意义检验不能通过的情况。造成这一结果的主要原因是：经济现象的统计数据无法象自然科学中的统计数据那样通过有控制的实验去取得，因而所观测的样本容量有可能偏小，不具有足够的代表性，或者不能满足标准线性回归所要求的假定条件。

2．统计检验

统计检验是利用统计学中的抽样理论来检验样本回归方程的可靠性，具体又分为拟合程度检验、相关系数检验、参数显著性检验（t 检验）和回归方程显著性检验（F 检验），是对所有现象进行回归分析时都必须通过的检验。

3．计量检验

计量检验是对标准线性回归模型的假定条件是否满足进行检验，具体包括序列相关检验、异方差性检验等。计量检验对于经济现象的定量分析具有特别重要的意义。

6.3 多元线性回归分析

例 7　某种商品的需求量 Y、价格 X1 和消费者收入 X2 的统计资料如表 9 所示，试估计 Y 对 X1 和 X2 的线性回归方程。

表 9　某商品的统计资料

年份	需求量 Y（吨）	价格 X1（元）	收入 X2（元）
1	59190	23.56	76200
2	65450	24.44	91200
3	62360	32.07	106700
4	64700	32.46	111600
5	67400	31.15	119000
6	64440	34.14	129200
7	68000	35.3	143400
8	72400	38.7	159600
9	75710	39.63	180000
10	70680	46.68	193000

用 SPSS 估计参数步骤如下：

1）在 SPSS 中输入变量数据，设变量名分别为 Y、X1、X2。

2）选择主菜单

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	62650.928	4013.010		15.612	.000
	x1	-979.057	319.784	-1.381	-3.062	.018
	x2	.286	.058	2.211	4.902	.002

a. Dependent Variable: y