

第一章 概述

§1.1 统计学和统计分析

统计学研究怎样以比较有效的方式收集、整理、分析带随机性的数据，并在此基础上，对所研究的问题作出统计性的推断，直至对可能作出的决策提供依据或建议[1]。简而言之，统计学研究数据的收集、整理、分析并做出推断[2]。

统计学的发展源远流长，社会各个领域和各门自然科学研究都离不开统计学方法。统计学在各个领域中的应用，不但促进了统计方法本身的发展，而且推动了各相关学科的发展。

统计分析是指用统计学的观点和方法对客观事物进行分析和研究，即数据的整理、描述和综合[2]。统计分析与信息提取存在密切关联，将收集的数据“去粗取精、去伪存真、由此及彼、由表及里”地加工处理，通过事物外在的数量表现，揭示事物可能存在的规律性，并结合各门专业知识加以解释。

统计分析常用的概念有：

1. 总体与样本。在实际应用中，把研究对象的全体叫做总体(population)，把每个研究单位叫做个体，把总体中个体的总数叫做总体容量。如果总体中包括有限个单位，就称为有限总体，否则为无限总体。从总体中随机抽出的一部分观察单位就称为概率样本(probability sample)，观察单位的数目为样本容量。统计分析的任务之一就是由概率样本推断总体。其它如自愿者组成的随意样本(haphazard sample)、指定的样本(representative samples)或分配样本(quota samples)等均是非概率样本。
2. 误差。总体中的个体之间存在着差异，这种差异可以由多方面原因引起。统计中的误差，是指测量值与真值之差或样本指标与总体指标之差。误差的出现，可以是实验仪器不准、研究对象不同质等导致的系统误差，也可以是由于疏忽造成的过失误差。由于观察单位间存在的个体差异，使得由样本指标与总体指标间存在差异，称为抽样误差。系统误差和过失误差是应该而且可以避免的。
3. 统计分布与统计量。对大量的带有随机性的统计资料的描述，我们得到资料波动变化的信息，统计学常用分布来刻画，如正态分布、二项分布及泊松分布等。设一个随机变量是一些独立同分布随机变量的函数，当样本给定时其值能够唯一确定，并且与总体的未知参数无关，则该随机变量称为统计量(statistics)。如果研究的总体只包含若干个未知参数则称其统计问题是参数性的，否则为非参的。不同的资料要用不同的统计方法来处理。总体与样本、参数与统计量的关系可用图 1.1来表示：
4. 变量分类。统计分析的原始数据通常是 n 行 p 列 $n \times p$ 的矩形数据矩阵 $X = \{x_{ij}\}, i = 1, \dots, n, j = 1, \dots, p$ ，它包含了关于 n 个对象的 p 个变量的信息，变量的统计学的意义是指一种可以测量的特征，统计分析时首先要对测量的方法加以考察。变量根据其测量的尺度不同，有表 1.1的分类[8]：

名义型(nominal)数据是一种纯粹的数据符号，没有量的概念，如婚姻状况中已婚、未婚、离婚、丧偶和分居，可任意记作1,2,3,4,5或A,B,C,D,E。有序型(ordinal)数据就是有先后次序，如由小到大的年龄分组。间隔型(interval)数据所包含的量不仅可以比较大

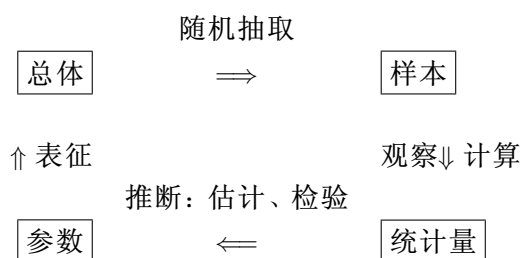


图 1.1 总体与样本、参数与统计量的关系[3]

表 1.1 各种变量的类型、意义与实例

名 称	意 义	举 例
名义型	类别间的地位相同	血型、是/否
有序型	类别可以排序	文化程度
间隔型	水平间的差值相同	温度、海拔高度
比率型	水平间比的等同	身高、体重

小，而且可以确定相差的量，温度是最典型的间隔型数据之一，即不仅可以比较温度的高低，而且可以说 10°C 比 20°C 低 10°C 。这一类数据可做复杂的四则运算，“间隔”的确定有一定的任意性，如摄氏和华氏两种温度的间隔就不相同。0点的确定是任意的，如 0°C 有确定的温度而不是没有温度。比率型(ratio)数据突出的特点是零点具有明确的含义，如重量尽管有不同的计量单位，但重量等于零时的概念很明确，而且任何一种计量单位可经一个比例常数简单地换算成另一种相应的单位。这些标度有几种情况[11]：a. 尺子上的值可表为比；b. 尺子上的距离有意义；c. 尺子上的元可以自小到大排列。间隔尺度无固定的原点，如海拔高度和温度的取值，仅仅是与一个相对的值比较，仅有b、c的性质。最后，仅有条件c成立的称做有序的，分类A,B,C,D不能说成立 $A-B=C-D$ 。名义变量不具有前面的三个性质。

变量的其它一些分类方法也与上述尺度有关。当变量测量的尺子无限可分时，称做连续的(continuous)，如长度由公里到米到厘米等等，否则是离散的(discrete)。名义的也称计数的(qualitative)，间隔和比率的也称计量的(quantitative)，有序的也称等级的。分类数据(categorical variable)度量尺度是不确定的，可以是名义的，或者是有序的，或者是间隔的。

统计学研究的课题有：

1. 抽样技术与实验设计(sampling technique and design of experiment)。统计分析中大量带有随机性的数据通常来自观察和实验，如医学中的临床试验、围绕某个专题进行的抽样调查等。统计抽样要保证收集到的样本数据在性质上和数量上对总体有代表性；实验设计包括专业设计和统计学设计。资料的收集至关重要，做不好这一环，则统计分析必是“垃圾进去，垃圾出来(garbage in and garbage out, GIGO)”。
2. 描述性分析(descriptive analysis)。简而言之，描述性分析是对数据的综合与提炼[6]。收集

的数据经过核实和归纳,计算出统计指标,或者结合统计图、统计表等手段表达出来,使对数据有一个概括的印象,就完成了数据的描述性分析。

复杂的分析常基于对数据基本的了解,统计原始数据的整理加工,传统上主要有三种方法: a.原始数据分组,从而获得各种统计图表; b.原始数据排序; c.按统计推断的要求,把原始数据归纳为一个或几个数字特征。D.R.Cox and E. J. Snell(1981)指出,典型的数据质量检查如: a.直观或自动检查在逻辑上不一致或与先验知识相悖的情况; b.检查主要变量的频数分布以找出少数有差异的数据; c.通过可能高度相关的变量对之间的散点图更敏感地检查; d.数据收集方法的检查找出测量上可能存在的偏差,如观察者间的差异,主要变量的编码出现的问题等; e.寻找缺失的记录,包括省略掉的那些记录,缺失值常用一种简便的方法标识,如99,999等,任何分析都不应包括这些数据。

探索性分析可以检测数据的错误、寻找数据存在的模式或关系。常见的数据错误有: a.数字位次搞错,1984写成了9184; b.某个变量的值用错,如用错变量,甚至数据集; c.把变量和观察关系搞混,录入不是按照观察而是按照变量; d.实验误差。常见的模式或关系有: a.线性关系; b.资料具有重复,宜采用方差分析; c.交叉或嵌套资料; d.时间趋势,如按时间的自然排序; e.边界点,在边界以外无观察; f.改变系统的点,即数据中出现一个冲激,此时应分别配合模型; g.异常点(outlier)。h.数据堆积(clump),即数据可分成两个或多个点的集合,从而把拟合的方程拉向自己。其他的方面如被研究量的缺失,实验随机化的影响、趋势外推等。在探索性数据分析中,常出现的概念有:局部稳定性(resistance)是指对于数据中局部出现的不正常变化的不敏感性,中位数相对于均数就是如此。稳健性(robustness)是指围绕特定概率模型假设下的偏离不敏感。

描述性分析对其它统计分析至关重要。统计指标在统计软件中往往以综合统计量(summary statistics)的形式出现,它是进一步分析的基础。如描述集中趋势的均数(如算术均数、几何均数、调和均数)与中位数,众数表示最大频数所在的位置。描述离散情况的标准差、全距、百分位差、偏度、峰度等,它们从不同的角度综合了数据的特性。图形分析(graphical analysis)如线图、圆图、直方图和直条图,茎叶图、箱式图、星形图、统计地图等,用于直观描述资料特征。对于多维的情况,可以用样本的相关系数、协方差阵描述它们之间的关系。数据的整理还包括一些变换,如Box-Cox转换等。探索性数据分析(exploratory data analysis, EDA)丰富了描述性分析的内容。

3. 统计推断(statistical inference)。基于收集的数据,以及对数据整理分析的结果,对数据所来自于总体的情况作出一定的论断,即是统计推断,也就是由样本作出关于总体结论的过程。简言之,统计推断就是由数据下结论[2]。统计推断可以用图 1.2来示意:

推断要解决三个问题,即由样本算出估计量,把它作为总体参数的点估计;对估计量算得一个确切的区间作为总体参数的区间估计,所估计的区间称为可信区间;应用统计检验来判断样本信息支持总体参数值是否理想。与推断有关的是统计决策问题,它是基于收集的数据,使用统计推断理论中提供的种种方法,结合经济上可能的后果进行分析。常用的估计方法有最大似然估计、最小二乘估计、贝叶斯估计等。

可信区间基于样本资料和 α 水平,指出在 $1-\alpha$ 的概率下该区间包含总体的参数, $1-\alpha$ 称为可信度,通常取为95%或99%。假设检验中把受到保护的假设定为原假设(null hy-

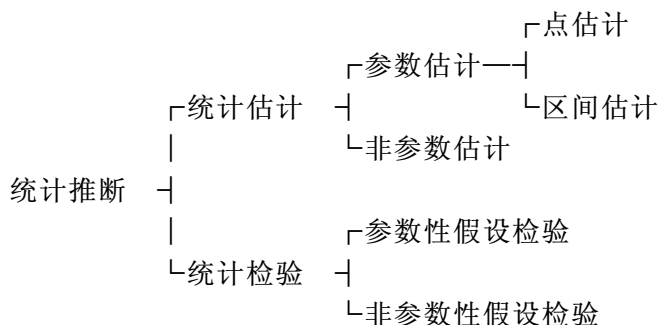


图 1.2 统计推断方法

pothesis, 也称无效假设), 对立的假设定为备择假设(alternative hypothesis)。根据样本提供的信息计算检验统计量, 在原假设下, 检验统计量取该值或比该值更极端的概率可以算出, 若概率值较小, 就拒绝原假设。一般, 取 $\alpha = 0.05$ 做为小概率事件。

假设检验可能犯两类错误: 当原假设为真但检验的结果否定其真实性, 称为 I 类错误, 常用 α 来表示; 当原假设为假而检验的结果不能够拒绝它, 称 II 类错误, 记其概率为 β 。检验的能力或功效可以用 $1 - \beta$ 表示。I 类错误由研究者控制, 等于设定的 α 水平。II 类错误受一系列因素影响: 备选假设离无效假设的距离较大时, 功效要高; 它也受 α 水平的影响, 常常要在二者之间有一个折衷。通常的做法是控制 I 类错误的条件下, 使 II 类错误尽可能地小。在多元分析中, 研究者倾向于对多个参数进行推断, 犯错误的概率就要大。若在给定的 α 水平下进行 k 个假设检验, 则当所有无效假设为真时, 以相同的水平完成 k 个检验, 接受所有假设的概率 $\geq 1 - k\alpha$; 若各检验独立, 则错误概率是 $1 - (1 - \alpha)^k$, 这样若在 0.05 水平上检验十个假设, 且这些检验所获 p 值的分布独立同分布, 则在原假设下得到一个特定检验为显著的的概率为 0.05 而声明十个检验至少一个为显著的概率为 0.401; 若是 20 个检验, 则后面的概率为 0.642。因此检验多个假设或求多维可信区间, 常常要控制任何一个 I 类错误的概率为 α , 这被称做是控制实验的(experimentwise) 或分析的(analysiswise) α 水平。有许多方法处理这种情形, 如 Bonferroni 检验(若 k 个假设在 α/k 水平检验, 则 I 类错误的概率不大于 α)、Scheffe 区间及其到 Roy 区间的推广、Tukey 检验及 omnibus 检验(对向量的检验)。在探索性研究或者当研究者想在资料中寻找有价值的东西时, 由于有许多机会出现阳性结果, 就应有控制 I 类错误的保守方法。当对所研究的领域有更多的把握时, 统计研究时的假设成为真正的假设, 可以冒风险对每个检验用 α 水平, 此时类似一般的假设检验。检验就是依照假设的(hypothesewise或comparisonwise 或parameterwise) 的 α 。

- 多元分析(multivariate analysis)。我们对客观现象进行观察和实验, 往往是多因素相互关联、相互影响的。对这些现象的分析, 便是多元分析研究的对象。由于这些方法的背景和处理对象不同而有多种分类方法。传统的多元与逐步回归、多元方差协方差分析、典型相关分析、主成分与因子分析、聚类分析、时间序列分析, 以及一些新的统计方法, 如回归中的有偏估计、Logit 分析与 Cox 回归和生存分析。多元方法处理问题的范围正逐渐扩大, 如计量经济学分析和群体遗传学分析。

多元分析是数理统计最重要的分支之一, 它的理论和方法在近半个世纪获得了飞

速的发展,尤其是最近十几年来由于计算机的普及,在许多领域和学科中,如生物、医学、地质、农业、工程技术、气象、社会经济等方面,得到日益广泛的应用,使得这个分支的发展更加活跃和深入。

统计研究主要的步骤是:

1. 初步的数据处理,即把数据归整为便于详细分析和数据检查的形式。2. 初步分析,目的在于搞清数据总的模式并提示下一步分析的方向,这可由简单的图表完成。3. 进一步分析,是论断的基础。4. 用准确、简明的形式表达得到的结论[5]。即明确问题、收集资料、分析资料、展示结果[6]。

根据表 1.1对变量的分类,统计分析大致有表 1.2的分类[8]:

表 1.2 统计分析的分类

因变量	自 变 量			
	名义的/有序的		间隔的/比率的	
	一个变量	多于一个	一个变量	多于一个
无因变量	χ^2 拟合度	关联度的测量 对数线性模型 χ^2 独立检验	一元统计量 描述统计 正态性检验	相关阵 主成分分析 因子分析 聚类分析
名义的/有序的 一个变量	χ^2 检验 Fisher 精确 检验	对数线性模型 LOGISTIC 回归	判别分析 LOGISTIC回归 一元统计量 (两组t)	判别分析 LOGISTIC回归
多于一个	对数线性模型	对数线性模型	判别分析	判别分析
间隔的/比率的 一个变量	t 检验 方差分析 生存分析	方差分析 多分类分析 生存分析	线性回归 相关 生存分析	多元回归 生存分析
多于一个	方差分析 主成分ANOVA Hotelling T^2 轮廓分析	方差分析 主成分ANOVA	典型相关	典型相关 通径分析 结构分析 (LIRESL,EQS)

多元分析中因变量(dependent)与自变量(explanatory)是由变量取值变化是否依其它变量而定,并没有尺度的区分。如临床上用新生儿月龄推算体重,月龄是自变量,体重是因变量。回归分析中有因变量,聚类分析无因变量,等等。A.J. Hartley(1983)按照变量的地位分成两类。第一类中变量的地位相同,有主成分分析、因子分析及聚类分析;第二类中各变量的地位不同,可以区分因变量和自变量,等等。

与探索性分析相应,确证性(confirmatory)分析是预先指示某种假设或模型,利用数据

检查模型的拟合情况，因此有：剩余(residual)=数据值(data)-模型拟合值(model)，数据表示基于样本的观测，模型表示假设观测所包含的结构，剩余是两者的差。模型的好坏常用拟合优度检验(goodness of fit test)，它是检验原始数据的样本分布是否与已知分布相符合的方法。拟合优度检验常用 χ^2 检验和Kolmogorov-Smirnov 检验。

总之，从实验设计到资料的收集、整理和分析，是有机的统一体，软件包也遵循了这个框架。常用内容有：

- 描述统计(descriptive statistics)
- 概率(Probability)
- 概率分布(probability distribution)
- 估计(estimation)
- 显著性检验(significance tests)
- 回归(regression)
- 方差分析(ANOVA)
- 线性模型(General Linear Models)
- 广义线性模型(Generalized Linear Models)
- 抽样调查(sample survey)
- 实验设计(design of experiments)
- 临床实验(clinical trials)
- 多元分析(multivariate analysis)
- 时间序列分析(time series analysis)
- 质量控制及可靠性(quality control and reliability)

从实际的分析来看，大多数时间常常用于数据的整理上。现场收集的数据必须进行逻辑检查，不同项目间交叉有许多缺失值，在处理上有其特色。

§1.2 统计分析软件包的种类

§1.2.1 计算机软件和软件包

计算机软件是计算机赖以完成其功能的一系列指令、程序等的统称。软件包不是子程序的集合，而是建立在基本功能之上的一个体系结构。统计软件包的应用开始于六十年代，主要有SAS、SPSS及BMDP，以批处理方式运行于大型机上，它们只有简单的数据类型。伴随着计算机软硬件等一系列新技术的发展，八十年代初出现了S语言，增加了图形、交互式处理和菜单驱动接口。目前，统计软件包是群星夺目，它们综合了计算机和统计学发展的最新成果。G.E.P. Box (1969) 生动描述了这种情形：现在的事情就这么简单，尽管可以对自己在做什么一无所知，但只肖借助于计算机，你仍能用令人难以置信的速度得到不论是正确的还是错误的所有的东西。

作为统计分析与计算机的有机结合体，统计软件包反映了计算机在统计中的作用：数据存贮及操作、综合数据特征、探索性数据分析、推断应用中的计算、模拟研究[6]。大量的数据处理和统计分析离开统计软件包是不可想象的，统计分析软件包的应用，使深奥的分析技术走出了统计的“象牙塔”，使人们从大量繁琐的手工计算、重复编程中解脱出来。它既能帮助统计工作者和各专业研究者把统计理论应用于实践，引导理论工作者走向应用领

域, 同时又在很大程度上帮助实际工作者从现实的直观背景中认识统计理论, 最终提高应用水平和理论水平, 可见软件包成了沟通理论和实际工作者、统计工作者和计算机人员的桥梁, 也遵循了“理论——实践——理论”循环往复的认识论规律。

§1.2.2 统计分析软件包的种类和特点

目前统计软件包的种类很多, 我们大致把它们归为通用和专用两类, 本书主要涉及以下几种:

SAS, 是一个通用的软件包, 它的数据处理和分析能力很强, 同时也是一个数据管理与撰写报告的工具。它拥有强大的编程能力和丰富的统计方法, 而且能够保存大部分中间结果, 尤其适于比较复杂的数据处理和建模。SAS系统庞大, 对非统计专业人员来说掌握比较困难。

SPSS, 是一个典型的统计分析软件包, 最初面向社会科学开发但其应用远不限于社会科学, 其特点是操作简便, 有良好的选择式菜单和运行提示。BMDP 是古老而著名的软件包, 它采用模块化调用, 各种统计分析功能由独立的模块实现, 也采用了方便调用的菜单操作方式, 但模块的独立调用也造成代码上的重复。SYSTAT 也是一个装卸方便、典型的模块化软件, 其图形功能很强。SYSTAT 的LOGIT和生存分析要用附加的LOGIT、SURVIVAL 模块。

Stata, 是一个数据处理、图形显示和统计分析高度集成化的软件包, 它同时也可以使用附加模块, 其源程序对用户透明, 从而使它生机勃勃。它提供了最好的软件运行使用说明。

S-Plus, 引进了目标编程, 它是一个功能强大的函数语言, 提供了丰富的数据分析和图形显示功能。

Minitab, 是一个特别适于教学使用的软件包, 包括了常规的统计方法, 其宏定义允许用户进行功能拓展。

GLIM, 是专为广义线性模型设计的软件包, 广义线性模型具有广泛的理论与实际意义。Genstat 功能与用法上与GLIM类似, 其开发宗旨在于通用, 故其功能扩大到处理一般统计问题。这两种软件对于进行新模型方法的研究很有用。

LISREL, 是专用于结构方程模型分析的软件, 该软件广泛应用于心理学、社会学、经济学等相关学科资料的分析。是确证型分析的典范。

MicroTSP, 是一个时序和经济分析软件包。

Epi Info, 特别适于现场资料的收集和流行病学分析, 它能与大多数统计软件包进行数据交换, 英文版的Epi Info 软件已由有关单位汉化成功。

综观这些统计软件包, 有如下功能特点: 1.丰富的函数和强大的编程能力, 通常多数操作可在软件包内完成; 2.系统的统计方法, 随着原有软件版本的更新, 现代统计方法不断被吸收。相互借鉴, 纳入一些专用的程序, 如SAS/STAT的CALIS 和SPSS 的LISREL; 3.数据处理功能强, 如适应和生成各种复杂的数据格式, 生成新的变量, 缺失值处理等; 4.丰富的图形功能, 标准图形的格式交换。5.用户界面友好, 可以交互式或批处理方式运行。如PC Windows版SPSS 和SAS 不仅考虑方便操作, 同时也接受命令输入, 用户可择其所好。另外, 它们有教学程序、样本程序、多窗口和在线帮助、运行提示等; 5. 程序可移植性, 有相应微机、小型机、巨型机上应用的软件。

计算机系统上的改进促进了软件包的更新换代, 如近年来PC Windows及软件包开发语言可移植性改善了软件包的可移植性。它们将随着统计分析技术和计算机技术而继续发展, 但在数据的智能化处理方面仍待突破, 如专家系统的开发。现有软件包强大的功能没有影

响新软件的出现,正说明一个软件包不能成为“万应灵药”,而是“兼收并蓄”,“吐故纳新”,不断发展。

统计软件包是方便研究与应用的手段,它也在理论的开拓与应用深度和广度方面,给统计工作者提出了更高的要求。统计软件包的应用离不开现实的背景,不断的探索,同时也要求具有一定的理论水平。正如达芬奇(Leonard da Vinci, 1452-1519)所说的那样:没有理论的实践,就象一个水手没有舵手和指南针,只能随波逐流。

§1.3 统计分析软件包的应用

软件包使用有几个前提条件[7]:首先要有适用的程序或软件包以及人力;其次应有把统计问题化为计算问题的能力,以保证正确使用这些程序或软件包;再次,能够正确地解释结果;最后还应有分析的资源,包括计算机、充足的经费和时间等。不考虑分析的研究是屡见不鲜的,这往往造成资源的浪费。

软件包使用的便利和效率之间常常需要一个折衷。菜单式直观简明,但行命令式纠错方便、运行速度快。灵活性包括数据的操作、宏调用、与其它软件的接口等。用户支持包括软件包提供的用户指南、示范程序的提供情况、教学与培训情况等。专用程序的好处是能够在软件包尚未包括所用的统计方法时采用,其缺点是书写费时,程序功能需要增强时,程序员往往不在场。

统计软件包有效的使用,必须借助扎实的统计理论基础和实际应用经验。我们推荐“以小见大”,即开始阶段用小的问题进行试验,进而“举一反三”。对于比较复杂的问题,阅读说明书、专著和向有经验的用户学习是很简捷的。一些要点如:

- 交互式与批处理方式的进入与退出、文件的操作方法。如程序文件的存贮方法:这些软件包的程序文件有习惯上的扩展名如SAS,INC,INP,DO,CMD,GLM。提交执行则是SUBMIT、DO以及INC。填充执行如SAS和Stata的MENU。又如系统外壳命令的使用,SAS用命令行的x命令、SPSS/PC+和SYSTAT用DOS.、Stata用!命令。
- 教学程序与样本程序的使用。软件包推出时,往往随软件提供了运行实例,利用上述文件输入输出方法可方便使用。建议开始用软件包统计分析时,用一些熟悉的例子进行一些演算和对比,明确其有关术语和体例,更主要是能从中获得结果的解释方法。
- 软件包手册、专著和专家请教,往往更可以事半功倍。
- 许多材料可以取自Internet,本书附录给出了一些地址可供参考。

选用软件包之前,应明确实际的需要和软件包分析的特点,以回归分析为例,多数软件包能够进行模型拟合和回归诊断,两阶段最小二乘、伪变量回归、有偏估计如岭回归等。软件包之间可以相互取长补短,如微机上的SAS 6.04没有COX生存分析过程,而Stata 2.0就已经提供了该项功能。在回归分析中对记录进行加权时,SPSS相对于SAS和Stata有其独到之处。

问题的程序化能力对于软件包功能的有效发挥是至关重要的,如掌握数据库管理系统技术和高级语言乃至有关计算方法。软件包的功能可以受限,但使用数据库和高级语处理问题是没有限度的。著名的Fortran语言专用程序库有如IMSL、NAG。使用Fortran时,要花很多时间搞清楚数据类型和格式,编程时的数据定义也要占据大量时间,调试也很费时。这些经验恰恰也是软件包程序化和调试的必要准备。SAS的编程采用模块调用,组合DATA

和PROC，思路非常清楚，调试也方便，出现错误时，系统以红色在记录窗口给出提示。系统能方便地与其它数据管理系统数据交换，如VAX/VMS下与SPSS、BMDP的交换。系统既有解释语言纠错的特点又有编译语言快速执行的优点。但如果用户已能熟练使用Fortran语言，则对于软件包的使用大有好处，建议从事数据处理和分析的人能够了解一些Fortran语言的有关知识。如Fortran对于记录的引用还是较SAS方便的，SYSTAT还提供了Fortran格式数据调用的子程序。

一开始接触软件包说明书，容易“只见树木，不见森林”，《指南》编写介绍其概貌，我们力求简明扼要，抓住共性。不同于传统的方式[8,9,10]，我们采用分别介绍的方式，各部分自成体系。这与新近出版的[12]体例相同但更充实。限于篇幅，我们仅对SAS和SPSS详细介绍。

