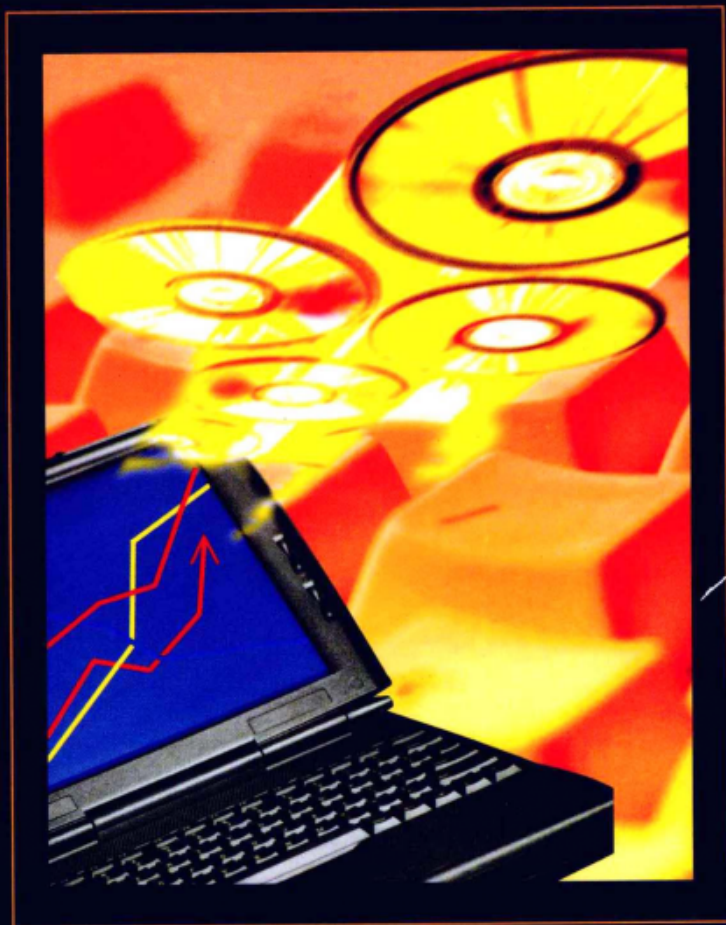


SAS软件与应用统计

王吉利 张尧庭 主编



中国统计出版社
China Statistics Press

责任编辑：吕军
封面设计：北京华运图文公司

ISBN 7-5037-3359-4



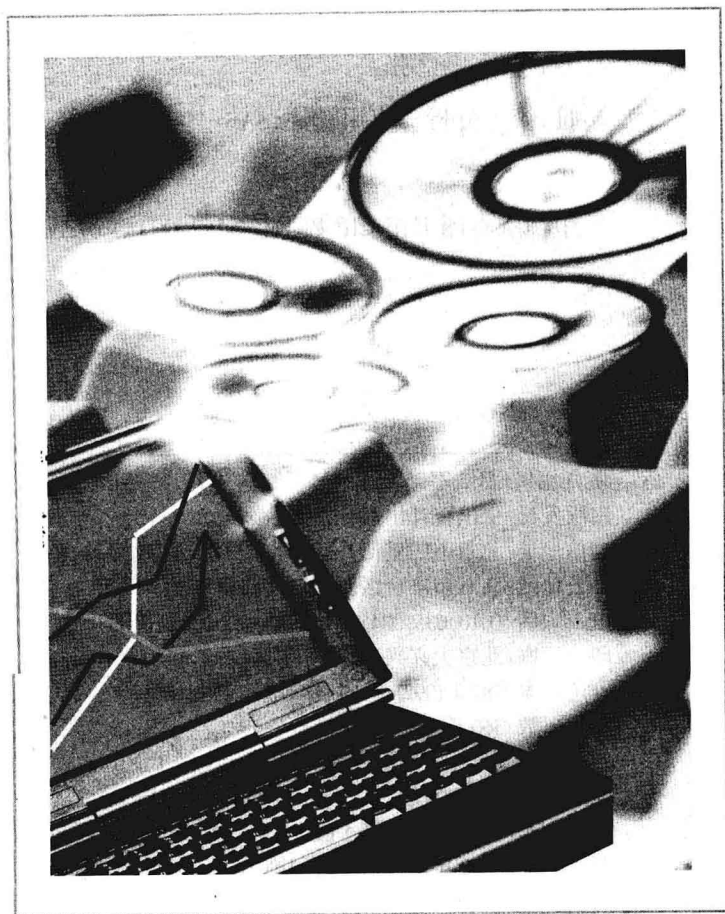
9 787503 733598 >

ISBN 7-5037-3359-4/C · 1830

定价：36.00元

SAS软件与应用统计

王吉利 张尧庭 主编



 中国统计出版社
China Statistics Press

(京)新登字 041 号

图书在版编目(CIP)数据

SAS 软件与应用统计/王吉利,张尧庭主编

—北京:中国统计出版社,2000.9

ISBN 7-5037-3359-4

I.S...

II.①王...②张...

III.统计分析—应用软件,SAS

IV.C819

中国版本图书馆 CIP 数据核字(2000)第 67282 号

责任编辑/吕 军

出版发行/中国统计出版社

通信地址/北京市三里河月坛南街 75 号 邮政编码/100826

办公地址/北京市丰台区西三环南路甲 6 号

电 话/(010)6359084,63266600-22500(发行部)

印 刷/北京大兴兴达印刷厂

开 本/787×1092mm 1/18

字 数/360 千字

印 张/16

印 数/1-4000 册

版 别/2000 年 9 月第 1 版

版 次/2000 年 9 月第 1 次印刷

书 号/ISBN 7-5037-3359-4/C·1830

定 价/36.00 元

中国统计版图书,版权所有,侵权必究。

中国统计版图书,如有印装错误,本社发行部负责调换。

序 言

SAS(Statistical Analysis System),在它面市的初期,就如它英语名字所说的是个统计分析软件.自它推出来以后,就是统计分析领域应用较为普及的一种软件,在世界各国统计部门中应用尤为广泛.国家统计局选定 SAS 作为一种在系统内推广应用的主体工作软件.

今天,欣喜地看到 SAS 经过二十多年的发展,功能更强大.以往成名的数据库管理系统一般都有与 SAS 软件的接口,而今 SAS 本身就具备了数据库的多重功能,并发展了数据仓库技术、数据挖掘技术和决策支持技术.

我们高兴地看到, SAS 公司最近积极发展实质性业务伙伴并指定了上海宏能计算机软件技术有限公司(简称“宏软”)作为其在统计教育领域及统计局相关 SAS 业务的独家代理,宏软公司可以比以前 SAS 中国公司投入更大的精力以更系统长远的方式支持和促进统计分析软件的教育与应用.应国家统计局的提议为方便普及 SAS 软件的推广应用,在原有《SAS 应用统计教程》教程的基础上,该书的原主要执笔者汪嘉冈教授,现宏软公司的高级顾问,在统计局教育中心与一些的应用统计界的同仁的支持与帮助下,更正了此书的一些错误及一些文字叙述后,将这本《SAS 软件与应用统计》由中国统计出版社正式出版.该书从实例入手,结合统计分析中常用的几种统计方法,步步深入,既阐述了 SAS 软件的基本应用,又介绍了这些方法的数学原理,将计算机应用与数理统计有机地结合起来,是一本比较实用的好书.

期望通过本书的发行,能促进我国统计部门统计分析水平的提高.同时希望能对凡是用到统计分析技术的领域都能有所裨益.

沈青华 2000 年 7 月

地介绍统计概念和原理的基础上,详细介绍了如何使用 SAS 软件实现常用的统计分析功能,并对软件提供的结果进行了必要的解说.

从目前广大用户的需要来看,首先需要的是如何使用软件实现最基本的统计功能方面的教材.所以尽管 SAS 软件的统计功能十分全面和完备,涉及的统计方法也非常广泛,我们还是以最基本和常用的统计概念、方法为线索,介绍 SAS 软件的使用和输出结果的含义.由于篇幅的限制,我们只能在介绍软件实现某种统计功能前,扼要地说明相应的统计原理、概念和一些统计指标的计算公式及含义.需要深入探讨统计方法原理的读者,可查阅专门的数理统计教科书.

本书在介绍软件的使用方面,兼顾了菜单系统和编程两个方面.这样可适应不同用户的需要.对每一种统计功能,在概要地介绍统计概念和原理的基础上,分别介绍用 SAS/INSIGHT,分析员应用和直接编程实现这一功能的做法.用户可根据自己的需要和基础,学会上述一种或多种手段.若从掌握软件的基本用法获得数据分析的结果这一要求来看,学会任何一种用法已是足够的了.

用户在使用本书时,若需要了解 PC 机上 WINDOWS 操作系统下 SAS 使用和编程最基本的事实,可参见参考文献的 [1].希望对编程的细节有进一步的了解,可使用联机帮助或参见 [3], [6] 和 [7].在软件实现各种统计功能方面,本书作为入门的教材,只能介绍最基本的用法和选项.若希望对这些方法深入的了解,除了使用联机帮助外,可参见 [4], [8], [9], [10] 和 [11].

本书在每章都附有习题,书中的例子和习题都是为了帮助读者在学习一般原理和做法的同时,进行实际操作.边做边学是最有效的学习方法.

最后要特别感谢 SAS 公司对宏软公司这项工作的支持,特别是北京办事处的栾世武博士仔细地校阅了该书的试用版的全文.感谢国家统计局沈清华总工程师为本书写的序言,他的热情支持使我们有信心从事这个工作.由于编写者水平、能力的限制以及经验的不足,本书必然还有许多不足和缺点,我们恳切地希望用户对我们提出批评和建议,共同为促进 SAS 软件的使用,提高我国数据分析应用的水平而努力.

上海宏能计算机软件技术有限公司
国家统计局统计教育中心

2000 年 7 月

目 录

第一章 引 论	1
§ 1.1 SAS 与统计分析	1
§ 1.2 SAS/INSIGHT	2
§ 1.3 分析员应用	11
§ 1.4 统计分析基本概念	21
习 题	30
第二章 描述性统计量计算	32
§ 2.1 频数统计与柱状图	32
§ 2.1.1 概述	32
§ 2.1.2 用 INSIGHT 作频数统计和直方图	35
§ 2.1.3 用分析员应用作频数统计和直方图	39
§ 2.1.4 用编程作频数统计	42
§ 2.2 常用描述性统计量计算	46
§ 2.2.1 概述	46
§ 2.2.2 用 Insight 计算常用描述性统计量	49
§ 2.2.3 用分析员应用计算常用描述性统计量	50
§ 2.2.4 用编程计算常用描述性统计量	53
§ 2.3 直方图、盒形图和分布拟合	63
§ 2.3.1 用 INSIGHT 作分布图形	66
§ 2.3.2 用分析员应用作分布图形	73
§ 2.3.3 用编程作分布图形	75
习 题	80
第三章 置信估计与假设检验	83
§ 3.1 置信区间	83
§ 3.1.1 概述	83
§ 3.1.2 用 Insight 作均值的置信区间	86
§ 3.1.3 用分析员应用作置信区间	87
§ 3.1.4 用编程作置信区间	89
§ 3.2 总体参数检验	91
§ 3.2.1 概述	91
§ 3.2.2 用 Insight 作均值检验	98

§ 3.2.3	用分析员应用作参数检验.....	99
§ 3.2.4	用编程作参数检验.....	102
§ 3.3	总体参数比较.....	104
§ 3.3.1	概述.....	104
§ 3.3.2	用分析员应用作参数比较.....	106
§ 3.3.3	用编程作分布检验.....	111
§ 3.4	分布置信带和分布拟合检验.....	112
§ 3.4.1	概述.....	112
§ 3.4.2	用 Insight 作分布检验和置信带.....	115
§ 3.4.3	用分析员应用作分布检验.....	117
§ 3.4.4	用编程作分布检验.....	119
习 题.....		122
第四章	方差分析	124
§ 4.1	单因素方差分析.....	124
§ 4.1.1	概述.....	124
§ 4.1.2	用 Insight 作单因素方差分析.....	128
§ 4.1.3	用分析员应用作单因素方差分析.....	132
§ 4.1.4	用编程作单因素方差分析.....	135
§ 4.2	多因素方差分析.....	140
§ 4.2.1	概述.....	140
§ 4.2.2	用 Insight 作多因素方差分析.....	144
§ 4.2.3	用分析员应用作多因素方差分析.....	148
§ 4.2.4	用编程作多因素方差分析.....	152
§ 4.3	均值估计与比较.....	156
§ 4.3.1	概述.....	156
§ 4.3.2	用分析员应用进行均值比较.....	157
§ 4.3.3	用编程进行均值比较.....	161
习 题.....		167
第五章	回归分析	169
§ 5.1	相关系数计算.....	169
§ 5.1.1	概述.....	169
§ 5.1.2	用 Insight 制作散点图和计算相关系数.....	171
§ 5.1.3	用分析员应用计算相关系数.....	178
§ 5.1.4	用编程计算相关系数.....	180

§ 5.2	简单线性回归.....	182
§ 5.2.1	概述.....	182
§ 5.2.2	用 Insight 拟合简单线性回归.....	186
§ 5.2.3	用分析员应用拟合简单线性回归.....	191
§ 5.2.4	用编程拟合简单线性回归.....	193
§ 5.3	多元线性回归.....	202
§ 5.3.1	概述.....	202
§ 5.3.2	用 Insight 拟合多元线性回归.....	207
§ 5.3.3	用分析员应用拟合多元线性回归.....	211
§ 5.3.4	用编程拟合多元线性回归.....	213
§ 5.4	回归诊断.....	221
§ 5.4.1	概述.....	221
§ 5.4.2	用 Insight 作回归诊断.....	225
§ 5.4.3	用分析员应用作回归诊断.....	231
§ 5.4.4	用编程作回归诊断.....	235
习 题.....		239
第六章	属性数据分析	224
§ 6.1	属性数据与列联表.....	244
§ 6.1.1	概述.....	244
§ 6.1.2	用 Insight 作拼花图.....	246
§ 6.1.3	用分析员应用作双向频数统计.....	247
§ 6.1.4	用编程作双向频数统计.....	250
§ 6.2	关联性分析.....	252
§ 6.2.1	概述.....	252
§ 6.2.2	用分析员应用作关联性分析.....	257
§ 6.2.3	用编程作关联性分析.....	260
§ 6.3	Logistic 回归分析.....	265
§ 6.3.1	概述.....	265
§ 6.3.2	用分析员应用作 Logistic 回归分析.....	267
§ 6.3.3	用编程作 Logistic 回归分析.....	272
§ 6.3.4	用 Insight 拟合 Logistic 回归.....	274
习 题.....		275
参考文献		278

第一章 引 论

§1.1 SAS 与统计分析

SAS 系统是美国 SAS 软件研究所的产品,它是一个用于决策支持的大型集成信息系统. 由于该系统在六十年代末及七十年代以统计分析及线性模型为主并且在当时以此闻名于世,所以该系统早期的全称为“统计分析系统 (STATISTICAL ANALYSIS SYSTEM)”. 这个全称的英文缩写“SAS”便很自然地成为开发此系统的公司的名称“SAS 研究所 (SAS INSTITUTE INC.)”及其产品的总商标.

SAS 系统经过二十多年的发展,以其卓越的数据处理能力,为在线数据分析、数据仓库、数据挖掘和决策支持提供了全面的解决方案. SAS 系统的发展始终离不开它的强大的数据分析功能,而且随着 SAS 系统的发展,其分析功能也与它在信息技术上的发展相辅相成,发展得更加深入、广泛和强大.

SAS 系统的分析功能是散布在几乎所有的模块之中. 较为集中的具有统计分析功能的是 SAS/STAT,SAS/QC,SAS/INSIGHT,SAS/ETS 等一些模块. BASE SAS 是运行 SAS 必须的模块,它也具有全面地管理 SAS 数据库的功能. 为了直接输入数据就需要 SAS/FSP. 若要访问异构数据库就需要相应的 SAS/ACCESS 模块. 为了将分析结果用各种高分辨率图形加以显示就需要 SAS/GRAPH. 如果分析的功能需要融合在一个计算机网络中运行,分析的结果要用超文本在企业网上发布,可用 SAS/INTRNET 和其它支持网络功能的模块.

SAS 系统的分析功能也在不断的发展之中. 它随时地把用户需要的和学术研究中得到的一些有效的实用分析方法加入到 SAS 的不同模块之中. 例如多变量分析中的偏最小二乘法便是一例.

在 SAS 系统分析功能的使用上,除了提供编程调用外, SAS 对一些常用的分析功能都提供了简便的菜单系统,使用户不用编程就可以方便地享用 SAS 的许多深入的分析功能. 目前已提供的有:

- SAS/INSIGHT: 是一个进行探索性数据分析的菜单操作系统. 它提供了所有的常用统计分析方法. 我们将在本书中着重地介绍它的使用.
- SAS/STAT—Analyst Application(分析员应用): 是一个为分析员设计的菜单系统. 它也提供了所有的常用统计分析方法. 在使用这一菜单系统完成每一个分析的同时还提供相应的 SAS 程序. 我们也将在这本书中介绍它的使用.
- SAS/STAT—Marketing Research: 是专为市场调查数据进行较深入的分析而开发的一个菜单系统. 包括对应分析 (Correspondence Analysis), 结合分析 (Conjoint Analysis), 离散选择分析 (Discrete Choice Analysis), 多维尺

度分析 (Multidimensional Scaling Analysis) 和多维偏好分析 (Multidimensional Preference Analysis) 等.

- SAS/QC—SQC: 是为企业使用各种质量控制方法而提供的菜单系统, 在使用这一菜单系统完成每一个分析的同时还提供了相应的 SAS 程序.
- SAS/QC—(X)ADX 是为安排各种试验和获得数据后进行分析的菜单系统.
- SAS/ETS—FORECAST 是为运用时间序列分析中各种建模和预测功能而开发的菜单系统.

在 SAS 的 OR 模块中还有 QSIM 和 XPROJMAN 等菜单系统也都有很直观的分析和安排计划的功能.

由于本书的目的是介绍运用 SAS 软件进行常用的统计分析, 所以在简要叙述各种统计分析方法后, 就分别介绍如何用 SAS/INSIGHT, 分析员应用 (Analyst Application) 和直接编程来完成这些分析功能, 以及说明 SAS 输出结果的含义. 对常用的一些统计分析方法而言, SAS/INSIGHT, 分析员应用和直接编程可以达到同样的目的. 所以掌握任何一种方法就可达到目的. 一般来说, SAS/INSIGHT 最为直观, 便于步步深入. 分析员应用可提供自动形成的程序, 而且在属性数据分析和功效函数计算方面较 INSIGHT 强. 编程是功能最强的, 尤其是一些特殊或深入的分析功能只能用编程实现. 所以各人可根据自己的需要和基础, 掌握一种或多种手段.

§1.2 SAS/INSIGHT

软件 SAS/INSIGHT 是一个交互式的数据探索和分析的工具. 用这一软件可以

- 通过多窗口连动的图象和分析结果, 对数据进行探索
- 分析单变量分布
- 用相关和主成分研究多变量间的关系
- 用方差分析和回归分析说明、拟合变量间关系的模型.

区间型变量和名义型变量

在 SAS 数据集中, 变量的两种类型为: 字符型变量和数值型变量. 在 SAS/INSIGHT 中, 为了区分变量在分析中不同作用, 变量又按其测量水平分为:

- 区间型的 (interval)
- 名义型 (或称列名型) 的 (nominal)

区间型变量必须是数值型变量, 可以对其观测值进行各种四则运算, 计算各种统计量. INSIGHT 中有些分析功能限定必须对区间型变量进行 (例如计算相关). 名义型变量可以是数值型的, 也可以是字符型的, 在 INSIGHT 中常起分

类的作用(例如在方差分析中). 字符型变量只能是名义型的, 对取值不多而用以区分不同类别的数值变量也可看作名义型的, 例如变量取不同数值表示性别、职业和品种, 就可将它定为名义型的.

在缺省情况下, INSIGHT 对所有数值变量冠以区间型测量水平, 对所有字符型变量冠以名义型测量水平. 在有需要时, 也可将数值型变量冠以名义型测量水平.

调用 SAS/INSIGHT

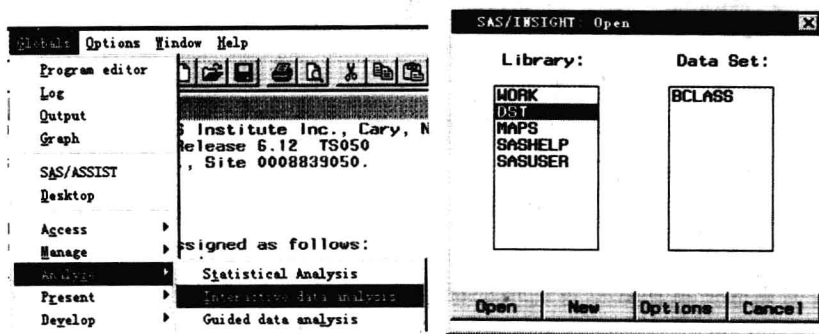
在宏软SAS学习版中调用 SAS/INSIGHT 只需在进入系统后点击“交互数据分析”图标即可.

在 SAS 显示管理系统中, 可用以下任一种方法进入 Insight:

- 可用菜单栏中的下拉菜单或弹出菜单;
- 在命令行中键入命令: **Insight**;
- 提交 PROC INSIGHT 语句.

用下拉菜单调用 INSIGHT 时选:

Globals ⇒ **Analyze** ► ⇒ **Interactive data analysis.**



打开的第一个窗提供可用的 SAS 数据库名表和数据集. 当选定了库名后, 该库所包含的数据集随之在右侧数据集 (Data Set) 框中列出.

打开数据集

以下我们采用 SAS 数据集 DST.BCLASS 作为例子. 这个数据集包含一个学校 40 个中学生的基本数据. 它包含的 5 个变量为:

- NAME 学生的名字
- SEX 学生的性别
- AGE 学生的年龄
- HEIGHT 以厘米为单位的身高数值
- WEIGHT 以公斤为单位的体重数值

对此例子，选

1. **DST** 为库名 (library)
2. **BCLASS** 为数据集 (data set)
3. **Open**.

数据窗打开，显示数据集 DST.BCLASS 的数据值。

ID	NAME	AGE	SEX	HEIGHT	WEIGHT
1	KATIE	12	女	145	43.1
2	LOUISE	12	女	149	55.8
3	JANE	12	女	135	33.6
4	JACLYN	12	女	162	65.8
5	LILLIE	12	女	127	29.1
6	TIM	12	男	147	38.1
7	JAMES	12	男	149	58.1
8	ROBERT	12	男	125	35.9
9	BARBARA	13	女	147	50.8
10	ALICE	13	女	149	48.6
11	SUSAN	13	女	137	30.4
12	JOHN	13	男	159	44.5
13	JOE	13	男	154	47.7
14	MICHAEL	13	男	142	43.1
15	DAVID	13	男	145	35.9
16	JUDY	14	女	149	36.8
17	ELIZABET	14	女	152	41.3
18	LESLIE	14	女	159	64.5
19	CAROL	14	女	154	38.1
20	PATTY	14	女	152	38.6

察看数据窗，它提供重要的信息：

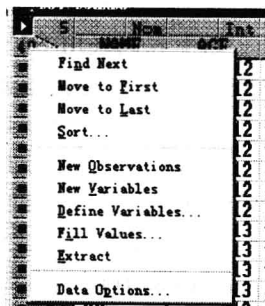
- 窗的左上角给出了变量数和观测值个数。数据集 DST.BCLASS 有 5 个变量和 40 个观测值。
- 每个观测值置以一个观测序号。观测序号前的符号表示该观测在作图时使用的符号和颜色。
- 数据集的变量在数据窗的顶部标明。每个变量都标以区间型 (Int) 或名义型的 (Nom)。

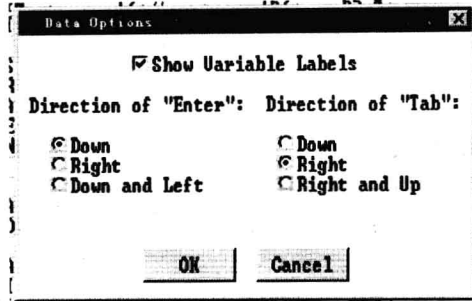
数据窗可像其它窗一样拖动边框改变其大小和使用滚动条滚动显示的数据。

若变量有标题名 (label)，可将标题名加到列的头

上：

1. 点击数据表左上角尖头，弹出菜单。
2. 从弹出菜单选 **Data Options...**
3. 在打开的窗点击 **Show Variable Labels** 前的方框，使之打勾。
4. 选 **OK**。



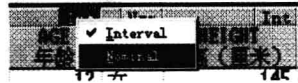


标题名即显示在变量名的下面。在这个例子中标题名就是中文的变量名。

设置测量水平

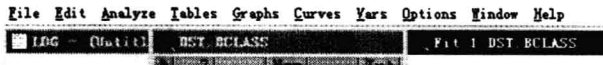
因为变量 AGE 是数值型的，INSIGHT 预定它也是区间型的数据。事实上以后常用它作分类变量的。若要改变 AGE 的测量水平：

1. 在 AGE 列的顶部选 Int。
2. 选 Nominal。



INSIGHT 功能概述

在 INSIGHT 的环境里，下拉（弹出）菜单包括 File, Edit, Analyze, Table, Graphs, Curves, Vars 等一些提供 INSIGHT 特殊功能的栏目。其中：



- File 栏目下提供读入数据集、存储数据集、分析结果的表格和图形的管理。
- Edit 栏目下提供对打开数据集的变量、观测进行增删、设定数据显示格式和窗口管理的功能。
- Analyze 栏目下提供了作图和分析的功能、其分析功能已归为分布拟合、线性模型（包括回归分析、方差分析等）、多变量分析三类。
- Tables, Graphs, Curves, Vars 栏目下随不同的分析项目提供不同的功能。Table 提供显示分析结果的各种表的功能，Graphs 提供生成与分析有关的各种图形的功能，Curves 提供与分析有关的各种曲线的功能，Vars 提供由分析结果形成的新的变量的功能。

在数据窗中移动列

对上面的例子若感兴趣每个学生的性别，希望把性别放在姓名之前的第一列。为此选：

1. 数据表左上角尖头，弹出菜单。

2. **Move to First**, 移至第一列 (Move to First) 对话框打开.

3. 选 **SEX**.

4. 选 **OK**.

SEX 就在数据窗的第一列.

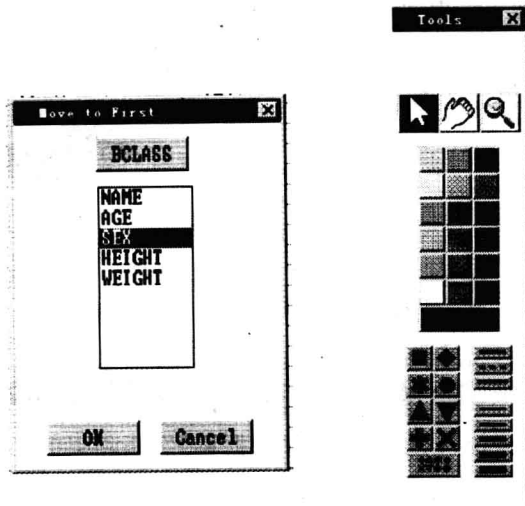
将 NAME 移至 HEIGHT 和 WEIGHT 之间:

1. 选 **Edit** ⇒ **Windows** ► ⇒ **Tools**, 打开工具窗 (Tools window).

2. 选其中的工具“手”.

3. 选变量 name 列的顶部. 按下鼠标器的选择 (左) 键, 拖动列到 HEIGHT 和 WEIGHT 之间的新位置.

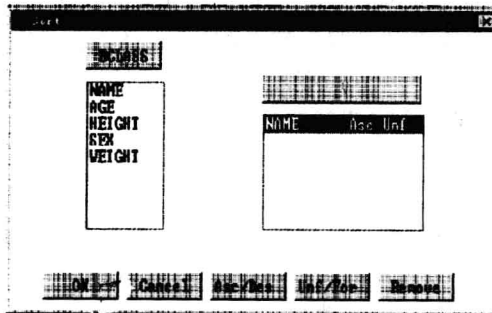
若要移动其它变量的位置也可做同样的操作.



对数据集排序

若为了便于查询, 希望数据集能按姓名的字母顺序进行排序. 选

1. 数据表左上角的尖头, 弹出菜单.



2. **Sort...**, 排序窗 (Sort window) 打开.

3. **NAME** \Rightarrow **Y**.

缺省情况下, 将按这些变量值的升序 (Asc) 排序. 若感兴趣的是按年龄排序, 年长的在前. 同年龄的学生, 按身高排序, 矮的在前. 为此可按 AGE 和 HEIGHT 排序, 而变量 AGE 按降序 (Des) 排序. 选:

1. **AGE** 至 Y 框中.

2. **HEIGHT** 至 Y 框中.

3. 在 Y 框中选中 AGE 使之变亮, 选 **Asc/Des**, Y 框中 AGE 后改为 Dec Unf(按变量值降序排序).

4. **OK**.

通过滚动数据集可看到, 年龄最大的为 17 岁, 最小的为 12 岁. 17 岁中的 Linda(女) 为最矮, Lawrence(男) 最高. 12 岁中 Robert(男) 最矮, Jaclyn(女) 最高.

创建新的变量

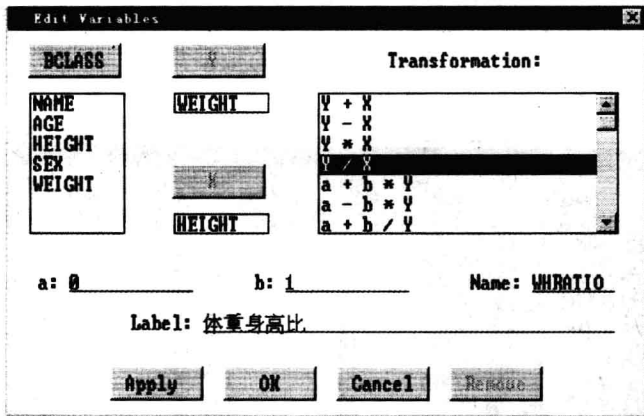
若感兴趣每个学生体重与身高之比可以定义新的一列, 为创建这一新的变量:

1. 选 **Edit** \Rightarrow **Variables** \Rightarrow **Other...**, 编辑变量窗 (Edit Variables window) 打开.

2. 从候选变换表中选 **Y/X**.

3. 选 **WEIGHT** \Rightarrow **Y**.

4. 选 **HEIGHT** \Rightarrow **X**.



5. 在 Name 字段, 键入 WHRATIO 作为新变量的名. 在 Label 字段, 键入 体重身高比.

6. 选 **OK**.

即关闭编辑变量窗，变量 WHRATIO 已加入数据表。这时也可按体重身高比的值的降序进行排序。选：

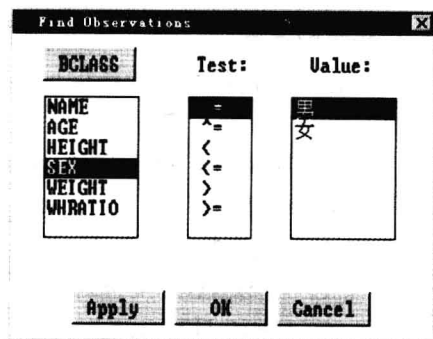
1. 数据表左上角尖头，弹出菜单。
2. **Sort...** 排序窗 (Sort window) 打开。
3. **WHRATIO** ⇒ **Y**。
4. 在 Y 框中使 WHRATIO 变亮，选 **Asc/Des**。
5. **OK**。

滚动数据集可看到，Lawrence 的体重身高比最大 (0.4541)。Susan 的体重身高比最小 (0.2219)。

建数据集的子集

若只要对数据中的所有男生进行分析，可以将数据中的所有男生的记录抽取出来，形成新的数据集。为此

1. 选 **Edit** ⇒ **Observations** ⇒ **Find...**，找观测值窗 (Find Observations window) 打开。
2. 选 **SEX** 作为变量。
3. 在 Test 字段选 **=**。
4. 在 Value 字段选 **男**。
5. 选 **OK**。



数据集中的某些观测 (行) 已被选中。被选中的观测其序号都会变亮。我们可以将这些选中的观测抽取出来建立一个新的数据集：

1. 点击数据表左上角尖头，弹出菜单。

Obs	NAME	AGE	HEIGHT	SEX	WEIGHT
姓名	年龄	身高 (厘米)	性别	体重 (公斤)	体重
28	JOHN	14	167	男	58.8
29	MARY	14	169	男	51.3
30	FREDRI	13	142	男	43.1
31	ELIZAB	13	162	男	48.1
32	LEWIS	15	162	男	47.7
33	TIM	15	125	男	35.9
34	PATTY	14	157	男	44.9
35	JANE	14	157	男	44.9
36	DAVID	13	159	男	44.5
37	CAROL	14	154	男	42.2
38	JUDY	14	157	男	41.8
39	LILLIE	12	147	男	38.1
40	SUSAN	13	145	男	35.9

2. 选 **Extract**。