

TURING

高等院校数学·统计学教材系列



统计软件教程

SAS系统与S语言

李东风 编著



人民邮电出版社
POSTS & TELECOM PRESS

统计软件教程

SAS系统与S语言

本书的内容曾以讲义形式在北京大学使用多年，主要讲述SAS系统和S语言。

SAS系统是一个集大型数据库管理、统计分析、报表图形、信息系统开发等多种强大功能为一体的大型软件系统，是国际上公认的优秀统计分析软件，但初学时有些困难。本书采取循序渐进的讲法，力图使读者更容易掌握SAS的要点。SPlus是一种以S语言为基础的面向对象的统计编程与统计分析系统，具有强大的探索性数据分析功能，实现了很多最新的统计方法。用S语言编写统计计算程序相对而言易学易用，可以极大地降低统计计算程序的开发时间。另外，还有一种开放的统计软件R，它和SPlus基本兼容。书中用S统称SPlus和R软件，介绍了SPlus和R都能使用的功能。

本书是统计软件的入门教程，适合作为高等院校统计学专业本科生、研究生的教材，也可作为从事统计相关工作的科研技术人员或对统计软件感兴趣的相关人员的自学教材。

本书相关信息请访问：图灵网站 <http://www.turingbook.com>

读者/作者热线：(010) 88593802

反馈/投稿/推荐信箱：contact@turingbook.com

上架建议

统计学/统计软件

ISBN 7-115-15220-9



9 787115 152206 >

ISBN7-115-15220-9/O1 · 4

定价：35.00 元

人民邮电出版社网址 www.ptpress.com.cn

TURING

高等院校数学·统计学教材系列

统计软件教程

SAS系统与S语言

李东风 编著



人民邮电出版社
POSTS & TELECOM PRESS

图书在版编目 (CIP) 数据

统计软件教程: SAS 系统与 S 语言 / 李东风编著.

—北京: 人民邮电出版社, 2006.11

(高等院校数学、统计学教材系列)

ISBN 7-115-15220-9

I. 统... II. 李... III. 统计分析—应用软件, SAS 高等学校—教材
IV. C812

中国版本图书馆 CIP 数据核字 (2006) 第 103013 号

内 容 提 要

本书是统计软件的入门教程, 分成两部分, 分别讲述 SAS 系统和 S 语言。主要包括: SAS 系统的基础知识、SAS 语言与数据管理、SAS 功能模块、基本统计分析、多元统计分析以及 S (SPlus 和 R 软件的统称) 语言。全书采取循序渐进的讲法, 不强调全面介绍软件的功能, 而是让读者掌握软件的基本用法。

本书适合作为高等院校统计学专业本科生、研究生的教材, 也可供从事统计相关工作的科研技术人员参考。

高等院校数学·统计学教材系列

统计软件教程: SAS 系统与 S 语言

-
- ◆ 编 著 李东风
责任编辑 王丽萍
 - ◆ 人民邮电出版社出版发行 北京市崇文区夕照寺街 14 号
邮编 100061 电子函件 315@ptpress.com.cn
网址 <http://www.ptpress.com.cn>
北京铭成印刷有限公司印刷
新华书店总店北京发行所经销
 - ◆ 开本: 700 × 1000 1/16
印张: 18
字数: 350 千字
印数: 1—4 000 册
 - 2006 年 11 月第 1 版
2006 年 11 月北京第 1 次印刷

ISBN 7-115-15220-9/O1 · 4

定价: 35.00 元

读者服务热线: (010)88593802 印装质量热线: (010)67129223

图灵公司致力于引进国外优秀数学统计学教材。出版的数学统计学系列丛书(含影印版和中文版)有的是国际各大名校竞相采用的顶级教材,久为国内学界瞩目;有的属于计算数学和金融数学等教学、科研和应用的热点领域,填补了国内教材的空缺;更有众多在国际学术界享有盛誉的大师名作,有些著作原版已经印行达40次以上,至今仍畅销不衰。丛书从选材、引进、翻译到出版,得到了国内数学统计学界老中青三代专家的鼎力支持。

目录、前言等出版物详细信息,请浏览www.turingbook.com



书 名: 矩阵分析 卷1 (英文版)
原书名: Matrix Analysis
作 者: Roger A. Horn, Charles R. Johnson
书 号: 7-115-13768-4
定 价: 59.00元
出版时间: 2005年9月



书 名: 矩阵分析 卷2 (英文版)
原书名: Matrix Analysis
作 者: Roger A. Horn, Charles R. Johnson
书 号: 7-115-14027-8
定 价: 69.00元
出版时间: 2005年10月



书 名: 应用随机过程: 概率模型导论 (英文版·第8版)
原书名: Introduction to Probability Models, Eighth Edition
作 者: Sheldon M. Ross
书 号: 7-115-14514-8
定 价: 88.00元
出版时间: 2006-03-22
教 辅: 习题答案



书 名: 金融数学教程 (英文版)
原书名: A Course in Financial Calculus
作 者: Alison Etheridge
书 号: 7-115-14090-1
定 价: 29.00元
出版时间: 2005年11月



书 名: 金融数学教程
原书名: A Course in Financial Calculus
作 者: Alison Etheridge
译 者: 张奇洲 等
书 号: 7-115-14892-9/O1·2
定 价: 29.00元
出版时间: 2006年8月



书 名: 金融数学: 衍生品定价引论 (英文版)
原书名: Financial Calculus
作 者: Martin Baxter, Andrew Rennie
书 号: 7-115-14089-8
定 价: 35.00元
出版时间: 2005年11月



书 名: 金融数学: 衍生品定价引论
原书名: Financial Calculus
作 者: Martin Baxter, Andrew Rennie
译 者: 叶中行 王桂兰 林建忠 (上海交通大学)
书 号: 7-115-14204-1
定 价: 29.00元
出版时间: 2006年1月



书 名: 时间序列分析: 预测与控制 (英文版·第3版)
原书名: Time Series Analysis
作 者: G. E. P. Box, G. M. Jenkins, G. C. Reinsel
书 号: 7-115-13772-2
定 价: 65.00元
出版时间: 2005年9月



书 名: 统计模拟 (英文版·第3版)
原书名: Simulation
作 者: Sheldon M. Ross
书 号: 7-115-14125-8
定 价: 39.00元
出版时间: 2005年11月
中文版2006年8月出版
译 者: 王兆军 (南开大学)



书 名: 集合论基础 (英文版)
原书名: Elements of Set Theory
作 者: Herbert B. Enderton
书 号: 7-115-14550-4
定 价: 39.00元
出版时间: 2006年3月



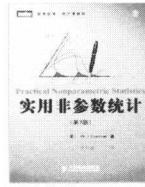
书 名: 偏微分方程数值解 (英文版·第2版)
原书名: Numerical Solution of PDEs
作 者: K. W. Morton, D. F. Mayers
书 号: 7-115-14075-8
定 价: 39.00元
出版时间: 2005年11月



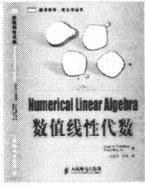
书 名: 偏微分方程数值解
原书名: Numerical Solution of PDEs
作 者: K. W. Morton, D. F. Mayers
译 者: 李治平 门大力等 (北京大学)
书 号: 7-115-14203-3
定 价: 29.00元
出版时间: 2006年1月



书 名: 流形上的微积分 (双语版)
原书名: Calculus on Manifolds
作 者: Michael Spivak
译 者: 齐民友 路见可 (武汉大学)
书 号: 7-115-14224-6
定 价: 29.00元
出版时间: 2006年2月



书 名: 实用非参数统计
原书名: Practical Nonparametric Statistics
作 者: W. J. Conover
译 者: 崔恒建 (北京师范大学)
书 号: 7-115-14616-0
定 价: 59.00元
出版时间: 2006年4月



书 名: 数值线性代数
原书名: Numerical Linear Algebra
作 者: Lloyd N. Trefethen, David Bau, III
译 者: 陆金甫 关治 译
书 号: 7-115-15168-7/O1·3
定 价: 39.00元
出版时间: 2006年9月



书 名: 概率论及其应用
原书名: An Introduction to Probability Theory and its Applications
作 者: William Feller
译 者: 胡迪鹤 (武汉大学)
书 号: 7-115-14729-9
定 价: 59.00元
出版时间: 2006年5月



书 名: 数理逻辑 (英文版·第2版)
原书名: A Mathematical Introduction to Logic
作 者: Herbert B. Enderton
书 号: 7-115-14145-2
定 价: 39.00元
出版时间: 2005年12月



书 名: 数据挖掘导论 (中文版)
原书名: Introduction to Data Mining
作 者: Pang-Ning Tan, Michael Steinbach, Vipin Kumar
译 者: 范明 范宏建 (郑州大学)
书 号: 7-115-14698-4
定 价: 49.00元
出版时间: 2006年5月



书 名: 数据挖掘导论 (英文版)
原书名: Introduction to Data Mining
作 者: Pang-Ning Tan, Michael Steinbach, Vipin Kumar
书 号: 7-115-14144-4
定 价: 59.00元
出版时间: 2005年12月

前 言

统计学是研究如何收集数据、分析数据并进行推断的学科。统计学的应用必然要涉及数据的收集、存储、整理以及各种统计方法的实际计算,这些都可以借助统计软件来完成。

从电子计算机出现至今,统计软件已经有了长足的发展。一方面,经典的统计方法都已在统计软件中实现;另一方面,帮助统计学家实现新的统计方法的软件也极大地推动了新的统计方法的研究与开发。国内统计软件的应用这些年来也有很大进步,先后引进了国外的 SAS、SPSS、SPlus、Gauss 等统计软件,并出版了一些相关的书籍。然而,市面上介绍统计软件的书多数是着重讲述软件本身的功能,没有考虑到初学者循序渐进的要求。此外,可供自学的教材性质的书也比较少。本书试图填补这一空白。以自学教材的形式讲述,不强调全面介绍软件功能,而是让读者通过阅读本书来掌握统计软件的基本用法,为读者进一步阅读更详细的资料以达到熟练使用统计软件完成统计应用的目的打下基础。

本书主要讲述 SAS 系统和 S 语言。SAS 系统是一个集大型数据库管理、统计分析、报表图形、信息系统开发等多种强大功能为一体的大型软件系统,是国际上公认的优秀统计分析软件,但初学时有些困难。本书力图采取循序渐进的讲法使读者更容易掌握 SAS 的要点。SPlus 是一种以 S 语言为基础的面向对象的统计编程与统计分析系统,具有强大的探索性数据分析功能,实现了很多最新的统计方法。用 S 语言编写统计计算程序相对而言易学易用,可以极大地降低统计计算程序的开发时间,方便了统计学家研究新的统计方法和进行统计计算。另外,还有一种开放的统计软件 R,它和 SPlus 基本兼容。在强调知识产权保护的今天,R 是一个不错的选择。

本书可用作大专院校统计软件课程的教材或用作自学教程。阅读本书需要读者有基本的数理统计知识。另外,读者最好已经学过计算机程序设计。书中讲述的是专用语言,不适合作为一般的程序设计入门教材。本书的重点是用统计软件进行数据管理、统计分析和统计计算编程,适用于统计软件的初学者。对于用统计软件开发商业化应用程序所需知识涉及较少,而希望进行统计应用开发的读者,可以在阅读本书后进一步学习有关知识。

本书来自作者长期在北京大学讲授统计软件课程的积累。在此作者感谢为此书写作提供了帮助和修改意见的老师和同学,特别要感谢高惠璇教授。她为了在国内推广 SAS 系统,做了许多开拓性的工作,也是她带领作者进入了 SAS 领域。书中引用了其他教材的一些例子和习题,不一一列举致谢,这里一并表示感谢。

本书用作 48 学时的教材,可以讲授第1~4章以及第6章的1~10节。书中用仿宋

体排印的部分技术性较强，可以不讲。如果用作60学时的教材，则可以讲授全部章节。

鉴于作者水平有限，书中可能还存在一些错误和不足之处，望广大读者批评指正。

编者

2006年3月于北京大学

目 录

第 1 章 SAS 初阶	1
1.1 初识 SAS	1
1.1.1 启动	1
1.1.2 SAS AWS	1
1.1.3 简单运行样例	3
1.2 SAS基本概念	5
1.2.1 SAS数据集	5
1.2.2 SAS数据库	6
1.2.3 SAS 管理器	7
1.3 SAS/INSIGHT	9
1.3.1 SAS/INSIGHT模块简介	9
1.3.2 数据窗口	10
1.3.3 一维数据探索	15
1.3.4 二维数据探索	19
1.3.5 三维数据探索	23
1.3.6 图形的调整	25
1.3.7 分布研究	27
习题	34
第 2 章 SAS语言与数据管理	37
2.1 SAS语言构成	37
2.1.1 SAS语句	37
2.1.2 SAS表达式	37
2.1.3 SAS程序规则	39
2.2 SAS用作一般高级语言	40
2.2.1 赋值语句	40
2.2.2 输出语句	41
2.2.3 分支结构	42
2.2.4 循环结构	44
2.2.5 数组	45
2.2.6 函数	48
2.2.7 SAS/IML矩阵功能简介	54
2.3 SAS语言的数据管理功能	55

2.3.1	SAS数据步的运行机制	55
2.3.2	用input语句输入数据	58
2.3.3	变量属性	61
2.3.4	读入外部数据	62
2.3.5	数据集的复制与修改	65
2.3.6	用set和output语句拆分数据集	66
2.3.7	数据集的纵向合并	67
2.3.8	数据集的横向合并	68
2.3.9	用update语句更新数据集	69
2.4	SAS宏介绍	70
2.4.1	实例	70
2.4.2	宏变量	71
2.4.3	宏	72
2.4.4	流程控制结构	73
2.4.5	宏与数据步的信息交换	73
2.4.6	宏语句和宏函数	74
2.5	用proc sql管理数据	75
	习题	79
第3章	SAS功能基础	81
3.1	SAS过程初步	81
3.1.1	SAS过程用法	81
3.1.2	SAS过程步常用语句	81
3.2	列表报告	84
3.2.1	基本用法	84
3.2.2	SAS中对输出结果的管理	85
3.2.3	使用中文列标题	87
3.2.4	标题及全程语句	88
3.2.5	用by语句分组处理	89
3.2.6	计算总计和小计	89
3.3	汇总表格	91
3.4	数据排序	95
3.5	数据集转置	95
3.6	描述统计	98
3.7	相关系数计算	101
3.8	用SAS/GRAPH绘图	102

3.8.1	散点图和曲线图	102
3.8.2	直方图和扇形图	104
3.8.3	三维曲面图和等高线图	106
3.8.4	图形的调整与输出	108
3.9	分析员模块介绍	109
3.9.1	数据管理	109
3.9.2	报表	113
3.9.3	描述性统计	115
3.9.4	画图	116
	习题	117
第4章	SAS的基本统计分析功能	119
4.1	几种假设检验	119
4.1.1	单总体t检验和p值	119
4.1.2	正态性检验	121
4.1.3	两独立样本的均值检验	121
4.1.4	成对总体均值检验	124
4.2	回归分析	126
4.2.1	用SAS/INSIGHT进行曲线拟合	126
4.2.2	用SAS/INSIGHT进行线性回归分析	128
4.2.3	用SAS/INSIGHT拟合广义线性模型	132
4.2.4	用reg过程进行回归分析	134
4.2.5	用Analyst进行回归分析	138
4.3	方差分析入门	138
4.3.1	用anova过程进行单因素方差分析	138
4.3.2	用npar1way进行非参数单因素方差分析	140
4.3.3	多重比较	141
4.3.4	多因素方差分析	145
4.3.5	用Analyst作方差分析	148
4.4	属性数据分析	149
4.4.1	拟合优度的卡方检验	150
4.4.2	列联表的输入与制表	151
4.4.3	列联表独立性检验	153
4.4.4	属性变量关联度计算	155
4.4.5	用Analyst作列联表分析	156
	习题	157

第 5 章 SAS多元统计分析	163
5.1 多变量分析	163
5.1.1 主分量分析	163
5.1.2 因子分析	172
5.2 判别分析	176
5.2.1 统计背景	176
5.2.2 <code>discrim</code> 过程的语句说明	177
5.2.3 例子	178
5.3 聚类分析	185
5.3.1 谱系聚类方法介绍	185
5.3.2 谱系聚类类数的确定	187
5.3.3 用 <code>cluster</code> 过程和 <code>tree</code> 过程进行谱系聚类	188
习题	193
第 6 章 S语言介绍	195
6.1 S快速入门	195
6.2 S向量	199
6.2.1 常量	199
6.2.2 向量与赋值	199
6.2.3 向量运算	200
6.2.4 产生有规律的数列	201
6.2.5 逻辑向量	202
6.2.6 字符型向量	203
6.2.7 复数向量	204
6.2.8 向量下标运算	204
6.3 矩阵与多维数组	206
6.3.1 矩阵	206
6.3.2 访问矩阵元素和子矩阵	208
6.3.3 矩阵运算	209
6.3.4 <code>apply</code> 函数	211
6.3.5 数组	211
6.3.6 数组的外积	213
6.4 因子	214
6.5 列表	216
6.5.1 列表定义	216
6.5.2 修改列表	217

6.5.3 几个返回列表的函数	218
6.6 数据框	219
6.6.1 数据框生成	220
6.6.2 数据框引用	220
6.6.3 <code>attach()</code> 函数	220
6.7 输出输入	221
6.7.1 输出	221
6.7.2 输入	223
6.8 程序控制结构	224
6.8.1 分支结构	224
6.8.2 循环结构	226
6.9 S程序设计	227
6.9.1 工作空间管理	228
6.9.2 函数定义	228
6.9.3 参数	229
6.9.4 作用域	230
6.9.5 程序调试	231
6.9.6 程序设计举例	232
6.10 图形	237
6.10.1 常用图形	237
6.10.2 高级图形函数	240
6.10.3 高级图形函数的常用选项	242
6.10.4 低级图形函数	243
6.10.5 交互图形函数	244
6.10.6 图形参数的使用	246
6.10.7 图形参数详解	247
6.10.8 图形设备	251
6.11 S初等统计	252
6.11.1 单变量数据分析	252
6.11.2 假设检验	255
6.12 S统计模型简介	256
6.12.1 统计模型的表示	256
6.12.2 线性回归模型	258
6.12.3 提取信息的通用函数	258
6.12.4 方差分析	259

6.13 统计分析实例	260
6.13.1 数据输入	261
6.13.2 探索性数据分析	261
6.13.3 组间比较	262
6.13.4 回归分析	263
6.14 用S作随机模拟计算	266
6.15 S常用函数参考	269
6.15.1 基本函数	269
6.15.2 数学	270
6.15.3 程序设计	271
6.15.4 统计计算	272
习题	274

第 1 章 SAS 初阶

1.1 初识 SAS

本节将介绍 SAS 应用工作空间，并通过一个简单的例子来了解 SAS 编程及程序运行情况。

1.1.1 启动

用如下方法可以进入 SAS 系统的窗口运行环境：在 Windows 环境中，从开始菜单的程序文件夹中找到 SAS 系统文件夹，双击其中的 SAS 运行程序启动 SAS 系统。或者生成 SAS.EXE 的快捷方式（把 SAS.EXE 用鼠标右键拖到桌面），双击 SAS.EXE 启动 SAS 系统。

1.1.2 SAS AWS

启动SAS系统后，出现如图1-1所示的SAS运行界面，称为“SAS应用工作空间（SAS Application Workspace）”，简称SAS AWS。这是SAS V8.1的界面。它像其他Windows应用程序一样，在一个主窗口内包含若干个子窗口，并有菜单栏、工具栏及状态栏等。

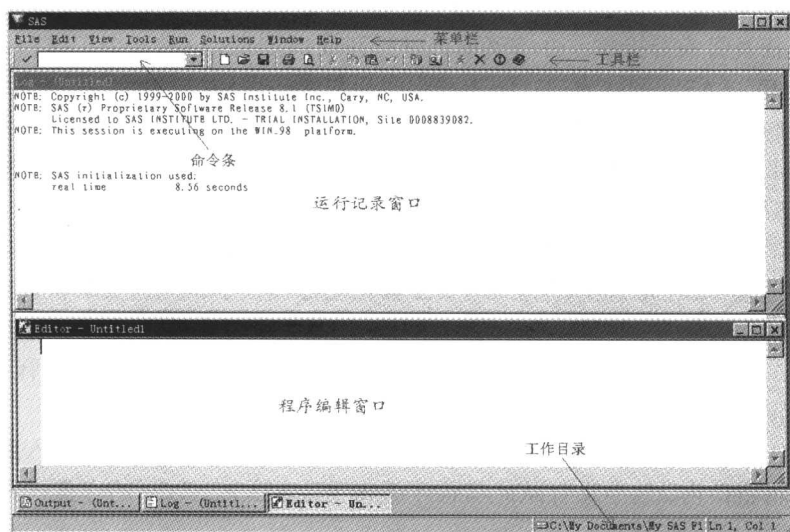


图 1-1 SAS AWS

SAS有三个最重要的子窗口：程序编辑窗口（Editor）、运行记录窗口（Log）和

输出窗口(Output)。

程序编辑窗口类似于Windows中的记事本程序,可以在其中编辑文本文件,主要是编辑SAS程序。SAS V8.1的程序编辑功能较以前版本有所增强。可以用不同颜色显示不同的SAS程序部分;也可以自动缩进排列程序文本,或者折叠一段程序。可以直接在窗口中键入程序,回车可插入新行,可以用光标键(上下左右箭头、Home或End)移动插入点光标(闪动的竖线)或用鼠标单击到某一处。按住Shift再按光标键就可以加亮显示一块文本,然后用复制、剪切、粘贴命令(Edit菜单中的Cut、Copy、Paste,或工具栏图标)来复制或移动加亮显示的文本。这些编辑操作可以参考Windows系统中记事本、Word等的用法。

运行记录窗口记录程序的运行情况:运行是成功还是出错;运行所用时间;如果出错,错在什么地方。运行记录窗口中以红色显示的是错误信息。

输出窗口显示SAS程序的文本型输出(图形输出单独有一个Graphics窗口)。输出分页显示。

要把光标移动到某一窗口,可以从菜单栏中的Window菜单中选择要显示的窗口。也可使用功能键,比如用功能键F5可以切换到程序编辑窗口,F6可以切换到运行记录窗口,F7可以切换到输出窗口。

SAS主窗口标题栏下是**菜单栏**。SAS菜单是动态的,其内容随上下文而不同,即光标在不同窗口其菜单也不同。其中,File(文件)菜单提供有关SAS文件调入、保存、转换及打印的功能;Edit(编辑)菜单用于窗口的编辑,如清空、复制、剪切、粘贴、查找及替换等操作;View菜单可以打开或切换到SAS的某个工作窗口,如程序编辑窗口、SAS管理器窗口;Tools菜单可以打开一些SAS提供的小工具如图像编辑器,还包含修改SAS运行选项的功能;Run菜单用于程序执行、远程调用等,只在当前窗口是程序窗时有效;Solutions菜单是SAS的一些图形界面操作的模块的入口,比如SAS/INSIGHT、Analyst、ASSIST等。

菜单栏下是一个命令条和工具栏。**命令条**主要是用于与SAS较早版本的兼容,可以在这里键入SAS的显示管理命令。**工具栏图标**提供了常见任务的快捷方式,比如保存、打印、帮助,等等。鼠标光标在某一工具栏图标上停留几秒钟可以显示一个说明。工具栏也是动态的,当光标在程序编辑窗口时工具栏图标的解释如下:

 New —— 建立新的程序编辑窗口。

 Open —— 打开文件到程序编辑窗口。用户将一个指定文件调入到程序编辑窗口内。这个文件从此与程序编辑窗口相关联,以后的存盘操作将自动存入这个文件。

 Save —— 存盘,保存程序编辑窗口中的内容。注意,如果此窗口已经与一个文件相联系,单击此图标将覆盖文件的原有内容而不给予提示。

 Print —— 打印当前窗口内容。

-  Print preview —— 打印预览.
-  Cut —— 剪切选定文本.
-  Copy —— 复制选定文本.
-  Paste —— 粘贴. 注意这些操作是对 Windows剪贴板进行的, 可以用来与其他 Windows应用程序交换文本、数据等. 剪切或复制到剪贴板的内容可以粘贴到其他应用程序中, 而其他应用程序放到剪贴板的内容也可以粘贴到SAS的编辑窗口中.
-  Undo —— 撤销刚才的编辑操作.
-  New Library —— 建立新的SAS库.
-  SAS Explorer —— 打开SAS管理器窗口查看、管理SAS的各个库和库中的文件. 这是较新版本 SAS系统的增强功能, 我们可以比较方便地利用这个窗口来管理SAS文件.
-  Submit —— 提交编辑窗口中的程序.
-  Clear All —— 清空当前窗口内容.
-  Break —— 中断正在运行的SAS程序.
-  Help —— 进入SAS的帮助界面. 因为SAS是一个十分庞大的系统, 我们不可能完全了解SAS的每一个细节, 所以善于利用SAS的帮助系统对于学好、用好SAS是极为重要的. 也可以从菜单栏中进入SAS帮助.

SAS界面的状态栏中包含了当前工作目录, 这是打开或保存文件的默认目录. 双击此处可以更改当前工作目录.

1.1.1.3 简单运行样例

假设我们有一个班学生的数学成绩和语文成绩, 数学满分为100, 语文满分为120, 希望计算学生的平均分数(按百分制)并按此排名. 可以在程序窗口输入此程序:

```

title '95级1班学生成绩排名';
data c9501;
    input name $ 1-10 sex $ math chinese;
    avg = math*0.5 + chinese/120*100*0.5;
    cards;
李明      男  92 98
张红艺    女  89 106
王思明    男  86 90
张聪      男  98 109
刘颖      女  80 110
;
run;
proc print; run;
proc sort data=c9501;

```