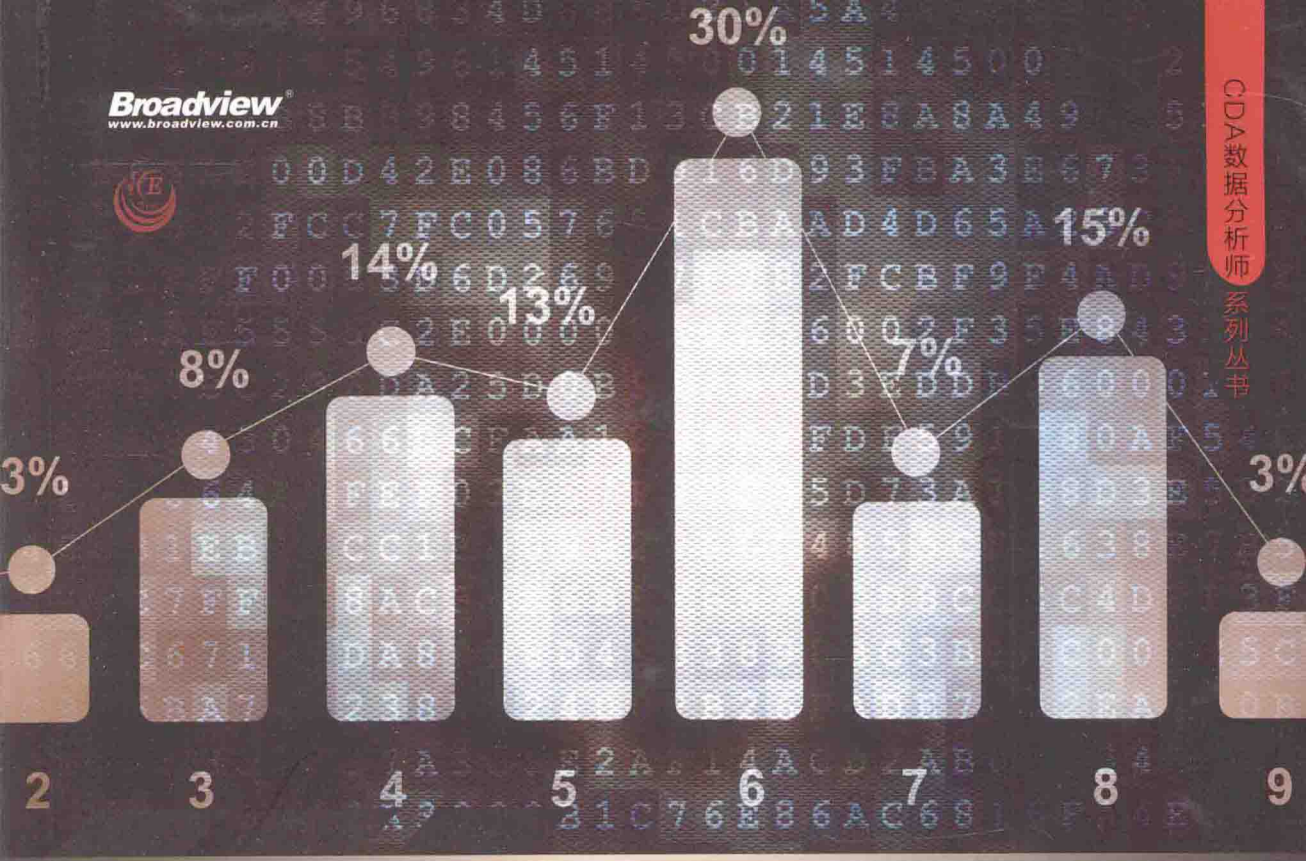

Broadview
www.broadview.com.cn

CDA数据分析师系列丛书

如虎添翼！

数据处理的SAS EG实现

人大经济论坛 主编　徐筱刚 编著

未来数据分析相关的就业岗位会有1000万人才缺口
CDA数据分析师系列丛书携你与时俱进！

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

CDA数据分析师系列丛书

如虎添翼！

数据处理的SAS EG实现

人大经济论坛 主编　徐筱刚 编著

電子工業出版社
Publishing House of Electronics Industry

内 容 简 介

作为 SAS EG 模块的首本中文教材，本书并非单纯的逐个讲解菜单的操作，而是将数据分析的基本思路、流程融入到软件的操作之中。每章通过设置商业背景，配以 SAS 理论的讲解形式更贴近读者的实际工作，使读者真正理解数据分析、数据处理的精髓。本书除讲解软件操作，还同时介绍了对应菜单操作的 SAS 程序语言实现过程，读者可以根据自己的需要逐步学习，进而走进用 SAS 程序处理数据的大门。

本书适合那些想了解数据预处理，或者被数据的预处理占去大部分时间而想提高效率，或者囿于菜单操作的局限性而希望通过程序实现的数据分析人员。

图书在版编目（CIP）数据

如虎添翼！数据处理的 SAS EG 实现 / 人大经济论坛主编；徐筱刚编著. —北京：电子工业出版社，2015.2
（CDA 数据分析师系列丛书）
ISBN 978-7-121-25245-7

Ⅰ. ①如… Ⅱ. ①人… ②徐… Ⅲ. ①数据处理—应用软件 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2014）第 307076 号

策划编辑：张慧敏
责任编辑：徐津平
印　　刷：三河市鑫金马印装有限公司
装　　订：三河市鑫金马印装有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编　100036
开　　本：787×980　1/16　印张：11.25　字数：275 千字
版　　次：2015 年 2 月第 1 版
印　　次：2015 年 2 月第 1 次印刷
印　　数：4 000 册　　定价：49.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

序言：这是一个用数据说话的时代

在 CDA（注册数据分析师）Level I 级教材付诸印刷之际，关于数据分析这个职业及其价值的报道就有很多。比如，下面两条报道就充分体现了在大数据时代下，数据分析的价值。这在以前是从来没有过的。

LinkedIn 的最新投票结果显示，‘统计分析和数据挖掘’是 2014 年最大的求职法宝。LinkedIn 对全球超过 3.3 亿用户的工作经历和技能进行分析，公布 2014 年最受雇主喜欢、最炙手可热的 25 项技能，其中位列榜首的是统计分析和数据挖掘。

麦肯锡公司的一份研究预测称，到 2018 年，在“具有深入分析能力的人才”方面，美国可能面临着 14 万到 19 万人的缺口，而“可以利用大数据分析来做出有效决策的经理和分析师” 缺口则会达到 150 万人。数据科学家将成为 2015 年最热门的职业。

早在 2010 年 2 月，肯尼斯•库克尔在《经济学人》上发表了一份关于管理信息的特别报告——《数据，无所不在的数据》，文中写道：“世界上有着无法想象的巨量数字信息，并以极快的速度增长……从经济界到科学界，从政府部门到艺术领域，很多地方都已感受到了这种巨量信息的影响。”2011 年，麦肯锡发布了《大数据：下一个具有创新力、竞争力与生产力的前沿领域》，使人们在这篇文章里认识到了数据的力量。于是，一夜之间，面向数据分析市场的新产品、新技术、新服务、新业态正在不断涌现。从个人、企业到国家层面，都把数据作为一种重要的战略资产，逐渐认识到了数据的价值，不同程度地渗透到每个行业领域和部门，大大提升了企业的经营利润，推动了经济的发展。

这是一个用数据说话的时代，也是一个依靠数据竞争的时代。目前世界 500 强企业中，有 90%以上都建立了数据分析部门。IBM、微软、Google 等知名公司都积极投资数据业务，建立数据部门，培养数据分析团队。各国政府和越来越多的企业意识到数据和信息已经成为企业的智力资产和资源，数据的分析和处理能力正在成为日益倚重的技术手段。

作为一个数学和统计学的强国，数据分析、数据挖掘和大数据价值挖掘行业在我国仍属于朝阳行业，数据分析人才仍然比较稀缺。各行各业在平常工作中积累的各种各样的数据分析问题仍然没有得到及时有效地解决，有些问题，还是关乎本行业发展的至关重要的问题。数据积累越来越多，期待解决分析的数据问题也越来越多，人们逐渐习惯使用数据作为决策的重要参考依据。据艾瑞的研究报告，未来与数据分析相关的就业岗位会在 1000 万人左右，而目前来说国内合格的数据分析师不足 5 万人，建立一个科学有效的数据分析师培训体系迫在眉睫。

在这样一个用数据说话的时代，积累了丰富的数据分析培训经验的人大经济论坛承担起使命，几番调查研究，几番反复推演论证，在 2013 年，这个大数据的“元年”，CDA 注册数据分析师应运而生！

2003 年，人大经济论坛依托中国人民大学成立，在金融、管理、统计领域已积淀 11 个年头，在国内享有良好声誉。

2006 年，人大经济论坛数据分析培训中心设立，至今经历 8 个春秋，建立了大陆、台湾一线师资团队，培养人才已达 3 万余人。

2013 年，“中国数据挖掘与数据分析俱乐部 CDMC”在人大经济论坛旗下成立，2014 年改名为“CDA 数据分析师俱乐部”。来自政府、金融、电信、零售、电商、互联网、教育等行业人士加入会员，成功举办了数十场行业聚会。紧接着，积累了数据分析培训丰富经验的人大经济论坛在国内展开 CDA 数据分析师系统培训和认证考试，成功见证了 1000 余名数据分析师的成长。

2015 年，人大经济论坛将提供高水平、多层次的数据分析培训服务，以在行业积累 多年的影响力，吸引更好更多的优秀师资，瞄准行业内重要的数据分析问题和难点，攻坚突破，建立更加规范的行业培训体系，引领数据分析培训行业向规范化、有效化和前瞻化方向发展，为数据分析培训做出应有的贡献。

其实，数学（含统计）和英语一样重要，都是人们不可或缺的重要技能。既然英语全民这么重视，数学及其数据分析的技能更加需求于方方面面，更应被做大做强。让我们共同期待人大经济论坛办成另一个数据的“新东方”！

覃智勇

2015 年 1 月 1 日

前　言

感谢您选择“CDA 数据分析师”Level I 学习系列丛书之《如虎添翼！数据处理的 SAS EG 实现》。

该丛书按照数据分析师规范化学习体系而定，对于一名初学者，应该先掌握必要的概率、统计理论基础，包括描述性分析、推断性分析、参数估计、假设检验、方差分析、回归分析等内容，这在第一本书《从零进阶！数据分析的统计基础》中进行了专业详细的讲解。其次，数据分析需要按照标准流程进行，即数据的获取、储存、整理、清洗、归约等系列数据处理技术，这在《如虎添翼！数据处理的 SAS EG 实现》中利用 SAS EG 和编程技术进行了操作过程的详解。最后，经过处理的数据需要根据业务问题，利用相关方法进行建模分析，得出结果，结果检验，绘制图表并解读数据，这在《胸有成竹！数据分析的 SAS EG 进阶》中进行了详细的讲解和操作分析。

CDA 数据分析师丛书整体风格是“理论>技术>应用”的一个学习过程，最终目的在于商业业务应用、职场数据分析，为欲从事于数据分析领域的各界人士提供了一个规范化数据分析师的学习体系。

读者对象

作为丛书中的一本，本书上承基础理论部分，下启最终建模及案例分析。本书将关注点集中到数据的探索及预处理上，通过本书的学习将会加深对基础理论部分的理解，为后续的建模分析做好数据上的准备。本书适合那些想了解数据预处理，或者被数据的预处理占去大部分时间而想提高效率，或者囿于菜单操作的局限性而希望通过程序实现的数据分析人员。

阅读指南

对数据分析师而言，合适的数据就像好的食材，对最终分析结果的影响不言而喻，但是在日常的工作中我们会经常遇到两个问题，一是数据的质量不高，数据在收集、存储等过程中不可避免地出现了脏数据、不一致数据、噪声数据、重复数据等，如果我们不做任何预处理而直接输入模型，就会出现“garbage in, garbage out”，即垃圾进垃圾出的情况。二是数据的形式不符，因为不同的模型，建模技术都有一定的前提假设，对数据的展现形式、分布状态等都有较为严格的要求，如果不做预处理，模型出来的结果很可能与数据底层真正蕴含的规律背道而驰，对这种形式的数据盲目地进行建模分析，极容易误人误己。

数据预处理占到整个数据挖掘的 60%～80%的时间，要想高效正确地完成数据的预处理工作其实不是一件容易的事情，本书作者根据在咨询公司、电信及金融行业的多年经验，将常用的数据预处理

思路融入到 SAS EG 的菜单操作中，并配以 SAS 程序的讲解，使得读者在熟悉菜单的同时，能用简单的 SAS 语言完成相对复杂的数据处理要求。

全书共分为九章：

第 1 章介绍了 SAS EG 软件，并介绍了三种常见的数据分析流程；

第 2 章介绍了如何通过多种方式使 SAS EG 可以轻松地访问多种形式的外部数据；、

第 3 章介绍了探索性数据分析的基本思路，以及数据清理的相关理论，并分别演示了如何对类别数据、数据数据进行清理；

第 4 章、第 5 章介绍了如何对数据观测进行筛选和排序、抽样，以及数据的分组和汇总，如何对数据进行转置，使用函数等，

第 6 章介绍了如何在整体上对数据集进行操作，包括如何对数据集进行横向连接和纵向连接，数据集之间的比较创建格式等。

第 7 章讲解了数据的可视化及图表、报告的编制方法。

第 8 章、第 9 章介绍了如何在 SAS EG 中运用提示、程序等来提高数据处理效率。

各部分相互独立，读者可以根据自己的需要选择性阅读。

本书特点

1. 关于 SAS EG 模块的首本中文教材；
2. 非单纯的逐个讲解菜单，而是将数据分析的基本思路、流程融入到软件的操作之中；
3. 每一章节通过设置商业背景，基本理论讲解的形式更贴近读者的实际工作；
4. 本书除讲解软件操作，还同时介绍了各种操作的 SAS 语言的实现过程，读者可以根据自己的基础逐步学习，进而走进 SAS 处理数据的大门。

学习方法

本书在编写上力求从读者的实战角度出发，每章基本上分为五部分：

1. 商业背景的介绍；
2. 相关的理论介绍；
3. EG 软件的解决方案；
4. 程序实现，包含实现菜单的程序的简单必要功能，读者可以轻松过渡到编程；
5. 扩展阅读，对于想深入学习 SAS 编程、数据准备的读者，进一步介绍了相关的学习内容及方向。

读者可以根据自己的需要来阅读，本书的菜单操作部分以 SAS Eenterprise guide 5.1 为例进行示范，代码部分以 SAS 9.3 为基础进行编写，读者在理解基本思路之后可以方便将其应用到其它版本之上。

售后服务

为方便读者学习，本书提供了书中实例的源文件下载，请读者进入人大经济论坛(http://bbs.pinggu.org/)，

注册后搜索“CDA 教材源文件”关键词下载相应的源文件。

本书读者可以在人大经济论坛的“数据挖掘与商业智能（http://bbs.pinggu.org/forum-133-1.html）”版块中就书中的问题进行提问，也欢迎大家就自己遇到的业务问题和大家讨论。同时，也可以向作者发邮件，作者邮箱为 xuxiaog_2003@163.com。

致谢

本书由人大经济论坛策划，徐筱刚负责编写和完成统稿。

丛书从策划到出版，倾注了电子工业出版社计算机图书分社张慧敏、石倩、官杨、张童等多位编辑的心血，特在此表示衷心地感谢！

为保证丛书的质量，使其更贴近读者，我们组织了人大经济论坛的多位版主和高级会员参与了本书的预读工作，他们是杨同梅、田佳、孙华枫、原瑜芬、叶阵雨、郑赟、李剑宇、江翊雪、陈鹏、刘莎莎、丁亚军。感谢各位预读员的辛勤、耐心与细致，使得本丛书能以更加完善的面目与各位读者见面，特别感谢覃智勇圆满地组织了本次预读工作和审校工作。

尽管作者们对书中的案例精益求精，但疏漏仍然在所难免，如果您发现书中的错误或某个案例有更好的解决方案，敬请登录社区网站向作者反馈，我们将尽快在社区中给出回复，且在本书再次印刷时修正。

再次感谢您的支持！

反侵权盗版声明

目　录

第 1 章

软件入门介绍

通过本章的学习，读者将对 SAS EG 的特点及窗口布局有个概略的了解，并熟悉常见的三种数据挖掘流程。

1.1 SAS EG 介绍

1.1.1 SAS EG 简介

SAS 系统全称为 Statistics Analysis System，最早由北卡罗来纳大学的两位生物统计学研究生编制，并于 1976 年成立了 SAS 软件研究所，正式推出了 SAS 软件。截至 2013 年，SAS 已在全球 130 多个国家拥有 7 万多公司和政府用户。SAS 具有强大的数据分析能力，SAS Enterprise Guide 是 SAS 其中的一个模块，简称为 SAS EG。

SAS EG 为用户提供了一个可视化的操作界面，以便快捷的管理数据，产生报告。SAS EG 提供如下功能：

- 便于管理 SAS 任务的图形化用户界面；
- 高度灵活和可扩展的编程语言；
- 丰富的立即可用的 SAS 过程；
- 灵活地运行在所有主流操作系统上，如 Windows、UNIX 和 z/ OS（OS/390）等；
- 能够访问几乎任何数据源，如 DB2、Oracle、Sybase、Teradata、SAP 和 Microsoft Excel 等；
- 支持全球最为广泛使用的字符编码方式。

使用 SAS Enterprise Guide 时，同时也在后台使用 SAS 软件。SAS Enterprise Guide 可以连接本地计算机上的 SAS，也可以连接其他计算机（即 SAS 服务器）上的 SAS。访问数据和创建任务时，SAS Enterprise Guide 将生成 SAS 代码。运行任务时，将生成的代码发送至 SAS 进行处理，然后将结果返回至 SAS Enterprise Guide，如图 1-1 所示。

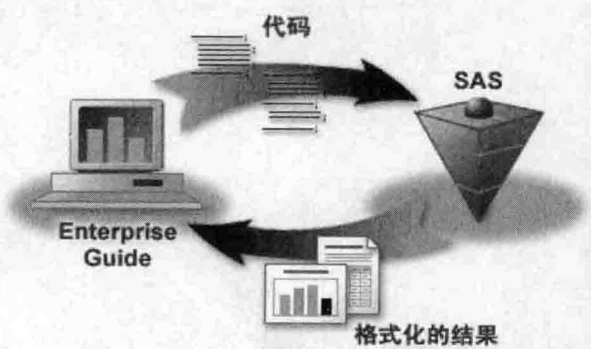

图 1-1

SAS Enterprise Guide 的核心是 Base SAS 软件。Base SAS 可以利用其他组件扩展功能。因而 SAS EG 可以方便快捷的调用 SAS STAT、SAS/ETS、SAS/ GRAPH 等模块。

为何使用 SAS Enterprise Guide ？

- SAS Enterprise Guide 是整理、分析和报告数据的利器，初学者可以使用菜单点击的方式完成丰富的数据处理和统计分析功能；
- SAS 程序员可以借助 EG 为其提供的优秀的编程界面和丰富的辅助工具，提高效率；
- SAS 商业分析（BI）可以方便使用 EG 的创建存储过程（Create Stored Processes），查看 OLAP 立方体（OLAP cubes），生成 SAS 报告。

1.1.2 SAS EG 的窗口及菜单

SAS EG 的窗口界面如图 1-2 所示。

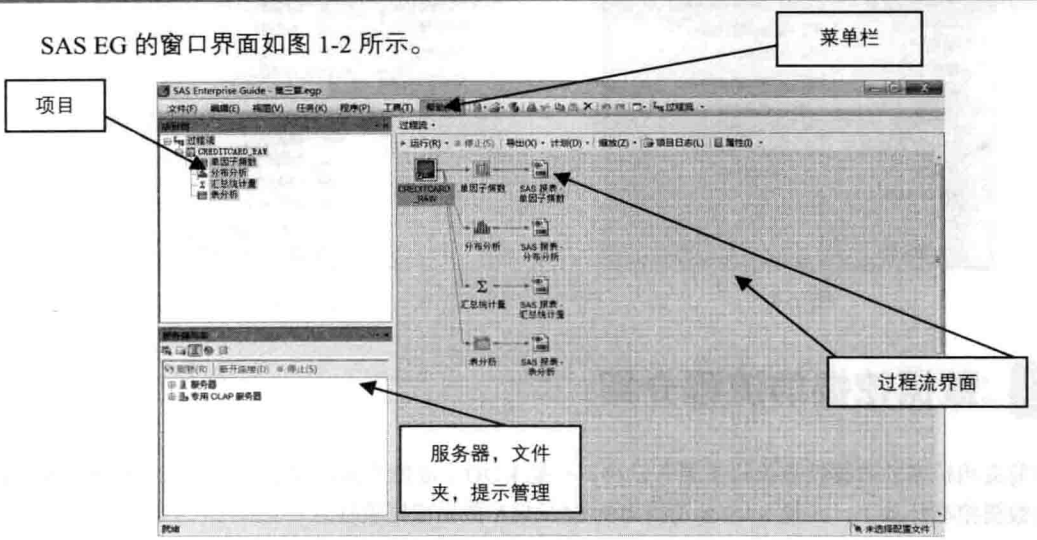

图 1-2

菜单

项目是指管理相关数据、任务、代码和结果的集合。“项目树”显示当前活动项目以及其相关项（数据、代码、注释和结果）的层次视图。可以删除、重命名和重新排序项目中的各个项。也可以运行项目，或预定某个项目在特定时间运行。

任务可以理解为一个分析过程，不同的任务运行之后可生成 SAS 代码、数据集或者报表等。

数据菜单主要描述了对数据行和列的基本处理，如图 1-3 所示。

描述菜单主要介绍了各种统计量以及各种数据特征的分析和展示，如图 1-4 所示。

图形菜单介绍了 SAS EG 所支持的图形的制作，如图 1-5 所示。

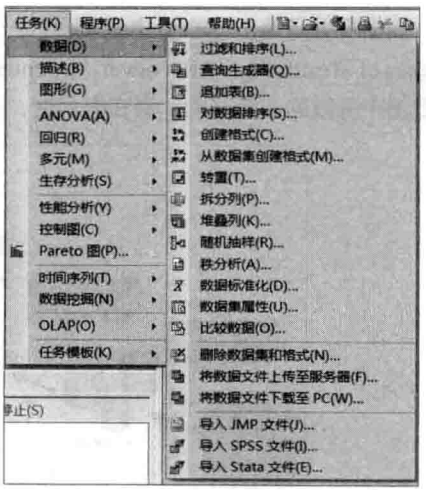

图 1-3