



本书是Stata入门必备书籍，全面、系统地讲解了软件的基础知识、数据访问及管理等内容，并配有大量来源于实践的案例

本书结构系统，讲解清晰，通俗易懂，实用性强，是统计人员、Stata初中级用户、医学及生物研究人员、数据管理员以及其他数据分析人员的必备参考书籍



技术丛书



Data Analysis and Statistics in Stata

# Stata数据统计 分析教程

廉启国◎编著



机械工业出版社  
China Machine Press

Data Analysis and Statistics in Stata

# Stata 数据统计 分析教程

廉启国◎编著



机械工业出版社  
China Machine Press

## 图书在版编目 ( CIP ) 数据

Stata 数据统计分析教程 / 廉启国编著. —北京: 机械工业出版社, 2015.4

ISBN 978-7-111-50028-5

I. S… II. 廉… III. 统计分析—应用软件—教材 IV. C819

中国版本图书馆 CIP 数据核字 ( 2015 ) 第 084296 号

## Stata数据统计分析教程

出版发行: 机械工业出版社 (北京市西城区百万庄大街22号 邮政编码: 100037)

责任编辑: 李华君

责任校对: 董纪丽

印 刷: 北京瑞德印刷有限公司

版 次: 2015年5月第1版第1次印刷

开 本: 185mm×260mm 1/16

印 张: 24

书 号: ISBN 978-7-111-50028-5

定 价: 69.00元

凡购本书, 如有缺页、倒页、脱页, 由本社发行部调换

客服热线: (010) 88378991 88361066

投稿热线: (010) 88379604

购书热线: (010) 68326294 88379649 68995259

读者信箱: [hzsj@hzbook.com](mailto:hzsj@hzbook.com)

版权所有·侵权必究

封底无防伪标均为盗版

本书法律顾问: 北京大成律师事务所 韩光/邹晓东

## 序

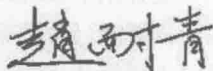
---

复旦大学公共卫生学院是国内最早开设卫生统计学课程的院校之一，也是最早提出用 Stata 软件进行教学的院校之一。该院从 1988 年起开设了统计软件课，通过教学实践，最终确定 Stata 软件是最为合适的教学软件。该软件功能强大，操作简便，输出结果针对性极强，并且其许多功能都是针对医学研究背景设计的。

从 2004 年开始，我们开始摸索新的教学方法，尝试以研究设计为主线，以数据类型为分类，以数据分析为教学目标，在统计学教学中融入统计软件 Stata 的基本操作，淡化统计计算公式的步骤，借助计算机辅助教学，强化统计学基本概念，强化学生数据分析的能力。

在学习了复旦大学公共卫生学院开设的卫生统计学课程后，本书作者廉启国熟悉并喜欢上了 Stata，且将 Stata 作为工作中数据分析的首选统计软件。Stata 软件的更新速度非常快，一般保持每两年一个大版本的更新速度。本书基于 Stata/MP 13.1，深入浅出地对 Stata 的常用功能及最新功能进行了介绍。本书的第 1 章介绍了 Stata 的入门知识，第 2 章至第 10 章介绍了 Stata 在数据访问、数据管理、数据呈现和数据分析 4 个重要方面的强大功能，第 11 章介绍了 Stata 的编程技巧。此外，本书还专门在第 12 章和第 13 章介绍了 Meta 分析和样本量估算的 Stata 实现。

我相信，此书的出版将为国内许多师生使用 Stata 提供很好的帮助。



复旦大学公共卫生学院

2015 年 1 月

# 前言

---

## 为什么要写这本书

Stata 是一款无与伦比的统计软件，它优雅、高效且易学。Stata 诞生于 1985 年 1 月，第一个版本只有 44 个命令和 175 页文档。2015 年是 Stata 诞生 30 周年，目前 Stata 已经发生了翻天覆地的变化，历经 30 个版本的迭代，已经升级至 14.0 版，支持矩阵编程和多核处理器，且提供 20 本合计超过 11 000 页的文档。Stata 日趋完美，用户遍布 200 多个国家和地区，已成为最重要的统计软件之一。

作者是 2004 年在复旦大学读硕士研究生时，在赵耐青教授和张文彤博士的课程中接触并喜欢上 Stata 的，十年弹指一挥间，也积累了一些 Stata 的使用经验。一个偶然的的机会，在机械工业出版社华章公司李华君编辑的建议下，开始尝试本书的撰写工作。写作的过程也是一个重新学习和梳理的过程，因为将自以为掌握的知识写出来并非易事。

Stata 8 是 Stata 发展的一个重要里程碑，但是在生物统计学和医用统计学领域，有关 Stata 的书籍还不是很丰富，很多书籍也都是介绍 Stata 8 之前的软件，未能紧跟 Stata 的发展步伐。本书定位为 Stata 入门级的书籍，以 Windows 平台下 Stata 13.1 MP 版本为基准，围绕数据访问、数据管理、数据呈现和数据分析 4 个核心问题进行了详细阐述（并介绍了大量实用且免费的第三方命令）。此外，本书还介绍了 Stata 编程以便提高用户日常科研工作效率并介绍了工作流程以有助于用户更好地实施项目管理（含数据的加密和恢复）。最后，本书通过两章对近年来比较热门的 Meta 分析和日常科研中频繁使用的样本量估计进行了介绍。

## 读者对象

本书适合的阅读对象如下：

- Stata 初中级用户
- 医务人员
- 医学/生物科研人员
- 数据管理员
- 生物统计师
- 其他行业有数据分析需求的人员

## 如何阅读本书

本书的每个章节都是一个相对独立的知识块，读者可以选择从头开始逐章阅读，如果有一定 Stata 基础和使用经验的话也可以根据目录跳转到感兴趣的章节。本书主要分为 6 部分，总计 13 章，另外还有 5 个附录。基本结构如下。

### 第一部分 软件入门（第 1 章）

第一部分介绍了 Stata 的基础知识，包括 Stata 软件概述（发展历史与版本选择、正确读写与文献引用、软件特点与优势）、Stata 操作入门（安装与激活、启动与退出、操作方式、结果输出、自定义设置）、Stata 使用基础（文件格式、变量类型、运算符、语法规则）、Stata 升级扩展（软件升级、第三方命令的查找与安装、帮助文件及学习资源）。

### 第二部分 数据访问（第 2 章）

第二部分介绍了 Stata 在数据导入/导出方面的功能，包括 Stata 自带的导入和导出功能、数据导入的第三方命令（`usespss`）、数据转换软件（`Stat/Transfer` 和 `EpiMate`）、数据录入软件（`EpiData Entry`）。

### 第三部分 数据管理（第 3~6 章）

准确的数据是以后进行科学分析的基础，所以在进行数据分析之前，确保数据已被清理干净是非常必要的。数据管理是个非常枯燥的工作，但它是一切后续工作的基础。第三部分介绍了 Stata 在数据管理方面的强大功能（尤其是时间戳功能），包括数据清理、标签和注释、变量加工和数据集加工 4 章。

### 第四部分 数据呈现篇（第 7~9 章）

第四部分介绍了 Stata 在数据呈现方面的功能，包括统计描述、报表制作和图形绘制 3 章。Stata 在数据呈现方面的功能非常强大，尤其是 Stata 的绘图效果堪称完美，建议对图形有特别要求的读者仔细阅读第 9 章的内容。

### 第五部分 数据分析（第 10 章）

Stata 的数据分析非常完善，由于篇幅所限，本书只介绍了统计分析中最常用的一些 Stata 的功能，包括正态性检验和变换、t 检验、方差分析、非参数检验、卡方检验、流行病学表格分析、相关分析、线性回归模型、logistic 回归模型、等效性检验。在多元回归模型里特别增加了流行病学家和统计学家对自变量纳入的不同考虑。

### 第六部分 科研必备（第 11~13 章）

第六部分是为从事科研工作和数据处理工作的读者准备的，包括 Stata 编程基础、Meta 分析、样本量与检验效能 3 章。建议对编程感兴趣的读者仔细阅读编程基础一章，医学/生物相关的专业人员需要特别关注的是 Meta 分析以及样本量与检验效能两章。

## 免责声明

本书中提及了大量的软件和服务，包括商业的（如：Stata、SAS、SPSS、Stat/Transfer、PASS、

Beyond Compare、FileWall、BitLocker、金山数据恢复软件、顶尖数据恢复软件)、免费的(如: EpiData、EpiMate、Power and sample size、uWall、魔方数据恢复、Eraser、金山快盘、有道云笔记、百度云盘、OneDrive、DropBox 和 Google Drive)和开源的(如: R) 3 类。作者声明自己与这些软件和服务无任何利益关系, 仅为提供更多信息。

此外, 本书借鉴了大量的参考资料, 并在所引用之处都尽可能地标注了文献出处, 若有疏漏之处, 作者在此深表歉意, 并请及时联系我们, 以便在后续版本中增加。

## 勘误和支持

由于作者的水平所限, 经验不足, 加之编写时间不尽宽裕, 书中疏漏、错误之处在所难免, 不妥之处恳请读者批评指正。

本书的勘误信息会发布在公卫人的 Stata 专版(网址: [www.epiman.cn/forum-9-1.html](http://www.epiman.cn/forum-9-1.html)), 作者会在 Stata 专版中不定期更新书中的遗漏。当然, 也欢迎读者将在阅读本书时遇到的疑惑、错误和建议在 Stata 专版中发帖提出。如果有些话题比较隐私, 可以发送邮件至作者的邮箱([qglian@fudan.edu.cn](mailto:qglian@fudan.edu.cn)), 期待能够收到各位的真挚反馈。

## 致 谢

---

首先要感谢的是我的父亲和母亲以及岳父和岳母，你们在生活上的支持让我远离各种琐事，才可能将业余时间集中投入到本书的创作中。

其次要感谢我的妻子石慧颖女士，若不是她的持续鼓励和背后默默的支持，按时交稿是一件无法完成的任务。特别感谢我刚满3岁的女儿廉孟洁，她每天都为家庭带来欢声笑语，也让我学会对世界始终保持一颗好奇的心。

还要感谢机械工业出版社华章公司李华君编辑的慧眼识才。因为平时的科研任务很重，所以虽然出书的想法产生已久，但若不是李编辑的鼓励，本书依然还在纸上谈兵的阶段。

本书内容参考了大量的专著、文章、文档和网站，此外，本书也介绍了众多 Stata 用户免费贡献的第三方命令，在此表示感谢。正因为有你们，Stata 社区才会如此繁荣。

最后，谨以此书献给 Stata 软件诞生 30 周年！

廉启国

2015 年 1 月于上海



# 目 录

---

序

前言

致谢

## 第一部分 软件入门

第 1 章 Stata 入门	1
1.1 Stata 软件概述	1
1.1.1 Stata 的发展历史与版本选择	1
1.1.2 Stata 的正确读写与文献引用	2
1.1.3 Stata 的软件特点与优势	3
1.1.4 Stata 的主要功能模块	5
1.2 Stata 操作入门	6
1.2.1 Stata 的安装与激活	6
1.2.2 Stata 的启动与退出	7
1.2.3 Stata 的操作方式	7
1.2.4 Stata 的结果输出	8
1.2.5 Stata 的自定义设置	11
1.3 Stata 使用基础	13
1.3.1 Stata 的文件格式	13
1.3.2 Stata 的变量类型	13
1.3.3 Stata 的运算符	14
1.3.4 Stata 的语法规则	15
1.3.5 Stata 的使用实例	16
1.4 Stata 升级扩展	16
1.4.1 Stata 软件升级	16
1.4.2 Stata 第三方命令的查找与安装	18
1.4.3 Stata 的帮助文件及学习资源	22

## 第二部分 数据访问

第 2 章 数据访问 .....	23
2.1 数据格式概述 .....	23
2.2 直接录入数据 .....	24
2.2.1 用 Stata 的数据编辑器录入 .....	24
2.2.2 用 Stata 的命令窗口录入 .....	26
2.3 数据的读取与保存 .....	26
2.3.1 直接读取和保存 Stata 格式的数据 .....	26
2.3.2 导入/导出 Excel 格式的数据 .....	27
2.3.3 导入/导出带分隔符的文本格式数据 .....	29
2.3.4 导入/导出自由格式的文本格式数据 .....	30
2.3.5 导入/导出固定格式的文本格式数据 .....	33
2.3.6 导入/导出 SAS XPORT 格式的数据 .....	34
2.3.7 导入/导出 XML 格式的数据 .....	35
2.3.8 导入/导出 SPSS 格式的数据 .....	36
2.4 数据格式转换软件 Stat/Transfer 简介 .....	37
2.4.1 Stat/Transfer 支持的数据类型 .....	37
2.4.2 Stat/Transfer 数据转换 .....	39
2.5 数据录入软件 EpiData Entry 简介 .....	39
2.5.1 建立调查表文件 .....	40
2.5.2 生成 REC 文件并建立 CHK 文件 .....	40
2.5.3 录入数据并导出 .....	41
2.5.4 EpiData Entry 伴侣 EpiMate 简介 .....	41

## 第三部分 数据管理

第 3 章 数据清理 .....	43
3.1 双次录入数据的一致性检验 .....	43
3.2 逐个变量对数据进行检查 .....	44
3.3 通过交叉表对数据进行检查 .....	48
3.4 通过分层对数据进行检查 .....	51
3.5 通过逻辑对数据进行检查 .....	52
3.6 更正数据 .....	53
3.7 识别重复记录 .....	54
3.8 对缺失值进行检查 .....	56
3.8.1 缺失值编码 .....	56
3.8.2 缺失值函数 .....	57
3.8.3 缺失值检查 .....	57

3.9 数据清理的注意事项	59
<b>第4章 标签和注释</b>	<b>60</b>
4.1 描述数据集	60
4.2 给数据集加标签	61
4.3 给变量加标签	62
4.4 给变量值加标签	65
4.5 管理标签	66
4.5.1 查看全部变量值标签的名称	66
4.5.2 查看变量值标签	66
4.5.3 查看变量值标签的详细内容	67
4.5.4 为变量值标签加数值前缀	67
4.5.5 复制变量值标签	68
4.5.6 移除变量值标签	68
4.5.7 删除变量值标签	68
4.5.8 保存变量值标签	69
4.6 给缺失值加标签	69
4.7 不同语言的标签	70
4.8 给数据集加注释	70
4.9 给数据集加时间戳	72
<b>第5章 变量加工</b>	<b>75</b>
5.1 查看变量与记录	75
5.1.1 查看变量	75
5.1.2 查看记录	77
5.2 删除变量与记录	77
5.2.1 保留/删除变量	77
5.2.2 保留/删除记录	77
5.3 新建变量	78
5.3.1 用 generate/replace 产生新变量	78
5.3.2 用 egen 产生新变量	79
5.3.3 克隆变量	80
5.3.4 新建分拆变量	81
5.3.5 新建指示变量	82
5.4 重命名变量	84
5.4.1 单个变量重命名	84
5.4.2 批量变量重命名	84
5.4.3 变量名大小写转换	85
5.4.4 批量变量名大小写转换	86

5.5 更改变量的格式	86
5.5.1 变量存储格式的起源	86
5.5.2 变量存储格式简介	86
5.5.3 更改变量的存储格式	87
5.5.4 更改变量的显示格式	87
5.6 调整变量的顺序	88
5.7 对变量的值进行排序	88
5.8 对变量进行编码	89
5.8.1 数值变量的编码	89
5.8.2 字符变量的编码	90
5.8.3 缺失值的编码	91
5.9 数值变量与字符变量的转换	91
5.9.1 字符变量转换为数值变量	91
5.9.2 数值变量转换为字符变量	92
5.10 日期时间变量	92
5.10.1 日期数据的导入	92
5.10.2 日期数据的运算	93
5.11 命名变量名的注意事项	93
<b>第6章 数据集加工</b>	<b>95</b>
6.1 数据集的合并	95
6.1.1 数据集的纵向追加	95
6.1.2 数据集的横向合并	97
6.1.3 数据集的交叉合并: 组内交叉	100
6.1.4 数据集的交叉合并: 一一交叉	100
6.2 数据结构的转换	101
6.2.1 数据的长型和宽型格式	101
6.2.2 宽型格式转换为长型格式	101
6.2.3 长型格式转换为宽型格式	102
6.3 数据转置	102
6.4 扩展数据	103
6.5 堆栈数据	104
6.6 压缩数据: 生成统计量	104
6.7 压缩数据: 生成频数或百分数	106
6.8 数据抽取	106
6.9 生成随机数据	107
6.10 缺失值填补	108

## 第四部分 数据呈现

第 7 章 统计描述	111
7.1 计量资料的统计描述	111
7.1.1 集中趋势的统计描述	111
7.1.2 离散趋势的统计描述	114
7.1.3 分布特征的统计描述	115
7.2 计量资料的参数估计	116
7.3 计数资料的统计描述	117
7.3.1 单个分类变量的统计描述	117
7.3.2 多个分类变量的统计描述	118
7.4 标准化法	119
7.4.1 标准化法的意义及基本思想	119
7.4.2 直接法	119
7.4.3 间接法	121
7.4.4 注意事项	123
第 8 章 报表制作	124
8.1 Stata 报表呈现	124
8.1.1 使用命令 tabulate	124
8.1.2 使用命令 table	124
8.1.3 使用命令 tabstat	127
8.1.4 使用命令 collapse	128
8.1.5 使用命令 contract	130
8.1.6 使用命令 statsby	131
8.2 Stata 报表呈现的第三方命令	132
8.3 Stata 报表导出	133
8.3.1 Stata 报表导出的官方命令	133
8.3.2 Stata 报表导出的第三方命令	135
第 9 章 图形绘制	136
9.1 Stata 图形概述	136
9.1.1 Stata 图形组成	136
9.1.2 Stata 绘图命令	136
9.1.3 Stata 图形格式	137
9.1.4 Stata 图形坐标轴选项	138
9.2 Stata 图形编辑器	138
9.2.1 使用菜单绘制图形	139
9.2.2 启用 Stata 图形编辑器并进行个性化设置	140

9.2.3	启用 Stata 图形编辑器的绘图记录仪	140
9.2.4	取舍: 图形编辑器与 Stata 命令	141
9.3	二维图	141
9.3.1	散点图	141
9.3.2	线图	145
9.3.3	面积图	146
9.3.4	条形图	147
9.3.5	区间图	149
9.3.6	分布图	149
9.4	散点图矩阵	150
9.5	条图	151
9.6	箱图	153
9.7	点图	155
9.8	饼图	156
9.9	图形的标准选项	158
9.9.1	创建和控制标题	158
9.9.2	使用图形格式控制图形外观	159
9.9.3	调整图形及其元素大小	159
9.9.4	调整图形区域的外观	160
9.10	修改图形的风格	160
9.10.1	角度	160
9.10.2	色彩	160
9.10.3	钟表方位	161
9.10.4	指南针方位	161
9.10.5	连接点	162
9.10.6	线条式样	163
9.10.7	线条宽度	163
9.10.8	页边	164
9.10.9	标记大小	164
9.10.10	标记符号	165
9.10.11	方向	165
9.10.12	文字大小	165
9.11	绘制地图数据	166
9.12	图形管理与控制	168
9.12.1	图形的存储	168
9.12.2	图形的重新展示	168
9.12.3	图形的合并	169
9.12.4	图形的输出	169

9.13 更多第三方绘图命令	170
9.13.1 小提琴图	170
9.13.2 雷达图	170
9.13.3 六图	171
9.13.4 更多资源	171

## 第五部分 数据分析

第 10 章 假设检验	173
10.1 正态性检验与正态性变换	173
10.1.1 正态性检验	173
10.1.2 正态性变换	176
10.2 t 检验	179
10.2.1 t 检验的基本原理	179
10.2.2 样本均数与总体均数的比较	179
10.2.3 成组设计两样本均数的比较	181
10.2.4 配对设计两样本均数的比较	184
10.2.5 两组间多个变量之间的均值比较	185
10.3 方差分析	186
10.3.1 方差分析的基本思想	186
10.3.2 单因素方差分析	186
10.3.3 两因素方差分析和多因素方差分析	188
10.3.4 协方差分析	188
10.4 非参数检验	190
10.4.1 非参数检验概述	190
10.4.2 样本中位数与总体中位数的比较	190
10.4.3 两个配对样本的非参数检验	191
10.4.4 两个独立样本的非参数检验	193
10.4.5 两个独立样本的非参数检验 (多个变量)	194
10.4.6 多个独立样本的非参数检验	195
10.4.7 配伍设计的多组秩和检验	196
10.5 卡方检验	198
10.5.1 卡方检验的基本原理	198
10.5.2 四格表的卡方检验	198
10.5.3 配对卡方检验	200
10.5.4 列联表分析	201
10.5.5 分层卡方分析	204
10.5.6 一致性检验	206
10.6 流行病学表格分析	207



10.6.1	成组病例对照研究	207
10.6.2	配对病例对照研究	212
10.6.3	队列研究	214
10.7	相关分析	217
10.7.1	相关分析的指标体系	217
10.7.2	Pearson 相关系数	217
10.7.3	Spearman 相关系数	218
10.7.4	Kendall 等级相关系数	219
10.8	线性回归模型	220
10.8.1	线性回归模型简介	220
10.8.2	线性回归模型分析步骤	220
10.8.3	自变量的筛选方法	222
10.8.4	衡量回归方程的标准	223
10.9	logistic 回归模型	224
10.9.1	logistic 回归模型简介	224
10.9.2	两分类 logistic 回归 (非条件 logistic 回归)	224
10.9.3	模型拟合效果的判断	226
10.9.4	两分类 logistic 回归 (条件 logistic 回归)	231
10.9.5	多分类无序 logistic 回归	232
10.9.6	多分类有序 logistic 回归	234
10.10	等效性检验	235
10.10.1	等效性检验和传统差异性检验的区别	235
10.10.2	均值等效性 t 检验	236
10.10.3	比例等效性 z 检验	237
10.10.4	配对数据的随机等效性检验	238
10.10.5	两样本的随机等效性秩和检验	238
10.10.6	配对二分类数据的随机等效性 z 检验	238
10.10.7	交叉设计的等效效应检验	239

## 第六部分 科研必备

第 11 章	Stata 编程基础	241
11.1	do 文件简介	241
11.2	do 文件的内容	242
11.2.1	版本控制	242
11.2.2	命令注释与空行	243
11.2.3	超长命令行	244
11.3	do 文件的运行	245
11.3.1	结果的保存	245



11.3.2	控制分页符	246
11.3.3	错误及调试	246
11.3.4	其他 do 文件的调用	247
11.4	do 文件的最优规则	247
11.4.1	稳健性	247
11.4.2	可读性	248
11.5	项目管理器	248
11.6	Stata 宏语句	249
11.6.1	宏的指定与引用	250
11.6.2	宏的扩展函数	250
11.6.3	调用 Stata 的计算结果	251
11.7	Stata 循环语句	252
11.7.1	forvalues 循环语句	252
11.7.2	foreach 循环语句	253
11.7.3	while 循环语句	255
11.8	include 命令	255
11.9	临时变量	256
11.10	编写 Stata 程序 (ado 文件)	257
11.10.1	ado 文件简介	257
11.10.2	编写一个简单的 ado 文件	258
11.10.3	为新建命令编写帮助文件	259
11.10.4	编写命令的注意事项	260
<b>第 12 章 Meta 分析</b>		<b>262</b>
12.1	Meta 分析简介	262
12.1.1	Meta 分析的起源	262
12.1.2	Meta 分析与系统评价的关系	263
12.1.3	Meta 分析的指征、特点、目的和优点	263
12.1.4	Meta 分析的制作步骤	264
12.1.5	Meta 分析的常见类型	265
12.1.6	Meta 分析的效应量和效应模型	265
12.1.7	Meta 分析的偏倚及控制	266
12.1.8	Meta 分析的报告规范	267
12.1.9	Meta 分析的注意事项	268
12.2	Stata 的 Meta 分析命令	268
12.2.1	Meta 分析命令的安装	269
12.2.2	Meta 分析命令简介	270
12.3	二分类数据的 Meta 分析	272