北京市版权局著作权合同登记号　　图字：01-2017-4222

Mehmet Mehmetoglu, Tor Georg Jakobsen

Applied Statistics Using Stata: A Guide for the Social Sciences

EISBN: 978-1473913233

图书在版编目(CIP)数据

Stata统计分析：社会科学应用指南 / (挪) 穆罕默德・梅赫梅托，(挪) 托尔・格奥尔格・雅各布森著；柏建岭，曾永艺译. —北京：清华大学出版社，2021.7

（新时代・技术新未来）

书名原文：Applied Statistics Using Stata: A Guide for the Social Sciences

ISBN 978-7-302-54600-9

Ⅰ. ①S…　Ⅱ. ①穆… ②托… ③柏… ④曾…　Ⅲ. ①统计分析—应用软件　Ⅳ. ①C819

中国版本图书馆 CIP 数据核字（2020）第 002560 号

责任编辑：刘　洋
封面设计：徐　超
版式设计：方加青
责任校对：宋玉莲
责任印制：沈　露

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编：**100084
社 总 机：010-62770175　　**邮　　购：**010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市金元印装有限公司
经　　销：全国新华书店
开　　本：187mm×235mm　　**印　　张：**23.5　　**字　　数：**426 千字
版　　次：2021 年 7 月第 1 版　　**印　　次：**2021 年 7 月第 1 次印刷
定　　价：118.00 元

产品编号：074345-01

内 容 简 介

本书基于社会学领域学生和学者的需求，将统计学的理论概念和详细的技术指导有机结合起来，通过众多来自社会学不同领域的有趣示例来呈现丰富的统计方法和模型，鼓励读者在了解理论的同时学习应用 Stata 软件来实现研究的目的。本书除了用 5 个章节渐进式地详细阐述线性回归模型之外，还进一步涵盖 logistic 回归、多层次分析、面板数据分析、探索性因子分析、结构方程模型和验证性因子分析等内容。本书通过配套网站提供各章配套的测试题、视频、数据集和 Stata 代码，方便读者学习并检查学习效果。

本书可作为社会学领域本科生或研究生定量研究课程的教材或参考书，也可作为想要学习应用 Stata 软件进行定量研究的社会学者的参考书。

致　　谢

我们想要感谢 SAGE 出版社聘请的匿名评审专家提供的很多有用且鼓舞人心的反馈意见。还要感谢我们在 StataCorp 的同事 Bill Rising，他对本书的每一个章节都进行了详尽的评论并提出了很多有价值的意见和建议。感谢另一位 StataCorp 的同事 Kristin MacDonald，她对结构方程模型一章提出了非常有用的意见。Rising 和 Kristin 正是 StataCorp 众所周知的专业水准声誉的明证。还要感谢我们的同事 Stein Are Sæther、Giovanni Cerulli、Sergio Venturini、Kjell Hines、Arild Blekesaune、Marthe L. Holum、Zan Strabac、Jon Olaf Olaussen、Jonathon W. Moses、Mikael Knutsson、Jo Jakobsen 和 Morten Blekesaune，感谢他们对本书不同章节提出的建议和意见，以及对本书写作的全程支持。此外，我们感谢 SAGE 出版社的专业支持团队，包括编辑 Jai Seaman、Alysha Owen、Ian Antcliff、Sally Ransom、Lily Mehrbod、Vanessa Harwood 以及文字编辑 Richard Leigh。最后，感谢我们各自的家庭（Mehmet 的 Rannvei 和 Deniz 以及 Tor Georg 的 Marthe 和 Sofie），感谢他们始终不渝的支持。

作者简介

Mehmet Mehmetoglu 是挪威科技大学（Norwegian University of Science and Technology，NTNU）心理学系的研究方法教授。他的研究兴趣包括消费者心理学、进化心理学以及统计方法。Mehmetoglu 已经在大约 30 份不同的同行评审国际刊物上发表或合作发表论文，这些刊物包括 *Scandinavian Journal of Psychology*（《斯堪的纳维亚心理学期刊》）、*Personality & Individual Differences*（《个性与个体差异》）、*Evolutionary Psychological Science*（《演化心理科学》）等。

Tor Georg Jakobsen 是挪威科技大学商学院的政治科学教授。他的研究兴趣包括政治行为、和平研究以及统计方法。Jakobsen 已经在包括 *European Sociological Review*（《欧洲社会学评论》）、*Work, Employment & Society*（《工作、雇佣与社会》）、*Conflict Management & Peace Science*（《冲突管理与和平科学》）等刊物上发表或合作发表了论文。

前　言

对学生和学者来说，知道如何运用统计学来解决社会问题是一项关键技能。想要得到和开发这样的技能需要理解不同统计技术（如线性回归模型、因子分析等）背后的原理，同时学会使用灵活且用户界面友好的软件进行分析。在本书中，我们通过揭示每项统计技术背后的原理并提供 Stata 软件的应用示例，试图帮助读者达到这两个重要目标。基于我们的座右铭“万物皆回归”，我们将线性回归模型作为解释不同统计技术的首要框架。

对线性回归模型的深入理解为学习其他统计技术打下了基础——不管这些技术是简单的（如 *t* 检验）还是高级的（如结构方程模型）。这也是我们选择深入剖析线性回归模型及其拓展技术的原因所在。在读过相关章节之后，你就会认识到线性回归模型可以很好地替代传统的独立样本组间比较的方差分析（ANOVA）。线性回归方法也成为理解多层次回归技术的纽带，而后者是分析重复测量数据时方差分析方法的有力替代。

若想从本书中有所获，读者最好具备关于基础统计学的背景知识，并对统计推断有所了解。读者不需要具备使用 Stata 的经验，在将 Stata 应用于不同统计技术之前，我们会用一章的内容来详细介绍该软件的使用方法。当你阅读本书的每一章时，我们强烈建议你打开 Stata 软件，以便复制和重现相关统计分析的过程与结果。在此之前，你需要登录网址 https://study.sagepub.com/login?destination=node/30193，浏览并下载本书的配套材料（如数据集、Stata 代码、期刊文章示例等）。

译者序

挪威科技大学穆罕默德 • 梅赫梅托和托尔 • 格奥尔格 • 雅各布森两位教授编写的《Stata 统计分析：社会科学应用指南》一书具备以下几点突出特色：①将统计学的理论概念和详细的技术指导有机结合起来，通过众多来自社会学不同领域的有趣示例来呈现丰富的统计方法和模型，鼓励读者在了解理论的同时学习应用 Stata 软件来实现研究的目的；②基于作者“万物皆回归”的座右铭，将线性回归模型作为统计建模和解释的首要框架，这有助于读者融会贯通 *t* 检验、ANOVA、结构方程模型等不同统计技术；③第 11 章和第 12 章对社会科学研究领域中常见的因子分析和结构方程模型展开循序渐进的讨论，结合具体的研究案例给出 Stata 代码和外部命令，让读者在 SPSS+AMOS 之外又多了一种选择。总之，这本优秀的著作不仅适用于学习社会科学的在校本科生和研究生，也适用于需要进行数据分析的研究人员。

由于本人在人大经济论坛做版主期间认识了厦门大学的曾永艺助理教授和北京林业大学的李强副教授，本人接受清华大学出版社责任编辑刘洋的邀请牵头翻译此书后便联系了他们，他们都很乐意参与此书的翻译工作，后来又有东南大学的余小金副教授、江苏省人民医院的张慧统计师、南京医科大学生物统计学系的仲子航博士相继加入，承担部分章节的初译工作。

本书共 13 章，参与初译的人员具体工作安排如下：柏建岭（第 3、5、6 章）、曾永艺（第 1、11、12、13 章）、李强（第 7、8、10 章）、余小金（第 2 章）、张慧（第 9 章）、仲子航（第 4 章）。初译完成后，柏建岭和曾永艺两位主译花了大量时间进行了多次审阅校对，最终定稿。

感谢清华大学出版社的刘洋和宋亚敏两位编辑以及排版校对人员的辛勤劳动与付出。

本书的翻译工作得到了南京医科大学学术著作出版资助项目的资助。在翻译过程中得到了南京医科大学公共卫生学院和生物统计学系领导及同事的关心帮助，译者表示衷心感谢。

本书的翻译一定还会有不妥之处，衷心希望得到各位专家、老师和同行读者的批评指正。

柏建岭

2021 年春于南京

目　录

2

Stata 简介

学习目标

- **熟悉 Stata 界面**
- **输入和导入数据到 Stata**
- **熟练使用 Stata 命令语言**
- **学习 Stata 常规数据管理命令**
- **使用 Stata 获取基本描述性统计量和图表**
- **使用 Stata 做一些简单的双变量分析**

本章首先介绍 Stata 界面及组成，接着解释如何在 Stata 中直接输入数据以及将外部数据导入 Stata。然后，介绍 Stata 中执行命令的 3 种主要方式，即菜单系统、命令系统和 do 文件编辑器。此外，介绍通过使用 do 文件编辑器执行最常用的数据管理命令（`recode`，`generate` 等）。最后系统介绍了用于描述性统计分析（频数、集中趋势指标等）和双变量统计分析（如相关分析、*t* 检验、方差分析和卡方检验等）的命令。

2.1 Stata 是什么?

Stata 是一款统计软件，包括内容丰富的可持续更新的内置分析方法（线性模型、纵向数据、多重填补等）、数据管理功能（输入 / 输出数据、合并数据集等）[①] 和用户编写命令的 Stata 程序语言开发功能的集合 [②]。在某种程度上，Stata 的内置功能可看作一个商业软件套件的一部分，用户通过 Stata 编写的功能 / 命令 [③] 可看作开源的组成部分。想要使用这些功能，用户需要购买 [④]Stata 并将其安装在电脑里。

2.1.1 Stata 界面

在展现和解释 Stata 如何运行之前，需要熟悉 Stata 的界面包含哪些部分。安装并打开 Stata 后，会出现图 2.1 所示的界面 [⑤]，该界面包括 5 个主要窗口（命令、回顾、结果、变量和属性）和 3 个额外的组成部分。

① 可通过 http://www.stata.com/features/ 进一步了解 Stata 的内置功能。
② 输入 net from http://fmwww.bc.edu/RePEc/bocode/ 获取 Stata 用户编写命令的列表。
③ 从 Stata 内下载这些命令需要用户连接互联网。
④ 登录 http://www.stata.com/order/ 了解如何购买及其价格。
⑤ 这可通过在结果窗口中右键单击并在出现的菜单中选择 Preferences 进行定制。

图 2.1 Stata 界面

命令窗口（command window）用于输入命令[①]让 Stata 执行相应的任务。例如，如果想要计算某个变量（如 price）的均数，可以输入 Stata 命令（`mean price`）来完成这项任务。如果我们还想做 Y 对 X_1、X_2 的回归，可以输入所需的命令（`reg Y X1 X2`）。

回顾窗口（review window）命令窗口运行的任何 Stata 命令，会立刻显示在回顾窗口中。这个列表的优点就是可以在同一会话中随时点击列表中的任意命令再次执行而不必重新输入。

结果窗口（results window）一旦一条命令被写在命令窗口中仅需按回车键就可以执行它。命令执行后的文本结果会显示在结果窗口中。

变量窗口（variables window）打开（创建或输入）的数据集中的变量和其对应标签会显示在变量窗口中。

① 根据统计软件的不同，命令也被称作代码、脚本或语法。

属性窗口（properties window）可以看到变量和数据集的属性。点击属性窗口上的“锁定”图标可以修改其属性。

通过点击和拖拽 Stata 窗口（如结果窗口、回顾窗口等），可以将其移到屏幕上的不同位置。

2.1.2 如何使用 Stata

我们可以用两种主要方法让 Stata 执行任务。第一种方法是使用下拉菜单。如图 2.1 所示，下拉菜单中包括数据管理（Data）、统计（Statistics）和绘图（Graphics）等功能。当点击某个菜单（例如“统计”）时，选择的统计任务（如 `regress`）的对话框就会弹出。然后可以使用对话框里的选项设置分析。这些菜单和对话框提供了 Stata 的绝大多数功能权限。第二个方法就是输入命令[①]。正如之前提到的，命令可以直接输入到命令窗口中（如图 2.1 所示）。在命令窗口中一次可以输入一条命令。

一种更为便捷的可选方法即使用图 2.2 所示的 do 文件编辑器。do 文件编辑器（通过点击）仅仅是一个 Stata 集成的文本编辑器，可在其中编写并且随时执行部分或全部命令。相较于使用命令窗口，使用文件编辑器的主要优点是可以将命令保存起来以便日后调用，从而容易实现研究的可重复性。试想，如果想在一次 Stata 会话中完成许多数据管理操作（如重编码、重命名等）以及分析（如列联表、回归等）任务，还想要记录以便重现相同的结果，你只需将这些命令保存在一个文件里并命名，就可以从 do 文件编辑器中打开这个文件，重新执行其部分或全部命令。

如图 2.2 所示，一个新的 do 文件被打开，显示的正是与图 2.1 所示的命令窗口执行过的相同的命令。在 do 文件中长命令可以用三个斜杠（///）断开以保证该命令能被正常执行。为了执行 do 文件中的命令，可单击工具栏上的指定按钮，或在电脑键盘上同时按“Ctrl+D”组合键。do 文件（扩展名 .do）可以在 do 文件编辑器中直接保存。

由于只需重新运行 do 文件中的所有命令即可得到相应的结果，我们似乎没有必要保存结果。但是，如果需要的话，可以将结果保存在日志文件中（文件扩展名 .smcl）[②]。现假如想要计算一些变量的均值并保存这些结果，如图 2.3 所示，我们首先打开日志文件（文件名为 mean_vars），使用 `sum` 命令来计算均值，然后关闭日志文件。这个日志文件会被保存在工作目录下面。

① 直到版本 8，Stata 还只是纯命令驱动的软件。

② 日志文件也可以 .log 为扩展名保存，这种日志文件可用文本编辑器（如 Notepad）打开和共享。

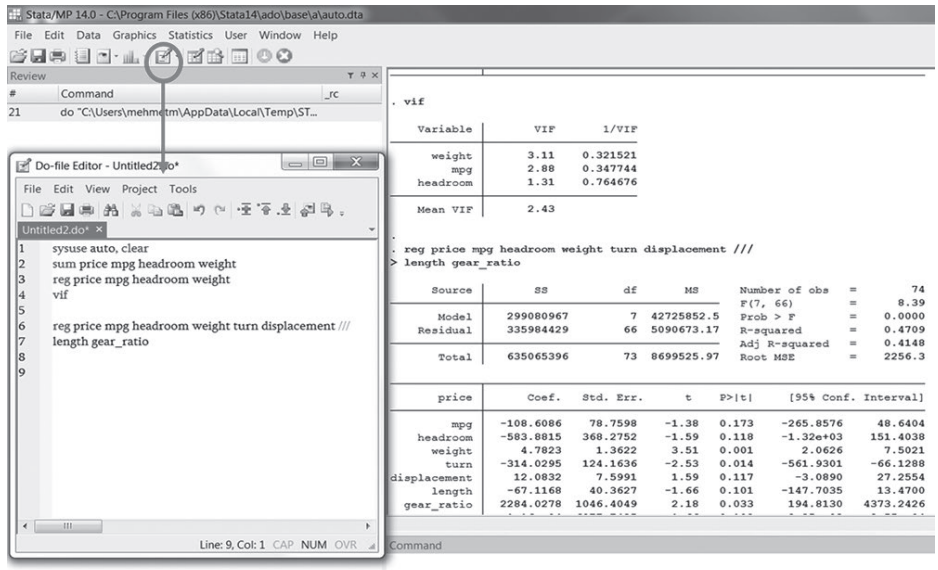

图 2.2　do 文件编辑器

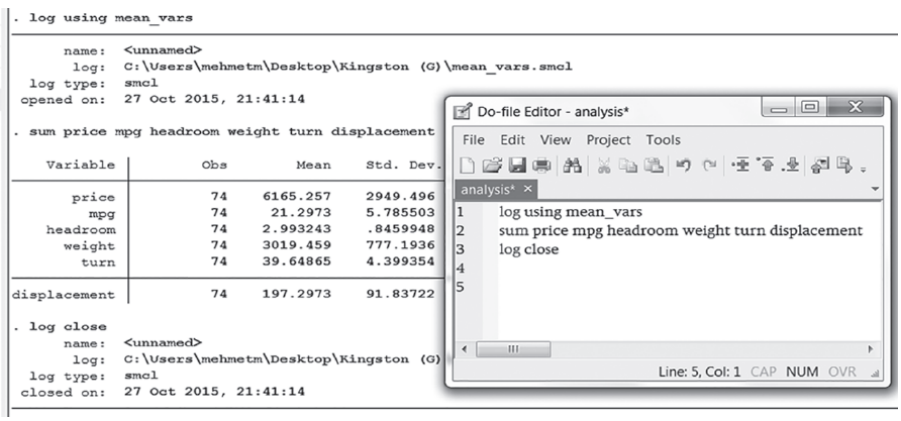

图 2.3　使用 log 文件的例子

工作目录显示在状态栏上（见图 2.1）。Stata 会话的默认工作目录路径如 C:\Users\mydir\Documents，该工作目录很容易被更改。例如，在图 2.1 中，我们通过输入 `cd "C:\Users\mehmetm\Desktop\Kingston(G)"` 更改工作目录。指定 Stata 会话工作目录可节约时间并且提高准确性，这时在 Stata 中不必写出目录的完整路径即可直接保存或搜索任何文档。例如，如果想在上述工作目录中保存一个数据文件，可以直接写 `save filename.dta`。如果没有指定为工作目录，就必须输入 `save "C:\Users\`

mehmetm\Desktop\Kingston(G)\filename.dta"。因此，无论做任何项目（例如文章、论文或者学术演讲等），尽可能为该项目创建一个文件夹，并指定为工作目录，以便随后保存所有 Stata 文件（包括数据、日志、do 文件等）。

Stata 命令（包括内置命令和用户编写命令）通常比默认方式提供更多的选项。Stata 为命令（如之前使用的 log、cd、regress、summarize 等）提供帮助文件，其中包含了对该命令的描述以及可以使用的选项。帮助文件可以通过输入 help 和命令的名字打开，例如，help reg。

帮助文件是学习理解 Stata 命令具体逻辑的钥匙。我们需要事先知道命令的名称以调用相应的帮助文件。但是，如果你不知道具体的命令，仍然可以使用关键词进行查找。例如，如果我们不知道 Stata 中存在 regress 命令，我们就可以输入 search regression，它会在搜索结果中给出 regress 命令的链接。

除了选项之外，帮助文件也显示很多命令可以和 if（条件选择）、in（观测选择）以及 weight（权重设定）语句连用。一个典型的内置 Stata 命令的基本结构如下面的 regress 命令所示：

```
regress depvar [indepvars] [if] [in] [weight] [,options]
```

下划线字符（reg）表示完整的命令 regress 是可以缩写的，具体例子如图 2.1 结果窗口所示。

2.2 数据输入和导入

2.2.1 输入数据

输入数据到 Stata 意味着你可以直接在 Stata 中创建数据集，可通过数据编辑器完成（见图 2.4）。通过单击工具栏上的图标或输入 edit 命令打开数据编辑器。接着在数据编辑器中以行（表示观测）或列（表示变量）的形式输入数据。当你第一次输入一个观测时对应的变量会被自动创建并命名，像 var1、var2、var3 等。可以使用数据编辑器窗口右侧的属性窗口更改变量名（以及标签、类型、格式等）。当输完所有数据之后，可以退出数据编辑器并返回 Stata 主窗口。这里，可以使用菜单，或输入 save filename.dta，或只输入 save filename 来保存数据文件。

在数据编辑器中，可以输入数值型和 / 或字符型变量。字符型变量以文本数据或字符的形式输入。当输入字符型数据，例如，male（男性）或 female（女性），作为性别变量时，这些字符型数据在数据编辑器中会显示为红色。数值型变量输入数值即可，这些值可以按从小到大依次采取字节（byte）、整数（int）、长整数（long）、浮点数（float）和双精度（double）的形式。Stata 默认所有的数值型变量均为浮点数。

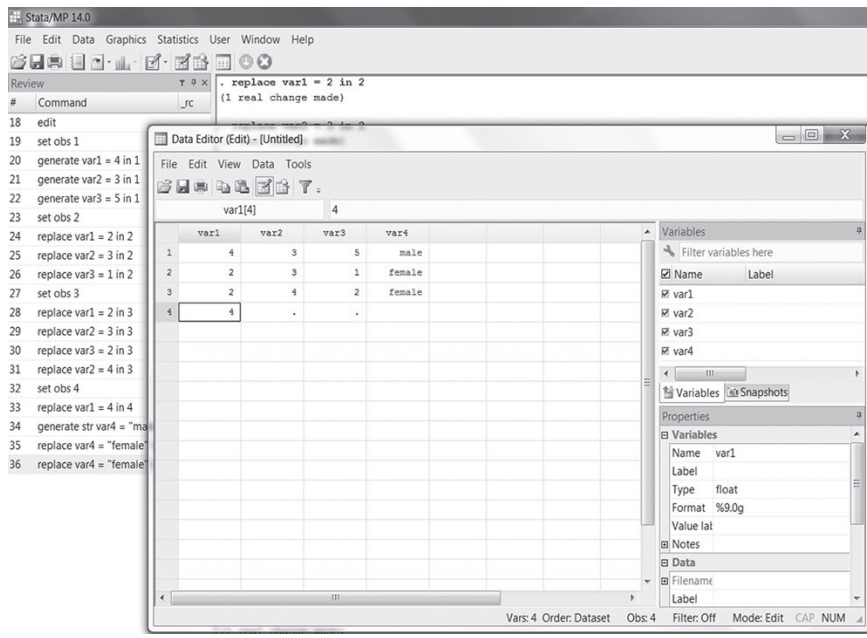

图 2.4　数据编辑器

2.2.2　导入数据

有几种方法用来将多种格式的外部数据文件导入 Stata 中。[①]Stata 可以有效地将一些常用格式转换为自己的格式（.dta）。[②] 尽管通常情况下我们推荐读者使用命令，但这里我们还是建议使用下拉菜单（如图 2.5 所示）将文件导入 Stata[③]。我们要做的就是保存文件到指定目录下（最好是工作目录），从下拉菜单中选择对应的选项，再从弹出的

① 在多数情况下，从其他软件（如 Excel、SPSS 等）传输数据的最简单的方法是直接复制数据矩阵进行粘贴。

② 对于 Stata 无法直接转换的数据格式，可用专业软件 Stat/Transfer 进行转换，更多信息请参见 https://stattransfer.com/。

③ 使用下拉菜单的一个优势是在结果窗口中会显示对应的命令，如果以后需要，你可在命令窗口或 do 文件编辑器中使用该命令。

对话框里做进一步的选择。

除了 Stata 官方的命令和下拉菜单选项，也可能有用户编写的命令便于将某种特定类型数据文件转换到 Stata。[①] 例如，用户编写的命令 `usespss`。在 Stata 中输入 `search usespss` 了解更多关于这个命令的信息。此外，一些软件可能允许自己的数据文件以 Stata 的 .dta 的形式保存，如可以在 SPSS 中将数据文件保存为 Stata 文件，从而可以直接在 Stata 中打开这个数据文件。

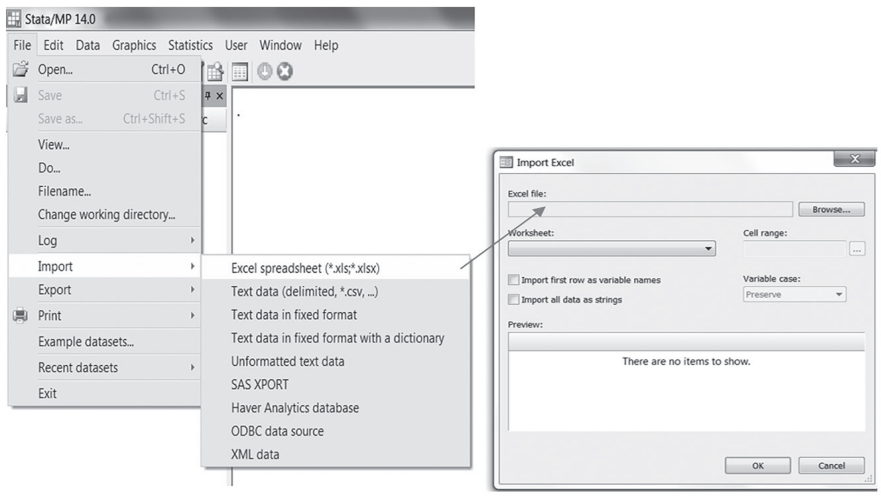

图 2.5　Stata 中用于文件转换的下拉菜单

2.3 数据管理

本节将使用一个真实的数据集 workout1.dta，该数据集收集自挪威某中等城市的一家健康中心的成员样本，包括与锻炼相关的动机和行为的信息以及社会人口统计数据。首先，我们把数据集下载到工作目录中。需要注意的是，本节提到了许多命令，[②] 但要在一章中详细介绍这些命令所有的选项是不现实的。因此，我们建议通过阅读相应的帮助文件来探索各个命令（例如，输入 `help codebook`）。

① 你也可考虑购买 Stat/Transfer 软件，借助该软件可以很容易地完成不同格式数据文件间的相互转换（如将 Excel、SPSS、SAS 等数据转换为 Stata 数据，反之亦然）。

② 我们建议读者结合本章配套的 do 文件学习本节内容，该文档可从本书的配套网站上下载。

2.3.1 打开数据

运行 Stata 之后，我们可以输入如下命令打开数据集 (workout1.dta)：[①]

```
.use "C:\Users\mehmetm\Desktop\Kingston (G)\ workout1.dta"
```

如果这个数据集如我们所建议的那样存储在工作目录下，就可以输入

```
.use workout1.dta
```

或者

```
.use workout1
```

如果在 Stata 会话里已经有一个数据集，应该添加 ,clear 到上述命令中。换句话说，该命令应该写成 use workout1,clear。但是，需要注意的是 clear 选项会移除在内存中的数据集（并不保存）然后再导入 use 命令中指定的数据集，因此建议用户在使用 clear 选项之前保存在内存中的数据集。

2.3.2 检查数据

可能我们想要使用的第一个命令是 describe。这个命令为我们提供了一些关于变量的简单信息，如图 2.6 所示。本例中，我们在该命令后添加变量（v03 v04）。如果我们想要数据集中所有变量的相关信息，只需输入 describe 即可。无须事先打开数据集也可以执行 describe 命令，此时输入 describe using workout1。

```
. describe v03 v04

              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------------
v03             byte    %28.0g     labelv03   Educational level
v04             byte    %17.0g     labelv04   What is your personal annual income
```

图 2.6 describe 命令

另一个需要知道的命令是 codebook。这个命令可以提供一些额外的信息（如频数分布、标签值等），这有助于我们快速熟悉和检查数据集中的所有变量，继而找到数据

① 在文中，我们在执行特定任务的 Stata 命令前加上一个 . 和一个空格，以此表示接下来的文本属于 Stata 命令，但当你在命令窗口或 do 文件中输入这些命令时，无须输入 . 和空格。此外，在 do 文件中可通过 /// 分割命令行，但在命令窗口中无法使用 ///，只能在一行中输入命令的全部文本。

集中可能的错误 / 矛盾（例如，输入错误）（见图 2.7）。

```
. codebook v03
-------------------------------------------------------------------------------
v03
-------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  labelv03

                 range:  [1,4]                        units:  1
         unique values:  4                        missing .:  1/246

            tabulation:  Freq.   Numeric  Label
                            11         1  Primary/Secondary school
                            74         2  High school
                            65         3  University up to 3 years
                            95         4  University more than 3 years
                             1         .
```

图 2.7　codebook 命令

第三个可用于检查数据的命令是 browse。它和上面两个命令一样，提供对数据矩阵中每个观察值的概览。如果我们想看一个或一组变量的实际值，可以输入 browse v01 v02 v03 v04,nolabel。这将打开图 2.8 所示的浏览窗口。nolabel 选项要求直接在浏览窗中显示变量的数值而不是标签值（例如，“2”代替“high school”）。

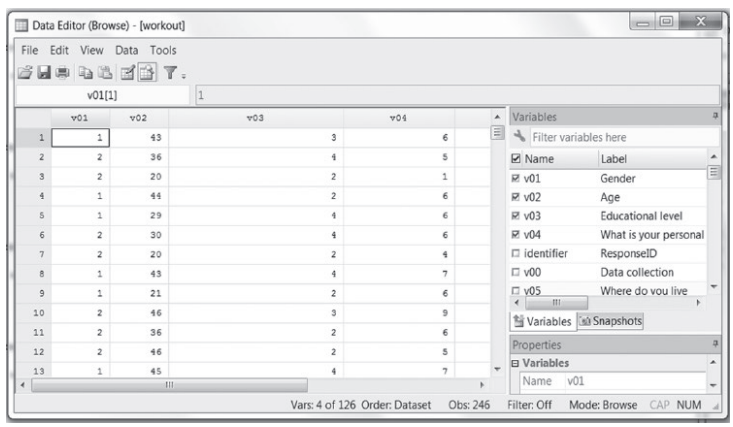

图 2.8　浏览窗口

注意，我们不能在浏览窗口中对这些值做任何更改。如果我们需要这样做，只需简单地用 edit 替换 browse。此外，如果想要这个数据概览出现在结果窗口，就用 list 替换 browse，即 list v01 v02 v03 v04,nolabel。

最后一个对检查数据有用的命令是 misstable sum。这个命令报告每个变量的缺失值的数量。注意，尽管在图 2.9 中所示的例子里我们输入了 v01 v02，但这些变量不会出现在结果窗口中。原因很简单，该命令从结果中删除数据完整的变量，只报告那些有缺失值的变量。在这个例子中，我们看到 v03 只有一个缺失值，而 v04 有 4 个缺失值。正如你观察到的那样，缺失值通常由“.”表示，在 Stata 中被认为是最大值（大于该变

量的最大值）。“.”是默认的缺失值，被称为系统缺失值，有时简称 sysmiss。

```
. misstable sum v01 v02 v03 v04
```

				Obs<.		
Variable	Obs=.	Obs>.	Obs<.	Unique values	Min	Max
v03	1		245	4	1	4
v04	4		242	9	1	9

图 2.9 misstable 命令

除了“.”，Stata 还能区分 26 种其他类型的缺失值。例如，我们可以插入“.a”代表那些拒绝回答问题的人的缺失值，“.b”代表那些认为问题不相关的人的缺失值，“.c”代表那些问题回答得不够清楚的人的缺失值，等等。当你特别想知道造成缺失值的原因时这种分类方法非常有用。

2.3.3 修改变量

recode 命令可以用于修改变量的取值。假如我们已经导入了一个外部数据集，该数据集包含变量 var1，其缺失值输入为 -999。既然我们知道这些在 Stata 中应该用 . 来代替，就可以使用以下命令做出所需的修改：

```
·recode var1 (-999=.)
```

或者

```
·recode var1 -999=.
```

如果你的一个数据集中所有的变量包含 -999，你就可以简单地用如下命令：

```
·recode _all (-999=.)
```

或者

```
·recode * (-999=.)
```

这里 _all 和 * 是指代数据集中所有变量的不同方法。

作为替代，你也可以使用一个专门为这个目的写的命令：

```
·mvdecode _all,mv(-999)
```

或者

```
·mvdecode *,mv(-999)
```

如果你想要达到相反的目的，可以这么写：

```
·mvencode _all,mv(-999)
```

或者

```
·mvencode *,mv(-999)
```

recode 命令可以用来做更多复杂的变换。假设我们有一个人群年收入分布数据（从最低是 1 到最高是 9），而且我们想合并 1、2、3 为第一组，4、5 为第二组以及 7、8、9 为第三组。图 2.10 所示的右边的分布图是很容易通过下面的命令实现的。但是要注意，使用 recode 会覆盖原始变量（v04）。事实上，另一种更好的方法是在重新编码原始变量的基础上生成一个新的变量，具体介绍见下一节。

```
. recode v04 (1/3=1) (4/6=2) (7/9=3)
```

编码前

v04	Freq.	Percent	Cum.
1	23	9.50	9.50
2	9	3.72	13.22
3	6	2.48	15.70
4	16	6.61	22.31
5	38	15.70	38.02
6	64	26.45	64.46
7	37	15.29	79.75
8	14	5.79	85.54
9	35	14.46	100.00
Total	242	100.00	

编码后

v04	Freq.	Percent	Cum.
1	38	15.70	15.70
2	118	48.76	64.46
3	86	35.54	100.00
Total	242	100.00	

图 2.10　recode 命令

recode 命令不能用于字符型变量，此时需要使用 replace 命令，该命令可用于字符型和数值型变量。例如，对于数值型变量 mark 得分高于或低于 90 的个体，要求在字符型变量 exammark 中分别插入文本“very good”和“good”（见图 2.11）。

```
.replace exammark="vary good" if mark>90
.replace exammark="good" if mark<90
```

替换前

	exammark	mark
1.		93
2.		92
3.		83
4.		76

替换后

	exammark	mark
1.	very good	93
2.	very good	92
3.	good	83
4.	good	76

图 2.11 replace 命令

rename 命令可在不同的情况下更改变量名。例如，如果我们要把上述变量名 v03 改成 education，就输入“rename v03 education”。该命令有时用来修改变量名的大小写。比如，我们输入“rename v03,lower”或“rename v03,upper”。如果想要将其应用到整个数据集，只需输入“rename _all,lower”或“rename _all,upper”即可。

2.3.4 生成变量

在学习如何生成新的变量之前，最好先了解下 Stata 中不同的数学运算符。[①] 这些运算符主要分成三类（见表 2.1）。

表 2.1 Stata 中三种类型的数学运算符

算术运算符	逻辑运算符	关系运算符（数字或字符串）
+ 加	& 与	> 大于
- 减	| 或	< 小于
* 乘	! 非	>= 大于等于
/ 除	~ 非	<= 小于等于
^ 乘方		== 等于
- 负号		!= 不等于
+ 字符串连接		~= 不等于

除了这些运算符，Stata 还有众多数学函数，比如 log(x)、sqrt(x) 和 exp(x)，可以通过输入 help functions 进一步了解。

区分用来赋值的单等号（=），如 generate lnprice = ln(price)，和用来比较的双等号（==），如 keep if gender == 2，是很重要的。使用单等号进行比较是新用户常犯的错误之一，所以务必小心！

① 在 Stata 中输入 help operator 了解更多关于这些运算符的信息。

Stata 中有两个方便且常见的命令 gen 和 egen，用来生成新的变量。前者代表 generate，后者代表 extension to generate。首先介绍 gen。这里用一些 gen 的例子来表明这个命令有多方便：

```
·gen age2 = age^2              // 年龄平方
·gen id=_n                     // 观测序号
·gen loghours=log(hours)       // 小时的对数值
·gen pdollar=price/6           // 挪威克朗换算成美元
·gen agecar=2015-year          // 2015年时的车龄
```

gen 和 recode 连用也是相当常见的。之前的 recode 命令会修改原有变量，但这里我们将看到在重新编码原有变量的基础上生成新变量的一个例子。变量 v04 是我们的现有变量，这被用来创建一个新的变量 inccat（收入分类）。

```
·recode v04 (1/3=1) (4/6=2) (7/9=3),gen(inccat)
```

执行该命令之后，数据集中包含了原有变量和新生成的变量。

得到相同结果的另一种方法是连用 gen 和 replace。

```
·gen inccat2 = .
·replace inccat2 = 1 if (v04 <= 3)
·replace inccat2 = 2 if(v04 >= 4)&(v04 <= 6)
·replace inccat2 = 3 if(v04 >= 7)&(v04 <= .)
```

这里需要提示两点：gen inccat2 = . 创建一个空白 / 缺失值向量，然后用实际值替换；第四行添加（v04 < .）[①] 以排除缺失值，因为缺失值是这个变量的最大取值。

egen 和 gen 一样是用来生成新变量的另一命令，但是与 gen 比起来，它更容易生成诸如均值、中位数、全距等的新变量。我们只给出几个 egen 的例子，更多的用法可以在它的帮助文件里找到（help egen）。

在图 2.12 的左图我们用 gen 生成 4 个变量（var1,…,var4）的平均数。看到 gen 排除了 4 个变量中有缺失值的个体。因此，新生成的变量（avg）中前两个观测的取值为缺失。在右图中，我们用 egen 生成一个平均值变量。此时，egen 根据其他 3 个变量的可用值计算前两个观测的平均值，例如第一个观测的平均值为（4+2+1）/3=2.33。

① 我们也可用 (!missing(v04)) 替换 (v04 <= .)。!missing 选项既可用于数值型变量，也可用于字符型变量。

```
. gen avg=(var1+var2+var3+var4)/4
(2 missing values generated)

. list
```

	var1	var2	var3	var4	avg
1.	4	.	2	1	.
2.	3	2	3	.	.
3.	5	3	5	3	4
4.	4	4	4	3	3.75
5.	5	5	5	5	5

```
. egen avg2=rowmean(var1 var2 var3 var4)

. list
```

	var1	var2	var3	var4	avg	avg2
1.	4	.	2	1	.	2.333333
2.	3	2	3	.	.	2.666667
3.	5	3	5	3	4	4
4.	4	4	4	3	3.75	3.75
5.	5	5	5	5	5	5

图 2.12　gen 和 egen 生成平均数

egen 常用于生成一个标准化变量，如：

```
·egen zvar1 = std(var1)
```

它还可以用于生成一系列变量（如 var1,…,var4）的和。正如图 2.13 所示，当计算总和时把缺失值当作零。

```
. egen tot = rowtotal(var1 var2 var3 var4)

. list
```

	var1	var2	var3	var4	tot
1.	4	.	2	1	7
2.	3	2	3	.	8
3.	5	3	5	3	16
4.	4	4	4	.	12
5.	5	5	5	5	20

图 2.13　egen 生成行合计

如果你想找出观测在一个或一系列变量中取值缺失的个数，可以使用如图 2.14 所示的命令。在左图中，我们看到前 5 个观测分别有 2、3、1、3、4 个非缺失值。在右图中，我们用 egen 生成缺失值的个数，是左图结果的反向情况。

```
. egen nonmiss=rownonmiss( var1  var2  var3  var4 )

. list
```

	var1	var2	var3	var4	nonmiss
1.	4	.	.	1	2
2.	3	2	3	.	3
3.	.	3	.	.	1
4.	4	4	4	.	3
5.	5	5	5	5	4

```
. egen rmiss= rowmiss( var1  var2  var3  var4 )

. list
```

	var1	var2	var3	var4	rmiss
1.	4	.	.	1	2
2.	3	2	3	.	1
3.	.	3	.	.	3
4.	4	4	4	.	1
5.	5	5	5	5	0

图 2.14　找出一系列变量中行取值非缺失个数和缺失个数

encode 命令将一个字符型变量转换成数值型变量（图 2.15）。正如下面的例子所示，encode 按照字母顺序（economics 1、political science 2 等）给 var1 的每个文本类别设定标签值。

```
. encode var1, gen(var1_num)

. list,nol
```

	var1	var1_num
1.	psychology	3
2.	economics	1
3.	sociology	4
4.	political science	2

图 2.15　encode 命令

还有另一个命令 decode，其作用正好与 encode 命令相反，是将数值型变量转换成字符型变量。这个命令的工作原理与 encode 相同。举例来说，我们可以用 decode 命令把 var1_num 重新转换为字符型变量（var2）。

```
.decode var1_num,  gen（var2）
```

2.3.5　数据子集

我们可以通过保留或删除变量和观测构造数据集的子集。构造子集有时候是必要或方便的，特别是遇到一个巨大数据集的时候。这通过 Stata 的两条命令 keep 和 drop 很容易实现。如果你只想要包含 4 个变量（v01、v02、v03 和 v04）的数据集而不是整个数据集，只需输入 keep v01 v02 v03 v04 即可。相反，如果想要从数据中移除这 4 个变量，可以输入 drop v01 v02 v03 v04。

为了保留和删除观测，需要在 if 和 in 语句的配合下使用 keep 和 drop。以下为相关示例。

```
·drop in 13                    // 删除观测13
·drop in 10/12                 // 删除观测10,11,12
·drop if missing(var5)         // 删除var5中有缺失值的观测
·drop if missing(var6,var7)    // 删除var6或var7中有缺失值的观测
·keep if !missing(var4)        // 保留var4中没有缺失值的观测
```

2.3.6　标记变量

标记变量需要两步。第一步我们要定义标签值，第二步我们应用这些标签值到

需要定义标签值的变量。为了执行这两步，我们使用命令 label define 和 label values。因为标记变量通常与 gen 命令一起使用，在此继续使用先前 gen 命令中用到的例子。

```
·gen inccat2  = .
·replace inccat2 = 1 if (v04 <= 3)
·replace inccat2 = 2 if(v04 >= 4)&(v04 <= 6)
·replace inccat2 = 3 if(v04 >= 7)&(v04 <= .)
```

如你所见，inccat2 变量有 3 个类别。假设我们想要标记三类分别为“low income”“medium income”和“high income”，首先我们定义一个标签，分别将这些数字分配到这三类中，并且储存在 labinc 下，示例如下：

```
·label define labinc 1 "low income" 2 "medium income" 3 "high income"
```

然后我们选择变量（inccat2）应用这个标签。换句话来说，labnic 被应用到 inccat2：

```
·label values inccat2 labinc
```

在这种情况下，labinc 被专门应用到变量 inccat2，但是有些情况下，你可能想要定义一个标签并且应用到不止一个变量。例如，把一个标签 lablikert 应用到 5 个不同的变量（var1, …, var5），可以采用如下命令：

```
·label define lablikert 1 "disagree" 6 "agree"
·label values var1-var5 lablikert
```

2.4 描述性统计和图

尽管许多读者可能更想用高级的统计方法，而不是比较简单的描述性统计，但后者仍然是任何实质性定量工作的第一步。根据变量测量水平的不同，描述性统计主要有两类：测量水平是无序或有序的频率分布和测量水平是区间或比率的集中趋势和变异度。

2.4.1 频率分布

频率分布显示变量的每个类别所包含的观测的个数。在 Stata 中用 tabulate（或

tab）命令可获得一个变量的频率分布。图2.16展示tab输出的标准频率分布。在左图中，我们看到在排除缺失值的个体后那些属于“No”和“Yes”类别的个体所占的百分比。在右图中，我们添加了miss选项，此时计算百分比就会包含那些缺失值的个体。

```
. tab v07_num

  Are you |
 divorced |      Freq.     Percent        Cum.
----------+-----------------------------------
       No |        225       93.36       93.36
      Yes |         16        6.64      100.00
----------+-----------------------------------
    Total |        241      100.00

. tab v07_num, miss

  Are you |
 divorced |      Freq.     Percent        Cum.
----------+-----------------------------------
       No |        225       91.46       91.46
      Yes |         16        6.50       97.97
        . |          5        2.03      100.00
----------+-----------------------------------
    Total |        246      100.00
```

图 2.16　tabulate 命令

为了得到一个排除缺失值计算的百分比和缺失值的个数的表格，可以用用户编写的fre[①]命令来代替tab。输出结果见图2.17，可将其和图2.16所示的tab输出的两个表格进行比较。

```
. fre v07_num

v07_num ── Are you divorced
------------------------------------------------------------------------------
                      |      Freq.     Percent       Valid        Cum.
----------------------+-------------------------------------------------------
Valid   1 No          |         225       91.46       93.36       93.36
        2 Yes         |          16        6.50        6.64      100.00
        Total         |         241       97.97      100.00
Missing .             |           5        2.03
Total                 |         246      100.00
------------------------------------------------------------------------------
```

图 2.17　用户编写的 fre 命令

图是描述性统计的重要组成。Stata中有两种绘图方法：一种是使用下拉菜单；另一种就是使用命令语言。这里，我们只展示基础的绘图命令，但是这些标准命令可以被进一步扩展。建议读者浏览一下每个绘图命令的帮助文件，并在自己的数据分析中应用命令的扩展功能。

我们可以用直方图和饼图来描述频率分布。命令分别为histogram（或只输入hist）和graph pie。要绘制一个变量v07_num频率分布的直方图（见图2.18）可以通过输入以下命令完成：

① 在Stata中输入ssc install fre来安装该命令。

```
·hist v07_num, discrete percent addlabel xlabel(1/2, valuelabel noticks)
```

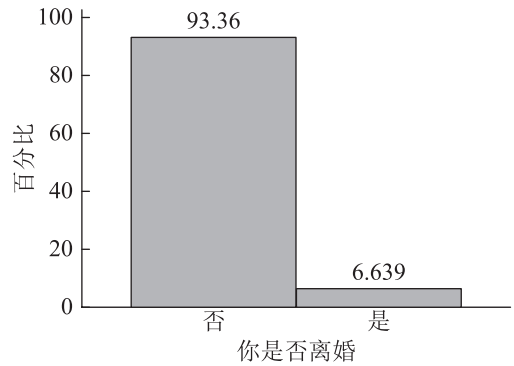

图 2.18 histogram 命令

可以用以下命令绘制如图 2.19 所示的饼图。

```
·graph pie, over(v07_num) plabel(_all percent)
```

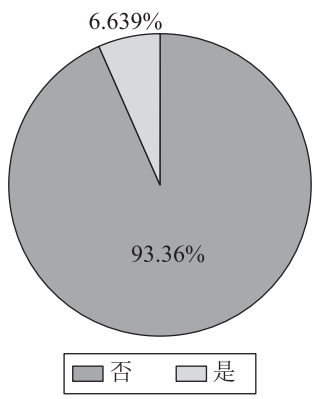

图 2.19 graph pie 命令

2.4.2 汇总统计

当我们研究的变量测量的是区间或比率数据时，采用集中趋势和变异指标（汇总统计）要比频率分布更恰当。最常见的集中趋势指标是算术均值，而典型的变异指标是标准差和全距。这些基本的汇总统计可以用 summarize（或 sum）命令来完成。在图 2.20 中以 Stata 自带的数据文件 auto.dta 中的 price 变量为例说明 sum 命令的用法。

```
. sysuse auto,clear
(1978 Automobile Data)

. sum price
```

Variable	Obs	Mean	Std. Dev.	Min	Max
price	74	6165.257	2949.496	3291	15906

图 2.20　sum 命令

其他集中趋势（如中位数）以及变异指标（方差）也可以在 sum 命令后添加 detail（或 d）选项来完成（见图 2.21）。

```
. sum price, d
```

Price

Percentiles		Smallest		
1%	3291	3291		
5%	3748	3299		
10%	3895	3667	Obs	74
25%	4195	3748	Sum of Wgt.	74
50%	5006.5		Mean	6165.257
		Largest	Std. Dev.	2949.496
75%	6342	13466		
90%	11385	13594	Variance	8699526
95%	13466	14500	Skewness	1.653434
99%	15906	15906	Kurtosis	4.819188

图 2.21　详细的 sum 命令（或 sum，d）

如果我们想要获得均值的标准误和置信区间，可以用命令 mean 来完成（见图 2.22）。

```
. mean price

Mean estimation                 Number of obs    =         74
```

	Mean	Std. Err.	[95% Conf.	Interval]
price	6165.2568	342.8719	5481.9140	6848.5995

图 2.22　mean 命令

尽管以上两个命令可以为多个变量和以另一个变量为条件生成汇总统计，但是要达到相同的目的，更有效的命令是 tabstat。我们还可以用 tabstat 命令自定义要计算并显示的汇总统计[①]。图 2.23 所示为国内外汽车的 price（价格）、weight（重量）、和 length（长度）的汇总统计。如果想水平显示汇总统计，可以在下面命令的末尾输入 col(stats) 选项。此外，如果想要删除合计的（total）汇总统计概述，可以在命令末尾添加 nototal。

① 输入 help tabstat 全面了解汇总统计量及其对应的 Stata 代码。

```
. tabstat price weight length, stats(mean sd range count) by(foreign)

Summary statistics: mean, sd, range, N
  by categories of: foreign (Car type)
```

foreign	price	weight	length
Domestic	6072.423	3317.115	196.1346
	3097.104	695.3637	20.04605
	12615	3040	86
	52	52	52
Foreign	6384.682	2315.909	168.5455
	2621.915	433.0035	13.68255
	9242	1660	51
	22	22	22
Total	6165.257	3019.459	187.9324
	2949.496	777.1936	22.26634
	12615	3080	91
	74	74	74

图 2.23 tabstat 命令

tabstat 制作一维汇总统计表，而另一个命令 tab，与选项 sum() 结合，却可以计算和展示多维汇总统计表。图 2.24 中我们举例展示基于 foreign 和 rep78 两个分类变量对变量 mpg 的二维汇总统计。

```
. tab rep78 foreign, sum(mpg)

        Means, Standard Deviations and Frequencies of Mileage (mpg)
```

Repair Record 1978	Car type Domestic	Foreign	Total
1	21	.	21
	4.2426407	.	4.2426407
	2	0	2
2	19.125	.	19.125
	3.7583241	.	3.7583241
	8	0	8
3	19	23.333333	19.433333
	4.0856221	2.5166115	4.1413252
	27	3	30
4	18.444444	24.888889	21.666667
	4.5856055	2.7131368	4.9348699
	9	9	18
5	32	26.333333	27.363636
	2.8284271	9.367497	8.7323849
	2	9	11
Total	19.541667	25.285714	21.289855
	4.7533116	6.3098562	5.8664085
	48	21	69

图 2.24 tab, sum 命令

对于汇总统计的图示，我们可以使用直方图和箱线图，命令分别是 histogram（或 hist）和 graph box。我们用 Stata 自带的 nlsw88 数据集来解释说明这两个命令。可以输入 sysuse nlsw88 或 sysuse nlsw88,clear（已打开其他数据集并需切换）打开这个数据集。如图 2.25 所示，首先制作 wage（度量人们的时薪）的直方图，输入以下命令：

```
·hist wage,frequency
```

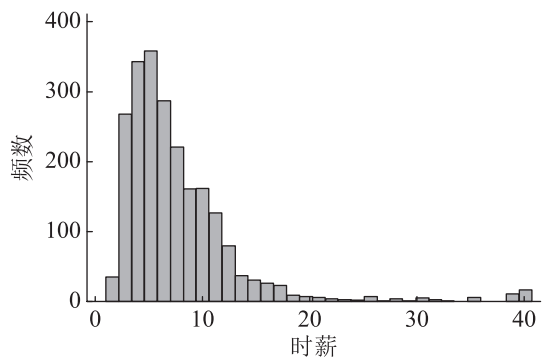

图 2.25 histogram,frequency 命令

我们接着制作变量 wage 的箱线图。箱线图通常用于比较不同子样本同一变量的分布情况。绘制变量 race 两个分类的变量 wage 的箱线图（见图 2.26），可以用以下命令：

```
·graph box wage,by(race)
```

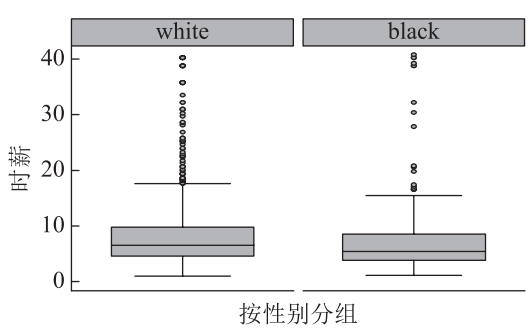

图 2.26 graph box 命令

2.4.3 纵向合并数据

纵向合并数据是指基于观测来合并两个数据集，即将一个数据集的观测添加到另一个

数据集中来完成（观测数量 N 增加），在 Stata 中可使用如下命令轻松实现，结果见图 2.27。

```
·append using dataset1 dataset2,gen(dataset3)
·save dataset3
```

Dataset1

id	data	var1	var2
1	1	3	2
2	1	4	3
3	1	5	1

Dataset2

id	data	var1	var2
4	2	3	1
5	2	5	3
6	2	5	4

Dataset3 (combined)

dataset3	id	var1	var2
1	1	3	2
1	2	4	3
1	3	5	1
2	4	3	1
2	5	5	3
2	6	5	4

图 2.27　纵向合并数据

图 2.27 中我们合并数据集 dataset1 和 dataset2，合并后的 dataset3 可以显示观测来自哪个原始数据集。分析时可能会用到这个信息 / 变量（例如，比较两个不同地方收集的数据）。

2.4.4　横向合并数据

横向合并数据是指基于变量合并两个数据集，即将一个数据集的变量添加到另一个数据集中（变量数量增加）。下面的命令是首先打开 data14，然后添加 data15 中的变量到 data14，最后保存结果数据集为 data1415（见图 2.28）：

```
·use data14,clear
·merge 1:1 id using data15
·save data1415
```

Data14

id	v1_14	v2_14
1	3	5
2	4	5
3	2	3
4	1	2
5	1	2

Data15

id	v1_15	v2_15
1	4	5
2	5	5
3	3	4
4	2	3
5	2	3

Data1415 (combined)

id	v1_14	v2_14	v1_15	v2_15
1	3	5	4	5
2	4	5	5	5
3	2	3	3	4
4	1	2	2	3
5	1	2	2	3

图 2.28　横向合并数据

这里有几点需要明确：第一，两个数据集中的 id 数量完全相同，但是变量名不同。这是因为我们在两个不同的年份（2014 年和 2015 年）测量相同的观测样本在两个相同变量上的取值情况。第二，1∶1 横向合并是指来自 data14 的一条观测与来自 data15 的一条观测合并。第三，Stata 生成一个变量 _merge。_merge 取值为 1 是指观测来自 data14，取值为 2 是指观测来自 data15，取值为 3 则是观测来自两个数据集（data14 和 data15），这意味着成功完成了合并。

2.4.5 数据变型

我们有时可能想要改变数据结构以便进行某种类型的分析（例如，多层次模型）。我们可以将宽型数据转成长型数据或将长型数据转成宽型数据格式。宽型数据结构中，每个观测只有一行，而长型数据中每个观测不只一行（见图 2.29）。最常见的格式转换是由宽型数据到长型数据。[①] 这通过以下命令完成：

```
·reshape long v1_ v2_, i(id) j(year)
```

宽型

id	v1_14	v2_14	v1_15	v2_15
1	3	5	4	5
2	4	5	5	5
3	2	3	3	4
4	1	2	2	3
5	1	2	2	3

长型

id	year	v1_	v2_
1	14	3	5
1	15	4	5
2	14	4	5
2	15	5	5
3	14	2	3
3	15	3	4
4	14	1	2
4	15	2	3
5	14	1	2
5	15	2	3

图 2.29　宽型和长型数据格式

reshape 命令的用法是我们首先指定想要转成的数据格式（long）。然后添加代表宽型变量（v1_14、v2_14、v1_15 和 v2_15）的共同前缀 v1_ 和 v2_。在逗号，之后我们把数据集中观测的唯一标识符 id 填入 i() 中。最后，在 j() 中填入 year 来代表宽型变量的数字后缀（14 和 15）。

① 在 Stata 中数据由长型变宽型同样容易实现，输入 help reshape 了解更多信息。

2.5 双变量统计推断

本节介绍如何用 Stata 完成一些最常用的基本或双变量统计推断，包括相关分析、*t* 检验、方差分析和卡方检验。我们的目的在于介绍如何在 Stata 中得到这些统计量并简要解释输出结果。因此我们跳过这些统计程序的技术 / 理论方面的处理，这里我们假设读者对基础统计学课程中的这些知识已经有所了解。本节以 Stata 内置的 `nlsw88` 数据集为例，输入如下命令加载该数据集到 Stata 中：

```
·sysuse nlsw88, clear
```

2.5.1 相关

图 2.30 所示为探索时薪（`wage`）和工作经验（`ttl_exp`）两者间关系的相关分析。我们发现工资与工作经验间有中等[①]正相关，并是显著的（r=0.27, N=2246, p<0.05）。

```
. pwcorr wage ttl_exp, star(0.05) obs

             |     wage  ttl_exp
-------------+------------------
        wage |   1.0000
             |     2246
             |
     ttl_exp |   0.2655*  1.0000
             |     2246     2246
```

图 2.30 `pwcorr` 命令

我们也可以输入 `corr wage ttl_exp`。但是，我们推荐 `pwcorr`，因为它对于缺失值采用成对删除（pairwise deletion），而 `corr` 是成列删除（listwise deletion）。后者对一对以上的变量进行相关分析时可能会导致信息丢失。对于完整的数据，两个命令的结果是一样的。

2.5.2 独立 *t* 检验

独立 *t* 检验用来检验两独立样本（组）的某个变量的（总体）均值 / 平均数是否存在差异。我们在图 2.31 中进行这个检验以弄清有大学学位的人与没有大学学位的人的平均时薪是否有显著差异。我们进行双侧检验（没有方向），Stata 输出基于原假设

① 0.1、0.3 和 0.5 的 *r* 值分别是弱相关、中等相关和强相关。

H_0：diff=0 的结果。这个简单来说就是均值之差为 0（均值相等）。正如图 2.31 所示，得到的 p 值是 0.0000，原假设应该被拒绝。据此，我们可以推断有无大学学历的人的时薪均数之差为 −3.62，双侧检验统计上显著 [t(2244)=−13, p<0.001]。

```
. ttest wage, by(collgrad)

Two-sample t test with equal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
not coll |   1,714    6.910561    .1276104    5.283132    6.660273    7.16085
 college |     532    10.52606    .2742596    6.325833    9.987296   11.06483
---------+--------------------------------------------------------------------
combined |   2,246    7.766949    .1214451    5.755523    7.528793   8.005105
---------+--------------------------------------------------------------------
    diff |           -3.615502    .2753268               -4.155424  -3.07558
------------------------------------------------------------------------------
    diff = mean(not coll) - mean(college)                         t = -13.1317
Ho: diff = 0                                     degrees of freedom =     2244

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

图 2.31　ttest 命令

图 2.31 所示的 t 检验结果是基于等方差的假设。但是，如果我们想要运行方差不等的 t 检验，我们可以在上述命令的结尾加上 unequal。另外，检验两样本中 wage 的方差是否相等，输入 sdtest wage, by(collgrad) 即可。

2.5.3　方差分析（ANOVA）

方差分析用于检验两个及以上独立样本均值间的差异。因此方差分析可以看成是独立 t 检验的扩展。在下面的例子中，我们想弄清白人、黑人和其他人种的时薪（总体）均值是否存在统计学差异。首先，我们可以用图 2.32 中的 tab 命令得到这三组的样本均值。

```
. tab race, sum(wage)

           |      Summary of hourly wage
      race |       Mean   Std. Dev.       Freq.
-----------+------------------------------------
     white |  8.0829994   5.9550691       1,637
     black |  6.8445578   5.0761866         583
     other |  8.5507813   5.2094301          26
-----------+------------------------------------
     Total |   7.766949   5.7555229       2,246
```

图 2.32　tab, sum 命令示例

根据输出结果，这些均值间存在一些差异。为了弄清这些差异是否统计上显著，我们进行如下的方差分析。这里，原假设是 H_0 ：$\mu_1=\mu_2=\mu_3$，也就是这三组的（总体）

均值是相等的。正如图 2.33 所示，该原假设应该被拒绝，因为总的 F 检验是显著的：$F(2,2243)=10.28$, $p<0.001$。

```
. anova wage race
```

```
                   Number of obs =      2,246    R-squared     =  0.0091
                   Root MSE      =    5.73188    Adj R-squared =  0.0082
```

Source	Partial SS	df	MS	F	Prob>F
Model	675.51028	2	337.75514	10.28	0.0000
race	675.51028	2	337.75514	10.28	0.0000
Residual	73692.457	2,243	32.854417		
Total	74367.967	2,245	33.126043		

图 2.33 anova 命令

我们据此接受备择假设，认为至少有一个组间均值差异是显著的。为了找出究竟哪两组的时薪均值间存在差异，我们接着进行两两比较（见图 2.34）。

```
. pwcompare race, pveffects asobserved

Pairwise comparisons of marginal linear predictions

Margins      : asobserved
```

	Contrast	Std. Err.	t	Unadjusted P>|t|
race				
black vs white	-1.238	0.276	-4.4798	0.000
other vs white	0.468	1.133	0.4129	0.680
other vs black	1.706	1.149	1.4851	0.138

图 2.34 pairwise 命令

2.5.4 卡方检验

卡方检验（χ^2 检验）用于检验两个分类变量之间的关系。图 2.35 所示的例子中，我们用 χ^2 检验弄清变量 union（0=non_union,1=union）是否与变量 collgrad（0=no college degree, 1=college degree）有关。我们想要检验该关系是基于我们假设有无大学学位会影响加入工会的可能性这一理论推理。更常见的是，在这种情况下我们考虑把 union 作为因变量，把 collgrad 作为自变量。[①]

① 另一个用于这个例子的命令为 prtest union, by(collgrad)。

从图 2.35 所示的结果中我们可以发现，union 和 collgrad 间是有统计学意义的：$\chi^2(1, N=1878)=17.97, p<0.001$。更确切地说，没有大学学历的人大约 22% 加入工会，有大学学历的人有大约 32% 加入工会，两者间存在约 10% 的差异。

```
·tab union collgrad, col chi2
```

```
| Key
|
|   frequency
|   column percentage
```

union worker	college graduate: not colle	college g	Total
non-union	1,101	316	1,417
	77.86	68.10	75.45
union	313	148	461
	22.14	31.90	24.55
Total	1,414	464	1,878
	100.00	100.00	100.00

```
Pearson chi2(1) =  17.9705   Pr = 0.000
```

图 2.35 tab, chi2 命令

2.6 总结

本章主要从 3 个方面对 Stata 做了基本介绍：Stata 界面、描述性统计和图表以及利用 Stata 完成一些常见的双变量统计分析。通过这 3 个方面的展示，我们试图给读者提供一个良好的基础，帮助读者理解 Stata 的工作原理和利用内、外部的资源更深入地自学 Stata。借助 Stata 获取这些资源的最简单的方法就是通过 help 和 search 命令。利用本章涉及的命令的帮助文件可以帮助你成为资深的 Stata 用户。尽管 search 命令是用来找到可获得的命令，它也有助于找到其他与 Stata 相关的资源（如在线教程、例子）。此外，我们强烈建议读者阅读 Stata 的入门手册（输入 help gsw 可调出该手册）并学习由 StataCorp 提供的视频教程（http://www.stata.com/links/video-tutorials/）。

关键术语

- **命令（Command）：** 由用户输入给 Stata 使其执行某些任务的指令。
- **Do 文件（Do-file）：** 包含一系列命令的文件。

- **日志文件（Log-file）：**包含 Stata 输出的文件。
- **工作目录（Working directory）：**Stata 当前会话的工作目录。
- **帮助文件（Help file）：**对给定 Stata 命令解释如何使用和列出更多特征。
- **数据编辑器（Data editor）：**有行（观测）和列（变量）的窗口。
- **sysuse：**打开 Stata 中内置的数据文件的命令。
- **webuse：**打开网络上的 Stata 数据文件的命令。
- **纵向合并（Appending）：**行方式扩展数据集的操作。
- **横向合并（Merging）：**列方式扩展数据集的操作。
- **变型（Reshaping）：**改变数据结构格式的操作（从宽型到长型或反之）。

问题

1. 你认为学习 Stata 的最好方法是什么？
2. 从任一数据集中选取 3 个连续变量并且运用“汇总统计”一节中所有的命令。
3. 从任一数据集中选取 3 个分类变量并且运用“频率分布”一节中所有的命令。
4. 使用任一数据集进行基本统计分析：相关、*t* 检验、方差分析、卡方（χ^2）检验。
5. 利用帮助文件进一步探索 anova 命令。

延伸阅读

Acock, A.C.(2014) *A Gentle Introduction to Stata*. College Station, TX: Stata Press.

这是一本用户容易掌握和使用的 Stata 书籍，涉及数据管理、图形、描述性统计、双变量分析以及一些在一般社会科学领域中常用的高级主题（如 logistic 回归、结构方程模型）。

StataCorp(2014) *Stata Manual:Relaese* 14. College Station,TX: Stata Press.

这是官方的 Stata 手册，里面包含了对 Stata 命令最全面和详细的概述和解释。