

УДК 519.23:[57:61]

ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ДЛЯ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ СТАТА: ВВЕДЕНИЕ


© 2014 г. Т. Н. Унгуряну, *А. М. Гржибовский

Северный государственный медицинский университет, г. Архангельск
*Норвежский институт общественного здравоохранения, г. Осло,
Норвегия

Настоящей статьей журнал «Экология человека» открывает серию публикаций по основам биostatистики с использованием пакета программ для статистической обработки данных STATA (Stata Corp, TX, USA). Более ранний выпуск практикума, появившийся на страницах журнала в 2008 году, был адаптирован под пакет программ SPSS (SPSS Inc. Chicago, IL, USA), однако существенное подорожание этого продукта привело к тому, что многие организации переходят с SPSS на STATA по причине более низкой стоимости лицензии последнего. Следует отметить, что в зарубежных университетах и научно-исследовательских институтах пакет STATA не менее популярен, чем SPSS. Кроме того, он позволяет применять более сложные методы статистической обработки, недоступные в SPSS. Однако на русском языке руководств по использованию пакета STATA в биомедицинских исследованиях авторами не обнаружено, что и обусловило необходимость написания серии статей по обработке данных с помощью STATA.

Ключевые слова: STATA, биostatистика, исследования

В 2008 году в журнале «Экология человека» был опубликован практикум, адаптированный под пакет программ SPSS (SPSS Inc. Chicago, IL, USA) [1, 2]. Настоящей статьей журнал открывает серию публикаций по основам биostatистики с использованием пакета программ для статистической обработки данных STATA (Stata Corp, TX, USA). STATA – это профессиональный статистический программный пакет для решения статистических задач в самых разных прикладных областях: экономике, медицине, биологии, социологии. STATA позволяет реализовывать большой спектр статистических методов [3–8]. Кроме того, программный пакет дает возможность программировать всю последовательность команд, начиная от загрузки данных в память и вплоть до всех деталей анализа. Официальный сайт разработчика программы – <http://www.stata.com>. Программа хорошо документирована, издается специальный журнал для пользователей системы. Журнал «Stata» (SJ) – это ежеквартальный рецензируемый журнал, являющийся основным методом распространения методик, разработанных для пользователя. Настоящая статья знакомит читателей с общими принципами работы в STATA, создания и сохранения файлов, правилами ввода и трансформации данных, а также импорта данных из других программ для обработки данных.

Меню STATA. Для того чтобы открыть программу STATA, необходимо нажать на ярлык . Открывается диалоговое окно, где представлена информация о программе (версия, разработчик, серийный номер) и на верхней панели – главное меню (рис. 1).

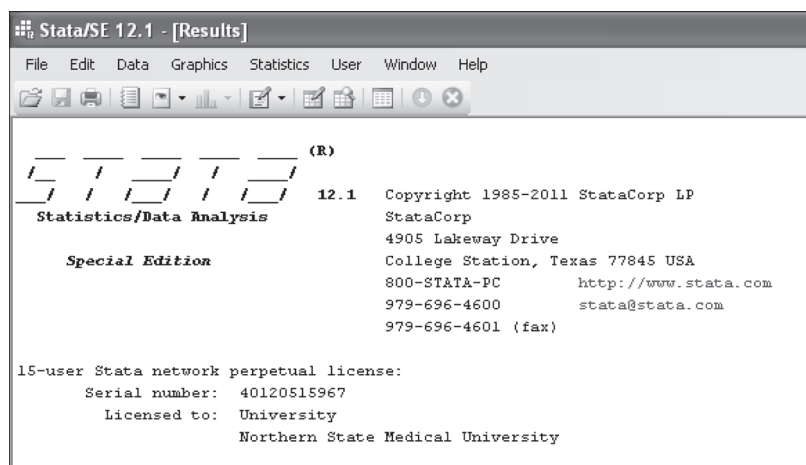


Рис. 1. Главное диалоговое окно и меню при запуске программы STATA

Меню STATA 12.1 for Windows состоит из следующих категорий:

File Edit Data Graphics Statistics User Window Help

- File (файл) позволяет открывать файлы с данными (Open), сохранять данные (Save, Save as), посмотреть файлы отчетов (View), создать собственную программу-процедуру пользователя do-file (Do), импортировать (Import) и экспортировать (Export) данные из/в других форматов, печатать полученные результаты (Print), выход из программы (Exit).
- Edit (редактировать) позволяет копировать как отдельные элементы вывода, так и таблицы, графики; дает возможность задать предпочтения (Preferences) пользователя: выбор цветов, шрифтов и т. д. для разных элементов вывода и визуального представления программы.
- Data (данные) позволяет редактировать данные (Data Editor (Edit)), создавать и изменять переменные (Create or change data).
- Graphics (графики) позволяет строить множество различных видов графических изображений.
- Statistics (статистика) содержит большой набор способов описания данных и видов статистического анализа данных.
- Window (окно) дает возможность увидеть информацию о проделанной работе (команды, результаты их выполнения, запрашиваемая информация из справки) (Results), представляет список проделанных команд вне зависимости от успешности выполнения (Review), список переменных (Variables), свойства переменных (Properties) и др.
- Help (помощь) включает справочник команд в PDF-формате, поиск по ключевым словам (Search), ссылку на веб-сайт программы (STATA website) и др.

Общие принципы работы с программой. Большинство опций может быть выполнено через меню программы или напрямую через различные команды. Использование команд и написание командного синтаксиса дает возможность воспроизводить полученные результаты. Воспроизводимость также означает, что можно легко сделать анализ данной модели в других условиях. Даже если выполнено большое количество шагов с момента начала анализа основной модели, очень легко вернуться назад и создать новую версию анализа, если вся работа сохранена как серия программных шагов.

Основной синтаксис всех команд STATA следует шаблону. Не все элементы данного шаблона используются всеми командами, а некоторые элементы имеют силу только для определенных команд. Но если элемент существует, он будет появляться в одном и том же месте, следуя

одним и тем же грамматическим правилам. В STATA все команды должны вводиться строчными буквами.

Создание базы данных. Для создания базы данных в формате Stata необходимо после запуска программы выбрать в меню Data > Data Editor > Data Editor (Edit). Появится диалоговое окно Data Editor [Edit] – [Untitled] (рис. 2). В появившемся макет базы можно непосредственно вносить данные или скопировать значения переменных, например, из базы данных, созданной в формате Excel.

Названия переменных. Названия переменных даются только на английском языке или обозначаются латинскими буквами. Для того чтобы внести название переменной, можно дважды щелкнуть по заголовку столбца (например, var 1) и затем в появившемся справа диалоговом окне Properties написать название переменной (Name). В заголовках столбцов отражаются полностью названия только из 8 или меньшего количества букв, допускается впечатать название из 32 букв, но оно не будет полностью отражено в заголовке столбца. Поэтому можно создать метку (Label) для переменной, которая будет содержать ее краткое описание. Например, переменная «Место жительства» может быть названа как «Place», а метка для нее обозначена как «Place of residence».

Тип переменных. Тип (Type) переменных может быть обозначен как качественный (string, str) или количественный (byte). В Data Editor качественные переменные представляются красным цветом, а количественные – черным или синим. STATA воспринимает непрерывные значения как количественные только в том случае, если в качестве разделителя используется точка (.), например, масса тела 72.3 кг. Если используется запятая, то данное значение будет восприниматься как качественное. Поэтому количественные непрерывные переменные должны быть представлены с использованием точки. Если программа ошибочно воспринимает количественные переменные как качественные, то для перевода качественного формата (string, str) в количественный (byte) нужно в меню Data выбрать Create or change data > Other variable-transformation commands > Convert variables from string to numeric. Для перевода количественного формата в качественный следует выбрать Create or change data > Other variable-transformation commands > Convert variables from numeric to string.

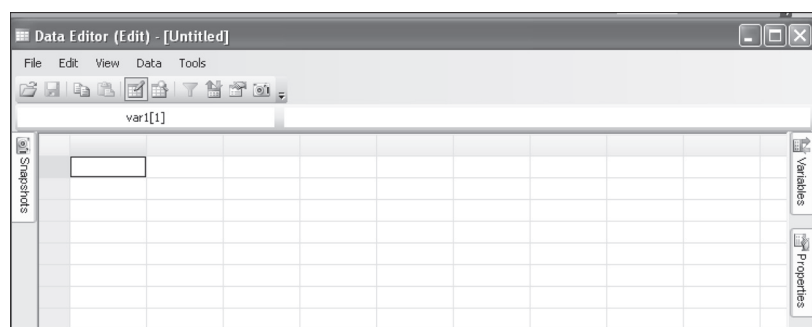


Рис. 2. Окно редактирования данных

Расшифровка кодировки качественных переменных. Для расшифровки значений качественных переменных необходимо открыть данные в Data Editor, выделить столбец с качественной переменной, например Sex (пол), затем нажать правую кнопку мыши и выбрать Value Labels > Manage Value Labels. В появившемся диалоговом окне нажать Create Label. Во вновь открывшемся диалоговом окне «Create Label» в строку «Label name» ввести название переменной (Sex), в поле справа «Value» ввести обозначение первой категории, например 1, а в поле ниже «Label» расшифровать ее, например Male (мужской), затем следует нажать Add. Потом в поле справа «Value» ввести обозначение второй категории, например 2, а в поле ниже «Label» расшифровать ее, например Female (женский), и также нажать Add. Полный список кодов и их расшифровка отображаются в окне слева (рис. 3).

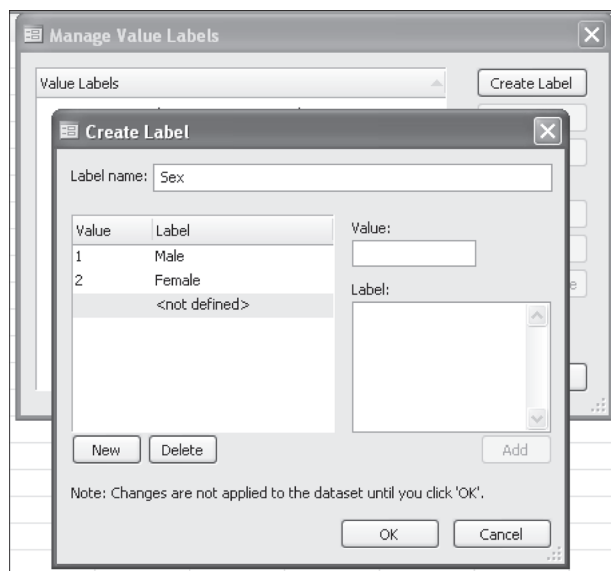


Рис. 3. Диалоговое окно для создания меток качественных переменных

Создание и замена переменных. Рассмотрим пример создания новой переменной «IBM» (body mass index, индекс массы тела) с использованием существующих в базе данных переменных Weight (масса) и Height (рост). Для создания новой переменной, получаемой путем арифметического вычисления, необходимо открыть Data > Create or change data > Create new variable. В появившемся окне «generate – Create a new variable» в поле «Variable type» указать тип создаваемой переменной, например byte, а в поле «Variable name» вписать название создаваемой переменной, например IBM. Затем нажать на кнопку Create и в появившемся окне «Expression builder» в поле Category выбрать Variables. Выбирая нужные переменные из списка (дважды кликнув на переменную) и арифметические действия, отображаемые в окне справа, вводим формулу расчета новой переменной, например Weight/(Height/100)² (рис. 4). После нажатия на «OK» в базе данных появится новый столбец с переменной IBM. Создать новую

переменную можно с помощью команды «generate». Для этого в нижней части главного диалогового окна STATA в поле Commands необходимо ввести: **generate byte IBM = Weight / (Height / 100)².**

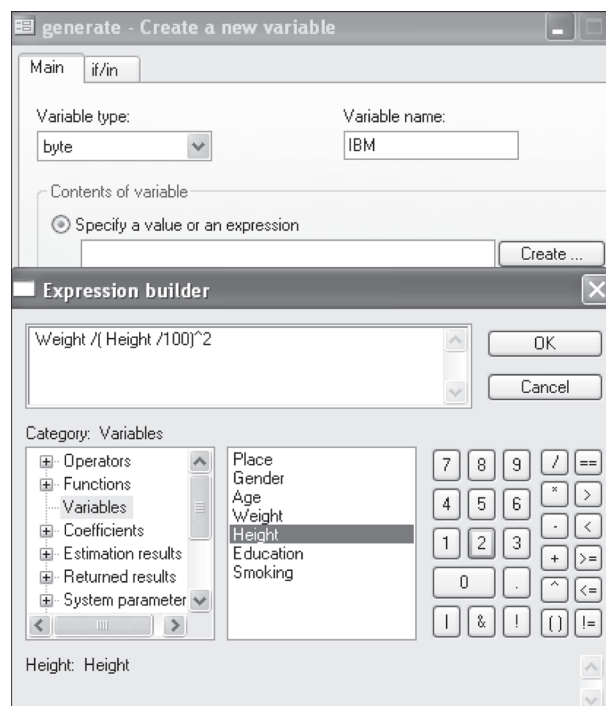


Рис. 4. Диалоговое окно для создания новой переменной путем вычисления

Создание новых номинальных и порядковых переменных. Создание новых номинальных или порядковых переменных осуществляется по следующим правилам:

Правило	Пример	Расшифровка
(# = #)	(3 = 1)	3 перекодировать в 1
(# # = #)	(2 ... = 9)	2 и ... перекодировать в 9
(# / # = #)	(1/5 = 4)	от 1 до 5 перекодировать в 4
(nonmissing = #)	(nonmiss = 8)	все другие не пропущенные в 8
(missing = #)	(miss = 9)	все другие пропущенные в 9

В нашем примере количественную переменную IBM (индекс массы тела) переведем в порядковую, обозначив значения IBM до 20 единиц как низкая масса, 20–25 единиц – нормальная масса и выше 25 единиц – высокая масса и присвоим данным группам коды 1, 2 и 3 соответственно. Для того чтобы выполнить трансформацию данных, необходимо открыть Data > Create or change data > Other variable-transformation commands > Recode categorical variable. В появившемся диалоговом окне «recode – Recode categorical variable» на вкладке «Main» выбрать переменную, подлежащую трансформации (в примере – IBM). Далее в Required и Optional следует указать, для каких групп значений какие коды присвоены, в рассматриваемом примере:

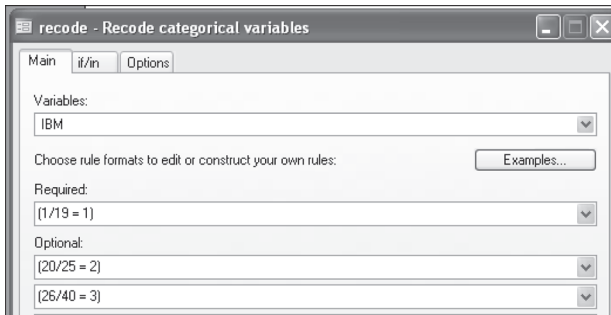


Рис. 5. Диалоговое окно для перекодировки переменных

(1/19 = 1), (20/25 = 2), (26/40 = 3) (рис. 5). На вкладке «Options» выбрать Generate new variable и указать название новой переменной, например IBM1. После нажатия «ОК» в базе данных появится новая порядковая переменная IBM1.

Создать новую номинальную или порядковую переменную можно с помощью команды «recode». Для этого в нижней части главного диалогового окна STATA в поле Commands необходимо ввести: **recode IBM (1/19 = 1) (20/25 = 2) (26/40 = 3), generate(IBM1).**

Удаление данных из базы. Для удаления ненужных столбцов и строк в базе данных необходимо их выделить, затем нажать правую кнопку мыши и выбрать Drop Selected Data.

Сохранение и открытие файла. Для того чтобы сохранить данные и дать название файлу, необходимо в меню выбрать File > Save as. STATA автоматически сохраняет файл с расширением *.dta (Stata Data *.dta). После сохранения и закрытия файла в следующий раз он может быть открыт через главное меню File > Open.

В STATA можно импортировать данные, сохраненные в формате Excel. Для этого следует в меню выбрать File > Import > Excel spreadsheet (*.xls, *.xlsx). База данных, созданная в программе SPSS и сохраненная с расширением *.spo, не может быть открыта в STATA. Базу данных нужно открыть в SPSS, сохранить в формате Excel и только затем импортировать файл с расширением *.xls, *.xlsx в STATA. Название файла с данными, сохраненными в STATA или импортируемыми из Excel, должно быть на английском языке.

В следующей статье мы познакомим читателей с типами данных и принципами проверки распределения количественных данных с использованием STATA.

Список литературы

1. Гржибовский А. М. Типы данных, проверка распределения и описательная статистика // Экология человека. 2008. № 1. С. 52–58.
2. Гржибовский А. М. Использование статистики в российской биомедицинской литературе // Экология человека. 2008. № 12. С. 55–64.
3. Acock A. C. Gentle Introduction to Stata. USA, Texas : Stata Press, 2006. 289 p.

4. Baum C. F. An Introduction to Modern Econometrics Using Stata. USA, Texas : Stata Press, 2006. 341 p.
5. Hamilton C. Statistics with Stata. USA, Belmont, CA : Brooks/Cole, 2006. 409 p.
6. Kohler U., Kreute F. Data Analysis Using Stata. USA, Texas : Stata Press, 2005. 378 p.
7. Mitchell M. A Visual Guide to Stata Graphics. USA, Texas : Stata Press, 2008. 471 p.
8. Rabe-Hesketh S., Everitt Brian. A Handbook of Statistical Analyses Using Stata. New York : Chapman & Hall, 2007. 352 p.

References

1. Grjibovski A. M. Types of data, distributions and descriptive statistics. *Ekologiya cheloveka* [Human Ecology]. 2008, 1, pp. 52-58.
2. Grjibovski A. M. Use and misuse of statistics in Russian biomedical literature. *Ekologiya cheloveka* [Human Ecology]. 2008, 2, pp. 55-64.
3. Acock A.C. Gentle Introduction to Stata. USA, Texas: Stata Press, 2006, 289 p.
4. Baum C. F. An Introduction to Modern Econometrics Using Stata. USA, Texas: Stata Press, 2006, 341 p.
5. Hamilton C. Statistics with Stata. USA, Belmont, CA: Brooks/Cole, 2006, 409 p.
6. Kohler U., Kreute F. Data Analysis Using Stata. USA, Texas: Stata Press, 2005. 378 p.
7. Mitchell M. A Visual Guide to Stata Graphics. USA, Texas: Stata Press, 2008. 471 p.
8. Rabe-Hesketh S., Everitt Brian. A Handbook of Statistical Analyses Using Stata. New York: Chapman & Hall, 2007, 352 p.

INTRODUCTION TO STATA - SOFTWARE FOR STATISTICAL DATA ANALYSIS

T. N. Unguryanu, *A. M. Grjibovski

Northern State Medical University, Arkhangelsk, Russia
**Norwegian Institute of Public Health, Oslo, Norway*

This is the first article of the series of publications in the Human Ecology journal on biostatistics using STATA software. Earlier series of methodological papers published in 2008 was adapted to SPSS software. However, the price for SPSS in Russia considerably increased in recent years that led to the fact that many institutions switch from SPSS to STATA. It is worth mentioning that STATA is as popular as SPSS in European universities and research institutions. Moreover, STATA has a broader range of advanced methods of data analysis. The authors are not aware of Russian textbooks on the use of STATA in biomedical research. This series is an attempt for a manual for Russian biomedical researchers on how to use STATA software.

Keywords: STATA, biostatistics, research

Контактная информация:

Гржибовский Андрей Мечиславович — профессор, доктор медицины, старший советник Норвежского института общественного здоровья, г. Осло, Норвегия

Адрес: Nasjonalt folkehelseinstitutt, Pb 4404 Nydalen, 0403 Oslo, Norway

E-mail: angr@fhi.no