Посвящаю эту книгу своим родителям.
Сергей Мастицкий

Всем заинтересованным читателям,
без которых книги вообще не имеют смысла...
Владимир Шитиков

Содержание

Предисловие 10

Глава 1. Основные компоненты статистической среды R 13

1.1. История возникновения и основные принципы организации среды R.....13
1.2. Работа с командной консолью.....17
1.3. Работа с меню R Commander.....20
1.4. Объекты, пакеты, функции, устройства.....24

Глава 2. Описание языка R 31

2.1. Типы данных.....31
2.2. Векторы и матрицы.....32
2.3. Факторы.....38
2.4. Списки и таблицы данных.....40
Заполнение пустых значений.....45
Сортировка таблиц.....46
Объединение таблиц.....46
2.5. Импортирование данных в R.....47
2.6. Представление дат и времени. Временные ряды.....51
Форматы представления дат и времени.....51
Вычисления с датами и временем.....52
Преобразование текстовых переменных в машинный формат времени.....53
Временные ряды.....54
2.7. Организация вычислений: функции, ветвления, циклы.....56
Написание собственных функций.....57
Условия и циклы.....59
2.8. Векторизованные вычисления в R.....61

Глава 3. Базовые графические возможности R 70

3.1. Функция `plot()` и ее параметры.....70
Управляющие параметры функции `plot()`.....73
Общие аргументы графических функций.....74
3.2. Гистограммы, функции ядерной плотности и функция `cdplot()`.....79
3.3. Диаграммы размахов.....87
3.4. Круговые и столбиковые диаграммы.....91

3.5. Диаграммы Кливленда и одномерные диаграммы рассеяния 99
3.6. Категоризованные графики 107

Глава 4. Описательная статистика, подгонка распределений и смежные задачи 114

4.1. Базовые функции для расчета параметров описательной статистики 114
4.2. `summary()` и функции из дополнительных пакетов 118
4.3. Анализ выбросов 121
4.4. Заполнение пропущенных значений в таблицах данных 125
4.5. Воспроизводимость результатов при использовании генератора случайных чисел 131
4.6. Законы распределения вероятностей, реализованные в R 134
4.7. Подбор закона и параметров распределения в R 136
4.8. Проверка на нормальность распределения 144
Графические способы 145
Формальные тесты 148

Глава 5. Классические методы статистики 151

5.1. Гипотеза о равенстве средних двух генеральных совокупностей 151
Одновыборочный t-критерий 151
Сравнение двух независимых выборок 153
Сравнение двух зависимых выборок 155
5.2. Ранговый критерий Уилкоксона-Манна-Уитни 157
Одновыборочный критерий Уилкоксона 157
Сравнение двух независимых выборок 158
Сравнение двух зависимых выборок 159
5.3. Рандомизация, бутстреп и оценка статистической мощности (на примере двухвыборочного t-критерия) 161
5.4. Гипотеза об однородности дисперсий 168
Проверка однородности дисперсии в двух группах 168
Проверка однородности дисперсии в нескольких группах 169
5.5. Введение в дисперсионный анализ 171
Постановка задачи 171
Две оценки генеральной дисперсии в дисперсионном анализе 174
Выполнение дисперсионного анализа в R 176
Двухфакторный дисперсионный анализ 176
5.6. Оценка корреляции двух случайных величин 180
5.7. Критерий хи-квадрат 184
Критерий хи-квадрат для таблиц сопряженности размером 2×2 184

Критерий хи-квадрат для таблиц сопряженности размером больше 2×2 187
5.8. Точный тест Фишера. Критерии Мак-Немара и Кохрана-Мантеля-Хензеля 187
Точный тест Фишера 187
Критерий Мак-Немара 190
Критерий Кохрана-Мантеля-Хензеля для таблиц сопряженности размером 2×2×K 193
5.9. Оценка статистической мощности при сравнении частот 197

Глава 6. Дисперсионный анализ 203

6.1. Протокол разведочного анализа данных 203
Выявление точек-выбросов 204
Проверка однородности групповых дисперсий 205
Проверка на нормальность распределения 206
Выявление избыточного числа нулевых значений 207
Выявление коллинеарности 207
Выявление формы связи между переменными 210
Выявление взаимодействий между предикторами 212
Влияние пространственно-временных факторов на анализируемую переменную 216
6.2. Дисперсионный анализ как линейная модель 219
6.3. Структура модельных объектов дисперсионного анализа 227
6.4. Оценка адекватности модели дисперсионного анализа 230
Проверка исходных предположений общей линейной модели 230
Проверка условия нормальности распределения 231
Проверка условия однородности групповых дисперсий 234
Что делать, когда однофакторный дисперсионный анализ неприменим? 237
6.5. Дисперсионный анализ по Краскелу-Уоллису 239
6.6. Модели двух- и многофакторного дисперсионного анализа 241
Синтаксис объекта «формула» 242
Выполнение двухфакторного дисперсионного анализа при помощи функции `lm()` 244
Порядок перечисления предикторов в формуле модели 246
Многофакторный дисперсионный анализ 248
6.7. Контрасты в линейных моделях, содержащих категориальные предикторы 249
Основные понятия 250
Контрасты комбинаций условий (treatment contrasts) 252
Контрасты сумм (sum contrasts) 254
Контрасты Хелмерта 255

Контрасты, задаваемые пользователем 257
6.8. Проблема множественных проверок статистических гипотез 258
Поправка Бонферрони 261
Метод Холма 262
Метод Беньямини-Хохберга 263
Метод Беньямини-Йекутили 266
6.9. Апостериорные сравнения групповых средних 267
Критерий Тьюки 268
Методы множественных проверок гипотез, реализованные в пакете `multcomp` 271

Глава 7. Регрессионные модели зависимостей между количественными переменными 279

7.1. О понятии «статистическая модель» 279
Пример простейшей статистической модели 279
Исследование свойств статистических моделей имитационными методами 282
Пример модели с одним количественным предиктором 287
Назначение регрессионных моделей 289
7.2. Простая линейная регрессия: каков возраст Вселенной? 290
Модель для оценки постоянной Хаббла 291
Доверительные интервалы 293
Оценка неопределенности в отношении параметров линейной регрессии 295
Оценка «качества» регрессионной модели 301
7.3. Стандартные методы диагностики линейных моделей 304
Проверка допущений в отношении остатков модели 304
Проверка адекватности структуры систематической части модели 308
Встроенные диагностические графики 313
Выявление необычных и влиятельных наблюдений 314
7.4. Модели регрессии при разных видах функции потерь 325
Два типа регрессионных моделей 325
Робастные процедуры 329
7.5. Критерии выбора моделей оптимальной сложности 331
7.6. Полиномиальные и нелинейные модели регрессии 335
Полиномиальная регрессия 335
Нелинейная регрессия 338
7.7. Модель множественной регрессии и выбор ее спецификации 344
Полная модель и обоснование необходимости ее оптимизации 345
Пошаговые алгоритмы селекции переменных 347

Построение «всех возможных моделей» 348
Пошаговое включение предикторов в сочетании с перекрестной проверкой 350
7.8. Диагностика моделей множественной регрессии 353
Сравнение нескольких альтернативных моделей 353
Диагностика допущений в отношении остатков модели 354
Учет нелинейного характера влияния предикторов на отклик 359
7.9. Регуляризация множественной регрессии 361
Гребневая регрессия 362
Лассо-регрессия Тибширани 364
7.10. Регрессия на главные компоненты 366
7.11. Сравнение эффективности различных моделей при прогнозировании 372
Формирование исходных данных для построения моделей 372
Общая линейная модель и ее тестирование на проверочной выборке 374
Выбор информативного комплекса предикторов 375
Модели с использованием регуляризации 377
Регрессия на главные компоненты 380
Результаты и некоторые выводы 382

Глава 8. Обобщенные, структурные и иные модели регрессии 384

8.1. Модели сглаживания 384
Ядерная модель сглаживания 389
Сплайны 393
8.2. Обобщенные модели регрессии 395
8.3. Модели пробит- и логит-регрессии 399
Пробит-регрессия для моделирования зависимости «доза–эффект» 400
Логистическая регрессия 407
8.4. Пример использования обобщенных моделей для оценки экологической толерантности 411
Модели с нормально распределенным откликом 412
Модели с бинарным откликом 416
8.5. Ковариационный анализ 419
8.6. Модели со смешанными эффектами для иерархически организованных данных 424
Основные идеи 424
Пример с морскими животными: несколько частных моделей 426
8.7. Индуктивные модели (метод группового учета аргументов) 433

8.8. Моделирование структурными уравнениями 440

Глава 9. Пространственный анализ и создание картограмм 451

9.1. Простая карта: использование растрового рисунка и расчет расстояний 451
Использование географических расстояний в статистическом анализе 452
Расчет расстояния между объектами по их географическим координатам 457
9.2. Анализ пространственного размещения точек 460
9.3. Использование сервисов картографической системы Google Maps 466
9.4. Создание картограмм при помощи R 469
Шейп-файлы 470
Функция `spplot()` из пакета `sp` 474
Создание картограмм при помощи пакета `ggplot2` 478

Библиография и интернет-ресурсы 484

Основные литературные ссылки по тексту книги 484
Литература по R 484
Общеметодическая литература по статистическому анализу 485
Библиографический указатель литературы по R 485
Рекомендуемые интернет-ресурсы 494
Русскоязычные ресурсы 494
Англоязычные ресурсы 495

В целях природы обуздания,
Чтобы рассеять незнания тьму,
Берем картину мироздания
И тупо смотрим, что к чему....

А. и Б. Стругацкие.
«Понедельник начинается в субботу»

ПРЕДИСЛОВИЕ

Одним из основных инструментов познания мира является обработка данных, получаемых человеком из различных источников. Суть современного статистического анализа состоит в интерактивном процессе, состоящем из исследования, визуализации и интерпретации потоков поступаемой информации.

История последних 50 лет – это и история развития технологии анализа данных. Один из авторов этой книги с умилением вспоминает конец 60-х годов и свою первую программу расчета парной корреляции, которая набиралась металлическими штыречками на «операционном поле» из 150 ячеек персональной ЭВМ «Промінь-2» весом более 200 кг. В наше время высокопроизводительные компьютеры и доступное программное обеспечение позволяют реализовать полный цикл информационно-технологического процесса, состоящего, в общем случае, из следующих шагов:

- доступ к обрабатываемым данным (их загрузка из разных источников и комплектация совокупности взаимосвязанных исходных таблиц);
- редактирование загруженных показателей (замена или удаление пропущенных значений, преобразование признаков в более удобный вид);
- аннотирование данных (чтобы помнить, что представляет собой каждый их фрагмент);
- получение общих сведений о структуре данных (вычисление описательных статистик);
- графическое представление данных и результатов вычислений в понятной информативной форме (одна картинка на самом деле иногда стоит тысячи слов);
- моделирование данных (математическое описание зависимостей и тестирование статистических гипотез);
- оформление результатов (подготовка таблиц и диаграмм приемлемого публикационного качества).

В условиях, когда в распоряжении пользователя имеются десятки пакетов прикладных программ, актуальна проблема выбора (иногда трагичная, если вспомнить «буриданова осла»): какое программное обеспечение анализа данных следует предпочесть для своей практической работы? Здесь обычно принимаются во внимание специфика решаемой задачи, эффективность настройки алгоритмов обработки, издержки на покупку программ, а также вкусы и личные предпочте-

ния исследователя. При этом, например, шаблонная Statistica с ее механическим комплексом кнопок меню далеко не всегда может удовлетворить творческого исследователя, предпочитающего самостоятельно контролировать ход вычислительного процесса. Комбинировать различные типы анализа, иметь доступ к промежуточным результатам, управлять стилем отображения данных, добавлять собственные расширения программных модулей и оформлять итоговые отчеты в необходимом виде позволяют коммерческие вычислительные системы, включающие высокоуровневые средства командного языка, такие как Matlab, SPSS и др. Прекрасной альтернативой им является бесплатная программная среда R, являющаяся современной и постоянно развивающейся статистической платформой общего назначения.

Сегодня R является безусловным лидером среди свободно распространяемых систем статистического анализа, о чем говорит, например, тот факт, что в 2010 году система R стала победителем ежегодного конкурса открытых программных продуктов Bossie Awards в нескольких номинациях. Ведущие университеты мира, аналитики крупнейших компаний и исследовательских центров регулярно используют R при проведении научно-технических расчетов и создании крупных информационных проектов. Широкое преподавание статистики на базе пакетов этой среды и всемерная поддержка научным сообществом обусловили то, что приведение скриптов R постепенно становится общепризнанным «стандартом» как в журнальных публикациях, так и при неформальном общении ученых всего мира.

Главным препятствием для русскоязычных пользователей при освоении R, безусловно, является то, что почти вся документация по этой среде существует на английском языке. Лишь с 2008 года усилиями А. В. Шипунова, Е. М. Балдина, С. В. Петрова, И. С. Зарядова, А. Г. Буховца, П. А. Волковой и других энтузиастов появились методические пособия и книги на русском языке (ссылки на них можно найти в списке литературы в конце этой книги; там же представлены и ссылки на образовательные ресурсы, авторами которых делается посильный вклад в продвижение R среди русскоязычных пользователей). Настоящая книга дополняет эту небольшую (пока) коллекцию работ по R на русском языке, обобщая совокупность методических сообщений, опубликованных одним из авторов с 2011 года в блоге «R: Анализ и визуализация данных» (http://r-analytics.blogspot.com). Нам показалась целесообразной идея представить для удобства читателей весь этот несколько разобщенный материал в концентрированной форме, а также расширить некоторые разделы для полноты изложения.

В первых трех главах содержатся подробные указания по работе с интерактивными компонентами R, детальное описание языка и базовых графических возможностей среды. Эта часть книги вполне доступна новичкам в области программирования, хотя читатель, уже знакомый с языком R, может найти там интересные фрагменты кода или использовать приведенные описания графических параметров как справочное пособие.

В последующих главах (4–8) приведено описание распространенных процедур обработки данных и построения статистических моделей, которое иллюстрировано несколькими десятками примеров. Все эти главы включают краткие описания соответствующих алгоритмов анализа, основные полученные в примерах резуль-

таты и их возможную интерпретацию. Мы старались, по возможности, обойтись без злоупотребления «ритуальными» словооборотами, характерными для многочисленных руководств по статистике, цитирования общеизвестных теорем и приведения многоэтажных расчетных формул. Акцент делался в первую очередь на практическое применение – на то, чтобы читатель, руководствуясь прочитанным, мог проанализировать свои данные и изложить результаты коллегам.

Материал выстроен по мере усложнения. Так, главы 4 и 5 ориентированы на читателя, интересующегося статистикой лишь в объеме начального университетского курса. В главах 6 и 7 в рамках единой теории общих линейных моделей представлены дисперсионный и регрессионный анализы и приведены различные алгоритмы исследования и структурной идентификации моделей. Глава 8 посвящена некоторым современным методам построения и анализа обобщенных линейных и иных типов моделей.

Поскольку неизменный интерес у исследователей вызывают пространственный анализ и визуализация данных на географических картах и схемах, в главе 9 приведены некоторые примеры соответствующих приемов.

Отдельно стоит упомянуть обозначения, принятые в книге. Все команды языка R выделены моноширинным шрифтом. При этом имена функций дополнительно выделены полужирным шрифтом, как в этом примере: **mean**(c(1, 2, 3)). При упоминании каких-либо функций в тексте мы всегда добавляем к их именам круглые кавычки, что позволяет отличать функции от других объектов R (например, **summary**()). Результаты вычислений, полученные при использовании той или иной функции, выделены шрифтом синего цвета. Наконец, комментарии к коду представлены наклонным шрифтом серого цвета и начинаются со знака #:

```
# Расчет среднего значения:
mean(c(1, 2, 3))
[1] 2
```

Мы адресуем эту книгу студентам, аспирантам, а также молодым и состоявшимся ученым, желающим освоить анализ и визуализацию данных с использованием среды R. Мы надеемся, что по окончании чтения этого руководства вы получите некоторое представление о том, как работает R, где можно получить дальнейшую информацию, а также как справиться с широким спектром задач анализа данных.

Файлы со скриптами кода R по всем главам книги и таблицы исходных данных, необходимые для выполнения примеров, свободно доступны для скачивания с GitHub-репозитория https://github.com/ranalytics/r-tutorials, а также с сайта Института экологии Волжского бассейна РАН по ссылке http://www.ievbras.ru/ecostat/Kiril/R/Scripts.zip.

Мы будем благодарны Вам, Читатель, за любые замечания и пожелания касательно этой работы, которые Вы можете направлять по электронной почте rtutorialsbook@gmail.com.

Сергей Мастицкий, Лондон
Владимир Шитиков, Тольятти
май 2015 года

Глава 1

Основные компоненты статистической среды R

1.1. История возникновения и основные принципы организации среды R

Система статистического анализа и визуализации данных R состоит из следующих основных частей:

- языка программирования высокого уровня R, позволяющего одной строкой реализовать различные операции с объектами, векторами, матрицами, списками и т. д.;
- большого набора функций обработки данных, собранных в отдельные так называемые «пакеты»;
- развитой системой поддержки, включающей обновление компонентов среды, интерактивную помощь и различные образовательные ресурсы, предназначенные как для начального изучения R, так и для последующих консультаций по возникающим затруднениям.

Начало пути относится к 1993 г., когда двое молодых новозеландских ученых Росс Ихака (Ross Ihaka) и Роберт Джентльмен (Robert Gentleman) анонсировали свою новую разработку, которую назвали R. Они взяли за основу язык программирования развитой коммерческой системы статистической обработки данных S-PLUS и создали его бесплатную свободную реализацию, отличающуюся от своего прародителя легко расширяемой модульной архитектурой. В скором времени возникла распределенная система хранения и распространения пакетов к R, известная под аббревиатурой «CRAN» (Comprehensive R Archive Network – http://cran.r-project.org), основная идея организации которой – постоянное расширение, коллективное тестирование и оперативное распространение прикладных средств обработки данных.

Оказалось, что такой продукт непрерывных и хорошо скоординированных усилий мощного «коллективного разума» тысяч бескорыстных разработчиков-интеллектуалов оказался значительно эффективнее коммерческих статистических программ, стоимость лицензии на которые может составлять несколько тысяч долларов. Поскольку R является любимым языком профессиональных статистиков, все последние достижения статистической науки очень быстро становятся

доступными для пользователей R во всем мире в виде дополнительных пакетов. Ни одна коммерческая система статистического анализа так быстро сегодня не развивается. У R есть многочисленная армия пользователей, которые сообщают авторам дополнительных пакетов и самой системы R об обнаруженных ошибках, которые оперативно исправляются.

Язык вычислений R, хотя и требует определенных усилий для своего освоения, недюжинных поисковых навыков и энциклопедической памяти, позволяет оперативно выполнить расчеты, по своему разнообразию практически «столь же неисчерпаемые, как атом». По состоянию на конец февраля 2015 г., энтузиастами со всего мира было создано 7336 дополнительных библиотек для R, включающих 149 688 функций (см. http://www.rdocumentation.org), которые существенно расширяют базовые возможности системы. Очень сложно представить какой-либо класс статистических методов, который еще не реализован сегодня в виде пакетов R, включая, разумеется, весь «джентльменский набор»: линейные и обобщенные линейные модели, нелинейные регрессионные модели, планирование эксперимента, анализ временных рядов, классические параметрические и непараметрические тесты, байесовская статистика, кластерный анализ и методы сглаживания. При помощи мощных средств визуализации результаты анализа можно обобщать в виде всевозможных графиков и диаграмм. Кроме традиционной статистики, разработанный функционал включает большой набор алгоритмов численной математики, методов оптимизации, решения дифференциальных уравнений, распознавания образов и др. Свои специфические методы обработки данных могут обнаружить в составе пакетов R генетики и социологи, лингвисты и психологи, химики и медики, специалисты по ГИС- и веб-технологиям.

«Фирменная» документация по R весьма объемна и далеко не всегда толково написана (по странной традиции англоязычной литературы, слишком много слов расходуется на описание тривиальных истин, тогда как важные моменты пробегаются скороговоркой). Однако в дополнение к этому ведущими мировыми издательствами (Springer, Cambridge University Press и Chapman & Hall/CRC) или просто отдельными коллективами энтузиастов выпущено огромное число книг, описывающих различные аспекты анализа данных в R (см., например, список литературы на сайте «Энциклопедия психодиагностики», http://psylab.info/R:Литература). Кроме того, существует несколько активно действующих международных и российских форумов пользователей R, где любой может попросить о помощи в возникшей проблеме. В списке литературы мы приводим пару сотен книг и интернет-ссылок, на которые советуем обратить особое внимание в ходе изучения R.

Непосредственное обучение практической работе в R состоит из *а*) освоения конструкций языка R и знакомства с особенностями вызова функций, выполняющих анализ данных, и *б*) приобретения навыков работы с программами, реализующими специфические методы анализа и визуализации данных.

Вопрос выбора средств пользовательского интерфейса R неоднозначен и сильно зависит от вкусов пользователей. Единого мнения нет даже у авто-

ритетных специалистов. Одни считают, что нет ничего лучше стандартного консольного интерфейса R. Другие полагают, что для удобной работы стоит инсталлировать какую-либо из имеющихся интегрированных сред разработки (IDE) с богатым набором кнопочных меню. Например, отличным вариантом является бесплатная интегрированная среда разработки RStudio (http://www.rstudio.com). Ниже мы остановимся на описании консольного варианта и на работе с R Commander, но дальнейшим исканиям читателя может помочь обзор различных версий IDE, представленный в приложении к книге А. Шипунова с соавторами (2014).

Один из R-экспертов, Джозеф Рикерт, считает, что процесс изучения R можно разделить на следующие этапы (подробнее см. его статью на сайте http://bit.ly/1FlOx2E):

1. Знакомство с общими принципами культуры R-сообщества и программной среды, в которой разрабатывался и функционирует язык R. Посещение основных и вспомогательных ресурсов и освоение хорошего вводного учебника. Инсталляция R на компьютере пользователя и выполнение первых тестовых скриптов.
2. Считывание данных из стандартных файлов операционной системы и уверенное использование R-функций для выполнения ограниченного набора привычных пользователю процедур статистического анализа.
3. Использование базовых структур языка R для написания простых программ. Написание собственных функций. Ознакомление со структурами данных, с которыми может работать R, и более сложными возможностями языка. Работа с базами данных, веб-страницами и внешними источниками данных.
4. Написание сложных программ на языке R. Самостоятельная разработка и глубокое понимание структуры объектов так называемых S3- и S4-классов.
5. Разработка профессиональных программ на языке R. Самостоятельное создание дополнительных модулей для R.

Большинство рядовых пользователей R останавливаются на стадии 3, так как полученных к этому времени знаний им вполне достаточно для выполнения статистических задач по профилю их основной профессиональной деятельности. Примерно в этом объеме мы и приводим описание языка R в рамках настоящего руководства.

Установить и настроить базовую комплектацию статистической среды R весьма просто. На момент написания этой книги актуальной была версия R 3.1.2 для Windows (доступны также дистрибутивы для всех других распространенных операционных систем). Скачать дистрибутив системы вместе с базовым набором из 29 пакетов можно совершенно бесплатно с основного сайта проекта http://cran.r-project.org или его русского «зеркала» http://cran.gis-lab.info. Процесс инсталляции системы из скачанного дистрибутива затруднений не вызывает и не требует никаких особых комментариев.

Для удобства хранения скриптов, исходных данных и результатов расчетов стоит выделить на пользовательском компьютере специальный рабочий каталог. Весьма нежелательно использовать в названии рабочего каталога символы кириллицы.

Путь к рабочему каталогу и некоторые другие опции настроек целесообразно разместить, изменив в любом текстовом редакторе системный файл C:\Program Files\R\R-3.1.2\etc\Rprofile.site (на вашем компьютере он может иметь иной адрес). В представленном ниже примере модифицированные строки отмечены синим цветом. Помимо указания рабочего каталога, эти строки определяют ссылку на российский источник загрузки пакетов R и автоматический запуск R Commander.

Листинг файла Rprofile.site

```
# Все, что следует за символом комментария "#", средой игнорируется
# options(papersize="a4")
# options(editor="notepad")
# options(pager="internal")

# установить тип отображения справочной информации
# options(help_type="text")
  options(help_type="html")

# установить место расположения локальной библиотеки
# .Library.site <- file.path(chartr("\\", "/", R.home()), "site-library")

# При загрузке среды запустить меню R Commander
# Поставить знаки #, если запуск Rcmdr не нужен
 local({
 old <- getOption("defaultPackages")
 options(defaultPackages = c(old, "Rcmdr"))
 })

# Определить зеркало CRAN
 local({r <- getOption("repos")
       r["CRAN"] <- "http://cran.gis-lab"
       options(repos = r)})

# Определить путь к рабочему каталогу (любой иной на вашем компьютере)
setwd("D:/R/Process/Resampling")
```

Что касается «хорошего вводного учебника», то любые наши рекомендации неизбежно будут носить субъективный оттенок. Тем не менее следует упомянуть официальное введение в R (Venables & Smith, 2014) и книгу Kabacoff (2011), отчасти еще и потому, что имеется их русский перевод. Отметим также традиционное «наставление для чайников» Meys & Vries (2012) и руководства Dalgaard (2008) и Lam (2010). Из русскоязычных вводных книг наиболее полными являются работы И. Зарядова (2010а, б) и А. Шипунова с соавторами (2014).

1.2. Работа с командной консолью

Статистическая среда R выполняет любой набор осмысленных инструкций языка R, содержащихся в файле скрипта или представленных последовательностью команд, задаваемых с консоли. Работа с консолью может показаться трудной для современных пользователей, привыкших к кнопочным меню, поскольку надо запоминать синтаксис отдельных команд. Однако после приобретения уже некоторых первичных навыков окажется, что многие процедуры обработки данных можно выполнять быстрее и с меньшим трудом, чем, предположим, в той же программе Statistica.

Консоль R представляет собой диалоговое окно, в котором пользователь вводит команды и результаты их выполнения. Это окно возникает сразу при запуске среды (например, после клика мышью на ярлыке R на рабочем столе). Кроме того, стандартный графический пользовательский интерфейс R (RGui) включает окно редактирования скриптов и всплывающие окна с графической информацией (рисунками, диаграммами и прочим).

В командном режиме R может работать, например, как обычный калькулятор (рис. 1). Справа от символа приглашения («*prompt*») > пользователь может ввести произвольное арифметическое выражение, нажать клавишу **Enter** и тут же получить результат. Например, во второй команде на рис. 1 мы использовали функции факториала и синуса, а также встроенное число π. Результаты, полученные в текстовой форме, можно выделить мышью и скопировать через буфер обмена в любой текстовый файл (например, документ Word).

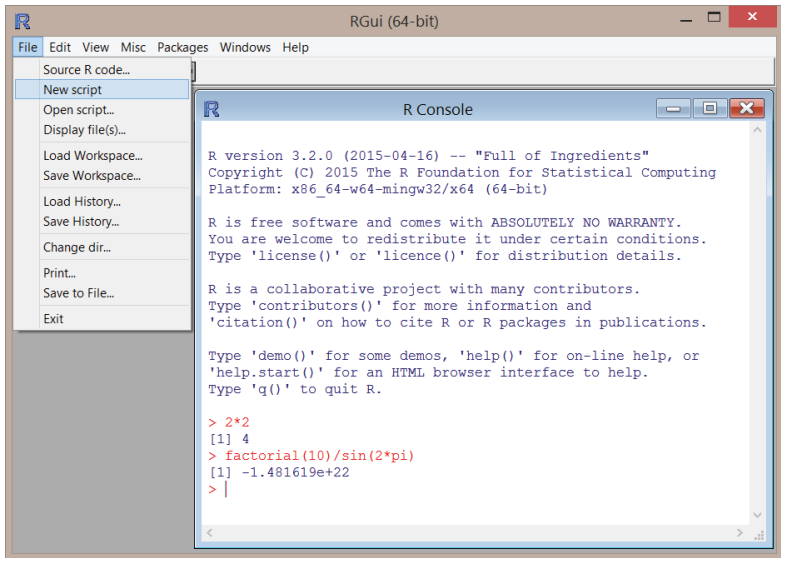


Рисунок 1

При работе с использованием RGui мы рекомендуем во всех случаях создавать скрипт (то есть файл с последовательностью команд языка R, выполняющей определенные действия). Как правило, это обычный текстовый файл с любым именем и, желательно, расширением `.r`, который можно создавать и редактировать обычным редактором типа «Блокнот». Если этот файл существует, его лучше всего поместить в рабочий каталог, и тогда после запуска R и выбора пункта меню **Файл** > **Открыть скрипт** (File > Open script) содержимое этого файла появится в окне **Редактор R** (R Editor). Выполнить последовательность команд скрипта можно из пункта меню **Правка** > **Запустить все** (Edit > Run all). Можно также выделить мышью осмысленный фрагмент из любого места подготовленного скрипта (от имени одной переменной до всего содержимого) и осуществить запуск этого блока на выполнение. Это можно сделать четырьмя возможными способами: из основного и контекстного меню, комбинацией клавиш **Ctrl**+**R** или кнопкой на панели инструментов.

Рисунок 2 описывает следующие действия:

- из бесплатного интернет-источника *Global Administrative Areas* (http://gadm.org/) был скачан R-объект `gadm` с данными по территориальному делению Республики Беларусь;
- переменная `NAME_1`, содержащая названия областных центров, преобразована в объект класса «фактор»;
- с использованием функции **`spplot()`** из пакета `sp` в графическое окно программы выведена административная карта, которую можно средствами

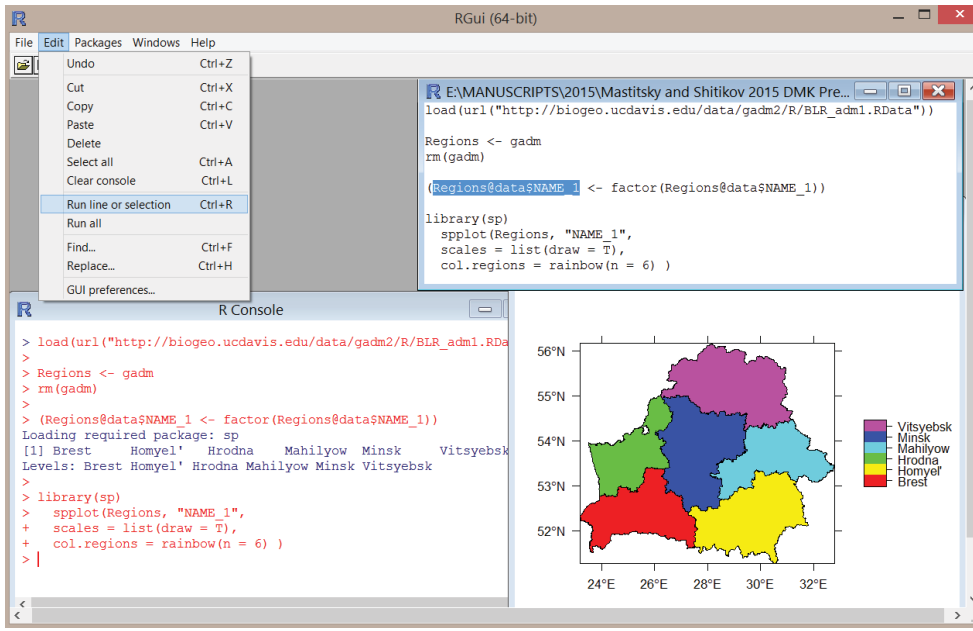


Рисунок 2

меню скопировать в буфер обмена или сохранить как стандартный мета- или растровый графический файл.

Подробнее смысл отдельных команд мы рассмотрим в последующих разделах, а здесь обратим внимание на то, что, выделив в скрипте и запустив на выполнение комбинацию символов `Regions@data`, мы получим в окне консоли весь набор данных `data` по объекту `gadm`, а команда, составленная из выделенных символов `Regions@data$NAME_1`, даст нам список наименований административных центров.

Таким образом, **Редактор R** позволяет легко выполнить навигацию по скрипту, редактирование и выполнение любой комбинации команд, поиск и замену определенных частей кода. Упомянутая выше интегрированная среда разработки RStudio позволяет дополнительно выполнять подсветку синтаксиса кода, его автоматическое завершение, «упаковку» последовательностей команд в функции для их последующего использования, работу с документами Sweave или TeX и другие операции, которые будут полезны продвинутому пользователю.

R обладает обширными встроенными справочными материалами, которые можно получить непосредственно в RGui. Если подать с консоли команду **`help.start()`**, то в вашем интернет-браузере откроется страница, предоставляющая доступ ко всем внутренним справочным ресурсам: основным руководствам, авторским материалам, ответам на распространенные вопросы, истории внесенных в R изменений и т. д. (рис. 3).

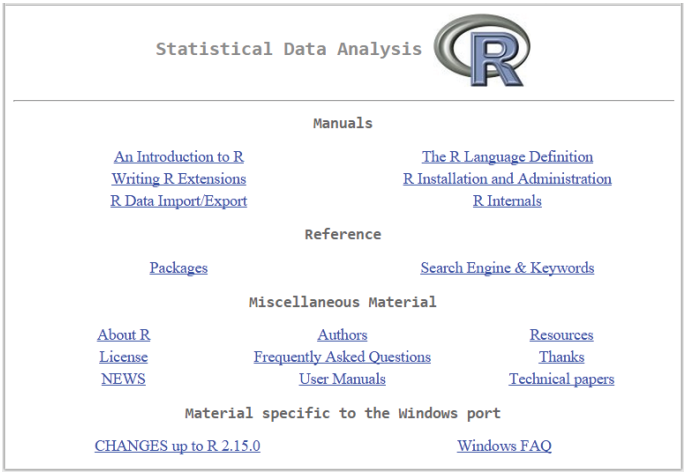


Рисунок 3

Справку по отдельным функциям можно получить с использованием следующих команд:

- **`help`**`("foo")` или `?foo` – справка по функции `foo` (кавычки необязательны);
- **`help.search`**`("foo")` или `??foo` – поиск всех справочных файлов, содержащих `foo`;

- `example("foo")` – примеры использования функции `foo`;
- `RSiteSearch("foo")` – поиск ссылок в онлайн-руководствах и архивах рассылок;
- `apropos("foo", mode = "function")` – список всех функций, в названии которых встречается `foo`;
- `vignette("foo")` – список руководств по теме `foo`.

1.3. Работа с меню R Commander

Удобным средством освоения вычислений в R для начинающего пользователя является R Commander – платформонезависимый графический интерфейс с кнопочным меню, реализованный в пакете `Rcmdr`. Он позволяет осуществить большой комплект процедур статистического анализа, не прибегая к предварительному заучиванию функций на командном языке, однако невольно способствует этому, поскольку отображает все выполняемые инструкции в консоли программы.

Установить `Rcmdr`, как и любые другие расширения, можно из меню **Пакеты** > **Установить пакет** (Packages > Install package(s)), но лучше выполнив команду:

```
install.packages("Rcmdr", dependencies = TRUE),
```

где включение опции `dependencies` вызовет гарантированную установку полного комплекта остальных пакетов, которые могут потребоваться при обработке данных через меню `Rcmdr`. Обратите внимание: выполнение приведенной команды предполагает, что ваш компьютер подключен к Интернету.

Запуск R Commander происходит при загрузке пакета `Rcmdr` через меню **Пакеты** > **Включить пакет** (Packages > Load package) или командой `library(Rcmdr)`.

Если по какой-то причине было принято решение анализировать данные исключительно с помощью R Commander, то для автоматической загрузки этой графической оболочки при запуске R необходимо отредактировать файл `Rprofile.site` (см. раздел 1.1).

Работу в R Commander рассмотрим на примере корреляционного анализа данных по уровню инвазированности двустворчатого моллюска *Dreissena polymorpha* (http://bit.ly/1C0WvOW) инфузорией *Conchophthirus acuminatus* в трех озерах Беларуси (Mastitsky, 2013, http://bit.ly/1wsDKm4). В таблице с исходными данными, которую скачаем с сайта *figshare* (http://bit.ly/1wMGDOQ), нас будут интересовать две переменные: длина раковины моллюска (`ZMlength`, мм) и число обнаруженных в моллюске инфузорий (`CAnumber`). Подробно этот пример будет рассмотрен в главах 4 и 5, поэтому здесь мы не будем детально останавливаться на смысле анализа, а сосредоточимся на технике работы с `Rcmdr`.

Первый этап – загрузка нового набора данных, и мы выбираем из меню **Данные** > **Импорт данных… из URL** (Data > Import data ... from URL) (рис. 4).

Далее определяем во всплывающих окнах режим загрузки данных и адрес ссылки в Интернете. Нетрудно заметить, что те же данные мы могли легко загрузить

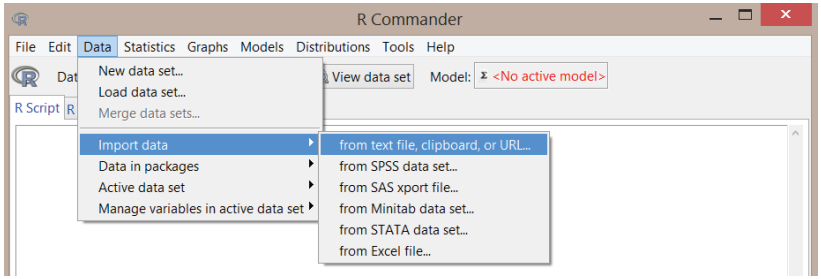
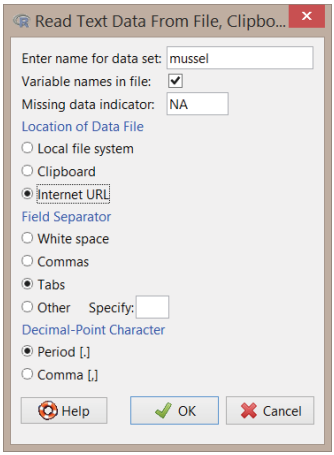
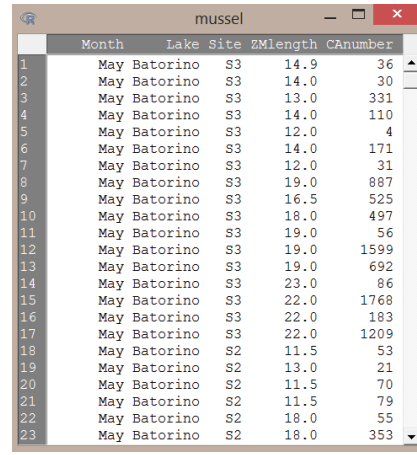


Рисунок 4

из локального текстового файла, книги Excel или таблицы базы данных. Чтобы убедиться в том, что наши данные загружены верно (и при необходимости их отредактировать), нажимаем кнопку **Посмотреть данные** (View data set) (рис. 5).



Окно спецификации данных



Фрагмент загруженной таблицы

Рисунок 5

На втором этапе в меню **Статистики** > **Итоги** (Statistics > Summaries) выбираем **Корреляционный тест** (Correlation test) (рис. 6).

Выделяем пару необходимых переменных, жмем **OK** и в окне вывода получаем коэффициент корреляции Пирсона (R = 0.467), уровень достигнутой статистической значимости (`p-value < 2.2e-16`) и 95%-ные доверительные пределы (рис. 7). Полученные результаты легко скопировать из окна вывода через буфер обмена.

Теперь получим графическое изображение корреляционной зависимости. В меню **Графики** (Graphs) выберем **Точечный график** (Scatterplot) зависимости `CAnumber` от `ZMlength` и снабдим ее краевыми диаграммами размахов, линией линейного тренда по методу наименьших квадратов (зеленым цветом) и сглаживающей линией по методу локальной регрессии (красным цветом), представленной с до-

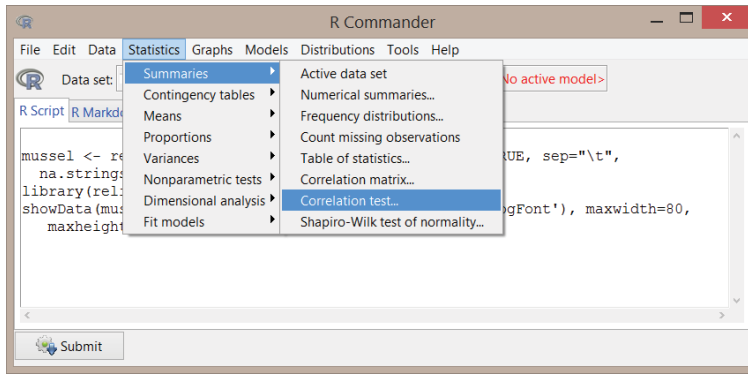
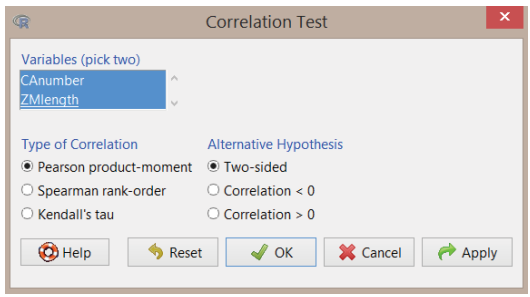


Рисунок 6



Окно спецификации вычислений

```
Pearson's product-moment correlation
data:  mussel$CAnumber and
mussel$ZMlength
t = 11.4964, df = 474
p-value < 2.2e-16
alternative hypothesis: true correlation
is not equal to 0
95 percent confidence interval:
 0.3935877 0.5343949
sample estimates:
     cor
0.466946
```

Полученные результаты

Рисунок 7

верительной областью (пунктир). Для каждого из трех озер (переменная `Lake`) экспериментальные точки будут представлены разными символами (рис. 8–9).

При нажатии кнопок меню R Commander в консоли R появляются эквивалентные инструкции языка R. В нашем случае они имеют следующий вид:

```
mussel <- read.table("http://bit.ly/1wMGDOQ", header=TRUE, sep="\t",
                     na.strings="NA", dec=".", strip.white=TRUE)
cor.test(mussel$CAnumber, mussel$ZMlength,
         alternative="two.sided", method="pearson")
scatterplot(CAnumber ~ ZMlength | Lake, reg.line=lm,
            smooth=TRUE, spread=TRUE, boxplots='xy', span=0.5,
            ylab="Численность инфузорий",
            xlab="Длина раковины",
            by.groups=FALSE, data=mussel)
```

Конец ознакомительного фрагмента.
Приобрести книгу можно
в интернет-магазине «Электронный универс»
(e-Univers.ru)