



ОТ ПЕРЕВОДЧИКА

По моему глубокому убеждению, на сегодняшний день это – лучшая книга, посвященная обработке данных в статистической среде R, для неспециалистов. Я рада, что теперь она стала доступной русскоязычным читателям. Надеюсь, мой перевод не сильно испортил эту книгу. По крайней мере, в некоторых местах она точно стала лучше, потому что я исправила довольно многочисленные и не всегда безобидные опечатки, обнаруженные в исходном издании мною и другими читателями, которые оставили свои отзывы на форуме издательства.

Я благодарю Александра Лободу, который рассказал мне о существовании этой книги, Дмитрия Мовчана, с энтузиазмом воспринявшего мое предложение опубликовать ее перевод, Бориса Демешева за консультации по переводу некоторых статистических терминов и Алексея Шипунова за техническую поддержку. Я особенно признательна Сергею Петрову и Сергею Мастицкому за внимательное прочтение рукописи перевода и конструктивные замечания.



ОГЛАВЛЕНИЕ

| | |
|---|-----------|
| От переводчика | 5 |
| Предисловие | 13 |
| Благодарности | 16 |
| Об этой книге | 18 |
| Об иллюстрации на обложке | 24 |
| | |
| ЧАСТЬ 1. | |
| Начало работы | 25 |
| Глава 1. Знакомство с R | 28 |
| 1.1. Зачем использовать R? | 30 |
| 1.2. Получение и установка R | 33 |
| 1.3. Работа в R | 33 |
| 1.3.1. Начало работы | 34 |
| 1.3.2. Как получить помощь | 37 |
| 1.3.3. Рабочее пространство | 38 |
| 1.3.4. Ввод и вывод | 41 |
| 1.4. Пакеты | 42 |
| 1.4.1. Что такое пакеты? | 42 |
| 1.4.2. Установка пакета | 44 |
| 1.4.3. Загрузка пакета | 44 |
| 1.4.4. Получение информации о пакете | 44 |
| 1.5. Пакетная обработка | 45 |
| 1.6. Использование вывода в качестве ввода – повторное использование результатов | 46 |
| 1.7. Работа с большими массивами данных | 47 |
| 1.8. Учимся на примере | 47 |
| 1.9. Резюме | 49 |
| Глава 2. Создание набора данных | 50 |
| 2.1. Что такое набор данных? | 51 |
| 2.2. Структуры данных | 52 |
| 2.2.1. Векторы | 53 |
| 2.2.2. Матрицы | 54 |

| | |
|--|------------|
| 2.2.3. Массивы данных | 56 |
| 2.2.4. Таблицы данных | 57 |
| 2.2.5. Факторы..... | 61 |
| 2.2.6. Списки | 63 |
| 2.3. Ввод данных..... | 65 |
| 2.3.1 Ввод данных с клавиатуры..... | 66 |
| 2.3.2. Импорт данных из текстового файла с разделителями..... | 67 |
| 2.3.3. Импорт данных из Excel..... | 69 |
| 2.3.4. Импорт данных из XML файлов..... | 70 |
| 2.3.5. Извлечение данных из веб-страниц..... | 70 |
| 2.3.6. Импорт данных из SPSS | 70 |
| 2.3.7. Импорт данных из SAS | 71 |
| 2.3.8. Импорт данных из Stata..... | 71 |
| 2.3.9. Импорт данных из netCDF | 72 |
| 2.3.10. Импорт данных из HDF5 | 72 |
| 2.3.11. Импорт данных из систем управления базами данных..... | 73 |
| 2.3.12. Импорт данных при помощи Stat/Transfer | 75 |
| 2.4. Аннотирование наборов данных..... | 75 |
| 2.4.1. Подписи для переменных..... | 76 |
| 2.4.2. Пояснение значений переменных | 76 |
| 2.5. Полезные функции для работы с объектами..... | 77 |
| 2.6. Резюме | 78 |
| Глава 3. Начало работы с диаграммами..... | 79 |
| 3.1. Работа с диаграммами..... | 80 |
| 3.2. Простой пример..... | 82 |
| 3.3. Графические параметры | 84 |
| 3.3.1. Символы и линии | 85 |
| 3.3.2. Цвета | 86 |
| 3.3.3. Характеристики текста | 88 |
| 3.3.4. Размеры диаграммы и полей | 91 |
| 3.4. Добавление текста, настройка параметров осей и условных обозначений..... | 93 |
| 3.4.1. Заголовки | 93 |
| 3.4.2. Оси | 94 |
| 3.4.3. Опорные линии | 97 |
| 3.4.4. Легенда..... | 98 |
| 3.4.5. Аннотации | 100 |
| 3.5. Объединение диаграмм | 103 |
| 3.5.1. Полный контроль над расположением диаграмм..... | 108 |
| 3.9. Резюме | 110 |
| Глава 4. Основы управления данными | 111 |
| 4.1. Рабочий пример..... | 111 |
| 4.2. Создание новых переменных | 114 |
| 4.3. Перекодировка переменных | 115 |

| | |
|--|-----|
| 4.4. Переименование переменных..... | 117 |
| 4.5. Пропущенные значения | 119 |
| 4.5.1. Перекодировка значений в отсутствующие | 120 |
| 4.5.2. Исключение пропущенных значений из анализа..... | 120 |
| 4.6. Календарные даты как данные | 122 |
| 4.6.1. Преобразование дат в текстовые переменные | 124 |
| 4.6.2. Получение дальнейшей информации | 124 |
| 4.7. Преобразования данных из одного типа в другой..... | 125 |
| 4.8. Сортировка данных..... | 126 |
| 4.9. Объединение наборов данных | 127 |
| 4.9.1. Добавление столбцов | 127 |
| 4.9.2. Добавление строк | 128 |
| 4.10. Разделение наборов данных на составляющие | 128 |
| 4.10.1. Выбор переменных | 128 |
| 4.10.2. Исключение переменных..... | 129 |
| 4.10.3. Выбор наблюдений | 130 |
| 4.10.4. Функция subset() | 131 |
| 4.10.5. Случайные выборки..... | 132 |
| 4.11. Использование команд SQL для преобразования таблиц данных | 133 |
| 4.12. Резюме | 134 |

Глава 5. Более сложные способы управления данными 135

| | |
|--|-----|
| 5.1. Задача по управлению данными, которую нужно решить ... | 136 |
| 5.2. Числовые и текстовые функции..... | 137 |
| 5.2.1. Математические функции | 137 |
| 5.2.2. Статистические функции..... | 138 |
| 5.2.3. Функции распределения | 141 |
| 5.2.4. Текстовые функции | 146 |
| 5.2.5. Другие полезные функции..... | 147 |
| 5.2.6. Применение функций к матрицам и таблицам данных | 149 |
| 5.3. Решение нашей задачи по управлению данными | 150 |
| 5.4. Управление выполнением команд..... | 155 |
| 5.4.1. Повторение и циклы | 156 |
| 5.4.2. Выполнение при условии | 157 |
| 5.5. Функции, написанные пользователем | 158 |
| 5.6. Агрегирование и изменение структуры данных | 161 |
| 5.6.1. Транспонирование | 161 |
| 5.6.2. Агрегирование данных | 162 |
| 5.6.3. Пакет reshape | 163 |
| 5.7. Резюме | 165 |

ЧАСТЬ 2.

Базовые методы 167

| | |
|---|----------------|
| Глава 6. Базовые диаграммы | 169 |
| 6.1. Столбчатые диаграммы | 170 |
| 6.1.1. Простые столбчатые диаграммы | 170 |
| 6.1.2. Столбчатые диаграммы: составные и с группировкой | 172 |
| 6.1.3. Столбчатые диаграммы для средних значений | 173 |
| 6.1.4. Оптимизация столбчатых диаграмм | 175 |
| 6.1.5. Спинограммы..... | 176 |
| 6.2. Круговые диаграммы | 177 |
| 6.3. Гистограммы..... | 180 |
| 6.4. Диаграммы ядерной оценки функции плотности | 183 |
| 6.5. Диаграммы размахов..... | 186 |
| 6.5.1. Использование диаграмм размахов для сравнения групп между собой..... | 187 |
| 6.5.2. Скрипичные диаграммы | 191 |
| 6.6. Точечные диаграммы | 192 |
| 6.7. Резюме | 195 |
| Глава 7. Основные методы статистической обработки данных | 196 |
| 7.1. Описательные статистики | 197 |
| 7.1.1. Калейдоскоп методов | 198 |
| 7.1.2. Вычисление описательных статистик для групп данных | 202 |
| 7.1.3. Визуализация результатов | 206 |
| 7.2. Таблицы частот и таблицы сопряженности | 206 |
| 7.2.1. Создание таблиц частот | 207 |
| 7.2.2. Тесты на независимость..... | 214 |
| 7.2.3. Показатели взаимосвязи..... | 216 |
| 7.2.4. Визуализация результатов | 217 |
| 7.2.5. Преобразование таблиц в неструктурированные файлы | 217 |
| 7.3. Корреляции | 219 |
| 7.3.1. Типы корреляций..... | 220 |
| 7.3.2. Проверка статистической значимости корреляций..... | 223 |
| 7.3.3. Визуализация корреляций | 226 |
| 7.4. Тесты Стьюдента..... | 226 |
| 7.4.1. Тест Стьюдента для независимых выборок..... | 227 |
| 7.4.2. Тест Стьюдента для зависимых выборок | 228 |
| 7.4.3. Когда имеется больше двух групп..... | 229 |
| 7.5. Непараметрические тесты межгрупповых различий..... | 229 |
| 7.5.1. Сравнение двух групп..... | 229 |
| 7.5.2. Сравнение более двух групп..... | 231 |
| 7.6. Визуализация групповых различий | 234 |
| 7.7. Резюме | 234 |

ЧАСТЬ 3.

| | |
|--|------------|
| Методы обработки данных средней сложности ... | 235 |
|--|------------|

| | |
|--|------------|
| Глава 8. Регрессия | 237 |
| 8.1. Многоликая регрессия..... | 239 |
| 8.1.1. Ситуации, в которых используется МНК-регрессия..... | 240 |
| 8.1.2. Что вам нужно знать..... | 242 |
| 8.2. МНК-регрессия..... | 242 |
| 8.2.1. Подгонка регрессионных моделей при помощи команды <code>lm()</code> | 243 |
| 8.2.2. Простая линейная регрессия..... | 245 |
| 8.2.3. Полиномиальная регрессия..... | 248 |
| 8.2.4. Множественная линейная регрессия..... | 251 |
| 8.2.5. Множественная линейная регрессия со взаимодействиями..... | 255 |
| 8.3. Диагностика регрессионных моделей..... | 257 |
| 8.3.1. Стандартный подход..... | 258 |
| 8.3.2. Усовершенствованный подход..... | 262 |
| 8.3.3. Общая проверка выполнения требований, предъявляемых к линейным моделям..... | 270 |
| 8.3.4. Мультиколлинеарность..... | 271 |
| 8.4. Необычные наблюдения..... | 272 |
| 8.4.1. Выбросы..... | 273 |
| 8.4.2. Точки высокой напряженности..... | 273 |
| 8.4.3. Влиятельные наблюдения..... | 275 |
| 8.5. Способы корректировки..... | 279 |
| 8.5.1. Удаление наблюдений..... | 279 |
| 8.5.2. Преобразование переменных..... | 279 |
| 8.5.3. Добавление или удаление переменных..... | 282 |
| 8.5.4. Попытка применить другой подход..... | 282 |
| 8.6. Выбор «лучшей» регрессионной модели..... | 283 |
| 8.6.1. Сравнение моделей..... | 283 |
| 8.6.2. Выбор переменных..... | 284 |
| 8.7. Продолжение анализа..... | 289 |
| 8.7.1. Кросс-валидация..... | 290 |
| 8.7.2. Относительная важность..... | 292 |
| 8.8. Резюме..... | 296 |
| Глава 9. Дисперсионный анализ | 297 |
| 9.1. Ускоренный курс терминологии..... | 298 |
| 9.2. Подгонка ANOVA-моделей..... | 302 |
| 9.2.1. Функция <code>aov()</code> | 302 |
| 9.2.2. Порядок членов в формуле..... | 303 |
| 9.3. Однофакторный дисперсионный анализ..... | 305 |
| 9.3.1. Множественные сравнения..... | 306 |
| 9.3.2. Проверка справедливости допущений, лежащих в основе теста..... | 310 |
| 9.4. Однофакторный ковариационный анализ..... | 312 |
| 9.4.1. Проверка допущений, лежащих в основе теста..... | 314 |

| | |
|---|------------|
| 9.4.2. Визуализация результатов | 315 |
| 9.5. Двухфакторный дисперсионный анализ | 316 |
| 9.6. Дисперсионный анализ для повторных измерений | 321 |
| 9.7. Многомерный дисперсионный анализ | 324 |
| 9.7.1. Проверка предположений, лежащих в основе теста | 326 |
| 9.7.2. Устойчивый многомерный дисперсионный анализ | 328 |
| 9.8. Дисперсионный анализ как регрессия | 329 |
| 9.9. Резюме | 331 |
| Глава 10. Анализ мощности | 333 |
| 10.1. Краткий обзор процедуры проверки гипотез | 334 |
| 10.2. Проведение анализа мощности при помощи пакета <code>pwf</code> | 337 |
| 10.2.1. Тесты Стьюдента | 338 |
| 10.2.2. Дисперсионный анализ | 340 |
| 10.2.3. Корреляции | 341 |
| 10.2.4. Линейные модели | 342 |
| 10.2.5. Сравнение пропорций | 343 |
| 10.2.6. Тесты хи-квадрат | 344 |
| 10.2.7. Выбор подходящего размера эффекта в незнакомых ситуациях | 346 |
| 10.3. Графический анализ мощности | 348 |
| 10.4. Другие пакеты | 350 |
| 10.5. Резюме | 352 |
| Глава 11. Диаграммы средней сложности | 354 |
| 11.1. Диаграммы рассеяния | 355 |
| 11.1.1. Матрицы диаграмм рассеяния | 359 |
| 11.1.2. Диаграммы рассеяния высокой плотности | 365 |
| 11.1.3. Трехмерные диаграммы рассеяния | 368 |
| 11.1.4. Пузырьковые диаграммы | 373 |
| 11.2. Линейные графики | 375 |
| 11.3. Кореллограммы | 380 |
| 11.4. Мозаичные диаграммы | 386 |
| 11.5. Резюме | 389 |
| Глава 12. Статистика повторных выборок и бутстреп-анализ | 390 |
| 12.1. Перестановочные тесты | 391 |
| 12.2. Перестановочные тесты в пакете <code>coin</code> | 393 |
| 12.2.1. Тесты на независимость для двух и k выборок | 395 |
| 12.2.2. Независимость в таблицах сопряженности | 397 |
| 12.2.3. Независимость между числовыми переменными | 398 |
| 12.2.4. Тесты для двух и k зависимых выборок | 398 |
| 12.2.5. Дополнительная информация | 399 |

| | |
|---|-----|
| 12.3. Перестановочные тесты, реализованные в пакете <code>lmPerm</code> | 399 |
| 12.3.1. Простая и полиномиальная регрессия | 400 |
| 12.3.2. Множественная регрессия | 401 |
| 12.3.3. Однофакторные дисперсионный и ковариационный анализы..... | 402 |
| 12.3.4. Двухфакторный дисперсионный анализ | 403 |
| 12.4. Дополнительные замечания о перестановочных тестах ... | 405 |
| 12.5. Бутстреп-анализ | 406 |
| 12.6. Бутстреп-анализ при помощи пакета <code>boot</code> | 407 |
| 12.6.1. Бутстреп-анализ для одной статистики | 409 |
| 12.6.2. Бутстреп-анализ для нескольких статистик | 411 |
| 12.7. Резюме | 414 |

ЧАСТЬ 4.

Продвинутые методы 415

Глава 13. Обобщенные линейные модели 417

| | |
|--|-----|
| 13.1. Обобщенные линейные модели и функция <code>glm()</code> | 418 |
| 13.1.1. Функция <code>glm()</code> | 419 |
| 13.1.2. Вспомогательные функции..... | 421 |
| 13.1.3. Соответствие модели данным и регрессионная диагностика..... | 422 |
| 13.2. Логистическая регрессия..... | 423 |
| 13.2.1. Интерпретация параметров модели | 426 |
| 13.2.2. Оценка влияния независимых переменных на вероятность исхода..... | 428 |
| 13.2.3. Избыточная дисперсия..... | 429 |
| 13.2.4. Дополнительные методы | 430 |
| 13.3. Пуассоновская регрессия | 431 |
| 13.3.1. Интерпретация параметров модели | 434 |
| 13.3.2. Избыточная дисперсия..... | 435 |
| 13.3.3. Дополнительные методы | 437 |
| 13.4. Резюме | 440 |

Глава 14. Главные компоненты и факторный анализ 441

| | |
|--|-----|
| 14.1. Выполнение анализа главных компонент и факторного анализа в R..... | 444 |
| 14.2. Главные компоненты | 445 |
| 14.2.1. Выбор необходимого числа компонент | 447 |
| 14.2.2. Выделение главных компонент | 449 |
| 14.2.3. Вращение главных компонент | 453 |
| 14.2.4. Вычисление значений главных компонент | 454 |
| 14.3. Разведочный факторный анализ | 457 |

| | |
|---|-----|
| 14.3.1. Определение числа извлекаемых факторов | 458 |
| 14.3.2. Выделение общих факторов | 460 |
| 14.3.3. Вращение факторов | 461 |
| 14.3.4. Значения факторов | 465 |
| 14.3.5. Другие пакеты для проведения факторного анализа | 466 |
| 14.4. Другие модели для латентных переменных | 466 |
| 14.5. Резюме | 468 |

Глава 15. Продвинутое методы работы

с пропущенными данными 470

| | |
|---|-----|
| 15.1. Этапы работы с пропущенными данными | 472 |
| 15.2. Обнаружение пропущенных значений | 474 |
| 15.3. Исследование структуры пропущенных данных | 475 |
| 15.3.1. Представление пропущенных значений в виде таблицы | 476 |
| 15.3.2. Визуальное исследование структуры пропущенных данных | 477 |
| 15.3.3. Использование корреляции для исследования пропущенных значений | 480 |
| 15.4. Выявление источников пропущенных данных и эффекта от них | 482 |
| 15.5. Рациональный подход | 484 |
| 15.6. Анализ полных строк (построчное удаление) | 485 |
| 15.7. Метод множественного восстановления пропущенных данных | 487 |
| 15.8. Другие подходы к пропущенным данным | 493 |
| 15.8.1. Парное удаление | 494 |
| 15.8.2. Простое (нестохастическое) восстановление данных | 494 |
| 15.9. Резюме | 495 |

Глава 16. Продвинутое графические методы 497

| | |
|--|-----|
| 16.1. Четыре графические системы R | 498 |
| 16.2. Пакет lattice | 499 |
| 16.2.1. Условные переменные | 505 |
| 16.2.2. Функции для изменения формата ячеек | 507 |
| 16.2.3. Группировка переменных | 510 |
| 16.2.4. Графические параметры | 516 |
| 16.2.5. Расположение диаграмм на странице | 517 |
| 16.3. Пакет ggplot2 | 518 |
| 16.4. Интерактивная графика | 524 |
| 16.4.1. Взаимодействие с диаграммами: идентификация точек ... | 525 |
| 16.4.2. Пакет playwith | 525 |
| 16.4.3. Пакет laticist | 527 |
| 16.4.4. Создание интерактивной графики при помощи пакета iplots | 528 |

| | |
|--|------------|
| 16.4.5. Пакет rggobi | 530 |
| 16.5. Резюме | 531 |
| Послесловие: В погоне за кроликом | 533 |
| Приложение А. Графические пользовательские интерфейсы | 537 |
| Приложение В. Настройка начальной конфигурации программы | 541 |
| Приложение С. Экспорт данных из R | 543 |
| С.1. Текстовый файл с разделителями..... | 543 |
| С.2. Таблица Excel..... | 543 |
| С.3. Другие статистические программы | 544 |
| Приложение D. Сохранение результатов в пригодном для публикации качестве | 545 |
| D.1. Подготовка отчета типографского качества при помощи пакета Sweave (R + LaTeX) | 546 |
| D.2. Объединение сил с OpenOffice при помощи пакета odfWeave | 552 |
| D.3. Комментарии..... | 555 |
| Приложение E. Матричная алгебра в R | 556 |
| Приложение F. Пакеты, упомянутые в этой книге ... | 559 |
| Приложение G. Работа с большими наборами данных | 568 |
| G.1. Эффективное программирование | 569 |
| G.2. Хранение данных вне оперативной памяти | 570 |
| G.3. Аналитические пакеты для больших объемов данных | 571 |
| Приложение H. Обновление версии R..... | 572 |
| Список литературы | 576 |



ПРЕДИСЛОВИЕ

Что толку в книжке, если в ней нет ни картинок, ни разговоров?

Алиса. «Алиса в Стране Чудес»¹

Оно чудесно и наделено сокровищами, способными удовлетворить всех от мала до велика, но не предназначено для робких духов.

Кью. Сериал «Звездный путь: следующее поколение»

Когда я начал писать эту книгу, я потратил достаточно много времени в поисках хорошего эпиграфа. В итоге я остановился на этих двух. R – это потрясающе гибкие приложение и язык для исследования, визуализации и понимания данных. Я выбрал цитату из «Алисы в Стране Чудес», чтобы передать суть современного статистического анализа – интерактивного процесса, состоящего из исследования, визуализации и интерпретации.

Вторая цитата отражает широко распространенное мнение о том, что работе в R сложно научиться. Я надеюсь показать вам, что это не обязательно должно быть так. R – мощная программа с таким большим числом доступных аналитических и графических функций (по последним подсчетам их более 50 000), что она может в одинаковой степени навести ужас и на новичков, и на опытных пользователей. Однако в этом мнимом безумии есть поэзия и логика. Вооружившись руководствами и инструкциями, вы сможете ориентироваться в огромном разнообразии возможностей, выбрав те инструменты, которые нужны для того, чтобы уверенно, эффективно и элегантно выполнить вашу задачу.

Я впервые познакомился с R несколько лет назад, когда хотел получить новую должность консультанта по статистике. Предполагаемый работодатель перед интервью спросил меня, владею ли я R. Следуя обычным советам специалистов по подбору персонала, я немедленно сказал «да» и стал учиться работать в этой программе. Я был опытным статистиком и исследователем с 25 годами опыта

1 Перевод Н. Демуровой

программирования в SAS и SPSS, свободно владевшим несколькими языками программирования. Чего же тут может быть сложного? Знаменитые последние слова.

По мере того как я пытался выучить язык программирования (как можно быстрее, ведь день собеседования приближался с угрожающей быстротой), я находил или тома, посвященные глубинной структуре языка или многочисленные трактаты об отдельных продвинутых статистических методах, написанных специалистами в данной области для своих коллег. Встроенная помощь была написана очень лаконично и служила скорее справочником, чем учебным пособием. Каждый раз, когда мне казалось, что я освоил общую логику и возможности R, находилось что-то новое, заставлявшее почувствовать себя невежественным и ничтожным.

При освоении R, я подошел к процессу с точки зрения исследователя, которому нужно обрабатывать данные. Я пытался понять, что нужно сделать, чтобы успешно обработать, проанализировать и понять данные, включая:

- доступ к данным (получение данных из разных источников);
- редактирование данных (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 данных (вычисление описательных статистик для того, чтобы охарактеризовать данные);
- визуализация данных (поскольку картинка на самом деле стоит тысячи слов);
- моделирование данных (нахождение зависимостей и тестирование гипотез);
- оформление результатов (подготовка таблиц и диаграмм достаточного для публикации качества).

Затем я постарался понять, как я могу использовать R, чтобы выполнить каждую из этих задач. Поскольку я лучше всего учусь, обучая других, со временем я создал сайт (www.statmethods.net), на котором рассказал все, что я узнал.

Затем, около года назад, Марьян Бейс (Marjan Base), издатель, позвонила и спросила, не хочу ли я написать книгу про R. К этому времени я уже написал 50 статей в научных журналах, четыре технических руководства, многочисленные главы в книгах и целую книгу по методологии исследования, так чего же тут могло быть сложного? Рискую повториться – знаменитые последние слова.

Книгу, которую вы держите в руках, я мечтал иметь много лет назад. Я постарался написать для вас путеводитель по R, который позволит быстро овладеть всей мощью этой замечательной программы с открытым кодом без разочарования и раздражения, которые пришлось испытать мне. Надеюсь, вам понравится.

P.S. Мне предложили ту должность, но я отказался. Однако знакомство с R развернуло мою карьеру в совершенно неожиданном направлении. Жизнь может быть забавной штукой.



ОБ ЭТОЙ КНИГЕ

Если вы выбрали эту книгу, скорее всего у вас есть какие-то данные, которые нужно собрать в единое целое, преобразовать, исследовать, смоделировать, визуализировать или представить коллегам. Если это так, то R создан для вас! R стал всемирно известным языком программирования для статистического анализа, предсказаний и визуализации данных. В этой программе реализовано множество методов анализа данных, от самых простых до самых сложных и современных.

Эта программа с открытым кодом работает под разными операционными системами, включая Windows, Mac OS X и Linux. Она развивается постоянно, новые методы появляются ежедневно. Кроме того, R поддерживается большим и разнородным сообществом ученых и программистов, которые охотно помогут новичку советами.

Хотя программа R, возможно, больше известна за способность создавать красивые и сложные диаграммы, она может справиться с любой статистической задачей. Базовая версия содержит сотни функций для статистического анализа, управления данными и построения диаграмм. Однако некоторые особенно мощные методы реализованы в дополнительных пакетах, созданных независимыми авторами.

Эта широта возможностей имеет свою цену. Для новичков бывает сложно понять, что такое R и как в ней работать. Даже самые опытные пользователи R с удивлением обнаруживают какие-то возможности, о которых они не подозревали.

«R в действии» представляет собой руководство-путеводитель по R, позволяя в общих чертах ознакомиться с самой программой и ее возможностями. В книге описаны наиболее полезные функции базовой версии и более 90 наиболее часто используемых дополнительных пакетов. На всем протяжении книги акцент делается на практическое применение – на то, чтобы вы, руководствуясь прочитанным, могли проанализировать ваши данные и изложить результаты коллегам. По окончании чтения этой книги вы будете иметь хорошее представление о том, как R работает, и где можно получить дополнительную информацию. Вы научитесь применять разнообразные методы для визуализации

зации данных и обретете достаточно умений, чтобы справиться как с простыми, так и со сложными задачами анализа данных.

Кому следует прочесть эту книгу

Книга «R в действии» предназначена для любого, кто имеет дело с данными. Опыт в статистическом программировании не требуется. Хотя эта книга доступна и новичкам, в ней содержится достаточно нового и полезного материала, чтобы удовлетворить запросы даже опытных специалистов по R.

Пользователи, не владеющие познаниями в области статистики, которые хотят использовать R для управления данными, их обобщения и представления в графическом виде, смогут легко понять главы 1–6, 11 и 16. Главы 7 и 10 подразумевают, что вы прослушали вводный курс статистики, а главы 8, 9 и 12–15 потребуют более глубоких познаний в этой области. Однако я старался написать каждую главу так, чтобы в ней было что-то интересное и полезное и для новичков и для опытных статистиков.

Структура книги

Эта книга создана как путеводитель по программе R, с акцентом на методы, которые можно сразу применить для управления данными, их визуализации и осмысления. Книга состоит из 16 глав, сгруппированных в четыре части: «Начало работы», «Базовые методы», «Методы средней сложности» и «Методы повышенной сложности». Дополнительные темы рассмотрены в восьми приложениях.

Глава 1 начинается с обзора программы в целом и характеристик, которые делают ее столь полезной для обработки данных. В главе рассказано, как установить программу и как расширить ее возможности путем установки доступных в Сети дополнительных пакетов. Оставшаяся часть главы посвящена описанию интерфейса программы и рассказу о том, как запускать ее в интерактивном и пакетном режимах.

В главе 2 описаны многие методы импорта данных в программу. Первая половина главы посвящена характеристике типов данных в R и тому, как вводить данные с клавиатуры. Во второй половине главы обсуждаются способы импорта данных из текстовых файлов, веб-страниц, электронных таблиц, других статистических программ и баз данных.

Многие пользователи изначально выбирают R потому, что они хотят создавать диаграммы, так что мы сразу переходим к этой теме в

главе 3. Вам не понадобится долго ждать. Мы обсуждаем, как создавать диаграммы, изменять их и сохранять в разных форматах.

Глава 4 посвящена основам управления данными, включая сортировку, объединение и разбиение наборов данных, а также преобразование, перекодировку и удаление переменных.

Глава 5 основана на главе 4 и содержит описание функций (математических, статистических, текстовых) и управляющих конструкций (циклы, выполнение при условии) для управления данными. Затем мы обсуждаем, как написать вашу собственную функцию в R и как сгруппировать данные различными способами.

В главе 6 рассказано, как создавать наиболее распространенные одномерные диаграммы, такие как столбчатая и круговая диаграмма, диаграмма распределения плотности, диаграмма размахов («ящик с усами») и точечная диаграмма. Все эти диаграммы полезны для изучения характера распределения значений одной переменной.

Глава 7 начинается с описания того, как находить общие характеристики данных, включая использование описательных статистик и сводных таблиц. Затем мы рассматриваем основные способы изучения взаимосвязи между двумя переменными, включая корреляцию, тест Стьюдента, тест хи-квадрат и непараметрические методы.

Глава 8 посвящена применению регрессионных методов для моделирования взаимосвязи между числовой переменной-откликом (outcome variable) и набором из одной или нескольких независимых переменных (predictor variables). Подробно рассмотрены методы подгонки этих моделей, оценки их адекватности и интерпретации их значений.

В главе 9 рассмотрены основные типы планов экспериментов при дисперсионном анализе и его разновидностях. В этой ситуации нас обычно интересует, как комбинации разных типов воздействия или разных условий влияют на числовую переменную-отклик. Также описаны методы оценки адекватности анализа и визуализации результатов.

Детальное описание анализа мощности статистических тестов – предмет главы 10. Она начинается с обсуждения проблемы проверки гипотез; далее описано, как определить объем выборки, необходимый для выявления эффекта заданной величины при заданном уровне достоверности. Это поможет вам повысить вероятность достижения желаемого результата при планировании экспериментов.

Глава 11 – это продолжение главы 5. В ней рассказано, как создать диаграммы для визуализации связей между двумя и более переменными. Обсуждаются разные типы двух- и трехмерных диаграмм

рассеяния, матриц диаграмм рассеяния, графиков, коррелограмм и мозаичных диаграмм.

В главе 12 представлены аналитические методы, которые хорошо работают, когда данные происходят из неизвестных или смешанных типов распределения, когда размеры выборок малы, когда выбросы представляют собой проблему или когда разработка статистического теста на основании наблюдаемого распределения слишком сложна. Это метод повторной выборки (resampling) и бутстреп-анализ (bootstrapping) – подходы, требующие большого объема вычислений и легко реализуемые в R.

Глава 13 описывает, как применять регрессионный анализ, рассмотренный в главе 8, к данным с распределением, отличным от нормального. Глава начинается с описания обобщенных линейных моделей. Затем более подробно рассматриваются случаи, когда нужно предсказать переменную-отклик, представленную либо категориальными (логистическая регрессия), либо счетными данными (пуассоновская регрессия).

Одна из сложностей, связанных с многомерными данными – это проблема снижения их размерности. В главе 14 описаны методы, с помощью которых большое число коррелирующих друг с другом переменных преобразуется в меньший набор независимых переменных (анализ главных компонент), а также методы обнаружения скрытой структуры в имеющемся наборе переменных (факторный анализ). Детально разобраны многочисленные этапы этих типов анализа.

В соответствии с нашим намерением описать актуальные методы анализа данных глава 15 посвящена современным подходам к решению распространенной проблемы пропущенных значений в данных. В R реализованы разнообразные изящные подходы к анализу неполных в силу разных причин данных. Здесь описаны лучшие из этих методов, вместе с разъяснениями, когда сто их применять каждый из них, а какие – лучше избегать.

Глава 16 завершает обсуждение диаграмм рассмотрением некоторых наиболее сложных и полезных методов визуализации данных. Рассмотрена визуализация очень сложных данных с использованием панельной (или категоризированной) графики, даны основные сведения о новом пакете ggplot2, также кратко описаны способы работы с диаграммами в режиме реального времени.

В послесловии перечислены многие из лучших сайтов, которые следует посетить, чтобы научиться работать в R, влиться в сообщество пользователей R, получить ответы на возникшие вопросы и от-

слеживать изменения в этом стремительно развивающемся программном продукте.

И, последнее, но не менее важное, восемь приложений (от А до Н) содержат дополнительные сведения по таким полезным темам, как пользовательский интерфейс, настройка и обновление программы, экспорт данных, получение результатов высокого полиграфического качества, использование R для матричной алгебры (по образцу MATLAB) и работа с большими объемами данных.

Примеры

Для того чтобы сделать книгу настолько широко применимой, насколько возможно, я выбрал примеры из разных областей знаний, включая психологию, социологию, медицину, биологию, бизнес и технические науки. Ни один из примеров не требует специальных знаний в соответствующей области.

Наборы данных, используемые в этих примерах, были выбраны потому, что они позволяют формулировать интересные вопросы и имеют небольшой размер. Это позволяет сосредоточиться на рассматриваемом методе и быстро понять происходящее. Когда учишься новым методам, меньше – значит лучше.

Наборы данных либо поставляются с базовой версией R, либо доступны в составе дополнительных пакетов, которые можно скачать из Интернета. Программный код для каждого примера размещен на сайте <http://www.manning.com/RinAction>. Для получения максимальной отдачи от этой книги я рекомендую выполнять примеры по ходу их прочтения.

В заключение нужно вспомнить известную сентенцию, которая гласит, что если спросить двух статистиков, как анализировать определенный набор данных, получишь три разных ответа. Можно понимать этот афоризм по-разному – ведь каждый ответ приблизит вас к пониманию данных. Я не утверждаю, что предлагаемый мной тот или иной способ анализа данных – лучший или единственный путь к решению конкретной задачи. Я предлагаю вам применить разные подходы к данным, используя знания, приобретенные во время чтения книги, и посмотреть, что вы сможете узнать. R – интерактивная программа, и лучший способ чему-то научиться в ней – это экспериментировать.

Принятые обозначения

В книге использованы следующие типографские обозначения:

- моноширинный шрифт использован для программного кода, который нужно вводить именно так, как указано в книге;

- моноширинный шрифт также использован внутри основного текста для обозначения фрагментов кода или ранее упомянутых объектов;
- *курсив* внутри программного кода – это указатель места заполнения. Его следует заменять подходящим текстом или значениями, соответствующими вашей задаче. Например, *путь_к_моему_файлу* должен быть заменен указанием пути к реальному файлу на вашем компьютере;
- R – это интерактивный язык, который информирует пользователя о готовности принять команду приглашением (> по умолчанию). Многие фрагменты программного кода в книге скопированы из интерактивных сессий. Если вы видите строки кода, которые начинаются с >, не набирайте этот символ приглашения к вводу команды;
- пояснения к программному коду приведены в виде внутритекстовых комментариев. В дополнение к этому некоторые пояснения обозначены нумерованными кружками, такими как ❶, которые отсылают к объяснению ниже по тексту;
- для того чтобы сэкономить место или сделать текст более понятным, мы иногда добавляли в вывод результатов интерактивных сессий дополнительные пробелы или удаляли текст, который напрямую не относился к обсуждаемой теме.

Об авторе

Доктор наук Роберт Кабаков – вице-президент по исследовательской работе (Vice President of Research) в Группе Исследований Менеджмента (Management Research Group – MRG), международной фирме, специализирующейся на организационном развитии и консалтинге. У него за спиной более 20 лет опыта в сфере исследовательских и статистических консультаций в областях заботы о здоровье, финансовых операций, производства, бихевиоризма, управления и академической науки. Прежде чем присоединиться к MRG, Р. Кабаков был профессором психологии в Юго-Восточном Университете Нова (Nova Southeastern University) во Флориде, где он преподавал количественные методы и статистическое программирование в магистратуре. В последние два года он поддерживает сайт Quick-R – учебное пособие по R.