Оглавление

Предисловие 7

Глава 1. Что такое данные и зачем их обрабатывать? . . . 10
1.1. Откуда берутся данные 10
1.2. Генеральная совокупность и выборка 12
1.3. Как получать данные 13
1.4. Что ищут в данных 17

Глава 2. Как обрабатывать данные 21
2.1. Неспециализированные программы 21
2.2. Специализированные статистические программы 22
2.2.1. Оконно-кнопочные системы 22
2.2.2. Статистические среды 24
2.3. Из истории S и R 24
2.4. Применение, преимущества и недостатки R 25
2.5. Как скачать и установить R 27
2.6. Как начать работать в R 28
2.6.1. Запуск 28
2.6.2. Первые шаги 29
2.7. R и работа с данными: вид снаружи 30
2.7.1. Как загружать данные 30
2.7.2. Как сохранять результаты 36
2.7.3. R как калькулятор 37
2.7.4. Графики 38
2.7.5. Графические устройства 40
2.7.6. Графические опции 42
2.7.7. Интерактивная графика 43

Глава 3. Типы данных 46
3.1. Градусы, часы и километры: интервальные данные . . . 46
3.2. «Садись, двойка»: шкальные данные 49
3.3. Красный, желтый, зеленый: номинальные данные 50
3.4. Доли, счет и ранги: вторичные данные 55
3.5. Пропущенные данные 59
3.6. Выбросы и как их найти 61

3.7. Меняем данные: основные принципы преобразования . . 61
3.8. Матрицы, списки и таблицы данных 63
3.8.1. Матрицы 63
3.8.2. Списки 65
3.8.3. Таблицы данных 68

Глава 4. Великое в малом: одномерные данные 72
4.1. Как оценивать общую тенденцию 72
4.2. Ошибочные данные 82
4.3. Одномерные статистические тесты 83
4.4. Как создавать свои функции 87
4.5. Всегда ли точны проценты 90

Глава 5. Анализ связей: двумерные данные 94
5.1. Что такое статистический тест 94
5.1.1. Статистические гипотезы 94
5.1.2. Статистические ошибки 95
5.2. Есть ли различие, или Тестирование двух выборок . . . 96
5.3. Есть ли соответствие, или Анализ таблиц 102
5.4. Есть ли взаимосвязь, или Анализ корреляций 109
5.5. Какая связь, или Регрессионный анализ 114
5.6. Вероятность успеха, или Логистическая регрессия . . . 123
5.7. Если выборок больше двух 127

Глава 6. Анализ структуры: data mining 142
6.1. Рисуем многомерные данные 142
6.1.1. Диаграммы рассеяния 143
6.1.2. Пиктограммы 146
6.2. Тени многомерных облаков: анализ главных компонент 149
6.3. Классификация без обучения,
или Кластерный анализ 155
6.4. Классификация с обучением,
или Дискриминантный анализ 164

Глава 7. Узнаем будущее: анализ временных рядов 173
7.1. Что такое временные ряды 173
7.2. Тренд и период колебаний 173
7.3. Построение временного ряда 174
7.4. Прогноз 181

Глава 8. Статистическая разведка 190
8.1. Первичная обработка данных 190
8.2. Окончательная обработка данных 190

8.3. Отчет 191

Приложение А. Пример работы в R 196

Приложение Б. Графический интерфейс (GUI) для R . . 207
Б.1. R Commander 207
Б.2. RStudio 209
Б.3. RKWard 211
Б.4. Revolution-R 211
Б.5. JGR 214
Б.6. Rattle 215
Б.7. `rpanel` 216
Б.8. ESS и другие IDE 218

Приложение В. Основы программирования в R 220
В.1. Базовые объекты языка R 220
В.1.1. Вектор 220
В.1.2. Список 221
В.1.3. Матрица и многомерная матрица 222
В.1.4. Факторы 223
В.1.5. Таблица данных 224
В.1.6. Выражение 224
В.2. Операторы доступа к данным 225
В.2.1. Оператор `[` с положительным аргументом 225
В.2.2. Оператор `[` с отрицательным аргументом 226
В.2.3. Оператор `[` со строковым аргументом 226
В.2.4. Оператор `[` с логическим аргументом 227
В.2.5. Оператор `$` 227
В.2.6. Оператор `[[` 228
В.2.7. Доступ к табличным данным 229
В.2.8. Пустые индексы 231
В.3. Функции и аргументы 231
В.4. Циклы и условные операторы 234
В.5. R как СУБД 235
В.6. Правила переписывания. Векторизация 238
В.7. Отладка 243
В.8. Элементы объектно-ориентированного программирования в R 246

Приложение Г. Выдержки из документации R 249
Г.1. Среда R 249
Г.2. R и S 250
Г.3. R и статистика 250

Г.4. Получение помощи 250
Г.5. Команды R 251
Г.6. Повтор и коррекция предыдущих команд 252
Г.7. Сохранение данных и удаление объектов 252
Г.8. Внешнее произведение двух матриц 253
Г.9. `c()` 254
Г.10. Присоединение 254
Г.11. `scan()` 255
Г.12. R как набор статистических таблиц 256
Г.13. Область действия 256
Г.14. Настройка окружения 260
Г.15. Графические функции 261
Г.15.1. `plot()` 262
Г.15.2. Отображение многомерных данных 263
Г.15.3. Другие графические функции высокого уровня 264
Г.15.4. Параметры функций высокого уровня 265
Г.15.5. Низкоуровневые графические команды 266
Г.15.6. Математические формулы 269
Г.15.7. Интерактивная графика 269
Г.15.8. `par()` 270
Г.15.9. Список графических параметров 272
Г.15.10. Края рисунка 275
Г.15.11. Составные изображения 276
Г.15.12. Устройства вывода 277
Г.15.13. Несколько устройств вывода одновременно 278
Г.16. Пакеты 279
Г.16.1. Стандартные и сторонние пакеты 280
Г.16.2. Пространство имен пакета 280

Приложение Д. Краткий словарь языка R 282

Приложение Е. Краткий словарь терминов 285

Литература 291

Об авторах 293

Предисловие

Эта книга написана для тех, кто хочет научиться обрабатывать данные. Такая задача возникает очень часто, особенно тогда, когда нужно выяснить ранее неизвестный факт. Например: есть ли эффект от нового лекарства? Или: различаются ли рейтинги двух политиков? Или: как будет меняться курс доллара на следующей неделе?

Многие люди думают, что этот неизвестный факт можно выяснить, если просто немного подумать над данными. К сожалению, часто это совершенно не так. Например, по опросу 262 человек, выходящих с избирательных участков, выяснилось, что 52% проголосовало за кандидата А, а 48% — за кандидата Б (естественно, мы упрощаем ситуацию). Значит ли это, что кандидат А победил? Подумав, многие сначала скажут «Да», а через некоторое время, возможно, «Кто его знает». Но есть простой (с точки зрения современных компьютерных программ) «тест пропорций», который позволяет не только ответить на вопрос (в данном случае «Нет»), но и вычислить, сколько надо было опросить человек, чтобы можно было бы ответить на такой вопрос. В описанном случае это примерно 5000 человек (см. объяснение в конце главы про одномерные данные)!

В общем, если бы люди знали, что можно сделать методами анализа данных, ошибок и неясностей в нашей жизни стало бы гораздо меньше. К сожалению, ситуация в этой области далека от благополучия. Тем из нас, кто заканчивал институты, часто читали курс «Теория вероятностей и математическая статистика», однако кроме ужаса и/или тоски от длинных математических формул, набитых греческими буквами, большинство ничего из этих курсов не помнит. А ведь на теории вероятностей основаны большинство методов анализа данных! С другой стороны, ведь совсем не обязательно знать радиофизику для того, чтобы слушать любимую радиостанцию по радиоприемнику. Значит, для того чтобы анализировать данные в практических целях, не обязательно свободно владеть математической статистикой и теорией вероятностей. Эту проблему давно уже почувствовали многие английские и американские авторы — названиями типа «Статистика без слез» пестрят книжные полки магазинов, посвященные книгам по анализу данных.

Тут, правда, следует быть осторожным как авторам, так и читателям таких книг: многие методы анализа данных имеют, если можно так

выразиться, двойное дно. Их (эти методы) можно применять, глубоко не вникая в сущность используемой там математики, получать результаты и обсуждать эти результаты в отчетах. Однако в один далеко не прекрасный день может выясниться, что данный метод совершенно не подходил для ваших данных, и поэтому полученные результаты и результатами-то назвать нельзя... В общем, будьте бдительны, внимательно читайте про все *ограничения* методов анализа, а при чтении примеров досконально сравнивайте их со своими данными.

Про примеры: мы постарались привести как можно больше примеров, как простых, так и сложных, и по возможности из разных областей жизни, поскольку читателями этой книги могут быть люди самых разных профессий. Еще мы попробовали снизить объем теоретического материала, потому что мы знаем — очень многие учатся только на примерах. Поскольку книга посвящена такой компьютерной программе, которая «работает на текстовом коде», логично было поместить эти самые коды в текстовый файл, а сам файл сделать общедоступным. Так мы и поступили — приведенные в книге примеры можно найти на веб-странице по адресу `http://ashipunov.info/shipunov/software/r/`. Там же находятся разные полезные ссылки и те файлы данных, которые не поставляются вместе с программой.

О структуре книги: первая глава, по сути, целиком теоретическая. Если лень читать общие рассуждения, можно сразу переходить ко второй главе. Однако в первой главе есть много такой информации, которая позволит в будущем не «наступать на грабли». В общем, решайте сами. Во второй главе самые важные — разделы, начиная с «Как скачать и установить R», в которых объясняется, как работать с программой R. Если не усвоить этих разделов, все остальное чтение будет почти бесполезным. Советуем внимательно прочитать и обязательно *проработать все примеры* из этого раздела. Последующие главы составляют ядро книги, там рассказывается про самые распространенные методы анализа данных. Глава «Статистическая разведка», в которой обсуждается общий порядок статистического анализа, подытоживает книгу; в ней еще раз рассказывается про методы, обсуждавшиеся в предыдущих главах. В приложениях к книге содержится много полезной информации: там рассказано о графических интерфейсах к R, приведен простой практический пример работы, описаны основы программирования в R, приведены выдержки из перевода официальной документации. По сути, каждое приложение — это отдельный небольшой справочник, который можно использовать более или менее независимо от остальной книги.

Конечно, множество статистических методов, в том числе и довольно популярных, в книгу не вошли. Мы почти не касаемся статистических моделей, ничего не пишем о контрастах, не рассказываем о стандартных распределениях (за исключением нормального), не объясняем,

как делать многофакторный и блочный дисперсионный анализ, планировать эксперимент, эффектах, кривых выживания, байесовых методах, факторном анализе, геостатистике и т. д., и т. п. Наша цель — научить основам статистического анализа. А если читатель хорошо освоит основы, то любой продвинутый метод он сможет одолеть без особого труда, опираясь на литературу, встроенную справку и Интернет.

Несколько технических замечаний: все десятичные дроби в книге представлены в виде чисел с разделителем-точкой (типа `10.4`), а не запятой (типа `10,4`). Это сделано потому, что программа R по умолчанию «понимает» только первый вариант дробей. И еще: многие приведенные в книге примеры можно (и нужно!) повторить самостоятельно. Такие примеры напечатаны **машинописным шрифтом** и начинаются со значка «больше» — «`>`». Если пример не умещается на одной строке, все последующие его строки начинаются со знака «плюс» — «`+`» (не набирайте эти знаки, когда будете выполнять примеры!). Если в книге идет речь о загрузке файлов данных, то предполагается, что все они находятся в поддиректории `data` в текущей директории. Если вы будете скачивать файлы данных с упомянутого выше сайта, не забудьте создать эту поддиректорию и скопировать туда файлы данных.

Глава 1

Что такое данные и зачем их обрабатывать?

В этой главе рассказывается о самых общих понятиях анализа данных.

1.1. Откуда берутся данные

«Без пруда не выловишь и рыбку из него»,— говорит народная мудрость. Действительно, если хочешь анализировать данные, надо их сначала получить. Способов получения данных много, а самые главные — *наблюдения* и *эксперименты*.

Наблюдением будем называть такой способ получения данных, при котором воздействие наблюдателя на наблюдаемый объект сведено к минимуму. *Эксперимент* тоже включает наблюдение, но сначала на наблюдаемый объект оказывается заранее рассчитанное воздействие. Для наблюдения очень важно это «сведение воздействия к минимуму». Если этого не сделать, мы получим данные, отражающие не свойства объекта, а его реакцию на наше воздействие.

Вот, например, встала задача исследовать, чем питается какое-то редкое животное. Оптимальная стратегия наблюдения здесь состоит в установке скрытых камер во всех местах, где это животное обитает. После этого останется только обработать снятое, чтобы определить вид пищи. Очень часто, однако, оптимальное решение совершенно невыполнимо, и тогда пытаются обойтись, скажем, наблюдением за животным в зоопарке. Ясно, что в последнем случае на объект оказывается воздействие, и немалое. В самом деле, животное поймали, привезли в совершенно нетипичные для него условия, да и корм, скорее всего, будет непохож на тот, каким оно питалось на родине. В общем, если наблюдения в зоопарке поставлены грамотно, то выяснено будет не то, чем вообще питается данное животное, а то, чем оно питается при содержании в определенном зоопарке. К сожалению, многие (и исследователи, и те, кто потом читает их отчеты) часто не видят разницы между этими двумя утверждениями.

Вернемся к примеру из предисловия. Предположим, мы опрашиваем выходящих с избирательных участков. Часть людей, конечно, вообще окажется отвечать. Часть ответит что-нибудь, не относящееся к делу. Часть вполне может намеренно или случайно исказить свой ответ. Часть ответит правду. И все это серьезным образом зависит от наблюдателя — человека, проводящего опрос.

Даже упомянутые выше скрытые камеры приведут к определенному воздействию: они же скрытые, но не невидимые и невесомые. Нет никакой гарантии, что наше животное или его добыча не отреагирует на них. А кто будет ставить камеры? Если это люди, то чем больше камер поставить, тем сильнее будет воздействие на окружающую среду. Сбрасывать с вертолета? Надеемся, что вам понятно, к чему это может привести.

В общем, из сказанного должно быть понятно, что наблюдение «в чистом виде» более или менее неосуществимо, поскольку всегда будет внесено какое-нибудь воздействие. Поэтому для того, чтобы адекватно работать с данными наблюдений, надо всегда четко представлять, как они проводились. Если воздействие было значительным, то надо представлять (хотя бы теоретически), какие оно могло повлечь изменения, а в отчете обязательно указать на те ограничения, которые были вызваны способом наблюдения. Не следует без необходимости применять экстраполяцию: если мы увидели, что А делает Б, нельзя писать «А всегда делает Б» и даже «А обычно делает Б». Можно лишь писать нечто вроде «в наших наблюдениях А делал Б, это позволяет с некоторой вероятностью предположить, что он может делать Б».

У эксперимента свои проблемы. Наиболее общие из них — это точный учет воздействия и наличие контроля. Например, мы исследуем действие нового лекарства. Классический эксперимент состоит в том, что выбираются две группы больных (*как* выбрать такие группы, *сколько* должно быть человек и прочее, рассмотрено дальше). Всем больным сообщают, что проводится исследование нового лекарства, но его дают только больным первой группы, остальные получают так называемое *плацебо*, внешне неотличимое от настоящего лекарства, но не содержащее ничего лекарственного. Зачем это делается? Дело в том, что если больной будет знать, что ему дают «ненастоящее» лекарство, то это скажется на эффективности лечения, потому что результат зависит не только от того, что больной пьет, но и от того, что он чувствует. Иными словами, психологическое состояние больного — это дополнительный фактор воздействия, от которого в эксперименте лучше избавиться. Очень часто не только больным, но и их врачам не сообщают, кому дают плацебо, а кому — настоящее лекарство (двойной слепой метод). Это позволяет гарантировать, что и психологическое состояние врача не повлияет на исход лечения.

Группа, которой дают плацебо (она называется *контроль*), нужна для того, чтобы отделить эффект, который может произвести лекарство, от эффекта какого-нибудь постороннего внешнего фактора. Известно, например, что уменьшение длины светового дня осенью и зимой провоцирует многие нервные заболевания. Если наше исследование придется на это время и у нас не будет контроля, то увеличение частоты заболеваний мы вполне можем принять за результат применения лекарства.

1.2. Генеральная совокупность и выборка

«Статистика знает все»,— писали Ильф и Петров в «Двенадцати стульях», имея в виду то, что обычно называют статистикой,— сбор всевозможной информации обо всем на свете. Чем полнее собрана информация, тем, как считается, лучше. Однако лучше ли?

Возьмем простой пример. Допустим, фирма-производитель решила выяснить, какой из двух сортов производимого мороженого предпочитают покупатели. Проблем бы не было, если бы все мороженое продавалось в одном магазине. На самом же деле продавцов очень много: это оптовые рынки и гипермаркеты, средние и малые магазины, киоски, отдельные мороженщики с тележками, те, кто торгует в пригородных поездах, и т. п. Можно попробовать учесть доход от продажи каждого из двух сортов. Если они сто́ят одинаково, то бо́льшая сумма дохода должна отразить больший спрос. Представим, однако, что спрос одинаков, но по каким-то причинам мороженое первого сорта тает быстрее. Тогда потерь при его транспортировке будет в среднем больше, продавцы будут покупать его несколько чаще, и получится, что доход от продажи первого сорта будет несколько выше, чем от второго. Это рассуждение, конечно, упрощает реальную ситуацию, но подумайте, сколько других неучтенных факторов стоит на пути такого способа подсчета! Анализ товарных чеков получше, однако многие конечные продавцы таких чеков не имеют и поэтому в анализ не попадут. А нам-то необходимо как раз учесть спрос покупателей, а не промежуточных продавцов.

Можно поступить иначе — раздать всем конечным продавцам анкеты, в которых попросить указать, сколько какого мороженого продано; а чтобы анкеты были обязательно заполнены, вести с этими продавцами дела только при наличии заполненных анкет. Только ведь никто не будет контролировать, *как* продавцы заполняют анкеты... Вот и получит фирма большую, подробную сводную таблицу о продажах мороженого, которая ровным счетом ничего отражать не будет.

Как же поступить? Здесь на помощь приходит идея *выборочных исследований*. Всех продавцов не проконтролируешь, но ведь нескольких-

то можно! Надо выбрать из общего множества несколько торговых точек (*как* выбирать — это особая наука, об этом ниже) и проконтролировать тамошние продажи силами самой фирмы или такими нанятыми людьми, которым можно доверять. В итоге мы получим результат, который является частью общей картины. Теперь самый главный вопрос: можно ли этот результат распространить на всю совокупность продаж? Оказывается, можно, поскольку на основе теории вероятностей уже много лет назад была создана *теория выборочных исследований*. Ее-то и называют чаще всего математической статистикой, или просто статистикой.

Пример с мороженым показывает важную вещь: выборочные исследования могут быть (и часто бывают) значительно более точными (в смысле соответствия реальности), чем сплошные.

Еще один хороший пример на эту же тему есть в результатах сплошной переписи населения России 1897 г. Если рассмотреть численность населения по возрастам, то получается, что максимальные численности (пики) имеют возрасты, кратные 5 и в особенности кратные 10. Понятно, как это получилось. Бóльшая часть населения в те времена была неграмотна и свой возраст помнила только приблизительно, с точностью до пяти или до десяти лет. Чтобы все-таки узнать, каково было распределение по возрастам на самом деле, нужно не увеличивать объем данных, а наоборот, создать выборку из нескольких процентов населения и провести комплексное исследование, основанное на перекрестном анализе нескольких источников: документов, свидетельств и личных показаний. Это даст гораздо более точную картину, нежели сплошная перепись.

Естественно, сам процесс создания выборки может являться источником ошибок. Их принято называть «ошибками репрезентативности». Однако правильная организация выборки позволяет их избежать. А поскольку с выборкой можно проводить гораздо более сложные исследования, чем со всеми данными (их называют *генеральной совокупностью*, или *популяцией*), те ошибки (ошибки точности), которые возникают при сплошном исследовании, в выборочном исследовании можно исключить.

1.3. Как получать данные

В предыдущих разделах неоднократно упоминалось, что от правильного подбора выборки серьезным образом будет зависеть качество получаемых данных. Собственно говоря, есть два основных принципа составления выборки: *повторность* и *рандомизация*. Повторности нужны для того, чтобы быть более уверенными в полученных результатах,

а рандомизация — для того, чтобы избежать отклонений, вызванных посторонними причинами.

Принцип повторностей предполагает, что один и тот же эффект будет исследован несколько раз. Собственно говоря, для этого мы в предыдущих примерах опрашивали *множество* избирателей, ловили в заповеднике *много* животных, подбирали группы из *нескольких десятков* больных и контролировали *различных* продавцов мороженого. Нужда в повторностях возникает оттого, что все объекты (даже только что изготовленные на фабрике изделия) пусть в мелочах, но отличаются друг от друга. Эти отличия способны затуманить общую картину, если мы станем изучать объекты поодиночке. И наоборот, если мы берем несколько объектов сразу, их различия часто «взаимно уничтожаются».

Не стоит думать, что создать повторности — простое дело. К сожалению, часто именно небрежное отношение к повторностям сводит на нет результаты вроде бы безупречных исследований. Главное правило — *повторности должны быть независимы друг от друга.* Это значит, например, что нельзя в качестве повторностей рассматривать данные, полученные в последовательные промежутки времени с одного и того же объекта или с одного и того же места. Предположим, что мы хотим определить размер лягушек какого-то вида. Для этого с интервалом в 15 минут (чтобы лягушки успокоились) ловим сачком по одной лягушке. Как только наберется двадцать лягушек, мы их меряем и вычисляем средний размер. Однако такое исследование не будет удовлетворять правилу независимости, потому что каждый отлов окажет влияние на последующее поведение лягушек (например, к концу лова будут попадаться самые смелые, или, наоборот, самые глупые). Еще хуже использовать в качестве повторностей последовательные наблюдения за объектом. Например, в некотором опыте выясняли скорость зрительной реакции, показывая человеку на доли секунды предмет, а затем спрашивая, что это было. Всего исследовали 10 человек, причем каждому показывали предмет пять раз. Авторы опыта посчитали, что у них было 50 повторностей, однако на самом деле — только десять. Это произошло потому, что каждый следующий показ не был независим от предыдущего (человек мог, например, научиться лучше распознавать предмет).

Надо быть осторожным не только с данными, собранными в последовательные промежутки времени, но и просто с данными, собранными с одного и того же места. Например, если мы определяем качество телевизоров, сходящих с конвейера, не годится в качестве выборки брать несколько штук подряд — с большой вероятностью они изготовлены в более близких условиях, чем телевизоры, взятые порознь, и, стало быть, их характеристики не независимы друг от друга.

Второй важный вопрос о повторностях: сколько надо собрать данных. Есть громадная литература по этому поводу, но ответа, в общем, два: (1) чем больше, тем лучше и (2) 30. Выглядящее несколько юмористически «правило 30» освящено десятилетиями опытной работы. Считается, что выборки, меньшие 30, следует называть малыми, а бóльшие — большими. Отсюда то значение, которое придают числу тридцать в анализе данных. Бывает так, что и тридцати собрать нельзя, однако огорчаться этому не стóит, поскольку многие процедуры анализа данных способны работать с очень малыми выборками, в том числе из пяти и даже из трех повторностей. Следует, однако, иметь в виду, что чем меньше повторностей, тем менее надежными будут выводы. Существуют, кроме того, специальные методы, которые позволяют посчитать, сколько надо собрать данных, для того чтобы с определенной вероятностью высказать некоторое утверждение. Это так называемые «тесты мощности» (см. пример в главе про одномерные данные).

Рандомизация — еще одно условие создания выборки, и также «с подвохом». Каждый объект генеральной совокупности должен иметь равные шансы попасть в выборку. Очень часто исследователи полагают, что выбрали свои объекты случайно (проделали рандомизацию), в то время как на самом деле их материал был подобран иначе. Предположим, нам поручено случайным образом отобрать сто деревьев в лесу, чтобы впоследствии померить степень накопления тяжелых металлов в листьях. Как мы будем выбирать деревья? Если просто ходить по лесу и собирать листья с разных деревьев, с большой вероятностью они не будут собранными случайно, потому что вольно или невольно мы будем собирать листья, чем-то привлекшие внимание (необычностью, окраской, доступностью). Этот метод, стало быть, не годится. Возьмем метод посложнее — для этого нужна карта леса с размеченными координатами. Мы выбираем случайным образом два числа, например 123 м к западу и 15 м к югу от точки, находящейся примерно посередине леса, затем высчитываем это расстояние на местности и выбираем дерево, которое ближе всего к нужному месту. Будет ли такое дерево выбрано случайно? Оказывается, нет. Ведь деревья разных пород растут неодинаково, поэтому у деревьев, растущих теснее (например, у елок), шанс быть выбранными окажется больше, чем у разреженно растущих дубов. Важным условием рандомизации, таким образом, является то, что *каждый объект должен иметь абсолютно те же самые шансы быть выбранным, что и все прочие объекты.*

Как же быть? Надо просто перенумеровать все деревья, а затем выбрать сто номеров по жребию. Но это только звучит просто, а попробуйте так сделать! А если надо сравнить 20 различных лесов?.. В общем, требование рандомизации часто оборачивается весьма серьезными затратами на проведение исследования. Естественно поэтому, что нередко

рандомизацию осуществляют лишь частично. Например, в нашем случае можно случайно выбрать направление, протянуть в этом направлении бечевку через весь лес, а затем посчитать, скольких деревьев касается бечевка, и выбрать каждое энное (пятое, пятнадцатое...) дерево, так чтобы всего в выборке оказалось 100 деревьев. Заметьте, что в данном случае метод рандомизации состоит в том, чтобы внести в исследуемую среду *такой порядок, которого там заведомо нет.* Конечно, у этого последнего метода есть недостатки, а какие — **попробуйте догадаться** сами (ответ см. в конце главы).

Теперь вы знаете достаточно, чтобы ответить на еще один вопрос. В одной лаборатории изучали эффективность действия ядохимикатов на жуков-долгоносиков (их еще называют «слоники»). Для этого химикат наносили на фильтровальную бумагу, а бумагу помещали в стеклянную чашку с крышкой (чашку Петри). Жуков выбирали из банки, в которой их разводили для опытов, очень простым способом: банку с жуками открывали, и первого выползшего на край жука пересаживали в чашку с ядохимикатом. Затем засекали, сколько пройдет времени от посадки жука в банку до его гибели. Потом брали другого жука и так повторяли 30 раз. Потом меняли ядохимикат и начинали опыт сначала. Но однажды один умный человек заметил, что в этом эксперименте самым сильным всегда оказывался тот химикат, который был взят для исследования первым. Как вы думаете, в чем тут дело? Какие нарушения принципов повторности и рандомизации были допущены? Как надо было поставить этот опыт? (См. ответ в конце главы).

Для рандомизации, конечно, существует предел. Если мы хотим выяснить возрастной состав посетителей какого-то магазина, не нужно во имя рандомизации опрашивать прохожих с улицы. Нужно четко представлять себе генеральную совокупность, с которой идет работа, и не выходить за ее границы. Помните пример с питанием животного? Если генеральная совокупность — это животные данного вида, содержащиеся *в зоопарках*, нет смысла добавлять к исследованию данные о питании этих животных *в домашних условиях.* Если же такие данные просто необходимо добавить (например, потому что данных из зоопарков очень мало), то тогда генеральная совокупность будет называться «множество животных данного вида, содержащихся *в неволе*».

Интересный вариант рандомизации используют, когда в эксперименте исследуются одновременно несколько взаимодействий. Например, мы хотим выяснить эффективность разных типов средств против обледенения тротуаров. Для этого логично выбрать (случайным образом) несколько разных (по возрасту застройки, плотности населения, расположению) участков города и внутри каждого участка случайным образом распределить разные типы этих средств. Потом можно, например, фиксировать (в баллах или как-нибудь еще) состояние тротуаров

каждый день после нанесения средства, можно также повторить опыт при разной погоде. Такой подход называется «блочный дизайн». Блоками здесь являются разные участки города, а повторность обеспечивается тем, что в каждом блоке повторяются одни и те же воздействия. При этом даже не обязательно повторять однотипные воздействия по нескольку раз внутри блоков, важно выбрать побольше отличающихся друг от друга блоков. Можно считать разными блоками и разные погодные условия, и тогда у нас получится «вложенный блочный дизайн»: в каждый погодный блок войдет несколько «городских» блоков, и уже внутри этих блоков будут повторены все возможные воздействия (типы средств).

В области рандомизации лежит еще одно коренное различие между наблюдением и экспериментом. Допустим, мы изучаем эффективность действия какого-то лекарства. Вместо того, чтобы подбирать две группы больных, использовать плацебо и т. п., можно просто порыться в архивах и подобрать соответствующие примеры (30 случаев применения лекарства и 30 случаев неприменения), а затем проанализировать разницу между группами (например, число смертей в первый год после окончания лечения). Однако такой подход сопряжен с опасностью того, что на наши выводы окажет влияние какой-то (или какие-то) неучтенный фактор, выяснить наличие которого из архивов невозможно. Мы просто не можем гарантировать, что соблюдали рандомизацию, анализируя архивные данные. К примеру, первая группа (случайно!) окажется состоящей почти целиком из пожилых людей, а вторая — из людей среднего возраста. Ясно, что это окажет воздействие на выводы. Поэтому в общем случае эксперимент всегда предпочтительней наблюдения.

1.4. Что ищут в данных

Прочитав предыдущие разделы, читатель, наверное, уже не раз задавался вопросом: «Если так все сложно, зачем он вообще, этот анализ данных? Неужели и *так* не видно, что в один магазин ходит больше народу, одно лекарство лучше другого и т. п.?» В общем, *так* бывает видно довольно часто, но обычно тогда, когда либо (1) данных и/или исследуемых факторов очень мало, либо (2) разница между ними очень резка. В этих случаях действительно запускать всю громоздкую машину анализа данных не сто́ит. Однако гораздо чаще встречаются случаи, когда названные выше условия не выполняются. Давно, например, доказано, что средний человек может одновременно удержать в памяти лишь 5–9 объектов. Стало быть, анализировать в уме данные, которые насчитывают больше 10 компонентов, уже нельзя. А значит, не обой-

тись без каких-нибудь, пусть и самых примитивных (типа вычисления процентов и средних величин), методов анализа данных.

Бывает и так, что внешне очевидные результаты не имеют под собой настоящего основания. Вот, например, одно из исследований насекомых-вредителей. Агрономы определяли, насколько сильно вредят кукурузе гусеницы кукурузного мотылька. Получились вполне приемлемые результаты: разница в урожае между пораженными и непораженными растениями почти вдвое. Казалось, что и обрабатывать ничего не надо — «и так все ясно». Однако нашелся вдумчивый исследователь, который заметил, что пораженные растения, различающиеся по степени поражения, не различаются по урожайности. Здесь, очевидно, что-то не так: если гусеницы вредят растению, то чем сильнее они вредят, тем меньше должен быть урожай. Стало быть, на какой-то стадии исследования произошла ошибка. Скорее всего, дело было так: для того чтобы измерять урожайность, среди здоровых растений отбирали самые здоровые (во всех смыслах этого слова), ну а среди больных старались подобрать самые хилые. Вот эта ошибка репрезентативности и привела к тому, что возникли такие «хорошие» результаты. Обратите внимание, что только анализ взаимосвязи «поражение—урожай» (на языке анализа данных он называется «регрессионный анализ», см. главу про двумерные данные) привел к выяснению истинной причины. А кукурузный мотылек, оказывается, почти и не вредит кукурузе...

Итак, анализ данных необходим всегда, когда результат неочевиден, и часто даже тогда, когда он *кажется очевидным*. Теперь разберемся, к каким последствиям может привести анализ, что он умеет.

1. Во-первых, анализ данных умеет давать общие характеристики для больших выборок. Эти характеристики могут отражать так называемую центральную тенденцию, то есть число (или ряд чисел), вокруг которых, как пули вокруг десятки в мишени, «разбросаны» данные. Всем известно, как считать среднее значение, но существует еще немало полезных характеристик «на ту же тему». Другая характеристика — это разброс, который отражает не *вокруг чего* «разбросаны» данные, а *насколько сильно* они разбросаны.

2. Во-вторых, можно проводить сравнения между разными выборками. Например, можно выяснить, в какой из групп больных инфарктом миокарда частота смертей в первый год после лечения выше — у тех, к кому применяли коронарное шунтирование, или у тех, к кому применяли только медикаментозные способы лечения. «На взгляд» этой разницы может и не быть, а если она и есть, то где гарантия того, что эти различия не вызваны случайными при-

чинами, не имеющими отношения к лечению? Скажем, заболел человек острым аппендицитом и умер после операции: к лечению инфаркта это может не иметь никакого отношения. Сравнение данных при помощи *статистических тестов* позволяет выяснить, насколько велика вероятность, что различия между группами вызваны случайными причинами. Заметьте, что гарантий анализ данных тоже не дает, зато позволяет оценить (численным образом) шансы. Анализ данных позволяет оценить и упомянутые выше общие характеристики.

3. Третий тип результата, который можно получить, анализируя данные,— это сведения о взаимосвязях. Изучение взаимосвязей — наверное, самый серьезный и самый развитый раздел анализа данных. Существует множество методик выяснения и, главное, проверки «качества» связей. В дальнейшем нам понадобятся сведения о том, какие бывают взаимосвязи. Есть так называемые *соответствия*, например когда два явления чаще встречаются вместе, нежели по отдельности (как гром и молния). Соответствия нетрудно найти, но силу их измерить трудно. Следующий тип взаимосвязей — это *корреляции*. Корреляции показывают *силу* взаимосвязи, но не могут определить ее *направления*. Другими словами, если выяснилась корреляция между качанием деревьев и ветром, то нельзя решить, дует ли ветер оттого, что деревья качаются, или наоборот. Наконец, есть *зависимости*, для которых можно измерить и силу, и направление, и оценить, насколько вероятно то, что они — результат случайных причин. Кстати говоря, последнее можно, как водится в анализе данных, сделать и для корреляций, и даже для соответствий. Еще одно свойство зависимостей состоит в том, что можно *предсказать*, как будет «вести» себя зависимая переменная в каких-нибудь до сих пор не опробованных условиях. Например, можно прогнозировать колебания спроса, устойчивость балок при землетрясении, интенсивность поступления больных и т. п.

4. И наконец, анализ данных можно использовать для установления *структуры*. Это самый сложный тип анализа, поскольку для выяснения структуры обычно используются *сразу несколько характеристик*. Есть и специальное название для такой работы — «многомерная статистика». Самое главное, на что способен многомерный анализ,— это создание и проверка качества *классификации* объектов. В умелых руках хорошая классификация очень полезна. Вот, например, мебельной фабрике потребовалось выяснить, какую мебель как лучше перевозить: в разобранном или в собранном виде. Рекомендации по перевозке зависят от уймы

причин (сложность сборки, хрупкость, стоимость, наличие стеклянных частей, наличие ящиков и полок и т. д.). Одновременно оценить эти факторы может лишь очень умелый человек. Однако существуют методы анализа, которые с легкостью разделят мебель на две группы, а заодно и проверят качество классификации, например ее соответствие сложившейся практике перевозок.

Существует и другой подход к результатам анализа данных. В нем все методы делятся на *предсказательные и описательные*. К первой группе методов относится все, что можно статистически оценить, то есть выяснить, *с какой вероятностью* может быть верным или неверным наш вывод. Ко второй — методы, которые «просто» сообщают информацию о данных без подтверждения какими-либо вероятностными методами. В последние годы все для большего числа методов находятся способы их вероятностной оценки, и поэтому первая группа все время увеличивается.

* * *

Ответ к задаче про случайный выбор деревьев в лесу. В этом случае шанс быть выбранными у елок выше, чем у дубов. Кроме того, лес может иметь какую-то структуру именно в выбранном направлении, и поэтому одной такой «диагонали» будет недостаточно для того, чтобы отобразить весь лес. Чтобы улучшить данный метод, надо провести несколько «диагоналей», а расстояния между выбираемыми деревьями по возможности увеличить.

Ответ к задаче про выбор жуков. Дело в том, что первыми вылезают самые активные особи, а чем активнее особь, тем быстрее она набирает на лапки смертельную дозу ядохимиката и, стало быть, быстрее гибнет. Это и было нарушением принципа рандомизации. Кроме того, нарушался принцип повторности: в чашку последовательно сажали жука за жуком, что не могло не повлиять на исход опыта. Для того чтобы поставить опыт правильно, надо было сначала подготовить (30 $\times$ количество ядохимикатов) чашек, столько же листочков с бумагой, *случайным образом* распределить ядохимикаты по чашкам, а затем перемешать жуков в банке, достать соответствующее количество и рассадить по чашкам.

Глава 2

Как обрабатывать данные

В принципе, обрабатывать данные можно и без компьютера. Так поступали в те годы, когда компьютерная техника была недоступна. Однако многие статистические расчеты настолько тяжеловесны, что уже в XIX веке стали придумывать способы их автоматизации.

2.1. Неспециализированные программы

Почти в любом компьютере с предустановленной системой есть программа-калькулятор. Такая программа обычно умеет выполнять четыре арифметических действия, часто — считать квадратные корни и степени, иногда логарифмы. В принципе, этого достаточно для того, чтобы делать простейшую обработку: считать среднее значение, стандартное отклонение, некоторые тесты.

Вообще говоря, для того чтобы делать тесты, кроме калькулятора потребуются еще и статистические таблицы, из которых можно узнать примерные значения так называемых статистик — величин, характеризующих данные в целом. Таблицы используются потому, что точный (по-английски «exact») подсчет многих статистик слишком сложен, порой даже для продвинутых компьютерных программ, поэтому используются оценочные («estimated») значения. Статистические таблицы можно найти во многих книгах по классической статистике, они также «встроены» во многие специализированные программы.

Главный недостаток калькулятора — сложность работы с сериями чисел, в то время как обычно данные как раз «идут» сериями (колонками, векторами). Чтобы работать с сериями более эффективно (и не только для этого), были придуманы электронные таблицы. Объяснять, как они устроены, наверное, не нужно. Сила электронных таблиц прежде всего в том, что они помогают визуализировать данные.

Глядя в электронную таблицу, можно сразу понять, как выглядят данные в целом, и «на глазок» оценить их основные параметры. Это очень полезно. Кроме визуализации, программы электронных таблиц снабжены развитым инструментарием для ввода и преобразования данных — автодополнением, автокопированием, сортировкой и т. п. Однако

большинство таких программ имеет своеобразное «родимое пятно» — они создавались в основном для офисного применения и были изначально ориентированы на бухгалтерские задачи. Для обработки данных нужен гораздо более специализированный инструмент. Конечно, развитые программы электронных таблиц, такие как MS Excel, Gnumeric или OpenOffice.org Calc, имеют, среди прочего, набор статистических функций. Но поскольку это — не основной компонент, на статистику в этих программах традиционно обращается мало внимания. Набор статистических тестов невелик, многие методы, особенно многомерные, отсутствуют, реализация (то есть как именно идут внутри программы сами вычисления) часто далека от оптимальной, нет специализированной системы отчетов, много неудобных ограничений, возможны ошибки, которые будут исправляться не слишком быстро, опять-таки потому, что статистика — не первоочередная функция электронных таблиц.

Кроме того, базовый принцип электронных таблиц — визуализация данных — имеет и свои оборотные стороны. Что, если данные не помещаются в окне? В этом случае их надо будет прокручивать или скрывать часть ячеек. И то, и другое оказывает пользователю медвежью услугу, потому что он рассчитывал на облегчение восприятия данных, в то время как программа, скорее, затрудняет его. Или другой пример — надо провести операции с тремя несмежными столбцами. Сделать это через выделение нельзя, потому что выделение всегда одного типа (да и буфер обмена очень часто всего один), приходится делать много движений мышкой с большим риском ошибиться. И уж совсем никуда не годится графическая система, если надо сочетать методы обработки каким-нибудь сложным образом.

Выход — в пользовании специализированными статистическими программами.

2.2. Специализированные статистические программы

2.2.1. Оконно-кнопочные системы

Есть две группы специализированных статистических программ. Первые не особенно отличаются внешне от электронных таблиц, однако снабжены значительно большим арсеналом доступных статистических приемов. Кроме того, у них традиционно мощная графическая часть (возможных графиков больше, и управление ими более гибкое), а часто и подсистема подготовки отчетов. Многие такие программы имеют значительно меньше ограничений, чем электронные таблицы.

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