Оглавление

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