Учебник: The Art of Data Science

Учебник для этого курса, [*The Art of Data Science*](https://leanpub.com/artofdatascience?utm_source=CourseraEDS&utm_medium=Lesson&utm_campaign=CourseraEDS), написан мной и Элизабет Мацуи. Книга охватывает все темы, затрагиваемые в курсе, и содержит дополнительные примеры. Она доступна в формате электронной книги и поддерживается на большинстве электронных книг (доступны версии для Kindle, iBooks, а также версия в формате PDF). Учебник послужит полезным справочным пособием при прохождении курса и даст возможность освежить знания по его завершении.

Роджер Пэн

# Книга Conversations on Data Science

Благодарим за интерес к курсу "Управление анализом данных". Как вы знаете, наука о данных — это быстро меняющаяся область. Стратегии и передовые методики выполнения анализа данных и управления этим процессом постоянно меняются.

Мы с доктором Хилари Паркер (специалист по данным в Stitch Fix) написали книгу [Conversations on Data Science](https://leanpub.com/conversationsondatascience?utm_source=Coursera&utm_medium=Announcement&utm_campaign=EDS3), в которой собраны многие из наших дискуссий по управлению и практике в области анализа данных в промышленности и науке. Эти дискуссии основаны на нашем подкасте Not So Standard Deviations, где мы обсуждаем отрасль науки о данных. Книга рассматривает обсуждения командного взаимодействия, автоматизации в науке о данных и многое другое. Книга публикуется с помощью Leanpub. Ее ключевой особенностью является **постоянное обновление** по мере записи новых эпизодов нашего подкаста. Читатели будут получать обновленные главы по мере их публикации.

Я думаю, это пособие станет хорошим дополнением к материалу, изучаемому на курсе, и надеюсь, что оно вам понравится.

Роджер Пэн

# Оценка распределения

Цель данного урока — в общих чертах описать понятие модели, объяснить ее назначение в контексте набора данных и описать процесс, с помощью которого специалисты по анализу данных создают, оценивают и уточняют модель. В самом общем смысле модель — это то, что мы создаем, чтобы понять реальность. Типичным примером модели является использование подопытного животного для имитации человеческого заболевания с целью понять и, если это возможно, предотвратить или излечить болезнь. То же самое относится и к набору данных. Зачастую данные используются для того, чтобы понять реальность.

В сфере политики у социолога есть набор данных о выборке потенциальных избирателей, и его задача — использовать эту выборку для прогнозирования результатов выборов. Специалист по анализу данных использует данные опроса для построения модели, прогнозирующей, что произойдет в день выборов. Процесс построения модели включает в себя наложение определенной структуры на данные и создание сводки данных.В примере с данными опроса у вас могут быть тысячи результатов наблюдений, поэтому модель представляет собой математическое уравнение, отражающее форму или структуру данных. Это уравнение позволяет обобщить тысячи результатов наблюдений до одного числа, которое может представлять собой, например, процент избирателей, которые проголосуют за вашего кандидата. Сейчас эти концепции могут показаться немного расплывчатыми, но если вы продолжите чтение, они станут понятнее.

В анализе данных модели выполняют две ключевые задачи: предоставление краткой количественной информации по данным и наложение определенной структуры на генеральную совокупность, из которой были отобраны данные. Примеры крайности — это хороший способ объяснить, что такое модель и как ее можно использовать. Самая тривиальная "модель" — **это отсутствие модели**.

Представьте, что вы хотите провести опрос 20 человек, чтобы узнать, сколько они готовы потратить на разрабатываемый вами продукт. Какова цель опроса? Скорее всего, если вы тратите время и деньги на разработку нового продукта, вы считаете, что существует большая совокупность людей, желающих купить этот продукт. Но спрашивать каждого в этой совокупности о том, сколько они готовы заплатить, слишком дорого и сложно. Поэтому вы делаете выборку из этой совокупности, чтобы понять, сколько готовы заплатить ее представители.

Недавно я опубликовал книгу под названием [R Programming for Data Science](https://leanpub.com/rprogramming). Перед публикацией заинтересованные читатели могли указать свое имя и адрес электронной почты на сайте книги, чтобы получить уведомление о её выходе. Кроме того, там была возможность указать, сколько они готовы заплатить за книгу. Ниже приведена случайная выборка из 20 ответов людей, которые добровольно предоставили эту информацию.

25 20 15 5 30 7 5 10 12 40 30 30 10 25 10 20 10 10 25 5

Теперь предположим, что кто-то спросил вас: "Что говорят данные?" Вы можете просто передать данные — все 20 чисел. Поскольку набор данных не такой уж большой, вряд ли это станет серьезной проблемой. В конечном счете, ответ на вопрос находится в этом наборе данных, но наличие всех данных — это не ответ. Иметь все данные очень важно, но часто не очень полезно. Это связано с тем, что тривиальная модель не обеспечивает редукции данных.

Первый ключевой элемент статистической модели — это редукция данных. Основная идея редукции заключается в том, чтобы взять исходный набор чисел, содержащийся в вашем наборе данных, и преобразовать их в меньший набор чисел. Если изначально было 20 чисел, модель должна дать сводку, содержащую меньше 20 чисел. Обычно процесс редукции данных дает результат в виде статистики. По большому счету, любая сводка данных — это статистика. Среднеарифметическое значение выборки — это статистика. Так же как и медианное, максимальное, минимальное значение, среднеквадратическое отклонение и диапазон. Одна статистика более полезна, другая менее, но все это сводные данные.

Вероятно, самый простой способ редукции, которым можно воспользоваться, это расчет среднего арифметического данных. В нашем случае это 17,2 доллара. Переход от 20 чисел к 1 — это максимальная редукция, которой вы можете достигнуть в этом случае, и она определенно удовлетворяет сводному элементу модели.

## Модели как ожидания

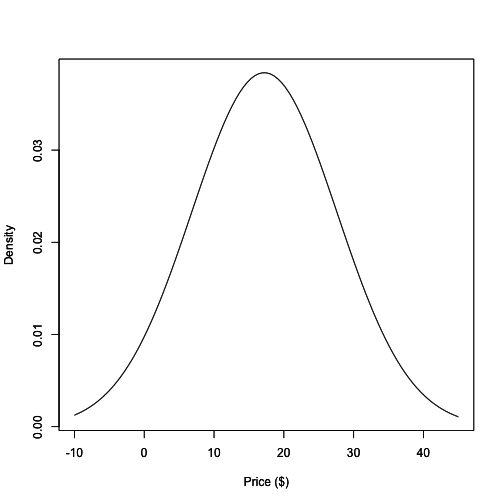
Но простой сводной статистики, такой как среднее арифметическое набора чисел, недостаточно для составления модели. Статистическая модель также должна накладывать определенную структуру на данные. По сути, **статистическая модель описывает реальность и то, как были сгенерированы данные**. Модель — это в первую очередь ожидание касательно зависимостей между различными факторами в реальности и в наборе данных. Модель является статистической, если она допускает случайность при генерации данных.

### Применение нормальной модели

Пожалуй, наиболее популярная статистическая модель в мире — это нормальная модель. Согласно этой модели, случайность в наборе данных может быть объяснена нормальным распределением, или колоколообразной кривой. Нормальное распределение полностью определяется двумя параметрами: средним арифметическим значением и среднеквадратическим отклонением.

Возьмем данные, которые мы рассмотрели в предыдущем разделе — сумму, которую 20 человек готовы заплатить за гипотетический новый продукт. Будем надеяться, что эти 20 человек представляют собой репрезентативную выборку из совокупности людей, которые могли бы купить его. В таком случае из информации, содержащейся в наборе данных, мы можем узнать что-нибудь о каждом человеке, входящем в совокупность.

Чтобы применить нормальную модель к этому набору данных, достаточно рассчитать среднеарифметическое значение и среднеквадратическое отклонение. В нашем случае среднеарифметическое значение составляет 17,2 доллара, а среднеквадратическое отклонение — 10,39 доллара. Учитывая эти параметры, мы ожидаем, что в рамках нормальной модели распределение цен, которые люди готовы платить, будет выглядеть примерно так.



Согласно модели, около 68% генеральной совокупности готовы платить за этот новый продукт от 6,81 до 27,59 долларов. Полезна эта информация или нет, зависит от специфики ситуации, которую мы пока не рассматриваем.

При необходимости можно использовать статистическую модель, чтобы ответить на более сложные вопросы. Например, представим, что вам нужно узнать, какая часть генеральной совокупности готова заплатить за эту книгу более 30 долларов.

Используя свойства нормального распределения (и небольшую помощь в вычислениях от среды R), можно легко выполнить этот расчет. Около 11% совокупности готовы заплатить за продукт более 30 долларов. Опять же, полезно ли это для вас, зависит от ваших целей.

Обратите внимание, что на графике выше отсутствует одна важная вещь — данные! Это не совсем так, поскольку мы использовали данные для создания графика (для вычисления среднеарифметического значения и среднеквадратического отклонения нормального распределения), но непосредственно данные на нем не отображены. В этом случае **мы используем нормальное распределение, чтобы узнать, как выглядит генеральная совокупность**, а не как выглядят данные.

Ключевым моментом здесь является то, что мы использовали нормальное распределение, чтобы задать форму распределения, которой, как мы ожидаем, будут следовать данные. Нормальное распределение — это ожидание в отношении того, как должны выглядеть данные.

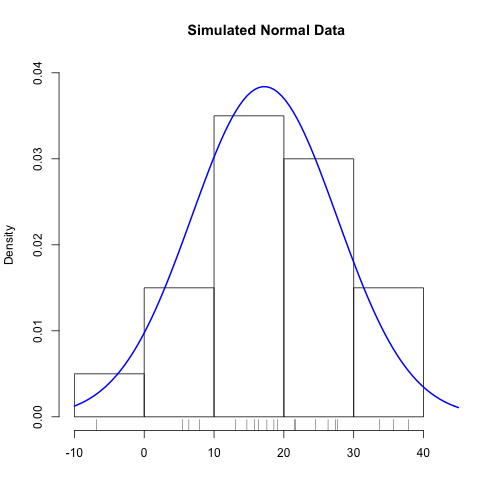
## Сравнение ожиданий с реальностью

Мы можем очень гордиться разработанной нами статистической моделью, но то, насколько она полезна, зависит от того, насколько точно она отражает реальные собранные данные. Как нам проверить, соответствуют ли наши ожидания реальности?

### Построение предварительного графика

Для начала мы можем построить несколько графиков, такие как гистограмма данных. Но прежде чем приступить к работе с данными, давайте выясним, что мы ожидаем из них узнать. Если генеральная совокупность примерно соответствовала нормальному распределению, а данные представляли собой случайную выборку из этой совокупности, то распределение, оцененное с помощью гистограммы, должно выглядеть как теоретическая модель, построенная на основе нормального распределения.

На изображении ниже я смоделировал 20 точек данных из нормального распределения и наложил теоретическую кривую нормального распределения поверх гистограммы.



Обратите внимание, насколько совпадают столбцы гистограммы и синяя кривая. Это именно то, что мы хотели увидеть. Если мы это видим, то можно сделать вывод, что нормальное распределение — это **хорошая статистическая модель для наших данных**.

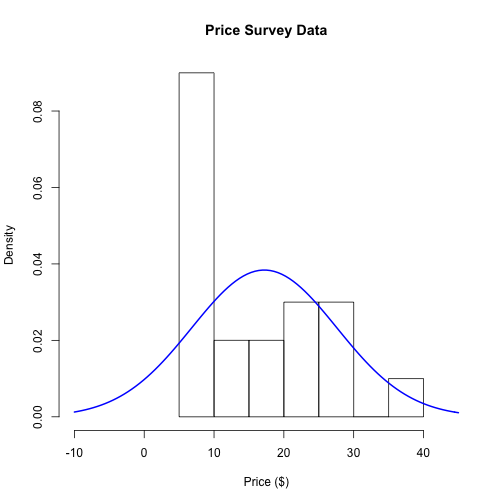
Моделирование данных на основе гипотетической модели — это хороший способ установить ожидания, прежде чем приступить к изучению данных. Построение предварительного графика (даже вручную, если необходимо) может быть очень полезным инструментом для начала обсуждения модели и ожиданий от реальности.

Например, ещё до того, как посмотреть на данные, мы можем догадаться, что нормальная модель не может обеспечить идеального представления генеральной совокупности. В частности, нормальное распределение допускает отрицательные значения, но вряд ли можно ожидать, что люди скажут, что они готовы заплатить за книгу отрицательную сумму.

У нас уже есть доказательство того, что нормальная модель не идеальный выбор, но идеальных моделей не существует. Вопрос в том, обеспечивает ли статистическая модель достаточное приближение, которое может быть каким-то образом полезно.

### Настоящий график

Взгляните на гистограмму данных выборки из 20 респондентов. Поверх гистограммы я наложил кривую нормального распределения с 20 точками данных, указывающими суммы, которые люди, по их словам, готовы заплатить за книгу.



Мы ожидаем, что гистограмма и синяя линия должны примерно совпадать друг с другом. Как соотносятся модель и реальность?

На первый взгляд кажется, что гистограмма и кривая нормального распределения не очень совпадают. На гистограмме есть большой скачок в области 10 долларов, которого нет на синей кривой. Кроме того, нормальное распределение допускает отрицательные значения в левой части графика, но в этой области нет точек данных.

Пока что данные показывают, что нормальная модель не очень хорошо представляет генеральную совокупность, учитывая выборку данных, взятую из этой совокупности. Похоже, что 20 респондентов решительно настроены заплатить примерно 10 долларов, хотя на самом деле есть несколько человек, готовых заплатить больше. Такие особенности данных плохо характеризуются нормальным распределением.

## Реакция на данные: уточнение ожиданий

Гистограмма показывает, что модель и данные не очень хорошо совпадают. Что же делать? Мы можем либо

1. выбрать другую модель,
2. либо взять другие данные,

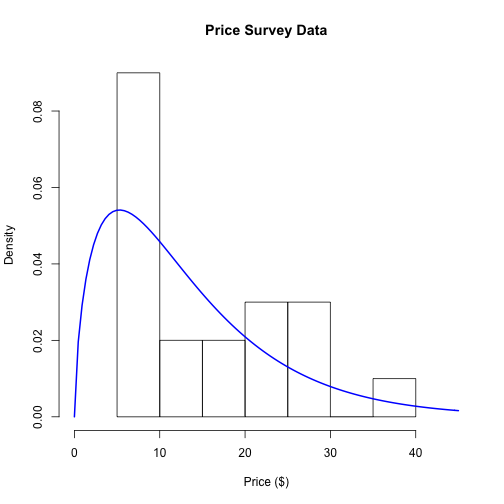
либо сделать и то, и другое. Наши действия в некоторой степени зависят от наших представлений о модели и нашего понимания процесса сбора данных. Если бы мы были твердо уверены в том, что генеральная совокупность сумм, которые готовы платить респонденты, должна соответствовать нормальному распределению, то вероятность внесения серьезных изменений в модель была бы ниже. Мы могли бы изучить процесс сбора данных, чтобы понять, не возникло ли при его выполнении какого-либо смещения данных. Но, если окажется, что процесс сбора данных прошел успешно, нам придется пересмотреть модель для генеральной совокупности и подумать, что можно изменить. В этом случае вполне вероятно, что наша модель не подходит. Трудно представить себе процесс сбора данных, который мог бы привести к отрицательным значениям (что допускает нормальное распределение).

Чтобы замкнуть цикл, мы выберем другую статистическую модель, представляющую генеральную совокупность, — гамма-распределение. Этот тип распределения имеет особенность: оно допускает только положительные значения. Это устраняет проблему с отрицательными значениями, с которой мы столкнулись, работая с нормальным распределением.

Теперь мы должны вернуться к началу нашего цикла и сделать следующее:

1. Сформируйте ожидания. Постройте предварительный график. Что мы ожидаем увидеть перед просмотром данных?
2. Сравните наши ожидания с данными
3. Уточните ожидания с учетом того, что показывают данные

Для справки здесь представлена гистограмма тех же данных с наложением гамма-распределения (оцениваемого по данным).



Насколько теперь данные соответствуют вашим ожиданиям?

Вы можете спросить, какая разница, какую модель я использую для представления генеральной совокупности, из которой сгенерированы данные? Во-первых, это может повлиять на прогнозы, которые вы делаете с помощью модели. Например, напомню, ранее вас интересовало, какая часть генеральной совокупности готова заплатить за книгу по крайней мере 30 долларов. Согласно новой модели лишь около 7% генеральной совокупности готовы заплатить хотя бы эту сумму (согласно нормальной модели 11% были готовы заплатить 30 долларов или больше). Таким образом, разные модели могут давать разные прогнозы на основе одних и тех же данных, что может повлиять на решения, принимаемые в дальнейшем.

Примечание. Отрывки этого материала для самостоятельного изучения взяты из книги Р. Пэна и Э. Мацуи [The Art of Data Science](https://leanpub.com/artofdatascience/).

# Оценка линейных зависимостей

Зачастую мы смотрим на данные и пытаемся понять линейные зависимости между интересующими нас переменными. Самый распространенный статистический метод для решения такой задачи — это линейная регрессия. Рассмотренные выше принципы, такие как формирование ожиданий, сопоставление ожиданий с данными и уточнение ожиданий, применимы и к линейной регрессии.

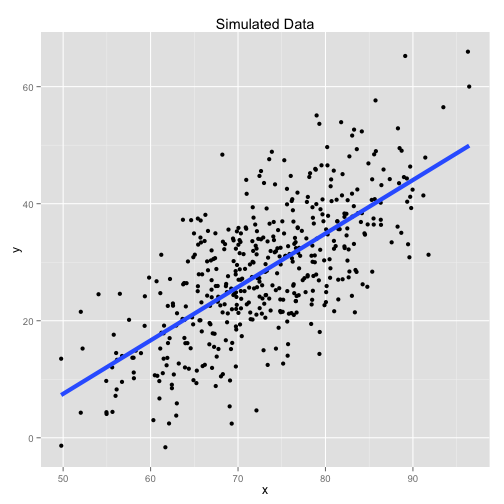
В этом примере мы рассмотрим простой набор данных о качестве воздуха, содержащий информацию об уровнях тропосферного озона в городе Нью-Йорк в 1999 году с мая до конца года. Данные содержат среднесуточные уровни озона (в частицах на миллиард) и температуры (в градусах по Фаренгейту). Один из вопросов, который может мотивировать сбор этого набора данных, — "Как температура воздуха связана с уровнем озона в атмосфере Нью-Йорка?"

### Ожидания

Почитав немного об [образовании озона в атмосфере](https://en.wikipedia.org/wiki/Tropospheric_ozone), мы узнаём, что образование озона существенно зависит от наличия солнечного света. Солнечный свет также связан с температурой, поскольку можно ожидать, что в дни, когда много солнечного света, средняя температура выше. В пасмурные дни, как правило, ниже и средняя температура, и уровень озона. Есть основания полагать, что в дни с более высокими температурами можно ожидать более высоких уровней озона. Это косвенная зависимость — мы используем температуру как показатель количества солнечного света.

Самая простая модель, которую можно составить, чтобы охарактеризовать зависимость между температурой и уровнем озона, — это линейная модель. Согласно этой модели, с повышением температуры количество озона в атмосфере линейно увеличивается. Как это будет выглядеть, согласно нашим ожиданиям?

Мы можем смоделировать данные, чтобы построить предварительный график того, как должна выглядеть зависимость между уровнем озона и температурой в линейной модели. Вот простая линейная зависимость со смоделированными данными на диаграмме рассеяния.

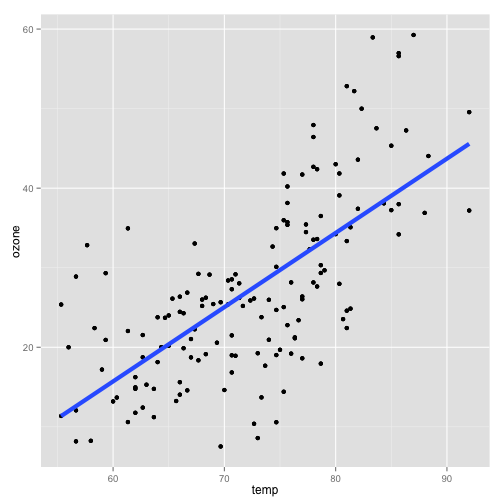


Обратите внимание, что если выбрать любую точку на синей линии, над ней будет примерно такое же количество точек, как и под ней (это также называется несистематическими ошибками). Кроме того, кажется, что количество точек на диаграмме рассеяния линейно увеличиваются по мере движения вправо по оси Х, даже несмотря на то, что вдоль линии наблюдается довольно много шума/рассеяния.

Если мы не ошиблись, выбрав линейную модель, и она действительно характеризует данные и зависимость между уровнем озона и температурой, то именно такую картину мы и должны увидеть, построив график данных.

### Сопоставление ожиданий с данными

Взгляните на график с фактическими данными по уровню озона и температуре в Нью-Йорке за 1999 год. Поверх диаграммы рассеяния данных мы построили подогнанную линию линейной регрессии, рассчитанную на основе данных.



Как этот график соотносится с тем, который вы ожидали увидеть?

Ясно одно: как мы и предполагали, наблюдается тенденция к увеличению содержания озона с повышением температуры. Тем не менее, есть несколько отклонений от предварительного графика, построенного нами ранее. Похоже, что точки не распределены равномерно вокруг синей линии регрессии.

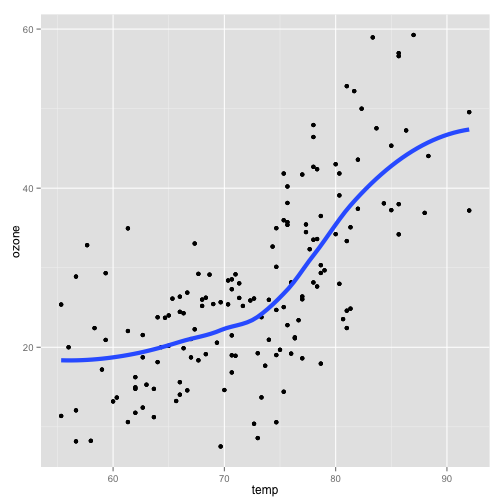
Если провести вертикальную линию возле температуры 85 градусов, вы заметите, что большинство точек находятся над линией. Начертив вертикальную линию возле 70 градусов, вы заметите, что большинство точек находятся под линией. Это означает, что при более высоких температурах наша модель смещается вниз (занижая оценку уровня озона), а при умеренных температурах модель, наоборот, смещается вверх. Это не очень хорошо. Было бы лучше, если бы наша модель не содержала никаких систематических ошибок.

Наша простая модель линейной регрессии, по-видимому, отражает общую возрастающую связь между температурой и уровнем озона, но при этом она смещена в определенных диапазонах температур. Похоже, эту модель можно улучшить, если необходимо лучше охарактеризовать взаимосвязь между температурой и уровнем озона в этом наборе данных.

### Корректировка ожиданий

Согласно изображению выше, зависимость уровней температуры и озона может быть нелинейной. В самом деле, по точкам данных мы видим, что она остается линейной примерно до 70 градусов, после чего оба показателя резко возрастают. Всё это указывает на нелинейную зависимость уровней температуры и озона.

Проще всего отразить это изменение ожиданий поможет сглаживание с помощью функции loess.



Новый график представляет другой результат — показатели медленно возрастает до 75 градусов, после чего резко повышаются. На уровне 90 градусов они снова выравниваются.

Функции сглаживания (например, loess) очень полезны, так как помогают быстро отразить закономерности в наборе данных, и для этого не нужно строить предположения о структуре данных. По своей сути они являются автоматизированными (компьютеризированными) инструментами, которые создают набросок кривой для определённых Однако сглаживатели редко помогают понять механизм связи, а потому их польза в этом деле ограничена. Чтобы подробнее изучить связь между температурой и уровнем озоном, возможно, понадобится применить более сложную модель, нежели обычная линейная регрессия.

Примечание. Отрывки этого материала для самостоятельного изучения взяты из книги Р. Пэна и Э. Мацуи [The Art of Data Science](https://leanpub.com/artofdatascience/).

# Исследовательский анализ данных. Когда прекращать анализ?

В примерах анализа данных, которые ранее приводились в этом курсе, мы выполняли одну итерацию. В некоторых случаях этого достаточно, но обычно нужно провести хотя бы несколько итераций. Вернёмся к примерам, которые ранее упоминались в этом разделе, ведь у нас остался ряд невыясненных вопросов.

* **Данные опроса о цене.** В этом примере мы остановились на модели гамма-распределения. Но как она используется вместе с данными? Каких результатов можно ожидать, если модель гамма-распределения действительно подходит (или если мы никогда ранее её не использовали)?Есть ли другой способ, который может лучше отразить резкий скачок распределения в районе 10 долларов?
* **Озон и температура.** Сглаживание помогло выявить нелинейную связь между уровнями температуры и озона, но в чём её причина? Существует ли эта связь на самом деле, или же это чистая случайность? Способны ли мы определить другие физические факторы, помимо температуры, которые объясняют такой резкий скачок показателей озона? Можем ли мы смоделировать их?

Возможно, понадобится провести ещё несколько итераций. Как правило, каждый ответ будет порождать больше вопросов и вынуждать вас ещё подробнее изучать данные. Когда же следует остановиться? В теории статистики существуют разные подходы к определению того, когда статистическая модель достаточно хороша и соответствует данным. Мы не будем обсуждать их в этой главе, а поговорим о нескольких важных критериях, которые помогут определить конец в цикле итераций.

### Есть ли у вас ещё данные?

Из-за цикличной природы анализа вы в итоге столкнётесь с теми вопросами, на которые просто не сможете дать ответ с помощью имеющихся данных. В примере с температурой и озоном моделирование указывает не на простую, а на нелинейную связь между двумя переменными. Но данные не могут точно объяснить, почему она возникает (хотя и могут стать основой для ряда гипотез). Более того, вам нужны дополнительные данные, чтобы проверить, являются ли ваши результаты истиной, случайностью или статистической ошибкой. В любом случае вам придётся отвлечься от расчётов и собрать новые данные, так как последующий анализ не даст вам ответы на вопросы.

Возможно, дополнительные данные понадобятся и тогда, когда вы завершаете анализ и получаете удовлетворительные и часто интересные результаты. В некоторых случаях важно попытаться воспроизвести полученные результаты с помощью другого, по возможности независимого набора данных. В примере с озоном и температурой мы установили между показателями нелинейную зависимость, но наши выводы могут стать куда весомее, если мы докажем наличие этой тенденции в других городах помимо Нью-Йорка. Кроме того, такое независимое подтверждение сделает обоснования весомее и сыграет более важную роль в принятии решения.

### Достаточно ли у вас информации для принятия решения?

Часто анализ данных проводят для того, чтобы принять конкретное решение в сферах бизнеса, образования, государственного управления и т. д. Учитывайте, что подход к анализу может отличаться в зависимости от того, хотите ли вы найти правильное решение, составить отчёт, написать публикацию или выпустить продукт.

Поэтому важно всегда помнить о своей цели и подстраивать под неё анализ, чтобы эффективно распределять ресурсы. Вы можете изменить её со временем или иметь одновременно несколько целей. Чтобы понять, хватает ли вам данных, учитывайте особенности своей ситуации, в частности издержки и выгоды. Если вам кажется, что полученных результатов недостаточно, чтобы принять решение, то вам нужно провести дополнительный анализ или получить больше данных.

### Важны ли ваши результаты в контексте прочих подобных исследований?

Или же по-другому: "Имеют ли смысл эти результаты?" Чтобы ответить на этот вопрос, изучите литературу по своей проблеме или узнайте, не приходил ли кто-то из ваших коллег к подобным выводам. Если результаты вашего анализа очень близки к тем, что получили ранее другие люди — это хороший знак, но далеко не единственный желаемый исход. Выводы, которые ранее не упоминались в исследованиях, могут проложить дорогу к новым открытиям. В любом случае найти правильный ответ бывает трудно, если не проводить дополнительных исследований.

Будьте осторожны с ответом на этот вопрос. Когда мы работаем с крупными и сложными наборами данных, очень просто остановиться на результатах, которые имеют хоть какой-то смысл и соответствуют нашему пониманию процесса. Поэтому важно относиться к результатам максимально критично и постоянно подвергать их сомнениям. Исходя из нашего опыта, если данные в значительной мере совпадают с ожиданиями, значит, вы неправильно провели анализ или сбор информации, либо же ошибочно истолковали результаты. Важно ставить под вопрос каждый аспект анализа, чтобы убедиться, что вы выполнили его правильно.

Если результаты не имеют смысла или не соответствуют ожиданиям, то этот случай вызывает куда больший интерес. Вы могли банально допустить ошибку в анализе или сборе данных. Скорее всего, именно в этом и причина. Образно говоря, на каждый алмаз приходится около 99 кусков угля. Убедитесь, что анализ выполнен правильно, и проверьте свои результаты на другом наборе данных на случай, если вы обнаружили что-то необычное и ранее неизвестное. Люди очень придирчивы к неожиданным результатам, поэтому будьте готовы отстаивать свою работу.

Если вы можете с уверенностью ответить на вопрос "Имеют ли смысл ваши результаты?" (и неважно, положительно или отрицательно), скорее всего, вам пора **прекратить анализ и перепроверить работу**.

### Действительно ли у нас нет времени?

Этот критерий кажется провокационным, но он играет важную роль в том, когда прекратить анализ. С ним также связан вопрос "Действительно ли у нас нет средств?" У каждого проекта есть бюджет времени и денег, которые вы можете вложить в анализ. Учитывайте бюджеты, даже если вы их не контролируете. Так вы сможете эффективно спланировать рабочий процесс. Возможно, вам понадобятся дополнительные ресурсы, и вы решите обратиться к своим коллегам. Чтобы чётко аргументировать необходимость продолжать анализ, важно понимать, когда пора остановить работу и подготовить полученные результаты.

ПРИМЕЧАНИЕ. Отрывки этого материала для самостоятельного изучения взяты из книги Р. Пэна и Э. Мацуи [The Art of Data Science](https://leanpub.com/artofdatascience/).

# Факторы, которые влияют на качество выводов

Основные факторы, которые могут повлиять на качество выводов, связаны с ошибочными представлениями о процессе создания выборки и модели генеральной совокупности. Очевидно, что если мы не определим чётко генеральную совокупность, наши выводы будут такими же неоднозначными.

Неправильное понимание принципов выборки может привести к тому, что данные, которые вы собрали, окажутся не настолько репрезентативными для генеральной совокупности, как вы думали. Из-за этого мы можем получить результаты, которые будут касаться не всей генеральной совокупности, а лишь её отдельной части. Это явление иногда называют **систематической ошибкой выборки**, так как итоговые показатели касаются только той совокупности, которую вы отобрали.

Если выбрать ошибочную модель для генеральной совокупности, возможно, вы неправильно оцените связь между признаками группы или недооцените неопределённость расчётов. Например, если пингвины действительно могут влиять на цвет шляп своих собратьев, то это противоречит предположению об их независимости друг от друга. Из-за этого все оценки относительно данных станут более неопределёнными. В общем, зависимость между элементами генеральной совокупности уменьшает эффективный размер выборки для набора данных. Причина в том, что изучаемые элементы не являются до конца независимыми, а потому не представляют отдельные фрагменты информации.

Последняя причина того, почему наши предположения о данных отличаются от реальных результатов — это **изменчивость выборки**. В нашем случае с пингвинами выборка составлялась случайно, а потому если бы мы повторили эксперимент и выбрали три других птицы, то получили бы другое приблизительное количество пингвинов в бирюзовых шляпах. Всё это обусловлено изменчивостью в процессе выборки. Она проявляется даже тогда, когда мы составляем точное описание процесса выборки и идеальную модель для генеральной совокупности.

В большинстве ситуаций разницу между ожиданиями и реальными результатами можно объяснить сочетанием всех трёх факторов. Определить влияние каждого из них на проблему может быть трудно из-за нехватки информации, но старайтесь учитывать каждый фактор в работе и определять для себя главный. Так вы сможете не допускать подобной ошибки в будущих исследованиях и экспериментах.

ПРИМЕЧАНИЕ. Отрывки этого материала для самостоятельного изучения взяты из книги Р. Пэна и Э. Мацуи [The Art of Data Science](https://leanpub.com/artofdatascience/).

# Примечание по генеральным совокупностям

Существует ряд стратегий, которые помогают установить формальные рамки для формулировки выводов. Обычно мы работаем буквально с совокупностью объектов — людей, пингвинов и т. д. В таким случаях мы чётко понимаем источник неопределённости (выборка из совокупности) и объект предположений (признак совокупности). Тем не менее в других ситуациях бывает не так просто определить, что является генеральной совокупностью и что мы пытаемся оценить. Тогда, чтобы избежать неопределённости, стоит более конкретно задавать генеральную совокупность.

## Временные ряды

Некоторые процессы измеряются с определённой регулярностью (ежеминутно, ежедневно и т. д.). Например, вы можете захотеть проанализировать ежедневные цены акций Apple в момент закрытия биржи за календарный 2014 год. Что будет генеральной совокупностью в этом случае? Есть несколько вариантов.

1. Мы можем утверждать, что 2014 год был случайно выбран из генеральной совокупности данных за все годы, а потому выводы, которые мы получим, применимы к ценам акций за другие периоды.
2. Мы также можем рассматривать акции Apple как выборку из всего рынка, а потому полученные выводы применимы и к другим акциям из этого набора.

Независимо от выбранного подхода, важно понимать генеральную совокупность, с которой вы работаете, прежде чем делать выводы.

## Естественные процессы

Природные явления, такие как землетрясения, пожары, ураганы и погодные изменения, часто измеряются во времени и пространстве. Для чисто временных измерений можно определять генеральную совокупность по тому же принципу, что и в примере с временными рядами. Тем не менее у нас могут быть только те данные, которые измеряются в пространстве. Например, карта эпицентров землетрясений в одном регионе. Как определить генеральную совокупность в этом случае? С одной стороны, мы можем утверждать о наличии ненаблюдаемого случайного процесса, который произвольно вызывает землетрясения в изучаемой области. В таком случае наши данные представляют случайную выборку для этого процесса и помогают его подробнее изучить.

## Данные как генеральная совокупность

Один из возможных, но не таких популярных методов — это использовать в качестве генеральной совокупности набор данных. В таком случае вам не придётся делать выводы, так как вы не проводите выборку. Поскольку набор данных и есть генеральная совокупность, у вас не возникнет неопределённости по поводу какой-либо из его характеристик. Такой подход может показаться неэффективным, но в некоторых ситуациях он способен дать ответы на важные вопросы. В частности, иногда нас не волнуют факторы, не входящие в изучаемый набор данных.

Например, в компаниях часто анализируют данные по зарплатам, чтобы убедиться, что женщины и мужчины, выполняющие одинаковую работу, имеют один уровень зарплат, или что этнический состав сотрудников достаточно сбалансирован. В таких условиях мы можем рассчитать разницу в зарплатах с помощью набора данных и оценить, представляет ли она повод для беспокойства. Данные напрямую отвечают на интересующий нас вопрос: "Существует ли значительная разница в уровне зарплат, которую нужно устранить?" В этом случае не нужно делать выводов о сотрудниках вне вашей компании (таковых нет по определению) или работниках других компаниях, которыми вы не управляете. Набор данных выступает как генеральная совокупность, и ответы на все вопросы по ней заложены в этом наборе.

ПРИМЕЧАНИЕ. Отрывки этого материала для самостоятельного изучения взяты из книги Р. Пэна и Э. Мацуи [The Art of Data Science](https://leanpub.com/artofdatascience/).

# Вывод и прогноз

Важно понимать, с каким типом вопроса вы работаете — индуктивным или предиктивным. От него будет зависеть ваш подход к моделированию. Если у вас нет чёткого понимания типа вопроса, вы рискуете использовать неправильную модель и в результате получить ошибочные выводы о данных. В этой главе мы хотим показать, что может случиться, если вы принимаете один тип вопроса за другой.

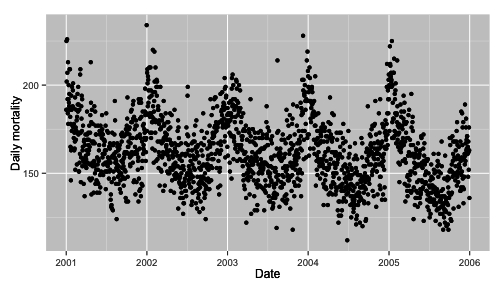
О чём стоит помнить?

1. Цель **индуктивных вопросов** — оценить связь между результатом и рассматриваемыми предикторами. Как правило, выбирается один или несколько предикторов и ряд потенциальных искажающих факторов. Основное назначение модели — оценить связь с учётом всех искажающих факторов. Часто проводят анализы чувствительности, чтобы определить, насколько изучаемые связи устойчивы к различным искажениям.
2. Цель **предиктивных вопросов** — выявить модель, которая лучше всего прогнозирует результат. Пока предикторы одинаково эффективно справляются с этой задачей, мы не отдаём приоритет какому-либо из них. Мы не выделяем искажающие факторы и ключевой предиктор, так как все переменные могут быть полезны в прогнозировании результата. Нас также не волнует принцип работы модели или подробные сведения о предикторах. Наша задача — разработать модель с высокой прогнозирующей способностью и определить допустимый коэффициент ошибок на основе данных.

## Уровень загрязнения воздуха и смертности в Нью-Йорке

Следующий пример показывает, как разные типы вопросов и соответствующие им модели влияют на выводы анализа. В качестве примера рассмотрим данные о смертности и загрязненности воздуха в Нью-Йорке. Показатели взяты из материалов Национального исследования смертности, заболеваемости и загрязнённости воздуха.

Ниже приведена диаграмма для уровня суточной смертности за период 2001–2005 гг.



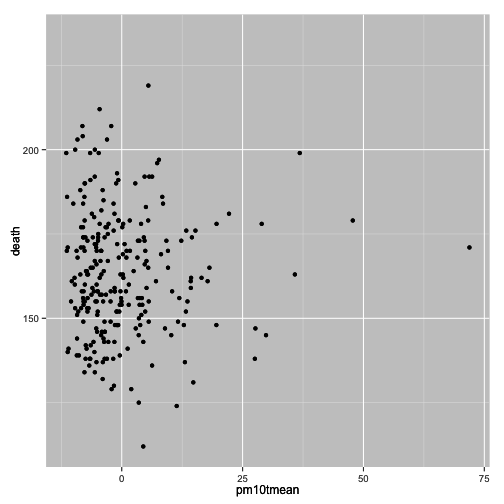
Перед вами показатели среднесуточного уровня твёрдых частиц, чей аэродинамический диаметр составляет 10 микронов или меньше (частицы PM10).

Заметьте, что на этой диаграмме куда меньше точек, чем на приведённой выше для уровня смертности, потому что уровень частиц PM10 не измеряется ежедневно. Также обратите внимание, что значения на этой диаграмме могут быть отрицательными, так как данные вычитались из среднего арифметического. В целом отрицательный уровень частиц PM10 невозможен.

## Выведение связи

Первым делом задаём себе вопрос: "Существует ли связь между среднесуточным уровнем частиц PM10 и суточной смертностью?" Мы хотим установить связь между показателями, а значит, это индуктивный тип вопроса. Помимо этого, мы также знаем, что нужно учесть ряд потенциальных искажающих факторов.

Давайте проведём двумерный анализ уровней смертности и частиц PM10 на основе графика ниже.



Выглядит не так сложно, и это подтверждает простая линейная регрессия на основе суточных показателей смертности и уровня частиц PM10.

1

2

3

                 Estimate   Std. Error      t value  Pr(>|t|)

(Intercept) 5.08884308354 0.0069353779 733.75138151 0.0000000

pm10tmean   0.00004033446 0.0006913941   0.05833786 0.9535247





Из таблицы выше видно, что коэффициент среднего уровня частиц PM10 (pm10tmean) очень низкий относительно значения его среднеквадратичной ошибки (Std. Error). Фактически связь нулевая.

Однако мы знаем один факт — время года сильно влияет как на уровень смертности, так и на показатели PM10. В частности, смертность возрастает зимой и снижается летом. С уровнем частиц PM10 ситуация обратная: он выше летом и ниже зимой. Так как время года связано с обоими переменными, мы можем ввести его в модель как искажающий фактор.

Перед вами результаты второй модели, которая включает показатели PM10 и время года. Время года представлено фиктивной переменной с 4 уровнями.

1

2

3

4

5

6

                Estimate   Std. Error    t value     Pr(>|t|)

(Intercept)  5.166484285 0.0112629532 458.714886 0.000000e+00

seasonQ2    -0.109271301 0.0166902948  -6.546996 3.209291e-10

seasonQ3    -0.155503242 0.0169729148  -9.161847 1.736346e-17

seasonQ4    -0.060317619 0.0167189714  -3.607735 3.716291e-04

pm10tmean    0.001499111 0.0006156902   2.434847 1.558453e-02





В этом случае коэффициент pm10tmean увеличился, а вместе с ним и его t-значение, что указывает на достаточно сильную зависимость. Как такое возможно?

В нашем случае мы имеем дело с классическим [парадоксом Симпсона](https://en.wikipedia.org/wiki/Simpson%27s_paradox). В общем показатели смертности и PM10 не связаны, но как только мы учитываем для обоих переменных время года, зависимость становится положительной. Такой неожиданный результат связан с тем, что смертность и уровень частиц PM10 по-разному соотносятся со временем года.

Пока что мы рассмотрели только один потенциальный искажающий фактор, но существуют и другие. Например, переменные для погоды, такие как температура и точка росы, также связаны с уровнем смертности и с частицами PM10.

Мы введём значения температуры (tmpd) и точки росы (dptp) в следующую модель, а также добавим переменную для даты (date) на случай, если нужно учесть долгосрочные тенденции.

1

2

3

4

5

6

7

8

9

10

                  Estimate    Std. Error    t value     Pr(>|t|)

(Intercept)  5.62066568788 0.16471183741 34.1242365 1.851690e-96

date        -0.00002984198 0.00001315212 -2.2689856 2.411521e-02

seasonQ2    -0.05805970053 0.02299356287 -2.5250415 1.218288e-02

seasonQ3    -0.07655519887 0.02904104658 -2.6361033 8.906912e-03

seasonQ4    -0.03154694305 0.01832712585 -1.7213252 8.641910e-02

tmpd        -0.00295931276 0.00128835065 -2.2969777 2.244054e-02

dptp         0.00068342228 0.00103489541  0.6603781 5.096144e-01

pm10tmean    0.00237049992 0.00065856022  3.5995189 3.837886e-04





Обратите внимание, что коэффициент pm10tmean вырос ещё больше по сравнению с предыдущей моделью. Связь между смертностью и уровнем PM10 всё ещё присутствует, хотя размер эффекта очень мал. Мы обсудим этот случай позже.

Наконец, ещё одним искажающим фактором могут выступать другие загрязняющие вещества. Прежде чем списывать весь вред на частицы PM10, важно изучить другие загрязнители, которые могут объяснить наши результаты. Хорошим кандидатом будет NO2, так как он выделяется теми же источниками, что твёрдые частицы, и влияет на смертность. Как будут выглядеть результаты, если мы включим его в модель?

1

2

3

4

5

6

7

8

9

10

                  Estimate    Std. Error    t value     Pr(>|t|)

(Intercept)  5.61378604085 0.16440280471 34.1465345 2.548704e-96

date        -0.00002973484 0.00001312231 -2.2659756 2.430503e-02

seasonQ2    -0.05143935218 0.02338034983 -2.2001105 2.871069e-02

seasonQ3    -0.06569205605 0.02990520457 -2.1966764 2.895825e-02

seasonQ4    -0.02750381423 0.01849165119 -1.4873639 1.381739e-01

tmpd        -0.00296833498 0.00128542535 -2.3092239 2.174371e-02

dptp         0.00070306996 0.00103262057  0.6808599 4.965877e-01

no2tmean     0.00126556418 0.00086229169  1.4676753 1.434444e-01

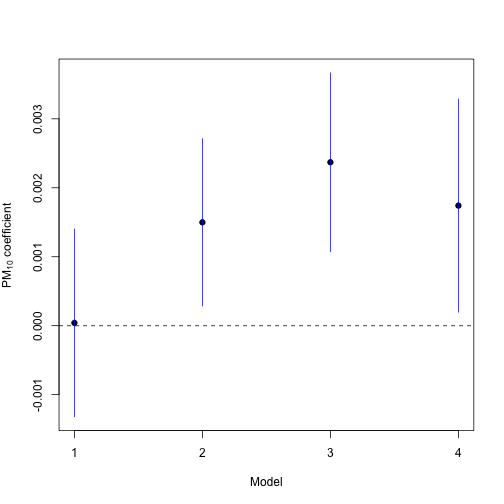
pm10tmean    0.00174189857 0.00078432327  2.2208937 2.725117e-02





Обратите внимание, что в таблице коэффициенты no2tmean и pm10tmean равны по величине, хотя и отличаются t-значениями. Коэффициент pm10tmean остаётся статистически значимым, несмотря на то, что стал меньше.

На графике ниже приведены значения PM10 для всех четырёх моделей, которые мы использовали.

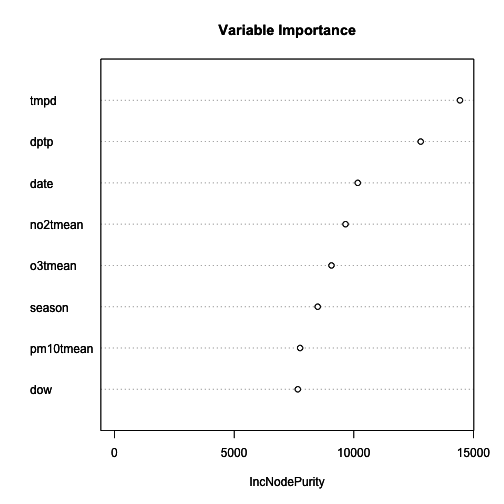


За исключением первой модели, где мы не учитывали потенциальные искажающие факторы, между показателями смертности и PM10 можно установить положительную связь. Как трактовать и использовать эти данные зависит от вашей конечной цели, но об этом мы поговорим в другой главе. Примечательно, что размер эффекта достаточно мал, особенно в сравнении с другими предикторами в модели. Однако важно учитывать, что каждый житель Нью-Йорка дышит этим воздухом, а потому даже маленький эффект может оказать большое влияние.

## Прогнозирование результата

Другой вопрос, который может нас интересовать: "Что эффективнее всего прогнозирует смертность в Нью-Йорке?" TЭто очевидно предиктивный вопрос. Воспользуемся имеющимися данными и построим модель. В этом случае мы воспользуемся моделью случайного леса — методом машинного обучения, который хорошо справляется с большим количеством предикторов. Благодаря такому подходу мы сможем оценить важность переменных. Грубо говоря, эта величина указывает, насколько каждая переменная влияет на прогнозирующую способность модели.

Ниже приведена диаграмма важности переменных, которую мы получили с помощью модели случайного леса. Чем больше значение предиктора по оси X, тем он важнее.



Обратите внимание, что переменная pm10tmean расположена по важности почти в самом низу списка. Всё потому, что она не эффективна в прогнозировании результата, то есть смертности. Вспомните из предыдущего раздела, насколько мал размер эффекта. Это значит, что уровень PM10 не очень объясняет вариативность показателей смертности. Такие предикторы, как температура и точка росы, куда эффективнее в её прогнозировании. Даже уровень NO2 выступает куда более значимым предиктором, нежели PM10.

Однако это совсем не значит, что уровень твёрдых частиц и смертность никак не связаны. Из-за всех компромиссов, на которые приходится идти при построении модели, мы не рассматриваем уровень частиц PM10 как потенциальный предиктор для использования. Мы просто не сможем включить в модель каждую переменную.

## Итоги

При любом анализе данных определяйте для себя тип интересующего вопроса — индуктивный или предиктивный. Это нужно сделать до начала анализа, так как от ответа будет зависеть ход моделирования .Если бы мы придерживались в нашем случае предиктивного вопроса, то пришли бы к ошибочному выводу, что уровень PM10 не связан со смертностью. При этом индуктивный подход указывает на статистически значимую связь этого показателя со смертностью. Выбор правильного вопроса и модели значительно влияет на выводы, которые вы можете получить из данных.

ПРИМЕЧАНИЕ. Отрывки этого материала для самостоятельного изучения взяты из книги Р. Пэна и Э. Мацуи [The Art of Data Science](https://leanpub.com/artofdatascience/).

# Интерпретация полученных результатов

Хотя мы посвятили интерпретации результатов отдельный урок, в действительности её выполняют на протяжении всего анализа данных. Опытные аналитики могут даже не осознавать, как часто они интерпретируют свои выводы, потому что у них это доведено до автоматизма. Вы уже знаете 3 этапа эпициклического процесса: установка ожиданий, сбор информации (данных) и сопоставление ожиданий с данными. Поэтому вы поймёте, что третий этап сам по себе является интерпретацией. В некоторой степени мы говорили об интерпретации на протяжении всей книги. Но эта тема заслуживает отдельного урока, так как интерпретация — это не только сопоставление ожиданий с результатами. Это отдельный важный этап анализа данных. Наиболее осознанно интерпретацию выполняют после того, как завершили основной и дополнительный анализ, но перед тем, как представлять результаты. Поэтому мы разместили эту лекцию между соответствующими уроками. Существует несколько принципов интерпретации результатов, которые мы рассмотрим в этом уроке. Вот эти принципы: 1. Вернитесь к первоначальному вопросу 2. Начните с первичной статистической модели, чтобы наметить ориентиры анализа, и сконцентрируйтесь на характере результата, а не на его бинарной оценке (например, является ли он статистически значимым или нет). Результат оценивают с помощью трёх показателей: направленности, силы связи и неопределённости. Последний выражает вероятность того, что результат получен случайно. 3. Разработайте общую интерпретацию, основанную на (а) совокупности результатов вашего анализа и (б) контексте того, что уже известно о предмете изучения. 4. Подумайте о последствиях вашего анализа. Так вы поймёте, какие действия принять в ответ на полученные результаты. Важно отметить, что эпицикл анализа применим и к интерпретации. Перед каждым этапом интерпретации следует установить ожидания, а затем проверить, соответствуют ли им полученные результаты. Ожидания основываются на том, что вы узнали в ходе анализа данных и формального моделирования. Если интерпретация не совпадает с ожиданиями, вам нужно определить, на каком из этих этапов вы допустили ошибку. Хотя формальная интерпретация — это один из последних этапов работы, возможно, вам будет нужно вернуться к исследовательскому анализу данных или моделированию, чтобы сопоставить ожидания с данными.

## Практический пример: употребление недиетических газированных напитков и индекс массы тела

Самый лёгкий способ научиться применять принципы интерпретации в собственном анализе данных — это увидеть, как они воплощаются в работе. Поэтому мы приведём для каждого из принципов практический пример.

### Вернитесь к изначальному вопросу

Первый принцип — помните об изначальном вопросе. Это может показаться несущественным, но часто аналитики слишком отходят от исходной задачи, когда выполняют исследовательский анализ или формальное моделирование. Обычно причина в том, что специалист слишком увлекается случайным результатом, который получил в ходе исследовательского анализа данных или формального моделирования. В результате итоговая модель отвечает не на исходный вопрос, а на тот, который возник во время анализа. Возвращаясь к исходному вопросу, вы также устанавливаете ориентиры для интерпретации. Например, изначальный вопрос мог быть таким: "Насколько в среднем увеличивается индекс массы тела (ИМТ) взрослого жителя США с каждой 12-унциевой банкой газировки в день?" Судя по вопросу, вы хотели узнать, насколько отличается ИМТ взрослых жителей США, которые ежедневно пьют в среднем две 12-унциевых банки газировки, от тех, кто выпивает только одну банку. Тогда интерпретация анализа должна звучать так: "Каждая 12-унциевая банка газировки, выпиваемая в день жителем США, увеличивает его ИМТ в среднем на X кг/м^2^". Но не так: "Каждая \*унция\* газировки, выпиваемая в день жителем США, увеличивает его ИМТ в среднем на X кг/м^2^". Более того, повторное рассмотрение вопроса помогает вспомнить его тип, что также важно в интерпретации. Например, вопрос звучит так: "Выше ли в среднем ИМТ тех взрослых жителей США, которые выпивают в день на одну 12-унциевую банку недиетической газировки больше?" Формулировка говорит, что это вопрос \*индуктивного\* типа, а значит ваша цель — установить средний эффект от употребления дополнительных 12 унций недиетической газировки в день на ИМТ взрослого населения США. Чтобы ответить на этот вопрос, вы, вероятно, проанализировали перекрёстные данные по репрезентативной выборке взрослого населения США. В таком случае цель интерпретации — объяснить связь между употреблением дополнительных 12 унций газировки в день и ИМТ взрослого населения США. Так как это не причинный вопрос, вы выполняли не причинно-следственный анализ, и, следовательно, интерпретация не должна решать проблему "Что произойдёт, если люди начнут выпивать на одну банку газировки в день больше?" \*Причинный\* вопрос мог бы звучать так: "Какое влияние на ИМТ оказывает употребление дополнительных 12 унций недиетической газировки в день?" Чтобы ответить на него, можно проанализировать данные клинического исследования, в котором участников случайным образом распределили в две группы. В первой люди выпивали по одной дополнительной банке газировки, а во второй — по банке напитка-плацебо. Результаты такого типа вопроса и анализа можно интерпретировать как причинно-следственный эффект на ИМТ от употребления дополнительных 12 унций газировки в день. Поскольку в анализе сравнивается средний эффект на ИМТ испытуемых в двух группах (употребляющих газировку и плацебо), результат можно интерпретировать как средний причинно-следственный эффект по выборке. Третья цель возвращения к изначальному вопросу заключается в том, чтобы сделать паузу и подумать, мог ли ваш подход к получению ответа на вопрос дать \*\*смещённый\*\* результат. Иногда в ходе исследовательского анализа данных и/или моделирования возникает новая информация, которая напрямую влияет на оценку смещённости результата. Напомню, что причиной этому может быть систематическая ошибка сбора или анализа данных, которая приводит к неправильному ответу. Рассмотрим ситуацию с газировкой и ИМТ как пример простейшей систематической ошибки. Предположим, ваш общий вопрос о зависимости между употреблением газировки и ИМТ включал изначальную проблему, которая звучит так: "Сколько недиетической газировки в день употребляют в среднем взрослые жители США?" Допустим, вы анализируете выборку, куда входит всё взрослое население США, и получаете такой результат: среднее количество 12-унциевых банок недиетической газировки, выпиваемых в день, составляет 0,5. Вы делаете вывод, что среднее количество 12-унциевых банок недиетической газировки, которое в день выпивают взрослые жители США, также равно 0,5. Так как всегда необходимо проверять результаты, подумайте о том, включает ли анализ неустранимую систематическую ошибку. Как это сделать? Для начала представьте, что ваш результат неверен, и обдумайте, как при сборе или анализе данных могла возникнуть систематическая ошибка, из-за которой вы неправильно оценили среднее количество 12-унциевых банок недиетической газировки, выпиваемых в день взрослыми жителями США. Хотя подобное упражнение используется в первую очередь, чтобы определить риск систематической ошибки отбора, искажающие факторы или технические проблемы анализа, оно будет полезным на \*\*каждом\*\* этапе работы. Этот мысленный эксперимент выполняется примерно так: представьте, что \*на самом деле\* взрослые жители США выпивают в среднем две 12-унциевые банки недиетической газировки в день. Теперь подумайте о том, почему ваш результат анализа выборки (0,5) настолько далёк от истины. По какой-то причине выборка, которую вы использовали в качестве набора данных, не является случайной для генеральной совокупности и включает непропорциональное число людей, не употребляющих недиетическую газировку. Это занижает оценку среднего количества 12-унциевых порций недиетической газировки, которое употребляется в день. Также можно представить, что результат вашей выборки равен 4, что значительно превышает реальное количество банок газировки, выпиваемых в день взрослыми в США. В таком случае ваша выборка включает непропорционально большое число людей, которые употребляют много недиетической газировки, а потому полученная оценка выше истинного значения. Так как же определить, что ваша выборка неслучайна? Чтобы понять, является ли ваша выборка неслучайной для целевой совокупности, подумайте, почему в ней оказалось больше тех людей, которые не употребляют недиетическую газировку (или употребляют её в больших количествах). Возможно, исследование проводили, чтобы прорекламировать подписку на журнал о фитнесе, а читатели таких журналов склонны употреблять больше диетических газированных напитков. Или же данные собраны в ходе опроса в интернете, и его респонденты реже пьют недиетическую газировку. Также в опросе могли предлагать список недиетических газированных напитков, а респонденты должны были указать, какие из них они пьют. Из-за того что в списке не было Mountain Dew и Cherry Coke, людей, которые предпочитают эти марки газировки, классифицировали как неупотребляющих недиетические напитки (или как употребляющих их меньше, чем в действительности). Причин может быть множество. Мы рассмотрели простейший случай систематической ошибки отбора, встречающийся при оценке доли или среднего значения, но, разумеется, смещённый результат можно получить и при оценке зависимости между двумя переменными. Например, методы опроса могут непреднамеренно завышать выборку людей, которые не употребляют недиетические газированные напитки, но имеют высокий ИМТ (например, люди с диабетом 2-го типа), поэтому результат будет ошибочно указывать на то, что употребление недиетической газировки не связано с высоким индексом массы тела. Очень важно сделать паузу, чтобы целенаправленно выполнить мысленный эксперимент для выявления источников смещения, так как это единственный способ оценить риск получения смещённого результата. Кроме того, этот мысленный эксперимент также следует проводить, когда вы формулируете и уточняете вопрос или выполняете исследовательский анализ и моделирование.

### Начните с первичной модели и оцените направленность, силу связи и неопределённость результата.

Второй принцип заключается в том, чтобы начать с одной модели и комплексно оценить результат, в частности его направленность, силу связи и степень неопределённости. Чем определённее результат анализа выборки, тем он правдивее для всей совокупности. Если сконцентрироваться на каком-либо одном аспекте результата (например, на р-значении), то можно упустить много информации, которая важна для интерпретации. Вы либо проигнорируете, либо не заметите другие важные сведения, выдаваемые моделью. Несмотря на то, что интерпретация не будет полной без результатов всех моделей, бывает полезно сначала сосредоточиться на первичных результатах, которые, по-вашему, полнее всего отвечают на вопрос и отражают полученные данные (см. раздел "Формальное моделирование"). Не тратьте много времени на размышления о том, с какой модели начать, потому что в итоге вы будете рассматривать все результаты. Эта первичная интерпретация нужна лишь для того, чтобы сориентировать вас и помочь сформулировать конечные выводы.

### Направленность

Вернёмся к примеру с газировкой и индексом массы тела и рассмотрим данные выборки с наложенной моделью.

![Sample Data for BMI-soda Example](images/interpret-unnamed-chunk-1-1.png)

Мы сосредоточимся на том, что модель говорит нам о \*\*направлении\*\* зависимости между употреблением газировки и ИМТ, \*\*силе связи\*\* и её \*\*неопределённости\*\*, то есть о том, с какой вероятностью модель отражает реальную зависимость между показателями, а не случайную вариацию, которая могла возникнуть при выборке из большей совокупности. Модель указывает на положительную направленность зависимости, а это значит, что чем больше употребление газированных напитков, тем выше ИМТ. Другим возможным результатом могла бы быть отрицательная направленность или отсутствие направленности (когда значение примерно равно нулю). Соответствует ли положительная направленность результата вашим ожиданиям на основе исследовательского анализа данных? Если да, то у вас всё в порядке и вы можете переходить к следующему этапу интерпретации. Если же нет, этому может быть несколько возможных объяснений. Во-первых, ваши ожидания могут быть ошибочны потому, что либо исследовательский анализ был выполнен неправильно, либо вы его неверно интерпретировали. Во-вторых, ваш исследовательский анализ и его интерпретация могут быть правильными, но формальное моделирование было выполнено неверно. Обратите внимание, что в этом процессе вы снова применяете эпицикл анализа данных.

### Сила связи

После того как вы выявили и устранили все несоответствия между ожидаемой направленностью зависимости и её интерпретацией, нужно проанализировать \*\*силу связи\*\*. Так как модель является линейной регрессией, видно, что наклон прямой (коэффициент бета) равен 0,28. Для интерпретации наклона необходимо знать единицы измерения переменной "газировка" (soda). Если единицами измерения были бы 12-унциевые банки газировки в день, то из интерпретации этого наклона следовало бы, что ИМТ возрастает на 0,28 кг/м^2^ с каждой выпитой в день банкой. Но единицами измерения являются унции газировки, поэтому интерпретация модели говорит о том, что ИМТ возрастает на 0,28 кг/м^2^ с каждой унцией недиетической газировки, выпиваемой в день. Хотя вы уверены, что правильно понимаете единицы измерения переменной газировки и правильно интерпретируете модель, вы всё ещё не можете ответить, как связан ИМТ с каждой дополнительной банкой, а не унцией недиетической газировки. Поэтому вам нужно так преобразовать наклон со значением 0,28, чтобы он относился к увеличению употребления газировки на 12 унций, а не на 1 унцию. Так как это линейная модель, можно просто умножить наклон (коэффициент бета) на 12 и получить 3,36. Результат говорит нам о том, что каждая 12-унциевая банка газировки в день повышает ИМТ на 3,36 кг/м^2^. Другой вариант — это создать для газировки новую переменную, единицей измерения которой будет 12 унций, а не 1 унция. Но умножение наклона — это простое математическое действие, что намного эффективнее. Опять же, после того как вы провели исследовательский анализ данных, у вас должны были появиться ожидания о силе связи между употреблением недиетической газировки и ИМТ. Поэтому нужно определить, соответствует ли им ваша интерпретация силы связи. Если нет, вам нужно определить, совершили ли вы ошибку в ожиданиях или интерпретации. После этого вы сможете предпринять нужные меры, чтобы добиться совпадения обоих точек зрения. Ещё один важный вопрос, который стоит учитывать в отношении силы связи — это её значимость. Например, увеличение ИМТ на 0,01 с каждыми дополнительными 20 унциями в день не особо значимо, так как большой объём газировки связан с очень маленьким увеличением ИМТ. С другой стороны, если речь идёт об увеличении ИМТ на 0,28 кг/м^2^ с каждой дополнительной унцией газировки, то это достаточно значимо. Так как вам известно, что индекс массы тела обычно колеблется в период от старшего подросткового возраста до 30 лет, изменение в 0,01 кг/м^2^ незначительно, а вот разница в 0,28 кг/м^2^ уже может значение. Если принимать во внимание возможное количество выпиваемых газированных напитков, то 0,01 кг/м^2^ за каждые 20 унций газировки — это мало, так как люди (будем надеяться) не выпивают десяток 20-унциевых порций в день, а именно столько нужно, чтобы ИМТ увеличился хотя бы на 0,1 кг/м^2^. С другой стороны, увеличение ИМТ на 0,28 кг/м^2^ с каждой дополнительной унцией быстро сказалось бы на людях, которые выпивают дополнительные 20 унций недиетической газировки в день, так как суммарно их ИМТ увеличился бы на 5,6 кг/м^2^. Таким образом, понимание того, как сила связи результата соотносится с тем, что известно об этом типе информации в интересующей вас генеральной совокупности, является ключевой частью интерпретации силы связи.

### Неопределённость

Теперь, когда вы понимаете, как модель оценивает направленность и силу связи между употреблением недиетических газированных напитков и ИМТ, можно переходить к следующему этапу — рассмотрению степени \*\*неопределённости\*\* вашего ответа. Напомню, что ваша модель была построена так, чтобы соответствовать данным \*выборки\* из общей совокупности. Цель этой модели — выяснить, как употребление недиетической газировки связано с ИМТ \*всего\* взрослого населения США. Давайте вернёмся к примеру с газировкой и индексом массы тела, где с помощью результатов по выборке нужно определить, как на самом деле связаны употребление газировки и ИМТ в общей совокупности взрослого населения США. Предположим, анализ данных показал, что \*в этой выборке\* у людей, выпивающих дополнительную унцию недиетической газировки в день, ИМТ на 0,28 кг/м^2^ выше, чем у тех, кто выпивает на унцию меньше. Но как узнать, является ли этот результат шумом случайной выборки или хорошим приближением истинной зависимости для всей совокупности? Чтобы понять, является ли результат случайным шумом, используются показатели неопределённости. Хотя некоторым кажется, что все случайные выборки идеально представляют генеральную совокупность, на самом деле это не так. Поясним эту идею на простом примере: представьте, что доля женщин среди всего взрослого населения США составляет 51%, и вы составляете случайную выборку из 100 взрослых. В этой выборке 45% женщин. Представьте, что составляете новую выборку из 100 взрослых, и в ней уже 53% женщин. Вы можете составить множество подобных выборок, и в некоторых из них доля женщин будет достигать даже 35 или 70%. Маловероятно, что доля женщин в составленной выборке будет настолько отличаться от показателей общего населения. Но шанс составить выборку, в которой доля женщин будет примерно равна 51%, куда более высок. \*\*Вероятность того, что выборка отражает ответ для общей совокупности, зависит от того, насколько результат для выборки близок к истинному результату для общей совокупности (или далёк от него)\*\*. В этом и заключается неопределённость. Так как мы не знаем ответа для общей совокупности (именно поэтому мы сперва проводим анализ), невозможно выразить неопределённость с точки зрения того, насколько вероятен или не вероятен результат для выборки, отражающий эту общую совокупность. Существуют другие подходы к измерению неопределённости на основе этой общей концепции, и ниже мы обсудим два широко распространённых. Один из способов даёт более непрерывную меру неопределённости — это доверительный интервал. Доверительный интервал — это диапазон значений, который включает результат выборки, и с определённой вероятностью истинный результат для общей совокупности. Чаще всего программное обеспечение для статистического моделирования выдаёт доверительные интервалы со значением 95%. Подобный интервал для полученной выше оценки выборки (0,28 кг/м^2^) составляет 0,15–0,42 кг/м^2^. Приблизительно это можно интерпретировать так: с вероятностью в 95% истинный результат для общей совокупности находится где-то между 0,15 и 0,42 кг/м^2^. Более точное определение доверительного интервала 95% звучит так: если неоднократно повторить эксперимент на основе повторно отобранных выборок (каждый раз собирая набор данных одинакового размера), то истинное значение входило бы в полученный интервал в 95% случаев. Важно понимать, что доверительный интервал строится на основе данных, а значит \*он сам является случайным\*. Следовательно, если собрать новые данные, то интервал на их основе будет немного отличаться. Но истинное значение, оно же значение параметра для совокупности, всегда остаётся неизменным. Другой показатель неопределённости — это р-значение, которое отражает вероятность получить результат выборки, равный 0,28 кг/м^2^ (или более экстремальному значению) при условии, что истинная зависимость между употреблением недиетической газировки и ИМТ в общей совокупности равна 0. Хотя р-значение является непрерывной мерой неопределённости, аналитики чаще работают с р-значением меньше 0,05. Оно означает, что вероятность получить текущий результат выборки (или более экстремальное значение) составляет менее 5% при условии, что в общей совокупности нет никакой статистически значимой зависимости. Эта точка сочленения произвольна и мало говорит о \*степени\* неопределённости или о том, где находится истинный ответ для всей совокупности. Не стоит оценивать неопределённость преимущественно по р-значению. Вы рискуете упустить более важную информацию, которая нужна для вдумчивой и точной интерпретации результатов. Доверительный интервал полезнее р-значения, потому что он формирует диапазон с количественной оценкой того, каким может быть действительный результат для общей совокупности. Кроме того, он выражает вероятность, с которой диапазон содержит результат для общей совокупности. Давайте рассмотрим, как можно использовать р-значение и доверительный интервал 95%, чтобы оценить неопределённость полученной зависимости между употреблением газировки и ИМТ. Предположим, мы получили результат, согласно которому у людей, выпивающих на одну унцию недиетической газировки в день больше, ИМТ в среднем выше на 0,28 кг/м^2^. В этом случае р-значение равно 0,03. Мы решаем измерить неопределённость с помощью р-значения и устанавливаем 0,05 как пороговый показатель статистической значимости. Вот как звучит наша интерпретация: если истинное значение совокупности равно 0 (другими словами, если на самом деле нет связи между употреблением газировки и ИМТ в общей совокупности), то вероятность получить этот (0,28) или более экстремальный результат составляет менее 5%. Теперь повторим это же упражнение с доверительным интервалом 95%. Доверительный интервал 95% для этого анализа составляет 0,15–0,42. Используя его как инструмент оценки неопределённости, мы можем утверждать с уверенностью 95%, что в среднем каждая дополнительная унция газировки увеличивает ИМТ взрослого населения США на некое значение в диапазоне от 0,15 до 0,42 кг/м^2^ с. С помощью такого подхода мы понимаем возможный диапазон эффекта, который оказывает употребление газировки на ИМТ, а также то, что эти два фактора скорее всего связаны между собой в общей совокупности. С другой стороны, использование р-значения как меры неопределённости показывает, что у нас есть только два варианта интерпретации результата: либо неопределённость очень большая, а значит нет никакой зависимости между употреблением газировки и ИМТ, либо что неопределённость результата крайне мала, и зависимость всё-таки существует. Использование р-значения ограничивает нас, потому что оно не отражает процесс взвешивания силы доказательств в пользу гипотезы или против неё. Ещё одно замечание по поводу неопределённости: мы обсуждали оценку этого показателя с помощью традиционных статистических подходов, наиболее распространённым из которых является принцип частотной вероятности. Альтернативой им является байесовский подход, при котором вы меняете свои прежние убеждения на основе результатов анализа. На практике чаще используется частотный подход, который мы обсуждали ранее, но в реальных условиях оба метода выдают в основном одинаковые результаты. Есть одна важная оговорка. Если целью вашего анализа не является формулирование выводов о генеральной совокупности, то оценивать неопределённость необязательно. Если, например, вы хотите понять зависимость между возрастом людей и тем, сколько долларов они тратят на продукцию вашей компании, то у вас уже могут быть все данные об общей совокупности, которая вас интересует (то есть, о клиентах вашей компании). В этом случае не придётся полагаться на выборку, так как ваша компания собирает данные о возрасте и покупках ВСЕХ своих клиентов. Нет необходимости учитывать неопределённость того, отражает ли результат истинное значение для общей совокупности, так как результат \*\*и так\*\* является истинным для общей совокупности.

### Разработайте общую интерпретацию, учитывая совокупность анализов и внешней информации

Теперь, когда вы постарались с интерпретацией результатов первичной модели, можно переходить к следующему этапу — разработке общих выводов с учётом совокупности анализов и сторонней информации. Обработка результатов первичной модели помогает установить ожидания для общей интерпретации, в которой учитываются все анализы. Давайте продолжим рассматривать пример с газировкой и индексом массы тела. Предположим, интерпретация первичной модели говорит о том, что у взрослых жителей США, употребляющий в среднем на одну унцию газировки в день больше, ИМТ обычно выше на 0,28 кг/м^2^. Напомню, что эта первичная модель построена после сбора информации с помощью исследовательского анализа и, возможно, вы уточнили её в ходе интерпретации результатов, оценивая направленность, силу связи и неопределённость результатов модели. В уроке "Формальное моделирование" говорилось, что нет ни одной модели, которая сама по себе даёт ответ на вопрос. Поэтому есть дополнительные модели, которые помогают проверить первичные результаты. Распространённым типом вторичной модели является тот, который определяет, насколько чувствительны первичные результаты к изменениям в данных. Классический пример — удаление выбросов, чтобы оценить степень изменения результата первичной модели. Если бы выяснилось, что на результат модели в основном повлияла небольшая группа людей, употребляющих очень много газировки, то можно было бы предположить, что между показателями нет линейной зависимости, и что употребление газированных напитков влияет на ИМТ только тех людей, кто употребляет их в очень больших количествах. При таком выводе нужно пересмотреть первичную модель. Второй пример — это оценка влияния потенциальных искажающих факторов на результаты первичной модели. Хотя первичная модель должна изначально содержать ключевые искажающие факторы, обычно есть и дополнительные параметры, которые нужно оценить. В примере с газировкой и ИМТ можно построить вторичную модель с учётом дохода, потому что зависимость, наблюдаемая в первичной модели, может полностью объясняться социально-экономическим положением. Люди с более высоким достатком могут употреблять меньше недиетических газированных напитков и иметь более низкий ИМТ, но это не связано с тем, что они пьют меньше газировки. Взамен на ИМТ влияет какой-то другой фактор, связанный с социально-экономическим положением. Чтобы проверить, так ли это, можно построить вторичную модель, в которой к изначальным показателям добавляется доход. Хотя есть и другие способы использования вторичных моделей, эти два одни из наиболее распространённых. Как интерпретировать влияние, которое результаты этих вторичных моделей оказывают на первоначальный результат? Можно вернуться к парадигме направленности, силы связи и неопределённости. Допустим, вы добавили показатели дохода в модель зависимости ИМТ от употребления газировки. Изменилась ли направленность ориентировочной связи между параметрами в сравнении с первичной моделью? Она стала отрицательной или нулевой? Если да, то это кардинальное изменение, которое означает, что либо что-то не так с данными (например, с переменной дохода), либо связь между употреблением газировки и ИМТ полностью обусловлена доходом. Предположим, добавление переменной дохода не повлияло на направленность, но изменило силу связи так, что оценка первичной модели уменьшилась с 0,28 кг/м^2^ до 0,12 кг/м^2^. Сила связи между употреблением газировки и ИМТ уменьшилась на 57%. Это значит, что доход объясняет чуть более половины, но не всю зависимость между употреблением газировки и ИМТ. Теперь переходим к неопределённости. Доверительный интервал 95% для модели с учётом дохода составляет 0,01–0,23, поэтому мы можем быть уверены, что истинная зависимость между употреблением газировки и ИМТ взрослого населения США, не зависящая от дохода, находится где-то в этом диапазоне. Что, если бы соответствующий доверительный интервал составлял от −0,02 до 0,26, но оценка по-прежнему была бы равна 0,12 кг/м^2^? Несмотря на то, что доверительный интервал теперь включает 0, результат первичной модели (0,12) не изменился. Значит, доход никак не объясняет связь между употреблением газировки и ИМТ, а лишь увеличивает неопределённость результата. Одна из причин, по которой добавление переменной дохода в модель могло увеличить неопределённость, заключается в том, что некоторые люди в выборке не предоставили данных о доходах, а поэтому её размер был уменьшен. Проверка числовых показателей поможет определить, так ли это. Также важно рассматривать общие результаты в контексте внешней информации. Внешняя информация — это общие знания по теме, которыми располагаете вы или ваша команда, а также результаты похожего анализа и информация о целевой совокупности. В одном из примеров выше говорилось, что мы можем лучше понять, насколько значима величина эффекта от употребления газировки на ИМТ, если понимаем, как много взрослый житель США её обычно пьёт и сколько может выпить в целом. Кроме того, нелишним будет узнать, какая доля взрослого населения США употребляет недиетическую газировку, и какая страдает от ожирения. Это поможет понять размер совокупности, для которой полученный результат будет потенциально актуальным. Один из интересных примеров того, насколько важно знать размер затрагиваемой совокупности — это загрязнение воздуха. Величина эффекта для связи между загрязнением атмосферного воздуха и критическими проблемами со здоровьем, такими как острые сердечно-сосудистые заболевания (инсульты, инфаркты) мала. Но так как загрязнение воздуха сказывается на сотнях миллионов жителей США, случаев острых сердечно-сосудистых заболеваний, связанных с загрязнениями, достаточно много. Кроме того, вы, вероятно, не первый человек, который пытается ответить на этот вопрос или связанные с ним. Возможно, другие люди уже проводили анализ, чтобы ответить на такой же или похожий вопрос для другой совокупности (скажем, подростков), например: "Какова связь между употреблением недиетической газировки и уровнем сахара в крови?" Очень важно понимать, как ваши результаты соотносятся со всеми знаниями по данной теме. Это поможет вам и другим аналитикам подтвердить или опровергнуть тот факт, что потребление недиетической газировки связано с высоким уровнем сахара в крови, инсулиновой резистентностью, ИМТ и диабетом 2-го типа. С другой стороны, если итоги вашего анализа отличаются от результатов других людей, это тоже важно. Хотя большинство таких случаев вызвано ошибкой, разными методами сбора данных или исследуемыми совокупностями, иногда кардинально отличающийся вывод представляет собой совершенно новое открытие.

### Последствия

Теперь, когда вы интерпретировали результаты и сделали выводы, нужно подумать о последствиях этой работы. В конце концов, смысл анализа обычно заключается в том, чтобы предоставить решение или предпринять меры. Иногда последствия очевидны, а иногда требуют некоторых размышлений. Пример очевидного последствия: вы провели анализ, чтобы определить, увеличила ли продажи покупка рекламы, и, если да, то принесли ли инвестиции в рекламу чистую прибыль. Вы можете узнать, была ли получена чистая прибыль, и если это так, то ваш вывод послужит аргументом в пользу продолжения рекламы. Пример с зависимостью между употреблением газировки и ИМТ, который мы использовали на этом уроке, более сложен. Если окажется, что употребление газировки связано с высоким индексом массы тела и что дополнительная 20-унциевая порция в день увеличивает ИМТ на 0,28 кг/м^2^, то значит, уменьшив потребление газировки, можно снизить и средний ИМТ всей совокупности. Так как вы не проводили причинно-следственный анализ, а лишь подтвердили наличие связи, возможно, понадобится провести дополнительное исследование. В нём вы случайно распределите людей в две группы: одна продолжит употреблять 20 унций недиетической газировки, другая — заменит её на диетический напиток. Если вы работаете в сфере общественного здравоохранения, ваша команда может решить, что пора начать социальную кампанию против газированных напитков и что дополнительные данные клинических исследований не нужны. Вместо этого вы можете отслеживать ИМТ людей в совокупности во время и после кампании, чтобы оценить, насколько сниженное потребление газировки повлияло на здоровье населения. Основной момент здесь заключается в том, что действия, вытекающие из последствий, зачастую зависят от цели организации, которая запросила анализ.

ПРИМЕЧАНИЕ. Отрывки этого материала для самостоятельного изучения взяты из книги Р. Пэна и Э. Мацуи [The Art of Data Science](https://leanpub.com/artofdatascience/).