3

Корреляция

Почему множество каузальных утверждений ошибочны

В 2009 году ученые обнаружили поразительную взаимосвязь между вирусом XMR* и синдромом хронической усталости (СХУ)[1]. Миллионы американцев страдают от этого заболевания с симптомами в виде сильной и постоянной утомляемости, однако причина его неизвестна, и это препятствует профилактике и лечению. Вирусы, недостаточность иммунной системы, генетические факторы и стресс — вот лишь единичные гипотезы, пытающиеся объяснить, что запускает механизм заболевания[2]. И в придачу ко всем соперничающим причинным объяснениям затруднительно даже просто поставить соответствующий диагноз, поскольку нет единого биологического маркера, достоверно тестируемого в лабораторных условиях. Многие случаи остаются незамеченными, и, возможно, СХУ — это в действительности целый букет различных болезней[3].

Группа исследователей во главе с доктором Джуди Миковитц обнаружила, что среди 101 пациента с СХУ вирус XMRV имеют 67% по сравнению со всего 3,7% из 218 контрольных подопытных. Вирус объяснял не все случаи заболевания; была подгруппа пациентов, у которых СХУ стал результатом его действия, у других болезнь не диагностировалась. Для проблемы, в которой оказалось так трудно разобраться, результаты выглядели просто потрясающими, вызвав к жизни массу попыток их подтвердить. Самые разные исследования не смогли обнаружить связь СХУ и XMRV[4]; но в 2010 году ученые выявили похожий вирус, который

---

* XMR (XMRV) — ретровирус; xenotropic murine leukaemia virus-related virus — ксенотропный вирус мышиной лейкемии. Ученым пока неизвестно точно, на самом ли деле вирус вызывает синдром хронической усталости или его размножение идет лучше в организме людей с нарушенной выработкой антивирусного фермента, но полагают, что именно вирус вызывает СХУ. *Прим. ред.*

также превалировал у пациентов с СХУ (86,5%: у 32 из 37) в сравнении со здоровыми донорами крови (6,8%: у 3 из 44)[5].

Эти результаты запустили новый виток гипотез и попыток подтвердить или опровергнуть обнаруженную взаимосвязь. Ученые предположили, что подобная мощная корреляция означает, что именно вирус XMR вызывает СХУ, то есть на этой основе стоит строить лечение. Кое-кто из пациентов, отчаянно желая выздороветь от изматывающей болезни, даже стал требовать у врачей лекарства против ретровируса на основе тестов XMRV.

Выявление у подавляющего большинства людей с СХУ этого вируса в крови — несомненно, интересная находка, которая помогла последующим экспериментам, но эта корреляция не доказывает, что вирус и есть виновник болезни или что антиретровирусное лечение будет эффективным. Вероятно, СХУ ослабляет иммунную систему, повышая подверженность вирусным заболеваниям. Даже если есть некая взаимосвязь, это не дает верного направления; иными словами, она не объясняет, что такое вирус для СХУ — причина или следствие, или же у всего есть общая причина.

В 2011 году оба исследования, выявившие корреляцию между вирусом и СХУ, были отвергнуты после яростных (часто публичных) дебатов. Что касается исследования доктора Миковитц, опровержение было частичным, а в одном случае журнал дал полное опровержение (правда, без согласия автора)[6]. Произошло следующее: пробы СХУ оказались заражены вирусом XMRV, выявив видимые отличия между двумя группами[7]. Помимо этого, был поставлен вопрос о возможной фальсификации данных, поскольку некоторая информация о методе приготовления образцов в подписях к рисункам была опущена, и кое-кто посчитал, что один и тот же рисунок был представлен с несхожими этикетками в разном контексте[8]. Наконец, исследование 2012 года, где различным группам (в том числе группам Миковитц) давались «слепые» образцы для анализа, не обнаружило связи СХУ и XMRV[9].

Интенсивные усилия, подогретые изначальными выводами, и накал страстей во время публичных дебатов между сторонниками и противниками новой теории — яркий пример того, насколько сильна может быть единственная корреляция, которую сочли убедительной.

* * *

Фраза «*корреляция не обязательно означает причинно-следственную связь*» прочно вбита в мозги любого студента, изучающего статистику;

но даже те, кто понимает это высказывание и согласен с ним, порой не могут удержаться от попыток трактовать связи как причинные зависимости. Ученые часто заявляют о корреляциях, много раз поясняя, почему эти соотношения не имеют каузальной взаимосвязи и какой информации для этого недостает. Однако корреляции по-прежнему интерпретируются и используются как причинные зависимости (достаточно лишь проанализировать порой весьма серьезные расхождения между научной статьей и ее популярным вариантом в прессе). Сильная взаимосвязь может показаться убедительной и инициировать ряд успешных прогнозов (хотя в случае с СХУ это не так). Но даже она не объясняет, как работают те или иные вещи и с помощью каких вмешательств их действие можно изменить. Видимая связь между XMR и СХУ не доказывает, что можно вылечить последний с помощью первого, однако пациенты интерпретировали это открытие именно так.

Видимые корреляции могут объясняться еще не измеренными причинами (исключение данных о курении может вызвать взаимосвязь между раком и множеством иных факторов), однако случайные соотношения способны существовать, даже когда две переменные вообще никак не связаны. Корреляции бывают результатом абсолютной случайности (например, вы много раз за неделю сталкиваетесь с подругой на улице), искусственных условий эксперимента (вопросы могут быть подстроены под конкретные реакции), ошибки или сбоя (баг в компьютерной программе).

Иными словами, корреляция — это одно из основополагающих заключений, которые мы способны сделать, и свидетельство в пользу наличия причинной взаимосвязи. В этой главе мы рассмотрим, что такое корреляции и для чего они используются, а также познакомимся с некоторыми из множества путей, посредством которых они возникают без каких бы то ни было причинно-следственных связей.

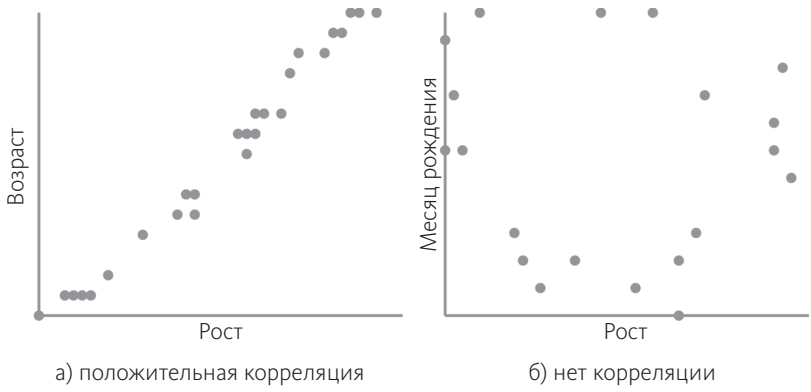
Что такое корреляция

X ассоциируется с раком, *Y связан* с припадками, а *Z привязан* к сердечным приступам. Каждый термин описывает корреляцию, сообщая, что эти явления соотносятся между собой. Хотя и не говоря, как именно.

Суть в том, что две переменные коррелируют, если изменения в одной из них ассоциируются с изменениями в другой. К примеру, рост и возраст детей коррелируют, потому что увеличение возраста

соответствует увеличению роста: дети, как правило, с годами растут. Эти соотношения могут быть выборочными (измерения множества детей различного возраста за один раз), временны́ми (измерения одного ребенка в течение жизни) или учитывать оба фактора (измерения разных людей в течение долгого срока). С другой стороны, между ростом и месяцем рождения нет долговременной корреляции. Это значит, что если месяц рождения варьируется, то рост так регулярно не меняется.

На рис. 3.1 (а) продемонстрировано, как возрастные изменения соотносятся с изменениями роста. Если увеличивается одна переменная, вместе с ней растет и другая. Напротив, на рис. 3.1 (б), где показаны рост и месяц рождения, мы видим набор случайно размещенных точек: месяц рождения варьируется, но соответствующего изменения в росте нет.



а) положительная корреляция б) нет корреляции

Рис. 3.1. Возраст и рост коррелируют, но рост и месяц рождения — нет

Это также означает, что, зная возраст ребенка, мы можем примерно предсказать его рост, а зная месяц рождения — нет. Чем ближе точки друг к другу, формируя линию, тем точнее наши прогнозы (поскольку при этом взаимосвязи теснее). Предсказание — одна из ключевых сфер применения корреляций, и в ряде случаев его можно сделать и без причинных взаимосвязей (хотя не всегда успешно).

Когда корреляции сильны, они могут приобретать видимые очертания, как на рис. 3.1 (а). Но нам необходимы методы измерения этой силы, чтобы провести количественное сравнение и оценку. Существует много единиц измерения корреляций, а одна из них наиболее употребительна — *коэффициент корреляции Пирсона* (обычно его

обозначают буквой *r*)[10]. Этот показатель может иметь значение от 1 до –1. При значении 1 переменные обладают абсолютной положительной корреляцией (положительное изменение одной переменной прямо соответствует положительному изменению другой), а значение –1 говорит об их абсолютной отрицательной корреляции (если одна переменная уменьшается, другая всегда увеличивается).

Получается, коэффициент корреляции Пирсона показывает, как варьируются вместе две переменные по сравнению с индивидуальными модуляциями (эти две меры называются «ковариация» и «вариация»). К примеру, мы можем отметить, сколько часов студенты в некой группе проводят за подготовкой к заключительному экзамену, чтобы посмотреть на соотношение показателей. Зная о наборе экзаменационных баллов и количестве часов, проведенных за подготовкой, но не имея возможности сопоставить итоговые оценки и соответствующие временны́е показатели, мы не определим, есть ли между ними корреляция. В этом случае получится наблюдать индивидуальные вариации каждой переменной, но не их взаимоизменения. То есть мы не можем выяснить, действительно ли большее время, потраченное на занятия, сопровождается более высокими оценками.

БЕЗ ВАРИАЦИИ НЕТ КОРРЕЛЯЦИИ

Скажем, вы хотите узнать, как получить грант, поэтому спрашиваете всех друзей, которые его имеют, что, по их мнению, помогло им. Все кандидаты оформляли заявку шрифтом Times New Roman; согласно мнению половины, важно, чтобы на каждой странице была как минимум одна иллюстрация; а треть рекомендуют представить заявку за 24 часа до установленного срока.

Означает ли это, что есть корреляция между названными условиями и получением гранта? Нет, не означает, потому что, не видя вариации исходного результата, нельзя определить, соотносится ли с ним какой-то иной фактор.

К примеру, если в течение некоей последовательности дней, когда температура доходила до 80 °F (примерно 26,6 °C), на углу улицы стояли две тележки с мороженым, трудно сказать о корреляции погоды и мороженщиков, поскольку нет вариации значения той или другой переменной (температуры или количества мороженщиков). То же справедливо и для случая, когда есть вариация только одной переменной — например, на улице всегда два мороженщика, а температура

изменяется от 80 до 90 градусов. Этот сценарий показан на рис. 3.2: отсутствие вариации ведет к тому, что данные скопились в одной точке, а модуляция единственной переменной дает горизонтальную линию[11]. Именно такой вариант в примере с грантом. Поскольку все результаты идентичны, нельзя сказать, что произойдет, если поменять шрифт или представить заявку за минуту до истечения срока.

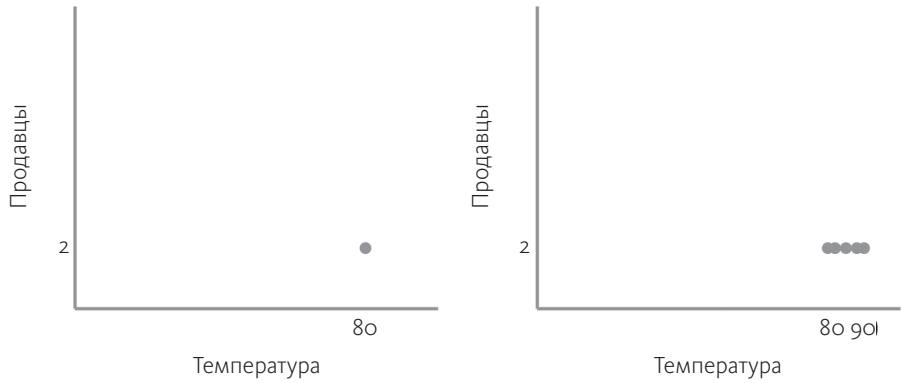


Рис. 3.2. Не наблюдая вариации обеих переменных, нельзя обнаружить корреляцию

И тем не менее широко распространена ситуация, когда анализируются только факторы, ведущие к определенному исходу. Только представьте, насколько часто победителей спрашивают, как именно они добились успеха, а потом стараются этот успех воспроизвести, выполняя в точности те же действия. Подобный подход полон недостатков по многим причинам, включая то, что люди просто не слишком хорошо умеют определять существенные факторы, недооценивают роль случайностей и переоценивают свои способности[12]. В результате мы не только путаем факторы, которые по чистой случайности сопутствуют желаемому эффекту, с теми, которые действительно его обеспечивают, но и видим иллюзорные корреляции там, где их нет.

К примеру, многие интересуются, действительно ли музыкальное образование соотносится с профессиональными успехами в других областях. Даже если мы обнаружим, что многие успешные люди (как бы мы ни определяли успех) играют на музыкальных инструментах, эти ничего не скажет о существовании корреляции — не говоря уже

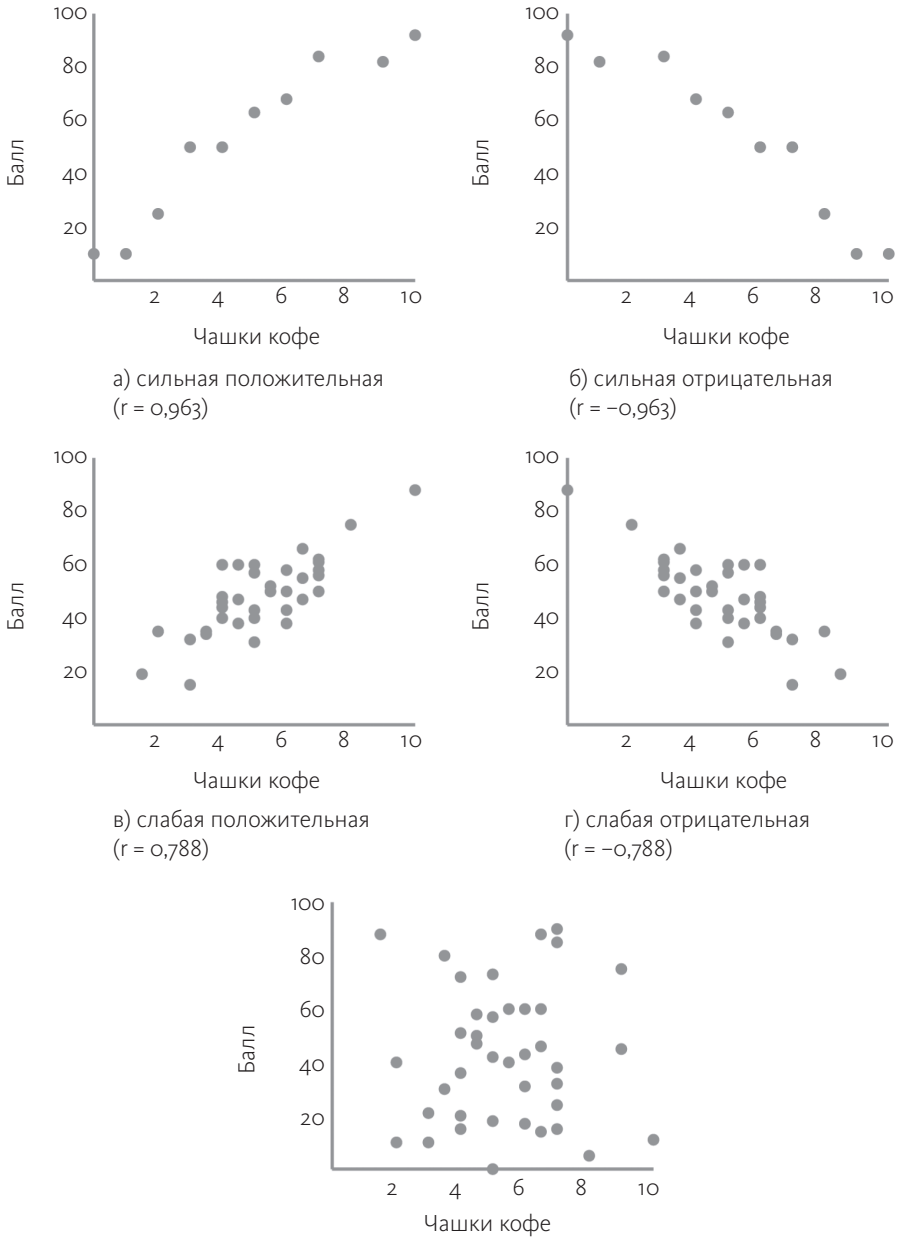
о причинно-следственной связи. Если напрямую спросить, верят ли они, что музыка помогает развивать и другие способности, многие, безусловно, отметят некую взаимосвязь. Но с гораздо меньшей вероятностью они сделают это, если интересоваться конкретно умением играть в шахматы, быстро бегать или тем, сколько кофе вы выпиваете каждый день.

Для целей этой книги важнее всего следующее: беседы с победителями бесполезны, поскольку можно сделать то же самое, но не преуспеть. Возможно, все кандидаты оформляют заявки на грант шрифтом Times New Roman (а значит, те, кто не получил гранты, порекомендуют использовать другой шрифт), а может, успешные кандидаты получили грант, несмотря на избыточное количество иллюстраций в документах. Не зная совокупности положительных и отрицательных примеров, мы не сможем даже предположить наличие корреляции.

КОРРЕЛЯЦИИ: ИЗМЕРЕНИЕ И ИНТЕРПРЕТАЦИЯ

Скажем, мы исследуем студенческий пул, чтобы выяснить, сколько чашек кофе молодые люди выпивают перед финальным экзаменом, а потом регистрируем полученные баллы. Гипотетические данные этого примера представлены на рис. 3.3 (а). Корреляция очень сильна и равна почти 1 (0,963, если быть точными), поэтому точки на графике тесно окружают некую невидимую линию. Если взять обратное отношение (0 чашек кофе соответствуют 92 экзаменационным баллам, а 10 чашек — 10 баллам), чтобы сформировать отрицательную ассоциацию, абсолютное значение окажется тем же, а единственное, что изменится, — знак коэффициента корреляции. Тогда показатель измерения будет равен почти –1 (–0,963), а кривая станет отраженным по горизонтали вариантом положительно коррелирующих данных, как показано на рис. 3.3 (б).

С другой стороны, если бы каждое из этих отношений стало слабее и имела место повышенная вариация результатов экзамена для каждого уровня потребления кофе, наблюдалась бы дисперсия точек, и корреляция была бы слабее. Это продемонстрировано на рис. 3.3 (в), где точки на графике по-прежнему имеют в основном линейную форму, но отклоняются от центра гораздо дальше.



а) сильная положительная (r = 0,963)

б) сильная отрицательная (r = –0,963)

в) слабая положительная (r = 0,788)

г) слабая отрицательная (r = –0,788)

д) нет корреляции (r = 0)

Рис. 3.3. Корреляции между потреблением кофе и экзаменационными баллами

Как и ранее, инверсия отношения (потребление кофе коррелирует с худшими оценками) формирует кривую на рис. 3.3 (г), где единственным отличием оказывается нисходящий уклон.

Заметим, что, если отношение слабое, гораздо труднее перейти от значения потребления кофе до экзаменационных баллов и обратно. Это четко видно, если в первых примерах выбор значения одной из переменных сильно ограничивает вероятные значения другой. Но если мы попытаемся предсказать экзаменационные баллы для 4 чашек кофе с более слабой корреляцией, прогноз будет гораздо менее точен, поскольку мы наблюдали более широкий диапазон баллов для такого уровня потребления кофе. Предел для этой возрастающей вариации — пара переменных, которые абсолютно не соотносятся (имеют нулевой коэффициент корреляции), как показано на рис. 3.3 (д), при этом нельзя вообще ничего сказать о результатах экзаменов на основе выпитого кофе.

Или мы захотели узнать, насколько сильна корреляция между тем, где человек живет, и его умением водить машину. Мера, о которой мы говорили до сих пор, применяется для неквантованных* данных, таких как цены на акции, а не дискретных, таких как местонахождение или киножанр. Если у нас всего две переменные, каждая из которых принимает только два значения, лучше взять упрощенный вариант коэффициента корреляции Пирсона — так называемый *фи-коэффициент***.

Например, можно проверить соотношение между местом, где люди живут, и их умением водить машину. Местом жительства может быть либо город, либо пригород / сельская местность, а факт вождения может либо иметь место (да), либо нет. Как и ранее, проверяем, как эти условия варьируются. Здесь вариация означает частоту, с которой они наблюдаются совместно (а не то, как значения увеличиваются или уменьшаются).

В табл. 3.1 показано, какой вид могут принимать данные. Фи-коэффициент для них составляет 0,81. Мы изначально смотрим, сосредоточено ли большинство измерений вдоль диагональной линии на таблице. Если значения в основном находятся в группах вождение/

* Неквантованный — то есть не преобразованный из непрерывной формы в дискретную (прерывную), не оцифрованный. *Прим. ред.*

** Фи-коэффициент применяется для анализа связи между двумя бинарными переменными. *Прим. науч. ред.*

не-город и не-вождение/город, можно говорить о положительной корреляции.

Если аккумулируются вдоль другой диагонали, корреляция имеет такую же силу, но другой знак.

Таблица 3.1. Различные комбинации местонахождения и вождения

	Пригород / сельская местность	**Город**
Водит машину	92	6
Не водит машину	11	73

Однако на основе этих измерений не каждая сильная корреляция будет иметь высокое значение. Применение коэффициента Пирсона предполагает, что это отношение линейно, а значит, если одна переменная (например, рост), увеличивается, другая (например, возраст) также увеличивается, причем с одинаковым темпом. Это не всегда справедливо, поскольку могут встречаться и более сложные, нелинейные отношения. К примеру, если из-за нехватки кофе человек становится вялым (и не способен показать хорошие результаты на экзамене), а избыток кофе его возбуждает (и тоже плохо влияет на результаты), то график, выстроенный на основе некоторых данных, может иметь вид, как на рис. 3.4. Здесь видно повышение балла в диапазоне от 0 до 5 чашек кофе, потом еще одно медленное падение. Хотя корреляция Пирсона для этого примера нулевая, данные показывают четкий паттерн.

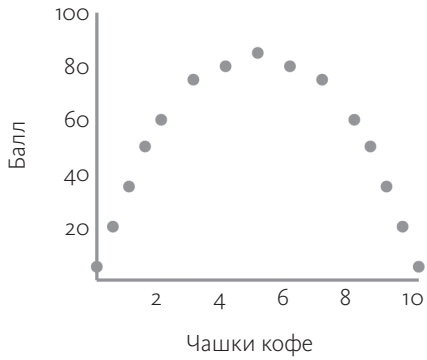


Рис. 3.4. Нелинейное отношение (r = 0,000)

Подобный тип отношений показывает неоднозначные результаты при многих методах причинных умозаключений. В последующих главах мы вернемся к этому. Его важно иметь в виду, поскольку он встречается в таких прикладных науках, как биомедицина (например, и недостаток, и передозировка витаминов могут иметь последствия для здоровья) и финансы (например, кривая Лаффера, которая показывает зависимость между доходами государства и динамикой налоговых ставок).

Аналогично, если вес детей всегда увеличивается с возрастом, но экспоненциально (дети растут, и их вес растет все сильнее), корреляция Пирсона будет ниже ожидаемой, так как она работает в линейных зависимостях. Это одна из опасностей, подстерегающая тех, кто бросает данные в «черный ящик» и просто принимает любые полученные результаты, не проводя дальнейших исследований. Поступив так, когда корреляция недооценивается или даже кажется равной нулю, мы упускаем потенциально интересные зависимости.

Это одна из причин, почему нельзя интерпретировать нулевую корреляцию (пирсоновскую или любую другую) как вообще незначимую (существуют и другие причины, например ошибки в измерениях или первичные данные, искажающие результаты). Еще одна важная причина заключается в том, что данные могут не быть репрезентативными с точки зрения исходного распределения. Если бы нам разрешили взглянуть на статистику смертей от гриппа, но предоставили только данные о количестве больных, поступивших в лечебные учреждения, и вызовов скорой помощи, мы наблюдали бы гораздо более высокий процент летальных исходов, чем в масштабах всего населения. Это происходит потому, что люди оказываются в стационаре, как правило, с более тяжелыми случаями или дополнительными заболеваниями (и с высокими шансами смерти от гриппа). Итак, мы снова сравниваем не все исходы, а только статистику для больных или обратившихся к врачам на фоне симптоматики гриппа.

Чтобы проиллюстрировать эту проблему в ограниченном диапазоне, возьмем, к примеру, две переменные: общий экзаменационный балл и часы, потраченные на подготовку. Однако вместо данных по всему спектру оценок за экзамен мы имеем только сведения о лицах, получивших общий балл за письменный и устный тест по математике выше 1400. На рис. 3.5 эта область показана серым цветом.

Согласно этим гипотетическим показателям, студенты с высокими баллами представляют собой комбинацию как лиц с природной

одаренностью (которые преуспевают, особо не утруждаясь), так и тех, кто получил лучшие оценки за счет интенсивных занятий. Если воспользоваться только данными из закрашенной области, мы не обнаружим никакой корреляции между переменными; но если применить информацию по всему спектру экзаменационных показателей, созависимость будет сильной (корреляция Пирсона оценки и упорных занятий для закрашенной области равна 0, а для всего набора данных — 0,85).

Оборотная сторона медали — это корреляции, которые мы порой находим между несвязанными переменными, опираясь только на следствия (то есть принимая во внимание только случаи, когда это следствие имеет место). К примеру, получение высокого экзаменационного балла и участие во множестве факультативных мероприятий обеспечивают прием в престижный университет. Значит, данные, взятые только в вузах, покажут корреляцию между высоким баллом и многочисленными факультативами, так как здесь эти показатели чаще всего в наличии.

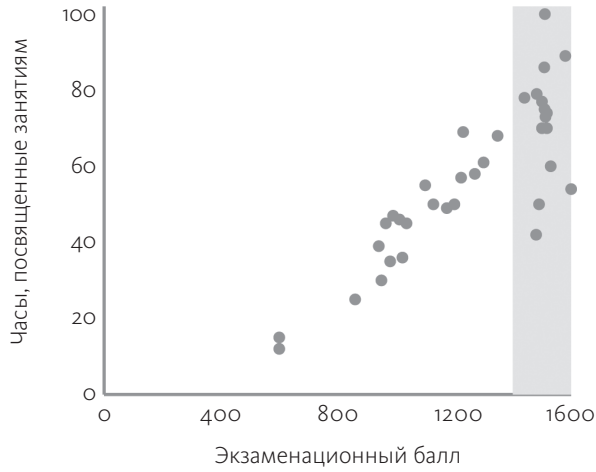


Рис. 3.5. Закрашенная область представляет ограниченный диапазон данных

Подобная тенденция отбора данных довольно типична. Возьмем, к примеру, сайты, опрашивающие посетителей насчет их политических взглядов. В интернете не получится отобрать участников опроса случайно в масштабах всего населения, а данные источников с сильным

политическим уклоном искажены еще сильнее. Если посетители конкретной страницы активно поддерживают действующего президента, то результаты по ним, возможно, покажут, что рейтинг главы государства растет каждый раз, когда он произносит важную речь. Однако это показывает лишь то, что есть корреляция одобрения президента и произнесения им речей перед сторонниками (поскольку на вопросы отвечают представители всего населения). Мы рассмотрим и эту, и другие формы трендов (например, смещение по выживаемости) в главе 7 и увидим, как они влияют на результаты анализа экспериментальных данных.

* * *

Важно помнить, что, помимо математических причин, по которым можно распознать ложные корреляции, есть еще наблюдение за данными, позволяющее найти ложные паттерны. Некоторые из когнитивных смещений, заставляющие нас видеть соотношение несвязанных факторов, также сходны с ошибкой отбора. К примеру, предвзятость подтверждения заставляет искать доказательства в пользу определенного убеждения. Иными словами, если вы верите, что лекарство вызывает некий побочный эффект, вы приметесь читать в интернете отзывы тех, кто уже принимал его и наблюдал это действие. Но таким образом вы игнорируете весь набор данных, не поддерживающих вашу гипотезу, вместо того чтобы искать свидетельства, которые, возможно, заставят ее переоценить. Предвзятость подтверждения также может заставить вас отказаться от свидетельств, противоречащих вашей гипотезе; вы можете предположить, что источник сведений ненадежен или что исследование основывалось на ошибочных экспериментальных методах.

Помимо предвзятости с точки зрения доказательств, может случиться ошибка интерпретации аргументов. Если в ходе «неслепого» тестирования нового лекарства доктор помнит, что пациент принимает это средство и считает, что оно ему помогает, то может начать искать признаки его эффективности. Поскольку многие параметры субъективны (например, подвижность или усталость), это может привести к отклонениям в оценке данных индикаторов и логическим заключениям о наличии несуществующих кореляций[13]. Этот пример взят из реального исследования, где доктора, выведенные из слепого метода, сделали вывод об эффективности препарата (мы подробнее обсудим ситуацию в главе 7). Таким образом, интерпретация данных может различаться в зависимости от убеждений, что приводит к отличиям в результатах[14].

Есть и специфическая форма предвзятости подтверждения — *иллюзорная корреляция*. Она означает поиск соотношения там, где его нет. Возможная взаимосвязь симптомов артрита и погоды настолько широко разрекламирована, что считается доказанной. Однако знание о ней может привести к тому, что пациенты будут говорить о корреляции просто из ожидания ее увидеть. Когда ученые попытались проанализировать эту проблему, взяв за основу обращения пациентов, клинические анализы и объективные показатели, то не обнаружили абсолютно никакой связи (а другие выяснили, что истинным виновником могла быть сырость, хотя и этот вывод не окончателен)[15]. А когда студентам колледжей показали данные из анкет пациентов, где отмечались одновременно болевые симптомы и атмосферное давление, те не только увидели корреляции там, где их не было, но и представили разные интерпретации одних и тех же последовательностей как положительно или отрицательно соотносящихся.

Это подобно ошибке отбора, поскольку одной из причин выявления неверной корреляции может быть концентрация на одном сегменте информации. Если вы прогнозируете отрицательное соотношение переменных, легко сосредоточите внимание на небольших сегментах целого, подтверждающих ваш прогноз. И такой случай относится к предвзятости подтверждения: можно сфокусировать внимание на определенных данных, повинуясь сформированным убеждениям. В случае с артритом и погодой люди, возможно, придают слишком большое значение определенным фактам (отбрасывая проявившиеся симптомы при хорошей погоде и придавая особое значение таким же при плохой) или видят доказательства там, где их нет (по-разному отмечают заболевание в зависимости от погоды и от ожидаемой связи того и другого).

Как пользоваться корреляциями

Скажем, мы действительно обнаружили соотношение между сроком представления заявки на грант и его получением. Действительно, чем раньше подана заявка, тем выше она будет оценена, поэтому коэффициент корреляции здесь и вправду будет равен единице. Значит, можно безошибочно предсказать, что некто получит грант, если подаст заявку за неделю, да?

Именно на это рассчитывают многие ретейлеры, пытаясь выявить индикаторы, которые спрогнозируют поведение покупателей. Реклама

компании Target не сходила с газетных полос, когда ее представители заявили, что «узнали» о беременности девочки-подростка раньше, чем ее семья[16]. Разумеется, в Target на самом деле понятия не имели об этом; просто воспользовались огромным пулом сведений, собранных от других покупателей (и из других источников), чтобы выяснить, какие факторы коррелируют с разными стадиями беременности. На основе приличного объема наблюдений компания смогла, например, выяснить, что покупка либо лосьона, либо ватных шариков сама по себе не значимый факт, но беременные женщины часто выбирают оба эти предмета вместе с определенными витаминными добавками. Имея достаточно данных о покупательных паттернах и соответствующих сроках (это можно выяснить из записей о рождениях или спрогнозировать на основе информации о приобретении тестов на беременность), компания может определить вероятность беременности покупательницы и даже оценить, на каком она сроке. Даже если просто знать, что девушка приобрела два теста один за другим, это позволит сделать вывод, что первый оказался положительным.

Корреляции используют, например, Amazon, Netflix и LinkedIn, предлагая дополнительные товары, фильмы, которые могут вам понравиться, или потенциальные контакты.

Netflix, к примеру, может найти людей, которым нравятся те же фильмы, что и вам, и предложить вам киноленты, на которые эти люди дали хорошие отзывы. Именно это позволило ученым повторно идентифицировать людей в деидентифицированном наборе данных Netflix, воспользовавшись информацией из другого источника — IMDb*[17]. Алгоритмы вообще-то сложнее, чем те, о которых мы рассказали, но основная идея именно такова. Правда, эти компании не обязательно волнуют причины, по которым вы совершаете некие действия. Netflix может порекомендовать достаточно фильмов, которые вам понравятся, не потрудившись выяснить, что после напряженного дня вы смотрите только сериалы.

Есть, однако, немало примеров, когда предсказания, основанные на корреляциях, не оправдываются — даже если не уточнять, соответствуют ли соотношения причинным зависимостям. Одна из опасностей в том, что для любой корреляции между двумя переменными можно с большой вероятностью придумать ситуацию, когда такая взаимосвязь возникнет, а это ведет к ложной вере в результат.

---

* IMDb (The Internet Movie Database) — интерактивная база данных, связанная с фильмами, телевизионными программами и видеоиграми, включая актеров, производство, биографии, сюжет и рецензии. *Прим. ред.*

Известен пример из области анализа данных, когда сведения о продажах в бакалейном магазине помогли выяснить, что люди часто покупают пиво и подгузники одновременно. Так возник миф, что мужчины, которые накануне выходных запасаются подгузниками, обязательно купят хоть немного пива в качестве награды за поход в магазин. Но, вернувшись в 2002 году к истокам этого случая, Дэниел Пауэр обнаружил, что изначальная корреляция ничего не говорила о гендерной принадлежности покупателей или в какой день недели совершались покупки. К тому же никогда не предпринимались попытки использовать ее для повышения прибыли — передвинув товары на полке магазина ближе друг к другу. Купленными товарами могли с тем же успехом оказаться попкорн и бумажные салфетки (для вечера перед телевизором) или яйца и таблетки от головной боли (для лечения похмелья).

Скажем, Amazon обнаружил сильную корреляцию между покупкой дисков с сериями телешоу, где действие происходит в колледже, и приобретением учебников для подготовки к экзамену по углубленной программе. Ясно, что продажи обоих товаров обеспечивают американские тинейджеры, но Amazon вполне может этого не выяснять, если единственная задача — дать рекомендации той же группе покупателей, на базе которой собирались маркетинговые данные. Если, однако, компания будет рекомендовать учебники покупателям из других стран, это не обеспечит вала продаж, поскольку такие экзамены сдают в основном ученики из США.

Итак, даже если корреляция истинна и надежна, она может оказаться бесполезной для прогнозирования, если мы попытаемся перенести ее на другую группу населения, не обладающую нужными свойствами для срабатывания взаимосвязи (подробнее об этом в главе 9). Корреляция ничего не говорит о том, почему эти предметы взаимосвязаны, то есть почему покупатели — именно конкретные подростки 16–17 лет, которые готовятся к экзаменам по углубленной программе, а также любят телешоу с персонажами такого же возраста. Значит, ее трудно применять для прогнозирования в других ситуациях.

Мы привели весьма однозначный пример, однако были и другие, с менее четким механизмом действия. В 1978 году спортивный журналист в шутку предложил новый индикатор фондового рынка: если команда Американской футбольной лиги выигрывает Супербоул*,

* Супербоул (*англ.* Super Bowl) — так в американском футболе называется матч за звание чемпиона Национальной футбольной лиги. Матч и сопутствующие ему торжественные мероприятия Super Bowl Sunday превратились в США в национальный праздник. *Прим. перев.*

к концу года рынок упадет; если нет — пойдет вверх[18]. Нет никакой специфической причины, по которой между этими событиями должна быть связь, но, если взять все возможные индикаторы поведения рынка, именно этот работает достаточно часто, убеждая некритично настроенную аудиторию. И все же без понимания того, почему это срабатывает, мы никогда не сумеем предсказать, в какие годы конкретный паттерн даст сбой. Может ведь оказаться, что с того момента, как этот индикатор получил широкую известность, знание о корреляции (пусть и безосновательно возведенной в ранг достоверных) влияет на поведение.

Аналогичные сомнения возникают, когда нужно использовать данные наблюдений (например, поисковые результаты в интернете или посты в соцсетях) для выявления трендов. Простое знание о том, что люди этим занимаются, приводит к изменениям в пользовательском поведении (возможно, благодаря освещению в СМИ), а также к злонамеренным азартным играм в системе.

Итак, хотя корреляции способны быть полезными для прогнозирования, прогнозы могут оказаться неверными, а измеренная корреляция — ложной.

Почему корреляция не причинно-следственная связь

Когда я читала лекцию о причинном осмыслении, один студент задал вопрос: «Разве Юм не утверждал, что причинность — всего лишь корреляция?»

И да, и нет. Да, причинно-следственная связь возможна, но мы не можем знать наверняка. А то, что мы способны наблюдать, — по сути, корреляция (или особый вид закономерности). Это, однако, не означает, что причинность представляет взаимосвязь только потому, что мы способны ее наблюдать. Это говорит еще и о том, что в большинстве работ, связанных с выявлением и оценкой причинных зависимостей, разрабатываются способы отличия каузальных корреляций от остальных.

Это можно проделать на основе экспериментов или статистических методов, но дело не только в том, чтобы выявить корреляцию. В этой книге мы проанализируем ситуации, в которых причинно-следственная связь кажется очевидной, но в реальности ее нет. В последующих

главах мы также рассмотрим некоторые случаи, когда соотношения возникают без соответствующей причинной зависимости.

Первое — меры корреляции симметричны. Соотношение роста и возраста в точности соответствует зависимости между возрастом и ростом. С другой стороны, причинно-следственная связь может быть асимметрична. Если кофе вызывает бессонницу, это не значит, что бессонница также должна стать причиной потребления кофе, хотя такое может случиться, когда не выспавшийся ночью человек утром вынужден пить больше кофе.

Точно так же любая мера значимости причин (например, условные вероятности) отличается в двух направлениях. Если мы выявили корреляцию, не имея никакой информации о том, какой фактор имеет место в начале, то с равной вероятностью каждый из них может оказаться причиной другого (или будет наличествовать петля обратной связи), а мера взаимосвязи сама по себе не дает представления о различиях между двумя (или тремя) возможностями.

Если мы попытаемся придумать историю причинной взаимосвязи для пары коррелирующих вещей, нам придется, основываясь на базовых знаниях, предположить, какая из них, вероятнее всего, повлечет за собой другую. Например, даже если пол человека связан с риском инсульта, трудно представить, чтобы инсульт определял пол. Но если мы выявили соотношение между набором веса и пассивным образом жизни, никакие данные о том, как коррелируют эти факторы, не скажут о направленности найденной взаимосвязи.

Ошибочные корреляции могут возникать по многим причинам. В случае с СХУ и вирусом XMR соотношение возникло из-за загрязнения экспериментальных образцов. В других ситуациях это мог быть баг в компьютерной программе, ошибки в расшифровке результатов или некорректный анализ данных. Видимая связь может также возникнуть из-за статистических отклонений или простого совпадения, как в примере с фондовым рынком и футболом. Но есть еще одна причина — необъективность. Иногда, если выборка нерепрезентативна, мы можем увидеть корреляцию там, где ее нет. Точно та же проблема приводит к обнаружению соотношения и без причинной зависимости.

Важно понимать, что причинно-следственные связи не единственное, хотя и возможное в ряде случаев, объяснение корреляций. К примеру, мы нашли соотношение в ситуации, когда человек, съевший плотный завтрак, вовремя успевает на работу; однако, вероятно, оба

фактора имеют общую причину: человек рано встал, а значит, у него было время хорошо позавтракать, вместо того чтобы в спешке бежать на службу. Выявив корреляцию между двумя переменными, нужно проверить, способен ли подобный неизмеренный фактор (общая причина) объяснить эту взаимосвязь.

В ряде случаев (о которых мы поговорим в главе 4) таким общим фактором оказывается время. Можно обнаружить множество ошибочных корреляций между факторами с устойчивыми по времени тенденциями. К примеру, если количество пользователей интернета всегда увеличивается и национальный долг — тоже, эти факторы будут взаимосвязаны. Но в целом мы ссылаемся на переменную или набор переменных, объясняющих корреляцию. Например, можно задуматься: действительно ли усердное учение обеспечивает лучшие оценки, или более вероятно, что лучшие студенты и усердно учатся, и получают высокие оценки. Возможно, врожденная способность становится общей причиной и оценок, и времени, проведенного за учебниками. Если бы была возможность изменить способность, это могло повлиять и на оценки, и на время обучения, но любое экспериментирование с оценками и усердием в учении не оказало бы никакого воздействия на два других фактора.

Аналогичная причина корреляции без прямой причинной зависимости — промежуточная переменная. Скажем, проживание в городе соотносится с низким индексом массы тела (ИМТ), поскольку горожане больше ходят, чем ездят на машине, и проявляют высокую физическую активность. Таким образом, жизнь в городе косвенно приводит к низкому ИМТ, однако переезд в город и постоянное использование транспорта — плохая стратегия для желающих похудеть. Бóльшую часть времени мы ищем косвенные причины (например, курение вызывает рак легких, а не особые биологические процессы, посредством которых и происходит воздействие), но, если знать механизм (как именно причина производит следствие), можно найти лучшие пути для вмешательства.

Наконец, агрегированные данные могут приводить к странным результатам. В статье за 2012 год в журнале New England Journal of Medicine рассказывалось о поразительном соотношении между количеством шоколада на душу населения и числом Нобелевских лауреатов на 10 000 000 жителей[19]. Коэффициент корреляции составлял 0,791. Этот показатель возрос до 0,862 после исключения статистики по Швеции — стране, давшей гораздо больше лауреатов

престижной премии, чем ожидалось, судя по статистике потребления шоколада.

Заметим, однако, что данные о шоколаде и Нобелевских премиях были взяты из различных источников, где каждая страна оценивалась отдельно. Это означает, что на самом деле мы не имеем ни малейшего представления, действительно ли потребители шоколада и лауреаты Нобелевки — представители одной и той же группы. Далее, количество награжденных — лишь малая доля населения, а значит, несколько дополнительных премий могли драматичным образом изменить расчеты. Большинство сообщений об отмеченной корреляции фокусировалось на потенциальном наличии причинной взаимосвязи между потреблением шоколада и получением награды, подавляя заголовками вроде «Шоколад делает нас умнее!»[20] и «Хотите Нобелевку? Ешьте больше шоколада!»[21]. Работа ученых, однако, не поддерживает ни одно из подобных утверждений, и страны с бóльшим числом лауреатов могли просто отметить это событие увеличенным количеством шоколада (не будем забывать, что коэффициент корреляции симметричен).

Более того, мы не способны ничего сказать о том, действительно ли любовь к шоколаду улучшит шансы на победу, если страны будут стимулировать его потребление у своих граждан, или этот продукт — просто индикатор иного фактора, к примеру экономического положения. Если нужны дополнительные причины, чтобы скептически отнестись к этой корреляции, вот еще факт.

Ученые, специально старавшиеся продемонстрировать всю глупость попыток интерпретировать взаимосвязь как причинно-следственную без дальнейших исследований, обнаружили статистически значимое соотношение между популяцией аистов и уровнем рождаемости[22].

Да, к исследованию про шоколад можно отнестись с юмором. Но подобный вид агрегированных данных часто используется для установления корреляции среди населения, и, по всем указанным причинам, эти данные особенно сложно использовать. Сведения за большой временной интервал несколько упростят задачу (например, росло ли потребление шоколада *перед* присуждением премий), но все равно придется учитывать разнообразные события, которые могут быть поводом для изменений (например, внезапный рост потребления шоколада и одновременная смена образовательной политики). Кроме того, Нобелевские премии часто присуждаются гораздо позже,

чем случаются соответствующие события. Может найтись огромное количество иных условий, которые сформируют аналогичные корреляции. Если говорить об этом исследовании, «анализ по горячим следам» выявил еще одну забавную связь — между Нобелевскими премиями и молоком[23].

Множественные сравнения и *p*-значения

Участника исследования помещают в аппарат МРТ и показывают фотографии различных социальных ситуаций. Он должен определить эмоции, которые выражает человек на каждом кадре. С помощью МРТ ученые измеряют ток крови в локальных областях мозга и часто пользуются этим измерением как показателем мозговой активности[24], чтобы определить, какие области мозга задействованы в решении различного рода задач. Итоговые цветные изображения отражают, в каких областях наблюдается усиленный кровоток: именно это имеют в виду авторы статей, говоря, что некая область мозга «светится», реагируя на определенный стимул. Выявление активируемых областей помогает понять взаимосвязи в мозге.

Исследование обнаружило, что некоторые области мозга участника эксперимента демонстрировали статистически значимые изменения тока крови. Действительно, при том, что значение 0,05 часто используется как пороговое для *p*-измерений* (меньшие показания более значимы), уровень активности, ассоциированный с одной областью, имел *p*-значение 0,001[25].

Может ли эта область мозга быть связана с представлением эмоций других существ («принятие перспективы»)?

Если учесть, что объектом исследования был пойманный лосось, это кажется невероятным.

Так как же дохлая рыбина могла реагировать на визуальный стимул?

Результаты могли бы считаться высокозначимыми с учетом любых обычных пороговых значений, поэтому дело не в попытке преувеличить

---

* *P*-значение (*англ.* p-value) — величина, используемая при тестировании статистических гипотез. Фактически это вероятность ошибки при отклонении нулевой гипотезы (ошибки первого рода). Проверка гипотез с помощью *p*-значения служит альтернативой классической процедуре проверки через критическое значение распределения. *Прим. ред.*

их важность. Чтобы понять, откуда они вообще могли взяться, сделаем небольшое отступление статистического характера.

Исследователи часто надеются определить, имеет ли некий эффект значимость (корреляция истинна, или это результат статистического отклонения), либо просто есть различие между двумя группами (активны ли разные области мозга, когда люди смотрят на людей или на животных). Но, чтобы объективно определить, какие выводы важны, необходима некая количественная мера. Одна из общепринятых мер — так называемое *p*-значение, которое используется для сравнения двух гипотез (нулевой и альтернативной).

P-значение показывает вероятность результата, который как минимум столь же нехарактерен, как и наблюдаемый, при условии истинности нулевой гипотезы.

Для наших целей такие гипотезы могут заключаться в следующем: между двумя вещами существует причинная зависимость (нулевая гипотеза) или нет (альтернативная гипотеза)*.

Еще одна нулевая гипотеза: монета симметрична (альтернативная гипотеза — монета со смещением). *P*-значения часто интерпретируются неверно — как вероятность того, что нулевая гипотеза истинна. Хотя обычно используется пороговое значение 0,05, нет никакого закона, по которому результаты с *p*-значениями меньше 0,05 значимы, а больше 0,05 — нет. Это просто договоренность, и показатель 0,05 редко вызывает возражения у других ученых[26]. Условные знания не соответствуют понятиям «истинно-ложно», поскольку незначимые результаты могут иметь очень маленькие *p*-показатели, а значимый результат иногда не достигает критического уровня.

Фильм «Розенкранц и Гильденстерн мертвы» начинается с эпизода, в котором герои бросают найденную монетку — и оказываются в полной растерянности, когда она 157 раз падает орлом вверх[27]. Вероятность того, что монетка упадет орлом вверх 157 раз подряд, действительно крайне мала ($1 : 2^{157}$, если быть точными), и единственный равно экстремальный результат для 157 бросков — это все решки. То, что наблюдали Розенкранц и Гильденстерн, в самом деле имело

---

* Вообще, более привычна обратная постановка: нуль-гипотеза — причинной зависимости нет, альтернативная — зависимость есть. Таким образом, стандартное исследование сводится к попытке отвергнуть нуль-гипотезу на некотором заранее выбранном уровне. То есть если получаем p=0,000001, значит можем отвергнуть нуль-гипотезу об отсутствии зависимости на уровне 0,001. Иными словами, p-величину можно рассматривать как вероятность получения нехарактерного результата при истинности нуль-гипотезы. *Прим. науч. ред.*

очень низкое *p*-значение. Но это не означает, что обязательно происходило нечто странное — только то, что подобный результат невероятен для симметричной монеты.

Для менее экстремального случая, скажем, мы подбросим монету 10 раз, и выпадут 9 орлов и 1 решка.

P-значение такого результата (здесь нулевая гипотеза — что монета симметрична, а альтернативная — что она смещена в любом направлении) — это вероятность тех самых 9 орлов и 1 решки + вероятность 9 решек и 1 орла + вероятность 10 орлов + вероятность 10 решек[28]. Причина, по которой сюда включены две серии со всеми орлами и всеми решками, в том, что мы рассчитываем вероятность события как минимум такого же экстремального, как и наблюдаемое, а эти серии — самые экстремальные. Наша альтернативная гипотеза — смещение монеты в любом направлении, а не просто в сторону орлов или решек; вот почему мы включили длинные серии решек.

На рис. 3.6 представлены гистограммы для орлов в серии из 10 бросков по 10 монет. Если бы результатом для каждой монеты было в точности 5 орлов и 5 решек, каждый график представлял бы одну черту длиной 10 пунктов с центром на отметке 5. Но в реальности случаются и бóльшие, и меньшие значения, и даже одна серия из всех решек (показанная маленькой чертой, которая пересекает один график справа налево).

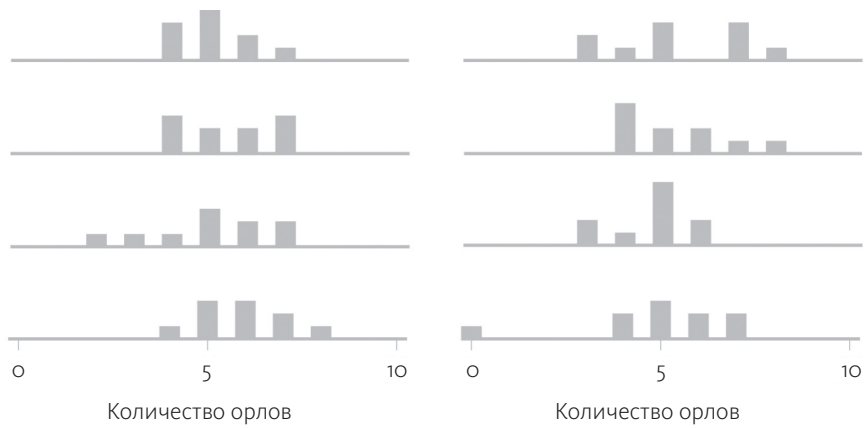


Рис. 3.6. Каждая гистограмма представляет эксперимент, где 10 монет подбрасывают 10 раз. Каждая серия из 10 монет образует точку данных на графике в зависимости от количества орлов. Показано 8 примерных экспериментов

Такое событие все равно невероятно при наличии одной симметричной монеты; но что будет, если мы подбросим 100 монет? Увеличивая число экспериментов, мы создаем больше возможностей, чтобы некое по видимости аномальное событие произошло случайно. К примеру, вероятность того, что конкретный человек выиграет в лотерею, на самом деле мала; но, если играют достаточно людей, можно гарантировать, что кто-нибудь победит. На рис. 3.7 показана такая же гистограмма, но уже для 100 монет. Действительно, будет странно, если мы не увидим как минимум одной серии из 9 или более орлов или решек, когда бросают так много монет (или лотерею, где не будет победителей, если шансы 1 : 1 000 000, а играют 100 000 000 человек).

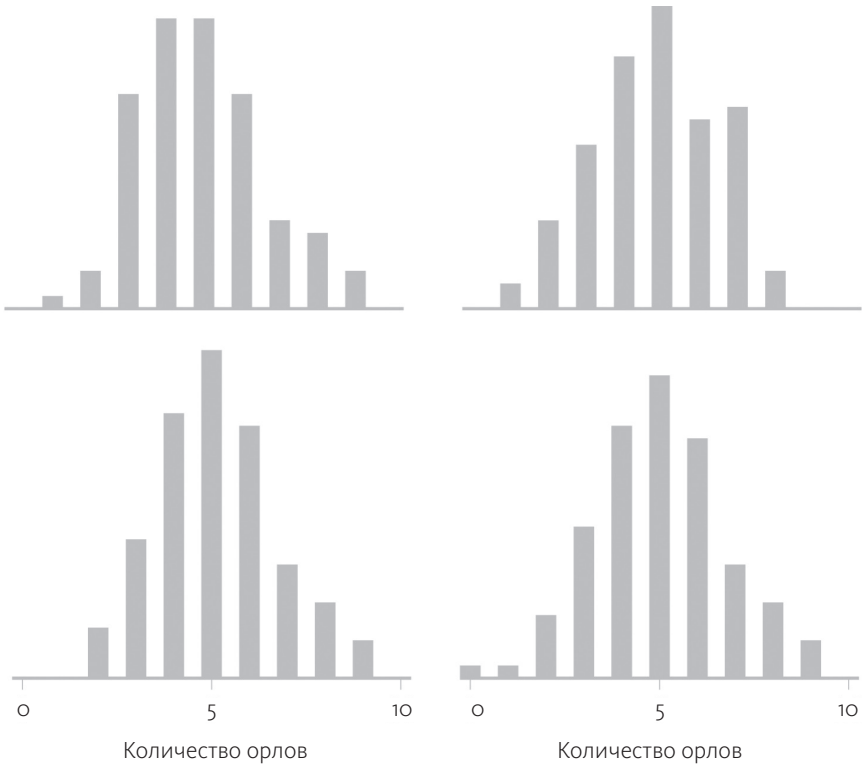


Рис. 3.7. Результаты подбрасывания 100 монет по 10 раз для каждой. Показано 4 эксперимента

Именно проблема одновременного проведения многочисленных тестов и оказалась во главе угла исследования МРТ, с рассказа

о котором мы начали разговор. Проверке подверглись тысячи малых областей мозга (а в исследованиях на людях их еще больше, потому что человеческий мозг включает множество областей), поэтому совсем неудивительно, что одна из них продемонстрировала значительный кровоток. Проблемы такого вида именуются *проверкой многомерной гипотезы*, что означает одновременную проверку большого количества гипотез. Вопрос становится еще более существенным с появлением нового метода, генерирующего громадные наборы информации (например, множества МРТ и экспрессии генов) с так называемыми большими данными. Ранее было возможно в рамках одного эксперимента проверить только одну гипотезу, теперь же, когда мы способны анализировать тысячи переменных, неудивительно, что между ними обнаруживаются корреляции в силу количества проведенных тестов.

В эксперименте с лососем ученые протестировали тысячи гипотез, и каждая утверждала, что некая область мозга проявит значительную активность.

В действительности же исследование доказало: все эти тесты могут дать кажущиеся значимыми результаты по чистой случайности. Было показано, что при использовании статистических методов, корректных для множества сравнений (фактически каждый тест требует более жесткого порогового показателя), значимой активности выявлено не было даже при очень нежестких порогах *p*-значений[29].

Важная вещь, которую стоит запомнить: читая отчет о некой необходимой находке, которая была взята из громадного набора одновременных тестов, обязательно обращайте внимание на то, как авторы решают проблему множественного сравнения. Статистики расходятся во мнении, как именно (и когда) корректировать этот фактор, но все дебаты в целом сводятся к тому, какой тип ошибки хуже. Корректируя множество сравнений, мы, по сути, заявляем о желании снизить количество ложных открытий и готовы мириться с возможностью пропустить из-за этого некие значимые находки (и генерировать ложноотрицательные результаты). С другой стороны, выступая против поправок, заявляем о нежелании упускать истинно положительные результаты за счет нескольких ложных открытий.

Между этими двумя типами ошибок всегда идет поиск компромисса, а предпочтения зависят от индивидуальных целей[30].

Возможно, для эксплораторного анализа, где поиск ведется экспериментальным образом до получения конечного результата, мы считаем нужным, образно говоря, раскинуть обширную сеть. С другой стороны, если мы стараемся отобрать узкоцелевую группу кандидатов для разработки дорогостоящего препарата, каждое ложное умозаключение способно привести к массе впустую потраченного времени и средств.

Причинность без корреляции

Мы часто спорим, почему корреляция может не иметь причинного характера, но важно признать, что также могут существовать истинные причинные взаимосвязи без видимого соотношения. То есть корреляцию нельзя считать демонстрацией причинности, и выявление взаимосвязи также не необходимое условие причинности.

Известен пример, именуемый *парадоксом Симпсона* (мы поговорим о нем в главе 5). В общем, даже если в рамках неких подгрупп есть взаимосвязь (скажем, тестируемый препарат в сравнении с известным лекарством улучшает результаты у некой группы населения), мы можем не обнаружить зависимости или найти, но обратную, если подгруппы объединить. Если новый препарат больше используют пациенты в наиболее тяжелом состоянии, а те, кто чувствует себя лучше, чаще получают обычное лекарство, то, если не принимать во внимание серьезность заболевания, может показаться, что тестовое лекарство приводит к худшим результатам для населения.

В качестве еще одного примера причинности без корреляции рассмотрим влияние длительных пробежек на вес. Да, пробежки могут снижать вес за счет траты калорий, но бег также приводит к повышению аппетита, что, в свою очередь, ведет к увеличению веса (и, таким образом, отрицательно влияет на его потерю). В зависимости от силы каждого конкретного воздействия или исследуемых данных положительный эффект пробежек может полностью нивелироваться отрицательным, а значит, между бегом и потерей веса соотношения не будет. Структура этого примера представлена на рис. 3.8. Причина обладает положительными и отрицательными воздействиями, которые осуществляются различными путями; вот почему мы можем либо не наблюдать корреляции вообще, либо наблюдать нечто близкое к ней (вспомним: любые меры не абсолютны).

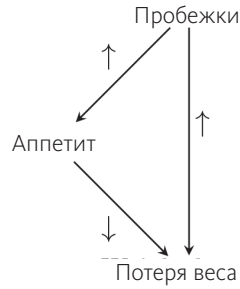
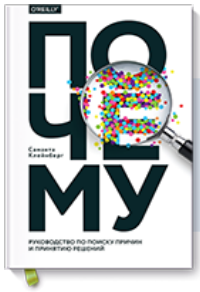


Рис. 3.8. Набор положительных (стрелка вверх) и отрицательных (стрелка вниз) причинных зависимостей. В разных группах населения они могут нивелироваться

Мы уже рассмотрели причины, по которым невозможно обнаружить существующую корреляцию (например, ошибка отбора, недостаточная вариация, предвзятость подтверждения, нелинейные зависимости и т. д.), и часто можно услышать, что соотношение не обязательно предполагает причинность. Но важно помнить об обратном: причинно-следственная связь не всегда подразумевает корреляцию[31].



издательство
МАНН, ИВАНОВ И ФЕРБЕР
Максимально полезные книги