Чарльз Уилан. Голая статистика

Статистика помогает принимать важные решения, находить скрытые взаимосвязи между явлениями, лучше понимать ситуацию в бизнесе и на рынке. Автор книги профессор Чарльз Уилан с юмором и блестящими наглядными примерами рассказывает о том, как это происходит. Эта книга будет полезной для студентов, которые не любят и не понимают статистику, но хотят в ней разобраться; маркетологов, менеджеров и аналитиков, которые хотят понимать статистические показатели и анализировать данные; а также для всех, кому интересно, как устроена статистика.

Чарльз Уилан. Голая статистика. Самая интересная книга о самой скучной науке. – М.: Манн, Иванов и Фербер, 2016. — 352 с.

Купить цифровую книгу в ЛитРес, бумажную книгу в Ozon или Лабиринте

Глава 1. В чем суть?

В чем суть? А в том, что статистика помогает нам обрабатывать данные, хотя на самом деле это всего лишь еще одно название информации. Статистика — самый мощный из имеющихся в нашем распоряжении инструментов для практического использования информации. Счет партии в боулинг является описательной (дескриптивной) статистикой. То же можно сказать и о каком-либо среднем показателе (например, в спорте). Чрезмерное увлечение любой из описательных статистик может привести к ошибочным умозаключениям и подтолкнуть к нежелательным действиям. Описательная статистика для того и существует, чтобы упрощать, что всегда подразумевает некоторую потерю нюансов и деталей (любопытно, что заметка Определение среднего значения, вариации и формы распределения. Описательные статистики является самой популярной в моем блоге; более 100 посещений в день).

Одна из ключевых функций статистики — использование имеющихся данных для выдвижения аргументированных предположений, касающихся вопросов, исчерпывающий ответ на которые невозможно дать из-за отсутствия полной информации. Короче говоря, мы можем использовать данные из «известного мира» для построения обоснованных гипотез относительно «неизвестного мира».

Например, точно подсчитать количество бездомных в крупном мегаполисе и дорого, и затруднительно. Одним из важных статистических методов является выборочное исследование — процесс сбора данных по какой-то небольшой области. По расчетам Американского института общественного мнения (Институт Гэллапа), методологически правильный опрос 1000 семей дает практически такие же результаты, как и опрос всех семей в Соединенных Штатах.

Даже в идеальных условиях статистический анализ лишь в редких случаях позволяет выявить «истину». Мы обычно выстраиваем некую версию, основанную на косвенных доказательствах, базирующихся на несовершенных данных. В результате появляются многочисленные причины, по которым интеллектуально честные люди не соглашаются со статистическими результатами или выводами. На самом фундаментальном уровне мы можем не соглашаться с самой постановкой рассматриваемого вопроса. Статистический анализ гораздо больше напоминает работу следователя. А умные и честные люди всегда будут спорить относительно того, о чем именно говорят нам те или иные данные.

Эта книга задумывалась как дань уважения классическому труду Дарелла Хаффа Как лгать при помощи статистики, который был впервые опубликован в 1954 году и разошелся тиражом свыше миллиона экземпляров (на русском языке книга впервые вышла в 2015 г. по моей рекомендации).

Математическая точность, сопутствующая статистическому анализу, может служить ширмой для откровенного бреда, которому пытаются придать некое наукообразие (см., например, Как с помощью диаграммы приукрасить действительность? или о факторе лжи Эдварда Тафти).

Глава 2. Описательная статистика

Самая фундаментальная задача при работе с данными — обобщить их огромные массивы. Чем большим количеством данных мы располагаем, тем труднее выделить в них главное. Поэтому мы вынуждены прибегать к упрощениям. Мы выполняем вычисления, которые сводят сложный массив данных к нескольким числам, описывающим эти данные.

Плюс состоит в том, что описательные статистики дают нам некое обобщенное и осмысленное представление исходного явления. Минус же в том, что любое упрощение порождает манипулирование. Два основных средних показателя – среднее арифметическое и медиана. Их вычисление не представляет особых трудностей; самое главное в этом случае — определить, какой именно показатель «середины» более точен в каждой конкретной ситуации (именно этот фактор нередко используется для манипулирования средними показателями).

Еще одной статистикой, которая позволяет описывать большие нагромождения данных, является среднеквадратическое (или, как его еще называют, стандартное) отклонение — показатель разброса данных по отношению к их среднему значению. Другими словами, среднеквадратическое отклонение представляет собой показатель рассредоточенности наблюдений.

Одним из наиболее важных, полезных и распространенных распределений в статистике является нормальное распределение, имеющее колоколообразную форму. Нормальное распределение описывает многие явления, часто встречающиеся в жизни. Красота нормального распределения — его мощь, изящество и элегантность — обусловлена тем, что нам по определению известно, какая именно доля наблюдений в нормальном распределении находится в пределах одного среднеквадратического отклонения от среднего значения (68,2%), двух среднеквадратических отклонений от среднего значения (95,4%), трех среднеквадратических отклонений от среднего значения (99,7%) и т.д. (рис. 1).

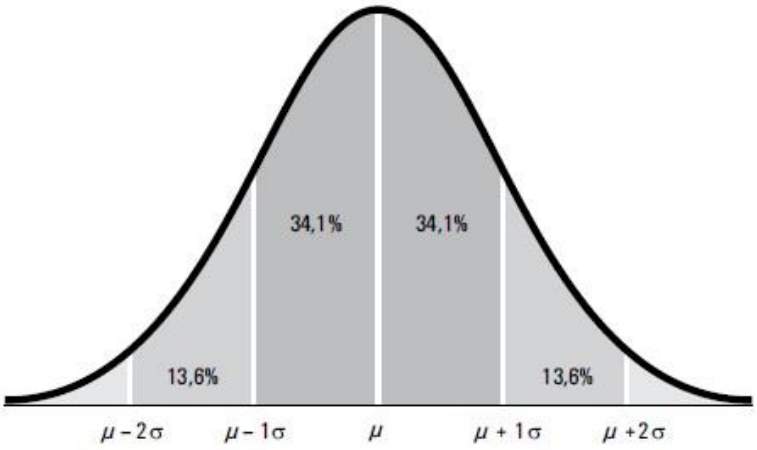

Рис. 1. Нормальное распределение

Глава 3. Дезориентирующее описание

Несмотря на то что статистика как область знаний коренится в математике, а математика, как известно, относится к числу точных наук, использование статистики для описания сложных явлений не может быть точным. Это оставляет немалый простор для манипуляций и искажения реального положения вещей. Марк Твен сказал однажды фразу, ставшую впоследствии знаменитой: «Есть три вида лжи: ложь, наглая ложь и статистика».

Словом «точность» мы обозначаем математическую точность того или иного явления. В описании протяженности вашего маршрута от дома до работы значение 41,6 мили будет более точным, чем «примерно 40 миль». Достоверность — это показатель того, соответствует ли истине рассматриваемое численное значение. Отсюда опасность путаницы между точностью и достоверностью. Если какой-либо ответ достоверный (правильный), то чем больше точность, тем, как правило, лучше. Однако даже самая высокая точность не в состоянии компенсировать недостоверности ответа. На самом деле точность может маскировать — случайно или вполне намеренно — недостоверность, вызывая у нас ложное ощущение определенности.

Рассмотрим пример. Многие из моделей управления рисками, использовавшиеся на Уоллстрит до финансового кризиса 2008 года, были довольно точными. Концепция «рисковой стоимости» (VaR) позволяла компаниям точно вычислить величину своего капитала, которая может быть потеряна в случае реализации тех или иных сценариев. Проблема состояла в том, что математические модели были сложными и запутанными. Ответы, которые можно было получить с их помощью, казались обнадеживающе точными. Однако предположения относительно того, что может случиться с глобальными рынками, встроенными в эти модели, были изначально неверными, в результате чего выводы, полученные с помощью этих моделей, были совершенно неправильными, что привело к дестабилизации не только Уоллстрит, но и всей мировой экономики.

Еще одна проблема может возникнуть, когда сравнивают несопоставимое. Например, нынешний доллар и доллар, каким он был шестьдесят лет назад, — это далеко не одно и то же: покупательная способность нынешнего доллара гораздо ниже. Вследствие инфляции товар, который стоил 1 доллар в 1950 году, стоил бы 9,37доллара в 2011-м. Это настолько важное явление, что экономисты придумали специальные термины, указывающие, была ли внесена поправка на инфляцию или нет. *Номинальные* величины не скорректированы с учетом поправки на инфляцию. Реальные величины, в отличие от номинальных, учитывают поправку на инфляцию. Чаще величины приводят к какой-то одной единице, например, долларам 2011 года, после чего становится возможным сравнение «яблок и апельсинов». На многих сайтах, включая сайт Бюро статистики труда (Министерства труда США), есть простые калькуляторы инфляции, которые позволяют сравнивать стоимость доллара в разные временны*е* периоды. Чтобы получить реальное представление о том, насколько может разниться статистика с поправкой и без поправки на инфляцию, рассмотрим приведенную ниже диаграмму изменения минимальной заработной платы на федеральном уровне США. На этой диаграмме представлены как номинальная величина минимальной заработной платы, так и ее реальная покупательная способность в долларах 2010 года.

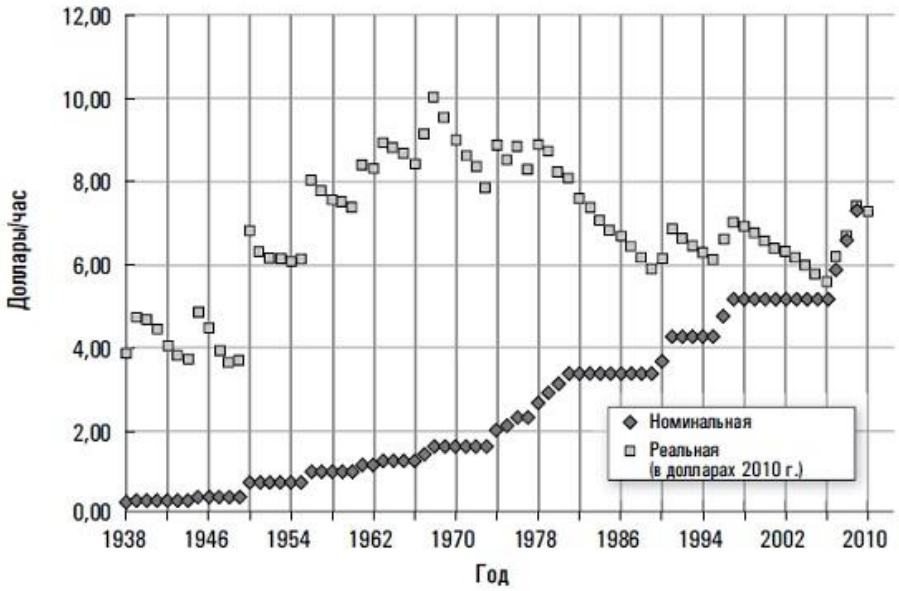

Рис. 2. Изменения минимальной заработной платы на федеральном уровне США

На мой взгляд, голливудские киностудии отличаются самым вопиющим игнорированием искажений, вносимых инфляцией, при сравнении доходов от разных фильмов в различные периоды времени (возможно, они делают это намеренно). Как, например, выглядит пятерка самых кассовых (на внутреннем рынке США) фильмов всех времен по состоянию на 2011 год?

1. «Аватар» (2009)
2. «Титаник» (1997)
3. «Темный рыцарь» (2008)
4. «Звездные войны. Эпизод IV» (1977)
5. «Шрек 2» (2004)

Голливуд хотел бы создать у нас впечатление, что каждый его очередной блокбастер грандиознее и прибыльнее предыдущего. Один из способов сделать это — подсчитывать кассовые поступление в номинальных ценах. Как выглядела бы пятерка самых успешных с коммерческой точки зрения американских фильмов за всю историю существования кино в США с поправкой на инфляцию?

1. «Унесенные ветром» (1939)
2. «Звездные войны. Эпизод IV» (1977)
3. «Звуки музыки» (1965)
4. «Инопланетянин» (1982)
5. «Десять заповедей» (1956)

В реальных величинах «Аватар» оказывается на 14-м месте, а «Шрек» опускается на 31-е.

Каждую осень несколько чикагских газет и журналов публикуют рейтинги лучших школ региона, основанные на результатах сдачи стандартизованного теста штата Иллинойс. Вот один из выводов, совершенно смехотворных с точки зрения статистики: поступление в несколько школ, постоянно занимающих высокие места в рейтинге, возможно лишь на конкурсной основе; для этого нужно предварительно подать соответствующие документы, причем в школу будет зачислена лишь малая часть из тех, кто их подал. Одним из важнейших критериев для поступления в такие школы являются результаты сдачи стандартизованных тестов. Итак, подведем итоги: 1) эти школы считаются «лучшими», потому что их ученики имеют высокие баллы на экзаменах; 2) чтобы попасть в такую школу, нужно иметь высокие баллы стандартизованных тестов.

Хорошей новостью будет то, что «управление посредством статистики» способно изменить к лучшему поведение соответствующего человека или учреждения. Если вы можете определить долю бракованных изделий, сходящих с производственного конвейера, и эти дефекты обусловлены ситуацией на заводе, то выплата работникам премии за сокращение количества бракованных изделий должна, по-видимому, надлежащим образом изменить их поведение. Каждый из нас реагирует на стимулы. Статистика измеряет важные для нас результаты; стимулы подталкивают нас к их улучшению… Или, в отдельных случаях, к приукрашиванию статистики.

Глава 4. Корреляция

Корреляция измеряет степень связи между двумя явлениями. Две переменные положительно коррелированы, если изменение одной переменной вызывает изменение другой в том же направлении (например, взаимосвязь между ростом и весом человека). Корреляция отрицательна, если положительное изменение одной переменной обусловливает отрицательное изменение другой (например, связь между регулярным выполнением физических упражнений и весом человека). Если построить диаграмму разброса данных, отражающих рост и вес произвольной выборки взрослых американцев, то получится примерно такая картина:

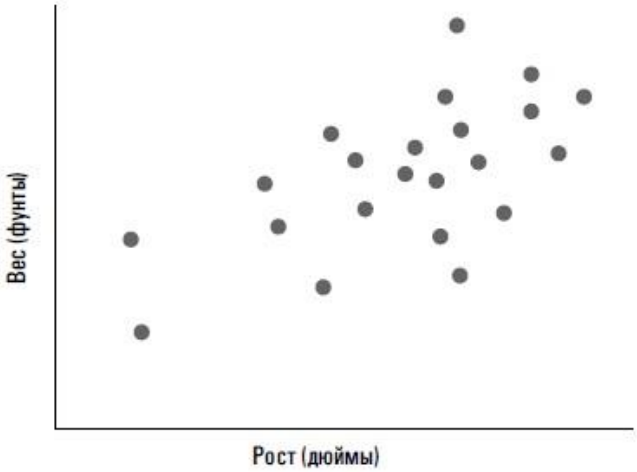

Рис. 3. Диаграмма разброса данных, отражающих рост и вес человека

Эффективность корреляции как статистического инструмента заключается в том, что мы можем выразить связь между двумя переменными с помощью одной описательной статистики — коэффициента корреляции. Он представляет собой число в диапазоне от –1 до 1. Чем ближе корреляция к 1 или –1, тем сильнее связь между переменными. Нулевая (или близкая к 0) корреляция говорит об отсутствии значимой связи между двумя переменными (например, между результатом экзамена по математике и размером обуви экзаменуемого).

Особенностью коэффициента корреляции является то, что с ним не связаны никакие единицы измерения. Мы можем рассчитать корреляцию между ростом и весом, несмотря на то что рост измеряется в дюймах, а вес — в фунтах. Коэффициент корреляции буквально творит чудеса: он сжимает сложное сочетание данных, измеряемых в разных единицах (наподобие наших диаграмм разброса роста и веса), в единственную элегантную описательную статистику.

Важным моментом в этом обсуждении является то, что корреляция не предполагает причинно-следственной связи: положительная или отрицательная корреляция между двумя переменными вовсе не обязательно означает, что изменения одной переменной вызывают изменения другой. Обе переменные могут быть обусловлены некой третьей переменной.

Коэффициента корреляции *r* для двух переменных *x* и *y* может вычислить с помощью формулы (подробнее см. Ковариация и коэффициент корреляции):

$$(1)\quad r = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{\sigma_x}\frac{(y_i-\bar{y})}{\sigma_y},$$

где

n — количество наблюдений;

$\bar{x}$ — среднее значение для переменной x;

$\bar{y}$ — среднее значение для переменной y;

σ_x — стандартное отклонение для переменной x;

σ_y — стандартное отклонение для переменной y.

В Excel для вычисления корреляции используется функция =КОРРЕЛ(массив1; массив2).

Глава 5. Основы теории вероятностей

Теория вероятностей — это наука о событиях и исходах, содержащих элемент неопределенности. Вероятности многих событий известны заранее. Вероятность выпадания орла при однократном подбрасывании «правильной» монетки равняется 1/2, а единицы при однократном подбрасывании игральной кости — 1/6. Выводы относительно вероятности наступления других событий можно сделать на основе прошлых данных. Так, например, начиная с 1960-х годов в гражданской авиации Австралии не зафиксировано ни одной катастрофы со смертельным исходом; таким образом, коэффициент смертности в расчете на каждые 100 миллионов километров «налета», по сути, равен нулю. Для автомобильного транспорта он составил 0,5, у мотоциклистов – 17,5 (в 35 раз больше!).

Вероятность *одновременного* наступления двух независимых событий представляет собой произведение их соответствующих вероятностей. Вероятность наступления события А *или* В равна сумме их индивидуальных вероятностей. Если вы играете в кости в Лас-Вегасе, то вероятность выпадания 7 или 11 в результате однократного подбрасывания равна количеству комбинаций, составляющих в сумме 7 или 11, поделенному на общее число вариантов, которые могут выпасть в результате подбрасывания двух игральных костей, или 8/36. Между прочим, значительная часть ранних исследований вероятности выполнялась именно любителями азартных игр в попытках точно определить свои шансы (см., например, Альфред Реньи. Письма о вероятности: письма Паскаля к Ферма).

Вероятность также позволяет подсчитать *математическое ожидание* — чрезвычайно полезный инструмент, используемый при принятии любых управленческих решений, особенно в сфере финансов. Математическое ожидание — это среднее значение случайной величины. Математическое ожидание представляет собой чрезвычайно мощный инструмент, поскольку он может сказать вам, является ли то или иное событие «справедливым», учитывая его цену и ожидаемый исход. Анализ математического ожидания выигрыша может показать, почему не стоит покупать лотерейные билеты. В отдельно взятом случае вам может повезти, но в среднем вы будете в проигрыше.

Важная теорема, известная как *закон больших чисел*, гласит, что по мере возрастания количества испытаний средний результат исходов все сильнее приближается к его математическому ожиданию. Вся страховая отрасль построена на вероятностях. Основная идея страхования заключается в том, что в обмен на регулярные и предсказуемые выплаты вы переносите на соответствующую страховую компанию риски. Почему страховые компании готовы взять на себя такие риски? Потому что в долгосрочном периоде они заработают большие прибыли — если, конечно, правильно рассчитают величину своих страховых взносов. Как потребитель, вы должны отдавать себе отчет, что в длительном периоде страховка не сэкономит вам деньги. Тем не менее это все же вполне разумный способ защиты от исходов, которые в противном случае могли бы вас просто разорить.

Глава 5½. Загадка Монти Холла

Это знаменитая задача по теории вероятностей, поставившая в тупик участников игрового шоу под названием Let’s Make a Deal («Совершим сделку»). Участник, добравшийся до финала, становился вместе с Монти Холлом перед тремя большими дверями: Дверью № 1, Дверью № 2 и Дверью № 3. Монти Холл объяснял финалисту, что за одной из этих дверей скрывается очень ценный приз — например новый автомобиль, а за двумя другими — козел. Финалист должен был выбрать одну из дверей и получить то, что за ней находилось.

После того как финалист шоу укажет на какую-то из трех дверей, Монти Холл открывает одну из двух оставшихся дверей, за которой всегда находится козел. Затем Монти Холл спрашивает финалиста, не желает ли он изменить свое решение, то есть отказаться от ранее выбранной им закрытой двери в пользу другой закрытой двери.

Следует ли финалисту отказаться от первоначального выбора в пользу Двери № 2? Отвечаю: да, следует. Если он будет придерживаться первоначального выбора, то вероятность выигрыша им ценного приза составит 1/3; если же передумает и укажет на Дверь № 2, то вероятность выигрыша ценного приза будет 2/3.

Первое мое объяснение — эмпирическое. Согласно Леонарду Млодинову (см. Леонард Млодинов. (Не)совершенная случайность. Как случай управляет нашей жизнью), те из финалистов, кто изменил свой первоначальный выбор, становились победителями примерно в два раза чаще, чем те, кто оставался при своем мнении.

Мое второе объяснение данного феномена основывается на интуиции. Допустим, правила игры слегка поменялись. Как и раньше, финалист начинает с выбора одной из трех дверей. Однако затем, прежде чем открыть какую-то из дверей, за которой скрывается козел, Монти Холл спрашивает: «Согласны ли вы отказаться от своего выбора в обмен на открывание двух оставшихся дверей?» Таким образом, если вы выбрали Дверь № 1, вы можете передумать в пользу Двери № 2 и Двери № 3. Если сперва указали на Дверь № 3, можете выбрать Дверь № 1 и Дверь № 2. И так далее.

Для вас это было бы не особо трудным решением: совершенно очевидно, что вам следует отказаться от первоначального выбора в пользу двух оставшихся дверей, поскольку это повышает шансы на выигрыш с 1/3 до 2/3. Самое интересное, что именно такой в сущности вариант предлагает вам Монти Холл в реальной игре, после того как откроет дверь, за которой скрывается козел. Принципиальный факт заключается в том, что, если бы вам была предоставлена возможность выбрать две двери, за одной из них в любом случае скрывался бы козел. Когда Монти Холл открывает дверь, за которой находится козел, и только после этого спрашивает вас, согласны ли вы изменить свой первоначальный выбор, он существенно повышает ваши шансы на выигрыш ценного приза! По сути, Монти Холл говорит вам: «Вероятность того, что ценный приз скрывается за одной из двух дверей, которые вы не выбрали с первого раза, составляет 2/3, а это все-таки больше чем 1/3.

Глава 6. Проблемы с вероятностью

Одним из самых безответственных случаев применения статистики за последнее время стал механизм оценивания рисков на Уолл-стрит перед финансовым кризисом 2008 года. Модель стоимости риска – Value-at-Risk (VaR) – сочетала в себе элегантность индикатора (совмещая обширную информацию в едином числовом показателе) с мощью вероятности (присоединяя ожидаемую прибыль или убыток к каждому из активов или торговым позициям соответствующей фирмы). На основе прошлых данных о движениях рынка «количественные» эксперты компании (в сфере финансов их часто называют «квантами», от слова quantitative, то есть «количественный») могли определить максимальную сумму в денежном выражении (например, 13 миллионов долларов), которую фирма может с 99процентной вероятностью потерять на данной позиции в течение рассматриваемого периода времени. Другими словами, в 99 случаях из 100 компания не потеряет более 13 миллионов долларов на конкретной торговой позиции; а в 1 случае из 100 потеряет.

К сожалению, с профилями риска, заложенными в моделях VaR, существовали две огромные проблемы. Во-первых, вероятности, на которых строились эти модели, исходили из прошлых движений рынка; однако на финансовых рынках будущее вовсе не обязательно должно быть похожим на прошлое. Кроме того, даже если бы исходные данные могли точно прогнозировать будущий риск, 99-процентная гарантия, обещанная моделью VaR, была опасно бесполезной, поскольку остающийся 1% *действительно вводит в заблуждение*. Колумнист The New York Times Джо Носер подытоживает мысли Талеба (см. Нассим Талеб. Черный лебедь. Под знаком непредсказуемости) и яростного критика VaR: «Самые опасные — отнюдь не риски, которые вы можете увидеть и измерить, а риски, которые вы не можете увидеть и, следовательно, измерить. Это риски, находящиеся настолько далеко за пределами нормальной вероятности, что невозможно даже себе представить, что они могут произойти в вашей жизни, — хотя, конечно же, они случаются, и даже чаще, чем вы могли бы предположить».

Вот несколько иных распространенных ошибок, заблуждений и этических дилемм, связанных с применением концепции вероятности.

Предполагается, что события независимы, тогда как на самом деле они зависимы друг от друга. Например, вероятность выхода из строя по тем или иным причинам авиадвигателя во время трансатлантического перелета составляет 1 шанс из 100 000. Учитывая количество трансатлантических перелетов, этот риск нельзя считать приемлемым. К счастью, каждый современный самолет, совершающий такие перелеты, оснащен по меньшей мере двумя

двигателями. Риск одновременного выхода из строя обоих во время трансатлантического перелета должен равняться $(1/100\,000)^2$, или 1 шансу из 10 миллиардов, что считается вполне приемлемым риском с точки зрения обеспечения безопасности полетов. Однако, поломка обоих авиадвигателей не относится к категории независимых событий. Если во время взлета самолет наталкивается на стаю гусей, то, вероятнее всего, оба двигателя выйдут из строя одинаковым образом. То же самое можно сказать о многих других факторах, влияющих на функционирование авиадвигателя, начиная с погодных условий и заканчивая небрежным выполнением своих обязанностей наземными службами техобслуживания.

Еще одна разновидность ошибок возникает, когда *события, действительно независимые друг от друга, рассматриваются как взаимосвязанные*. Если вы когда-либо окажетесь в казино, то обязательно увидите людей, вперившихся взглядом в игральные кости или карты и заявляющих, что они «ожидают должное». Если шарик рулетки пять раз подряд остановился на черном поле, то всякому здравомыслящему человеку понятно, что на следующий раз должно выпасть красное. Нет, нет и еще раз нет!

> Вместе с тем в теории вероятностей доказан факт, что если достаточно долго подбрасывать монету, то будут наблюдаться периоды преобладания выпадания орла или решки. Это так называемый первый закон арксинуса. Этот закон не отменяет сказанного автором, а только показывает структуру исходов в испытаниях Бернулли. О данном феномене см., например, классическую книгу В.Феллер. Введение в теорию вероятностей и ее приложения. Т.1. Глава III. – *Прим. ред.*

Кластеры действительно встречаются. Т.е., даже случайные события могут выстраиваться в некоторые закономерности. Я провожу следующий эксперимент со своими студентами, чтобы подтвердить этот базовый постулат. Чем больше аудитория, тем лучше. я предлагаю каждому из присутствующих вынуть монетку и встать. Затем все подбрасывают монетку, и те, у кого выпадает решка, садятся. Допустим, в аудитории находится 100 студентов; примерно 50 из них займут свое место после первого подбрасывания. Потом мы выполняем это упражнение еще раз, в результате чего останутся стоять примерно 25 студентов. И так далее. Чаще всего после пяти или шести подбрасываний остается всего один человек, у которого пять или шесть раз подряд выпал орел. я спрашиваю этого уникума: «Как вам это удалось?». Все присутствующие, конечно, воспринимают это как шутку. Однако каждый раз, когда мы видим какое-либо аномальное событие вне конкретного контекста, в котором оно произошло, у нас поневоле возникает подозрение, что здесь, помимо чистой случайности, замешано что-то еще.

Регресс к среднему. Возможно, вы слышали о так называемом проклятии Sports Illustrated, в результате которого спортсмены или команды, фотографии которых помещались на обложке журнала Sports Illustrated, впоследствии снижали свои спортивные достижения. Одно из объяснений этого феномена заключалось в том, что размещение фотографии спортсмена на обложке издания неблагоприятно сказывается на его последующих спортивных показателях. Более правдоподобным, с точки зрения статистики, будет объяснение, что команды и спортсмены обычно появляются на обложке Sports Illustrated после того, как добьются выдающихся успехов (например, станут олимпийскими чемпионами), поэтому вполне естественно, что, пройдя пик физической формы, они демонстрируют результаты, близкие к средним (подробнее см. Канеман, Словик, Тверски. Принятие решений в неопределенности: Правила и предубеждения).

Статистическая дискриминация (установление различия в статистическом смысле). Мужчины обычно платят больше за автостраховку, поскольку чаще, чем женщины, попадают в аварии. Для страховых компаний – это всего лишь статистика. Однако, Еврокомиссия в 2012 году запретила ставить страховые надбавки в зависимость от пола человека.

Глава 7. Почему так важны данные

Данные для статистики очень важны. Даже самый изощренный анализ не принесет никакой пользы, если за основу взяты сомнительные данные. Отсюда выражение: «Мусор на входе — мусор на выходе». Как правило, данные выполняют одну из трех функций. Во-первых, нам может потребоваться определенная выборка данных, соответствующая характеристикам генеральной совокупности (так называемая репрезентативная выборка). Получить хорошую выборку гораздо

сложнее, чем может показаться на первый взгляд. Многие из самых ошибочных статистических утверждений обусловлены применением совершенно правильных статистических методов к плохим выборкам, а вовсе не наоборот. Размер выборки имеет значение — чем она больше, тем лучше.

Второе, что нам зачастую требуется от данных, — это чтобы они служили нам источником сравнения. Новое лекарство эффективнее нынешнего? В подобных случаях наша задача — найти две группы субъектов, в целом похожих между собой — за исключением интересующего нас «параметра».

Третья причина сбора данных. Иногда у нас нет четкого представления о том, для чего нам может понадобиться та или иная информация, но интуитивно мы предполагаем, что в какой-то момент она обязательно пригодится. Город Фрамингем, пригород Бостона, для ученых ассоциируется с исследованием под названием Framingham Heart Study — одним из самых успешных в истории современной науки, оказавшим огромное влияние на развитие медицины. В 1948 году ученые собрали информацию о 5209 взрослых жителях города: их рост, вес, кровяное давление, уровень образования, состав семьи, типичные продукты питания, склонность к курению, употребление наркотиков и т.п. Далее эти люди периодически повторно обследовались, а также собирались данные об их потомстве, чтобы выявить генетические факторы, связанные с развитием сердечно-сосудистых заболеваний. Начиная с 1950 года фрамингемские данные использовались при написании более чем двух тысяч научных статей.

Эти исследования позволили получить чрезвычайно важные для понимания механизмов развития сердечно-сосудистых заболеваний результаты, многие из которых кажутся нам сейчас очевидными: курение сигарет увеличивает риск сердечно-сосудистых заболеваний (1960 год); физическая активность снижает риск сердечно-сосудистых заболеваний, а ожирение, наоборот, повышает (1967 год); высокое кровяное давление увеличивает риск инсульта (1970 год); высокий уровень холестерина альфа-липопротеинов высокой плотности (известного с тех пор как «полезный холестерин») снижает риск смертельного исхода (1988 год); у лиц, родители и близкие родственники (родные братья и сестры) которых страдали сердечно-сосудистыми заболеваниями, риск их развития значительно выше (2004–2005 годы).

Вот несколько типичных примеров из категории «мусор на входе — мусор на выходе».

Систематическая ошибка выбора. Вопрос, который всегда нас должен интересовать: как была сформирована выборка для оценивания? Если каждому члену генеральной совокупности не предоставлены равные шансы на включение в выборку, у нас наверняка возникнут проблемы с результатами, полученными на ее основе.

Систематическая ошибка публикации. Позитивные результаты обнародуют охотнее, чем негативные. Согласно важному положению в статистике, необычные явления происходят довольно редко и, как правило, в результате случайного стечения обстоятельств. Например, в одном из 100 исследований обнаружатся нелепые результаты типа взаимозависимости между увлечением видеоиграми и меньшей заболеваемостью раком толстой кишки. Проблема в том, что результаты 99 исследований, которые не выявили такую связь, опубликованы не будут, поскольку малоинтересны. А вот единственное исследование, которое ее обнаружит, попадет в печать и привлечет к себе повышенное внимание. Источником данной систематической ошибки является не исследование как таковое, а сомнительная информация, которая фактически становится достоянием широкого круга читателей.

Чтобы справиться с данной проблемой, теперь медицинские журналы, как правило, требуют зарегистрировать любое исследование в самом начале проекта, если предполагается последующая публикация его результатов. Это предоставляет редакторам определенные свидетельства о соотношении позитивных и негативных исходов. Если, например, зарегистрировано 100 исследований по анализу влияния катания на роликовой доске (скейтборде) на развитие сердечно-сосудистых заболеваний, и лишь одно из них будет в конечном счете представлено для публикации с положительными результатами, то редакторы могут заключить, что в ходе других исследований получены отрицательные результаты (или по крайней мере проверить такую вероятность).

Систематическая ошибка памяти. Человек интуитивно пытается находить причинно-следственные связи. Проблема в том, что наша память оказывается «систематически хрупкой», когда мы пытаемся объяснить какой-либо особенно хороший или плохой результат в настоящем. Наличие такой

систематической ошибки памяти — одна из причин, почему ученые чаще предпочитают проводить повторные исследования, а не исследования типа «поперечный срез». В случае повторного исследования сбор данных выполняется на протяжении всего времени его проведения. В пятилетнем возрасте участника спрашивают о его отношении к школе. Затем, спустя тринадцать лет, мы можем наведаться к нему и выяснить, не бросил ли он школу досрочно. При проведении исследования «поперечный срез» все данные собираются одномоментно, и, спрашивая восемнадцатилетнего парня, бросившего школу, как он к ней относился в пятилетнем возрасте, мы вряд ли получим правдивый ответ.

Систематическая ошибка доживаемости до определенного возраста возникает, когда какие-то из наблюдений выпадают из выборки, изменяя состав оставшихся наблюдений и тем самым сказываясь на результатах того или иного анализа. Индустрия взаимных фондов охотно ухватилась за систематическую ошибку доживаемости до определенного возраста, воспользовавшись ею для того, чтобы их прибыльность выглядела для инвесторов гораздо привлекательнее, чем на самом деле. Менеджеры взаимных фондов убеждают нас в своей дальновидности и умении использовать знания для выбора таких ценных бумаг, которые обеспечивают более высокую прибыльность, чем какой-нибудь простой индексный фонд. В действительности превзойти S&P 500 на достаточно продолжительном отрезке времени довольно трудно.

Чем занимается традиционная компания типа взаимного фонда? Открывает много новых активно управляемых фондов. Допустим, она открывает двадцать новых фондов, каждый из которых с 50процентной вероятностью может в данном году превзойти S&P 500. Согласно теории вероятностей, в первый год лишь десять новых фондов компании превзойдут S&P 500; пять фондов превзойдут S&P 500 в течение двух лет подряд; а два или три фонда — в течение трех лет подряд. В этот момент новые взаимные фонды, которые продемонстрировали не особо впечатляющие результаты по сравнению с S&P 500, по-тихому прикрываются (их активы включаются в другие существующие фонды). Затем компания может запустить массированную рекламу двух или трех новых фондов, которые «год за годом превосходят S&P 500», — даже если результат, достигнутый ими, такая же случайность, как выпадание решки три раза подряд. Дальнейшие показатели эффективности этих фондов наверняка приблизятся к среднему значению — правда, по пути они привлекут к себе толпы новых инвесторов. На самом деле количество взаимных фондов или инвестиционных гуру, которые на протяжении достаточно продолжительного времени превосходят S&P 500, удручающе мало. (С очень интересным обсуждением того, почему следует отдать предпочтение покупке индексных фондов, вместо того чтобы пытаться превзойти рынок, можно ознакомиться в книге профессора Бертона Малкиела Случайная прогулка по Уолл-стрит. — Минск: Попурри, 2006.)

Глава 8. Центральная предельная теорема

Базовый принцип, лежащий в основе центральной предельной теоремы, заключается в том, что большая, надлежащим образом сформированная выборка будет похожа на совокупность, из которой она извлечена. Центральная предельная теорема позволяет нам сделать следующие выводы:

1. Располагая подробными сведениями о какой-то совокупности, мы можем сделать далекоидущие выводы о любой надлежащим образом сформированной из нее выборке.
2. И, наоборот, располагая подробными сведениями о какой-либо надлежащим образом сформированной выборке (среднее значение и среднеквадратическое отклонение), мы можем сделать чрезвычайно точные выводы относительно совокупности, из которой эта выборка была получена.
3. Наличие данных о какой-то конкретной выборке и данных о какой-то конкретной совокупности позволяет определить, с какой вероятностью выборка получена из совокупности.
4. Наконец, если нам известны исходные характеристики двух выборок, то мы можем определить, сформированы ли они из одной и той же совокупности.

Согласно центральной предельной теореме, средние значения выборок для любой совокупности будут распределены относительно ее среднего значения примерно по нормальному закону. При этом, вовсе не обязательно, чтобы совокупность, из которой получены эти выборки, имела

нормальное распределение. Например, распределение семейного дохода в Соединенных Штатах характеризуется значительным скосом (рис.4), однако у распределения средних значений выборок скос отсутствует. Чем больше количество выборок, тем точнее это распределение аппроксимируется нормальным распределением. А чем больше размер каждой выборки, тем такое распределение будет *уже*. Чтобы обеспечить применимость центральной предельной теоремы, желательно, чтобы размер выборки был не менее 30.

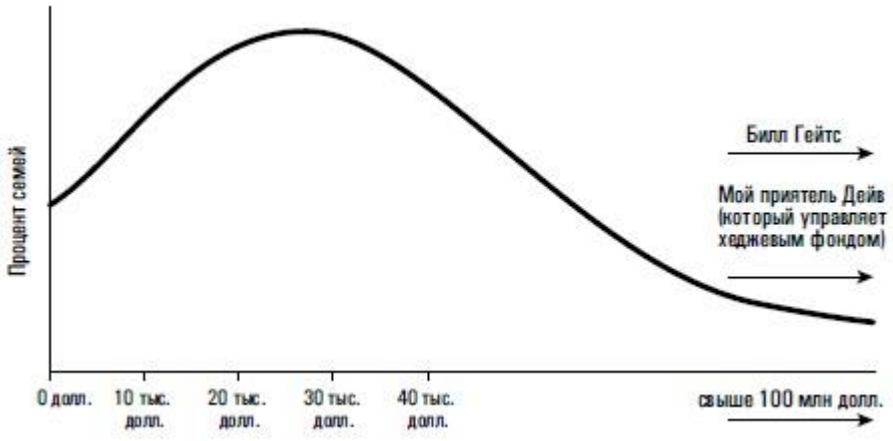

Рис. 4. Годовой семейный доход

Если среднеквадратическое отклонение измеряет разброс в исходной совокупности, то стандартная ошибка измеряет разброс средних значений выборок. Стандартная ошибка является среднеквадратическим отклонением средних значений выборок. Формула для стандартной ошибки:

$$(2)\ \ SE = \frac{s}{\sqrt{n}}$$

где s — среднеквадратическое отклонение для совокупности, из которой сформирована данная выборка, а n — размер выборки.

Чем меньше вероятность того, что какой-то исход оказался чисто случайным, тем больше мы можем быть уверены в том, что здесь не обошлось без воздействия какого-то другого фактора. В этом по большому счету и заключается сущность статистического вывода.

Глава 9. Статистические выводы

Статистика не может ничего утверждать с определенностью. Напротив, сила статистического вывода проистекает из наблюдения некой картины или исхода и последующего использования теории вероятностей для получения его (ее) самого вероятного объяснения.

Одним из самых распространенных инструментов в статистическом выводе является проверка гипотез. Любой статистический вывод начинается с подразумеваемой или явно сформулированной основной (так называемой нулевой) гипотезы. Это наша начальная гипотеза, которая будет отвергнута или принята исходя из последующего статистического анализа. Может показаться нелогичным, но исследователи часто формулируют нулевую гипотезу в надежде, что им удастся отвергнуть ее. Исследователи обычно спрашивают: если нулевая гипотеза истинна, то какова вероятность того, что мы наблюдаем такую картину данных по чистой случайности?

Одно из самых распространенных пороговых значений, используемых исследователями для отклонения нулевой гипотезы, — 5%. Данная вероятность известна как уровень значимости. Если вероятность возникновения наблюдаемой картины меньше этого порогового значения, мы отвергаем нулевую гипотезу.

Также используют *p-значение* — вероятность получения наблюдаемого результата, если бы нулевая гипотеза была верна. Если уровень значимости 0,05 кажется вам в какой-то мере произвольным, то вы абсолютно правы: так оно и есть! Не существует единого стандартизированного статистического порога для отказа от нулевой гипотезы. Когда мы можем отвергнуть основную гипотезу с некоторым разумным уровнем значимости, соответствующие результаты считаются «статистически значимыми».

Уровень значимости исследователи должны устанавливать **до** выполнения статистического анализа, чтобы избежать его выбора постфактум, что бывает очень удобно, когда полученным результатам требуется придать значимость.

Доверительным интервалом называется диапазон от $\mu - A$ до $\mu + A$, где μ – среднее значение в выборке, A – отклонение, обусловленное выбранным пороговым значением. Для 5%-ного порога $A \approx 2\sigma$. Если порог, позволяющий отвергнуть основную гипотезу, будет чересчур большим (например, 0,1), то нам придется периодически отклонять нулевую гипотезу, хотя на самом деле она верна. На языке статистики это называется *ошибкой первого рода*. Ошибка первого рода заключается в ошибочном отказе от основной гипотезы. Иногда это называют «ложным позитивом». Например, когда вы приходите к врачу, чтобы выяснить, не страдаете ли вы некой болезнью, основная гипотеза заключается в том, что вы ею не страдаете. Если результаты анализов позволяют отвергнуть нулевую гипотезу, то врач говорит, что у вас положительный результат анализов. А если у вас положительный результат анализов, хотя в действительности вы не больны, то это и есть случай «ложного позитива».

Однако, чем ниже порог для отказа от нулевой гипотезы (например, 0,001), тем вероятнее, что нам не удастся отвергнуть ту нулевую гипотезу, которую на самом деле следовало было бы отвергнуть. На языке статистики это называется ошибкой второго рода, или «ложным негативом». Еще раз, ошибка второго рода – это вероятность принятия нулевой гипотезы тогда, когда она неверна.

Рассмотрим две ситуации, предполагающие достижение определенного компромисса между ошибками первого и второго рода. *Спам-фильтры.* Основная гипотеза: любое конкретное сообщение, приходящее по электронной почте, не спам. Ваш спам-фильтр отыскивает признаки, которые могут использоваться для отказа от нулевой гипотезы для того или иного конкретного сообщения. Ошибка первого рода предполагает отбраковку сообщения, которое на самом деле не является спамом. Ошибка второго рода предполагает пропуск спама через фильтр и его попадание в ваш почтовый ящик. Сравнивая последствия от потери важного сообщения и незначительное раздражение, вызванное получением спама, большинство людей, предпочтут терпеть неудобства, обусловленные ошибками второго рода. Оптимально разработанный спам-фильтр должен требовать относительно высокой степени определенности, прежде чем отвергнуть нулевую гипотезу и заблокировать соответствующее сообщение.

Проверка на наличие раковых заболеваний. Существуют многочисленные тесты для раннего выявления раковых заболеваний, например, маммография. Основная гипотеза для каждого, кто проходит такое обследование, заключается в том, что он не болен раком. Ошибка первого рода («ложный позитив», что в конечном счете означает отсутствие заболевания) безусловно предпочтительнее ошибки второго рода («ложный негатив», который означает, что диагностирование не выявило заболевания, которое на самом деле имеется). Врачи и пациенты готовы мириться с умеренным количеством ошибок первого рода, чтобы избежать вероятности появления ошибок второго рода, когда пациенту не диагностируется раковое заболевание, хотя в действительности он болен. Впрочем, в последнее время специалисты в области политики охраны здоровья подвергают сомнению такой подход из-за высоких издержек и побочных эффектов, связанных с «ложными позитивами».

Формула для сравнения двух средних значений

$$(3) \quad \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

где $\bar{x}$ — среднее значение выборки x

$\bar{y}$ — среднее значение выборки y

s_x — среднеквадратическое отклонение выборки x

s_y — среднеквадратическое отклонение выборки y

n_x — количество наблюдений в выборке x

n_y — количество наблюдений в выборке y

В числителе вычисляется разность двух средних значений; в знаменателе — стандартная ошибка для разности двух средних значений разных выборок. Нулевая гипотеза: средние значения этих двух

выборок одинаковы. Приведенная выше формула вычисляет наблюдаемую разность средних значений относительно величины стандартной ошибки для разности средних значений. Как и прежде, мы предполагаем, что имеем дело с нормальным распределением. Если средние значения исходной совокупности действительно одинаковы, то можно ожидать, что разность средних значений двух выборок окажется меньше одной стандартной ошибки в 68 случаях из 100 и меньше двух стандартных ошибок в 95 случаях из 100.

Если, например, значение (3) получилось равным 3,15, то существует лишь 0,002 шанса, что эти выборки сделаны из одной совокупности данных (рис. 5).

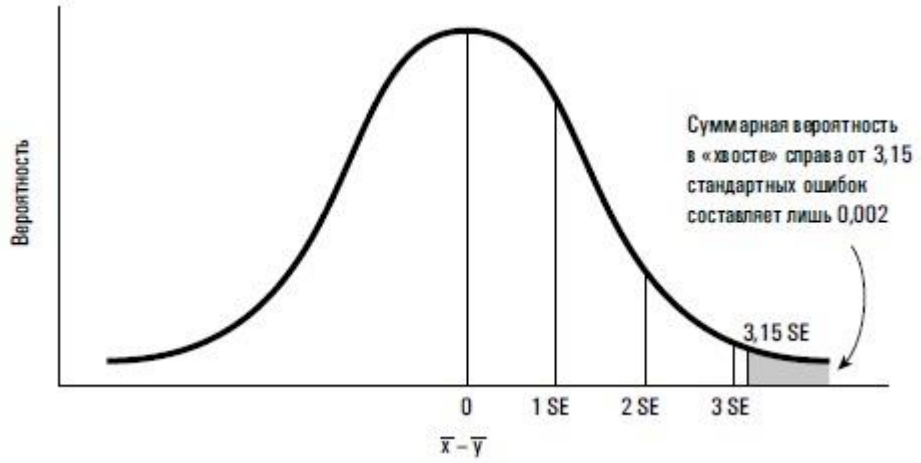

Рис. 5. Разность средних значений выборок

Проверка гипотез с одно- и двусторонним критерием. Допустим, что нулевая гипотеза утверждает, что два средних одинаковы, и мы исходим из того, что будем выполнять проверку значимости на уровне 0,05. Когда альтернативная гипотеза гласит, что одно среднее больше другого, мы будем выполнять *проверку гипотез с односторонним критерием*. Мы знаем, что в случае, если наша нулевая гипотеза верна, мы будем наблюдать разницу не меньше 1,64 стандартной ошибки лишь в 5 случаях из 100. Мы отвергнем нулевую гипотезу, если полученный результат попадает в диапазон, указанный на приведенном ниже графике (рис. 6).

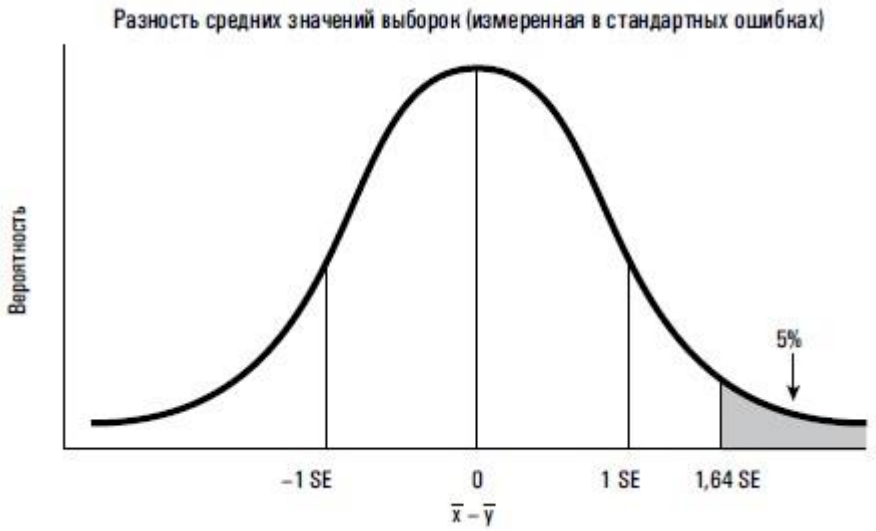

Рис. 6. Проверка гипотезы односторонним критерием

Если же альтернативная гипотеза гласит, что одно среднее больше **или меньше** другого, нам понадобится выполнять проверку гипотез с двусторонним критерием. Граница, по достижении которой мы отклоняем нулевую гипотезу, будет другой, поскольку на сей раз мы должны учитывать вероятность большой разницы в средних значениях выборок в обоих направлениях: положительном и отрицательном (рис. 7).

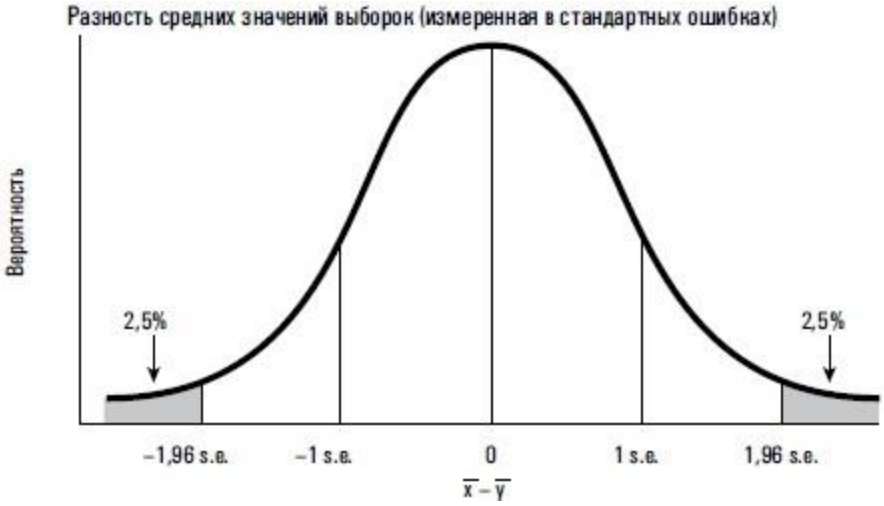

Рис. 7. Проверка гипотезы двусторонним критерием

Глава 10. Опрос общественного мнения

Опрос общественного мнения — это вывод о настроениях определенной совокупности людей, основанный на мнениях, высказанных выборкой, сформированной из генеральной совокупности. Эффективность опросов обусловливается использованием центральной предельной теоремы. Формула расчета стандартной ошибки в случае, когда речь идет о процентной величине или доле:

$$(4)\ SE = \sqrt{\frac{p(1-p)}{n}}$$

где *p* — доля респондентов, выражающих определенную точку зрения, *(1 – p)* — доля респондентов, имеющих противоположную точку зрения, а *n* — общее количество респондентов в выборке.

Проблемы с результатами опросов обычно обусловлены не ошибкой в математических расчетах при вычислении стандартных ошибок, а являются следствием некорректно сформированной выборки, или неправильно сформулированных вопросов, или того и другого. Выражение «мусор на входе — мусор на выходе» полностью применимо к проведению социологических опросов. Ниже перечислены ключевые методологические вопросы, которые необходимо задать при проведении любого опроса общественного мнения или оценивании чьей-то работы.

- *Действительно ли данная выборка является репрезентативной (представительной) из совокупности, настроения которой мы пытаемся выяснить?*
- *Позволяет ли формулировка вопросов получить точную информацию по интересующим нас темам?* Согласно опросу, проведенному Институтом Гэллапа, начиная с 2002 года свыше 60% американцев ежегодно заявляют, что поддерживают применение смертной казни в отношении лиц, осужденных за убийство. Однако, поддержка американцами смертной казни падает, когда в качестве альтернативы предлагается пожизненное тюремное заключение без права условно-досрочного освобождения. Опрос, проведенный Институтом Гэллапа в 2006 году, показал, что лишь 47% американцев считают смертную казнь справедливой карой за убийство, тогда как 48% высказываются за пожизненное тюремное заключение.
- *Говорят ли респонденты правду?*

Почему стандартная ошибка оказывается больше, когда p и (1 – p) близки к 50%? По-видимому, самое простое доказательство, что функция $f(p) = p(1 – p) = p – p^2$ принимает максимальное значение при p = 0,5, — это математическое доказательство. Находим производную $f'(p) = 1 – 2p$, приравниваем ее к нулю, и получаем уравнение $1 – 2p = 0$. Решением этого уравнения будет p = 0,5. Что и требовалось доказать. О том, что это максимум, свидетельствует вторая производная $f''(p) = – 2$.

Глава 11. Регрессионный анализ

Может ли стресс на работе стать причиной вашей смерти? Да, вполне. Существуют убедительные доказательства того, что суровые условия на работе могут привести к преждевременной смерти,

особенно в результате развития сердечно-сосудистых заболеваний. Однако это не тот вид стресса, о котором вы, наверное, подумали. Главы компаний, которым буквально каждый день приходится принимать чрезвычайно сложные и ответственные решения, определяющие дальнейшую судьбу их бизнеса, рискуют значительно меньше, чем их секретарши, бесконечно отвечающие на телефонные звонки, параллельно выполняя множество других задач, предусмотренных должностной инструкцией. Как такое может быть? Оказывается, самый опасный вид стресса на работе обусловлен невозможностью человека в достаточной степени контролировать способы и условия выполнения поставленных задач.

Статистический инструмент под названием *регрессионный анализ* позволяет нам измерить величину зависимости между какой-то переменной и интересующим нас исходом, зафиксировав действие всех прочих факторов. Другими словами, мы можем вычленить влияние одной переменной (например, занятие определенным родом деятельности), сохраняя на постоянном уровне действие других переменных. Регрессионный анализ использовался и при проведении упоминавшегося выше исследования.

По своей сути регрессионный анализ стремится найти «наилучшее приближение» линейной зависимости между двумя переменными. Простой пример — зависимость между ростом и весом людей (см. рис. 3). Регрессионный анализ позволяет «провести линию», которая точнее всего отражает линейную зависимость между этими двумя переменными. Для этого используется метод наименьших квадратов.

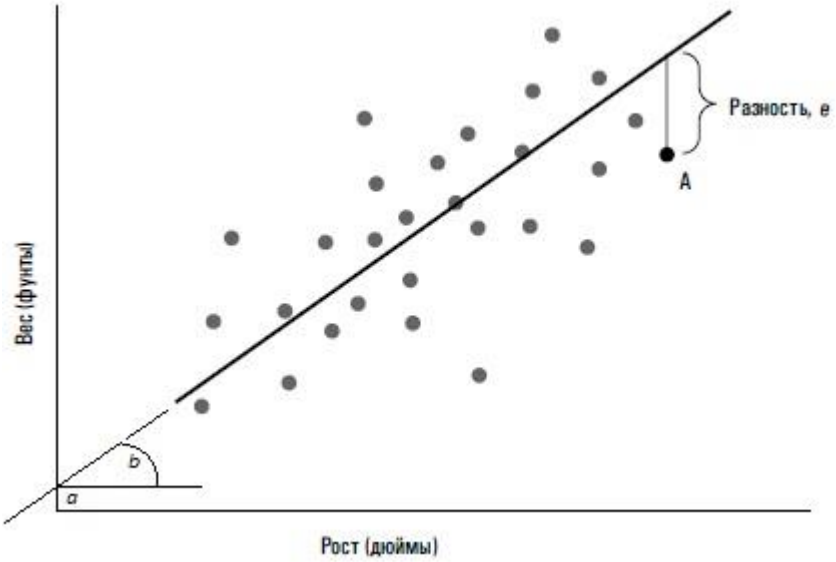

Рис. 8. Линия наилучшего приближения для роста и веса

В результате мы получаем не только линию как таковую, но и уравнение, описывающее ее. Оно известно, как уравнение регрессии и имеет следующий вид: $y = a + bx$, где y — вес в фунтах, a — отрезок, отсекаемый этой линией на оси Y (то есть значение y, когда $x = 0$), b — коэффициент наклона линии, x — рост в дюймах.

Переменная, которая подлежит объяснению, — в нашем случае это вес — называется зависимой переменной, так как она зависит от других факторов. Переменные, используемые для объяснения зависимой переменной, называются объясняющими переменными, поскольку они объясняют интересующий нас результат:

(5) Вес = –135 + 4,5 x (Рост в дюймах)

Если мы повторим исследование для нескольких крупных выборок (по 2000–300 человек), то получим несколько иные коэффициенты a и b, но все они будут распределены по нормальному закону вокруг «истинных» a_0 и b_0. Однако, при использовании небольшой выборки данных — например группы из 20 взрослых человек, полученные нами разные коэффициенты регрессии будут распределены вблизи «истинной» зависимости между ростом и весом в совокупности взрослых американцев по закону, известному как t-распределение, или распределение Стьюдента. t-

распределение характеризуется большей степенью разброса, чем нормальное распределение, и, следовательно, имеет «более толстые хвосты».

Регрессионный анализ дает еще одну статистику, заслуживающую внимания, R^2, которая представляет собой показатель суммарной величины разброса, объясняемого уравнением регрессии. Величина R^2 говорит нам, какая доля этого разброса вокруг среднего значения ассоциируется лишь с различиями в росте. В нашем случае эта доля составляет 0,25, или 25%. Более значимым может быть то обстоятельство, что 75% этого разброса в весе для нашей выборки остаются необъясненными. Есть очевидные факторы, помимо роста, которые могут нам помочь их объяснить.

Посредством регрессионного анализа (часто называемого множественным регрессионным анализом, если в нем задействовано несколько объясняющих переменных, или многофакторным регрессионным анализом) можно вычислить некий коэффициент регрессии для каждой объясняющей переменной, задействованной в уравнении регрессии. Скажем, какова зависимость между возрастом и весом среди людей одного и того же пола, и роста. Когда нам приходится иметь дело с несколькими объясняющими переменными, соответствующие данные уже невозможно отобразить на двумерной диаграмме. Когда мы вычислим уравнение регрессии, включающее рост и возраст в качестве объясняющих переменных, то получим вот что:

(6) Вес = –145 + 4,6 x (Рост в дюймах) + 0,1 x (Возраст в годах)

Коэффициент для роста несколько увеличился. После того как мы включили в нашу регрессию возраст, у нас появилось уточненное понимание зависимости между ростом и весом.

Теперь давайте добавим еще одну переменную — пол. Тут есть один нюанс: пол может принимать лишь два значения (мужской и женский). Используем так называемую двоичную, или фиктивную переменную: 1 для женщин и 0 — для мужчин:

(7) Вес = –118 + 4,3 x (Рост в дюймах) + 0,12 x (Возраст в годах) – 4,8 (Если пол женский)

Значение R^2 для этой регрессии повысилось по сравнению с (5) с 0,25 до 0,29. Напомню: нулевая величина R^2 означает, что уравнение регрессии прогнозирует вес любого человека в данной выборке ничуть не лучше, чем среднее значение; если же R^2 равно 1, то наше уравнение регрессии идеально прогнозирует вес каждого человека в данной выборке. Существенная доля разброса величин веса среди членов данной выборки остается необъясненной.

Множественный регрессионный анализ — лучший из имеющихся в нашем распоряжении инструмент для поиска существенных закономерностей в больших и сложных совокупностях данных.

t-распределение представляет собой некую совокупность, или «семейство», функций плотности вероятности, которые варьируются в зависимости от величины выборки. В частности, чем больше данных содержится в выборке, тем больше «степеней свободы» у нас имеется при определении подходящего распределения, которое служит нам эталоном для оценки результатов (степень свободы и в русской статистической литературе обозначается как *df* от англ., degrees of freedom).

Степень свободы примерно равны количеству наблюдений в выборке. Например, регрессионный анализ с выборкой, размер которой составляет 10, и с единственной объясняющей переменной, имеет 9 степеней свободы. Чем больше степеней свободы, тем больше уверенность, что выборка представляет истинную совокупность, и тем «плотнее» будет распределение, как следует из приведенной ниже диаграммы. Когда число степеней свободы увеличивается, t-распределение сходится к нормальному распределению. Чем меньше степеней свободы (и, следовательно, чем «толще» хвосты у соответствующего t-распределения), тем больше должна быть t-статистика, чтобы мы могли отвергнуть нулевую гипотезу на некотором заданном уровне статистической значимости.

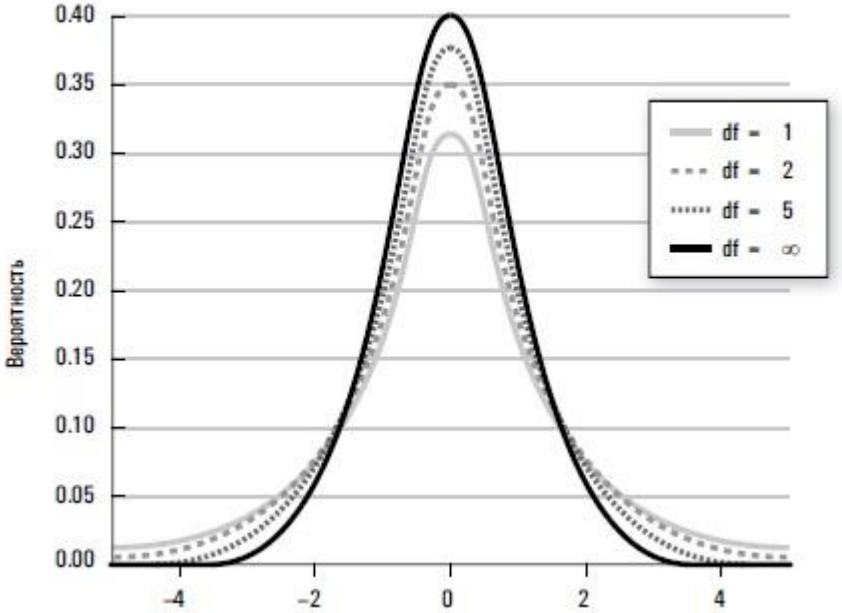

Рис. 9. Семейство t-распределений

Глава 12. Типичные регрессионные ошибки

Регрессионный анализ — это своего рода водородная бомба в арсенале статистики. В чем же причина проблем с регрессионным анализом? Таких причин очень много. Регрессионный анализ позволяет получить точные ответы на сложные вопросы, но они могут быть правильными или неправильными. Вот семь самых типичных злоупотреблений этим замечательным инструментом.

Использование регрессии для анализа нелинейной связи. Запомните: коэффициент регрессии описывает степень наклона «линии наилучшего приближения» для рассматриваемых вами данных; непрямая линия будет характеризоваться разными степенями наклона в разных точках.

Корреляция и причинно-следственные зависимости — не одно и то же. Регрессионный анализ может лишь продемонстрировать взаимосвязь между двумя переменными. С помощью только статистики невозможно доказать, что изменение одной переменной *обусловило* изменение другой переменной.

Обратная причинно-следственная зависимость. Причинно-следственные связи могут действовать в обоих направлениях. У нас должны быть все основания полагать, что наши объясняющие переменные влияют на зависимую переменную, а не наоборот.

Систематическая ошибка, вызванная пропущенной переменной. Результаты регрессии будут вводить нас в заблуждение и страдать неточностью в случае отсутствия в уравнении регрессии какой-либо важной объясняющей переменной, особенно если другие переменные в этом уравнении «подхватывают» данный эффект.

Сильно коррелированные объясняющие переменные. Когда две объясняющие переменные сильно коррелированны между собой, исследователи обычно используют в уравнении регрессии какую-то одну из них.

Экстраполяция за границы имеющихся данных.

Ваши результаты могут быть поставлены под угрозу, если вы включите в уравнение регрессии *чересчур большое число переменных*, особенно если речь идет о дополнительных объясняющих переменных. Если вы включите в уравнение регрессии достаточно большое число лишних переменных, то одна из них, по чистой случайности, обязательно достигнет порога статистической значимости. Еще одна опасность заключается в том, что лишние переменные порой не так-то легко распознать именно как лишние. Опытные исследователи могут всегда обосновать теоретически, постфактум, почему та или иная необычная переменная, которая в действительности совершенно бессмысленна, оказывается статистически значимой.

Приложение. Статистическое программное обеспечение

Подозреваю, что вы не будете выполнять статистический анализ с помощью карандаша, бумаги и карманного калькулятора. *Microsoft Excel* — пожалуй, самая широко используемая программа для вычисления простых статистических показателей, таких как среднее значение и среднеквадратическое (стандартное) отклонение. Кроме того, с помощью Excel можно выполнять базовый регрессионный анализ.

Stata — статистический пакет, используемый специалистами-исследователя ми во всем мире; его интерфейс отличается серьезным, научным видом. Впрочем, Stata окажется не самым идеальным инструментом, если ваша цель — оперативно строить графики на основе имеющихся данных.

SAS. В этом пакете предусмотрены хорошие инструменты визуализации данных.

R — это бесплатный (с открытым исходным кодом) популярный статистический пакет. Прелесть пакета R не только в его бесплатности, но и в необычайной пластичности и гибкости.

IBM SPSS

На отечественном рынке представлены: универсальный статистический пакет STATISTICA, который может служить не только эффективным инструментом для научных исследований, но и чрезвычайно удобной средой для обучения методам статистического анализа. Из российских разработок отметим пакеты STADIA, ЭВРИСТА, МЕЗОЗАВР, САНИ, СТАТЭксперт.

По теме см. также

Левин. Статистика для менеджеров с использованием Microsoft Excel

Идеи Байеса для менеджеров

Как с помощью диаграммы приукрасить действительность? или о факторе лжи Эдварда Тафти

Дарелл Хафф. Как лгать при помощи статистики.