Голая статистика. Самая интересная книга о самой скучной науке

Автор:
Чарльз Уилан

Голая статистика. Самая интересная книга о самой скучной науке

Чарльз Уилан

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

Charles Wheelan

Naked Statistics: Stripping the Dread from the Data

Научный редактор Александр Минько

Издано с разрешения Janklow & Nesbit Associates и литературного агентства Prava I Perevodi

Книга рекомендована к изданию Федором Царевым

Правовую поддержку издательства обеспечивает юридическая фирма «Вегас-Лекс».

* * *

Посвящается Кэтрин

Введение

Почему я ненавидел вычисления, но обожал статистику

Я всегда недолюбливал математику. Мне вообще не нравятся числа как таковые. На меня не производят впечатления заумные формулы, не имеющие реального практического применения. Но особенно, учась в средней школе, я не любил алгебру, по той простой причине, что никто так и не смог мне толком объяснить, почему я должен изучать ее. Как вычислить площадь под параболой? Кому это

нужно?

Кстати, один из самых значимых моментов в моей жизни пришелся на время учебы в выпускном классе. Это было в конце первого семестра; я готовился к сдаче последнего экзамена, однако чувствовал, что шансов на высокий результат мало. (Должен сказать, что к тому времени меня уже приняли в колледж, в который я давно мечтал поступить, поэтому какая-либо мотивация особо усердствовать при подготовке к школьным экзаменам у меня отсутствовала.) Вытянув экзаменационный билет и взглянув на вопросы, я понял, что быть беде. Причем даже не потому, что я не знал правильных ответов, а потому, что я вообще не понимал, о чем идет речь. Я не впервые приходил на экзамены плохо подготовленным, но по крайней мере, как правило, знал, в каких вопросах «мелко плаваю». Однако на сей раз я, похоже, не знал почти ничего. Поломав какое-то время над вопросами экзаменационного билета голову и поняв, что катастрофа неизбежна, я подошел к столу, за которым сидела наша преподавательница (помню, ее звали Кэрол Смит). «Миссис Смит, – произнес я, – я вообще не понимаю, о чем говорится в моем экзаменационном билете».

Должен сказать, что я не нравился миссис Смит гораздо больше, чем она нравилась мне. Да, сейчас я могу сознаться, что иногда злоупотреблял своими правами председателя ученической ассоциации и планировал общешкольные собрания таким образом, чтобы время их проведения совпадало с уроками по началам анализа, которые вела миссис Смит (уроки приходилось отменять). Да, мы с одноклассниками время от времени клали букет цветов на стол миссис Смит перед ее приходом в класс (предполагалось, что это были цветы от некоего «тайного обожателя») и буквально давились от смеха, наблюдая, как она, войдя в класс и заметив букет, ужасно смущалась и краснела. И еще: поступив в колледж, я сразу же перестал выполнять домашние задания по математике.

Поэтому, когда я подошел к миссис Смит и сообщил, что не понимаю вопросов в экзаменационном билете, она не посочувствовала мне. «Чарльз, – сказала она громко, обращаясь, по-видимому, не только ко мне, но и ко всем присутствующим в классе, – если бы вы работали в течение семестра и добросовестно готовились к экзамену, то вопросы не показались бы вам непонятными». Это был железный аргумент.

Я молча вернулся на место. Через несколько минут Брайан Арбеттер, гораздо лучше меня разбирающийся в математическом анализе, подошел к миссис Смит и что-то прошептал ей на ухо. Она что-то тихо ответила ему, а затем произошло нечто неожиданное. «Попрошу минутку внимания, – обратилась миссис Смит к классу. – Оказалось, что по ошибке я принесла на экзамен билеты для второго семестра». С момента начала экзамена прошло уже достаточно много времени, поэтому было решено прервать его и перенести на другой день.

Не могу описать эйфорию, охватившую меня тогда. Одним словом, все закончилось как нельзя лучше. Со временем я женился на замечательной девушке. У нас родилось трое детей. Я опубликовал несколько книг и побывал в таких местах, как Тадж-Махал и храмовый комплекс Ангкор-Ват. Тем не менее день, когда моя преподавательница математики понесла заслуженное наказание, остается одним из самых памятных в моей жизни.
(То обстоятельство, что в тот день я чуть не провалил экзамен, не оказало существенного влияния на мою дальнейшую счастливую жизнь.)

Инцидент, случившийся на экзамене по математике, весьма красноречиво (но не до конца) иллюстрирует мои отношения с этим предметом.
Что любопытно, к школьному курсу физики я не испытывал такой неприязни. Более того, физика мне нравилась, несмотря на то что она тоже относится к точным наукам и широко использует математический аппарат. Как это объяснить? Дело в том, что физика гораздо ближе к жизни и практике, чем математика. Я прекрасно помню, как учитель физики показывал нам во время ежегодного чемпионата США по бейсболу, как использовать базовую формулу ускорения, чтобы оценить дальность хоумрана[1 - Хоумран – удар в бейсболе, при котором мяч перелетает через все игровое поле; дает право совершить перебежку по всем базам и принести своей команде очко. Прим. перев.]. Это здорово, притом что у той же формулы есть множество других сфер применения.

Во время учебы в колледже одним из моих любимых предметов была теория вероятностей – опять же потому, что она позволяет лучше понять ряд интересных реальных ситуаций. Теперь я знаю, что моя неприязнь к математическому анализу, который мы изучали в старших классах школы, объясняется тем, что никто нам так и не растолковал, какое отношение этот предмет имеет к реальной жизни. Если вас не приводит в восхищение элегантность самих математических формул, – а меня, безусловно, нет, – то ничего, кроме смертельной скуки, они у вас не вызывают. Не исключаю,

что в этом во многом виноваты наши школьные учителя, которые не сумели привить нам любовь к математике.

Теперь настало время поговорить собственно о статистике (в рассказе о которой не обойтись без теории вероятностей). Я обожаю статистику: ее можно использовать для объяснения очень многих вещей, от тестирования ДНК до бессмысленности участия в разного рода лотереях. Статистика способна помочь в выявлении факторов, связанных с такими недугами, как рак и заболевания сердца, а также в обнаружении манипуляций с проведением стандартизованных тестов. Благодаря ей вы даже можете выиграть некоторые игровые шоу. В детстве я любил смотреть знаменитую телепрограмму под названием Let’s Make a Deal («Совершим сделку») с ее не менее знаменитым ведущим Монти Холлом. В конце каждого выпуска передачи участник, добравшийся до финала, становился вместе с Монти Холлом перед тремя большими дверьми – Дверью № 1, Дверью № 2 и Дверью № 3, – и Монти Холл объяснял ему, что за одной из них скрывается очень ценный приз – скажем, новый автомобиль, а за двумя другими – козел. Финалист должен был выбрать одну из дверей и получить то, что находилось за нею.

Вероятность того, что финалист выберет дверь, за которой скрывался самый ценный приз, составляла 1 к 3. Однако в игре Let’s Make a Deal был предусмотрен интересный трюк, приводивший в восхищение статистиков и ставивший в тупик остальных. После того как финалист указывал на какую-то из трех дверей, Монти Холл открывал одну из двух оставшихся дверей, за которой всегда оказывался козел. Допустим, к примеру, что финалист выбрал Дверь № 1. После этого Монти Холл открывал Дверь № 3 – за ней находился козел. При этом две другие двери – Дверь № 1 и Дверь № 2 – оставались закрытыми. Если ценный приз скрывался за Дверью № 1, то финалист становился победителем игры, если же за Дверью № 2, то считался проигравшим. Но далее ситуация становилась еще более интригующей: Монти Холл спрашивал у финалиста, не передумал ли он и не считает ли, что ценный приз находится не за Дверью № 1, а за Дверью № 2. Напоминаю, что к этому времени Дверь № 1 и Дверь № 2 остаются закрытыми, и единственная новая информация, которой располагает финалист, состоит в том, что за одной из них скрывается козел.

Следует ли финалисту отказаться от своего прежнего выбора и указать на Дверь № 2?

Отвечаю: да, следует. Почему? Объяснение найдете в главе 5? (#litres_trial_promo).

Парадокс статистики в том, что она вездесуща – начиная с так называемых средних показателей и заканчивая голосованием на выборах президента, – но при этом пользуется репутацией неинтересной и малопонятной. Многие книги и курсы по статистике перегружены математическими формулами и специальным жаргоном. Поверьте, все эти технические подробности важны и по-своему привлекательны, но для человека, который не страдает избытком интуиции и воображения, выглядят как абракадабра, способная вызвать исключительно отторжение. Если вы не понимаете, зачем изучать статистику, то лучше не беритесь. Именно поэтому в каждой главе книги я пытаюсь ответить на основной вопрос, который безуспешно задавал в школе своему преподавателю математики: зачем все это нужно лично мне?

Эта книга об интуиции. Я старался по возможности избегать употребления математических формул, уравнений и графиков, в тех же случаях, когда без них нельзя было обойтись, я преследовал четкую конкретную цель. Множество приведенных мною примеров призваны убедить вас в целесообразности изучения этой дисциплины. Статистика может быть действительно интересной и по большей части не так сложна, как кажется поначалу.

Идея написать эту книгу родилась через несколько лет после моей неудавшейся попытки постичь сущность математического анализа под чутким руководством миссис Смит. В магистратуре мне предстояло изучать экономику и политологию. Но прежде чем читать нам курс экономики, меня (что неудивительно) и большинство моих сокурсников направили в так называемый математический лагерь, чтобы мы ликвидировали там свои многочисленные пробелы в познании этого предмета. На протяжении трех недель мы чуть ли не круглосуточно изучали математику в плохо проветриваемом полуподвальном помещении.

В какой-то из таких дней я как никогда был близок к тому, что принято называть прозрением. Преподаватель пытался объяснить нам условия, при которых сумма бесконечного ряда сходится к конечному числу. Постарайтесь следить за ходом моих рассуждений, а я попробую описать суть данной концепции. (Возможно, сейчас вы испытываете те же ощущения, что и я, сидя в душном полуподвальном помещении.) Бесконечный ряд представляет собой последовательность чисел, уходящую куда-то в… бесконечность, например 1 + ? + ? + ? + … Многоточие означает, что эта последовательность продолжается

до бесконечности.

На этом месте мы впали в ступор. Используя какое-то доказательство (какое именно, уже не помню), преподаватель пытался убедить нас, что хоть такая последовательность чисел и может продолжаться до бесконечности, тем не менее она все равно сойдется (приблизительно) к какому-то конечному числу. Один из моих одноклассников, Уилл Уоршоер, сильно в этом сомневался (собственно, как и я). Разве так бывает?

Затем меня осенило: мне показалось, я понял, что именно пытается втолковать нам преподаватель. Я повернулся к Уиллу и изложил ему версию, которая только что возникла у меня в голове.

Допустим, вы стали ровно в двух футах от стены. Теперь придвиньтесь к стене на половину этого расстояния (1 фут). В результате вы окажетесь в одном футе от стены.

Еще раз придвиньтесь к стене на половину оставшегося расстояния (6 дюймов, или ? фута). Находясь в 6 дюймах от стены, повторите описанные выше действия (придвиньтесь к стене на 3 дюйма, или ? фута). Выполните их еще раз (придвиньтесь к стене на 1? дюйма, или ? фута). И так далее.

Постепенно вы почти упретесь в стену. (Например, окажетесь на расстоянии 1/1024 дюйма от нее, а затем придвинетесь еще на половину этого пути, или на 1/2048 дюйма.) Но ключевым здесь является слово почти: сколько бы раз вы ни повторяли это действие, расстояние между вами и стеной никогда не станет в точности равно нулю, поскольку, по определению, каждое такое продвижение приближает вас к стене лишь на половину оставшегося расстояния. Иными словами, вы все время будете оказываться бесконечно близко к стене, но никогда не упретесь в нее. Если измерять ваши продвижения в футах, то соответствующую последовательность можно описать как 1 + ? + ? + ? ...

Именно в этом и заключалось мое прозрение. Сколько бы вы ни продвигались таким способом к стене (а вы будете делать это до бесконечности), совокупное расстояние, пройденное вами, не может превышать 2 футов, то есть вашего исходного расстояния от стены. С математической точки зрения, совокупное расстояние, пройденное вами, можно приравнять к 2 футам, что весьма удобно

в плане вычислений. Математик сказал бы, что сумма бесконечного ряда 1 фут + ? фута + ? фута + ? фута … сходится к 2 футам, то есть именно то, что пытался объяснить нам преподаватель.

Что показательно, в процессе объяснения мне удалось убедить в правильности моей версии не только Уилла, но и самого себя. Я уже не помню дословно математического доказательства того, что сумма бесконечного ряда при определенных условиях может сходиться к конечному числу (хотя могу найти его в соответствующем учебнике по математике), но исходя из собственного опыта готов утверждать, что благодаря интуиции математика и другие технические детали становятся гораздо понятнее (но необязательно наоборот).

Задача этой книги – доходчиво объяснить самые важные статистические концепции не только тем, кому приходится осваивать их в плохо проветриваемых, душных помещениях, но и тем, кого влечет магия чисел.

Хотя выше я был вынужден признать, что базовые инструменты статистики, к сожалению, менее интуитивно понятны и доступны, чем следовало бы, сейчас я намерен сделать несколько на первый взгляд противоречащее этому заявление, а именно: статистика может быть более чем доступной для понимания в том смысле, что каждый из нас, вооружившись исходными данными и компьютером, способен выполнить сложные статистические выкладки, нажав буквально несколько клавиш. Однако в случае, если исходных данных недостаточно или статистические методы используются некорректно, появляется риск, что наши выводы не только могут ввести нас в заблуждение, но и оказаться потенциально опасными. Рассмотрим следующую гипотетическую новость из интернета: «Люди, которые делают короткие перерывы в работе в течение дня, имеют гораздо больше шансов умереть от рака». Представьте появление на экране такого сообщения, когда вы занимаетесь веб-серфингом. Согласно весьма впечатляющим результатам обследования 36 000 работников (огромный массив данных, не правда ли?!), у тех, кто выходил из офиса на регулярные десятиминутные перерывы в течение каждого рабочего дня, вероятность заболевания раком в последующие пять лет оказалась на 41 % выше, чем у тех, кто офисы не покидал. Понятно, что узнав такую новость, мы обязаны как-то на нее реагировать: возможно, провести общенациональную кампанию за запрет коротких перерывов в течение рабочего дня.

А может, следует подойти к проблеме с другой стороны и задуматься над тем, чем именно обычно занимаются работники во время таких десятиминуток? Не мне вам рассказывать, что многие кучкуются неподалеку от входа в офисное помещение, покуривая сигареты (и создавая при этом облако дыма, через которое вынуждены проходить те, кто входит или выходит из здания). Смею предположить, что именно сигареты, а не кратковременные перерывы в работе, являются основной причиной раковых заболеваний. Большинству читателей этот пример покажется абсурдным, но могу вас заверить, что многие статистические умозаключения, встречающиеся в реальной жизни, оказываются не менее абсурдными после их тщательного анализа.

Статистика подобна мощному оружию, полезному в случае его правильного применения и потенциально разрушительному в неумелых руках. Прочитав эту книгу, вы, конечно, не станете профессиональным статистиком, но по крайней мере она научит вас осторожному обращению со статистическими данными и убережет от их неверной интерпретации, которая может иметь непредсказуемые последствия.

Книга, которую вы держите в руках, – не учебник, и это обеспечило мне достаточно высокую степень свободы в выборе тем и способов изложения материала. Цель этой книги – ознакомить читателей со статистическими концепциями в их непосредственной связи с повседневной жизнью. Как ученые приходят к выводу о том, что некий фактор служит причиной раковых заболеваний? Каков механизм опросов общественного мнения (и что может исказить их результаты)? Кто «лжет, манипулируя статистическими данными», и как им это удается? Как компания, выпустившая вашу кредитную карточку, использует информацию о совершаемых вами покупках, чтобы прогнозировать вероятность пропуска вами платежа? (Да-да, они и такое умеют!)

Если вы хотите правильно интерпретировать числа, озвученные в новостях, и использовать необычайную (и все более возрастающую) силу данных, то материал этой книги – именно то, что вам нужно. В конечном счете я надеюсь убедить вас в справедливости мысли, высказанной шведским математиком и писателем Андрейсом Дункельсом: «Опираясь на статистику, легко лгать, но без статистики очень трудно выяснить истину».

Но я мечтаю о большем. Мне хочется, чтобы вы начали получать наслаждение от статистики. Идеи, положенные в ее основу, чрезвычайно интересны и актуальны. Главное – уметь отделять по-настоящему важные идеи

от технических подробностей, которые способны стать для вас непреодолимым препятствием. Этому я и стараюсь вас научить на страницах данной книги.

1. В чем суть?

Я заметил один любопытный феномен. Хотя студенты часто жалуются, что статистика – неинтересная и малопонятная наука, тем не менее, выйдя из аудитории, они охотно обсуждают свои спортивные достижения и средние результаты, которых добились летом, или коэффициент изменчивости погоды (в холодное время года), или свои баллы в колледже (этот вопрос не волнует их только во время каникул). Они признают, что «рейтинг распасовщика» – статистический показатель, выражающий в одном числе эффективность действий куортербека[2 - Куортербек – распасовщик, играющий помощник тренера в американском футболе. Прим. перев.], – весьма некорректно отражает качество его игры. Те же самые исходные данные (коэффициент удачного завершения, среднее число ярдов на каждую попытку паса, процент тачдаун-пасов[3 - Тачдаун – в американском футболе: пересечение мячом или игроком с мячом линии зачетного поля соперника. Прим. перев.] на каждую попытку паса и коэффициент перехватов мяча) можно было бы скомбинировать как-то по-другому, например присвоить каждой составляющей определенный весовой коэффициент и в результате создать другой, не менее надежный показатель эффективности действий куортербека. Однако все, кто интересуется американским футболом, должны признать, что наличие рейтинга распасовщика весьма удобно.

Является ли данный рейтинг идеальным? Разумеется нет. Статистика крайне редко предлагает единственно верный вариант оценивания чего бы то ни было. Предоставляет ли данный показатель возможность получить важную информацию? Разумеется да. Это превосходный инструмент, позволяющий быстро сравнивать эффективность действий двух куортербеков в один и тот же день. Я болею за команду Chicago Bears. Во время серии плей-офф 2011 года Chicago Bears играли с Packers (Packers одержали победу). Я мог бы описать этот матч множеством способов, потратив не одну страницу на его анализ. Но вот более сжатый вариант: рейтинг распасовщика куортербека Chicago Bears Джея Катлера составил в тот день 31,8, а куортербека Green Bay Аарона Роджерса – 55,4. Аналогично мы можем сравнить эффективность действий Джея Катлера

с эффективностью его же действий в одной из предыдущих игр того же сезона против команды Green Bay, когда его рейтинг распасовщика равнялся 85,6. Эти показатели способны многое сказать тому, кто хочет понять, почему ранее в том сезоне Chicago Bears выиграли у Packers, а затем потерпели поражение в серии плей-офф.

Это может служить весьма поучительным – и достаточно лаконичным – объяснением итогов футбольного сезона 2011 года. Однако нет ли здесь чрезмерного упрощения? Да, именно в этом и заключается сила и слабость любой описательной статистики. Один-единственный показатель говорит вам, что Джей Катлер продемонстрировал в играх плей-офф с участием Chicago Bears худшую эффективность, чем Аарон Роджерс. С другой стороны, тот же показатель ничего не скажет вам о том, потерпел ли тот или иной куортербек в ходе игры досадную неудачу (например, его идеальная передача не была поймана принимающим, а затем перехвачена), удавалось ли ему действовать с максимальной отдачей в определяющих с точки зрения конечного результата ключевых розыгрышах (поскольку весовые коэффициенты всех розыгрышей одинаковы и не зависят от их важности для конечного результата), насколько успешно действовала защита и т. д.

Парадоксально, что те же люди, которые свободно рассуждают о статистике в контексте спорта, погоды или академической успеваемости, начинают теряться, когда исследователь переходит к объяснению чего-нибудь наподобие коэффициента Джини – стандартного инструмента в экономике, демонстрирующего степень неравенства доходов. Ниже я объясню суть данного коэффициента, сейчас же для нас главное – признать, что между коэффициентом Джини и рейтингом распасовщика нет принципиальных отличий. Оба позволяют представить сложную информацию в виде единственного числового показателя. Как таковой коэффициент Джини обладает достоинствами большинства описательных статистик, а именно: обеспечивает удобный способ сравнения распределения дохода в двух странах или в одной стране в разные моменты времени.

Коэффициент Джини помогает оценить по шкале от 0 до 1, насколько равномерно распределяется в стране совокупный доход. Этот статистический показатель можно вычислить для материального благосостояния или годового дохода, причем он может быть рассчитан на индивидуальном или семейном уровне. (Все эти значения будут сильно коррелированны, но не идентичны.) У коэффициента Джини, подобно рейтингу распасовщика, нет какого-либо

собственного, внутренне присущего ему смысла – это всего лишь инструмент для сравнения. У страны, в которой все семьи имеют одинаковый уровень благосостояния, был бы нулевой коэффициент Джини. А в той стране, где все богатство сосредоточено в руках одной семьи, он равнялся бы единице. Как вы, наверное, догадались, чем ближе значение к единице, тем выше степень расслоения общества. Согласно данным Центрального разведывательного управления (между прочим, ЦРУ активно занимается сбором статистических данных)[1 - Central Intelligence Agency, The World Factbook, https://www.cia.gov/library/publications/the-world-factbook/ (https://www.cia.gov/library/publications/the-world-factbook/).], коэффициент Джини для Соединенных Штатов равен 0,45. И что?

Если этот показатель поместить в определенный контекст, он может многое нам рассказать. Например, коэффициент Джини для Швеции составляет 0,23; для Канады – 0,32; для Китая – 0,42; для Южной Африки 0,65[4 - Коэффициент Джини иногда умножают на 100, чтобы он выражался целым числом. В таком случае для Соединенных Штатов он равнялся бы 45.]. Анализ этих значений позволяет получить представление о том, какое место в мире занимают Соединенные Штаты с точки зрения неравенства распределения доходов. Можно также проанализировать, как коэффициент Джини изменяется со временем в одной и той же стране. Например, в 1997 году для Соединенных Штатов он равнялся 0,41, а в следующем десятилетии достиг 0,45 (самые последние данные ЦРУ относятся к 2007 году). Это дает возможность составить объективную картину нарастания неравенства в распределении богатства по мере процветания Соединенных Штатов (во всяком случае на рассматриваемом отрезке времени). Кроме того, мы можем сравнить изменения коэффициента Джини в разных странах примерно за один и тот же период времени. Скажем, в Канаде за указанный период он практически остался прежним. Швеция на протяжении двух последних десятилетий переживала фазу значительного экономического роста, однако коэффициент Джини в ней фактически снизился с 0,25 в 1992 году до 0,23 в 2005-м; это означает, что за указанный период Швеция не только стала богаче, но и доходы в ней начали распределяться более равномерно.

Можно ли считать коэффициент Джини идеальным показателем неравенства? Отнюдь нет – точно так же как рейтинг распасовщика нельзя считать идеальным показателем эффективности действий куортербека. Но несомненно одно: он позволяет нам получить весьма ценную информацию о социально значимом явлении – неравенстве в распределении богатства – в достаточно удобном формате.

Итак, мы медленно продвигаемся к получению ответа на вопрос, поставленный в названии этой главы: в чем суть? А в том, что статистика помогает нам обрабатывать данные, хотя на самом деле это всего лишь еще одно название информации. Подчас эти данные тривиальны, как в случае спортивной статистики, а подчас проливают свет на природу человеческого общества, как в случае коэффициента Джини.

Но, как любят повторять в телевизионных рекламных роликах, это еще не все! Хол Вариан, главный экономист компании Google, в интервью The New York Times сказал, что в следующем десятилетии работа со статистическими данными станет «модной профессией», а точнее «сексуальной» (дословное выражение Хола Вариана: the sexy job)[2 - Steve Lohr, For Today’s Graduate, Just One Word: Statistics, New York Times, August 6, 2009.]. Я, наверное, окажусь первым, кто пришел к выводу о весьма превратном представлении некоторых экономистов о том, что следует считать «сексуальным». Тем не менее предлагаю рассмотреть несколько никак не связанных между собой вопросов.

• Как уличить учебные заведения в подтасовке результатов стандартизированных тестов?

• Откуда Netflix[5 - Netflix – американская компания, поставщик фильмов и сериалов на основе потокового мультимедиа. Прим. перев.] известно о том, какого рода фильмы вам нравятся?

• Как определить, какие вещества и образ жизни вызывают раковые заболевания, учитывая, что мы не можем проводить над людьми экспериментов, приводящих к заболеванию раком?

• Можно ли рассчитывать на более успешный исход хирургической операции, если молиться за пациента?

• Существует ли реальная экономическая выгода в получении диплома какого-либо из престижных колледжей или университетов?

• Что является причиной роста заболеваемости аутизмом?

Статистика способна помочь нам (или, как мы рассчитываем, поможет в ближайшем будущем) получить ответы на эти вопросы.

Наш мир все быстрее и быстрее генерирует все большие и большие объемы данных. Тем не менее, как справедливо отметила The New York Times, «данные – всего лишь исходный материал знаний»[3 - Steve Lohr, For Today’s Graduate, Just One Word: Statistics, New York Times, August 6, 2009.],[6 - Исторически так сложилось, что слово «данные» (data) используется во множественном числе (например, «эти данные являются весьма обнадеживающими»). Это слово можно употреблять и в единственном числе: «данное» (datum); в этом случае речь идет о каком-то отдельно взятом элементе данных (например, ответ одного человека на какой-то один вопрос анкеты, используемой при опросе общественного мнения). Употребление слова «данные» во множественном числе сигнализирует каждому, кто занимается серьезными исследованиями, о том, что вы знаете толк в статистике. С учетом сказанного многие специалисты по грамматике, а также многие издания, такие как The New York Times, в настоящее время согласны с тем, что слово «данные» может означать как единственное, так и множественное число, как свидетельствует приведенная мной цитата из The New York Times.]. Статистика – самый мощный из имеющихся в нашем распоряжении инструментов для практического использования информации, например для оценивания эффективности действий бейсболистов или более справедливой оплаты труда преподавателей. Ниже приведен краткий обзор того, как статистика способна придать смысл исходным данным.

Описание и сравнение

Счет партии в боулинг является описательной (дескриптивной) статистикой. То же можно сказать и о каком-либо среднем показателе (например, в спорте). Большинство американских спортивных болельщиков в возрасте старше пяти лет неплохо разбираются в описательной статистике. Мы используем численные показатели в спорте и других сферах жизни для подытоживания информации. Насколько Микки Мэнтл был хорош как бейсболист? Его итоговый рейтинг как хиттера составил 0,298. Для бейсбольных болельщиков это весьма красноречивое число. Итоговый рейтинг 0,298 – выдающийся показатель, если принять во внимание, что в нем учитываются результаты Микки Мэнтла за восемнадцать лет карьеры профессионального бейсболиста[4 - Baseball-Reference.com (http://www.baseball-reference.com/), http://www.baseball-

reference.com/players/ (http://www.baseball-reference.com/players/)]. (Хотя, согласитесь, если итог жизни человека можно выразить одним-единственным числом, это несколько разочаровывает и настраивает на мысли о бренности человеческого бытия.) Разумеется, фанаты бейсбола должны помнить о существовании другой описательной статистики, которая, возможно, отражает ценность того или иного бейсболиста гораздо лучше, чем пресловутый средний показатель.

Академическая успеваемость учащихся школ и колледжей в США оценивается с помощью среднего балла. В стране используется шкала с буквенными обозначениями, где каждой букве соответствует определенный балл: как правило, A = 4 балла, B = 3 балла, C = 2 балла и т. д. По окончании учебного заведения, когда абитуриенты поступают в колледжи, а выпускники колледжей подыскивают себе работу, средний балл становится удобным инструментом для оценивания их академического потенциала. Тот, у кого средний балл 3,7, явно сильнее выпускника со средним баллом 2,5. Таким образом, средний балл является весьма полезной описательной статистикой. Его легко вычислить, понять и сравнивать с баллами других учащихся.

Тем не менее данный показатель не идеален. В нем не учитывается сложность учебных программ, которые проходят разные ученики. Как можно сравнивать знания учащегося со средним баллом 3,4, обучавшегося по относительно легкой программе, и его сверстника со средним баллом 2,5, изучавшего математику, физику, химию и другие сложные предметы? В свое время я посещал школу, которая пыталась решить эту проблему, присваивая таким дисциплинам дополнительные весовые коэффициенты, в результате чего оценка A по предмету повышенной трудности соответствовала пяти баллам, а по обычному предмету приравнивалась к четырем. Однако у данного подхода были существенные минусы. Моя мать довольно быстро уяснила, как эта «поправка» влияет на средний балл. Дело в том, что для таких учеников, как я (изучавших много сложных предметов), максимальная оценка A по любому из обычных предметов (например, по физкультуре или основам безопасности жизнедеятельности) не могла превышать 4 баллов, что снижало средний балл, как бы хорошо мы ни учились. В результате родители запретили мне посещать в школе курсы вождения автомобиля, поскольку даже самые высокие оценки по этому курсу уменьшали мои шансы на поступление в какой-либо престижный колледж и последующие занятия писательским трудом. Поэтому они отправили меня в частную (платную) школу вождения, которую мне пришлось посещать летом.

Глупость? Конечно! Но одной из тем, которые я затрону в этой книге, будет опасность чрезмерного увлечения любой из описательных статистик, поскольку это может привести к ошибочным умозаключениям и подтолкнуть к нежелательным действиям. В первоначальном варианте книги я использовал выражение «упрощенная описательная статистика», однако в конечном счете выбросил слово «упрощенная», поскольку оно показалось мне заведомо избыточным. Описательная статистика для того и существует, чтобы упрощать, что всегда подразумевает некоторую потерю нюансов и деталей. Каждый, кто работает с числами, должен воспринимать это как данность.

Умозаключения

Сколько бездомных живет на улицах Чикаго? Как часто женатые пары занимаются сексом? На первый взгляд у этих вопросов нет ничего общего. На самом же деле на каждый из них можно ответить (правда, не с абсолютной точностью) с помощью базовых статистических инструментов. 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».

Начнем с вопроса о бездомных. Точно подсчитать их количество в крупном мегаполисе и дорого, и затруднительно. Тем не менее располагать численной оценкой этой группы населения необходимо с целью предоставления социальных услуг, обоснования права на получение части доходов штата и федеральных доходов и соответствующего представительства в Конгрессе. Одним из важных статистических методов является выборочное исследование – процесс сбора данных по какой-то небольшой области, например нескольких районов, где проводилась перепись населения, чтобы на их основе сделать умозаключение о количестве бездомных в городе в целом. Такой подход требует значительно меньших ресурсов, чем попытка сосчитать всех бездомных; к тому же при правильном проведении выборочного исследования можно получить очень близкий к точному результат.

Опрос общественного мнения – еще одна форма статистической выборки. Скажем, исследовательская организация опрашивает членов среднестатистических семей, чтобы выяснить их точку зрения на ту или иную проблему или их мнение о том или ином политическом деятеле. Сделать это, естественно, гораздо проще, дешевле и быстрее, чем обойти все домохозяйства в соответствующем штате или стране в целом. 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.

Именно таким способом нам удалось выяснить, как часто, с кем и как американцы занимаются сексом. В середине 1990-х годов Национальный центр изучения общественного мнения при Чикагском университете провел масштабное исследование сексуального поведения населения страны. Результаты основывались на детальных опросах крупной репрезентативной выборки взрослых американцев. Если вы продолжите чтение этой книги, то в главе 10 узнаете подробности. В каких еще книгах, посвященных статистике, вы могли бы почерпнуть подобные сведения?

Оценивание риска и событий, имеющих вероятностный характер

Казино никогда не бывают внакладе в долгосрочной перспективе. Это не означает, что они зарабатывают деньги в любой момент, но в конечном итоге остаются прибыльными, как бы ни складывалась каждая отдельно взятая игра. Весь игорный бизнес построен на азартных играх, поэтому исход каждой из них непредсказуем. В то же время базовые вероятности наступления соответствующих событий – выпадения двадцати одного очка в блек-джек или зеро при игре в рулетку – известны. И когда эти базовые вероятности выступают в пользу казино (а это происходит всегда), можно не сомневаться, что по мере увеличения количества ставок вероятность того, что истинным победителем окажется игорное заведение, повышается, несмотря на мелкие «досадные недоразумения», случающиеся по ходу дела.

Данный феномен характерен не только для казино, но и для многих других сфер нашей жизни. Компаниям постоянно приходится оценивать риски, связанные со всевозможными неблагоприятными факторами. Полностью исключить такие

риски невозможно – точно так же как казино не может гарантировать, что, сделав ставку, вы не сорвете крупный куш, доставив тем самым владельцам заведения немалое огорчение. Однако любой бизнес, сталкивающийся с неопределенностью, может управлять рисками, организовав соответствующие процессы таким образом, чтобы снизить вероятность того или иного неблагоприятного исхода (начиная со стихийного бедствия и заканчивая выпуском бракованного изделия) до приемлемого уровня. Компании на Уолл-стрит зачастую пытаются оценивать риски, связанные с их портфелями при разных сценариях, причем каждому из этих сценариев в зависимости от вероятности его реализации присваивается определенный вес. Финансовый кризис 2008 года отчасти спровоцировали события на рынке, наступление которых считалось крайне маловероятным (например, как если бы все игроки в казино за один вечер оказались в крупном выигрыше). Далее в этой книге я попытаюсь доказать, что модели, которыми руководствовались компании на Уолл-стрит, были изначально ущербными, а данные, использовавшиеся для оценивания ключевых рисков, – слишком ограниченными, однако сейчас я лишь хочу сказать, что в основу любой модели, имеющей дело с рисками, должны быть положены вероятности.

Когда отдельные люди и фирмы не в состоянии полностью устранить неприемлемые для них риски, они пытаются обезопасить себя другими способами. Вся страховая индустрия построена на требовании клиентов защитить их от того или иного негативного события, такого как автомобильная авария, пожар и т. п. Страховая отрасль зарабатывает деньги отнюдь не на устранении подобных случаев: ДТП происходят каждый день, собственно, как и пожары. (Бывает даже так, что автомобиль, врезавшись в дом, становится причиной пожара.) Она процветает за счет взносов владельцев страховых полисов, которых оказывается более чем достаточно, чтобы покрыть ожидаемые страховые выплаты в случае автомобильной аварии или пожара в доме. (Страховая компания может также попытаться снизить ожидаемые страховые выплаты путем поощрения методов безопасного вождения, установки детекторов дыма в каждой спальне, ограждений вокруг водоемов и т. п.)

В определенных случаях концепцию вероятности можно даже использовать для поимки мошенников. Фирма Caveon Test Security специализируется на так называемой экспертизе данных, позволяющей выявить некие закономерности, которые предполагают обман[5 - Trip Gabriel, Cheats Find an Adversary in Technology, New York Times, December 28, 2010.]. Например, эта компания (между прочим, основанная бывшим разработчиком тестов SAT[7 - Scholastic Aptitude Test – стандартизированный тест для поступающих в американские

высшие учебные заведения. Прим. ред.]) обратит внимание общественности на результаты экзаменов в том или ином учебном заведении или каком-либо другом месте их проведения, если обнаруженное количество идентичных неправильных ответов окажется крайне маловероятным (обычно речь идет о картине, которая складывается реже чем один раз на миллион). При этом она руководствуется следующей математической логикой: когда большая группа учащихся правильно отвечает на какой-то вопрос, из этого нельзя сделать однозначный вывод. Здесь возможны два варианта: либо они дружно списали правильный ответ у кого-то из своих товарищей, либо все как один очень умные ребята. Но когда большая группа учащихся отвечает на какой-то вопрос неправильно, это настораживает: все не могут ответить одинаково неправильно – по крайней мере вероятность такого сценария чрезвычайно мала. Это говорит о том, что они списали неправильный ответ у кого-то из одноклассников. Кроме того, Caveon Test Security выявляет экзамены, в ходе которых экзаменуемые отвечают на сложные вопросы значительно лучше, чем на простые (в таком случае предполагается, что ответы им были известны заранее), или количество исправлений неправильного ответа на правильный существенно превышает количество исправлений правильного ответа на неправильный (в таком случае предполагается, что после экзамена преподаватель или экзаменатор подменил листы с ответами).

Разумеется, нетрудно заметить ограничения, присущие использованию вероятностей. Достаточно большая группа экзаменуемых может абсолютно случайно дать одинаково неправильные ответы на какой-то вопрос; к тому же чем больше учебных заведений будет проверяться, тем выше вероятность натолкнуться на подобную картину. Однако никакая статистическая аномалия не опровергает принципиальную правильность предлагаемого подхода.
В 2008 году Делма Кинни, пятидесятилетний житель города Атланта, выиграл в мгновенную лотерею миллион долларов, а затем, в 2011-м, еще миллион[6 - Eyder Peralta, Atlanta Man Wins Lottery for Second Time in Three Years, NPR News (блог), November 29, 2011.]. Вероятность такого совпадения равна примерно один к 25 триллионам. Естественно, оснований арестовывать г-на Кинни за мошенничество, опираясь исключительно на аналогичные математические выкладки, нет (правда, не мешало бы проверить, не работает ли кто-то из его родственников в лотерейной комиссии штата). Вероятность – лишь один из инструментов в арсенале статистики, и этот инструмент требует умелого обращения.

Выявление важных зависимостей (работа статистика-детектива)

Действительно ли курение вызывает рак? У нас есть ответ на этот вопрос, однако процесс его получения был не так прост, как может показаться на первый взгляд. Научный метод диктует, что при проверке той или иной гипотезы необходимо провести управляемый эксперимент, в ходе которого именно интересующая нас переменная (например, курение) должна определять разницу между экспериментальной и контрольной группами. Если между двумя этими группами в чем-то (в нашем случае – в частоте возникновения рака легких) прослеживается заметная разница, то можно с уверенностью заключить, что к такому результату привела именно искомая переменная. Однако мы не имеем права ставить над людьми подобные эксперименты. Если, согласно нашей рабочей гипотезе, курение является причиной раковых заболеваний, то было бы неэтично, скажем, разделить недавних выпускников колледжа на две группы, курящих и некурящих, и спустя двадцать лет со дня окончания колледжа, когда они соберутся отметить эту круглую дату, выяснять, кто из них заболел раком легких, а кто – нет. (Управляемые эксперименты над людьми оправданны, если нужно проверить, поможет ли новое лекарство или метод лечения улучшить состояние их здоровья. Но когда речь идет о вероятности летального исхода и нам это хорошо известно, мы не имеем права подвергать людей опасности лишь ради того, чтобы подтвердить или опровергнуть свое предположение.)[8 - Разумеется, я заведомо упрощаю здесь многогранные и чрезвычайно сложные проблемы, которые ставит перед нами медицинская этика.]

Итак, нам не стоит проводить весьма сомнительный в этическом плане эксперимент, чтобы изучить последствия курения. А не проще ли вместо всей этой заумной методологии взять и сравнить во время встречи по случаю двадцатилетнего юбилея со дня окончания колледжа процент заболевания раком у бывших выпускников – курильщиков и некурильщиков?

Не проще! Курильщики и некурильщики, скорее всего, будут отличаться не только своим отношением к курению. Например, не исключено, что у курильщиков выработался ряд специфических привычек, таких как тяга к алкоголю или склонность к перееданию, что тоже негативно сказывается на их здоровье. Поэтому мы не можем быть твердо убеждены, что их нездоровый вид – следствие именно курения, а не каких-либо других пагубных пристрастий. Кроме того, у нас возникла бы серьезная проблема с данными, на которых основывается наш анализ. Курильщики, действительно заболевшие раком

(не говоря уже о тех, кто к тому времени от него умер), вряд ли придут на празднование юбилея. В результате на точности любого анализа состояния здоровья тех, кто пришел (касается ли этот анализ вреда курения или чего-либо другого), существенно скажется то обстоятельство, что в этом праздновании, скорее всего, примут участие лишь те, кто не испытывает особых проблем со здоровьем. Чем больше лет пройдет с момента окончания учебы в колледже (скажем, будет отмечаться сорокалетний или пятидесятилетний юбилей), тем меньшей будет точность анализа.

Мы не можем относиться к людям как к подопытным кроликам. В итоге статистика оказывается сродни профессии детектива. Исходные данные могут подсказать нам модели, которые в конечном счете способны привести к правильным выводам. Вы наверняка смотрели увлекательные полицейские сериалы наподобие CSI: New York, где очень симпатичные детективы и эксперты-криминалисты скрупулезно исследуют всевозможные «мелочи»: ДНК из остатков слюны на сигаретном окурке, отпечатки зубов на яблоке, кусочек волокна из автомобильного коврика, – а затем используют полученные улики для поимки преступника. «Изюминка» сериала заключается в том, что поначалу эксперты не располагают традиционными вещественными доказательствами (например видеозаписью камер наружного наблюдения или живым свидетелем преступления), позволяющими им изобличить «плохого парня», поэтому им приходится прибегать к научным методам и логическим умозаключениям. Статистика, по сути, идет тем же путем. Исходные данные дают нам некое хаотическое нагромождение подсказок и намеков – так сказать, сцену преступления. А статистический анализ их упорядочивает и систематизирует таким образом, чтобы на их основе можно было сделать логический вывод.

После прочтения главы 11 вы сможете по достоинству оценить телевизионное шоу, которое я планирую предложить какому-либо из телеканалов: CSI: Regression Analysis («CSI: регрессионный анализ»). Это шоу лишь немного отличалось бы от множества других остросюжетных полицейских сериалов. Регрессионный анализ – инструмент, позволяющий исследователям вычленить взаимосвязь между двумя переменными, такими как курение и раковые заболевания, удерживая при этом постоянным (или «учитывая») влияние других важных переменных, таких как режим питания, физические упражнения, вес и т. п. Когда вы читаете в газете о том, что ежедневное употребление в пищу хлеба из отрубей снижает риск заболевания раком толстой кишки, вы не должны думать, что группу несчастных испытуемых насильно кормили хлебом из отрубей в подвале какой-то федеральной лаборатории, в то время как контрольная группа, находившаяся в соседнем здании, с удовольствием

уплетала яичницу с беконом. Вовсе нет! Исследователи собирают подробные сведения о тысячах людей (в том числе как часто они едят хлеб из отрубей), а затем используют регрессионный анализ, чтобы сделать две важные вещи: во-первых, выразить в количественной форме связь между употреблением в пищу хлеба из отрубей и снижением вероятности заболевания раком толстой кишки (например, гипотетический вывод о том, что у тех, кто ежедневно ест хлеб из отрубей, рак толстой кишки встречается на 9 % реже, с учетом других факторов, которые могут вызывать это заболевание); и во-вторых, вычислить вероятность того, что связь между ежедневным поеданием хлеба из отрубей и снижением заболеваемости раком толстой кишки, наблюдаемая в этом исследовании, является простым совпадением – случайностью в данных именно для этой выборки людей, – а не устойчивой закономерностью: связью между режимом питания и состоянием здоровья человека.

Разумеется, в телешоу CSI: Regression Analysis будут участвовать профессиональные актеры, которые выглядят на экране гораздо лучше реальных ученых, исследующих такие данные. Этим актерам и актрисам (многие из которых, между прочим, несмотря на молодой возраст, будут иметь ученые степени) предстоит изучить огромные массивы данных и использовать новейшие статистические инструменты для ответа на важные социальные вопросы (например, каковы самые эффективные методы борьбы с преступностью и насилием и какие социальные типы чаще всего становятся террористами). Далее в этой книге мы обсудим концепцию «статистически значимого» вывода, то есть когда в результате анализа выявляется связь между двумя переменными, которая не является случайной. Ученые рассматривают такой статистический вывод как «явную улику». Я предполагаю, что в телешоу CSI: Regression Analysis героиней будет девушка-исследователь, работающая поздно вечером в компьютерной лаборатории, поскольку днем она интенсивно тренируется в составе олимпийской сборной США по пляжному волейболу. Получив распечатку со статистическим анализом, девушка видит именно то, на что и рассчитывала: ярко выраженную статистически значимую связь между некой, по ее мнению, важной переменной и развитием аутизма. Естественно, она тут же спешит поделиться своим открытием с коллегами!

Девушка берет распечатку и бежит по коридору; скорость ее передвижения замедляют лишь высокие каблуки и очень узкая короткая черная юбка. Моя героиня вбегает в комнату к коллеге, симпатичному загорелому парню (и когда он только успел так загореть, ежедневно просиживая по четырнадцать часов за компьютером?), и демонстрирует ему распечатку. Он задумчиво теребит пальцами свою аккуратно подстриженную эспаньолку, вынимает

из ящика письменного стола пистолет калибра 9 мм марки Glock и сует его в боковой карман своего костюма от Hugo Boss за 5000 долларов (и откуда, интересно, взялся у него такой костюмчик, учитывая, что размер его годовой заработной платы составляет примерно 38 000 долларов?). Затем они быстрым шагом направляются в кабинет к боссу, прожженному ветерану сыска, которому уже удалось наладить отношения со своей женой и вылечиться от алкоголизма…

Ладно, вам вовсе не обязательно смотреть телевизор, чтобы оценить важность подобных статистических исследований, практически все важнейшие социальные проблемы решаются с помощью систематического анализа огромных массивов данных. (Во многих случаях их сбор – весьма дорогостоящий и трудоемкий – играет решающую роль в этом процессе, что я постараюсь продемонстрировать в главе 7.) Возможно, я несколько приукрасил своих героев в CSI: Regression Analysis, но это отнюдь не снижает актуальности решаемых ими вопросов. Существует научная литература о террористах и террористах-смертниках – теме, которую было бы очень трудно изучать на живых примерах, используя добровольцев в качестве подопытных кроликов. Одну из таких книг, What Makes a Terrorist («Как человек становится террористом»), написал мой преподаватель статистики в магистратуре. Материал книги основан на данных, собранных по результатам террористических актов в разных странах. Вот один из важных выводов, сделанных ее автором, экономистом Принстонского университета Аланом Крюгером: «Террористы отнюдь не всегда оказываются выходцами из беднейших слоев населения или малообразованными людьми, наоборот, обычно они принадлежат к среднему классу; уровень их образования также достаточно высок»[7 - Alan B. Krueger, What Makes a Terrorist: Economics and the Roots of Terrorism (Princeton: Princeton University Press, 2008).].

В чем тут дело? В этой ситуации проявляется одно из ограничений регрессионного анализа. С помощью статистического анализа мы можем изолировать сильную связь между двумя переменными, но далеко не всегда можем объяснить причину ее существования, а в некоторых случаях даже не знаем наверняка, носит ли она причинно-следственный характер (то есть что изменение одной переменной действительно влечет за собой изменение другой переменной). Что касается терроризма, то профессор Крюгер считает, что, поскольку террористы мотивированы определенными политическими целями, те, кто наиболее образован и богат, движимы сильным желанием изменить общество. Особенно таких людей возмущает подавление свободы – еще один фактор, связанный с терроризмом. Согласно исследованию, выполненному Крюгером, странам с высоким уровнем политических репрессий присущ более

высокий уровень террористической деятельности (при условии и неизменности прочих факторов).

Это обсуждение возвращает меня к вопросу, поставленному в названии главы: в чем суть? Точно не в том, чтобы заниматься сложными математическими выкладками или поражать друзей и коллег мудреными статистическими методами. Суть в том, чтобы узнать вещи, которые позволяют нам лучше понимать свою жизнь.

Ложь, наглая ложь и статистика

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. Любители спорта будут до бесконечности спорить по поводу «лучшего бейсболиста всех времен и народов» ввиду отсутствия четкого определения того, что именно следует считать «самым лучшим». Изощренные описательные статистики могут в той или иной степени проливать свет на этот вопрос, но они никогда не дадут на него исчерпывающего ответа. Как указывается в следующей главе, гораздо более значимые социальные вопросы пали жертвой той же фундаментальной проблемы. Что происходит с экономическим благополучием американского среднего класса? Ответ на этот вопрос зависит от того, как мы трактуем понятия «средний класс» и «экономическое благополучие».

Существуют определенные ограничения на данные, которые мы в состоянии собрать, и на виды эксперимента, который можем провести. Исследование корней терроризма, выполненное Аланом Крюгером, не могло охватить жизни тысяч молодых людей на протяжении нескольких десятилетий, чтобы проследить, кто из них стал террористом. Это физически невозможно. Не можем мы и создать две идентичные страны, отличающиеся лишь наличием в одной из них мощного репрессивного аппарата, а затем сравнить количество террористов-смертников, появившихся в каждой из них. Даже когда

крупномасштабные контролируемые эксперименты на людях проводятся, они оказываются чрезвычайно трудоемкими, сложными и дорогостоящими. Ученые выполнили одно такое исследование, чтобы выяснить, помогают ли молитвы снизить количество и тяжесть послехирургических осложнений (вы, наверное, помните, что это был один из вопросов, поднимавшихся ранее в настоящей главе), и оно обошлось в 2,4 миллиона долларов (его результаты обсуждаются в главе 13).

Министр обороны США Дональд Рамсфелд однажды сделал заявление, ставшее знаменитым: «Вы начинаете войну с армией, которая у вас на данный момент есть, а не которую вы хотели бы или можете иметь в будущем». Каким бы ни было ваше мнение о Дональде Рамсфелде (и о войне в Ираке, результаты которой он пытался объяснить), этот афоризм относится не только к армии, но и к исследованиям. Мы выполняем статистический анализ, используя доступные нам данные, методологии и ресурсы. Такой подход не похож на операции сложения или деления в столбик, когда применение правильного метода дает правильный ответ, а компьютер всегда обеспечивает более высокую точность и намного реже ошибается, чем человек. Статистический анализ гораздо больше напоминает работу следователя (что может служить гарантией высокого коммерческого потенциала телешоу CSI: Regression Analysis). А умные и честные люди всегда будут спорить относительно того, о чем именно говорят нам те или иные данные.

Но кто возьмется утверждать, что каждый, кто использует статистику, непременно умный и честный человек? Эта книга задумывалась как дань уважения классическому труду Дарелла Хаффа How to Lie with Statistics («Как лгать при помощи статистики»), который был впервые опубликован в 1954 году и разошелся тиражом свыше миллиона экземпляров. Да, реальность такова, что с помощью статистики можно вводить людей в заблуждение или совершать непреднамеренные ошибки. В любом случае математическая точность, сопутствующая статистическому анализу, может служить ширмой для откровенного бреда, которому пытаются придать некое наукообразие. В своей книге я расскажу о наиболее характерных статистических ошибках и искажении фактов, чтобы вы могли распознать подобные случаи манипулирования статистикой (надеюсь, вы не станете сами пытаться ею манипулировать).

Итак, возвращаясь к названию этой главы, зачем нам изучать статистику?

Это необходимо для того чтобы:

• обобщать огромные массивы данных;

• принимать более эффективные решения;

• находить ответы на важные социальные вопросы;

• распознавать ситуации, которые позволяют уточнить метод решения тех или иных задач, от продажи подгузников до поимки преступников;

• выслеживать мошенников и находить доказательства, помогающие изобличать преступников;

• оценивать эффективность полиции, тех или иных социальных программ, лекарственных препаратов, медицинских процедур и прочих инноваций;

• а также «вычислять» негодяев, которые используют мощные статистические инструменты для достижения своих неблаговидных целей.

Если вам удается делать все это и при этом превосходно выглядеть в костюме от Hugo Boss или черной мини-юбке, то вам ничто не мешает стать очередной звездой телешоу CSI: Regression Analysis.

2. Описательная статистика

Кто же все-таки лучший бейсболист всех времен и народов?

Давайте подумаем над двумя на первый взгляд не связанными между собой вопросами:

1. Что происходит с экономическим благополучием американского среднего класса?

2. Кого же все-таки считать лучшим бейсболистом всех времен и народов?

Первый вопрос крайне важен и, как правило, ложится в основу президентских кампаний и других социальных движений. Средний класс, если можно так выразиться, – это сердце Америки, поэтому его экономическое благополучие является индикатором общего экономического благосостояния страны. Второй вопрос тривиален (в буквальном смысле этого слова), однако любители бейсбола готовы до бесконечности спорить по этому поводу. Объединяет оба вопроса то, что они позволяют проиллюстрировать сильные и слабые стороны описательной статистики, которая представляет собой числа и вычисления, используемые для обобщения исходных данных.

Если я захочу продемонстрировать вам, что Дерек Джетер является великим игроком в бейсбол, то смогу описать каждый удачно посланный им мяч в каждом матче Высшей бейсбольной лиги, в котором он принимал участие. Это будут исходные данные, и, чтобы упорядочить их, потребуется какое-то время (с учетом того, что Джетер провел семнадцать сезонов в составе New York Yankees и за это время совершил 9868 удачных бросков).

Или я просто могу вам сказать, что к концу сезона 2011 года средний результат Дерека Джетера за всю его карьеру составлял 0,313. Это описательная, или «сводная» статистика.

Однако такой средний показатель – явное упрощение достижений Джетера за семнадцать сезонов игры в Высшей бейсбольной лиге. Да, он весьма элегантен в своей простоте, но не отражает всех нюансов спортивной карьеры Джетера. В распоряжении экспертов по бейсболу есть целый арсенал описательных статистик, которые они считают более ценными, чем данный показатель. Я позвонил Стиву Мойеру, президенту Baseball Info Solutions (фирмы, которая предоставила большой объем исходных данных для спортивной драмы Moneyball[9 - В российском прокате этот фильм вышел под названием «Человек, который изменил все». Фильм снят по книге Майкла М. Льюиса, изданной в 2003 году, о бейсбольной команде «Окленд Атлетикс» и ее генеральном менеджере Билли Бине. Его цель – создать конкурентоспособную бейсбольную

команду, несмотря на отсутствие больших финансовых возможностей. Главную роль исполняет Брэд Питт. Прим. перев.]), чтобы задать ему два вопроса: 1) каковы самые важные статистические показатели для оценки бейсбольного таланта и 2) кто, по его мнению, величайший бейсболист всех времен и народов? Я познакомлю вас с ответами Стива, когда мы получим больше контекста.

А пока вернемся к менее тривиальному предмету – экономическому благополучию среднего класса. В идеале было бы желательно найти экономический эквивалент среднего показателя (или что-нибудь получше). Нас устроил бы какой-либо простой, но точный показатель того, как за последние годы изменилось экономическое благосостояние типичного американского рабочего. Стали ли люди, которых мы определяем как средний класс, богаче, беднее или в их финансовом положении ничего не изменилось? Подходящий вариант ответа на этот вопрос – который ни в коем случае нельзя рассматривать как «правильный» – рассчитать изменение дохода на душу населения в Соединенных Штатах на протяжении жизни одного поколения (примерно тридцать лет). Доход на душу населения вычисляется путем деления совокупного дохода на численность населения. Согласно этому показателю, средний доход в США повысился с 7787 долларов в 1980 году до 26 487 долларов в 2010-м (последний год, за который правительство располагает соответствующими данными)[8 - U.S. Census Bureau, Current Population Survey, Annual Social and Economic Supplements, http://www.census.gov/en.html (http://www.census.gov/en.html).]. Вот так-то! Принимайте поздравления.

Есть, правда, одна проблема. Мой быстрый подсчет технически правилен и совершенно неверен с точки зрения ответа на интересующий нас вопрос. Начнем хотя бы с того, что в приведенных выше цифрах отсутствует поправка на инфляцию. (Величина дохода на душу населения 7787 долларов в 1980 году составляет примерно 19 600 долларов в 2010-м.) Такой корректив внести относительно просто. Более серьезная проблема заключается в том, что средний доход в Америке не равняется доходу среднего американца. Попытаемся расшифровать это утверждение.

Чтобы вычислить величину дохода на душу населения, мы берем весь национальный доход и делим его на численность населения. Однако полученный таким образом показатель абсолютно ничего не говорит нам о том, кто и сколько при этом зарабатывает – хоть в 1980 году, хоть в 2010-м. Как сказали бы участники акции Occupy Wall Street, взрывообразный рост доходов 1 % самых

богатых людей Америки способен существенно повысить значение дохода на душу населения, ничего при этом не изменив в карманах остальных 99 % американцев. Иными словами, средний доход может повышаться без помощи среднего класса.

Как и в случае бейсбольной статистики, мне хотелось узнать мнение авторитетного эксперта о том, как нам следовало бы измерять экономическое благосостояние американского среднего класса. Я спросил у двух известных специалистов по трудовым отношениям, в том числе у ведущего экономического советника президента Обамы, какие описательные статистики они использовали бы для оценки экономического благополучия типичного американца. Вы узнаете их ответы после того, как ознакомитесь с кратким обзором описательных статистик и лучше уясните их смысл.

Будь то бейсбол, доход или что-то еще, самая фундаментальная задача при работе с данными – обобщить их огромные массивы. Численность населения Соединенных Штатов составляет примерно 330 миллионов человек. Электронная таблица, в которой указывались бы фамилия и история доходов каждого американца, содержала бы всю информацию, которая могла потребоваться для оценки экономического благосостояния страны, однако эта информация была бы настолько громоздкой, что извлечь из нее хоть какую-то пользу было бы практически невозможно. Ирония судьбы заключается в том, что 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, точно так же как пытаемся оценить разноплановую программу выступления гимнаста на Олимпийских играх одним числом: 9,8 балла.

Плюс состоит в том, что описательные статистики дают нам некое обобщенное и осмысленное представление исходного явления. О чем, собственно, и идет речь в этой главе. Минус же в том, что любое упрощение порождает манипулирование. Описательные статистики можно сравнить с анкетами на сайтах знакомств: технически они точны и тем не менее сильно вводят в заблуждение.

Допустим, сидя на работе, вы от нечего делать бродите по интернету и наталкиваетесь на онлайн-дневник известной светской львицы Ким Кардашьян, в котором она рассказывает о своей «долгой» (целых семьдесят два дня!) супружеской жизни с профессиональным баскетболистом Крисом Хэмфри.

И вот в тот самый момент, когда вы добрались до описания седьмого дня их супружеской жизни, в комнату неожиданно заходит ваш босс с двумя огромными папками данных. В одной из папок собрана информация о гарантийных претензиях по каждому из 57 334 лазерных принтеров, которые ваша фирма продала в прошлом году. (По каждому из проданных лазерных принтеров перечисляются все проблемы с качеством, зафиксированные в течение гарантийного периода.) В другой содержится такая же информация по каждому из 994 773 лазерных принтеров, которые продал за тот же период ваш главный конкурент. Босс хотел бы сравнить качество принтеров вашей компании с качеством принтеров конкурента.

К счастью, на компьютере, на котором вы почитывали дневник Кардашьян, установлен пакет основных статистических методов, но с чего в данном случае начать? Ваша интуиция, по-видимому, подсказывает вам правильное решение: первой описательной задачей зачастую становится поиск некоего показателя «середины» совокупности данных, или того, что статистики называют «центральной тенденцией». Что является типичным показателем качества для ваших принтеров по сравнению с принтерами конкурента? Обычно самым фундаментальным показателем «середины» какого-либо распределения считается среднее значение. В данном случае нам нужно определить среднее количество проблем с качеством на каждый проданный принтер для вашей фирмы и фирмы вашего конкурента. Вы могли бы просто подсчитать общее число выявленных проблем с качеством для всех принтеров в течение гарантийного периода, а затем разделить его на общее количество проданных принтеров. (Учтите, что в течение гарантийного периода в одном и том же принтере может возникнуть несколько проблем с качеством.) Эту операцию можно проделать для каждой компании, создав важную описательную статистику: среднее количество проблем с качеством на каждый проданный принтер.

Предположим, выяснилось, что среднее количество проблем с качеством в течение гарантийного периода у принтеров вашего конкурента равно 2,8 на каждый проданный принтер, тогда как соответствующий показатель для вашей фирмы составляет 9,1. Как видите, вывести среднее значение совсем не сложно. Вы просто использовали информацию для миллиона принтеров, проданных двумя разными компаниями, и извлекли из нее суть интересующей вас проблемы: ваши принтеры ломаются слишком часто. Похоже, самое время отправить боссу по электронной почте краткое уведомление с численным подтверждением столь тревожного факта, а затем вернуться к более увлекательному занятию: чтению дневника Ким Кардашьян.

А может, не стоит торопиться? Я ведь не зря выразился довольно туманно, упомянув о какой-то там «середине» распределения. В этом отношении у среднего значения есть определенные проблемы, а именно: оно подвержено существенным искажениям со стороны «отщепенцев», то есть значений, резко отклоняющихся от центра. Чтобы вам было легче уяснить эту концепцию, вообразите десяток парней, сидящих у стойки бара какого-нибудь питейного заведения в Сиэтле, рассчитанного на представителей среднего класса. Каждый из парней зарабатывает по 35 000 долларов в год; стало быть, средний годовой доход этой группы составляет 35 000 долларов. Внезапно в заведение входит Билл Гейтс с говорящим попугаем на плече (вообще-то в данном примере говорящий попугай не играет никакой особой роли; это не более чем деталь, призванная несколько оживить повествование и придать ему определенный колорит) и усаживается на одиннадцатый стул за стойкой бара; при этом средний годовой доход его завсегдатаев резко повышается до 91 миллиона долларов. Очевидно, что первые десять посетителей бара могут лишь мечтать о таком уровне годового дохода (хотя все они, наверное, надеются, что Билл Гейтс расщедрится и угостит их стаканчиком-другим). Если бы я написал, что средний годовой доход посетителей заведения составляет 91 миллион долларов, то данный вывод был бы статистически правильным, однако не имел бы ничего общего с реальным положением вещей. Этот бар отнюдь не относится к числу заведений, где коротают свободное время мультимиллионеры, – здесь обычно отдыхают молодые люди с относительно невысоким уровнем годовых доходов. Просто сегодня им повезло оказаться в компании с Биллом Гейтсом и его говорящим попугаем. Именно высокая чувствительность среднего значения к значениям, резко отклоняющимся от центра, не позволяет нам измерять экономическое благополучие среднего класса с помощью такого показателя, как величина дохода на душу населения. Поскольку в последнее время наблюдается резкий рост доходов в верхней части распределения – глав компаний, управляющих хедж-фондами и выдающихся спортсменов, таких как Дерек Джетер, – величина среднего дохода в США может быть сильно искажена, как в вышеупомянутом баре, где несколько парней с относительно скромными доходами случайно оказались в компании Билла Гейтса.

По этой причине нам приходится пользоваться еще одной статистикой, которая также является отражением «середины» распределения, однако делает это несколько иначе. Речь идет о так называемой медиане. Медиана – это точка, которая делит распределение пополам таким образом, что одна половина наблюдений располагается выше медианы, а другая половина – ниже.

(При наличии четного количества наблюдений медиана представляет собой среднюю точку между двумя средними наблюдениями.) Если мы вернемся к примеру с баром, то срединный (медианный) годовой доход для десяти человек, сидевших поначалу за стойкой, равняется 35 000 долларов. Когда в заведении появился – и уселся на одиннадцатый стул – Билл Гейтс с говорящим попугаем, срединный годовой доход для одиннадцати человек по-прежнему составлял 35 000 долларов. Если представить, что посетители бара расселись за его стойкой в порядке возрастания их доходов, то доход посетителя, сидящего на шестом стуле, будет срединным для данной группы людей. Даже если бы в заведение зашел Уоррен Баффет и уселся рядом с Биллом Гейтсом на двенадцатый стул, медиана все равно осталась бы неизменной[10 - После того как в баре оказалось бы двенадцать посетителей, медианой была бы средняя точка между доходом посетителя, сидящего на шестом стуле, и доходом посетителя, сидящего на седьмом стуле. Поскольку доход того и другого составляет 35 000 долларов, медиана равняется 35 000 долларов. Если бы доход одного из них равнялся 35 000, а доход другого – 36 000, то медиана для этой группы в целом равнялась бы 35 500 долларов.].

В случае распределений без «отщепенцев» срединное (медиана) и среднее значения совпадают. Выше говорилось о гипотетической сводке данных, отражающих качество принтеров конкурирующей фирмы. В частности, я представил эти данные в виде так называемого частотного распределения (гистограммы). Число проблем с качеством на один принтер представлено на горизонтальной оси (внизу); высота каждого вертикального столбца соответствует проценту проданных принтеров, у которых наблюдалось такое число проблем с качеством. Например, у 36 % принтеров конкурента в течение гарантийного периода возникало по две проблемы с качеством. Поскольку это распределение включает все возможные случаи проблем с качеством (в том числе и их отсутствие), сумма всех долей (процентов) должна равняться 1 (или 100 %).

Поскольку такое распределение почти симметрично, среднее и срединное значения довольно близки друг к другу. Распределение слегка скошено вправо, что объясняется малым количеством принтеров, имеющих множественные

дефекты. Эти «отщепенцы» слегка смещают среднее значение вправо, однако на медиану это не влияет. Допустим, что перед тем как составить для босса отчет о качестве принтеров, вы принимаете решение вычислить медианы, то есть число проблем с качеством для принтеров, проданных вашей и конкурирующей компанией. Нажав всего несколько клавиш, вы получите результат. Медиана проблем с качеством для принтеров конкурента равняется 2; а для принтеров вашей фирмы – 1.

Что из этого следует? Оказывается, медиана проблем с качеством на каждый принтер вашей фирмы фактически меньше, чем у вашего конкурента. Поскольку супружеская жизнь Ким Кардашьян становится однообразной, а полученный результат вас заинтриговал, вы распечатываете распределение частот проблем с качеством у принтеров, проданных вашей компанией.

Из приведенных выше гистограмм становится ясно, что для вашей компании нехарактерно равномерное распределение проблем с качеством. Напротив, у вас налицо проблема «лимона»[11 - «Лимонами» на американском сленге называют устройства с дефектами, которые проявляются уже после покупки. Прим. ред.]: у малого числа ваших принтеров наблюдается большое количество дефектов. Эти «отщепенцы» способствуют наращиванию среднего значения, тогда как медиана остается неизменной. Более важным с производственной точки зрения является то обстоятельство, что вам нет необходимости переоснащать весь производственный процесс; достаточно лишь определить, какое из предприятий компании выпускает некачественную продукцию, и исправить ситуацию[12 - Вот что удалось выяснить в ходе дальнейшего исследования проблемы. Оказалось, что почти все бракованные принтеры производились на заводе в Кентукки, где рабочие разобрали часть сборочного конвейера, чтобы создать подпольное предприятие по изготовлению виски. Постоянно пьяные рабочие и частично разобранный сборочный конвейер стали причиной резкого ухудшения качества выпускаемых заводом принтеров.].

Вычисление среднего и медианы 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). Между тем у медианы имеются

весьма полезные «родственники». Как указывалось выше, медиана делит любое распределение пополам. Затем его можно разбить на четверти, или, как их еще называют, квартили. Первый квартиль состоит из нижних 25 % наблюдений; второй из следующих 25 % наблюдений и т. д. Еще один вариант – разделить распределение на децили, каждый из которых заключает в себе 10 % наблюдений. (Если ваш доход находится в верхнем дециле американского распределения доходов, то это означает, что вы зарабатываете больше, чем 90 % ваших коллег-рабочих.) Можно пойти еще дальше и разбить распределение на сотые доли, или процентили. Каждый процентиль представляет 1 % распределения; таким образом, первый процентиль представляет нижний 1 % данного распределения, а 99-й – его верхний 1 %.

Преимущество описательных статистик такого рода заключается в том, что они указывают, где именно располагается то или иное конкретное наблюдение по сравнению с остальными. Например, информация, что ваш ребенок по результатам теста на понимание прочитанного материала получил третий процентиль, должна сказать вам о том, что вы уделяете недостаточно внимания совместному обсуждению книг, прочитанных вашим ребенком. Вам вовсе не обязательно знать подробности самого теста или точное количество вопросов, на которые ваш ребенок ответил правильно. Однако его попадание в определенный процентиль в любом случае говорит о том, насколько успешно ваш ребенок сдал этот тест по сравнению с другими его участниками. Если тест был сравнительно легким, то большинство его участников правильно ответят на подавляющее число вопросов, при этом количество правильных ответов у вашего ребенка все равно будет меньшим, чем у большинства других участников тестирования. Если же тест был очень трудным, то у всех его участников окажется малое число правильных ответов, однако и в этом случае «рейтинг» вашего ребенка будет несколько ниже, чем у остальных.

Конец ознакомительного фрагмента.

Безопасно оплатить книгу можно банковской картой Visa, MasterCard, Maestro, со счета мобильного телефона, с платежного терминала, в салоне МТС или Связной, через PayPal, WebMoney, Яндекс.Деньги, QIWI Кошелек, бонусными картами или другим удобным Вам способом.

notes

Сноски

1

Хоумран – удар в бейсболе, при котором мяч перелетает через все игровое поле; дает право совершить перебежку по всем базам и принести своей команде очко. Прим. перев.

2

Куортербек – распасовщик, играющий помощник тренера в американском футболе. Прим. перев.

3

Тачдаун – в американском футболе: пересечение мячом или игроком с мячом линии зачетного поля соперника. Прим. перев.

4

Коэффициент Джини иногда умножают на 100, чтобы он выражался целым числом. В таком случае для Соединенных Штатов он равнялся бы 45.

5

Netflix – американская компания, поставщик фильмов и сериалов на основе потокового мультимедиа. Прим. перев.

6

Исторически так сложилось, что слово «данные» (data) используется во множественном числе (например, «эти данные являются весьма обнадеживающими»). Это слово можно употреблять и в единственном числе: «данное» (datum); в этом случае речь идет о каком-то отдельно взятом элементе данных (например, ответ одного человека на какой-то один вопрос анкеты, используемой при опросе общественного мнения). Употребление слова «данные» во множественном числе сигнализирует каждому, кто занимается серьезными исследованиями, о том, что вы знаете толк в статистике. С учетом сказанного многие специалисты по грамматике, а также многие издания, такие как The New York Times, в настоящее время согласны с тем, что слово «данные» может означать как единственное, так и множественное число, как свидетельствует приведенная мной цитата из The New York Times.

7

Scholastic Aptitude Test – стандартизированный тест для поступающих в американские высшие учебные заведения. Прим. ред.

8

Разумеется, я заведомо упрощаю здесь многогранные и чрезвычайно сложные проблемы, которые ставит перед нами медицинская этика.

9

В российском прокате этот фильм вышел под названием «Человек, который изменил все». Фильм снят по книге Майкла М. Льюиса, изданной в 2003 году, о бейсбольной команде «Окленд Атлетикс» и ее генеральном менеджере Билли Бине. Его цель – создать конкурентоспособную бейсбольную команду, несмотря на отсутствие больших финансовых возможностей. Главную роль исполняет Брэд Питт. Прим. перев.

10

После того как в баре оказалось бы двенадцать посетителей, медианой была бы средняя точка между доходом посетителя, сидящего на шестом стуле, и доходом посетителя, сидящего на седьмом стуле. Поскольку доход того и другого составляет 35 000 долларов, медиана равняется 35 000 долларов. Если бы доход одного из них равнялся 35 000, а доход другого – 36 000, то медиана для этой группы в целом равнялась бы 35 500 долларов.

11

«Лимонами» на американском сленге называют устройства с дефектами, которые проявляются уже после покупки. Прим. ред.

12

Вот что удалось выяснить в ходе дальнейшего исследования проблемы. Оказалось, что почти все бракованные принтеры производились на заводе в Кентукки, где рабочие разобрали часть сборочного конвейера, чтобы создать подпольное предприятие по изготовлению виски. Постоянно пьяные рабочие и частично разобранный сборочный конвейер стали причиной резкого ухудшения качества выпускаемых заводом принтеров.

Комментарии

1

Central Intelligence Agency, The World Factbook, https://www.cia.gov/library/publications/the-world-factbook/ (https://www.cia.gov/library/publications/the-world-factbook/).

2

Steve Lohr, For Today’s Graduate, Just One Word: Statistics, New York Times, August 6, 2009.

3

Steve Lohr, For Today’s Graduate, Just One Word: Statistics, New York Times, August 6, 2009.

4

Baseball-Reference.com (http://www.baseball-reference.com/), http://www.baseball-reference.com/players/ (http://www.baseball-reference.com/players/)

5

Trip Gabriel, Cheats Find an Adversary in Technology, New York Times, December 28, 2010.

6

Eyder Peralta, Atlanta Man Wins Lottery for Second Time in Three Years, NPR News (блог), November 29, 2011.

7

Alan B. Krueger, What Makes a Terrorist: Economics and the Roots of Terrorism (Princeton: Princeton University Press, 2008).

8

U.S. Census Bureau, Current Population Survey, Annual Social and Economic Supplements, http://www.census.gov/en.html (http://www.census.gov/en.html).
