individuum

Tom Chivers
and David Chivers

How to Read Numbers

A Guide to Statistics in the News
(and Knowing When to Trust Them)

Том Чиверс,
Дэвид Чиверс

Цифры врут

Как не дать статистике обмануть себя

Перевод с английского
Натальи Шаховой

Individuum
Москва, 2024

УДК 31
ББК 60.6
Ч58

Напечатано с разрешения Anna Jarota Agency

Чиверс, Том; Чиверс, Дэвид.
Ч58 Цифры врут. Как не дать статистике обмануть себя / Том Чиверс, Дэвид Чиверс [пер. с английского Н. Шаховой]. — М. : Individuum, 2024. — 224 с.

ISBN 978-5-6046877-9-6

Чтение на ночь сокращает жизнь. Видеоигры провоцируют массовые убийства. Газировка делает людей агрессивными. Что?! Нет!
Каждый день медиа пишут о сенсационных открытиях и шокирующих результатах исследований. Но не всем им можно верить: статистические погрешности, намеренные фальсификации и неочевидные огрехи никто не отменял.
Из-за пандемии COVID-19 человечество было вынуждено пройти ускоренный курс статистики: теперь мы неплохо разбираемся в графиках, кое-что слышали о нормальном распределении и знакомы с ошибкой выжившего. Но нам еще многое предстоит узнать: как работают математические модели? Чем отличаются абсолютные и относительные риски? О чем говорят рейтинги? Что такое ошибка техасского стрелка? Научный журналист Том Чиверс и преподаватель экономики в Даремском университете Дэвид Чиверс на примерах громких заголовков ковидного времени показывают, как не дать себя обмануть с помощью чисел.

УДК 31
ББК 60.6

ISBN 978-5-6046877-9-6

Оглавление

Посвящается нашим бабушке и дедушке —
Джин и Питеру Чиверсам

Введение

Цифрам неведомы чувства. Цифры не истекают кровью,
не проливают слез, не питают надежд. Им не знакомы
отвага и самопожертвование, любовь и преданность.
На пике черствости вы найдете лишь нули и единицы.
Эми Кауфман и Джей Кристофф. «Иллюминэ»*

Цифры холодны и бесчувственны. Поэтому зачастую они вызывают неприязнь, и это вполне объяснимо. Во время написания этой книги газеты ежедневно сообщали о количестве умерших от COVID-19, пандемия которого в первой половине 2020 года завладела миром. Когда в Великобритании число погибших упало с тысяч всего до сотен, показалось, что виден свет в конце туннеля.

Но ведь каждый из скончавшихся от коронавируса был индивидуальностью, каждый был уникален. Можно говорить об их числе — к августу это 41 369 человек в Великобритании или 28 646 в Испании** — или о том, сколько всего людей умрет к тому моменту, когда (если) пандемия закончится. Только сухие цифры ничего не сообщают нам об этих людях. А ведь у каждого из них своя история: кем они были, что делали, кого любили и кем были любимы. Их будут оплакивать.

Представление всех погибших одним числом — «сегодня умерло X человек» — кажется грубым и бездушным. Игнорируются печаль и горе. Устраняются индивидуальности и судьбы.

---

* Перевод С. Рюмина. — *Прим. ред.*

** По состоянию на 09.07.2024 от коронавируса в России умерло 403 321 человек. См.: https://russian-trade.com/coronavirus-russia/ — *Прим. ред.*

Но если бы мы не вели ежедневный учет смертей, не отслеживали распространение болезни, весьма вероятно, погибло бы еще больше людей. Еще больше уникальных личных историй оборвалось бы преждевременно. Просто мы бы не знали числа жертв.

В этой книге мы будем много говорить о числах: как их используют СМИ, что может пойти не так и как это может исказить реальную картину. Но по ходу дела постараемся не забывать: числа обозначают что-то конкретное. Часто — людей или что-то для людей важное.

Эта книжка в некотором роде математическая. Вы можете опасаться, что ничего не поймете, если вам кажется, что вы не в ладах с математикой. Но вы не одиноки. Похоже, чуть ли не все думают, что не разбираются в ней.

Дэвид преподает экономику в Даремском университете. Все его студенты получили высшую оценку (A) на школьном выпускном экзамене по математике, и тем не менее многие из них считают, что плохо разбираются в этом предмете. Том думает, что довольно плохо знает математику, хотя и выиграл две награды Королевского статистического общества за «статистическое совершенство в журналистике» (он любит время от времени невзначай упомянуть об этом). Дэвид тоже иногда думает, что плохо разбирается в математике, хотя и *учит математике* тех, кто уже неплохо ее освоил.

Возможно, и вы знаете математику лучше, чем вам кажется. Просто плохо считаете в уме. Когда мы думаем о тех, кто разбирается в математике, первыми в голову приходят люди вроде Кэрол Вордерман или Рэйчел Райли — ведущих телепередачи «Обратный отсчет», которые быстро считают в уме. Они-то, конечно, хорошие математики, но, если вы так не умеете, это еще не значит, что вы — плохой.

Принято думать, что в этой науке есть ответы верные и неверные. Зачастую это не так, по крайней мере в той математике,

о которой мы говорим. Возьмем, к примеру, с виду простое, но такое печальное число — количество людей, умерших от коронавируса. Как его определить? Нужно ли учитывать только тех, у кого диагноз «COVID-19» был подтвержден тестом? Или просто вычислить количество «лишних» смертей, сравнив число умерших в этом году со среднегодовым показателем за последние несколько лет? Это будут два очень разных числа, и какое из них нам подходит, зависит от вопроса, на который мы хотим ответить. Ни одно из них не является неверным, но и правильным его не назовешь.

Важно понимать, почему эти числа неоднозначны и почему то, что порой кажется очевидным, на самом деле куда сложнее. Ведь числами легко затуманить смысл и сбить с толку, и многие (в особенности политики, но не они одни) пользуются этим. Различия в трактовках влияют на нашу жизнь, на способность участвовать демократических процессах. Тут так же, как с грамотностью. Демократическому государству трудно функционировать без грамотного населения: чтобы осознанно голосовать, избиратели должны понимать политические решения властей.

Но недостаточно понимать слова — нужно еще разбираться в цифрах. Новости всё чаще принимают числовую форму: число зарегистрированных полицией преступлений то увеличивается, то уменьшается; экономика страны растет или идет на спад; публикуются всё новые данные об умерших от ковида. Чтобы ориентироваться во всем этом, необязательно быть математиком, но нужно понимать, как числа подсчитываются, для чего применяются и какие с ними бывают подвохи. Иначе мы — как отдельные индивидуумы и как общество в целом — будем принимать неверные решения.

Иногда предельно ясно, как неверное истолкование статистики приводит к плохим решениям. Так, нельзя оценить адекватность антикоронавирусных мер, не зная точного числа заболевших. В других случаях — например, далее мы рассмотрим,

вызывает ли бекон рак и повышает ли потребление газировки склонность к насилию, — опасность не так очевидна. При этом все мы, чтобы ориентироваться в мире, постоянно осознанно или неосознанно опираемся на числа. Пьем красное вино, занимаемся спортом, вкладываем средства — и всё это исходя из предположения, что преимущества (с точки зрения удовольствия, здоровья или богатства) перевешивают риски. Мы должны знать о них и оценивать их, чтобы делать разумный выбор. А представления о преимуществах и рисках мы зачастую получаем из СМИ.

Не стоит полагаться на то, что СМИ всегда дают точные числа без преувеличений и выбора эффектных ракурсов. И дело не в том, что медиа стремятся вас обмануть, — просто им нужно рассказывать об удивительных, интересных и поразительных вещах, чтобы вы покупали газеты и смотрели передачи. А еще потому, что они — и мы — жаждут историй, где у проблем есть очевидные причины и решения. Если же выбирать самые удивительные, интересные и поразительные числа, то многие из них вполне могут оказаться неверными или сбивающими с толку.

Кроме того, хотя журналисты обычно умны и (вопреки стереотипам) имеют добрые намерения, они, как правило, не очень ладят с числами. Поэтому числа, которые вы видите в СМИ, нередко неверны. Не всегда, но достаточно часто — не теряйте бдительности.

К счастью, пути искажения чисел бывают вполне предсказуемыми. Например, эффектный результат можно получить, выбрав какую-то экстремальную точку или удачное начало отсчета, а также многократно перебирая данные, пока не найдется что-то интересное. Результат можно преувеличить, если говорить не о реальном изменении, а о процентном. С помощью чисел создается видимость причинно-следственной связи там, где есть простая корреляция. Существует и масса других способов. Эта книга научит вас замечать некоторые из них.

Мы вовсе не утверждаем, что никаким цифрам из СМИ нельзя верить. Мы просто хотим научить вас разбираться, каким и когда верить можно.

Математику мы постарались свести к минимуму. Почти все, что похоже на уравнение, вынесено из основного текста в отдельные врезки. Их читать необязательно — вы и так все поймете.

Но мы не могли совсем обойтись без технических понятий, поэтому изредка в книге будут попадаться выражения типа $p = 0{,}049$ или $r = —0{,}4$; пусть они вас не пугают. Это лишь краткие формы записи совершенно простых житейских понятий — вы их, несомненно, легко поймете.

Книга разделена на 22 короткие главы. В каждой — на примерах, взятых из СМИ, — рассматривается какой-то один способ неправильной интерпретации чисел. Мы надеемся, что к концу каждой главы вы поймете, в чем проблема, и научитесь ее распознавать. Нам кажется, что лучше всего начать с чтения первых восьми глав — в них изложены идеи, которые помогут понять остальное. Но если вам нравится перескакивать с одного на другое — так тоже можно. Если мы опираемся на что-то уже описанное, то указываем на это.

В конце книги мы излагаем ряд предложений по совершенствованию работы СМИ — то, как можно избежать ошибок, которые мы обсуждаем. Мы надеемся, что эта книга станет своего рода руководством по правильной подаче статистики. Будет здорово, если вы посоветуете следовать ему тем СМИ, которые читаете или смотрите.

А теперь вперед.

Глава 1

Как числа могут вводить в заблуждение

Со статистикой врать легко, а без — еще легче.
Приписывается статистику Фредерику Мостеллеру

Из-за COVID-19 человечество прошло ускоренный (и весьма дорогостоящий!) курс статистики. Все были вынуждены в сжатые сроки познакомиться с экспоненциальными кривыми и интервалами неопределенности, ложноположительностью и ложноотрицательностью, усвоить разницу между уровнем инфекционной смертности и показателем летальности. Некоторые из этих понятий, бесспорно, сложны, но даже те, что на первый взгляд кажутся простыми, — например, количество умерших от вируса — на поверку вызывают затруднения. В первой главе мы рассмотрим, как обычные с виду числа могут удивительным образом сбивать с толку.

Одним из первых люди усвоили коэффициент распространения (R). Если еще в декабре 2019 года вряд ли хотя бы один человек из пятидесяти знал о нем, то уже к концу марта 2020-го этот показатель упоминался в новостях практически без всяких пояснений. Но поскольку числа могут вести себя очень коварно, искренние попытки сообщить аудитории об изменениях R вводили читателей и зрителей в заблуждение.

Напомним: R — это *репродуктивное число* чего-либо. Оно применимо ко всему, что распространяется или воспроизводится: мемам, людям, зевоте и новым технологиям. В эпидемиологии

инфекционных болезней R — это число людей, которых в среднем заражает один заболевший. Если у инфекции коэффициент распространения равен пяти, то каждый инфицированный заражает в среднем пятерых.

Конечно, этот показатель не так прост: это всего лишь среднее. При R = 5 каждый из сотни человек может заразить ровно пятерых, но может случиться и так, что 99 человек не заразят никого, а один заразит 500 человек. Возможен и любой промежуточный вариант.

Причем с течением времени коэффициент распространения меняется. R может быть сильно больше в самом начале эпидемии, когда ни у кого еще нет иммунитета и никакие превентивные меры — социальное дистанцирование или ношение масок, — скорее всего, еще не приняты. Одна из задач здравоохранения в этот момент — с помощью вакцинации или выработки у населения новых привычек снизить R. Ведь если он выше единицы, инфекция будет распространяться экспоненциально, а если ниже — эпидемия сойдет на нет.

Но даже с учетом всех этих тонкостей можно было бы ожидать, что в случае вируса есть одно простое правило: если R растет, это плохо. Поэтому в начале мая 2020 года никого не удивлял тон сообщений, заполонивших британскую прессу: «коэффициент распространения вируса снова превысил единицу»[1], вероятно из-за «скачка заболеваемости в домах престарелых»[2].

Но, как обычно, всё несколько сложнее.

С 2000 по 2013 год медианная заработная плата в США выросла примерно на 1% в реальном выражении (то есть с учетом инфляции)[3].

> *Эту врезку читать необязательно, но, если вы не помните разницу между медианой и средним арифметическим, не пропускайте ее.*

Понятия среднего арифметического, медианы и моды вы могли узнать в школе. Что такое среднее арифметическое, наверное, даже помните — нужно сумму нескольких чисел разделить на их количество. А медиана — это среднее число в последовательности чисел.

Разница вот в чем. Пусть население — 7 человек, причем один из них зарабатывает 1 фунт в год, один — 2 фунта и так далее — до 7. Если все эти числа сложить, получится 1 + 2 + 3 + 4 + 5 + 6 + 7 = 28. Разделив 28 на число людей (7), получим 4 фунта. Среднее арифметическое — 4 фунта.

А чтобы узнать медиану, числа не складывают, а располагают по возрастанию: с левого края заработок в 1 фунт, потом — 2, и так до 7 с правого края. Так вы увидите, кто оказался в середине — человек, получающий 4 фунта. Так что и медиана у нас равна 4 фунтам.

Теперь представим, что тот, кто зарабатывает 7 фунтов, продает свой технический стартап компании Facebook* за миллиард. Наше среднее арифметическое внезапно становится равно (1 + 2 + 3 + 4 + 5 + 6 + 1 000 000 000) / 7 = 142 857 146 фунтам. Таким образом, хотя положение 6 из 7 человек никак не изменилось, «среднестатистический гражданин» стал мультимиллионером.

В подобных случаях неравномерного распределения статистики часто предпочитают иметь дело с медианой. Если мы снова выстроим людей по порядку возрастания их зарплат, то в середине опять окажется тот, кто зарабатывает 4 фунта. При изучении реального населения, состоящего из миллионов человек, медиана дает лучшее представление о ситуации, чем среднее арифметическое, особенно если оно искажено зарплатами нескольких суперпреуспевающих работников.

А *мода* — это самое частое значение. Поэтому, если у вас есть 17 человек, зарабатывающих по 1 фунту, 25 — по 2 и 42 — по 3, то мода — 3 фунта. Все несколько усложняется, когда статистики принимаются с помощью моды описывать непрерывные величины вроде высоты, но об этом мы пока постараемся не думать...

---

* Упоминаемый здесь и далее Facebook принадлежит компании Meta, которая признана экстремистской организацией и запрещена в РФ. — *Прим. ред.*

Кажется, что рост медианной заработной платы — это хорошо. Но если рассмотреть отдельные группы населения США, то можно обнаружить нечто странное. Медианный заработок тех, кто окончил только среднюю школу, снизился на 7,9%; тех, кто окончил старшие классы, — на 4,7%. Медианный заработок людей с неполным высшим образованием снизился на 7,6%, а с высшим образованием — на 1,2%.

Окончившие и не окончившие старшие классы, окончившие и не окончившие колледж — медианная зарплата во всех группах с определенным уровнем образования снизилась, хотя медианная зарплата населения в целом повысилась.

Как так?

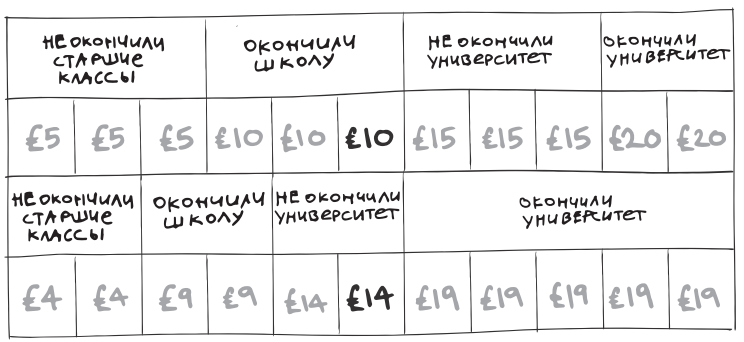
Дело в том, что количество людей с высшим образованием увеличилось, а их медианный заработок снизился. В результате с медианой происходят странности. Это называется парадоксом Симпсона — в 1951 году его впервые описал британский дешифровщик и статистик Эдвард Симпсон[4]. Парадокс распространяется не только на медианы, но и на среднее арифметическое — однако в нашем примере мы поговорим о медианах.

Предположим, что население — 11 человек. Трое из них не пошли в старшие классы и зарабатывают по 5 фунтов в год; трое окончили школу и зарабатывают по 10; трое бросили университет и зарабатывают по 15; а двое закончили бакалавриат и зарабатывают по 20 фунтов. Медианная зарплата такой популяции в целом (то есть зарплата среднего человека при таком распределении доходов, см. врезку на предыдущей странице) составляет 10 фунтов.

Потом правительство проводит кампанию по стимуляции населения к продолжению учебы в старших классах и в университетах. При этом медианная зарплата в каждой группе уменьшается на 1 фунт. Внезапно оказывается, что школу не закончили двое и они получают по 4 фунта, двое выпускников школы зарабатывают по 9, двое бросивших университет — по 14, а пять

выпускников университета — по 19. В каждой группе медианная зарплата уменьшилась на 1 фунт, но у населения в целом она выросла с 10 фунтов до 14. Вот и в американской экономике в период с 2000 по 2013 год случилось нечто подобное, только в более крупных масштабах.

ПАРАДОКС СИМПСОНА



Такое происходит на удивление часто. Например, чернокожие американцы курят чаще, чем белые, но если разбить их на группы по уровню образования, то оказывается, что в *каждой из них* чернокожие курят реже. А все потому, что среди более образованных граждан, где процент курящих меньше, ниже доля чернокожих[5].

Или вот еще один широко известный пример. В сентябре 1973 года в аспирантуру Калифорнийского университета в Беркли подали заявки 8000 мужчин и 4000 женщин. Из них было принято 44% мужчин и только 35% женщин.

Но если посмотреть повнимательнее, то можно заметить: почти на всех факультетах у женщин было *больше шансов* поступить. Самый популярный факультет принял 82% подавших заявки женщин и лишь 62% мужчин; второй по популярности — 68% женщин и 65% мужчин.

Тут дело в том, что женщины подавали заявки на факультеты с самым большим конкурсом. На один из факультетов было подано 933 заявки, из которых 108 подали женщины. Зачислили 82% женщин и 62% мужчин.

В то же время на шестой по популярности факультет было подано 714 заявок, из них 341 от женщин. Здесь поступили 7% женщин и 6% мужчин.

Но если сложить данные по этим двум факультетам, то на них поступало 449 женщин и 1199 мужчин. Было принято 111 женщин (25%) и 533 мужчины (44%).

Еще раз: на каждом из факультетов в отдельности у женщин было больше шансов поступить, а на двух вместе — *меньше*.

Как это лучше всего представлять? Зависит от обстоятельств. В случае с зарплатами американцев можно считать медианы более информативными, потому что медианный американец стал зарабатывать больше (поскольку теперь больше американцев оканчивают колледжи и школы). А в случае с аспирантами можно говорить о том, что, какой бы факультет ни выбрала женщина, у нее больше, чем у мужчины, шансов поступить в аспирантуру. Но с таким же успехом можно говорить о том, что для людей, не окончивших школу, ситуация ухудшилась; и можно отметить, что тем факультетам, на которые хотят поступать женщины, явно не хватает ресурсов: они могут принять лишь небольшую долю подавших заявки. Беда в том, что в ситуациях парадокса Симпсона можно высказывать противоположные точки зрения — в зависимости от вашей политической позиции. Честнее всего тут было бы сообщать о наличии этого парадокса.

А теперь вернемся к коэффициенту распространения COVID-19. Он вырос, стало быть, вирус поражает больше людей, а это плохо.

Только все не так просто. Одновременно происходили две как бы отдельные эпидемии: в домах престарелых и больницах болезнь распространялась не так, как в стране в целом.