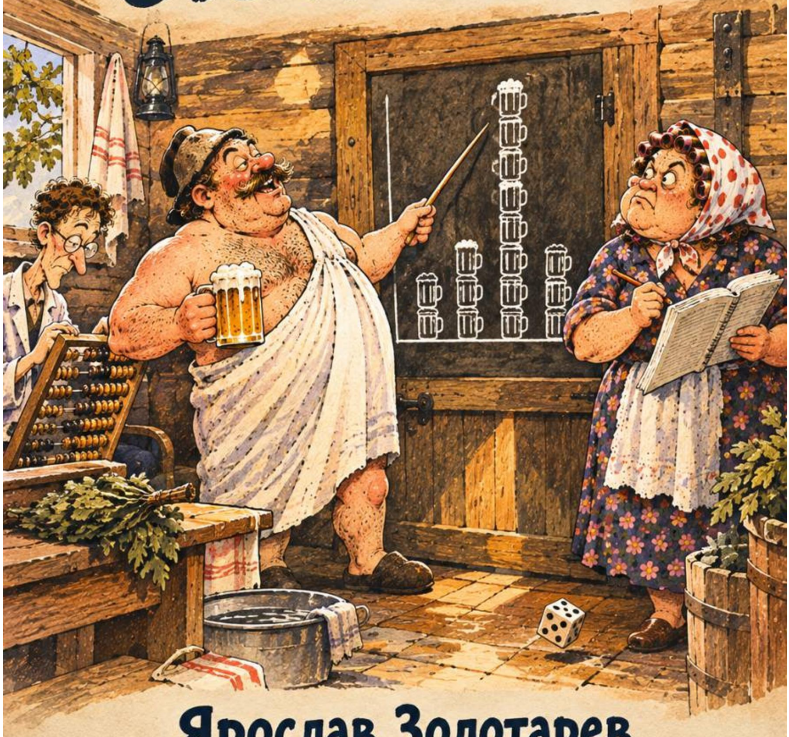
Прикольная статистика



Ярослав Золотарев

18+

Ярослав Золотарёв

Прикольная статистика

https://litres.ru/74715463
SelfPub; 2026

Аннотация

Средняя температура по больнице 36,6: у восьми больных жар, двое в морге. Врёт ли главврач? Ни на одну десятую градуса. Так как же цифры обманывают, не солгав ни разу?

Это книга о математической статистике для тех, кто её боялся. Здесь Петрович считает кружки в бане, тёща ведёт тетрадь, а зять ставит опыты с самогоном. Рядом с ними настоящая наука: прусские кони, немецкие танки, дама с чаем, мёртвый лосось на томографе, парадокс Беркли и спор об авторе «Тихого Дона».

Тридцать три главы проводят от среднего и медианы до доверительных интервалов, p-значений, регрессии и испытаний лекарств. Каждый метод назван своим научным именем, разобран на простых числах и показан там, где он ломается. После глав идут смешные задачи с ответами.

Подойдёт студентам, журналистам и всем, кому важно отличать доказательство от красивой цифры. В конце вас ждут десять вопросов к любому числу из новостей.

Содержание

Глава 1. Средняя температура по больнице 6
Всё отнять и поделить 7
Формула, которую вы умели в третьем классе 8
Три способа опозориться 11
Как среднее работает в настоящей науке 13
Задачи 15
Глава 2. Зарплата, которой никто не получает 18
Середняк и модник 20
Три способа опозориться снова 23
Как медиана работает в настоящей науке 25
Задачи 27
Глава 3. Автобус, который в среднем приходит вовремя 30
Квадратные минуты 32
Где разброс даёт промах 36
Как разброс работает в настоящей науке 37
Задачи 39
Глава 4. Портрет толпы 42
Столбики и колокол 43
Как картинка вводит в заблуждение 47
Как гистограмма работает в настоящей науке 49
Задачи 51
Глава 5. Один бестселлер и тысяча книг без 53

читателей
Закон для стенографистов 54
Где закон даёт осечку 57
Как тяжёлые хвосты работают в настоящей науке 59
Задачи 61
Глава 6. Дождь с вероятностью 70% и двое с одним днём рождения 63
Три смысла одного числа 65
Три любимых заблуждения 69
Как совпадения работают в настоящей науке 71
Задачи 73
Глава 7. Положительный тест у здорового человека 75
Формула преподобного 77
Когда вопрос переворачивают 79
Как Байес работает в настоящей науке 82
Задачи 84
Глава 8. Прусские кони и опечатки 87
Метод имени копыта 88
Когда конь даёт осечку 92
Где это применяют всерьёз 93
Задачи 95
Глава 9. Почему лучшие школы часто маленькие 98
Закон, который знает каждый дурак 100
Чего закон больших чисел не обещал 105

Где это применяют всерьёз 107
Задачи 110
Глава 10. Откуда берётся колокол 113
Доска с гвоздями 115
Когда колокол треснет 120
Где это применяют всерьёз 122
Задачи 124
Глава 11. Суп пробуют ложкой 127
Ложка и кастрюля 129
Когда ложка врёт 133
Где это применяют всерьёз 135
Конец ознакомительного фрагмента. 136

Ярослав Золотарёв
Прикольная статистика

Глава 1. Средняя температура по больнице

Главврач докладывает министру: «Средняя температура по больнице тридцать шесть и шесть, товарищ министр. Норма». Министр доволен, главврачу премия, больным компот.

Врёт ли главврач? Ни на одну десятую градуса. В его заведении десять постояльцев. Восемь лежат в палатах с температурой 40 и уже беседуют с покойной бабушкой. Ещё двое лежат в морге, где холодильник сломался при прошлом главврача, и потому там ровно 23 градуса. Считаем: $8 \cdot 40 + 2 \cdot 23 = 366$. Делим на десять постояльцев и получаем 36,6.

Шутка старая, заезженная, и я начинаю с неё книгу нарочно. В ней уже лежит вся статистика: и её сила, и её подлость. Число посчитано безупречно. Ни один человек в больнице такой температуры не имеет. Министр при этом узнал о положении дел меньше, чем знал до доклада.

Так зачем вообще нужно среднее, если им так легко морочить голову?

Всё отнять и поделить

Представьте, что вас спрашивают, сколько пьют в бане по субботам. Можно ответить честно и подробно: Коля две кружки, Вася три, Серёга три, Михалыч четыре, Петрович тринадцать. Для пятерых это ещё терпимо. Для пяти тысяч человек такой ответ никто не дослушает, а для ста сорока миллионов его не дослушает даже Росстат.

Здравый смысл предлагает простой выход: заменить всю толпу одним числом. Каким? Тем, которое досталось бы каждому, если бы всё выпитое слили в одну бочку и разлили поровну.

Философы со времён Платона спорят о справедливом распределении благ. Полиграф Полиграфович Шариков решил вопрос одной фразой: «Взять всё, да и поделить». Именно этим и занимается среднее арифметическое. Это единственное место на Земле, где коммунизм построен полностью, причём без жертв, поскольку всё происходит на бумаге.

Формула, которую вы умели в третьем классе

Считаем баню. Всего выпито 2 + 3 + 3 + 4 + 13 = 25 кружек. Народу пять человек. На брата приходится по пять кружек. Вся наука умещается в одну строчку.



(2 + 3 + 3 + 4 + 13) : 5 = 25 : 5 = 5 кружек

Буквы здесь занимаются тем же, чем фамилии в ведомости. Икс с цифрой означает показание конкретного участника: x1 у Коли, x5 у Петровича. Буква n означает количество участников. Черта над иксом говорит, что перед нами среднее. Учёные записывают понятные вещи непонятно вовсе не из вредности. Так короче, а ещё так солиднее выглядит за-

явка на грант.

У среднего есть свойство, ради которого его и носят на руках три века. Посмотрим, кто насколько отклонился от положенных пяти кружек. Коля недопил три, Вася и Серёга по две, Михалыч одну. Итого недобор восемь. Петрович перебрал ровно восемь. Совпадения тут нет, так выходит всегда и при любых числах.



Значит, среднее можно понимать как точку опоры качелей. Рассадите всех по длинной доске соответственно их кружкам, и доска уравновесится именно на пятёрке. Петрович держит четверых, потому что сидит далеко, и в этом простая школьная физика. Примет он ещё пять кружек, и сред-

нее по бане вырастет до шести, хотя остальные не сделали ни глотка.

Здесь сразу видны оба лица метода. Среднее честно учитывает каждого, в этом его сила. И среднее честно учитывает каждого, в этом его беда.

Три способа опозориться

Смешать несмешиваемое. Среднее не спрашивает, что вы в него свалили. Живые с мёртвыми, грудные дети с пенсионерами, дворник с олигархом: всё перемелет без жалоб. Получится, например, что у среднего человека примерно одна грудь и одно яичко. Арифметика безупречна, а на пляже такого гражданина пока не встречали. Вывод прост: прежде чем усреднять, стоит выяснить, одной ли породы усредняемое.

Усреднить средние. Возьмём пример для филологов, числа в нём условные. У автора есть роман в 10 000 предложений, по 20 слов в среднем, и рассказ в 100 предложений, по 6 слов. Молодой исследователь пишет в диссертации: (20 + 6) : 2 = 13 слов, и выводит отсюда лапидарность мастера. Но всего тут 10 100 предложений, и в них 200 600 слов. Делим и получаем почти 20 слов на предложение. Лапидарность мастера оказалась болезнью исследователя. Большой кусок должен весить больше маленького, и никакая учёная степень этого не отменяет.

Искать среднего человека. В начале 1950-х годов американские ВВС обмерили 4063 лётчика всеми мыслимыми способами: денег на сантиметры у них всегда хватало. Кабины строили под среднего пилота, и было бы неплохо узнать его размеры поточнее. Лейтенант Гилберт Дэниелс за-

дал простой вопрос: а сколько там вообще средних? Он взял десять главных параметров и посчитал средним каждого, кто попал в средние 30% диапазона по каждому из них. Условие щедрое. Средних по всем десяти нашлось ровно ноль человек. Воздушный флот сверхдержавы готовился сажать в кабину призрака.

Среднее описывает толпу и не обязано описывать ни одного человека в ней. Потому кресло в вашей машине и ездит взад-вперёд.

Как среднее работает в настоящей науке

После такого позора среднее стоило бы выбросить. Но придумали его в современном виде для дела вполне возвышенного: для звёзд. Десять астрономов меряют положение одной планеты и получают десять разных чисел, потому что глаз устал, труба дрожит, а жена зовёт ужинать. Планета при этом одна. Ошибки бывают в обе стороны, а наши качели показали, что в сумме отклонения от среднего гасят друг друга. Значит, среднее из десяти плохих измерений обычно надёжнее каждого из них в отдельности.

В XIX веке бельгийский астроном Адольф Кетле перенёс этот приём с небес на землю. Он взял обхваты груди 5738 шотландских солдат и получил в среднем чуть больше 39 3/4 дюйма, это около 101 сантиметра. Дальше восторг его унёс. Кетле рассуждал примерно так: у астрономов среднее даёт истину, а всё остальное ошибки, стало быть, и с людьми то же самое. Так появился «средний человек», замысел природы, а живой солдат с его тещёй оказались погрешностью измерения. Идея вздорная, ведь планета одна, а людей много. Зато сам подход оказался гениальным: общество можно измерять почти так же, как небо.

А теперь главное чудо. В 1907 году восьмидесятипятилетний Фрэнсис Гальтон, двоюродный брат Дарвина, напечатал

в журнале Nature заметку о выставке скота в Плимуте. Там за шесть пенсов можно было угадать, сколько будет весить откормленный вол после забоя и разделки. Гальтону досталось 787 годных билетов. Угадывали мясники и фермеры, но и публика, которая, по выражению самого Гальтона, разбиралась в волах примерно так же, как средний избиратель в государственных делах. В опубликованной заметке вол весил 1198 фунтов, около 543 килограммов. Среднее по толпе вышло 1197.

Толпа бездельников ошиблась на один фунт, потому что завышения и занижения погасили друг друга. Более того, позже статистик Кеннет Уоллис поднял рабочие листы Гальтона и нашёл там вес вола 1197 фунтов. Выходит, на фунт ошибся сам великий учёный, когда переписывал цифру, а бездельники угадали точно.

Тот же приём работает сегодня повсюду. Физик усредняет повторные измерения, врач сравнивает средние показатели в группах с лекарством и без него, лингвист считает среднюю длину слова, чтобы понять, кто написал текст. Большая часть этой книги так или иначе растёт из одного действия: сложить и поделить. Важно только помнить, что усреднить можно что угодно, но осмысленное число выйдет не всегда. Вол у Гальтона был один. Больных у главврача десять, и двое из них уже ни на что не жалуются.

Задачи

1. Ремонт в морге. В больнице 30 больных с температурой 38 и 6 постояльцев морга при 20 градусах. Какую среднюю температуру получит министр?

Ответ: $(30 \cdot 38 + 6 \cdot 20) : 36 = 1260 : 36 = 35$ градусов. Теперь главврачу придётся объяснять, почему вся больница переохлаждена, при том что в палатах все горят.

2. Пельмени равенства. Пятеро съели 10, 12, 14, 16 и 48 пельменей. Сколько съел в среднем каждый? Сколько человек съело столько?

Ответ: в среднем $100 : 5 = 20$ пельменей. Съел столько никто. Четверых обделили, один опять сидит далеко на качелях.

3. Спасение Вовочки. У Вовочки в журнале две двойки и две тройки. Сколько пятёрок ему нужно получить, чтобы средний балл стал 3? А чтобы стал 4?

Ответ: для тройки хватит одной: $(10 + 5) : 5 = 3$. Для четвёрки нужно шесть: $(10 + 30) : 10 = 4$. Старые грехи потому и тяжелы, что у среднего хорошая память. Раскаяние оно учитывает только по расценке.

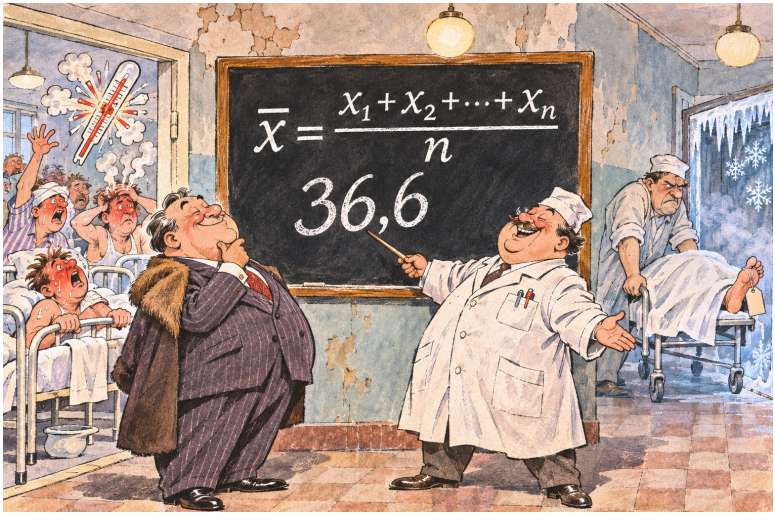
4. Молодой район. В Верхних Выселках 900 жителей со средним возрастом 40 лет, в Нижних 100 жителей со средним возрастом 80 лет. Глава района докладывает: средний возраст 60 лет, дайте денег на соцзащиту. Сколько на самом

деле?

Ответ: (900 · 40 + 100 · 80) : 1000 = 44 года. Район молодой, глава просто усреднил средние. Денег ему всё равно дадут: там тоже усредняют.

5. Гальтон на колхозном дворе. Четыре соседа на глаз оценили кабана в 70, 90, 110 и 130 килограммов. На весах оказалось 100. На сколько ошибся самый точный сосед и на сколько ошиблось среднее?

Ответ: лучшие двое ошиблись на 10 килограммов, а среднее 400 : 4 = 100 не ошиблось вовсе. По отдельности все не правы, вместе правы. Жаль, при таком результате кабана всё равно делить придётся на четверых.


$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$
36,6

Глава 2. Зарплата, которой никто не получает

Вы читаете объявление: «Дружный коллектив, средняя зарплата 125 тысяч». Надеваете единственный приличный пиджак, едете на собеседование, выходите на работу, и в первый же день зарплаты получаете 50. Идёте в отдел кадров бить посуду.

Кадровичка спокойно достаёт ведомость. В фирме шесть человек. Пятеро, включая вас, получают по 50 тысяч. Директор, он же супруг владелицы, получает 500. Всего 750 тысяч на шестерых, то есть по 125. По математике пятёрка, по криминальному кодексу претензий нет. Зарплату в 125 тысяч в этой фирме никогда никто не получал, но так ведь и температура 36,6 в прошлой главе ни у кого не наблюдалась.

При этом ваш вопрос был совсем простой: сколько получает обычный человек в этой конторе? В прошлой главе мы уже видели, как один Петрович на дальнем конце качелей уводит среднее за собой. Значит, нужен показатель, который на Петровича плевать хотел.

Здравый смысл опять справляется без университета, и навык тут требуется школьный. Выстройте всех по зарплате, как на физкультуре по росту, и спросите того, кто стоит посередине. Богатый правый фланг может держать на счету хоть

весь бюджет волости, середина шеренги от этого не сдвинется ни на шаг.

Середняк и модник

Того, кто стоит посередине, учёные называют медианой. Вернёмся в баню из прошлой главы: 2, 3, 3, 4 и 13 кружек. Числа уже стоят по порядку, посередине Серёга с тремя кружками. Медиана равна трём, хотя среднее равнялось пяти.



Теперь пусть Петрович выпьет не 13 кружек, а 130. Вопросы к медицине оставим, статистика отвечает сразу. Среднее подпрыгнет с пяти до 142 : 5 = 28,4 кружки на брата. Медиана останется равна трём. В науке такое равнодушие на-

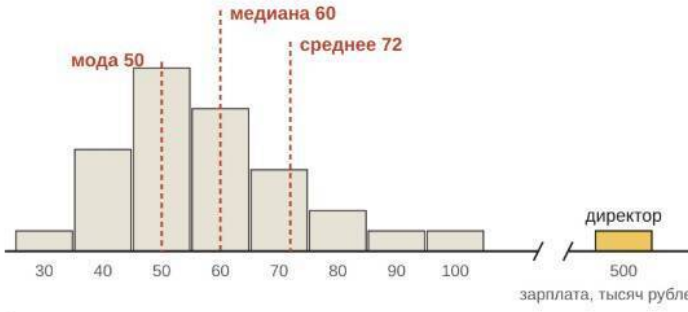
зывают устойчивостью, а Петровича и подобных ему выбросами. Слово неласковое, но точное: он действительно далеко выброшен от всех остальных.

Если людей чётное число, посередине стоит пустое место между двумя соседями. Берём обоих и находим их полусумму. В нашей фирме на третьем и четвёртом местах стоят два бедолаги по 50 тысяч, так что медиана равна 50. Именно эту цифру и стоило написать в вакансии, если бы кто-нибудь хотел сказать правду.

Есть и третий показатель, самый ленивый. Мода означает просто то, что встречается чаще всего. Слово ввёл Карл Пирсон в 1895 году, и никакой ошибки тут нет: модно то, что носят все, даже если оно ужасно. В бане мода равна трём кружкам, в конторе 50 тысячам. Формула для моды не нужна, достаточно уметь считать до десяти и оглядываться по сторонам.

На большой толпе все три показателя расходятся в понятном порядке. Возьмём условную контору побольше, где деньги тоже стекаются вверх, к начальству.

Условная контора: 30 сотрудников и директор на 500 тысяч



$$Me = \frac{x_{n/2} + x_{n/2+1}}{2}$$

если людей чётное число

50, 50, 50, 50, 50, 500: (50 + 50) : 2 = 50 тысяч

Мода показывает, что получает самая большая кучка народу, 50 тысяч. Медиана делит коллектив пополам, на более и менее удачливых, это 60. Среднее утащил вправо директор, и оно равно примерно 72. Притом никто не соврал. Когда видите такой порядок, это примета: у распределения справа длинный хвост, на конце которого кто-то хорошо устроился.

Три способа опозориться снова

Не спешите объявлять медиану царицей полей. У неё и моды есть свои способы осрамиться.

Считать деньги по медиане. Та самая глухота к Петровичу, за которую мы её похвалили, иногда обходится дорого. Пошлите в баню закупать пиво по медиане: пять человек на три кружки, итого 15. Выпьют 25, и дальше вас будут ведать не научные соображения. Потому соискателю нужна медиана, а бухгалтеру среднее. Только среднее, умноженное на число людей, даёт общую сумму, и ни один другой показатель этого не умеет.

Довериться медиане на малой выборке. Пятеро ценителей поэзии поставили сборнику оценки 2, 2, 2, 5, 5. Медиана равна двум, гоните поэта копать картошку. Один человек, выпив с автором, меняет двойку на пятёрку, получается 2, 2, 5, 5, 5, и медиана сразу равна пяти. Среднее при этом скромно сдвинулось с 3,2 до 3,8. Медиана не замечает краёв, зато у самой середины может прыгнуть через пропасть.

Поверить моде. Мод бывает несколько, и нигде это не видно лучше, чем в обувном магазине. Женщины возьмут 37-й размер, мужчины 43-й, и посередине окажутся склады никому не нужных сороковых, если директор заказывал по среднему. Бывает и наоборот: когда все значения разные, моды нет вообще. А на малых цифрах она принципиально не

внушает доверия, ведь два одинаковых ответа из семи случаются просто так.

Единственного правильного показателя не существует. Вопрос всегда в том, что вы собираетесь делать с ответом.

Как медиана работает в настоящей науке

Помните Гальтона с волом? В той же самой истории старик спорил с прошлой главой этой книги, сам того не зная. В статье он счёл главным не среднее, а серединную оценку, и в заметке «Один голос, одно значение» объяснил почему. Думайте о комиссии или присяжных, которые назначают сумму. Среднее, писал он, даёт чудакам власть пропорционально их чудачеству. Один помешанный напишет сумму с десятью нулями и потянет всё решение за собой. Серединная же оценка единственная, против которой нет большинства: любую другую большинство сочтёт либо завышенной, либо заниженной. В том опыте медиана была 1207 фунтов, а среднее 1197. Среднее оказалось ближе к истине, но правота Гальтона не в одном воле. Когда среди голосующих бывают чудаки, а они бывают, медиана защищает от них без всякой цензуры.

Вот почему медиану так любят там, где велик одиночный размах. В апреле 2025 года средняя зарплата в России, по данным Росстата, подобралась к 100 тысячам рублей, а медианная составила 73,9 тысячи. Оба числа честные, просто отвечают на разные вопросы. И когда в телевизоре сообщают об одном из них, полезно уточнить, о каком. За пределами России медиана работает даже надёжнее: Евросоюз определяет порог риска бедности в 60% от медианного дохода.

Разбогатей пара миллиардеров, такой порог не шелохнётся, тогда как порог от среднего бы уполз вверх вместе с яхтой.

Но самая сильная история про медиану совсем не смешная, и шутить над ней не буду. В 1982 году палеонтолог Стивен Джей Гулд узнал, что у него мезотелиома, редкий рак. В библиотеке он вычитал, что медиана выживания восемь месяцев. Обычный человек услышит «через восемь месяцев ты умрёшь». Гулд был учёным и услышал другое. Медиана означает, что половина больных умирает раньше, а половина позже. Слева шкала упирается в ноль, справа же хвост может тянуться на годы. Он был молод, диагноз поставлен рано, и были основания надеяться на правую часть. В 1985 году он написал об этом эссе «The Median Isn't the Message», по смыслу «Медиана ещё не приговор», и его читают до сих пор. Прожил он ещё около двадцати лет и умер в 2002 году от другого рака.

Разница между средним, медианой и модой кажется школьной забавой ровно до того дня, когда от неё зависит ваша зарплата, ваше лечение или ваша надежда. При этом одного числа мало любому из них. Надежду Гулду дал разброс, и про разброс будет следующая глава.

Задачи

1. Автобус равенства. В автобусе едут 40 человек с доходом 25 тысяч рублей в месяц. На остановке входит человек с доходом 40 миллионов, потому что у его машины спустило колесо. Какими стали средний доход и медиана?

Ответ: среднее 41 000 000 : 41 = 1 миллион, медиана 25 тысяч. Рост благосостояния населения в сорок раз достигнут на одной остановке. Отметить можно, но за свои 25 тысяч.

2. Село трезвенников. В селе 100 жителей. Девяносто девять не пьют, Петрович выпивает 100 литров в год. Найдите среднее, медиану и моду.

Ответ: среднее 1 литр на душу, медиана и мода 0. Районная наркология запишет потребление на всё село, хотя отвечает один подвижник.

3. Честная вакансия. В фирме директор получает 1 миллион, два зама по 200 тысяч, семеро рабочих по 40 тысяч. Какую цифру напишет отдел кадров, а какую стоило бы услышать соискателю?

Ответ: отдел кадров напишет среднее, 1680 : 10 = 168 тысяч. Соискателю нужна медиана, 40 тысяч. Но тогда он не придёт, а кто-то должен работать.

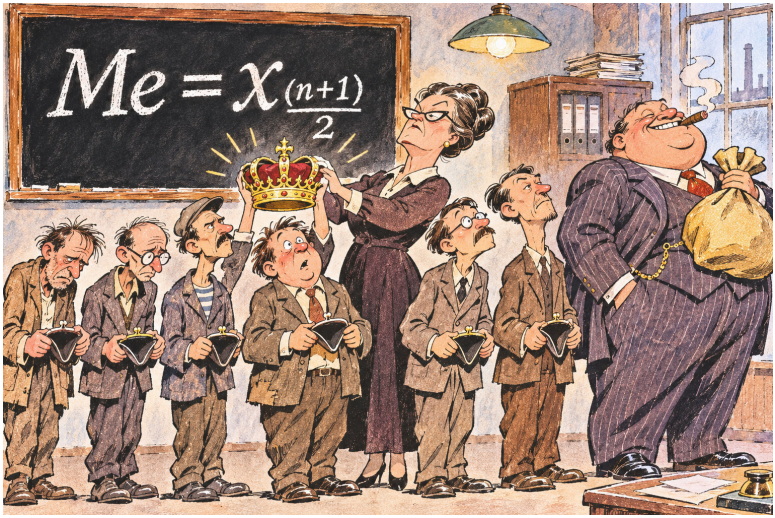
4. Тёща в жюри. На конкурсе самодеятельности по десятибалльной шкале четверо судей поставили 7, 8, 8 и 9 бал-

лов. Пятый судья, тёща участника, написала 100, потому что мальчик старался. Ваши среднее и медиана?

Ответ: среднее 132 : 5 = 26,4 из десяти возможных, медиана 8. Именно о таких судьях и писал Гальтон. Только он не знал вашу тёщу.

5. Ботинки по среднему. За день проданы ботинки размеров 37, 37, 37, 38, 42, 43, 43, 43. Найдите среднее, медиану и моду. Ботинки какого размера закупать?

Ответ: среднее 320 : 8 = 40, медиана (38 + 42) : 2 = 40. Мод две: 37 и 43. Закупать надо по модам. Среднее и медиана единодушно советуют 40-й, который за день не купил никто.


$Me = x_{\frac{(n+1)}{2}}$

Глава 3. Автобус, который в среднем приходит вовремя

Начальник автопарка выходит к разгневанным гражданам с отчётом. За неделю тринадцатый маршрут пришёл на двадцать минут раньше расписания, на двадцать позже, на десять раньше, на десять позже и один раз, по всей видимости случайно, точно вовремя. Складываем: -20 + 20 - 10 + 10 + 0 = 0. В среднем автобус не опаздывает ни на секунду. Начальнику грамота, водителю премия, вам полчаса на морозе с выводом о несовершенстве мироздания.

Рядом ходит пятый маршрут. За ту же неделю у него одна минута раньше, одна позже и три раза точно вовремя. В среднем тот же самый ноль. В отчёте оба автобуса неразличимы. В жизни один возит людей на работу, а другой приучает их к мысли о бренности сущего.

Значит, одного среднего мало. И медианы тоже мало, она у обоих равна нулю. Всё, что мы уже выучили, отвечает на вопрос «где середина». Теперь нужен второй вопрос: насколько далеко от неё люди разбрелись.

Здравый смысл предлагает для начала посмотреть на крайних: у тринадцатого от минус 20 до плюс 20, у пятого от минус 1 до плюс 1. Это уже неплохо, и называется размахом. Беда в том, что размах знает только двух человек. Один раз

в году запил водитель пятого, автобус пришёл на час позже, и по размаху он вмиг стал хуже тринадцатого. Нужен показатель, который спросит каждого.

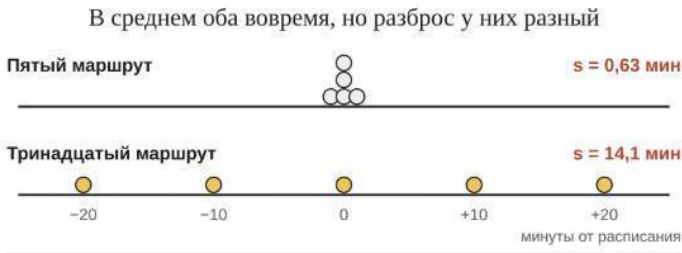
Квадратные минуты

Возьмём тринадцатый маршрут. Хочется узнать, на сколько он в среднем отклоняется от среднего. Тут нас встречает наша же первая глава с дубиной. На качелях мы выяснили, что сумма отклонений от среднего всегда равна нулю. Значит, по такому показателю любой бардак на свете безупречен. Плюсы съедают минусы, и доклад начальника автопарка опять готов.

Значит, от минусов надо избавиться. Можно просто отбросить знак, и такой показатель в науке тоже есть. А можно возвести в квадрат, ведь минус на минус даёт плюс. Статистики пошли вторым путём, потому что с квадратами дальше намного удобнее считать. Квадраты отклонений: 400, 400, 100, 100 и 0. Всего 1000. Делим на пять поездок и получаем 200. Это число называется дисперсией, и оно ещё всплывёт в нашей книге.

У дисперсии есть только один недостаток: она измеряется в квадратных минутах. Что это такое, не знают ни Кант, ни автопарк. Чтобы вернуться к людям, извлекаем корень: корень из 200 составляет примерно 14,1 минуты. Это и есть стандартное отклонение. Сам термин первым употребил письменно в 1894 году всё тот же Карл Пирсон, главный поставщик слов для этой книги. У пятого маршрута тот же расчёт даёт 0,63 минуты. Автобусы наконец различаются, и

в двадцать с лишним раз.



$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n}}$$

кто насколько промахнулся, в квадрате;
поделили на всех и взяли корень

(400 + 400 + 100 + 100 + 0) : 5 = 200, корень из 200 ≈ 14,1 минуты

Здесь стоит сделать одну оговорку для совести. Когда перед нами не все данные, а лишь выборка, в учебниках делят не на n, а на n - 1. Почему, объясню, когда доберёмся до пивовара Госсета и малых выборок. На большой толпе разницы почти нет, а в этой главе нас интересует сама идея.

Идея такая: стандартное отклонение показывает, на сколько типично отбиваются от среднего. Тринадцатый маршрут типично угуливает на 14 минут, пятый меньше чем на минуту. Разброс стоит настоящих денег и настоящего вре-

мени, и всякий, кто ждал на остановке, уже оплатил его своими ногами. Пусть по расписанию автобус ходит раз в 20 минут. Придёте наугад, и будете ждать в среднем 10 минут. А теперь пусть автобусы ходят парой, один за другим, как настоящие друзья, и потом оба пропадают на 40 минут. Средний интервал всё те же 20 минут, и парк машин тот же, а ждать вы будете в среднем 20 минут, вдвое дольше. Расписание не изменилось в среднем, ваша жизнь сократилась тоже не в среднем.

У стандартного отклонения есть и вторая служба. Оно работает единой линейкой для вещей, которые сравнивать нельзя. Что удивительнее: 13 кружек пива или рост 190 сантиметров? На вид вопрос для сумасшедшего дома. Но можно спросить по-другому: на сколько своих стандартных отклонений человек отошёл от своего среднего. Это называется z-оценкой, и единиц измерения у неё нет: кружки сокращаются с кружками, сантиметры с сантиметрами.



$$z = \frac{x - \bar{x}}{s}$$

отступ от среднего, измеренный в разбросах

$(13 - 5) : 4 \approx 2$ $(190 - 176) : 7 = 2$

Среди своих банных товарищей Петрович отклонился примерно на два стандартных отклонения. Ровно так же странен условный дылда ростом 190 при среднем 176 и отклонении 7. Пиво и рост приведены к одному знаменателю, чего не удавалось даже великим философам.

Где разброс даёт промах

Петрович бьёт по новому показателю ещё больнее. Среднее он утаскивал за собой, а тут он ещё и возводится в квадрат. Четверо в бане, без него, дают стандартное отклонение 0,71 кружки. С ним 4,05. Один человек увеличил разброс почти вшестеро. Поэтому прежде чем считать квадраты, полезно посмотреть, кто там сидит на дальнем конце скамьи.

Слон и мышь. Пусть у мышей вес колеблется на 5 граммов при среднем 20, а у слонов на 500 килограммов при среднем 5 тонн (числа условные). По стандартному отклонению слоны разнороднее мышей в сто тысяч раз. Делим отклонение на среднее и получаем 10% у слонов и 25% у мышей. Стало быть, слоны как раз пооднороднее. Показатель так и называется, коэффициент вариации, и разномасштабные вещи без него сравнивать нельзя.

Одинаковое число, разная жизнь. Автобус приходит с отклонениями -10, -10, +10, +10. Стандартное отклонение равно десяти, и он никогда не бывает вовремя. Другой приходит с отклонениями -14, 0, 0, +14. Отклонение почти то же, 9,9, но половину поездок он приходит точно по расписанию. Два числа, среднее и разброс, гораздо лучше одного, но форму они не описывают. Формой мы займёмся в следующей главе.

Как разброс работает в настоящей науке

Разброс долго считали помехой. Затем оказалось, что в нём можно читать, как в книге.

В мае 1924 года Уолтер Шухарт из американских Bell Labs написал начальству служебную записку с графиком. Мысль была простая. Любой станок делает детали чуть-чуть разными, и на каждую такую разницу бегать к наладчику не надо, только хуже сделаете. Другое дело, если деталь вышла за три стандартных отклонения: такое обычным шумом уже не объяснишь, тут пора искать причину. Эти границы на контрольной карте по сей день рисуют на заводах. Ступень мудрости начальника в том и состоит, чтобы отличать проблему от обычного дрожания жизни.

Той же меркой физики проверяют открытия. 4 июля 2012 года в Женеве объявили о находке бозона Хиггса на уровне пяти сигм, они же пять стандартных отклонений. Это означает: если частицы нет, такое или более сильное отклонение случайно выпадает примерно один раз из 3,5 миллиона. Обратите внимание на формулировку. Это не вероятность того, что бозона нет, на этом спотыкались тогда многие газеты. К этой тонкости мы ещё вернёмся, когда будем разбирать даму с чаем.

z-оценка пригодилась и литературоведам. Любой автор

оставляет отпечатки в самых частых словах, всяких «и», «не» и «в», которые сознательно не контролирует. Но частоты у этих слов сильно разные, и просто складывать разницы нельзя. В 2002 году Джон Берроуз опубликовал метод, названный дельтой: частоту каждого слова переводят в z-оценки, а уже потом сравнивают тексты. Пушкин и какой-нибудь графоман становятся подсудимыми той же формулы, что Петрович. Подробнее об этом в главе об авторстве.

Задачи

1. На ком жениться. Первый жених за четыре месяца приносит домой 50, 50, 50 и 50 тысяч. Второй 0, 0, 0 и 200 тысяч. Найдите среднее и стандартное отклонение у обоих.

Ответ: среднее у обоих 50. Отклонение у первого 0, у второго корень из 7500, то есть около 87 тысяч. Первый скучен, у второго душа шире ведомости. Статистика на обоих только указывает, а решать придётся вам.

2. Заяц в среднем убит. Два охотника стреляют по зайцу. Один промахнулся на метр влево, другой на метр вправо. Найдите среднее отклонение и стандартное отклонение.

Ответ: среднее 0, заяц убит точно в центр. Стандартное отклонение 1 метр, заяц убежал. На ужин в среднем жаркое.

3. Брод статистика. Глубина реки в пяти точках: 0,1, 0,2, 0,3, 0,4 и 4 метра. Найдите среднюю глубину и стандартное отклонение. Может ли статистик ростом 170 сантиметров спокойно перейти брод?

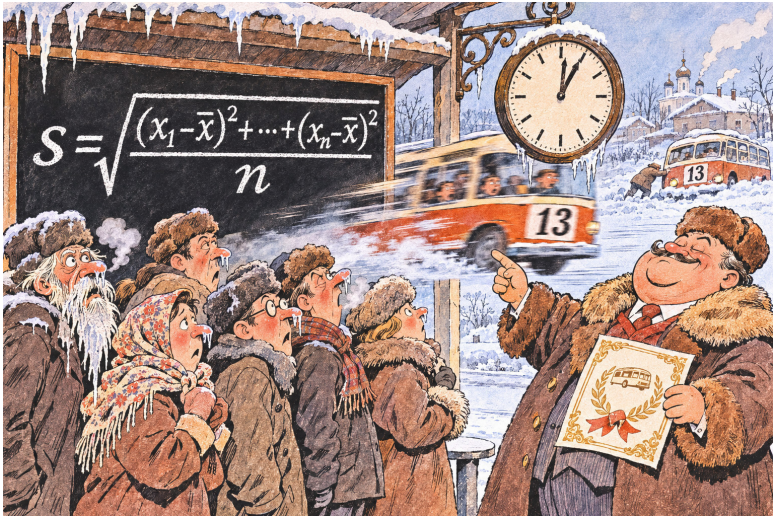
Ответ: средняя глубина 1 метр, отклонение около 1,5 метра. Если статистик посмотрел только на среднее, его труды выйдут посмертно.

4. Кто страннее. Вася съел 30 пельменей при среднем 20 и отклонении 5. Петя выпил 7 кружек при среднем 4 и отклонении 1. Кто сильнее выбился из коллектива?

Ответ: у Васи z = (30 - 20) : 5 = 2, у Пети z = (7 - 4) : 1 = 3. Петя страннее, хотя разница у него всего три кружки. Десять пельменей сверху тут никого бы не удивили.

5. Ждун. Автобусы приходят поочерёдно через 10 и через 30 минут. Вы приходите наугад. Как долго ждать в среднем? Подсказка: в длинный промежуток вы попадёте втрое чаще, чем в короткий.

Ответ: 3/4 · 15 + 1/4 · 5 = 12,5 минуты, хотя при ровном интервале в 20 минут вы ждали бы 10. Начальник автопарка может не беспокоиться, в его отчёте оба расписания дают ровно 20.



Глава 4. Портрет толпы

На плац вывели сотню новобранцев, и старшине надо выписать на них обмундирование. Из прошлых глав он уже знает: средний рост 176, стандартное отклонение 7. Полезные цифры, только штаны по ним всё равно не закажешь. Сколько нужно штанов на длинных, сколько на коротких? Два числа об этом молчат. В прошлой главе мы уже видели, что толпы с одинаковым средним и одинаковым разбросом могут быть устроены совершенно по-разному.

Получить всё сразу можно, только если посмотреть на толпу целиком. Только таблицу из ста фамилий глаз не воспринимает, а из миллиона тем более.

Здравому смыслу старшины никакой математики не нужно. Он закричит: те, кто ниже 165, стройсь у первого столба! От 165 до 170 у второго! И так далее, в затылок. Через минуту на плацу видно несколько колонн разной длины. Самой длинной окажется какая-нибудь средняя, а по краям будут торчать по два-три оригинала. Старшина только что построил гистограмму из живых людей, и от статистика его отличает только то, что он при этом матерится.

Столбики и колокол

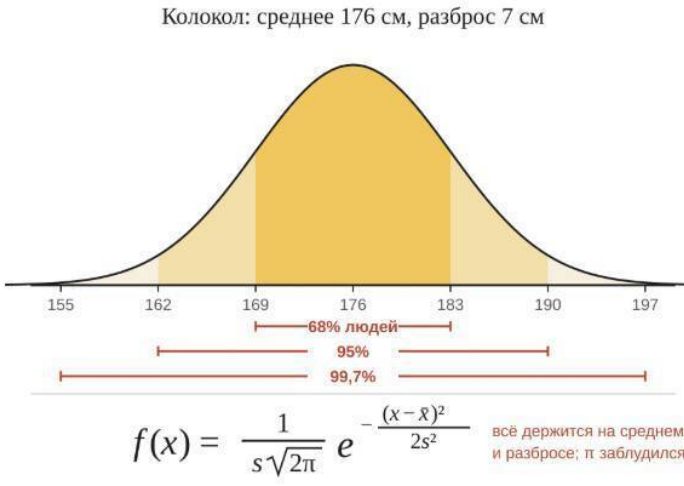
Перенесём плац на бумагу. Разобьём рост на промежутки по пять сантиметров, над каждым поставим столбик высотой в число людей. Это и есть гистограмма. Слово ввёл в лекциях 1892 года, как вы могли догадаться, Карл Пирсон. Расчёт к ней нужен ровно один: сколько людей попало в столбик и какую долю от всех это составляет.



Теперь старшина может заказывать. Размер на рост 175–180 носят 30% бойцов, значит, на тысячу возьмём 300 штанов. А ещё по картинке видно то, чего никаким числом не

скажешь: горб один, справа и слева он пологий примерно одинаково, и никакого директора на дальнем конце не притаилось. Старшина наш придуман, но приём настоящий. В 1914 году ботаник и генетик Альберт Блейксли выстроил по росту 175 студентов и сфотографировал их. Так появилась знаменитая живая гистограмма в студенческих пиджаках.

Постройте такую гистограмму не для ста человек, а для миллиона, да сделайте столбики поуже, и ступеньки сгладятся в плавную кривую, похожую на колокол. Называется она нормальным распределением. Откуда она берётся, расскажет десятая глава, а пока хватит практического правила. Зная два числа из прошлых глав, можно сразу сказать, сколько народу где стоит.



Примерно 68% людей лежит в одном стандартном отклонении от среднего, 95% в двух, 99,7% в трёх. Значит, дылда на 190 сантиметров из прошлой главы уже стоит на границе двух отклонений: выше него только около 2,5% толпы. Старшина может не заказывать длинных штанов вагонами.

У самого колокола на картинке вы видите формулу, пугаться которой не нужно, считать по ней вручную никто не будет. Вся она держится на тех же двух числах, среднем и стандартном отклонении. Рассмешить может другое: туда пролезло число пи. Нобелевский лауреат Юджин Вигнер начал с этого знаменитую статью о непостижимой эффектив-

ности математики. Статистик показывает однокласснику работу о населении, одноклассник тычет пальцем в знак пи и слышит: это отношение длины окружности к диаметру. Тот решает, что его разыгрывают: население тут ведь ни при чём окружность. Тем не менее она тут. В штанах старшины тоже.

Как картинка вводит в заблуждение

Ширина столбиков. Гистограмма выглядит объективной, но размер промежутка выбирает человек. Возьмём тех же новобранцев и сделаем два широких столбика: от 160 до 180 и от 180 до 200. Выйдет 75 и 25 человек, обрыв вместо колокола. Или возьмём столбики по одному сантиметру, и получим случайную расчёску. Данные при этом одни, никто не соврал, а форма разная. Полезная привычка: увидели график с подозрительно удобной формой, спросите, какой ширины столбики и кто их резал.

Один горб ещё ничего не доказывает. Казалось бы, если смешать в одной толпе мужчин и женщин, у гистограммы роста вырастет два горба, как у верблюда. Такой пример кочует по учебникам. В 2002 году три американских автора решили это проверить и выяснили, что второго горба нет. Для него средние должны расходиться больше чем на два стандартных отклонения, а у людей не дотягивает. Возьмём условно мужчин 176 и женщин 163 при отклонении 7: разница 13, нужно больше 14. Выходит один полный горб. Так что толпа может быть очевидной смесью, а гистограмма смолчит. О половой принадлежности среднего человека из первой главы мы уже немного знаем.

Колокол не везде. Правило 68–95–99,7 годится только для колокола. Пусть у вас средняя зарплата 100 тысяч с от-

клонением 100 тысяч. По правилу выходит, что шестая часть зарплат отрицательная, даже если такое и бывает, то не в таких количествах. Зарплаты, тиражи и частоты слов колоколом не звонят вообще, и следующая глава целиком про это.

Как гистограмма работает в настоящей науке

Самое ценное в гистограмме то, что она умеет ловить жуликов, даже когда никто из них не попался. Одним из первых этим приёмом воспользовался наш знакомый Кетле. Он взял данные о росте французских новобранцев. В службу брали от 62 дюймов, то есть примерно от 157 сантиметров. Колокол был хорош, но как раз у этой отметки в нём зияла прореха: чуть выше людей подозрительно мало, чуть ниже подозрительно много. Такой формы природа сама не делает. Кетле заключил, что часть молодых французов на замере чудесным образом уменьшалась в росте. То ли сутулились, то ли проставлялись замерщику, этого уже не узнать. Зато учёный вывел их на чистую воду, не зная ни одной фамилии.

Прошло больше века, и метод сработал там же. В 2011 году блог Freakonomics показал гистограмму баллов польского выпускного экзамена по родному языку. Проходной балл там около 21. Именно в этом месте кривая проваливается и взлетает пиком. Проверяющие здесь, похоже, притягивали бедолаг за уши к тройке. Притом никто не листал работы. Это уже целый жанр: статистик не видит нарушителей, он видит вмятину на кривой.

А настоящая польза ближе к вам, чем кажется. Возьмите бланк из лаборатории с графой «норма». Эта графа обыч-

но строится так: обследуют здоровых людей и берут серединные 95% их значений. Значит, 5% здоровых людей выходят за норму по определению. Это не ошибка и не болезнь, а так проведена граница. Теперь сдайте двадцать анализов. Если считать их независимыми (на деле они связаны, поэтому это оценка), то все двадцать в норме окажутся с вероятностью 0,95 в двадцатой степени, примерно 36%. Значит, в 64% случаев совершенно здоровый человек получит хотя бы один показатель с пометкой «вне нормы». Старая врачебная шутка о том, что здоровый человек это недообследованный больной, имеет строгую математическую основу. Поэтому с одной пометкой идут к врачу, а не к нотариусу: он смотрит на картину целиком.

Рассматривать толпу целиком только на вид просто. Статистики говорят: прежде чем посчитать любое среднее, постройте гистограмму. Это та же старшина на плацу, только без мата и с учёной степенью.

Задачи

1. Столбики из пельменей. На свадьбе 50 гостей. От 0 до 9 пельменей съели 5 человек, от 10 до 19 съели 20, от 20 до 29 съели 15, от 30 до 39 съели 10. Какая доля гостей в самом высоком столбике? Какая доля съела 30 и больше?

Ответ: 20 : 50 = 40% и 10 : 50 = 20%. Эти 20% обычно и сидят ближе к кухне.

2. Жених выбирает невесту. Рост женщин условно подчиняется колоколу со средним 165 и отклонением 6 сантиметров. Какая доля женщин лежит между 159 и 171? Какая выше 177?

Ответ: примерно 68%, это одно отклонение в каждую сторону. Выше 177 около 2,5%, это уже два отклонения. Придётся жениху расти самому.

3. Здоровый больной. Здоровый человек сдал пять независимых анализов, где в норму входят 95% здоровых. С какой вероятностью хоть один выйдет за норму?

Ответ: все пять в норме с вероятностью 0,95 в пятой степени, то есть около 77%. Значит, примерно 23%. С каждым новым анализом здоровье тает, хотя человек тот же.

4. Верблюд. В посёлке 50 шахтёров получают по 100 тысяч и 50 учителей по 30 тысяч. Найдите среднюю зарплату. Как выглядит гистограмма?

Ответ: среднее 65 тысяч. Гистограмма из двух столби-

ков по краям и пустоты посередине. На месте среднего там никого нет, учителя и шахтёры на нём не встречаются.

5. Вмятина на шкале. На школьном экзамене с порогом 30 баллов ученики набрали так: 28 баллов у четырёх, 29 у одного, 30 у двадцати шести, 31 у пятерых, 32 у четырёх. Сколько человек на отметке 30, похоже, попали туда не своим умом?

Ответ: у соседних отметок по 4–5 человек, так что лишних около двадцати. Что такое материнский инстинкт учителя, гистограмма не знает, она видит только столбик.



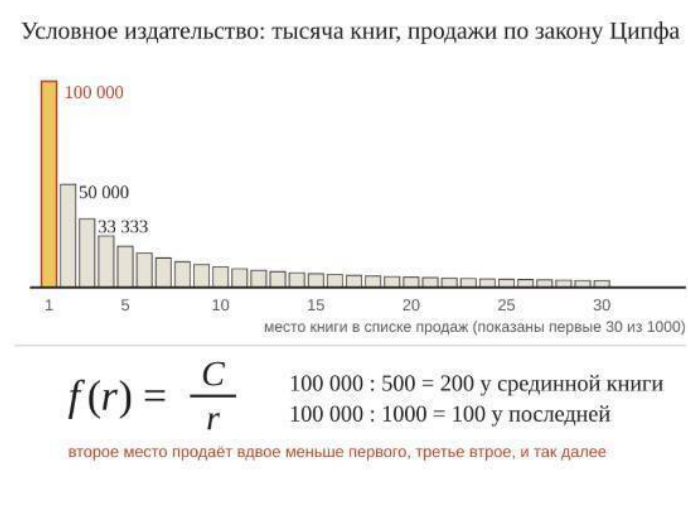
Глава 5. Один бестселлер и тысяча книг без читателей

Молодой автор приходит в издательство и спрашивает, сколько в среднем продаёт книга. Редактор берёт калькулятор: тысяча названий в каталоге, чуть больше 700 тысяч проданных экземпляров, стало быть, примерно по 750 на книгу. Автор радуется и бежит покупать новую ручку. Через год в отчёте у него стоит 200, и он обижается на читателя, жену и министерство культуры. Редактор при этом не соврал, а автор оказался на самом деле очень и очень типичным.

Здравый смысл тут подводит. Он насмотрелся на новобранцев из прошлой главы и думает, что мир устроен так же: большинство около среднего, герои и неудачники по краям. Для роста это верно. Самый высокий человек в истории примерно в полтора раза выше среднего. А теперь представьте колонну новобранцев, в которой последний боец достаёт головой до луны. С книгами, деньгами и словами дело обстоит именно так. Такую колонну старшине в строй не поставить, тут нужна другая картинка.

Закон для стенографистов

Расставим нашу условную тысячу книг по продажам, от первой до последней. Пусть первая продала 100 тысяч, вторая вдвое меньше, третья втрое, десятая в десять раз. Такая зависимость называется законом Ципфа, и в ней помещается вся несправедливость литературного процесса.

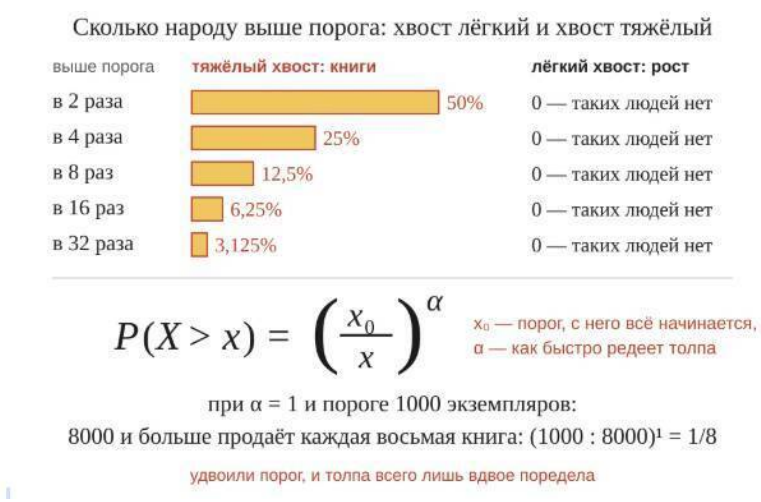


Буква r здесь занимаемое место, C продажи первого места. Теперь смотрим, что вышло. Всего продано около 750 тысяч книг. Первое место забрало 13%, первая десятка 39%, первая сотня 69%. Первые 200 книг, то есть 20% каталога,

держат 79% продаж. А нижняя половина каталога добирает 9% на всех пятьсот. Средние продажи составляют 750 экземпляров, медиана 200. Так мы и получили нашего обиженного автора.

Закон придумали вовсе не для книг и не статистики. Его заметили в словах. В 1916 году французский стенографист Жан-Батист Эсту считал частоты слов, чтобы улучшить стенографию, и увидел эту закономерность. В 1930-х её рассмотрел и прославил американский лингвист Джордж Ципф, который честно признавался, что открытие не его. Это его не спасло: закон всё равно назвали в его честь. Проверим на большом английском корпусе Брауна, в миллион слов. Самое частое слово the встречается 69 971 раз, второе, of, 36 411 раз, третье, and, 28 852 раза. Закон обещает примерно 35 и 23 тысячи. Не точно, но для шаманской формулы из одной дроби прилично. А половину всего корпуса составляют всего 135 слов. Выходит, что в языке царит тот же дикий капитализм, что и на книжном рынке. Предлоги и союзы жируют, а слово «палимпсест» за всю жизнь не зарабатывает на компот.

Теперь главное. В прошлой главе колокол у краёв обрывался почти мгновенно: такие края зовут лёгкими хвостами. Здесь всё иначе. Ведь что такое выброс для колокола? Беда и позор. А для книг выброс это и есть вся история. Такие края называются тяжёлыми хвостами.



Разницу между ними трудно преувеличить. Возьмём тысячу человек в зале со средним ростом 176 и приведём самого высокого человека в истории, ростом около 272 сантиметров. Среднее подрастёт на 0,1 сантиметра. Теперь пусть у каждого в зале по миллиону рублей, и приведём человека со 100 миллиардами. Среднее подскочит с миллиона примерно до 101 миллиона. При колоколе один человек ничего не решает. В тяжёлом хвосте один человек может оказаться важнее всего зала. Нассим Талеб в книге о чёрном лебеде назвал эти два мира Среднестаном и Крайнестаном, и разные вопросы там решаются по-разному.

Где закон даёт осечку

Не всё на свете Ципф. Формула так красива, что её находили везде, где график спускается похожим образом. В 2009 году Аарон Клаусет с коллегами взял 24 набора данных, про которые говорили, что они подчиняются степенному закону, и проверил их строгими методами. Хорошо сошёлся один набор, частоты слов, ещё 10 с оговорками. Ещё 10 лучше описывались степенным законом с обрезанным концом, а три не подошли вообще. Причём среди этих трёх оказалось богатство, главный пример почти во всех рассказах о тяжёлых хвостах.

Российские города. В учебниках ципфовский закон любят показывать на городах. Попробуем дома. На начало 2025 года в Москве 13,3 миллиона человек. Закон обещает на втором месте 6,6 миллиона, и Петербург с 5,7 ещё как-то справляется. На третьем месте должно быть 4,4 миллиона. Новосибирск даёт 1,6. На четвёртом должно быть 3,3, а Екатеринбург даёт 1,5. Либо между Петербургом и Новосибирском потерялись три города-миллионера, либо у нас свои законы географии, что вообще-то не новость.

Вирусная цифра. Тяжёлые хвосты так удивительны, что в любую дикость про них верят сразу. После антимонопольного суда над крупными американскими издателями по сети разошлось: 50% книг продают меньше дюжины экземпля-

ров. Писатели рыдали и пересылали это друг другу. Затем Кристен Маклин из BookScan попробовала получить такую цифру на настоящих данных и не смогла. По её подсчёту для новинок десятки крупнейших издателей меньше дюжины продаёт 14,7% книг, а меньше тысячи 51,4%. Хвост там тот же, только цифра не та была совсем. Писатели не прекратили рыдать, но перешли на более достоверную цифру.

Среднее без дна. При колоколе выборка быстро даёт надёжное среднее. При тяжёлом хвосте ответ зависит от того, попался ли в неё гигант. Не попался, и вы уверены, что книги продаются плохо. Попался, и среднее подскочило вдесятеро. А при альфа не больше 1, как в нашей картинке, среднего не существует даже теоретически: оно уходит в бесконечность, и всякий новый миллиардер его подталкивает. Таким толпам первые три главы этой книги не помощники.

Как тяжёлые хвосты работают в настоящей науке

Итальянец Вильфредо Парето заметил, что около 80% земли в Италии принадлежит примерно 20% населения. В 1941 году специалист по качеству Джозеф Джуран увидел ту же картину в браке на производстве: большая часть сбоев идёт от небольшого числа причин. Так возникло правило 80/20, которое сегодня продают на тренингах за большие деньги. Заметьте, наша формула из одной дроби на тысяче книг сама дала 79% на 20%. Само число 80 тут священного смысла не имеет, но идея дельная. Когда хвост тяжёлый, лечить надо не среднего, а самого толстого.

Хвосты пригодились и дома, в филологии, откуда всё началось. Раз 135 слов дают половину английского текста, то не все слова нужно учить с одинаковым рвением. На частотных словарях стоят учебники иностранного, стенография и поиск в сети. Зато слова из дальнего конца хвоста создают смысл и проблемы для переводчиков. Всё, как в жизни: работают одни, а интересны другие.

А теперь вернёмся к обиженному автору. У него остался главный вопрос: бестселлер лучше его книги или просто удачливее? Это удалось проверить опытом. В 2006 году в журнале Science вышла работа Мэттью Салганика, Питера Доддса и Дункана Уоттса. Их 14 341 участник слушали и

скачивали 48 песен неизвестных групп. Часть участников не знала чужих оценок. Остальных разбили на восемь параллельных миров, где видно было, что уже скачали другие в этом же мире. Песни везде одни и те же, старт одинаковый. Результат для писателя одновременно утешителен и ужасен. Чужое мнение усилило неравенство и сделало успех непредсказуемым: одна и та же песня в разных мирах жила совсем по-разному. Качество кое-что решало: лучшие песни редко проваливались, худшие редко взлетали. Но всё остальное могло случиться.

Так что напишите хорошую книгу, она повышает шансы. Бестселлера она не обещает, а если он случится у соседа, сосед вряд ли намного гениальнее вас. Это честное утешение, и за него статистику можно любить.

На этом заканчивается первая часть. Мы научились описывать толпу, где её середина, как широко она расселась, какой формы у неё горб и что тащится за ней хвостом. Нас ни разу не затронул вопрос, откуда она такая взялась. Для этого нужен случай.

Задачи

1. Лестница славы. По закону Ципфа первая книга продала 60 тысяч. Сколько продала третья, шестая? На каком месте стоит книга со 100 проданными экземплярами?

Ответ: 20 тысяч, 10 тысяч, на 600-м месте. Тем, кто на 600-м, найти некому благодарность, кроме мамы.

2. Клуб авторов. Девять авторов продали по 100 книг, десятый 99 100. Найдите среднее и медиану. Что напишут в рекламе клуба?

Ответ: среднее 100 000 : 10 = 10 тысяч, медиана 100. В рекламе напишут про средний тираж в десять тысяч. Девять членов клуба будут кивать, потому что им взносы платить.

3. Слова по себестоимости. В условном языке самое частое слово занимает 4% текста. Сколько раз в книге из 50 тысяч слов встретится первое и десятое по частоте слово?

Ответ: первое 2000 раз, десятое 200 раз. Слово «палимпсест» встретится один раз, да и то если автор захочет произвести впечатление.

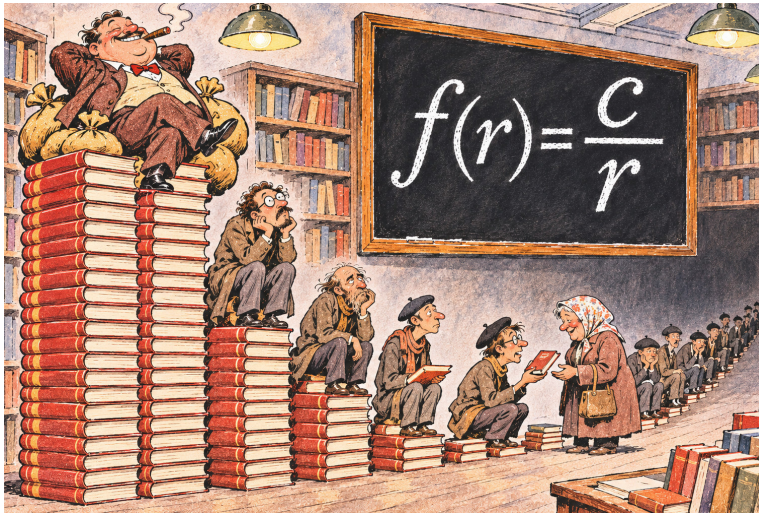
4. Таблица удвоений. Берём тяжёлый хвост с альфа = 1 и порогом в 1000 продаж. Какая доля книг продаёт 5000 и больше? А 100 тысяч?

Ответ: 1000 : 5000 = 20%, 1000 : 100 000 = 1%. У новобранца шанс быть в сто раз выше среднего равен нулю. На

книжном рынке один процент. Вот почему писатели всё ещё пишут.

5. Ципф едет по России. В Москве 13,3 миллиона жителей. Сколько должно жить в пятом по величине городе по закону? На начало 2025 года пятое место занимает Казань с 1,33 миллиона. Во сколько раз она не дотягивает?

Ответ: 13,3 : 5 = 2,66 миллиона, Казань меньше вдвое. Претензии направляйте Ципфу, а не Казани.



Глава 6. Дождь с вероятностью 70% и двое с одним днём рождения

Синоптик обещает дождь с вероятностью 70%. Вы берёте зонт, целый день таскаете его по жаре, дождя нет, и вечером вы материте синоптика словами, которых нет в словаре. В полной мере заслуженно? А что, собственно, он обещал? Что дождь пойдёт 70% времени? Что намокнет 70% города? Что семь синоптиков из десяти за дождь, а трое против? Или он просто немного уверен?

Здравый смысл и тут подсказывает неплохой ответ, если спросить его вежливо. Соберите дни, про которые синоптик говорил «70%», и посчитайте. Если из десяти таких дней дождь был в семи, прогноз честен. И три сухих дня с зонтом в руке входили в цену. Синоптик, обещавший 70%, не врёт, когда дождя нет. Он врёт, если дождь бывает каждый раз или раз в месяц. Потому проверить его можно только на серии дней. Один день ничего не доказывает и ничего не опровергает.

Именно здесь лежит пропасть между толпой из первой части книги и тем, что нас ждёт дальше. Толпа уже собрана, её можно посчитать. Завтрашний дождь ещё не случился, и про него приходится говорить уже на другом языке. Язык этот называется теорией вероятностей, и нам нужен его ми-

нимум, примерно на уровне «где тут туалет». Родился он, кстати, в игорном притоне: им занялись Паскаль и Ферма, когда разбирали задачу о дележе ставки в прерванной игре. Так что у науки вполне криминальное происхождение.

Три смысла одного числа

Вероятность обозначается буквой P и лежит между нулём и единицей. Ноль значит «не бывает», единица «обязательно». Строго говоря, математике всё равно, что она значит, лишь бы складывалась по правилам. А вот люди спорят до сих пор, и толкований здесь по меньшей мере три.

Первое толкование для игр. У кубика шесть граней, все равноправные, значит, на шестёрку приходится 1/6. Второе для повторяющихся случаев, им пользуется наш синоптик: просто доля таких дней среди похожих. Третье для случаев, которые не повторяются: «издатель возьмёт мою рукопись с вероятностью 10%». Серии тут нет, тут мера вашей уверенности. Первые два смысла укладываются в одну дробь, а третий мы оставим на следующую главу.

«70%» значит: семь дождливых дней из десяти таких дней



$$P(A) = \frac{m}{n}$$

m — сколько раз случилось то, что нам нужно,
n — сколько всего подходящих случаев

кубик: шестёрка — 1 грань из 6, P = 1/6 ≈ 0,17
погода: 7 дождливых из 10 похожих дней, P = 7/10 = 0,7
ни один отдельный день прогноз не проверяет, проверяет только серия

Путаница тут не выдумана. В 2005 году Герд Гигеренцер с коллегами опросил прохожих в Амстердаме, Афинах, Берлине, Милане и Нью-Йорке, что значит «вероятность дождя 30%». Законный метеорологический ответ, три дня из десяти похожих, выбрало большинство только в Нью-Йорке. Европейцы чаще считали, что дождь будет идти 30% времени или над 30% города, и все оставались при своём, в том числе в мокрых штанах. Вывод авторов полезен всем, кто слышит проценты: всегда спрашивайте «процент от чего».

Теперь проверим, насколько надёжно наша интуиция про вероятности. На свадьбе двое гостей выясняют, что у них один день рождения, и пьют за перст судьбы. Сколько гостей нужно, чтобы такое встречалось хотя бы в половине свадеб?

Здравый смысл называет обычно несколько сотен, потому что дней в году 365. Правильный ответ: 23.



$$P(n) = 1 - \frac{365 \cdot 364 \cdot 363 \cdot \ldots \cdot (365 - n + 1)}{365^n}$$

дробь: у всех дни разные. Вычли её из единицы: совпала хоть одна пара

Считать проще наоборот, вероятность того, что у всех дни рождения разные. Второму гостю доступно 364 дня из 365, третьему 363, и так далее. Перемножаем и вычитаем из единицы. Для 23 человек получаем 50,7%, для 41 уже 90%, для 70 совпадение почти обязательно.

Отчего так быстро? Наша голова ставит вопрос не так. Она спрашивает: кто-то родился в мой день? Для этого и вправду нужно около 253 человек. Но у нас не один участ-

ник с одним днём, а любая пара. Пар из 23 человек можно собрать 253, потому что $23 \cdot 22 : 2 = 253$. Число возможных совпадений растёт как квадрат числа людей, и этот смысл стоит запомнить надолго. Тот, кто удивляется совпадению, обычно считает один шанс. Тот, кто не удивляется, считает все пары.

Три любимых заблуждения

Динозавр пополам. Старый анекдот: какова вероятность встретить на улице динозавра? Пятьдесят на пятьдесят: либо встретишь, либо нет. Наша дробь работает только тогда, когда все случаи равноправны. Грани кубика равноправны, динозавр и его отсутствие нет. Смеётесь? А потом тот же человек говорит про операцию, выборы или бизнес: ну там фифти-фифти, или получится, или нет. Тот же динозавр, только уже в галстуке.

Снаряд дважды в одну воронку не падает. В 1913 году в Монте-Карло на рулетке чёрное выпало 26 раз подряд. После десятого чёрного публика поняла: теперь-то красное обязано выпасть, справедливость требует. Ставки росли с каждым оборотом, состояния переходили к казино. Но у колеса нет памяти, и оно никому ничего не должно. А русская солдатская статистика давно придумала для этого свою формулу: снаряд дважды в одну воронку не падает. У неё та же болезнь. Расчёту орудия нет дела до вашей воронки, стало быть, туда падает по обычным расценкам. Наука зовёт это ошибкой игрока. Независимые события не ведут бухгалтерии.

Мишень пририсована после выстрела. Вы подумали про старого друга, и он позвонил. Какова была вероятность? Очень малая, говорите вы и делаете выводы о тонких связях.

Но до звонка вы не заказывали именно этого друга и именно этот час. Годилось бы и другое: одноклассница на улице, песня по радио, сон об экзамене. Так стрелок сначала палит по стене сарая, а потом рисует мишень вокруг дырки и получает приз за меткость. Вероятность нужно считать до выстрела, а не после. К этой теме мы ещё вернёмся, когда дойдём до мёртвого лосося.

Как совпадения работают в настоящей науке

Задачу о днях рождения проверила жизнь. В составе футбольной команды на чемпионате мира как раз 23 человека. Би-би-си в 2014 году проверила 32 заявки и нашла общий день рождения в 16 командах. Ровно половина, как в учебнике, и болельщики ни о чём не догадывались.

Та же математика разоблачает сенсации. В сентябре 2009 года болгарская лотерея два тиража подряд, 6 и 10 числа, выдала одни и те же шесть чисел: 4, 15, 23, 24, 35, 42. Министр спорта назначил расследование, газеты заговорили о махинациях. Заданная шестёрка из 49 чисел выпадает с вероятностью 1 из 13 983 816. Но действует наше правило пар. Как посчитал статистик Дэвид Хэнд, чтобы повтор в одной лотерее стал вероятнее обратного, достаточно 4404 тиражей, а это меньше 43 лет при двух тиражах в неделю. А лотерей на свете много. Невероятно тут только то, что тиражи шли подряд, но такое попадает в газеты, а повтор через семь лет никто и не заметит.

Математик Джон Литлвуд закончил про это раз и навсегда. Назовём чудом то, что имеет для человека особое значение и случается с вероятностью 1 на миллион. Восемь часов в день вы бодрствуете и замечаете примерно по событию в секунду. За месяц набегает около миллиона событий. Зна-

чит, чудо с вами случается примерно раз в месяц. Эту мысль физик Фримен Дайсон назвал законом чудес Литлвуда. Церковь тут ни при чём, закон про другое: при большом числе попыток невероятное обязательно.

А вот тут не до шуток. В 2001 году Кэтрин Тройер, аналитик криминалистической лаборатории Аризоны, обнаружила в базе ДНК двух преступников с поразительно похожими профилями. Потом проверили все около 65 тысяч профилей и нашли 122 пары, совпадающие по 9 участкам из 13. А в судах при этом звучали цифры страшно внушительные, одна на миллиарды. Скандал? Стивен Левитт посчитал, и вышло по тому же правилу. В базе около двух миллиардов пар и множество способов выбрать девять участков, так что около сотни совпадений ожидалось чисто случайно. То есть судебные оценки не обязательно лживы, просто они отвечают на вопрос об одной паре. Когда подозреваемого ищут прогоном по базе, совпадение доказывает уже гораздо меньше, и суду стоит это понимать. Та же задача о днях рождения стоит за расчётом длины криптографических ключей: взломщики так и называют поиск совпадений атакой дней рождения.

Запомним одну вещь на будущее. В суде всё теперь держится на том, какой именно вопрос мы задали к вероятности. Следующая глава как раз об этом, и ценой ошибки там будет чужая жизнь.

Задачи

1. Зонт. Синоптик десять раз подряд обещал дождь с вероятностью 70%. Сколько дождливых дней в среднем получится, если он честен? Какова вероятность, что дождь пойдёт все десять дней?

Ответ: в среднем семь дней. Все десять: 0,7 в десятой степени, около 2,8%. В остальные три дня синоптика будут ругать по расписанию.

2. Школьный класс. В классе 30 человек. Сколько в нём пар учеников? Больше или меньше половины вероятность совпадения дней рождения?

Ответ: пар 30 · 29 : 2 = 435. Вероятность около 71%, так что на пари с одноклассниками можно заработать на обед.

3. Мой день. В конторе 30 ваших коллег. Какова вероятность, что хоть один из них родился именно в ваш день?

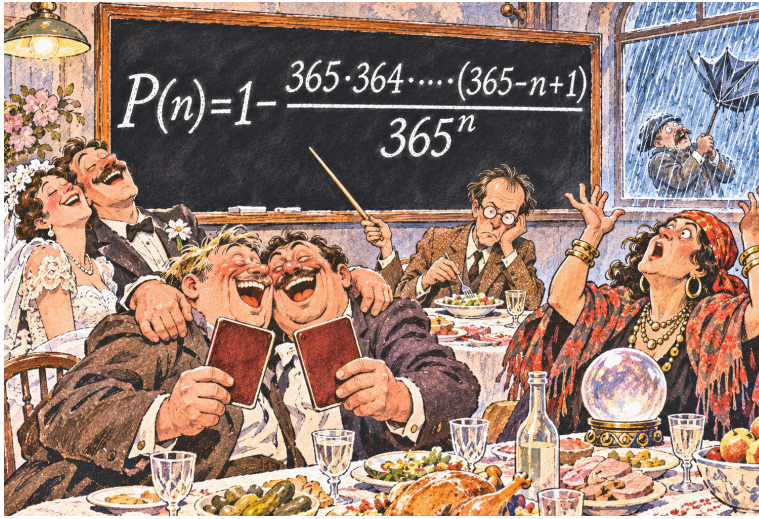
Ответ: 1 - (364/365) в тридцатой степени, около 8%. Только восемь процентов, что кто-то разделит с вами торт. Ваша пара там одна из тридцати, потому так мало.

4. Справедливость колеса. На европейской рулетке 18 чёрных ячеек, 18 красных и зеро. Чёрное уже выпало 26 раз подряд. Какова вероятность чёрного в следующий раз?

Ответ: 18/37, около 48,6%. Как и двадцать шесть оборотов назад. Колесо не читало эту главу и про серию не знает.

5. Вещий сон. В городе 100 тысяч жителей. Каждому каждую ночь что-то снится, и пусть сон случайно сбывается наутро с вероятностью 1 на 10 тысяч. Сколько вещих снов будет в городе за ночь?

Ответ: 100 000 · 1/10 000 = 10. Десять человек каждое утро рассказывают о своём даре. Девяносто девять тысяч девятьсот девяносто ни о чём таком не рассказывают, их в передачи не зовут.



Глава 7. Положительный тест у здорового человека

Вы прошли обследование просто так, за компанию, и вдруг тест на редкую болезнь положительный. Болезнь условная, но у вас она уже звучит как приговор. Тест точный: больных угадывает в 90% случаев, здоровых ошибочно помечает лишь в 9%. Вы считаете в уме и получаете: примерно 90% за то, что вы больны. Дальше пишется завещание, жена узнаёт, что такое гараж, тёща надевает лучшее платье.

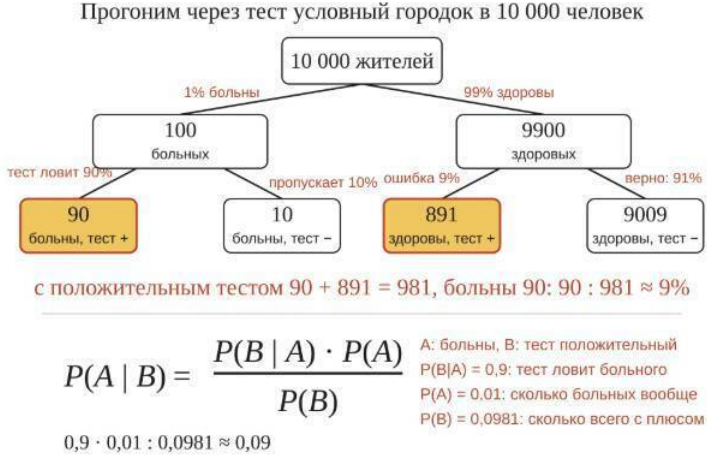
Здесь стоит спросить одну вещь, которую паника спрашивать не даёт: а насколько эта болезнь вообще редкая? Пусть ей болеет один человек из ста. Здравый смысл с процентами не дружит, зато отлично умеет считать людей. Соберём целый городок в десять тысяч жителей и прогоним их через этот тест.

Больных в городке 100 человек. Тест пометит 90% из них, то есть 90 человек. Здоровых 9900. Тест ошибочно пометит 9% из них, то есть 891 человека. Итого положительный тест получили 981 человек, и больны из них 90. Вы в этой очереди стоите среди почти тысячи человек, и девять из десяти там просто за компанию. Вероятность того, что вы больны, не 90%, а 90 : 981, то есть около 9%. Завещание пока можно убрать в стол, а врачу позвонить обязательно: девять

процентов это уже в девять раз больше, чем было утром, и проверяться надо. Только уже без тёщи в платье.

Формула преподобного

То, что мы сейчас посчитали на людях, в буквах называется теоремой Байеса. Надо только привыкнуть к одному значку. Запись P(A | B) читается так: вероятность A при условии, что B уже случилось. Палочка значит «среди тех, у кого». Среди тех, у кого тест положительный, сколько больных?



Ужас в начале главы происходил из того, что мы спутали две вещи. 90% это P(тест положителен | болен). А вас волнует P(болен | тест положителен). Это разные числа, и од-

но не выводится из другого, пока не знаешь, насколько болезнь распространена вообще. Английский преподобный отделил мух от котлет: вторая дробь выражается через первую и через долю больных. Та доля P(A), которая была известна до теста, называется априорной вероятностью. То, что получилось после, апостериорной. Слова пугающие, а смысл домашний: что думал до и что думаешь после.

Здесь как раз живёт третий смысл вероятности из прошлой главы. Байес позволяет менять уверенность под давлением фактов. Сдали второй такой же независимый тест, и он снова положительный? Теперь ваши априорные не 1%, а уже 9%. Считаем по той же формуле и получаем около 50%. Третий даст уже около 91%. Умный человек не прыгает от «верю» к «не верю», он доворачивает своё мнение по чуть-чуть, как гайку. Так делает байесовский подход в науке, да и хороший следователь.

Когда вопрос переворачивают

Попутать P(A | B) и P(B | A) очень легко, и на примере очень смешно. Какова вероятность, что папа римский мужчина? Единица. А что случайный мужчина окажется папой римским? Примерно один к четырём миллиардам, и место уже занято.



Тот же самый фокус в суде называется ошибкой прокурора. На неё мы и посмотрим, и дальше шуток не будет.

Салли Кларк, английский юрист, в 1996 и 1998 годах по-

теряла двух маленьких сыновей, обоих на первых месяцах жизни. Её обвинили в убийстве. На процессе 1999 года педиатр Рой Мидоу заявил, что в такой семье внезапная детская смерть случается примерно в одном случае из 8543. А две подряд, значит, 8543 в квадрате: один шанс из 73 миллионов. Суд присяжных большинством в 10 голосов против 2 признал её виновной, приговор был пожизненным.

В этом рассуждении две грубые ошибки. Первая: возводить в квадрат можно только независимые события. А в одной семье смерти младенцев могут быть связаны общими генами и общими условиями, и после первого такого случая шанс второго оценивают примерно в 1 к 100. Вторая и главная: присяжные слышали «такое почти не бывает» и делали вывод «значит, виновна». Это тот самый перевёрнутый вопрос. Двойное убийство собственных младенцев тоже вещь чрезвычайно редкая. Сравнивать надо не редкость с нулём, а одну редкость с другой. Математик Рэй Хилл внёс обе поправки и заключил: если брать только сам факт двух необъяснённых смертей, вероятность её вины могла составлять всего около десяти процентов.

В октябре 2001 года Королевское статистическое общество публично заявило, что цифра «1 из 73 миллионов» не имеет статистического основания. В январе 2003 года приговор отменили по второй апелляции, главным образом потому, что всплыли скрытые патологом результаты микробиологического анализа. Больше трёх лет она провела в тюрьме

как убийца своих детей. Салли Кларк не оправилась от пережитого и умерла в 2007 году. Вот что бывает, когда присяжные не знают, что означает палочка в формуле.

Ну и последнее предупреждение, уже в другую сторону. Байес ничего не знает от себя, он выдаёт только то, что вложили. Начальную цифру нужно брать для своего случая. Проверяющийся за компанию и человек, которого врач отправил на тот же тест из-за симптомов, стоят в разных толпах. У второго та же положительная полоска значит совсем другое. Так что эта глава не повод отмахнуться от плохого анализа, нужна повторная проверка и врач.

Как Байес работает в настоящей науке

Томас Байес был пресвитерианским священником в Танбридж-Уэллсе, умер в 1761 году и своёго главного труда не напечатал. Рукопись разобрал и представил Королевскому обществу его друг, тоже священник, Ричард Прайс, в 1763 году. Так метод, который потом будут то боготворить, то проклинать поколения статистиков, вошёл в мир в посмертной публикации с помощью двух священников, что неплохо звучит для теории об уточнении своих верований.

В 1940 году в Блетчли-Парке Алан Тьюринг и Джек Гуд взламывали морскую «Энигму» с помощью собственной единицы измерения улик, бана и децибана. Улики складывались, пока не перевешивали сомнения. Гуд потом настаивал, что это по сути байесовский вывод, хотя нашлись и философы с другим мнением.

Настоящий триумф ждал в почтовом ящике. В августе 2002 года Пол Грэм опубликовал эссе «План против спама». Его фильтр поступал так, как давно поступают лингвисты. Он считал, как часто слово встречается в спаме и в обычной почте, и по Байесу обновлял уверенность слово за словом. В первом тесте на собственной почте автора он словил около 99,5% спама и не ошибся на нормальных письмах ни разу. Пресвитерианский священник из XVIII века вошёл в

каждый почтовый ящик на борьбу с рекламой средств для потенции.

Но самая важная практика сегодня другая. Герд Гигеренцер, тот самый исследователь дождя из прошлой главы, дал группе из 150 гинекологов похожую задачу про маммограмму: 1% больных, тест угадывает 80% больных, 9,6% ложных тревог. Верный ответ примерно 8%. До занятия справились 21% врачей. После часовой лекции, на которой он научил их просто считать людей вместо процентов, как мы в начале этой главы, справились 87%. Так что, если ваш врач или адвокат торжественно называет процент, спросите его: сколько это человек из тысячи? За лекцию врачи научились, а вы уже умеете.

Задачи

1. Прибор не пьёт. Из 1000 водителей на посту пьяных 20. Трубка ловит 95% пьяных и ошибочно помечает 5% трезвых. Трубка загорелась. Какова вероятность, что водитель пьян?

Ответ: пьяных поймано 19, трезвых ошибочно 49. Получаем 19 : 68, около 28%. Обидно, конечно, но инспектору этот расчёт лучше не предлагать.

2. Помада на воротнике. Жена до находки оценивала шанс измены мужа в 5%. При измене помада появляется в половине случаев, без измены в 5% (маршрутка, тёща, корпоратив). Что она должна думать после находки?

Ответ: $0{,}5 \cdot 0{,}05 : (0{,}5 \cdot 0{,}05 + 0{,}05 \cdot 0{,}95)$ примерно 34%. Сковородку можно взять, но пока без удара.

3. Такси. В городе 85% такси зелёные, 15% синие. Свидетель ночного ДТП видел синее такси, при этом цвета он различает верно в 80% случаев. Какова вероятность, что такси действительно синее?

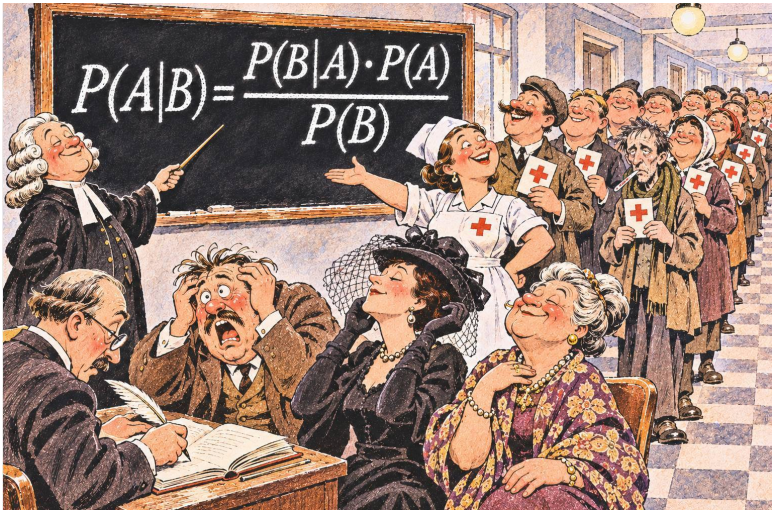
Ответ: из 100 машин синих верно назовут 12, зелёных ошибочно 17. Получаем 12 : 29, около 41%. Свидетель честен, а такси всё равно скорее зелёное.

4. Слово-предатель. Спама в почте половина. Слово «подарок» есть в 60% спама и в 2% обычных писем. Пришло письмо со словом «подарок». Это спам?

Ответ: 0,3 : (0,3 + 0,01) примерно 97%. Только тёща пишет про подарок искренне, и фильтр уводит её в корзину вместе со всеми.

5. Ошибка прокурора. На месте преступления найден след, который случайно совпадает с 1 человеком из миллиона. В базе города на 5 миллионов нашли Петровича с таким следом, других улик нет. Прокурор говорит: шанс невиновности 1 на миллион. Каков он на самом деле?

Ответ: в городе ожидается около пяти случайных обладателей следа и один преступник. Виновен Петрович с вероятностью 1/6, и невиновен с вероятностью около 83%. За старое он, конечно, ответит, но тут оправдан.


$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$

Глава 8. Прусские кони и опечатки

Главный редактор листает свежую книгу и вдруг багровеет: на сто семьдесят третьей странице три опечатки! На одной! В книге на 300 страниц их всего полтораста, в среднем половина опечатки на страницу. Редактор зовёт корректора: ты что, на сто семьдесят третьей заснул? Там диверсия, туда беги, всю перепроверь! Корректор, человек сухоносый и пьющий, отвечает, что такие страницы тут обязаны быть, и что таких страниц в книге должно быть около четырёх. Редактор выгоняет его из кабинета, но зря: корректор прав.

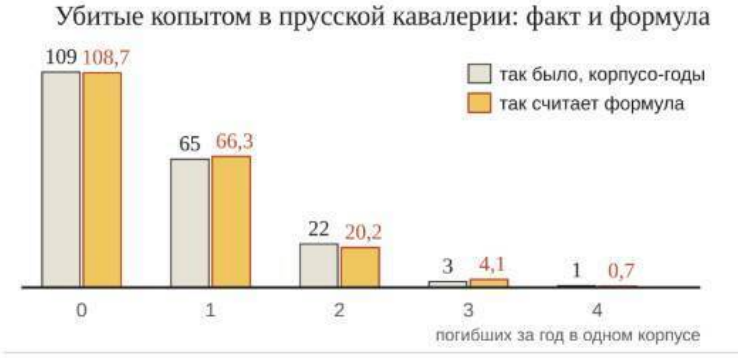
Здравый смысл думает про такие вещи примерно так: раз в среднем половина, значит, на каждой странице должно быть так и сяк, ну ноль, ну одна. Три на одной это уже кто-то виноват. На самом деле случай редко кладёт поровну. Он совсем не шариковец из первой главы. Случай любит кучки, дырки и пустыри.

Вопрос можно задать точнее. Букв в книге огромное множество, ошибка на каждой случается очень редко, и соседние ошибки про друг друга не знают. Три условия: шансов много, каждый мал, все независимы. Окажется ли при этих условиях картина одной и той же и у опечаток, и у смертей от конского копыта? Оказывается, да.

Метод имени копыта

В 1898 году в Лейпциге вышла книжка «Закон малых чисел». Её написал Ладислав Борткевич, уроженец Петербурга из польских дворян, успевший поработать и в германских университетах, и на российских железных дорогах. Великие умы искали в ту пору законы истории и нравственности. Борткевич взял другое: сколько прусских кавалеристов за год погибло от удара конским копытом. Без всякого почтения к высокому он сложил 200 корпусо-лет и 122 покойника, то есть по 0,61 на корпус в год. Каждый конь мог лягнуть в любой день, лягал очень редко, с соседним корпусом не совещался. Все три условия соблюдены.

Математика на этот случай готова была уже шестьдесят лет: в 1837 году француз Симеон Дени Пуассон вывел формулу для таких редких событий, и нужна она была ему, как и следовало ожидать, для судебных приговоров. Для неё нужно знать одно число, среднее количество событий. Греческая буква лямбда выглядит в ней грозно, но на деле это те же наши 0,61.



$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

λ — среднее число событий (лямбда),
k — сколько случилось на самом деле,
k! = 1 · 2 · … · k, e ≈ 2,718

λ = 0,61; P(0) = e в степени −0,61 ≈ 0,543; 200 · 0,543 ≈ 109

Смотрите, как сошлось. В 109 случаях из 200 кони за год никого не убили, формула обещала 108,7. По одному убитому: 65 случаев, обещано 66,3. Четыре убитых за год случились один раз, и формула насчитала 0,6 такого случая. Объяснять этот чёрный год плохим начальством или происками врагов не нужно. Это просто очень большая неудача, и в таблице она предусмотрена.

Теперь вернёмся к нашему главному редактору с его тремя опечатками. Подставляем лямбда = 0,5 для 300 страниц. Совсем чистыми окажутся около 182 страниц, с одной опечаткой около 91, с двумя около 23, с тремя около четырёх.

Заметьте, больше половины книги без единого греха, и никто за это корректора премию не выдаст. Четыре страницы с тремя опечатками не диверсия, а расписание.

У такого распределения есть удобное свойство. Его разброс совсем не надо измерять отдельно, он задан самим средним.



Эта картинка стоит целой главы. Я попросил компьютер бросить 225 точек на поле наугад, без всяких намерений. Глаз тут же находит сгустки и пустоты и начинает объяснять: почему тут так густо? Кто-то постарался? Никто не старался. Настоящая случайность всегда выглядит подозрительно,

потому что в ней обязаны быть скопления. Ровненькая сетка, по одной точке в клетке, вот что было бы настоящим заговором.

Когда конь даёт осечку

Формула держится на трёх ногах, и каждую можно подшибить.

События знают друг о друге. Если наборщик до обеда трезв, а после обеда нет, опечатки будут кучковаться страницами. Больной гриппом заражает соседей, и провалившийся студент манит за собой остальных. Тогда пустых клеток и переполненных становится больше, чем обещает формула, и дисперсия обгоняет среднее. И вот это уже повод искать причину.

Среднее плывёт. Пусть придорожный кафе принимает в среднем десять человек в час. В обед оно видит тридцать, в три ночи одного. Формула с лямбда = 10 будет плоха и для обеда, и для ночи. Каждому часу нужна своя лямбда.

Границы рисуют после выстрела. На карте всегда найдётся деревня с подозрительным числом редких болезней. Если обвести её кружком уже после того, как увидел, любой случайный сгусток превратится в сенсацию. Это наша мишень из шестой главы. Скопление может быть настоящим, и проверять его надо, но посчитать при этом стоит, а не кричать.

Где это применяют всерьёз

Летом 1944 года на Лондон падали немецкие самолёты-снаряды Фау-1, и в городе часто говорили, что они ложатся кучно, будто некоторые кварталы выбраны нарочно. Для людей под обстрелом это был вопрос совсем не отвлечённый. В 1946 году английский актуарий Р. Д. Кларк разделил 144 квадратных километра юга Лондона на 576 квадратов и подсчитал 537 попаданий. Без единого попадания остались 229 квадратов, формула Пуассона предсказывала 226,7. С одним попаданием оказался 211 квадратов при расчёте 211,4. С четырьмя 7 квадратов при расчёте 7,1. Скопления были настоящими, люди их не выдумали, но именно такие скопления и даёт случайное падение. Кларк не нашёл в этих данных признаков прицеливания по отдельным кварталам внутри обследованного района.

В 1907 году сотрудник пивоварни Гиннесса Уильям Госсет, подписывавшийся псевдонимом Стьюдент, считал под микроскопом дрожжевые клетки на сетке из 400 квадратиков. От этого подсчёта зависело пиво, а значит, дело было важное. Распределение и тут оказалось пуассоновским, а небольшое отклонение в сторону многолюдных квадратиков он объяснил: клетки почкуются и слипаются, то есть нарушают независимость. Вот вам наша первая нога прямо под микроскопом. С самим Госсетом мы ещё встретимся в два-

дцать первой главе.

В те же годы датский инженер Агнер Эрланг, работавший в Копенгагенской телефонной компании и лично лазивший в уличные люки, показал, что звонки на станцию подчиняются тому же закону. В 1909 году он опубликовал об этом статью. Отсюда пошла теория очередей, по которой сегодня считают, сколько нужно операторов в колл-центре, касс в магазине и серверов на сайте. Если вас пятнадцать минут уговаривают оставаться на линии, знайте: кто-то сознательно подставил в формулу маленькое число операторов.

А самая красивая история здесь о том, как формула не сошлась. В 1943 году Сальвадор Лурия и Макс Дельбрюк заражали колонии кишечной палочки вирусом и считали, сколько бактерий выжило. Если бы устойчивость появлялась в ответ на встречу с вирусом, то числа выживших от чашки к чашке шли бы по Пуассону, с дисперсией, равной среднему. Дисперсия оказалась намного больше среднего: то почти никого, то джекпот. Значит, мутации возникли заранее и случайно, ещё до всякого вируса, а потом размножились вместе с потомством. Нобелевская премия 1969 года досталась в том числе и за это. Сравнение дисперсии со средним, которое вы только что освоили за одну страницу, помогло установить, как устроена изменчивость живого.

Задачи

Задача 1. В книге 200 страниц и 100 опечаток. Сколько страниц окажется совсем чистыми, если опечатки легли случайно?

Ответ. лямбда = 0,5, вероятность чистой страницы e^-0,5 примерно 0,61, значит, около 121 страницы. Остальные 79 страниц делят между собой 100 опечаток, и автор, открыв любую из них, скажет, что корректора надо гнать. Гнать не за что: корректор работал ровно как случайный процесс, хотя платили ему как человеку.

Задача 2. Тёща звонит в среднем два раза в день, в случайные моменты. Как часто выпадает день совсем без звонка?

Ответ. e^-2 примерно 13,5 %, то есть примерно 49 дней в году. Считать эти дни знаком примирения не следует, как и знаком заговора. Это просто хвост распределения, и на следующий день среднее восстановит справедливость.

Задача 3. На перекрёстке случается в среднем одно ДТП в месяц. Какова вероятность, что в данном месяце их будет три? Что случится, если такой месяц наступит?

Ответ. Формула даёт около 6,1 %, то есть такой месяц бывает примерно раз в полтора года. Когда он настанет, мэр потребует найти виновного и уволит начальника дорожной службы. В следующем месяце ДТП будет ноль или одно, и

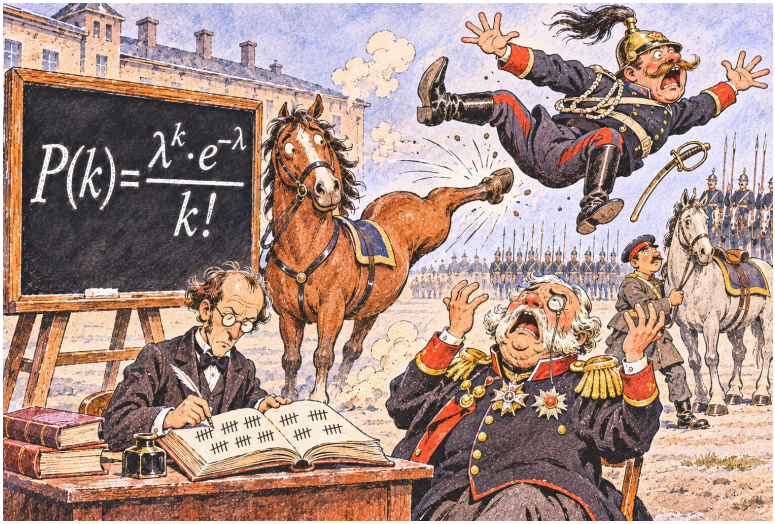
мэр запомнит, что железная рука работает.

Задача 4. Автор популярной книги получает от читателей в среднем одно письмо за десять дней. Какова вероятность, что за десять дней не придёт ни одного?

Ответ. лямбда = 1, вероятность e^-1 примерно 37 %. Больше чем в трети таких периодов автор будет заключать, что его разлюбила страна. В среднем страна его и прежде любила на 0,1 письма в день.

Задача 5. В тысяче деревень сообщения о снежном человеке приходят в среднем по 0,5 за год на деревню. В скольких деревнях за год наберётся четыре сообщения и больше?

Ответ. Примерно в двух. Этого хватит на документальный фильм, областную экспедицию и магазин сувениров. Снежного человека для такого результата не требуется, требуется только тысяча деревень.


$P(k)=\frac{\lambda^k \cdot e^{-\lambda}}{k!}$

Глава 9. Почему лучшие школы часто маленькие

Областная газета печатает рейтинг школ по среднему баллу за ЕГЭ. Пример условный, адресов не ищите. На первом месте школа в селе Верхние Кочки: девять выпускников, средний балл 81. Знаменитая городская гимназия с девятьюстами выпускниками, бассейном и директором в таких часах, что на них можно купить Верхние Кочки, стоит где-то посередине, с баллом 66.

Здравый смысл справляется с этим за минуту и плачет от умиления. В маленькой школе учитель знает каждого ребёнка и каждую его корову. Никто не теряется в толпе, печка трещит, за окном берёза. Большая школа же фабрика бездушия, где ребёнок в одиннадцатом классе всё ещё ищет столовую. Вывод просится сам: гимназию поделить на сто Верхних Кочек, бассейн засыпать, берёзу посадить. Министр областного образования уже собирает совещание.

На совещании никто не дочитывает газету до конца. А на последнем месте в рейтинге стоит школа в селе Нижние Кочки: семь выпускников, средний балл 49. Та же печка, та же берёза, та же корова.

Эта история условна только в названиях сёл. В конце девяностых в Америке заметили, что среди школ с лучшими

баллами маленьких подозрительно много, и пошло движение за разделение больших школ на малые. В него вложили миллиарды долларов. Фонд Билла и Мелинды Гейтс увлёкся этой идеей всерьёзно, и к 2001 году раздал образовательных грантов примерно на 1,7 миллиарда. Цифры вроде бы подтверждали: в Пенсильвании из 1662 школ в первую полсотню по баллам вошли шесть школ из полсотни самых маленьких, то есть 12 % вместо положенных трёх.

Затем статистики Ховард Уэйнер и Харрис Зверлинг сделали то, что не пришло в голову нашему министру. Они дочитали таблицу до конца. Среди пятидесяти худших школ Пенсильвании девять оказались из той же полсотни самых маленьких: 18 % вместо трёх. Маленькие школы толпятся и наверху, и внизу. Так лучше они или хуже? И почему миллиарды пошли в дело только по верхней строчке?

Закон, который знает каждый дурак

Швейцарский математик Якоб Бернулли великодушно уведомлял Лейбница в 1703 году: даже самый тупой человек без всякого обучения, каким-то чудным природным чутьём знает, что чем больше наблюдений, тем меньше опасность промахнуться. После чего потратил на доказательство этой очевидности около двадцати лет и назвал её золотой теоремой. Книгу «Искусство предположений» племянник выпустил в 1713 году, через восемь лет после смерти автора. Имя «закон больших чисел» дал в 1837 году Пуассон, наш знакомый по прусскому копыту.

Зачем двадцать лет, если дурак и так знает? Дурак знает, что среднее по большому числу надёжнее. Он не знает, стремится ли оно к истине или какому-то своему пределу, можно ли подобраться к истине сколь угодно близко и сколько наблюдений для этого нужно. Бернулли доказал, что можно. Точное же правило нашёл к 1730 году француз Абраам де Муавр, живший в Лондоне. Разброс среднего падает не вместе с числом наблюдений, а вместе с корнем из него. Эту небольшую формулу статистик Уэйнер назвал самым опасным уравнением в истории: не потому, что она взрывается, а потому, что очень дорого обходится тем, кто её не знает.



$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

s — разброс оценок одного ученика,
n — сколько учеников в выборке,
разброс среднего падает как корень из n

s = 15; n = 9: 15 / 3 = 5; n = 900: 15 / 30 = 0,5
в сто раз больше учеников, в десять раз точнее среднее

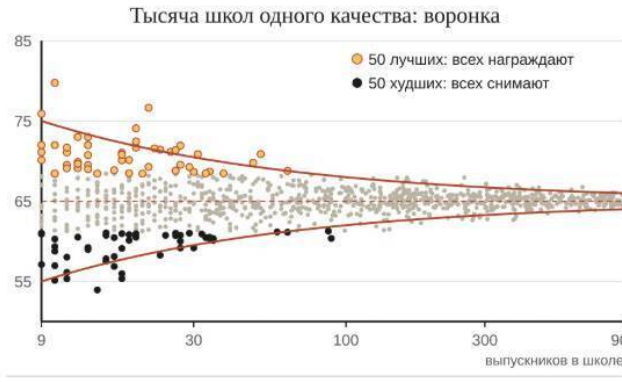
Здесь s это уже знакомое вам из третьей главы стандартное отклонение. Но теперь речь о двух разных разбросах: среди отдельных учеников и среди средних по целым школам. Возьмём условно, что отдельные дети пишут ЕГЭ с разбросом 15 баллов, а среднее везде 65. У верхнекоченских девяти среднее гуляет с разбросом 5, у девятисот гимназистов с разбросом 0,5. В пределах двух разбросов село будет показывать то 55, то 75 при неизменной печке и берёзе, а гимназия всегда от 64 до 66. Из девяти детей один выдающийся мальчик или один полный отморозок двигает среднее всего села на несколько баллов. В гимназии тот же отморозок растворяется без следа.

То же самое с Петровичем. Помните его пять банных за-

ходов: 2, 3, 3, 4 и 13 кружек? Один загул поднял среднее с трёх до пяти. Будь у нас не пять заходов, а сто, тот же загул добавил бы к среднему всего десятую долю кружки. Петрович, безусловно, не стал бы трезвее, но его среднее стало бы.

Отсюда же принципиально неприятное следствие. Точность стоит дорого. Чтобы уточнить среднее вдвое, наблюдений нужно вчетверо больше. Вдесятеро? В сто раз больше. Помещённый в войну в лагерь для интернированных в Дании математик Джон Керрих вместе с товарищем по заключению бросил монету 10 000 раз и получил 5067 орлов. Доля 50,67 % при ожидаемом разбросе полпроцента. Закон можно проверить и дома, нужна только монета и очень много времени.

А теперь главная картинка. Я попросил компьютер создать тысячу условных школ одинакового качества. Вообще одинакового, всем детям балл выпадает случайно вокруг 65. Отличается только число выпускников, от девяти до девятисот.



$$65 \pm 2 \cdot 15 / \sqrt{n}$$

школы разные только размером; разницы в качестве на картинке нет

Вот она, воронка. Из пятидесяти лучших школ 32 попали в пятую часть самых маленьких, где не больше двадцати выпускников. При равном счёте их было бы десять. Из пятидесяти худших маленьких оказались 31. А школы с двумя сотнями и больше выпускников, которых тут треть, не попали ни в верхнюю полсотню, ни в нижнюю. Ни одна. Они плотно стоят в середине, как средний класс на плакате. При этом министр проведёт совещание, вручит премию школе с баллом 80 на десять учеников и снимет директора школы с баллом 54 на четырнадцать. Разница между этими директорами в нашем опыте равна нулю, их вообще не существует.

Ответ на вопрос главы теперь прост. Лучшие школы часто

маленькие, потому что и худшие часто маленькие. Маленькие школы не лучше и не хуже, они просто дёргаются сильнее, как лодка на волне рядом с ледоколом.

Чего закон больших чисел не обещал

Он не возвращает долги. 18 августа 1913 года в казино Монте-Карло чёрное выпало 26 раз подряд. Шанс на такую серию примерно один на 137 миллионов. Уже после десятого чёрного уважаемая публика попёрла ставить на красное: по закону больших чисел должно же выравняться! Оставили на столе миллионы франков. Ни колесо, ни монета не ведут бухгалтерии. Закон работает не возвратом долга, а разбавлением. У Керриха над ровной половиной остался излишек в 67 орлов, и никто его не отдал. Он просто утонул в десяти тысячах бросков.

Он не чинит кривые весы. Закон убирает случайную ошибку и не трогает постоянную. Если в кружке Петровича не поллитра, а три четверти без пены, то через миллион заходов мы узнаем точное неверное число, до третьего знака. Большая выборка из не тех людей просто врёт очень уверенно. Как журнал с двумя миллионами анкет простудился на президентских выборах, мы увидим в одиннадцатой главе.

Он требует независимости. Корень из n в формуле стоит потому, что случайные отклонения у разных людей друг друга гасят. Если девятьсот гимназистов пишут в одном здании, где в этот год случилась утечка заданий или уволился хороший математик, то отклонения накладываются, и девятьсот человек дают не девятьсот свидетельств, а гораздо

меньше. Так же тысяча опрошенных из одного подъезда выражают мнение подъезда, а не тысячи.

Он может тянуть вечность. В пятой главе мы видели тяжёлые хвосты, где один бестселлер перевешивает тысячу книг без читателей. Там среднее может долго стоять, а потом подпрыгнуть от одного наблюдения. А в крайнем случае, например у распределения Коши, среднего не существует вообще: среднее из миллиона таких чисел не надёжнее одного. Закон не действует, потому что у него нет предмета.

И он не отменяет настоящих различий. Воронка не доказывает, что все маленькие школы средние. Вполне может быть гениальная учительница в Верхних Кочках. Воронка говорит, что один балл за один год этого не показывает. Хотите знать про учительницу, смотрите много лет подряд и сравнивайте с такими же детьми. В 2009 году Билл Гейтс сам написал в годовом письме фонда, что многие маленькие школы, куда они вложились, успеваемость сколько-нибудь заметно не подняли, хотя посещаемость и доля выпускников там часто были выше, чем у соседей. Маленькая школа может быть в чём-то лучше, только узнавать это по верхней строчке рейтинга нельзя.

Где это применяют всерьёз

Начнём с цены незнания. Примерно с 1150 года в Англии проводился суд над монетами, так называемый Trial of the Pyx, по имени ящика, куда откладывали образцы. Мастеру монетного двора разрешали небольшой допуск. На примере более поздней гинеи допуск составлял 0,28 грана на монету и, совершенно естественно для здравого смысла, 28 гранов на сотню монет. А случайные ошибки складываются по корню: на сотню хватило бы в десять раз меньше, 2,8 грана. Мастер, снимавший по чуточке с каждой монеты, спокойно проходил суд. Историк статистики Стивен Стиглер осторожно пишет, что есть некоторые свидетельства, что мастера так и делали. Близко шести веков, до работы Муавра, английская корона охраняла своё золото правилом, неверным в десять раз.

Затем карты болезней. Когда в США нанесли на карту округа с самой низкой смертностью от рака почки, оказалось, что это почти сплошь сельская глубинка. Объяснение готово: чистый воздух, простая еда, своё молоко. Но округа с самой высокой смертностью тоже оказались сельской глубинкой, и нередко тут же, по соседству. Объяснение тоже готово: бедность, табак, больница в трёх часах езды. А дело в арифметике. В округе из ста человек ноль смертей дают нижнюю строчку таблицы, а одна смерть верхнюю. Вот по-

чему перед тем, как искать причину болезни, эпидемиолог смотрит, сколько там вообще людей.

Затем психология. Амос Тверски и Даниэль Канеман спросили людей: есть два роддома, в одном рождается около 45 детей в день, в другом около 15. Где за год будет больше дней, когда мальчиков родилось больше 60 %? 56 % ответили что примерно поровну, и лишь 22 % назвали маленький. Расчёт даёт около 15 % таких дней в маленьком роддоме и около 7 % в большом. Бернулли был слишком оптимистичен. Чутьё срабатывает, когда просишь угадать среднее, а когда речь о разбросе средних, оно у большинства спит.

Затем больницы. Сравнивать клиники по смертности простой таблицей нельзя: небольшое отделение случайно окажется и первым, и последним. Поэтому кембриджский статистик Дэвид Шпигельхалтер в 2005 году предложил для сравнения учреждений ту самую воронку, что вы видели выше. Тот, кто внутри неё, не отличим от остальных. Тот, кто вышел за край, требует проверки. Воронка стала стандартным способом сравнения клинических исходов в британской службе здравоохранения, в том числе для детской хирургии сердца. Здесь от формулы может зависеть и честь хирурга, и жизнь пациента.

Наконец, вот как на законе больших чисел зарабатывают. Страховая компания не знает, у кого из вас сгорит дом, но при ста тысячах независимых домов она довольно точно знает, сколько их сгорит. Закон превращает чужую катастрофу

в статью бюджета. И рушится это, как и следует, только когда пропадает независимость: наводнение затапливает дома всех клиентов сразу.

Тот же закон сидел и в Монте-Карло в 1913 году. Публика ждала от него услуги за двадцать шесть вращений колеса. Казино с его одним зелёным зеро из тридцати семи делово ждало миллионы вращений и своих 2,7 %. Один и тот же закон, просто по разные стороны стола.

Задачи

Задача 1. В сельской школе 25 выпускников, разброс баллов у детей 15, обычный средний балл 65. В этом году средний балл 71, директор готовит пиджак под орден. Чудо ли это?

Ответ. Разброс среднего 15 / корень из 25 = 3, а 71 лежит в двух разбросах от 65. Чудо ожидается примерно раз в сорок лет, а школ таких в области достаточно, так что орденами можно запасаться ежегодно. На следующий год будет 66, и директору придётся объясняться.

Задача 2. Среднее Петровича по пяти заходам в баню известно с разбросом 4 / корень из 5 примерно 1,8 кружки. Сколько заходов нужно, чтобы узнать среднее с разбросом 0,4 кружки?

Ответ. 4 / корень из n = 0,4, то есть корень из n = 10, n = 100. Петрович готов послужить науке и ожидает грант. При независимости наблюдений, правда, сотню заходов нельзя делать подряд: следующая пятница после загула не независима.

Задача 3. У одной шашлычной на картах рейтинг 5,0 по двум отзывам, у другой 4,7 по трём тысячам. Пусть разброс оценок одного гостя равен единице. Куда идти?

Ответ. Разброс среднего у первой около 0,7, у второй около 0,02. Про вторую мы почти всё знаем, про первую

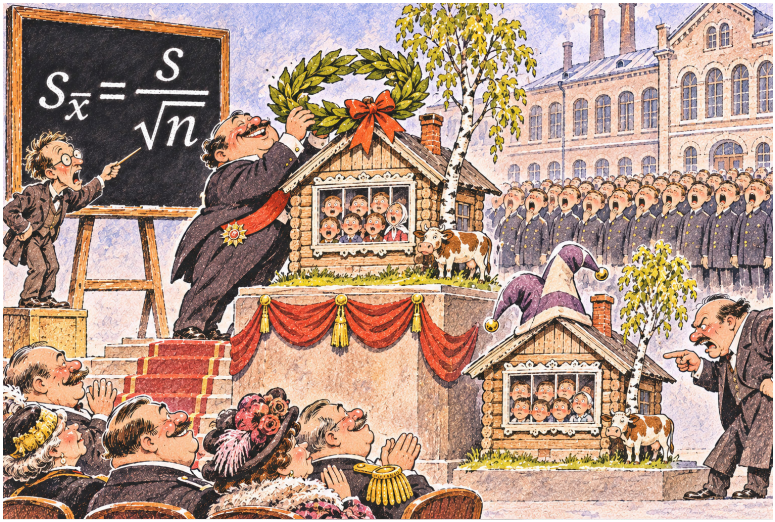
только то, что у шашлычника есть тёща и брат.

Задача 4. На рулетке десять раз подряд выпало чёрное. Сосед по столу в галстуке шепчет, что теперь по закону больших чисел точно красное. Какова вероятность красного?

Ответ. Те же 18 из 37, около 48,6 %. Колесо памяти не имеет и про десять чёрных не знает. Знает про них только сосед, и галстук он скоро продаст.

Задача 5. Начальник автопарка опросил тысячу пассажиров и недоволен точностью. Он хочет втрое точнее. Сколько пассажиров опросить?

Ответ. Девять тысяч. За тройную точность девятикратные деньги, таков курс корня. Автобус при этом по-прежнему не придёт, но теперь мы точно знаем, насколько это всех бесит.


$S_{\bar{x}} = \frac{S}{\sqrt{n}}$

Глава 10. Откуда берётся колокол

Старшина выстраивает роту по росту. Формул он не держит, но глаз намётан: две-три каланчи на правом фланге, два-три заморыша на левом, а посередине главная масса людей одного калибра, на которых и шьют форму. Расставьте эту роту столбиками по сантиметрам, и получится холм с пологими склонами, похожий на колокол.

Пока ничего странного. Странное начинается, когда тот же холм получается у буханок на хлебозаводе, у ошибок землемера, у числа орлов при сотне бросков монеты. Буханка, землемер и монета друг с другом не совещались, и старшина им не приказывал. Откуда всюду одна и та же кривая?

Здравый смысл отвечает красиво: у природы есть замысел. Так думал и Адольф Кетле, знакомый нам по первой главе. В обхватах груди тех самых 5738 шотландских солдат он увидел в 1846 году колокол, тот самый, что у астрономов получался при ошибках измерения звёзд. Вывод его получился потрясающим. Природа целится в идеального среднего человека, а каждый живой шотландец есть промах в ту или иную сторону. То есть вы, уважаемый читатель, тоже результат косой стрельбы.

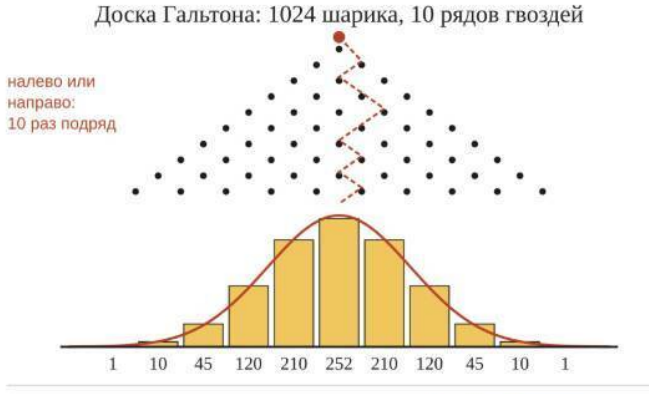
Звучит возвышенно, но один вопрос портит всю картину. Во что целится монета? У неё нет идеала в пятьдесят орлов. Дело, очевидно, не в замысле, а в чём-то гораздо более про-

стом, и увидеть это можно за пять минут, имея доску, гвозди и горсть дроби.

Доска с гвоздями

Фрэнсис Гальтон, тоже знакомый нам по первой главе, был человеком, который измерял всё, что не успело убежать. (Он же придумал евгенику, и она обошлась человечеству очень дорого, шутить тут не о чем.) 27 февраля 1874 года он показал в Королевском институте в Лондоне нехитрое устройство. Доска стоит наклонно, в неё вбиты ряды гвоздей, сверху сыплется дробь. На каждом гвозде дробина отскакивает налево или направо, как получится. Внизу карманы. И в карманах каждый раз вырастает колокол, хотя никакой идеальной дробины природа в виду не имела.

Почему? Посчитайте пути. Пусть рядов десять. В крайний правый карман ведёт одна-единственная дорога: десять раз подряд направо. А в средний карман, где поровну правых и левых, ведут 252 разных дороги. Вот и вся тайна середины. Середина не лучше краёв, просто в неё больше дорог. Чтобы стать средним человеком, годится почти любая биография, а чтобы стать крайним, требуется десять одинаковых ударов судьбы подряд.



$$f(x) = \frac{1}{s\sqrt{(2\pi)}} e^{-(x-\mu)^2 / 2s^2}$$

μ — центр колокола,
s — его ширина (разброс),
π ≈ 3,14; e ≈ 2,718

μ = 5; s = √(10 · ¼) ≈ 1,58; f(5) = 1 / (1,58 · 2,507) ≈ 0,252

точный подсчёт путей: 252 / 1024 ≈ 0,246

Эту формулу зубрить не надо. В ней всего два числа, которые вы выбираете: где центр и какая ширина. Всё остальное константы, причём потрясающей красоты: для расчёта, сколько дробин упадёт в карман, понадобилось число пи, которое про круги. Где тут круг, не спрашивайте, на пальцах этого не объяснить. Но что формула не врёт, проверяется на калькуляторе: для среднего кармана она даёт 0,252, а точный подсчёт путей 0,246. Десять гвоздей, а уже почти совпало.

Гальтон не скрывал восторга. В книге 1889 года он назвал

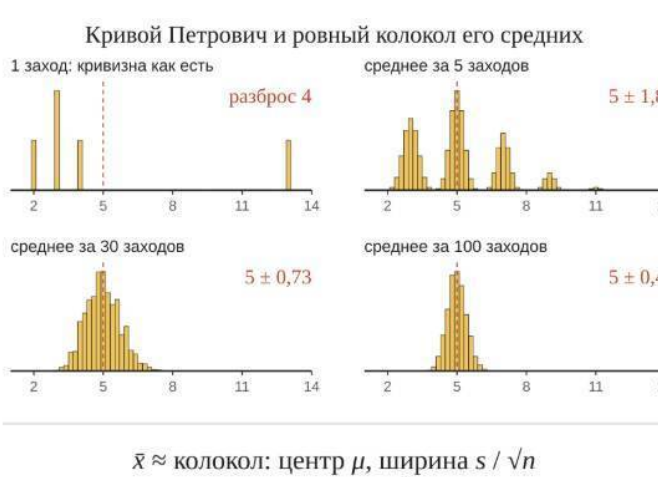
это чудесной формой космического порядка и высшим законом Неразумия, а греки, писал он, олицетворили бы и обожествили его, знай они о нём. Греки действительно нашли бы место на Олимпе, они обожествляли и пьянку.

Теперь главное. Дробина на доске есть сумма многих маленьких независимых толчков. В 1733 году де Муавр, уже знакомый вам по прошлой главе, понял, что для монет такая сумма даёт колокол. Потом Лаплас расширил это на более широкие случаи, а настоящую общность дала русская школа: Чебышёв, Марков и Ляпунов, который около 1900 года доказал теорему для почти любых толчков. В 1920 году Дьёрдь Пойа назвал её центральной предельной теоремой, потому что она центральная для всей теории вероятностей. Смысл в одну строку: складывайте много мелких независимых причин, и получится колокол, какими бы ни были сами причины.

Здесь тайна буханки и тайна старшины. Буханка недобрала веса из-за муки, печи, сменщицы, семи мелочей. Рост складывается из тысяч генов с малым эффектом и маминых котлет, что Рональд Фишер показал в 1918 году. Кетле совершенно верно увидел факт, но природа тут никуда не целилась. Сам по себе идеальный средний шотландец тоже просто сумма.

А теперь проверим теорему на самом неудобном из наших героев. Банные заходы Петровича никакого колокола не образуют: 2, 3, 3, 4 и 13 кружек, три торчка слева и один па-

мятник справа. И вот что будет со средним за несколько заходов, если каждый раз тянуть наугад из этих пяти чисел.



За пять заходов колокола ещё нет. Есть горбы, и каждый означает число загулов: ни одного, один, два, три. За тридцать заходов горбы сливаются, хотя справа кривизна ещё заметна. За сотню перед нами колокол, с центром в тех же пяти кружках и шириной посчитанной в прошлой главе: четыре, делённые на корень из ста. Закон больших чисел сообщал, куда стягивается среднее и как быстро. Центральная предельная теорема сообщает, какую форму оно имеет по доро-

ге. Данные могут быть сколь угодно кривыми, но их среднее ведёт себя прилично. Петрович остаётся кривым, но среднее его при этом идёт прямо.

У колокола есть практическое правило, которое стоит помнить наизусть: в пределах одной ширины от центра лежит 68 % всех значений, в пределах двух 95 %, в пределах трёх 99,7 %. Проверим на шотландцах Кетле. Средний обхват груди у них 39,8 дюйма, разброс 2,05. В пределах одного разброса лежит 66,8 % солдат, в пределах двух 95,3 %. Шотландцы теоремы не читали, но строй держат.

Когда колокол треснет

Причины не мелкие. Если среди толчков есть один огромный, он и будет диктовать форму. Загул Петровича на пяти заходах вы только что видели: вместо колокола стоят горбы. Теорема про предел, а не про пять бутылок, и чем кривее исходные данные, тем дольше ждать.

Причины не складываются, а перемножаются. У кого уже есть миллион, тот зарабатывает процентами от миллиона, а у кого тысяча, тот процентами от тысячи. Книгу, которую уже читают, читают ещё больше. Тут получается не колокол, а перекошенная кривая с длинным хвостом из пятой главы. Зарплата, доход и население городов колоколом не бывают, сколько бы людей и городов ни перебрать.

Причины не независимы. Если все одновременно бегут продавать, потому что все бегут продавать, теорема молчит. Вдобавок колокол хорош в середине и плох в хвостах. 19 октября 1987 года биржевой индекс S&P 500 упал за день на 20,5 %, то есть примерно на 25 разбросов обычного дня. В августе 2007-го финансовый директор Goldman Sachs Дэвид Виньяр пожаловался, что они наблюдают события в 25 стандартных отклонений, несколько дней подряд. По формуле колокола вероятность одного такого события примерно 3 · 10^-138. Для понимания: возьмём условно средний рост 176 сантиметров и разброс 7. Человек в двадцать пять разбросов

имеет рост 3 метра 51 сантиметр. И вот такой заходит в банк несколько дней подряд. Тут прилично уже не удивляться великану, а поменять модель.

Не всё то, что названо колоколом, колокол. Физик Габриэль Липпман зло заметил: все верят в нормальный закон, экспериментаторы потому, что думают, будто это математическая теорема, а математики потому, что думают, будто это экспериментальный факт. В 1989 году Теодор Микчери проверил 440 больших наборов данных по успеваемости и психологическим тестам. Статью он назвал «Единорог, нормальная кривая и другие неправдоподобные существа». Ни один из 440 наборов нормальным тест не прошёл. А колокол в IQ вообще сделан по приказу: баллы пересчитывают так, чтобы получился центр 100 и разброс 15. Это как старшина, который выстроил роту и доложил, что она ровная.

И слово не то. В учебниках колокол зовут нормальным распределением. Карл Пирсон, который это название ввёл в обиход, в 1920 году сам жаловался, что оно заставляет людей думать, будто все остальные распределения ненормальные. Так оно и вышло. Запомните: колокол не норма, а ответ арифметики. Кому в роте не досталось середины, тот не брак.

Где это применяют всерьёз

Самое главное применение вы увидите в следующих главах, и оно занимает половину книги. Когда социолог говорит «плюс-минус три процента», а фармаколог «лекарство помогает», они не знают формы исходных данных. Они знают форму среднего. Теорема дарит науке настоящую роскошь: судить о точности по одной выборке, не зная, как устроена исследуемая толпа. Без неё не было бы ни доверительных интервалов, ни большинства критериев, которыми проверяют гипотезы.

Астрономия пользовалась колоколом раньше всех. В начале XIX века Карл Фридрих Гаусс вывел закон ошибок наблюдений и по нескольким неточным отсчётам вычислил орбиту пропавшей Цереры. Оттого кривую зовут ещё гауссианой, и попала она на немецкую банкноту в десять марок рядом с портретом Гаусса. Не без иронии: формула, по которой финансисты потом ошибутся на двадцать пять разбросов, так и висела у них в кошельке.

Заводы взяли колокол на службу в двадцатом веке. Служебная записка Уолтера Шухарта из третьей главы, поданная 16 мая 1924 года, занимала одну страницу, и треть её занимал чертёж. Так появилась контрольная карта. Теперь видно, на чём она стоит: мелкие причины дают колокол, поэтому точка за тремя разбросами означает уже особую причину. В

1986 году Билл Смит из Motorola довёл это до программы «Шесть сигм», где целью стало 3,4 брака на миллион. Ровно шесть сигм при этом там нет: в расчёт заложен сдвиг в полторы сигмы, который статистик Дональд Уилер назвал дурацким. Зато специалистам по внедрению там выдают чёрные пояса, как в каратэ. Петрович бы оценил.

И физика. В 1905 году Эйнштейн увидел в беспорядочной танце пылинки в воде доску Гальтона без доски. Пылинку толкают невидимые молекулы, толчков много, каждый мал, и сумма получается колоколом с шириной, растущей как корень из времени. Когда Жан Перрен это измерил, сомнения в существовании атомов стали неприличными. Тот же самый корень из n, что у школы в Верхних Кочках, помог доказать, что мир состоит из атомов.

Задачи

Задача 1. На доску в десять рядов бросили 1024 дробины. Сколько из них попадут в крайний правый карман, и что об этом напишет газета?

Ответ. В среднем одна, шанс 1 на 1024. Газета напишет о секретах успеха этой дробины и о её стальном характере. Дробина согласится дать мастер-класс.

Задача 2. Сосед бросил монету 100 раз на пиво и получил 60 орлов в свою пользу. Бить ли соседа?

Ответ. Центр 50, ширина корень из $(100 \cdot 1/4) = 5$, то есть 60 отстоит на две ширины. Вероятность получить 60 и больше случайно около 2,8 %, один спор из тридцати пяти. Подозрительно, но на рукоприкладство не тянет. Предложите соседу бросить ещё сотню. Заодно проведёте вечер.

Задача 3. Рост новобранцев условно имеет среднее 176 и разброс 7 сантиметров. Сколько двухметровых на десять тысяч новобранцев?

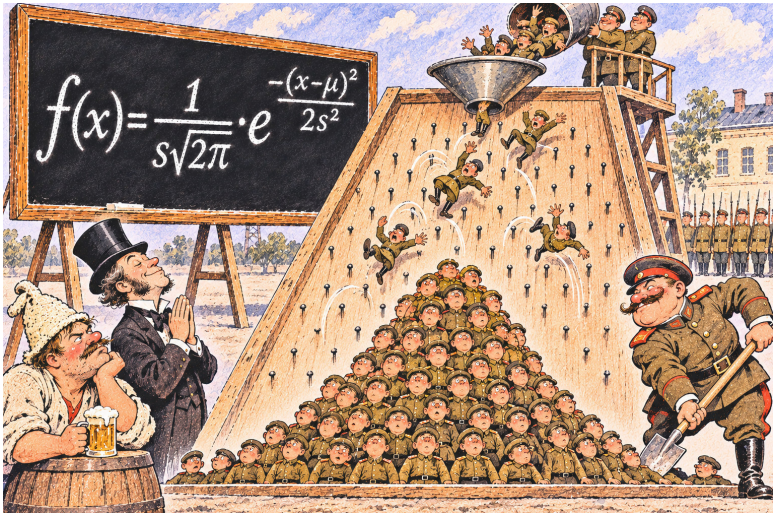
Ответ. От 176 до 200 тут 3,4 разброса, по таблице колокола примерно трое. Если в вашей роте двухметровых весь правый фланг, колокол не виноват: вы попали в почётный караул, где отбирали.

Задача 4. В бутылке поллитра с разбросом 10 миллилитров. Петрович слил весь ящик из двадцати бутылок и намерил 9,8 литра вместо десяти. Случайно ли?

Ответ. Разброс суммы $10 \cdot$ корень из 20 примерно 45 миллилитров, а недолив 200, почти 4,5 разброса. Шанс примерно 4 на миллион. Это уже не случай, а либо завод недоливает, либо Петрович успел приложиться по дороге к мерной кружке.

Задача 5. За последние сто заходов в баню среднее у Петровича 6,2 кружки вместо пяти, а он уверяет жену, что ничего не изменилось, а это так случайно вышло. Прав ли Петрович?

Ответ. Ширина 0,4, сдвиг 1,2, ровно три ширины. Шанс на такой случай около 0,13 %, один на 740. Жена права: в процессе появилась особая причина, по Шухарту. Скорее всего, приехал зять.


$f(x)=\frac{1}{s\sqrt{2\pi}}\cdot e^{\frac{-(x-\mu)^2}{2s^2}}$

Глава 11. Суп пробуют ложкой

В новостях сообщают: по данным опроса, сорок семь процентов граждан страны считают то-то и то-то. Опрошено 1600 человек. Петрович отрывается от кружки и возмущается: в стране сто сорок миллионов, меня никто не спрашивал, соседа не спрашивали, и вообще никого никогда не спрашивали. Как можно по тысяче с чем-то судить о всей России? Это же один человек почти на девяносто тысяч!

Здравый смысл стоит на стороне Петровича и предлагает простой выход. Чем больше страна, тем больше надо спрашивать. Тысяча годится для района, а для России нужен миллион, а лучше два. Хороший опрос это большой опрос.

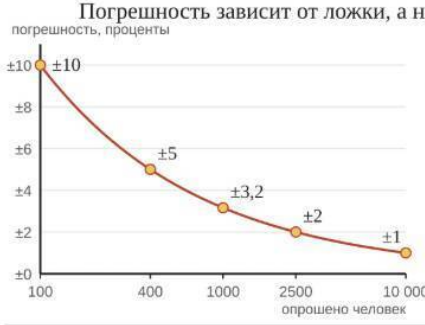
Предложим Петровичу задуматься о супе. Жена варит борщ. Чтобы узнать, хватает ли соли, она не съедает кастрюлю. Она пробует ложку. А в армейской столовой, где борщ варят в котле на триста литров, повар пробует такой же ложкой. Никто не требует от него пробовать ведром, хотя котёл в сто раз больше. Но и тот и другой перед этим делают одно и то же: мешают. В неперемешанном борще сверху жир, а внизу гуща. Зачерпните хоть ведром сверху, и вы точно узнаете, каков на вкус жир.

В этом борще содержится вся глава, и проверили его на масштабах целой страны. Осенью 1936 года американский журнал The Literary Digest зачерпнул поварёшком в два

с лишним миллиона бюллетеней и предсказал победу республиканца Альфа Лэндона. Никому не известный молодой Джордж Гэллап зачерпнул более скромной ложкой, около пятидесяти тысяч человек, но помешанной гораздо лучше, и назвал победителем Франклина Рузвельта. Рузвельт получил 60,8 % голосов, взял все штаты, кроме двух, а журнал после этого протянул меньше двух лет.

Ложка и кастрюля

Каждый случайно выбранный человек отвечает «за» или «против», как монета падает орлом или решкой. Значит, доля «за» в выборке есть среднее, и для него у нас уже всё готово. Из девятой главы берём корень из n, из десятой колокол и две ширины, которые дают нам 95 процентов. Разброс одного ответа да-нет равен корню из p(1 - p).



Опрошена тысяча, а население разное:

5 тысяч	±2,8
1 миллион	±3,2
140 миллионов	±3,2

$$m = 2\sqrt{\frac{p(1-p)}{n}}$$

m — погрешность (95 %),
p — доля «за» в выборке,
n — размер выборки

p = 0,5; n = 1000: $2 \cdot \sqrt{(0{,}25 / 1000)} \approx 0{,}032 = 3{,}2$ %

населения в формуле нет; работает только для случайной выборки

Теперь посмотрите на эту формулу глазами Петровича. Где тут сто сорок миллионов? Нигде. Размера страны в ней нет вообще, как и размера котла в голове повара с ложкой. Тысяча случайных человек даёт погрешность около трёх процентов и в миллионном городе, и в стосорокамиллионной стране. Небольшая прибавка к точности набегает лишь тогда, когда выборка составляет заметную долю населения, как тысяча из пяти тысяч: получается 2,8 вместо 3,2. Привыкнуть к этому трудно, голова сопротивляется, но у борща всё точно так же.

Заодно понятно, почему социологи так любят числа в тысячу-две. Это не лень, а бухгалтерия. Переход от тысячи к десяти тысячам умножает расходы на десять, а точность повышает чуть больше чем втрое. А наши 1600 человек из новостей дают для сорока семи процентов точность плюс-минус два с половиной. Этим плюсом-минусом мы займёмся подробно в шестнадцатой главе.

У формулы есть одно маленькое условие, и оно то и дело ломает всё: выборка должна быть случайной. Случайной не в житейском смысле, когда кого встретил, того и спросил, а в строгом: у каждого жителя должен быть известный шанс попасть в ложку. Эту мысль пришлось долго пробивать. Статистики XIX века считали настоящим знанием только сплошной счёт. Норвежец Андерс Кьер, предлагавший пробовать ложкой, потратил на уговоры десятилетия. Англичанин Артур Боули в 1912 году показал на примере города Рединга,

что это вообще возможно. А в 1934 году Ежи Нейман, учившийся в Харькове, убедил сообщество, что судить о точности можно только по случайной выборке, какой бы умной ни казалась тщательно подобранная вручную.

Теперь вернёмся к журналу и понюхаем его борщ. Редакция разослала около десяти миллионов бюллетеней, и брала адреса в телефонных книгах, списках владельцев автомобилей и среди своих читателей. Это в разгар Великой депрессии. Вернулось 2,4 миллиона. Лэндону обещали 57,1 процента, а он получил 36,5. По нашей формуле погрешность такого опроса несколько сотых процента, а ошибка больше двадцати. Тут работает другая статья расходов.



Разброс убирается деньгами, смещение деньгами не убирается. Журнал купил за большие деньги самый точный неправильный ответ в истории Америки.

Откуда же смещение? Во всех учебниках пишут: телефоны и автомобили были у богатых, а богатые голосовали за правого. История красивая, и она даже отчасти правда, но не вся правда. В мае 1937 года Гэллап сам расспросил людей, и современный разбор этих данных показал другое. Даже владельцы телефонов и автомобилей склонялись к Рузвельту: около 54 процентов. Если бы ответили все десять миллионов, Рузвельт набрал бы около 57 процентов и журнал назвал бы победителя верно. Главное случилось у почтового ящика. Бюллетени вернула примерно треть сторонников Лэндона и лишь пятая часть сторонников Рузвельта. Довольный человек сидит дома, а сердитый идёт покупать марку. Примерно две трети ошибки журнала пошли отсюда.

Что же делать, если люди не мешаются сами? Зачерпывать жребием, добиваться ответа и делить котёл на слои. Расслоение это когда берут отдельно из города и деревни, из мужчин и женщин, в тех долях, в каких они есть в стране, а недобранных потом довешивают весами. Это помогает, но только пока вы знаете, кто именно от вас убежал.

Когда ложка врёт

Борщ сам пришёл в ложку. Зрители присылают смс на короткий номер, читатели голосуют на сайте газеты, покупатели пишут отзывы. Измеряется не общественное мнение, а готовность потратить вечер. Классический пример дала Шер Хайт, уважаемый автор книг о любви и браке: она разослала сто тысяч анкет по женским обществам и получила 4500 ответов, то есть 4,5 %. Женщина, заполняющая длинную и очень личную анкету о своём браке, отличается от остальных девяноста пяти процентов как минимум этим самым. Формула погрешности к таким числам не применима вообще, и никакие проценты с десятыми долями не сделают её применимой.

Ложка не достаёт до дна. Есть люди, до которых трудно добраться совсем. Если вы звоните днём на домашний телефон, то узнаёте мнение бабушек. Если опрашиваете на улице, то не встретите ни лежачих больных, ни вахтовиков, ни просто тех, кто увидел вас издалека и перешёл на другую сторону. Так журнал не достал до тех, у кого в 1936 году не было ни телефона, ни автомобиля.

Повар выбирает кусочки получше. В 1948 году Гэллап, герой 1936 года, вместе с Роупером и Кроссли обещали победу Томасу Дьюи. У Гэллапа было 50 на 44, а Гарри Трумэн выиграл 50 на 45. Сохранилась знаменитая фотогра-

фия, где сияющий Трумэн держит газету с заголовком «Дьюи победил Трумэна». Все трое пользовались квотами: интервьюер обязан опросить столько-то горожан, столько-то женщин, а кого именно, выбирает сам. Сам он выбирал тех, кого легче найти и с кем приятнее говорить, а это оказались республиканцы. Так герой тридцать шестого года наступил в тот же борщ. После этого случайный отбор перестал быть научной придиркой.

Котёл цифровой, борщ тот же. В 2021 году в журнале Nature сравнили оценки доли привитых от ковида американцев. Опрос через крупную соцсеть собирал около 250 тысяч ответов в неделю и завышал на 17 процентных пунктов по сравнению с данными службы здравоохранения. Скромный распланированный опрос на тысячу человек ошибался на 4,2. Авторы посчитали, что четверть миллиона таких ответов была не точнее случайной выборки из десяти человек. Десяти, не тысяч. Сообщайте это каждому владельцу больших данных, который обещает вам правду, потому что у него много строк.

Есть ещё два рода порчи борща, но они так важны, что им отведены следующие две главы: в ложку попадают только выжившие, и попавшие в ложку врут.

Где это применяют всерьёз

Вы живёте в мире, устроенном по ложкам. Безработицу не пересчитывают повально. Американская статистика каждый месяц опрашивает около шестидесяти тысяч домохозяйств, и этого хватает, чтобы увидеть сдвиг уровня безработицы на две десятых процента. По этой цифре потом двигаются биржи и центральные банки. У Росстата есть такое же обследование рабочей силы, а инфляцию считают по выборке товаров и магазинов. Обойти все магазины страны за месяц невозможно, да и бессмысленно.

Конец ознакомительного фрагмента.

Текст предоставлен ООО «Литрес».

Прочитайте эту книгу целиком, купив полную легальную версию на Литрес.

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.