# Введение в мерфологию

Это, одна из первых глав, в которой на примере велосипедиста рассматриваются нужные нам инструменты для измерения несправедливости: кривая Лоренца и индекс Джини, а также упоминаются пресловутый Парето и грозный инспектор. 

## Закон есть закон

В этой книжке речь пойдёт о различных неприятностях. Привычных, ожидаемых и настолько предсказуемых, что они получили статус законов. Их уже сформулировано множество: это и закон падающего бутерброда, и закон Мерфи: "*Если какая-нибудь неприятность может произойти, она случится*" и законы Чизхолма на тему: "*Когда дела идут хорошо, что-то должно случиться в самом ближайшем будущем.*" и наблюдение Этторе: "*Соседняя очередь всегда движется быстрее.*" Большая их часть вполне тривиальна, но согласно закону Муира "*Когда мы пытаемся вытащить что-нибудь одно, оказывается, что оно связано со всем остальным.*" Мы постараемся найти рациональное зерно этих закономерностей, но не для того, чтобы с ними бороться, а для удовольствия. И поскольку при этом мы будем использовать математику, то удовольствие будет своеобразным и полезным, в отличие от самого результата. Ну, а если наши рассуждения заведут нас слишком далеко, мы можем взять на вооружение постулат Персига: "*Число разумных гипотез, объясняющих любое данное явление, бесконечно*". В конце концов, Гроссман, цитируя Х. Л. Менкина верно указал, что "*Сложные проблемы всегда имеют простые, легкие для понимания неправильные решения.*" Со всеми этими глубокомысленными фразами и законами мы и станем разбираться.

Какие-то случающиеся с нами неприятности закономерны, как говорят математики, детерминированы, то есть, не зависят от случайностей. Например, если вам понизили зарплату на 10%, а потом извинились и повысили на 10%, то в итоге этом махинаций, вы останетесь в убытке, поскольку

Более того, если зарплату сначала повысят, а потом, не извинившись даже, понизят на те же самые 10%, результат выйдет такой же, поскольку неважно в каком порядке перемножать коэффициенты. Это очень просто, обидно, но к удаче отношения не имеет.

Другой пример детерминированной неприятности — волшебство, случающееся в наших карманах с наушниками: кладём аккуратно сложенные наушники в карман, а через полчаса там происходит чудо, и из кармана мы вынимаем дикий узел проводов. В 2007 году вышла серьёзная научная статья двух учёных из солнечного и безмятежного Сан-Диего «Спонтанное образование узлов на возбуждаемой нити»[[1]](#footnote-1), в которой детально анализируется и моделируется запутывание наушников в кармане. Авторы, основываясь на теории узлов, теории вероятностей и физических экспериментах, убедительно показывают, что при стандартном способе сматывания, наушники, действительно, должны запутываться, причём, спустя лишь несколько секунд тряски. Впрочем, это мы и так наблюдаем практически каждый день. Неожиданной здесь может оказаться только ожидаемая скорость запутывания. С этой неприятностью вполне можно бороться математическим способом: нужно поменять способ складывания наушников — не кольцами, которым свойственно образовывать узлы, а чередой петель взаимно-обратного направления, как например, показано на рисунке. При таком способе складывания петли разных «знаков» взаимно уничтожают друг друга, и узлы не формируются. Уже много лет я складываю наушники именно таким образом, чувствуя себя крутым топологом, и всякий раз радуюсь, как фокусу, когда они разматываются сами от одного небрежного встряхивания рукой.



Один из способов складывания проводов, не приводящий к их запутыванию. Он хорош ещё и тем, что попутно вы складываете пальцы в мудру любви.

Но и среди стохастических по своей природе законов не все одинаково интересны. Например, закон Бука: «ключи всегда находишь в последнем кармане» не имеет под собой какого-либо рационального основания. Простой подсчёт показывает, что при равной вероятности отыскать ключи для всех карманов, последний ничем не отличается от прочих. Впрочем, этот закон можно трактовать разве что как забавный софизм: утверждение Бука верно всегда, поскольку тот карман, в котором ключи будут обнаружены, завершит процесс поиска и, следовательно, будет последним. Однако и здесь есть о чем поговорить. В процессе перебора карманов, так называемая условная вероятность того, что ключи лежат в последнем кармане, действительно повышается. Но это, уже нельзя просто трактовать, как вероятность того, что ключи находятся в последнем кармане, это уже другая задача. С тем, что же здесь имеется в виду, мы разберёмся в третьей главе, когда и понятие вероятности станет нам ближе и можно будет его несколько усложнить.

Нас будут интересовать законы парадоксальные и поучительные, законы, которые выглядят злым роком, выбирающим из множества вариантов самые досадные и неприятные, наперекор интуиции подсказывающей, что этот выбор не должен быть самым вероятным. И прежде чем приступить к детальным и точным рассуждениям о случайностях и о вероятностях, давайте предположим, что какая-то интуиция в отношении случайных процессов и вероятностей у нас уже есть. Это вполне допустимо даже в математической книжке -- до какого-то момента использовать интуитивное представление о предмете, а потом дать строгое определение. Тем самым, в первую очередь, мы определяем границы применимости нашей интуиции, а во-вторых, расширяем эти границы в правильном с научной точки зрения направлении.

Начнём мы с некоторого простого инструментария, который будем использовать на протяжении всей книжки, а для этого нам потребуется… велосипед.

## Если долго, долго, долго, если долго по тропинке

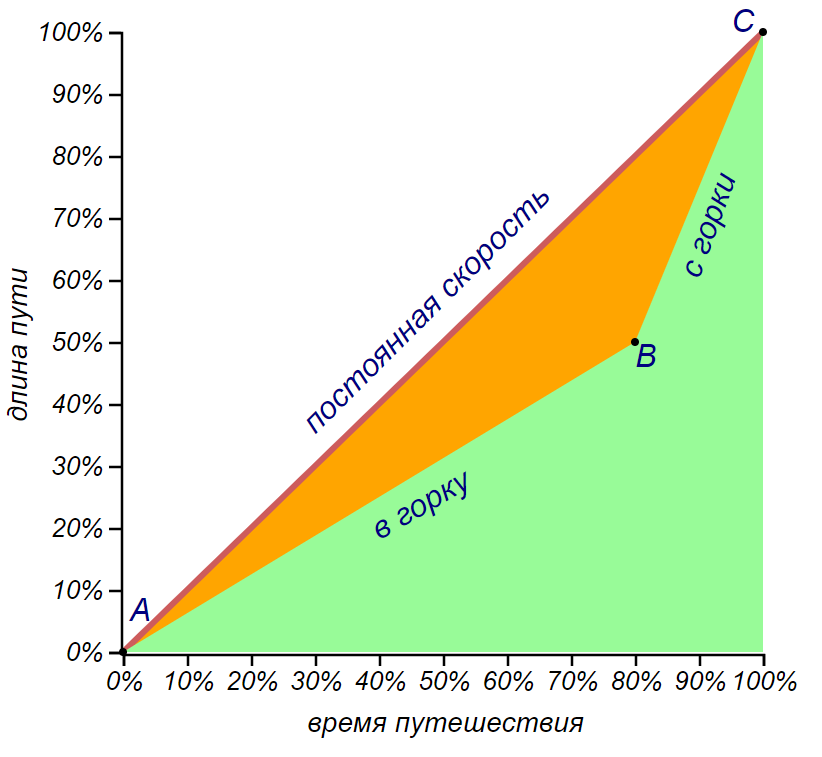
Я большой энтузиаст велосипедного любительского спорта. Многие задачи, ставшие примерами для этой книжки я обмозговывал в седле, вертя их мысленно и так и эдак, пытаясь найти наиболее наглядный и простой подход к ним. Но, вообще, что может быть лучше, чем мчаться по трассе ранним утром, по холодку, скатываясь с лёгкого склона… это ощущение стоит того, чтобы ради него преодолевать бесконечные подъёмы или сопротивление встречному ветру! Правда, порой кажется, что подъёмов как-будто бы больше, чем спусков, а ветер норовит быть встречным, куда ни поверни. В книгах по мерфологии в этой связи приводится **закон велосипедиста**:

Независимо от того, куда вы едете — это в гору и против ветра.

Живу я на Камчатке, в Петропавловске много горок, и катаясь по городу, их не миновать. Однако меня должна успокаивать мысль, что начиная свой путь из дома, я возвращаюсь снова домой, значит, суммарный спуск должен быть равен суммарному подъёму. Особенно честным будет радиальный маршрут. Представим себе 2-километровую трассу, состоящую из одной симметричной горки: километр вверх, километр вниз. Вверх по склону я могу достаточно долго ехать со скоростью 10 км/ч, а на спуске стараюсь держать скорость в 40 км/ч (да, я осторожный и езжу в шлеме). Значит, на подъём я буду тратить в четыре раза больше времени, чем на спуск, и общая картина будет такая: 4/5 времени путешествия уйдёт на тягучий подъём, и лишь 1/5 — на приятный спуск. Получается обидно — 80% времени прогулки составляют сложные участки пути! Если я выкачусь из нашего холмистого города, в сторону океана или в долину реки Авачи, горок почти не будет, но в моём распоряжении остаются встречный и попутный ветер, или участки с плохой дорогой.

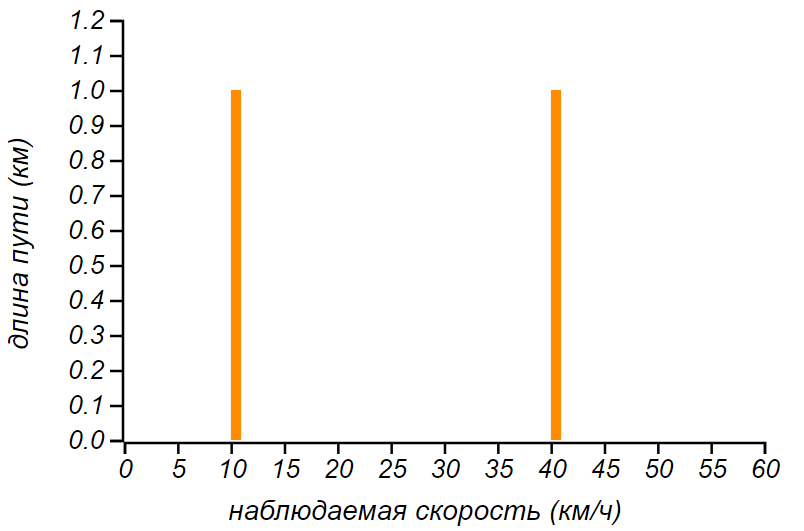
Давайте взглянем на закон велосипедиста со стороны теории вероятности. Если я сделаю множество селфи на протяжении своей велопрогулки, а потом стану доставать их, не глядя, из перемешанной пачки, то значительная часть картинок покажет мне согбенную фигуру в оранжевом шлеме, смиренно ползущую вверх по склону или против ветра. Вероятность увидеть на снимке летящего и сияющего велосипедиста, с рекламной картинки, увы, составит лишь около 20%. А что скажет статистика? Если мы выпустим на холмистую трассу большую толпу велосипедистов, подождём немного, и пронаблюдаем за их плотностью, то увидим, как большая часть спортсменов толпится на трудных участках, и вероятность обнаружить безмятежно улыбающееся лицо в общей массе окажется не так уж и велика!

*Результат имитационного моделирования движения ансамбля велосипедистов на холмистой трассе. Для каждого из участников движения задана его мощность, она определяет его предельную скорость, как на подъёме, так и на спуске (учитывается сопротивление воздуха). Видно, как вскоре после начала движения, на подъёмах сосредотачивается большая часть всего ансамбля.*  
  
Давайте, как когда-то в школе, покажем на графике зависимость перемещения велосипедиста от времени, при движении по симметричной треугольной горке. Только сделаем всё по-взрослому, в собственных масштабах задачи: расстояние будем измерять не в километрах, а в долях общего пути, так же поступим и со временем путешествия. Первую половину пути (отрезок AB) велосипедист двигался медленно и долго — 4/5 всего времени пути, а вторую (отрезок BC) преодолел быстро — за 1/5 времени.



*График перемещения велосипедиста в долях от общего пути и времени.*

Существует один вполне универсальный способ суждения о несправедливости этого мира, принятый у эконометристов, демографов, экологов или маркетологов — *кривая Лоренца* и связанный с ней *индекс Джини*. Для известного распределения чего-нибудь ценного, например, денег, в некоторой популяции, можно, предварительно отсортировав членов множества по возрастанию уровня богатства, построить кумулятивную кривую, нормируя ось X на численность популяции, а ось Y — на общее её благосостояние. Получится кривая, носящая имя американского экономиста Макса Отто Лоренца. Когда мы строили график перемещения велосипедиста, мы, по существу, построили кривую Лоренца для распределения скоростей по отрезкам пути, состоящего всего из двух столбцов.

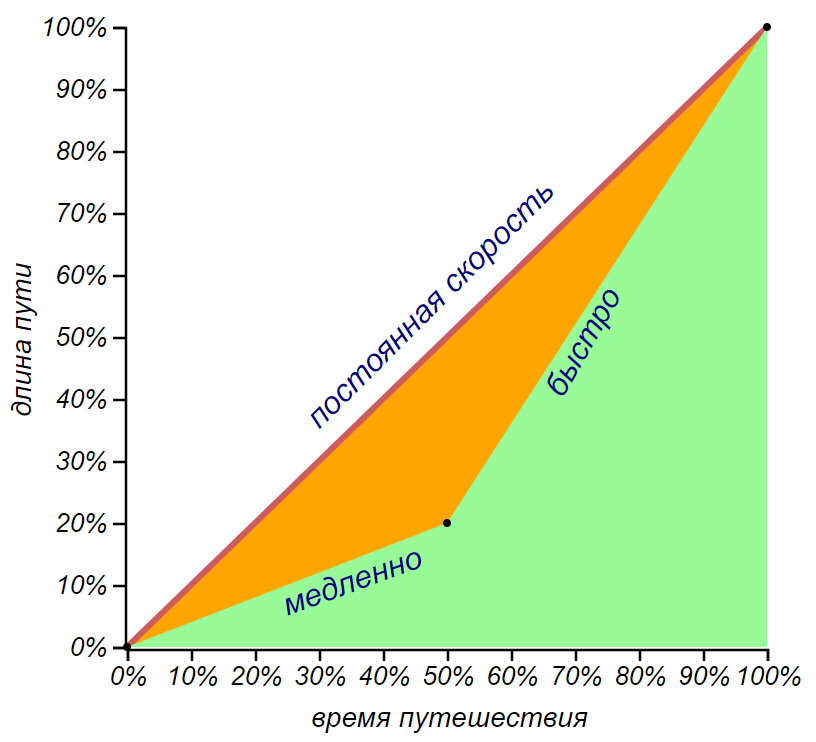


*Распределение скорости велосипедиста по пройденному пути.*

Конечно же, не всякий график перемещения можно воспринимать, как кривую Лоренца. Перед тем как её строить, нужно отсортировать периоды путешествия по возрастанию скорости, после чего уже приступать к построению. Иными словами, сначала нужно построить гистограмму скоростей, после чего последовательно складывать вклады всех столбиков гистограммы, начиная со вклада малых значений, заканчивая самыми большими. Результатом должна явиться всюду вогнутая кривая, которая проходит ниже диагонали (AC). Эта диагональ называется *кривой равенства*, она, в нашем случае, соответствует постоянной (средней) скорости на всём пути, или гистограмме с одним единственным столбиком (дельтообразной функции плотности вероятности). А в экономическом смысле — всеобщему равенству благосостояния. Чем больше кривая Лоренца отклоняется от кривой равенства, тем менее «справедливым» можно считать распределение. Коль скоро мы изучаем законы подлости и несправедливости нашего мира, разумно использовать и терминологию и инструменты, используемые для исследования справедливости.

Площадь под кривой Лоренца для любого распределения, отличного от дельтаобразного, будет меньше площади под кривой равенства. Их разница может служить формальной характеристикой неравенства или «несправедливости» распределения. Эту характеристику отражает *индекс Джини*. Он вычисляется, как удвоенная площадь фигуры, образованной кривой равенства и кривой Лоренца. Для идеального мира индекс Джини равен 0, в самом кошмарном варианте он стремится к единице. В рассмотренном нами примере, он равен 0.35. Это вполне неплохой показатель. Скажем, распределение богатства среди населения в России сейчас имеет индекс Джини 0.39, в США — 0.49, в Австрии и Швеции он не превышает 0.3, а для всего Мира он в 2017 г. составил 0.66. Так что ситуация с велосипедистами, конечно, обидна и несправедлива, но вполне терпима.

Мы рассматривали распределение скоростей по расстоянию, а что будет, если нам дано распределение скоростей по времени (делим время пути на интервалы и подсчитываем количество интервалов с той или иной скоростью). Благодаря безразмерности диаграммы Лоренца, мы снова сможем изобразить соответствующую кривую, и даже сравнить с предыдущим результатом. Например, пусть половину времени путешествия, скажем, час, велосипедист ехал со скоростью 10 км/ч, а час — со скоростью 40 км/час (при этом не важно, в каком порядке). Тогда на малую скорость придётся 1/5 всего пути, а на большую — 4/5. Кривая Лоренца, в случае распределения скорости по времени, будет отражением кривой Лоренца для распределения скоростей по расстоянию, относительно диагонали, перпендикулярной линии равенства. При этом индекс Джини будет тем же, ведь при отражении кривой, площадь под ней не изменится. Так что по уровню несправедливости эти два разных условия, оказываются одинаковыми, хотя по ощущениям, второй случай гораздо приятнее!



*График перемещения (кривая Лоренца) велосипедиста в случае равного времени следования с двумя различными скоростями.*

Обратите внимание, с помощью некоторого формального индекса мы стали сравнивать совершенно разные и не сравнимые вещи, это одновременно и заманчиво и опасно. Нужно отдавать себе отчёт в том, что формальные индексы и критерии всегда чему-то равны, не зависимо от того есть в этом смысл, или нет. Мы сравниваем распределение богатства среди населения стран и распределение времени затрачиваемого на преодоление пути с точки зрения отличия от некоторого варианта, которое сочли бы справедливым. Покуда мы ведём фривольные и, подчас, хулиганские разговоры о законах подлости, пожалуй, это оправданное сравнение, но в математике так, конечно же, делать нельзя. Кривую Лоренца, а по ней и индекс Джини можно формально рассчитать и для гистограммы яркости пикселей на картинке или для частотности слов в живой речи, к справедливости это не будет иметь никакого отношения, да и смысла останется совсем немного. Поэтому, имея в виду индекс Джини для чего попало, мы будем его называть **индексом подлости**, чтобы не вводить читателя в заблуждение наукообразностью терминов.

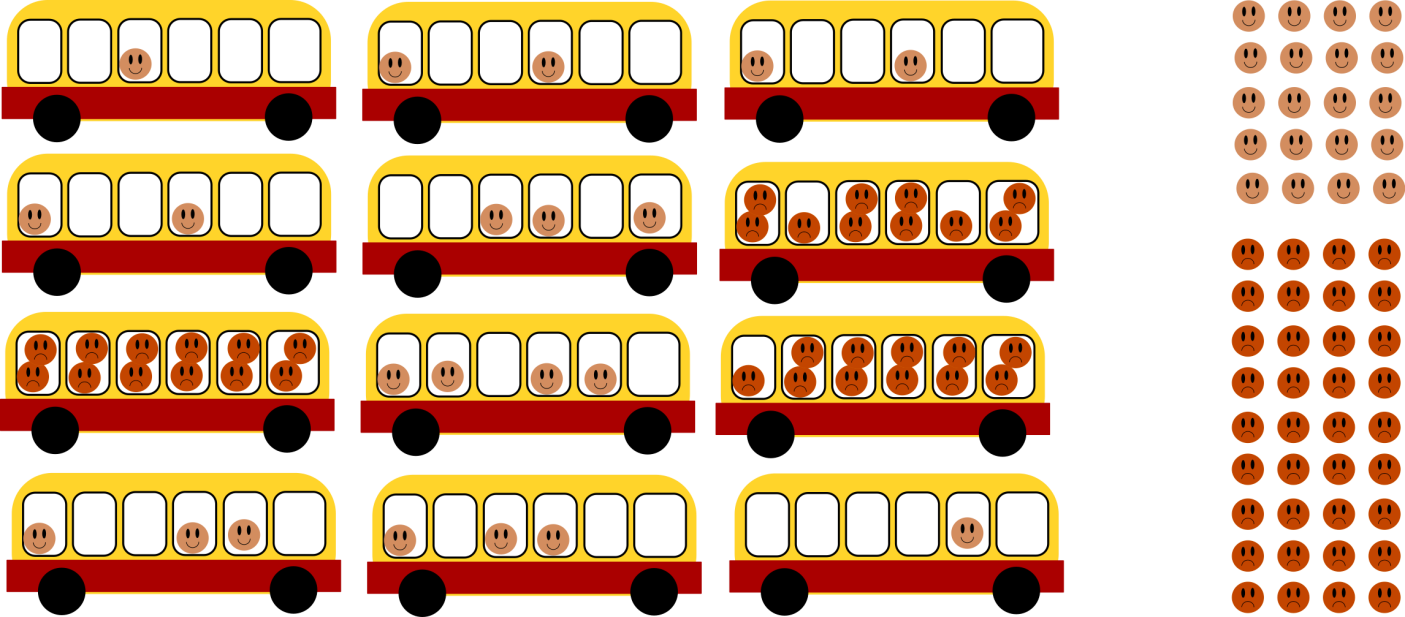
∗ ∗ ∗

Вывод, который делает велосипедист, пыхтя на пониженной передаче: «мир несправедлив и большую часть сил отнимает самая дурацкая часть работы», часто именуют *принципом Парето* или *принципом «80/20»*. Это абсолютная эмпирика, принцип Парето никто не доказывал, но его так часто цитируют, что он уже производит впечатление истины. Его используют, как оправдание и как инструкцию, обнаруживают в самых разных проявлениях и иногда это работает, например, принципу «80/20» соответствует индекс подлости порядка 0.6 — как для распределения богатства во всем мире. Понимая, что это не козни судьбы, а простейшая математика, с которой бороться смысла нет, можно научиться получать удовольствие и от затяжных подъёмов и от нудных, но неизбежных этапов работы, хотя бы, решая в уме задачки, или медитируя. Даосы стремились жить вечно, и правильно рассудили, что вместе с работой над телом, для достижения их цели, требуется подготовка ума. Ведь для вечной жизни нужно не только умение отпускать привязанности, но и терпение, а также умение получать удовольствие от затяжных участков.

У принципа Парето есть полезное для понимания более строгое обобщение. Закон подлости, названный в честь безымянного велосипедиста, имеет официальное научное звание: **парадокс инспекции**. Это хорошо известное явление встречается в самых разных исследованиях, связанных с социологическими опросами, тестированием в теории отказов (разделе прикладной математики, занимающемся надёжностью сложных систем), неявно, но систематически смещая наблюдаемые результаты в сторону более часто наблюдаемых явлений.

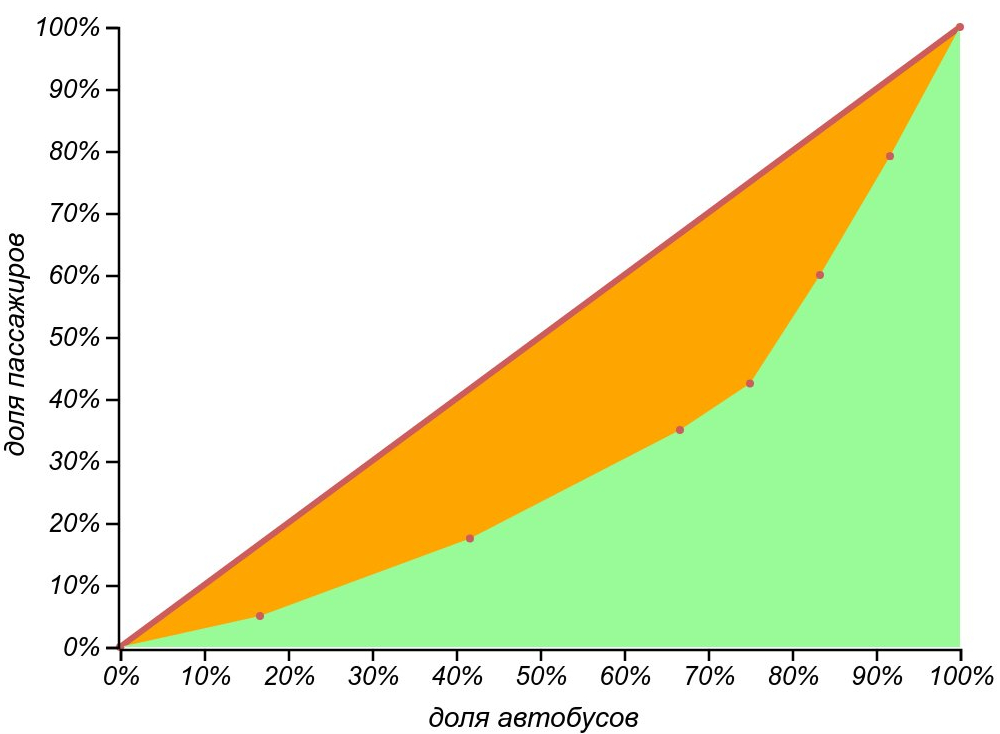
Приведём классический пример, с опросом пассажиров общественного транспорта. На линии в день работает множество автобусов, в относительно короткий час пик автобусы переполняются, а всё остальное время они ходят почти пустыми. Если мы станем опрашивать пассажиров, то значительная их часть окажется именно в переполненном автобусе (там попросту больше людей), и получим выражение общего недовольства. Если же мы опросим водителей, то они пожалуются на незаполненность значительной части маршрутов и неразумность начальства, гоняющего их попусту. Гибкий график сгладит ситуацию, но, в любом случае, кривая Лоренца будет отклоняться от кривой равенства, соответствующей невероятной ситуации всегда одинакового числа пассажиров во всех автобусах.

Во введениях в теорию вероятностей часто встречается специальный непрозрачный мешок, в который математики складывают разнообразные объекты, а потом наугад вытаскивают, делая, подчас, очень глубокомысленные выводы. Разрешение парадокса состоит в том, что анализируем мы систему пассажиропотока в целом и кладём в мешок автобусы, а проводя опрос, мы достаём из него наугад (инспектируем) пассажиров, и по их данным пытаемся делать выводы. Картинка показывает в чём тут разница:



*Статистика по автобусам говорит, что 75% из них свободна и ездит впустую. В то же время, опрос пассажиров обнаружит, что 64% пассажиров, проехавших в этот день, оказались в переполненном транспорте.*

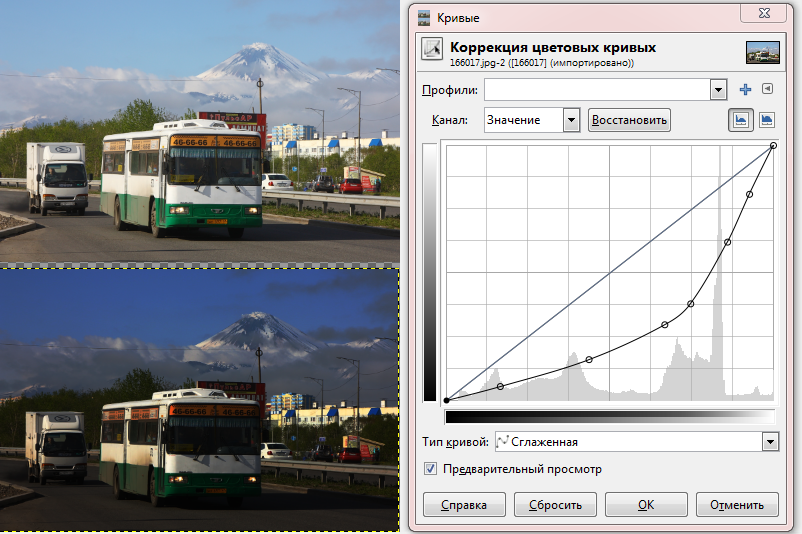
Давайте рассмотрим эту ситуацию, построив кривую Лоренца, на этот раз, настоящую, для числа пассажиров в автобусах из предыдущего рисунка. Для этого нужно отсортировать автобусы по числу пассажиров и последовательно суммировать вклад каждого из них в общий пассажиропоток:



*Кривая Лоренца хорошо иллюстрирует наблюдаемую несправедливость ситуации с автобусами: половина автобусов возит лишь пятую часть пассажиропотока.*

Кривая Лоренца, в данном случае, показывает как квантили распределения числа элементов в некоторых группах (горизонтальная ось) смещаются при анализе распределения элементов по принадлежности к группам (вертикальная ось). В этом, собственно, и состоит парадокс инспекции: картинка, которую наблюдает инспектор, оказывается искажённой, ведь он анализирует не группы, а элементы групп, а при этом наблюдаемые среднее значение и медиана смещаются в сторону более «весомого хвоста» распределения.

Сам по себе, наш закон велосипедиста очень прост, но он то и дело будет усугублять другие законы подлости, добавляя им угрюмую эмоциональную окраску. Размышляя о законах подлости, мне нравится думать об искажении восприятия мира инспектором в терминах изменения цветовых кривых какого-либо изображения. В растровых графических редакторах мы с помощью инструмента «Кривые» изменяем картинки, смещая распределение числа пикселов по яркости. Вот, например, как меняет восприятие реальности кривая Лоренца, полученная нами для автобусов. Картина мира становится мрачнее, как мы и ожидаем.



*Кривая Лоренца из примера, применённая в качестве фильтра «Кривая» в растровом графическом редакторе, делает видимую картину камчатского автобуса мрачнее. Сетуя на то, что автобусы «вечно опаздывают» и «вечно полны народу», утешайтесь тем, что, это всего лишь иллюзия, связанная с парадоксом инспекции!*

Парадокс инспекции может проявляться в своей крайности: если среди групп элементов, помещённых в наш теоретический мешок, есть такие, элементы которых не просто редки, но ненаблюдаемы вовсе, мы получаем *систематическую ошибку выжившего*. Об этом явлении часто рассказывают в различных демотивирующих статьях, для начинающих бизнесменов и программистов, уверяя их в том, что успешный путь, описываемый в книгах, скорее всего не для них, ибо, дескать, неуспешные книг не пишут. Впрочем, к законам подлости это отношения не имеет, так что оставим эти рассуждения. По большому счёту, описанные парадоксы являются методическими ошибками, допускаемыми при получении и обработке данных, о них знать полезно, но, к сожалению, они приводят к расхожему мнению о статистике, как о нечестном манипулировании фактическими данными, среди людей весьма далёких от этих методик.

Мы встретимся с законом велосипедиста и его влиянием ещё не раз: стоя в очереди или на автобусной остановке, наблюдая несправедливость распределения богатства. А кривые Лоренца и индекс подлости позволят нам смело сравнивать между собой возмутительно разные вещи. Математика — точная наука, но никто не запрещает математикам хулиганить. В своём, конечно, кругу и без драк.

1. Dorian M. Raymer and Douglas E. Smith, Spontaneous knotting of an agitated string, PNAS October 16, 2007 104 (42) 16432-16437; [↑](#footnote-ref-1)