Прелести чужой очереди

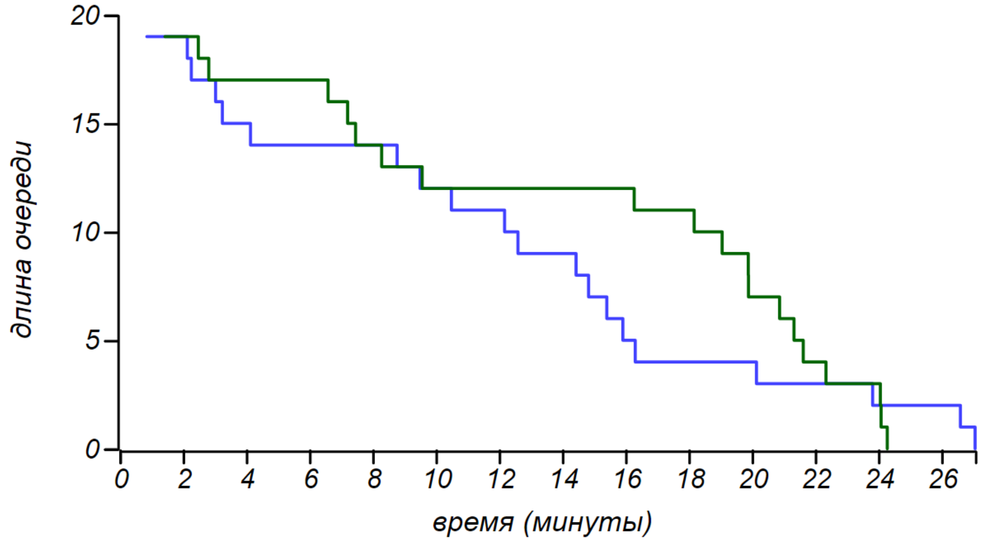
В этой главе мы научимся не скучать, стоя в длинном хвосте у стойки регистрации, узнаем о том, как застревают бумаги в кабинете у чиновника и каким образом можно испортить работу конторы, уволив нерасторопного работника.

Я размышляю о законах подлости, стоя в аэропорту в очереди на регистрацию пассажиров и оформление багажа. Очередь длинная, люди разные и заметные со всеми своими сумками, детьми или клетками. Сзади слышу ворчание: «Как обычно, наша очередь тормозит. Вон, гляди, тот усатый в кепке наравне с нами стоял, а теперь вон где... Вот ведь закон подлости!» Этот закон зовётся **наблюдением Этторе**:

Соседняя очередь всегда движется быстрее.

Что же это, психологический эффект или причуды математики?

Мы уже достаточно вооружены знаниями, чтобы немного проанализировать очередь, в которой стоим. За неимением других данных разумно предположить, что выход из очереди происходит по-пуассоновски, то есть через экспоненциально распределённые промежутки времени. Перемещения наблюдателя, стоящего в очереди, будет иметь вид монотонно изменяющейся ступенчатой линии, с одинаковыми шагами, случающимися через случайные промежутки времени. Такая зависимость от времени называется *пуассоновским процессом*. Пара примеров пуассоновских процессов приведена на рисунке внизу. Обычно, пуассоновский процесс накапливает количество событий и его изображение выглядит, как «лесенка», растущая со временем. Но стоя в очереди, мы заинтересованы в её скорейшем уменьшении, так что шаги нашего процесса ведут вниз.

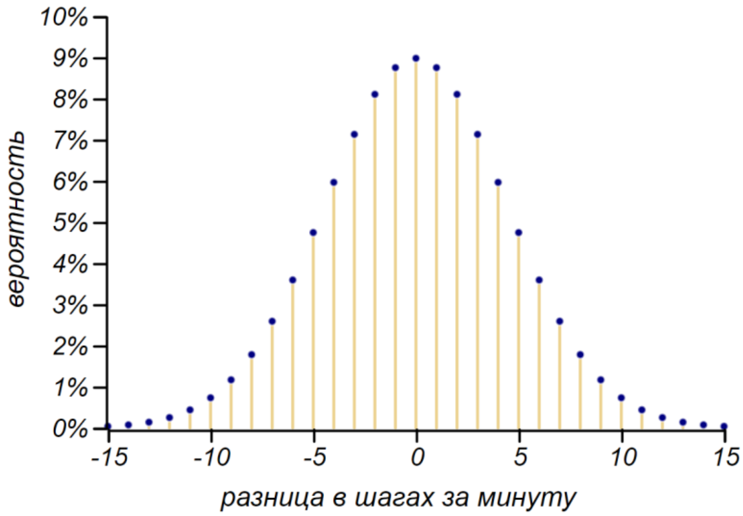


Перемещения двух очередей, как пуассоновских процессов с равной интенсивностью.

Разница двух одинаковых пуассоновских процессов, а именно её наблюдает человек, скучающий в хвосте и исследующий соседнюю очередь, представляет собой своеобразное случайное блуждание. В описанном нами случае, величина отставания одной очереди от другой подчиняется *распределению Скеллама*. Для двух одинаковых очередей, пропускающих μ человек в единицу времени, вероятность отставания одной из них на k шагов равна:

$$\mathrm{P}(k) = e^{-2\mu} I_{|k|}(2\mu),$$

где $I_k(x)$ – встречавшаяся нам в предыдущей главе модифицированная функция Бесселя. Эта функция возникла здесь не из-за круговой симметрии, а как результат сложения двух случайных величин, подчиняющихся распределению Пуассона.



Вероятность накопления разницы между двумя одинаковыми очередями со средней скоростью в 5 шагов в минуту.

Распределение Скеллама имеет симметричный колоколообразный вид, практически не отличимый от биномиального распределения, ну, а раз так, мы уже готовы сделать некоторые качественные выводы, основываясь на опыте, полученном в прошлой главе.

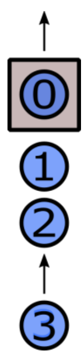
Вывод первый: расстояние между одновременно вставшими в одинаковые очереди людьми будет то увеличиваться, то уменьшаться, при этом, будут образовываться характерные меандры с постоянно меняющейся длительностью. Второй вывод: из-за самоподобия случайного блуждания, как для коротких очередей, так и для длинных, меандры будут иметь длительность, соизмеримую со временем стояния в очереди, и значит, они будут заметны, а меандры — это уже повод для недовольства. Третий вывод: заранее неизвестно какая очередь пройдёт быстрее, ведь случайное блуждание равновероятно уходит как вверх, так и вниз. И, наконец, четвёртое заключение: очереди движутся независимо, то и дело опережая и нагоняя друг друга, но в среднем, они движутся одинаково, и ожидаемая разница между ними стремится к нулю, однако разброс вокруг среднего со временем растёт пропорционально квадратному корню из времени.

Выходит, нет никаких подлых штучек со стороны злодейки судьбы, а есть одно только честное случайное блуждание. Правда, если нам не повезло оказаться во временно отстающей очереди, то мы в ней проведём больше времени, и, согласно закону велосипедиста, у нас будет больше возможностей посетовать на судьбу! А теперь, внимание, хорошие новости: в любой выбранный интервал времени тех, кому повезёт попасть в быструю очередь, будет больше чем невезунчиков, ведь быстрая очередь может пропустить больше людей! Но, увы, это вовсе не утешит того, кто надолго застрял в хвосте.

Теория для заскучавших в коридоре

Тем и хороша математика, что она способна сделать даже стояние в очереди увлекательным процессом. Например, можно прикинуть, сколько ещё предстоит ждать своей очереди, но для этого, как ни странно, надо посмотреть не вперёд, а назад, на растущий хвост. Если подождать какое-то время, скажем, 10 минут и посчитать, сколько человек выстроилось за вами, то разделив количество людей, стоящих перед вами на полученное число вы получите среднее время ожидания в десятках минут. Например, пусть за десять минут хвост вырос на 5 человек, если в момент подсчёта перед вами стоит семь человек, то ожидаемое время ожидания составит $10 \cdot 7/5 = 14$ минут. Понятно, что эта оценка будет весьма грубой, но любопытно, что она действительно соответствует среднему времени ожидания, об этом говорит теорема Литтла – один из самых ранних и самых общих результатов *теории очередей*, известной в России как *теории массового обслуживания*.

Теория очередей берёт своё начало в самом начале XX века, с первых работ Агнера Эрланга, который занимался только-только зарождающейся областью телекоммуникаций. За сотню лет результаты исследований Эрланга прочно вошли в нашу жизнь, настолько, что возникает ощущение того, что это мы вошли в мир телекоммуникаций. Несколько позже, большой вклад в развитие этой науки внёс советский математик Александр Яковлевич Хинчин, который вместе с А.Н. Колмогоровым заложил основы современной теории вероятностей. Результаты теории массового обслуживания важны для проектирования магазинов и залов ожидания, оптимального управления операционной системой в компьютере и операционным залом в банке, для грамотной разработки бюрократической машины, для управления дорожной сетью и в оценке рисков страховой компании. В очередях могут стоять люди (покупатели, клиенты, пассажиры), автотранспорт и грузы, задачи и документы; а обрабатывать их могут кассиры, операторы, регистраторы, серверы и бюрократы. Чтобы не путаться и не утопать в деталях, мы будем называть стоящих в очереди *клиентами*, а того, кто их обслуживает – *оператором*.

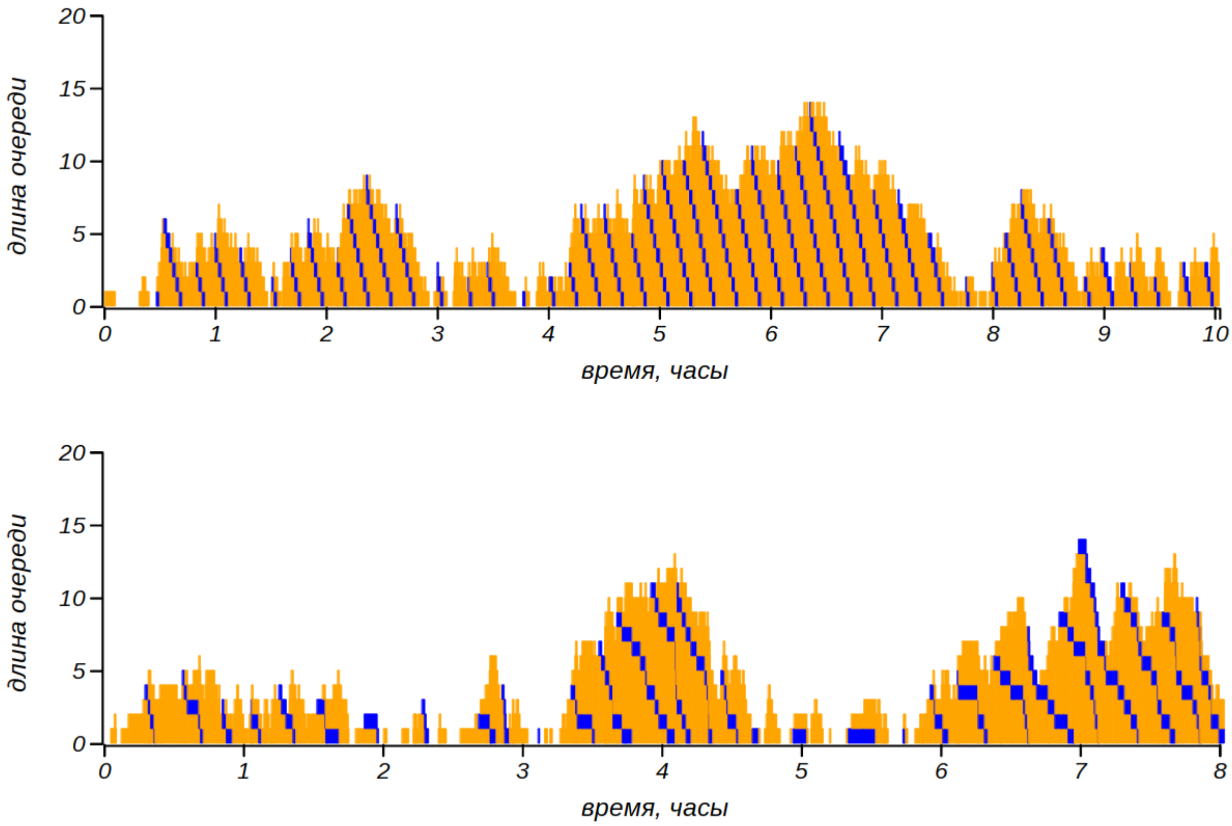


Представьте себе очередь, в которую люди встают, согласно некоторому распределению временных интервалов $p_{in}(t)$, со средним значением $1/\lambda$. Время, которое оператор тратит на работу с клиентами подчинено распределению $p_{out}(t)$ со средним значением $1/\mu$. Такую очередь описывают как марковский процесс в котором состояние определяется длиной очереди: состояние 0 — в очереди нет никого, состояние 1 — стоит один клиент, состояние 2 — два клиента и так далее… В идеальном мире, ничто не запрещает очереди вырасти сколь угодно длинной, значит, мы получаем цепь с *бесконечным числом состояний* и для анализа очереди придётся иметь дело с матрицей переходов, содержащей бесконечное число строк и столбцов. В предыдущей главе мы уже имели дело с марковскими процессами, и для анализа стационарного состояния цепи нам понадобилось возводить матрицу переходов в бесконечную степень. Так что же надо будет вычислить *бесконечную матрицу, возведённую в бесконечную степень*? Математиков эта задача не испугала и уже в тридцатые годы XX века были придуманы методы для таких вычислений. Результатом этого анализа будут свойства *стационарного состояния* очереди. Это состояние не меняется со временем, но все параметры очереди, такие как её длина или время ожидания в ней представляют собой случайные величины. То есть, могут постоянно меняться, но при этом постоянно остаются в рамках каких-то распределений, от времени не зависящих. И чего только не придумаешь, скучая в зале ожидания!

Свойства очереди сильно зависят от соотношения λ и μ. Если $\lambda > \mu$, то хвост будет расти неограниченно, как пробка на дороге, в которую въезжает больше автомобилей, чем может выехать. Такая очередь попросту перекрывает поток клиентов, накапливая их внутри себя. Для $\lambda < \mu$ очередь является *устойчивой*. Она может расти или уменьшаться, но клиенты в ней не исчезают и не появляются, а значит, сколько их вошло в зону ожидания, столько и выйдет. То есть, устойчивая очередь может затормозить тех, кто в ней стоит, но она не может изменить интенсивность потока людей, проходящих сквозь неё. И если на входе мы имеем в среднем λ человек в единицу времени, то и на выходе мы должны получить такой же поток, независимо от скорости работы оператора. Случай $\lambda \approx \mu$ рассматривается отдельно, такая метастабильная очередь ведёт себя неустойчиво и моделируется даже не марковским процессом, а процессом случайного блуждания, с той только разницей, что длина очереди не может быть отрицательной. У блуждающей таким образом системы есть непроницаемая стенка снизу, впрочем, она не мешает практически неограниченному росту длины очереди. И, хотя, рано или поздно, она сократится и даже исчезнет, отклонения времени ожидания и времени работы оператора от среднего будут столь велики, что счесть такое обслуживание удовлетворительным никак не получится.

Далее мы будем рассматривать только учтойчивые очереди. От характера распределений $p_{in}(t)$ и $p_{out}(t)$ зависит динамика очереди и такие её характеристики как распределения для её длины, времени ожидания клиентом и времени занятости оператора. Для очередей создана целая номенклатура, называемая *формулой Кэндэлла*. Например, простая очередь, в которую люди подходят равномерно и равномерно же уходят, как, например, в аэропорту при посадке на рейс, обозначается D/D/1 (D здесь обозначает дельта-распределение, а единица – одного оператора). Въезд и выезд автомашин на территорию аэропорта через три автоматических шлагбаума можно описать очередью M/D/3 – буквой M обозначается пуассоновский (марковский) процесс, то есть случайный процесс без памяти. В очередь на регистрацию билетов и оформление багажа новые люди приходят по-пуассоновски, и багаж у всех разный, так что клиенты будут выходить из очереди тоже по-пуассоновски, для пяти стоек такая очередь обозначается M/M/5. Если же на входе или выходе процесс перестаёт быть марковским, он обозначается буквой G (general).

В этой главе мы будем исследовать неприятности и неожиданности, наблюдаемые в очередях на примере очереди с $\lambda = 30$ и $\mu = 34\,\text{человек/час}$. В среднем, новые клиенты будут поступать в неё с интервалом в 2 мин, а обрабатываться оператором за 1 минуту 45 секунд. Это похоже на очередь у стойки регистрации в аэропорту. На рисунке показан пример того, как могут «жить» M/D/1- и M/M/1-очереди с такими параметрами.



Динамика MD1 и MM1 очередей. Синим цветом выделены траектории каждого седьмого клиента в очереди. Длина очереди склонна к своеобразным колебаниям, она «дышит», то удлинняясь, то сокращаясь, оставаясь при этом стационарной.

В стационарном, состоянии длина M/M/1-очереди, описывается *геометрическим распределением*:

$$P(\text{длина очереди} = n) = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n,$$

мы встречали его в предыдущей главе, рассматривая простейшую несимметричную марковскую цепь. Зная это распределение, можно вычислить ожидаемую длину $L = \lambda/(\mu - \lambda)$. Для нашего примера средняя длина очереди составит 7.5 человек. *Время обслуживания клиента* (то есть сумма времени ожидания своей очереди и, собственно, времени работы с оператором), в M/M/1-очереди описывается экспоненциальным распределением с параметром $\mu - \lambda$. Это приводит к значению среднего времени ожидания $W = 1/(\mu - \lambda)$. Не смотря на то, что среднее время работы с каждым клиентом не превышает двух минут, среднее время ожидания на нашего примера равно 15 минутам. Как видно, для стационарной M/M/1-очереди выполняется равенство:

$$\lambda W = L.$$

Это и есть формула Литтла, которой мы воспользовались, стоя в очереди и от нечего делать, занявшись подсчётами. Будучи очень простой, эта формула на удивление сильна, она выполняется для очень широкого класса очередей и в самых разных задачах. То, что в формулу Литтла входит только λ, а не μ отражает основное свойство стабильной очереди, она может задерживать клиентов, но не меняет их поток, который определяется λ. И даже если скорость работы оператора μ будет очень велика, среднее время ожидания будет всё равно определяться входным потоком и уже скопившимся числом клиентов. А так как для устойчивых очередей $\lambda < \mu$, то мы получаем ещё один закон подлости:

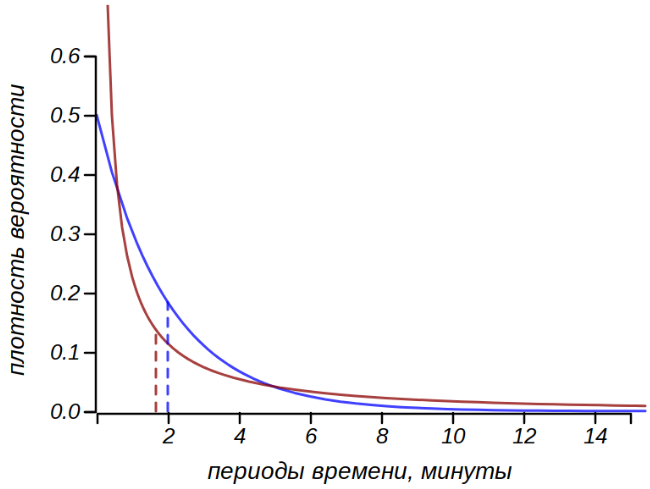
Даже с идеальным кассиром, время стояния в очереди в кассу определяется бестолковыми покупателями.

Важной характеристикой очереди является *время занятости оператора* B, то есть длительность непрерывных периодов времени, в которые оператор обслуживает клиентов. Такие периоды перемежаются *периодами простоя*, когда по какой-то причине клиентов в очереди не оказывается. Клиенты приходят, ждут и уходят, а оператор остаётся работать, так что разумно предположить, что $B > W$. В действительности, ожидаемое, то есть, среднее время занятости для M/M/1-очередей равно среднему времени ожидания, то есть, $B = W$. Это уже кажется не вполне интуитивным результатом, однако, и это ещё не всё: при той же интенсивности работы оператора, среднее время обслуживания клиента может стать существенно больше среднего времени работы оператора!

Тут всё дело в дисперсии распределения времени обслуживания одного клиента $p_{out}(t)$. Ещё в 1930-е годы австрийскому математику Феликсу Поллачеку удалось в общем виде вычислить отношение W/B для произвольной M/G/1-очереди:

$$\frac{W}{B} = 1 + \frac{\lambda}{2\mu}(\mu^2\sigma^2 - 1).$$

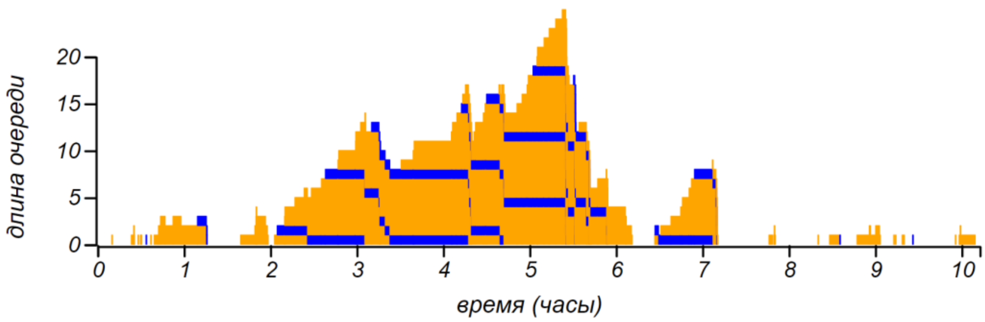
здесь σ – дисперсия распределения $p_{out}(t)$. В случае M/M/1-очереди $\sigma = 1/\mu$ и это отношение равно единице. Но может так случиться, что при том же значении среднего распределение $p_{out}(t)$ будет иметь большую дисперсию, и тогда W окажется больше B. На рисунке показан пример, в котором $p_{in}(t)$ распределено экспоненциально с $\lambda = 30\,\text{человек/час}$, а $p_{out}(t)$ описывается гамма-распределением, соответствующим интенсивности $\mu = 34\,\text{человек/час}$, с дисперсией $\sigma = 2/\mu$.



Распределения для периодов времени между появлением новых клиентов (синяя линия, экспоненциальное распределение) и времени обслуживания одного клиента (красная линия, гамма-распределение).

Очередь остаётся стабильной: клиенты в среднем обслуживаются быстрее, чем приходят новые. Оператор работает хорошо: большая часть клиентов обслуживается очень быстро, но обратите внимание на то, что велика доля «трудных» клиентов, которые формируют достаточно толстый хвост распределения. Их немного, но каждый из них отнимает много времени и все в очереди вынуждены их ждать. Для примера, приведённого на рисунке, среднее время ожидания оказалось равно 35 минутам, тогда как среднее время занятости оператора осталось прежним (15 минут). Получается, не переставая работать, оператор, в среднем, филонит, пока мы страдаем в очереди от безделья!

Динамика такой очереди отличается от динамики M/M/1. Для неё характерен несимметричный пилообразный рисунок с плавной восходящей линией и резким сбросом. Пока оператор занят «трудным» клиентом постепенно вырастает длинный хвост, а потом, освободившись, оператор очень быстро с ним справляется.

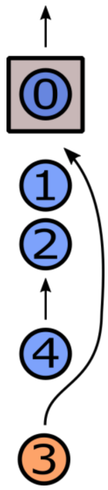


Динамика M/G/1-очереди, в которой время ожидания клиентов вдвое превосходит время занятости оператора. Горизонтальные полосы показывают периоды долгого ожидания очередного трудного клиента.

Мне только спросить!

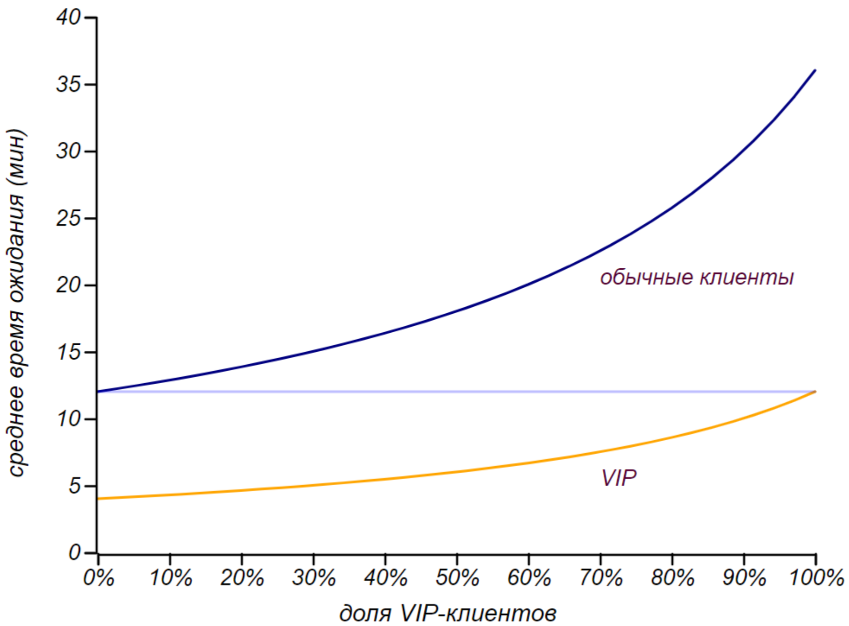
Есть в нашей жизни такое досадное явление – «обочечники» – ушлые водители, объезжающие пробку по обочине и потом встревающие в поток. Есть настырные посетители поликлиник и касс, норовящие просочиться к окошку или двери с заветной формулой

«Мне только спросить…». В любую отлаженную бюрократическую систему то и дело врываются неотложные дела, не терпящие промедления. Понятно, что иногда без таких случаев не обойтись: в больницах бывают неотложные пациенты, в операционной системе компьютера есть задачи с очень высоким приоритетом, наконец, на дороге мы обязаны пропускать спецтранспорт, едущий по экстренному случаю. Но как внеочередники влияют на всю очередь? Подобные случаи моделируются *очередями с приоритетом* и для них тоже есть развитая теория, поскольку в жизни они встречаются чуть ли не чаще простых очередей.



Пусть в нашей M/M/1-очереди с вероятностью ε могут появляться особые клиенты, назовём их VIP (very impatient person – очень нетерпеливые персоны), которые встают не в конец очереди, а вклиниваются в её начало, заставляя ждать всех стоящих позади. При этом, они, всё же, дают оператору завершить работу с текущим клиентом, не прерывая его. Если внеочередников наберётся несколько, они могут образовать свою VIP-очередь. Вспомним, что пуассоновский поток можно представить как случайное «расбрасывание» по временному интервалу какого-то известного количества событий. Так как все клиенты приходят совершенно независимо, то согласно нашему условию, мы получим поток нетерпеливых клиентов $\varepsilon\lambda$ и поток обычных клиентов $(1-\varepsilon)\lambda$, при этом, общий их поток останется неизменным. Среднее время ожидания для VIP будет равно $W_{VIP} = 1/(\mu - \varepsilon\lambda)$, как в простой M/M/1-очереди, поскольку они в своей VIP-очереди «не замечают» присутствия обычных клиентов. Для того, кто ждёт на общих основаниях, время ожидания вырастет, и составит уже:

$$W_0 = \frac{\mu}{(\mu-\varepsilon\lambda)(\mu-\lambda)} = \frac{\mu}{(\mu-\lambda)} W_{VIP}$$



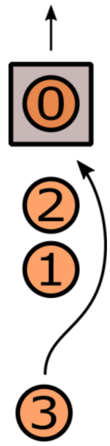
Соотношение средних времён ожидания для очереди с нетерпеливыми VIP-клиентами.

Пока VIP-ов немного, очереди они мешают не сильно, но если доля внеочередников оказывается близкой к единице, то никакого преимущества они уже не имеют, зато немногочисленным скромным очередникам приходится ждать существенно дольше. При

$\varepsilon = 1$ среднее время ожидания становится равным $\mu/(\mu - \lambda)^2$ (больше двух часов в нашем случае!) и, вообще, если μ лишь немного превышает λ, очередь остаётся устойчивой, однако время ожидания в ней вырастает катастрофически!

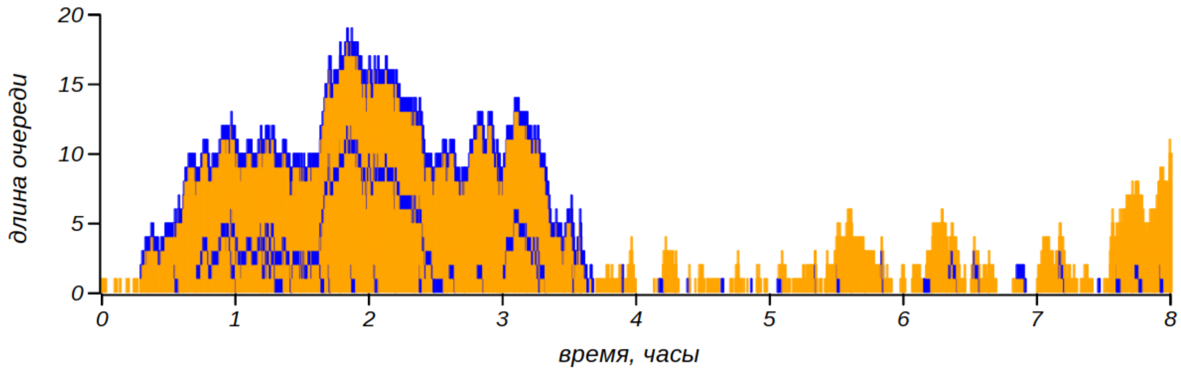
Но вот, что любопытно. Можно найти среднее время ожидания для всей группы клиентов, как взвешенную сумму $\varepsilon W_{VIP} + (1 - \varepsilon) W_0$ и она окажется равной $1/(\mu - \lambda)$, то есть, такой же, как для обыкновенной M/M/1-очереди без всяких VIP-ов. Выходит, системе в целом внеочередники не мешают. На время занятости оператора они тоже не влияют и распределения времён ожидания остаётся экспоненциальным. Мы уже говорили в предыдущей главе, что для экспоненциального распределения кривая Лоренца и, соответственно, коэффциент Джини не зависят от параметра распределения. А значит, все M/M/1-очереди имеют одинаковую степень несправедливости – 0.5. Отсюда следует, что наш обобщённый критерий несправедливости для всех ожидающих в очереди также останется равным 0.5.

Стационарный бардак



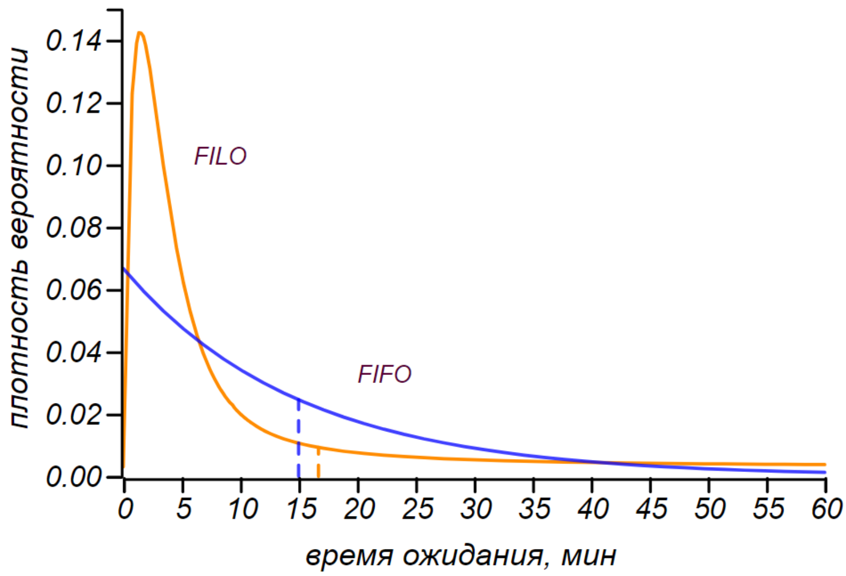
А теперь немного изменим политику очерёдности. Пусть внеочередники будут сверхнаглыми, и если так случится, что один такой клиент придёт вслед за другим, то вместо формирования нормальной очереди, второй вклинится перед первым. Эта задача уже отличается от классического подхода к очередям с приоритетом. Давайте сразу рассмотрим предельный случай, когда доля наглых клиентов равна единице. Тогда наша очередь превращается в то, что программисты называют *стеком* – в последовательность элементов, подчиняющуюся правилу «*первым вошёл, последним вышел*» (FILO – first in, last out) в противовес очереди, для которой выполняется правило «*первым вошёл, первым вышел*» (FIFO – first in, first out).

Такая «очередь наоборот» выглядит неестественно, но если вместо людей мы рассмотрим пачку документов, тогда узнаем знакомую картину на рабочем столе, когда входящие документы не сортируются по времени, а просто складываются в стопку по мере поступления, а потом обрабатываются, начиная сверху. Удивительно, но в стационарном состоянии все средние значения основных параметров: и длины очереди и времени ожидания и времени занятости оператора, будут точно такими же, как и в FIFO-очереди. Что же поменяется? Давайте посмотрим на пример работы такой очереди, он показан на рисунке. Мы видим, что вместо целенаправленного движения к оператору, клиенты могут двигаться, то к нему, то от него. Время ожидания для самого последнего клиента существенно удлиняется, однако пока он ждёт, через оператора проходит большое число вновь поступающих клиентов, которые обрабатываются почти мгновенно. В среднем же, мы получаем время ожидания, примерно такое же, как для «нормальной» очереди. Но мы уже много раз убеждались в том, что среднее значение не может характеризовать случайную величину в полной мере.



Динамика FILO-очереди, или стопки документов, которые при поступлении, кладутся наверх и обрабатываются, начиная сверху.Как и прежде, цветом выделен каждый седьмой клиент.

Глядя на динамику FILO-очереди легко понять, что время ожидания клиента должно быть близким к времени занятости оператора. Действительно, время занятости определяется как период от момента прихода первого клиента до момента выхода последнего, но в стеке первый клиент и является последним. Нужно ещё учесть, что очередной клиент не прерывает оператора, и поэтому к времени его ожидания добавится время обслуживания клиента, с которым уже работает оператор. Если это время распределено экспоненциально, то, как уже обсуждалось в связи со временем ожидания автобуса, добавочное время будет распределено точно также. В конечном итоге, время ожидания будет распределено как сумма времени занятости оператора и периода работы с одним клиентом. На рисунке показаны распределения времени ожидания для M/M/1-очередей, обыкновенной -- с политикой FIFO и придерживающейся правила FILO. В обоих случаях $\lambda = 30$ и $\mu = 34$ человека в час.

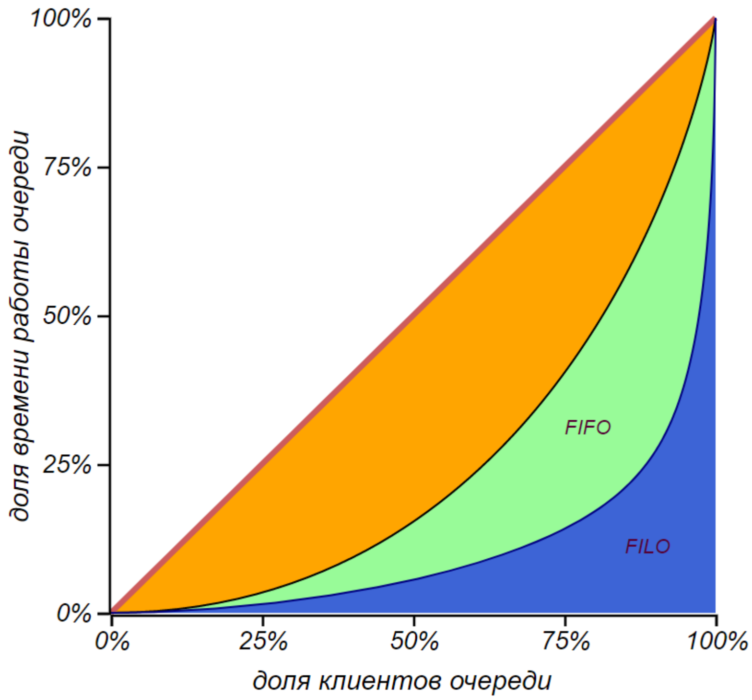


Распределения времени ожидания для M/M/1-очередей с различной политикой.

Распределения сильно отличаются, но средние значения у них практически одинаковые. Несмотря на то, что распределение времени ожидания для FILO-очереди кажется сконцентрированным около моды (близкой к $1/\mu$, то есть, времени работы с одним клиентом) у него длинный тяжёлый хвост, который сильно увеличивает дисперсию и увеличивает среднее значение. Медиана этого распределения равна 3 минутам, это значит, что в половине случаев клиент будет ждать совсем немного, но если уж застрянет, то застрянет: 5% клиентов потратят на ожидание больше часа, а самые невезучие два процента вместо 2 минут вынуждены будет ждать своей очереди больше 2 часов! Для FIFO-очереди с такими же параметрами вероятность застрять на 2 часа составляет не более 0.04%.

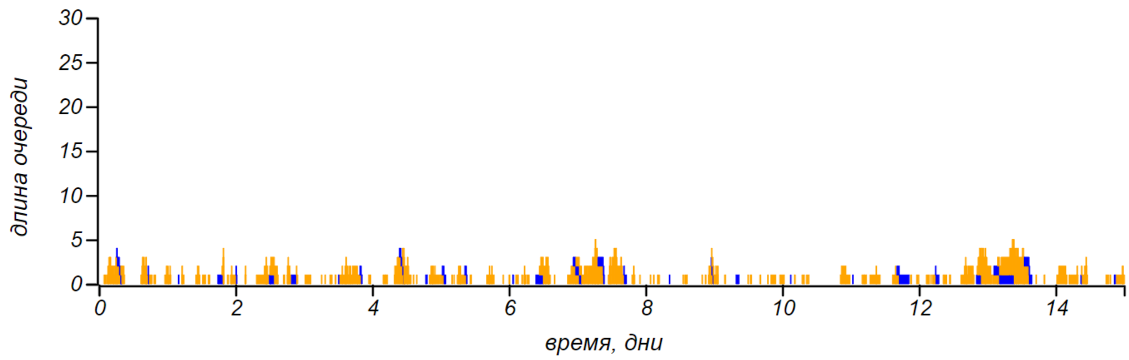
При этом оператор, то есть, бюрократ, обрабатывающий бумаги не заметит разницы между очередью и стеком, распределение его времени занятости не изменится. Начальник бюрократа тоже увидит, что из кабинета подчинённого бумаги выходят с нормальной интенсивностью в силу устойчивости очереди. И даже большая часть документов окажется обработанной очень оперативно. Но то и дело какая-то часть документов внезапно «проваливается» на дно стопки и задерживается там очень и очень надолго. Такая же картина наблюдается и в шкафу, в который мы складываем вещи с мыслью разобрать потом. Но потом, мы задвигаем то, что уже лежит в шкафу поглубже и добавляем в него новые вещи. Так что даже если мы, всё же, станем их постепенно разбирать, до «ископаемых» у самой стенки руки дойдут нескоро.

Для демонстрации несправедливости распределения времени между различными делами в ведомстве бюрократа (или вещами в нашем шкафу), изобразим кривые Лоренца для FIFO и FILO очередей, описываемых формулой M/M/1. Если для всех FIFO-очередей кривая Лоренца одинакова, то несправедливость FILO-очередей зависит от соотношения λ и μ. Чем их отношение ближе к единице, тем ближе к единице оказывается и индекс Джини. На рисунке показаны кривые Лоренца для этих двух типов очередей.



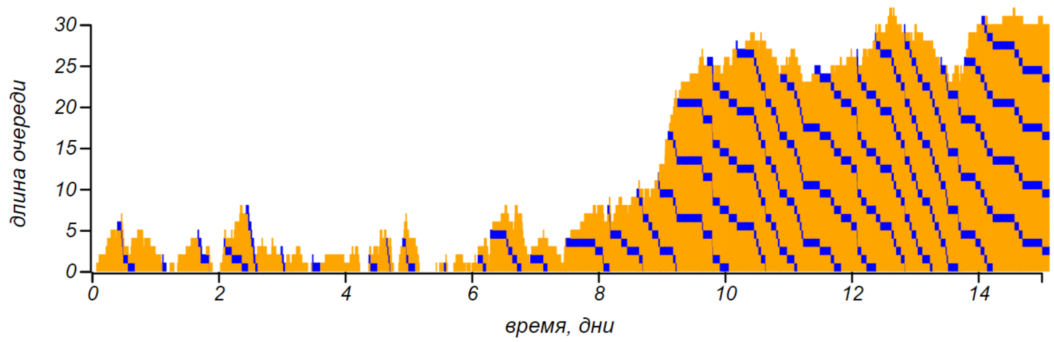
Кривые Лоренца для времени ожидания в двух типах очередей. Коэффициент Джини для FIFO-очереди равен 0.5, а для FILO-очереди – 0.78.

Закон сохранения клиентов, то есть, равенство входного и выходного потока может сыграть ещё одну злую шутку. Представьте себе контору, через которую проходит в рабочий день, скажем, 15 человек, при этом, на работу с каждым клиентом, в среднем, уходит полчаса. В конторе работают два клерка и каждый, в принципе, в состоянии в одиночку справиться с таким потоком. Один трудится с интенсивностью 16 человек в день, а второй – 14 человек в день. Вместе они могли бы обслужить человек 30, но, кажется, столько и не нужно. Люди в такой конторе почти не ждут своей очереди, среднее время ожидания для клиента составляет 18 минут, средняя длина очереди всего 1.2 человека[1], из которых с одним клиентом уже работает кто-либо из клерков, второй клерк в это время отдыхает. На рисунке показан пример динамики очереди в конторе. При этом надо иметь в виду, что этот незначительный поток распределён между двумя операторами, то есть каждый из них наблюдает поток в ещё два раза менее интенсивный. Про такую работу говорят: «Не бей лежачего».



Две недели в конторе с двумя клерками.

Начальство, проведя все эти замеры и наблюдения, полагает, что клерки уж больно вольготно живут, тратя на работу только половину рабочего времени, и в стремлении оптимизировать работу учреждения, увольняет нерасторопного клерка. И вот всё стало совсем по-другому! Система приблизилась к опасному состоянию, когда $\mu \approx \lambda$. При таких условиях очередь становится метастабильной, она может какое-то время вести себя «хорошо», а потом внезапно превратиться в совершеннейший коллапс.

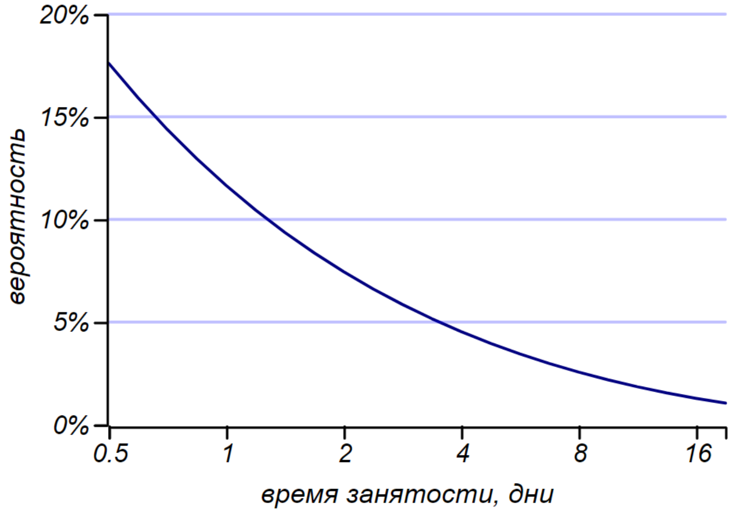


В течение недели один клерк вполне справлялся с объемом работ,
но потом всё превратилось в кошмар.

При $\lambda = 15$ и $\mu = 16$ средняя длина очереди будет как раз равна 15 клиентам, а среднее время занятости оператора составит 1 день. Начальство может быть довольно своей

[1] Параметры стационарной M/M/2 очереди можно рассчитать по общим формулам, которые мы здесь не приводим из-за их громоздкости.

оптимизацией. Но мы-то знаем, что средние показатели не показывают почти ничего. Посмотрите, с какой вероятностью время занятости клерка превысит указанное количество дней.



Вероятность для одного клерка не уложиться с текущими делами в указанный период времени.

Что ещё хуже, среднее время ожидания одного клиента вырастает тоже до одного дня! Вместо 18 минут, клиент застрянет со своим делом на час с вероятностью 88%, вероятность проторчать в конторе полдня составит 60%, а ухлопать на это весь день – 37%. Таким образом, разумное, как кажется, решение, может иметь неожиданно неприятные последствия.