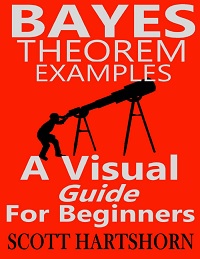
**Скотт Хартшорн. Теорема Байеса для начинающих с примерами в Excel**

Понимание (изучение) вероятностей начинается там, где заканчивается классический курс теории вероятностей. Почему-то в школе и вузе преподают частотную (комбинаторную) вероятность, или вероятность того, что определено. Человеческий мозг устроен иначе. У нас имеются теории (мнения) по поводу всего на свете. Мы субъективно оцениваем вероятность тех или иных событий. Мы также можем изменить свое мнение, если произошло нечто неожиданное. Это то, что мы делаем каждый день. Например, если вы встречаетесь с подругой у памятника Пушкину, вы понимаете, будет ли она вовремя, опоздает на 15 минут или полчаса. Но выйдя на площадь из метро, и увидев 20 см свежего снега, вы обновите свои вероятности, чтобы учесть новые данные.

Такой подход был впервые описан Байесом и Лапласом. Хотя Лаплас [сделал это позже](http://baguzin.ru/wp/?p=16017), я думаю, что он не был знаком с работой Байеса. По непонятной мне причине байесовский подход довольно слабо представлен в русскоязычной литературе. Для сравнения отмечу, что по запросу Байес [Ozon](https://www.ozon.ru/?context=search&text=%e1%e0%e9%e5%f1&group=div_book) выдает 4 ссылки, а [Amazon](https://www.amazon.com/s/ref=nb_sb_ss_i_5_5?url=search-alias%3Dstripbooks&field-keywords=bayes&sprefix=Bayes%2Cstripbooks%2C254&crid=23Y2QS4E4LKJ0&rh=n%3A283155%2Ck%3Abayes) – около 1000.

Настоящая заметка является переводом небольшой английской книги, и даст вам интуитивное понимание того, как использовать теорему Байеса. Она начинается с определения, а далее использует примеры в Excel, которые позволят отслеживать весь ход рассуждений.

Scott Hartshorn. Bayes' Theorem Examples: A Visual Guide For Beginners. – 2016, 82 p.



### Определение теоремы Байеса и интуитивное объяснение

Теорема Байеса

где A и B – события, P(A) и P(B) – вероятности A и B без учета друг друга, P(A|B) – условная вероятность события А при условии, что B истинно, P (B|A) – условная вероятность B, если А истинно.

На самом деле, уравнение несколько сложнее, но для большинства применений достаточно и этого. Результат вычислений – это просто нормализованное взвешенное значение на основе первоначального предположения. Итак, возьмите первоначальное предположение, взвесьте его по отношению к другим первоначальным возможностям, нормализуйте на основе наблюдения:



В ходе решения проблем мы будем выполнять следующие шаги (далее они станут понятнее):

1. Определите, какую из вероятностей мы хотим вычислить, а какую мы наблюдаем.
2. Оцените начальные вероятности для всех возможных вариантов.
3. Предположив истинность некоего начального варианта, рассчитайте вероятность нашего наблюдения; и так для всех начальных вариантов.
4. Найдите взвешенную величину, как произведение начальной вероятности (шаг 2) и условной вероятности (шаг 3), и так для каждого из начальных вариантов.
5. Нормализуйте результаты: разделите каждую взвешенную вероятность (шаг 4) на сумму всех взвешенных вероятностей; сумма нормализованных вероятностей = 1.
6. Повторите шаги 2–5 для каждого нового наблюдения.

### Пример 1. Простой пример с костями

Предположим, у вашего друга есть 3 кости: с 4, 6 и 8 гранями. Он случайным образом выбирает одну из них, не показывает вам, бросает и сообщает результат – 2. Вычислите вероятность того, что был выбран 4-гранник, 6-гранник, 8-гранник.

Шаг 1. Мы хотим вычислить вероятность выбора 4-гранника, 6-гранника или 8-гранника. Мы наблюдаем выпавшее число – 2.

Шаг 2. Поскольку костей было 3, исходная вероятность выбора каждой из них – 1/3.

Шаг 3. Наблюдение – кость упала гранью 2. Если был взят 4-гранник, шансы этого равны 1/4. Для 6-гранника шансы выпадения 2-ки – 1/6. Для 8-гранника – 1/8.

Шаг 4. Выпадение 2-ки для 4-гранника = 1/3 \* 1/4 = 1/12, для 6-гранника = 1/3 \* 1/6 = 1/18, для 8-гранника = 1/3 \* 1/8 = 1/24.

Шаг 5. Общая вероятность выпадении 2-ки = 1/12 + 1/18 + 1/24 = 13/72. Это число меньше 1, потому что шансы бросить 2-ку меньше 1. Но мы знаем, что уже бросили именно 2-ку. Таким образом, нам нужно разделить шансы каждого варианта из шага 4 на 13/72, чтобы сумма всех шансов для всех костей лечь 2-ой равнялась 1. Этот процесс называется нормализацией.

Нормализуя каждую взвешенную вероятность, мы находим вероятность того, что именно эта кость была выбрана:

* 4-гранник = (1/12) / (13/72) = 6/13
* 6-гранник = (1/18) / (13/72) = 4/13
* 8-гранник = (1/24) / (13/72) = 3/13

И это ответ.

Когда мы начали решать задачу, мы предположили, что вероятность выбрать определенную кость равна 33,3%. После выпадения 2-ки, мы рассчитали, что шансы, что первоначально был выбран 4-гранник выросли до 46,1%, шансы выбора 6-гранника снизились до 30,8%, а шансы, что был выбран 8-гранник и вовсе упали до 23,1%.

Если сделать еще один бросок, мы могли бы использовать новые рассчитанные проценты в качестве наших начальных предположений и уточнить вероятности на основе второго наблюдения.

Если у вас единственное наблюдение, все шаги удобно представить в виде таблицы:



Таблица. 1. Пошаговое решение в виде таблицы (формулы см. в файле Excel на листе *Пример 1*)

Обратите внимание:

* Если бы вместо 2-ки выпала, например, 7-ка, то шансы на шаге 3 у 4- и 6-гранника равнялись бы нулю, и после нормализации шансы 8-гранника составили бы 100%.
* Поскольку пример включает лишь три кости и один бросок, мы использовали простые дроби. Для большинства проблем с большим количеством вариантов и событий легче работать с десятичными дробями.

### Пример 2. Больше костей. Больше бросков

На этот раз у нас 6 костей с 4, 6, 8, 10, 12 и 20 гранями. Мы выбираем одну из них случайным образом и бросаем 15 раз. Какова вероятность того, что была выбрана определенная кость?

Я использую модель в Excel (рис. 1; см. лист *Пример 2*). Случайные числа генерируются в столбце B с помощью функции =СЛУЧМЕЖДУ(1;$B$9). В данном случае в ячейке В9 выбран 8-гранник, поэтому случайные числа могут принимать значения от 1 до 8. Поскольку Excel обновляет случайные числа после каждого изменения на листе, я скопировал столбец В в буфер и вставил только значения в столбец C. Теперь значения не меняются и будут использоваться для последующих рисунков. (Я добавил вам возможность «поиграть» с выбором числа граней и случайными бросками на листе *Пример 2 игровой*. Особенно любопытные результаты получаются, если в ячейке В9 установить число 13 ☺ – *Прим. Багузина*.)

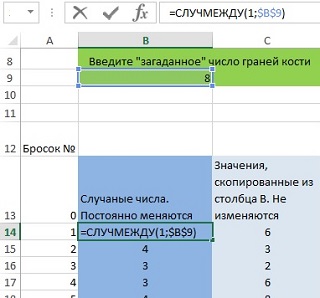


Рис. 1. Генератор случайных чисел

Шаг 2. Поскольку всего шесть кубиков, то вероятность выбрать один случайным образом равна 1/6 или 0,167.

Шаги 3 и 4. Запишем уравнение для вероятности первоначального выбора определенной кости после соответствующего броска. Как мы видели в конце примера 1, некоторые броски могут не соответствовать тем или иным костям. Например, выпадение 9-ки делает вероятность 4-, 6- и 8-гранной кости равной нулю. Если же выпало «легитимное» число, то его вероятность для данной кости равна единице, деленной на число граней. Для удобства мы объединили шаги 3 и 4, поэтому мы сразу запишем формулу для вероятности броска, умноженной на нормализованную вероятность после предыдущего броска (рис. 2):

= ЕСЛИ(бросок > числа граней; 0; 1/число граней \* предыдущая нормализованная вероятность)

Если вы аккуратно воспользуетесь [относительными и абсолютными ссылками](http://baguzin.ru/wp/?p=2251), то сможете протащить эту формулу на все строки.



Рис. 2. Уравнение вероятности

Шаг 5. Последним шагом является нормализация результатов после каждого броска (область L11:R28 на рис. 3).

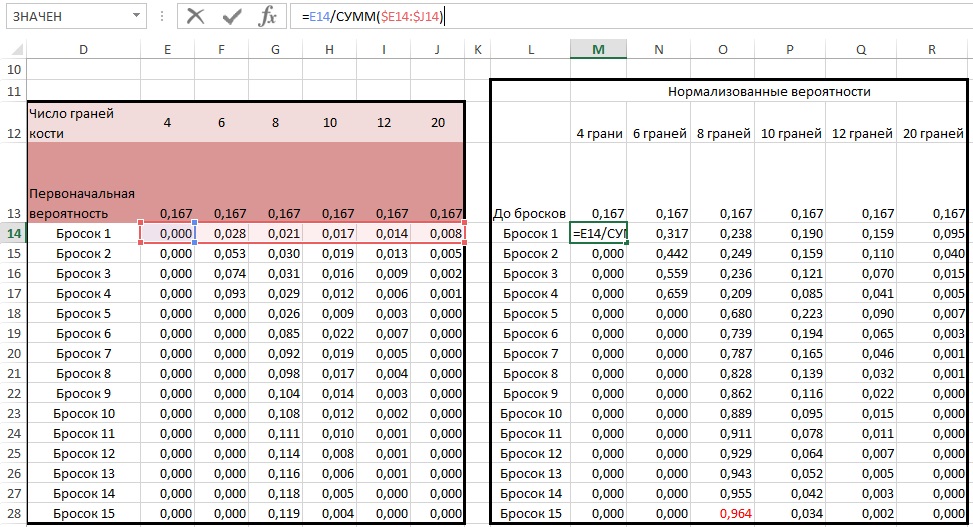


Рис. 3. Нормализация результатов

Итак, после 15 бросков с вероятностью 96,4% мы можем считать, что первоначально выбрали 8-гранную кость. Хотя остаются шансы, что была выбрана кость с б***о***льшим числом граней: 3,4% – за 10-гранную кость, 0,2% – за 12-гранную, 0,0001% – за 20-гранную. А вот вероятность 4- и 6-гранных костей равна нулю, так как среди выпавших чисел были 7 и 8. Это, естественно, соответствует тому, что мы ввели число 8 в ячейку В9, ограничив значения для генератора случайных чисел.

Если мы построим график вероятности каждого варианта первоначального выбора кости, бросок за броском, то увидим (рис. 4):

* После первого броска вероятность выбора 4-гранной кости падает до нуля, так как сразу же выпала 6-ка. Поэтому лидерство захватил вариант 6-гранной кости.
* Для нескольких первых бросков 6-гранная кость имеет наибольшую вероятность, так как она содержит меньше всего граней среди костей, которые могут отвечать выпавшим значениям.
* На пятом броске выпала 8-ка, вероятность 6-гранника падает до нуля, и 8-гранник становится лидером.
* Вероятности 10-, 12- и 20-гранных костей при первых бросках плавно уменьшались, а затем испытали всплеск, когда 6-гранная кость выпала из гонки. Это связано с тем, что результаты были нормализованы по гораздо меньшей выборке.

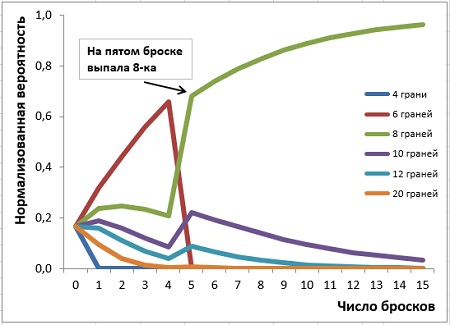


Рис. 4. Изменение вероятностей бросок за броском

Обратите внимание:

* Теорема Байеса для нескольких событий – просто повторное умножение на последовательно обновляемых данных. Окончательный ответ не зависит от того, в каком порядке наступали события.
* Не обязательно нормализовать вероятности после каждого события. Можете сделать это один раз в самом конце. Проблема в том, что, если не заниматься нормализацией постоянно, вероятности становятся такими маленькими, что Excel может работать некорректно из-за ошибок округления. Таким образом, практичнее нормализовывать на каждом шаге, чем проверять, не подошли ли вы к границе точности Excel.

### Теорема Байеса. Терминология

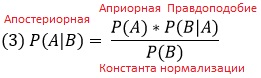
* Начальная вероятность, вероятность каждой возможности до того, как произошло наблюдение, называется *априорной*.
* Нормализованный ответ после вычисления вероятности для каждой точки данных (для каждого наблюдения) называется *апостериорным*.
* Суммарная вероятность, используемая для нормализации ответа, является *константой нормализации*.
* Условная вероятность, т.е. вероятность каждого события, называется *правдоподобием*.

Вот как эти термины выглядят для первого примера (сравни с рис. 1).



Рис. 5. Термины теоремы Байеса

Сама теорема Байеса в новых определениях выглядит так (сравни с формулой 2):



### Пример 3. Нечестная монета

У вас есть монета, которая, как вы подозреваете, не является честной. Вы кидаете ее 100 раз. Вычислите вероятность того, что нечестная монета упадет орлом вверх с вероятностью 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%.

Обратимся к файлу Excel, лист *Пример 3*. В ячейках В13:В112 я сгенерировал случайное число от 0 до 1, и с помощью специальной вставки перенес значения в столбец С. В ячейке В8 я указал ожидаемый процент выпадений орла для этой нечестной монеты. В столбце D с помощью функции ЕСЛИ я превратил вероятности в единицы (орлы, для вероятности *р* от 0,35 до 1) или в нули (решки, для *р* от 0 до 0,35).

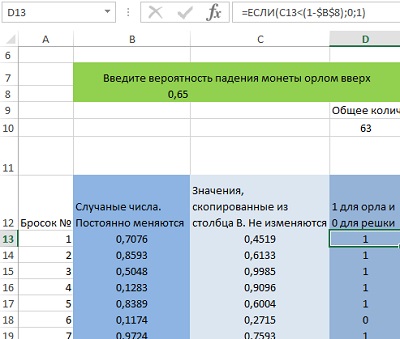


Рис. 6. Исходные данные для подбрасываний нечестной монеты

У меня получилось 63 орла и 37 решек, что хорошо соответствует генератору случайных чисел, если на входе мы установили вероятность орлов 65%.

Шаг 1. Мы хотим вычислить вероятности того, что орлы относятся к корзинам 0%, 10%, … 100%, наблюдая 63 орла и 37 решки при 100 бросках.

Шаг 2. Есть 11 начальных возможностей: вероятности 0%, 10%, … 100%. Будем наивно полагать, что все начальные возможности имеют равную вероятность, то есть 1 шанс из 11 (рис. 7). (Более реалистично мы могли бы придать начальным вероятностям, располагающимся в районе 50% большие веса, чем вероятностям на краях – 0% и 100%. Но самое замечательное заключается в том, что, поскольку у нас целых 100 подбрасываний, первоначальные вероятности не так уж важны!)

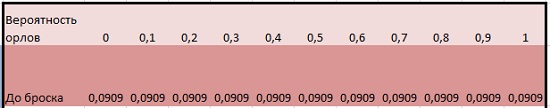


Рис. 7. Априорные вероятности

Шаг 3 и 4. Расчет правдоподобия. Чтобы рассчитать вероятность после каждого подбрасывания в Excel используется функция ЕСЛИ. В случае, если выпал орел, правдоподобие равно произведению возможности на предыдущую нормированную вероятность. Если выпала решка, правдоподобие равно (1 минус возможность) \* предыдущую нормированную вероятность (рис. 8).

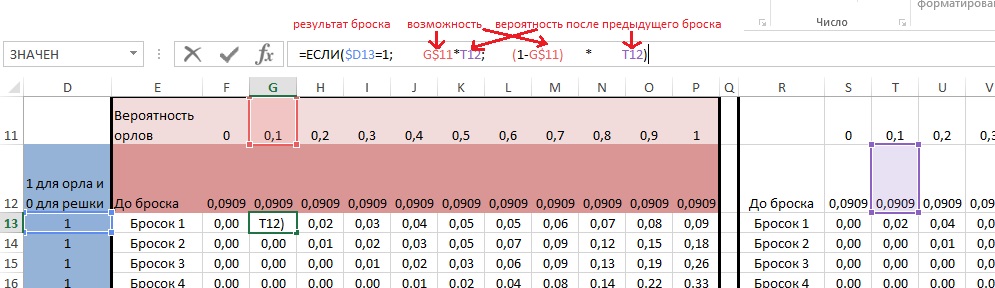


Рис. 8. Правдоподобие

Шаг 5. Нормализация выполняется, как и в предыдущем примере.

Результаты наиболее наглядно представить в виде серии гистограмм. Начальный график – это априорная вероятность. Затем каждый новый график – ситуация после очередных 25 бросков (рис. 9). Поскольку мы задали на входе вероятность орла 65%, представленные графики не вызывают удивления.

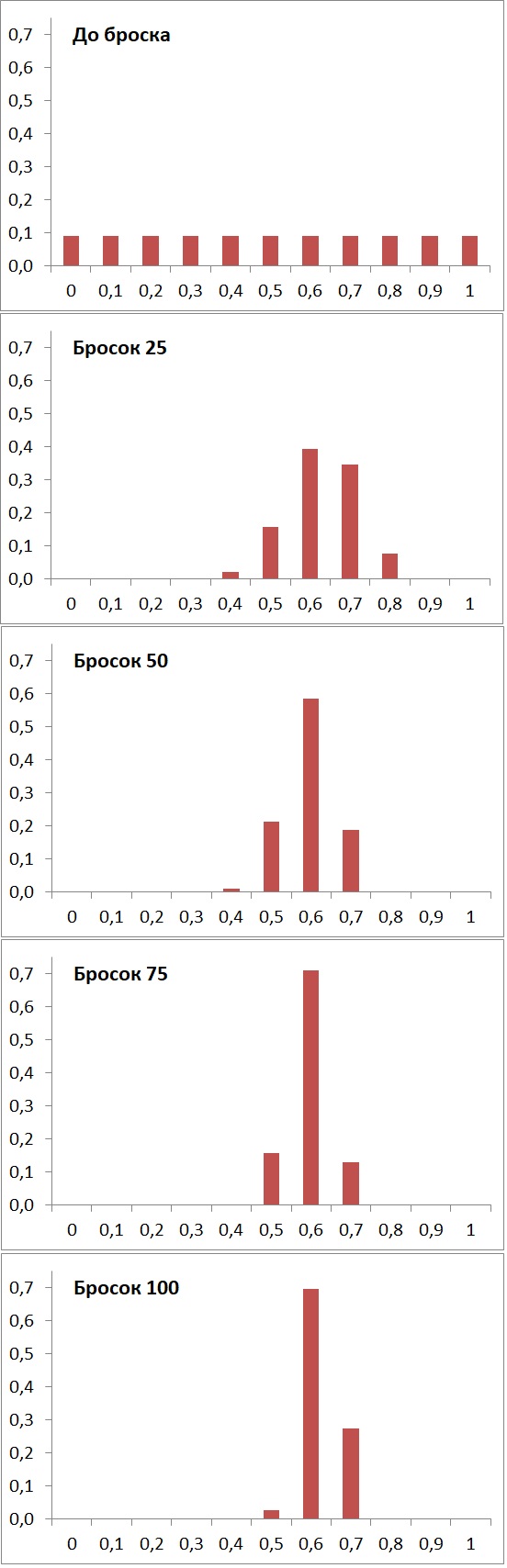


Рис. 9. Вероятности вариантов после серии бросков

Что на самом деле означает 70%-ный шанс для возможности 0,6? Это не 70%-ный шанс, что монета точно попадает на 60%. Поскольку у нас был шаг размером 10% между вариантами, мы оцениваем, что есть 70%-ный шанс, что эта монета попадет в диапазон между 55 и 65%. Решение использовать 11 начальных вариантов, с шагом 10% было полностью произвольным. Мы могли бы использовать 101 начальную возможность с шагом 1%. В этом случае мы бы получили результат с максимумом при 63% (так как у нас было 63 орла) и более плавное падение на графике.

Обратите внимание, в этом примере мы наблюдали более медленную сходимость по сравнению с Примером 2. Это связано с тем, что разница между монетой, переворачивающейся 60% против 70%, меньше, чем между кубиками с 8 и 10 гранями.

Пример 4. Еще кости. Но с ошибками в потоке данных

Вернемся к примеру 2. У друга в мешке кости с 4, 6, 8, 10, 12, 20 гранями. Он вынимает одну кость случайным образом и бросает ее 80 раз. Он записывает выпавшие числа, но в 5% случаев ошибается. В этом случае появляется случайное число от 1 и 20 вместо фактического результата броска. После 80 бросков, как вы думаете, какая кость была выбрана?

В качестве входных данных в Excel (лист *Пример 4*) я ввел количество сторон (8), а также вероятность того, что данные содержат ошибку (0,05). Формула для значения броска (рис. 10):

=ЕСЛИ (СЛЧИС() > вероятности ошибки; СЛУЧМЕЖДУ(1; число граней); СЛУЧМЕЖДУ(1;20))

Если случайное число больше вероятности ошибки (0,05), то при этом броске ошибки не было, так что генератор случайных чисел выбирает значение между 1 и «загаданным» количеством сторон кубика, в противном случае следует сгенерировать случайное целое число между 1 и 20.

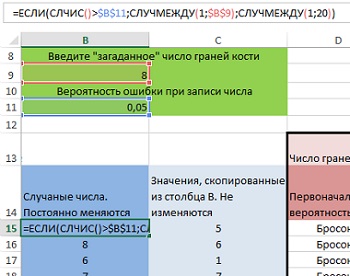


Рис. 10. Расчет значения броска

На первый взгляд, мы могли бы решить эту проблему так же, как и в примере 2. Но, если не учитывать вероятность ошибки, мы получим график вероятностей как на рис. 11. (Самый простой способ получить его в EXCEL – сначала сгенерировать броски в столбце В при значении ошибки 0,05; затем перенести значения бросков в столбец С, и наконец, поменять значение в ячейке В11 на 0; поскольку формулы расчета правдоподобия в диапазоне D14:J94 ссылаются на ячейку В11, эффект не учета ошибок будет достигнут.)

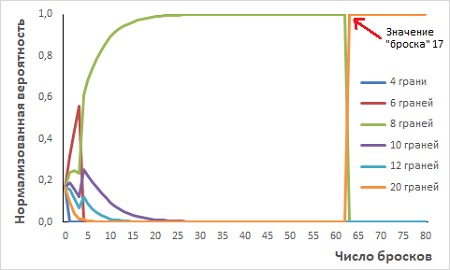


Рис. 11. Обработка значения бросков без учета вероятности присутствия ошибок

Поскольку вероятность ошибки мала, а генератор случайных чисел настроен на 8-гранник, вероятность последнего с каждым броском становится доминирующей. Более того, так как ошибка может с вероятностью 40% (восемь из двадцати) дать значение в пределах 8, то значение ошибки, повлиявшее на результат, появилось лишь на 63-ем броске. Однако, если ошибки не берутся в расчет, вероятность 8-гранника обратится в ноль, а 100% получит 20-гранник. Заметим, что к 63-му броску вероятность 20-гранника составляла всего 2\*10–25.

Шансы получить ошибку – 5%, а вероятность того, что ошибка даст значение больше 8, составляет 60%. Т.е., 3% бросков дадут ошибку со значением более 8, которая и случилась на броске 63, когда была сделана запись 17. Если формула правдоподобия не будет учитывать возможные ошибки, мы получим взлет вероятности 20-гранника с 2\*10–25 до 1, как на рис. 11.

Если человек скрупулезно наблюдает за данными, он может обнаружить эту ошибку и не принимать в расчет ошибочные значения. Для автоматизации процесса дополните уравнение правдоподобия проверкой на ошибки. Никогда не устанавливайте нулевые вероятности ошибок, если вы допускаете, что их нельзя полностью исключить. Если вы учтете вероятности ошибок, то сотни «правильных» данных не позволят отдельным ошибочным значениям испортить картину.

Дополняем уравнение функции правдоподобия проверкой на ошибки (рис. 12):

=ЕСЛИ($C15>F$13;$B$11\*1/20\*N14;($B$11\*1/20+(1-$B$11)/F$13)\*N14)

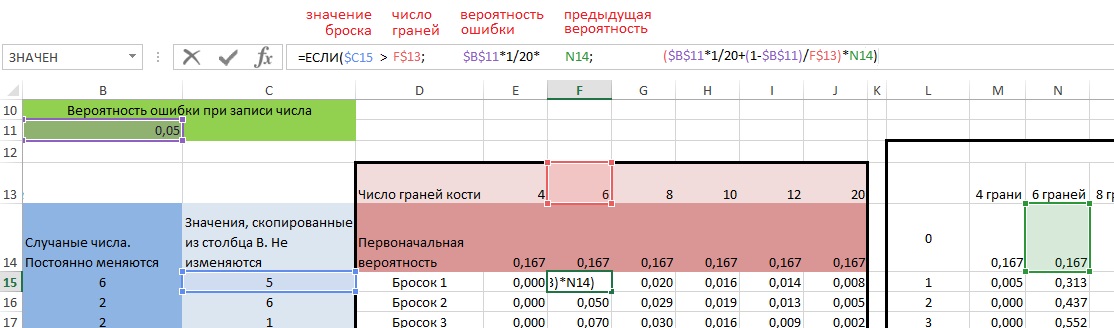


Рис. 12. Функция правдоподобия с учетом ошибок

Если записанное значение броска больше числа граней ($C15>F$13) условную вероятность не обнуляем, а уменьшаем с учетом вероятности ошибки ($B$11\*1/20\*N14). Если записанное число меньше числа граней, условную вероятность увеличиваем не в полном объеме, а также с учетом возможной ошибки ($B$11\*1/20+(1-$B$11)/F$13)\*N14). В последнем случае считаем, что записанное число могло явиться как следствием ошибки ($B$11\*1/20), так и результатом правильной записи (1-$B$11)/F$13).

Изменение нормализованной вероятности становится более устойчивым к возможным ошибкам (рис. 13).



Рис. 13. Изменение нормализованной вероятности от броска к броску

В этом примере 6-гранная кость изначально является фаворитом, потому что первые 3 броска – 5, 6, 1. Потом выпадет 7-ка и вероятность 8-гранника идет вверх. Однако, появление 7-ки не обнуляет вероятность 6-гранника, потому что 7-ка может быть ошибкой. И следующие девять бросков вроде бы подтверждают это, когда выпадают значения не более 6: вероятность 6-гранника снова начинает расти. Тем не менее, на 14-м и 15-м бросках опять выпадают 7-ки, и вероятность 6-гранной кости приближается к нулю. Позже, появляются значения 17 и 19, которые «система» определяет, как явно ошибочные.

### Пример 4A. Что делать, если у вас действительно высокая частота ошибок?

Этот пример аналогичен предыдущему, но частота ошибок увеличена с 5% до 75%. Поскольку данные стали менее релевантными, мы увеличили число бросков до 250. Применяя те же уравнения, что и в примере 4 получим следующий график:

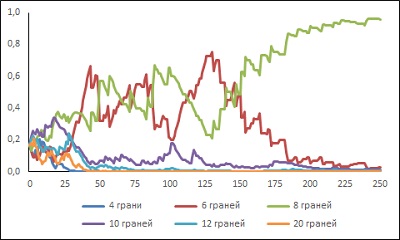


Рис. 14. Нормализованная вероятность при 75% ошибочных записей

Со столь высокой частотой ошибок потребовалось гораздо больше бросков. К тому же результат менее определен, и 6-гранник периодически становится более вероятным. Если у вас еще более высокая частота ошибок, например, 99%, все равно можно получить правильный ответ. Очевидно, чем выше частота ошибок, тем больше бросков нужно сделать. Для 75% ошибок мы получаем одно правильное значение из четырех. Если же вероятность ошибки составит 99%, мы бы получили лишь одно правильное значение из ста. Нам, вероятно, понадобится в 25 раз больше данных, чтобы выявить доминирующий вариант.

А что если вы не знаете вероятность ошибки? Рекомендую «поиграть» с примерами 4 и 4А, устанавливая в ячейке В11 различные значения от очень маленьких (например, 2\*10–25 для примера 4) до очень больших (например, 90% для примера 4А). Вот основные выводы:

* Если оценка частоты ошибок выше, чем фактическая частота ошибок, результаты будут сходиться медленнее, но все равно сходятся к правильному ответу.
* Если вы оцениваете частоту ошибок слишком низко, существует риск того, что результаты не будут правильными.
* Чем меньше фактическая частота ошибок, тем больше места для маневра у вас есть в угадывании частоты ошибок.
* Чем выше фактическая частота ошибок, тем больше данных вам нужно.

### Пример 5. Проблема немецкого танка

В этой задаче вы пытаетесь оценить, сколько танков было произведено, исходя из серийных номеров захваченных танков. Теорема Байеса была использована союзниками во время второй мировой войны, и в конечном итоге дала результаты более низкие, чем те, о которых сообщала разведка. После войны записи показали, что статистические оценки с использованием теоремы Байеса были более точными. (Любопытно, что я написал заметку по этой теме, еще не зная, что такое вероятности по Байесу; см. [Малые выборки в конкурентной разведке](http://baguzin.ru/wp/?p=2880). – *Прим. Багузина*.)

Итак, вы анализируете серийные номера, снятые с разбитых или захваченных танков. Цель – оценить, сколько танков было произведено. Вот что вы знаете о серийных номерах танков:

* Они начинаются с 1.
* Это целые числа без пропусков.
* Вы нашли следующие серийные номера: 30, 70, 140, 125.

Нас интересует ответ на вопрос: каково максимальное число танков? Я начну с 1000 танков. Но кто-то другой мог начать с 500 танков или 2000 танков, и мы можем получить разные результаты. Я собираюсь анализировать каждые 20 танков, что означает, что у меня есть 50 начальных возможностей для количества танков. Можно усложнить модель, и проанализировать для каждого отдельного числа в Excel, но ответ сильно не изменится, а анализ значительно усложнится.

Я предполагаю, что все возможности количества танков равны (т.е. вероятность наличия 50 танков, такая же, как и 500). Обратите внимание, что в файле Excel больше столбцов, чем показано на рисунке. Условная вероятность для функции правдоподобия очень похожа на условную вероятность из Примера 2:

* Если наблюдаемый серийный номер больше максимального серийного номера для этой группы, то вероятность наличия такого количества танков равна 0.
* Если наблюдаемый серийный номер меньше максимального серийного номера для этой группы, вероятность есть единица, деленная на число танков, умноженная на нормализованную вероятность на предыдущем шаге (рис. 15).

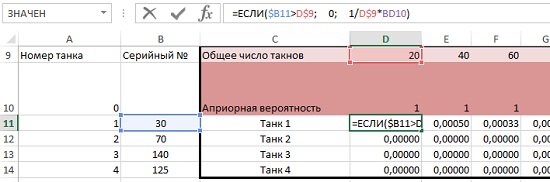


Рис. 15. Условные вероятности распределения танков по группам

Нормализованные вероятности выглядят следующим образом (рис. 16).

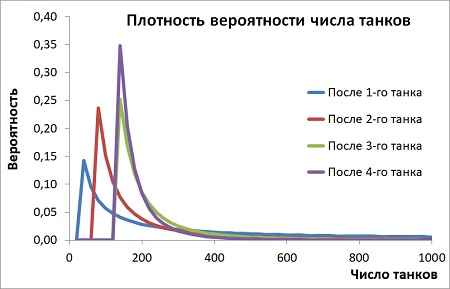


Рис. 16. Нормализованные вероятности количества танков

Наблюдается большой всплеск вероятности для максимально наблюдаемого серийного номера. После этого происходит асимптотическое снижение к нулю. Для 4 обнаруженных серийных номеров максимум отвечает 140 танкам. Но, несмотря на то, что это число является наиболее вероятным ответом, это не лучшая оценка, так как она почти наверняка недооценивает количество танков.

Если взять средневзвешенное количество танков, т.е. суммировать попарно перемноженные группы и их вероятности для четырех танков, применив формулу:

=ОКРУГЛ(СУММПРОИЗВ(BD9:DA9;BD14:DA14);0)

мы получаем наилучшую оценку равную 193.

Если бы мы первоначально исходили из 2000 танков, средневзвешенное значение было бы 195 танков, что по существу ничего не меняет.

### Пример 6. Тестирование на наркотики

Вы знаете, что 0,5% населения употребляет наркотики. У вас есть тест, который дает 99% истинных положительных результатов для употребляющих наркотик, и 98% истинных отрицательных результатов для не употребляющих. Вы случайным образом выбираете человека, проводите тест и получаете положительный результат. Какова вероятность того, что человек на самом деле употребляет наркотики?

Для нашего случайного индивидуума *первоначальная вероятность* того, что он является потребителем наркотиков, равна 0,5%, и вероятность того, что он не является потребителем наркотиков составляет 99,5%.

Следующий шаг – расчет условной вероятности:

* Если испытуемый употребляет наркотики, то тест будет положительным в 99% случаев и отрицательным в 1% случаев.
* Если испытуемый не употребляет наркотики, то тест будет положительным в 2% случаев и отрицательным в 98% случаев.

Функции правдоподобия для употребляющих и не употребляющих наркотики представлены на рис. 17.

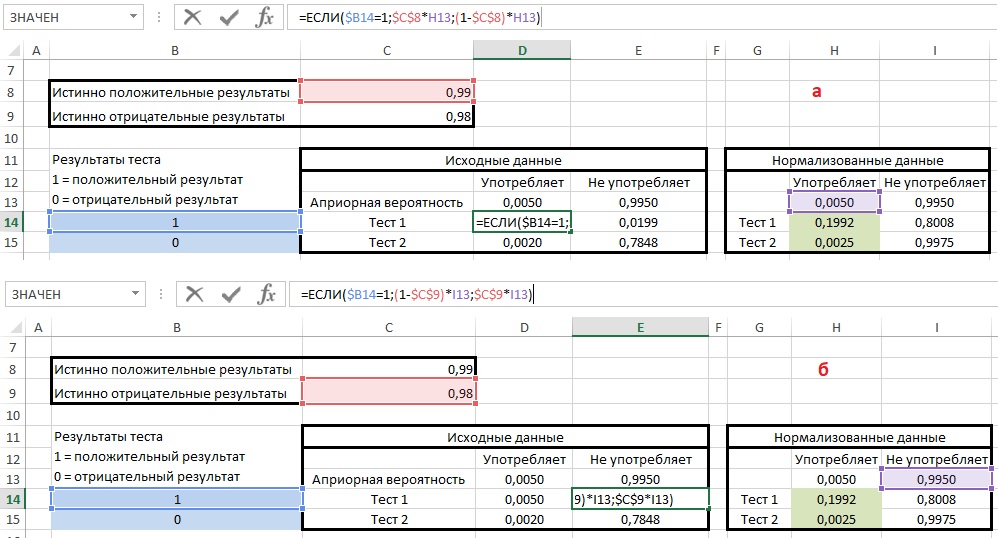


Рис. 17. Функции правдоподобия: (а) для употребляющих наркотики; (б) для не употребляющих наркотики

После нормализации, мы видим, что, несмотря на положительный результат теста, вероятность того, что этот случайный человек, употребляет наркотики, составляет всего 0,1992 или 19,9%. Этот результат удивляет многих людей, потому что в конце концов, точность теста довольно высока – целых 99%. Поскольку начальная вероятность была лишь 0,5%, даже большого увеличения этой вероятности было недостаточно, чтобы сделать отклик действительно большим.

Интуиция большинства людей не учитывает начальную вероятность. Даже если условная вероятность действительно высока, очень низкая начальная вероятность может привести к низкой конечной вероятности. Интуиция большинства людей настроена вокруг начальной вероятности 50/50. Если это так, и результат теста положителен, то нормализованная вероятность составит ожидаемые 98%, подтверждая, что человек употребляет наркотики (рис. 18).

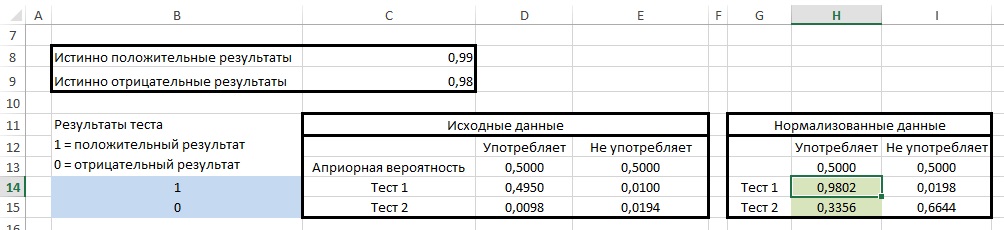


Рис. 18. Результат теста при исходной вероятности 50/50

Альтернативный подход к объяснению подобных ситуаций см. [Естественная частота против байесовского подхода](http://baguzin.ru/wp/?p=17160).

Библиографию по теореме Байеса смотри в конце заметки [Идеи Байеса для менеджеров](http://baguzin.ru/wp/?p=6355).