## Естественная частота против байесовского подхода

Начну с того, что мне очень интересен байесовский подход. Много читал, чтобы осмыслить его (см. литературу в конце заметки), и могу засвидетельствовать, что постигнуть его непросто… А недавно, прочитав книгу Герд Гигеренцер. Понимать риски. Как выбирать правильный курс, обнаружил, что байесовский подход можно дополнить более интуитивным (хотя и менее строгим) методом, который автор назвал *естественной частотой*.

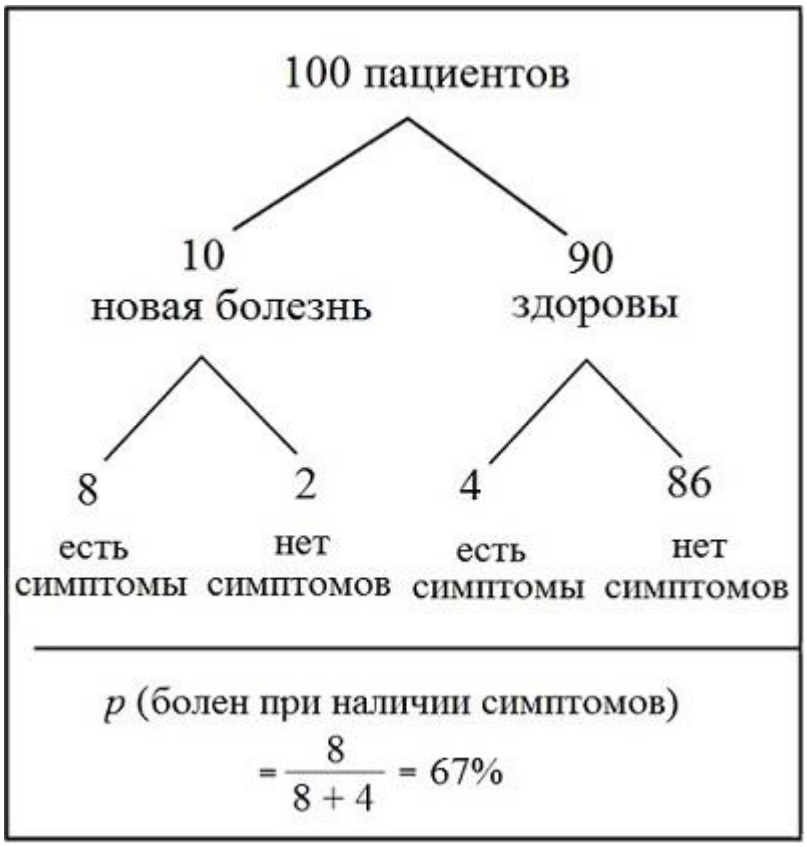


Рис. 1. Иллюстрация суждений в стиле естественной частоты

### Естественные частоты

*Естественные* (*natural frequencies*) – частоты, соответствующие способу, при помощи которого люди имели дело с информацией до изобретения книгопечатания и теории вероятностей. В отличие от вероятностей и относительных частот они являются «сырыми» наблюдениями, которые не были нормализованы относительно базовых показателей рассматриваемого события. Например, врачи наблюдали 100 пациентов, у 10 из которых была обнаружена новая болезнь (рис. 1). Из этих 10 у 8 есть симптомы болезни, но у 4 из не признанных больными также имеются эти симптомы. Разделение этих 100 случаев на 4 группы (болезнь и симптомы – 8, болезнь при отсутствии симптомов – 2, нет болезни при наличии симптомов – 4, нет болезни и нет симптомов – 86) дает значение 4 естественных частот: 8, 2, 4 и 86.

Естественные частоты облегчают получение байесовских выводов. Например, врач, который наблюдает нового пациента с симптомами болезни, легко может увидеть, что шансы на то, что данный пациент действительно болен, составляют 8/(8 + 4), то есть два к трем. Такая вероятность называется апостериорной. Однако если наблюдения врача трансформировать в условные вероятности или относительные частоты (например, посредством деления естественной частоты 4 на базовый показатель 90, что даст долю ложных положительных результатов в 4,4%), то вычисление становится более трудным. Естественные частоты позволяют увидеть апостериорные вероятности, в то время как условные частоты затуманивают людям мозги.

### Задача Монти Холла

Я уже не первый раз встречаю описание игрового телешоу «Давайте заключим сделку» (см., например, Леонард Млодинов. (Не)совершенная случайность. Как случай управляет нашей жизнью, Чарльз Уилан. Голая статистика). Телешоу впервые было показано на NBC в 1963 г. Один из

центральных эпизодов шоу назывался «Большая сделка дня». В нем ведущий, Монти Холл, показывал участнику три двери. За одной из них находился большой приз – новый «кадиллак» или другое чудо, способное вызвать вопли восторга, а за другими – совершенно бесполезные дары, например, живые козы (рис. 2). Колумнистка журнала *Parade* Мэрилин Савант, которая пять лет подряд упоминалась в книге рекордов Гиннесса как женщина с самым высоким IQ, сделала «задачу Монти Холла» весьма популярной. Она описывала проблему, с которой сталкивался участник передачи, следующим образом:

> *Предположим, вы участвуете в игровом шоу, и вам предлагают выбрать одну из трех дверей. За одной находится автомобиль, а за двумя другими – обыкновенные козы. Допустим, вы выбираете дверь номер 1. Ведущий знает, что находится за этой дверью, но открывает другую дверь, например, дверь номер 3, за которой вы видите козу, и обращается к вам: «А вы не хотите выбрать дверь номер 2?» Выгодно ли вам изменить свое первоначальное решение?*


Рис. 2. Задача Монти Холла

Вы будете менять выбранную ранее дверь на другую? Если нет, то вы поступите так, как в подобной ситуации поступает большинство людей. Ведь осталось всего две двери, поэтому шансы на выигрыш и проигрыш выглядят одинаковыми, и, если вы совершите ошибку, изменив свой изначальный выбор, и укажете на дверь, за которой стоит коза, это может вызвать горькие сожаления.

Мэрилин советовала изменить первоначальный выбор. Этот совет вызвал шквал писем, не утихавший в течение целого года. Около тысячи писем написали люди с учеными степенями, высказав свое несогласие с ней. Профессор математики Роберт Сакс из Университета Джорджа Мейсона писал: «Вы говорите ерунду! Позвольте мне объяснить. Если вам показывают, что одна из дверей не скрывает главного приза, то эта информация изменяет вероятность каждого из оставшихся вариантов выбора до 1/2, так что ни один вариант не становится более вероятным». Одна обладательница Y хромосомы доказывала: «Вы не можете применять женскую логику к вероятности. Новая ситуация предлагает всего лишь выбор одного из двух равновероятных вариантов». Еще один автор высказывался до неприличия резко: «Вы просто коза!» Наконец страсти поутихли, и почти все согласились с Мэрилин в том, что теория вероятности указывает на изменение первоначального выбора как на лучший способ действий в данной ситуации. Профессор Сакс написал ей письмо с извинениями, оказавшись одним из немногих, кто открыто признал свою ошибку.

## Решение с помощью естественных частот

Доктор Сакс, как и многие другие, запутался в вероятностях. Типичный ход размышлений выглядит так: «Вероятность, что машина находится за любой из трех дверей, равна одной трети. Одна дверь была открыта, что устраняет из рассмотрения и ее, и одну треть вероятности. Теперь, когда машина находится за одной из двух дверей, шансы на выигрыш нужно разделить поровну между этими двумя дверями, то есть они составят 50:50». Это одна из известных «когнитивных иллюзий», которая прочно засела в нашем мозгу.

Разобраться во всем этом нам поможет простой метод: метод, в основе которого заложено использование значений естественной частоты. Позвольте мне пояснить, что это такое, применительно к задаче Монти Холла. Очень важно в данном случае учитывать, что в конкурсе принимают участие сразу несколько человек, а не один. Допустим, их трое, и все они выбирают разные двери. Пусть машина находится за дверью 2 (рис. 3).


Рис. 3. Иллюстрация решения задачи Монти Холла методом естественных частот

Первый участник выбирает дверь 1. В этом случае Монти ничего не остается, кроме как открыть дверь 3 и предложить участнику изменить свой первоначальный выбор. Изменение выбранной двери на дверь 2 будет выигрышным. Допустим, второй участник выбирает дверь 3. На этот раз Монти должен открыть дверь 1, и если игрок изменит свой выбор и предпочтет дверь 2, то это позволит ему получить главный приз. Только третий участник, сразу выбравший дверь 2, проиграет, если изменит первоначальный выбор. Такой подход помогает понять, что изменить первоначально выбранный вариант чаще выгоднее, чем его сохранить. Можно точно рассчитать, как часто это выгодно: в двух случаях из трех. Вот почему Мэрилин рекомендовала изменять первоначальный выбор.

Задача Монти Холла обсуждалась на вечеринках, в учебных аудиториях и на первой странице New York Times, заставляя людей вести споры о вероятностях событий. За долгое время показа этого игрового шоу за дверями Монти Холла могли быть оставлены миллионы долларов. Здесь я лишь постарался показать, что все эти споры легко могут быть улажены при рассуждении в терминах естественных частот. Проблема находится не просто в человеческом разуме, но и в том способе, прибегая к которому используется информация.

## Решение с помощью формулы Байеса

Сравните это решение со стандартным решением в терминах вероятностей, используя формулу Байеса (если вы не знакомы с ней, рекомендую начать с заметки Формула Байеса). Рассмотрим ситуацию, когда участник сначала выбирает дверь 1, а Монти Холл открывает дверь 3 и показывает козу. Здесь мы хотим узнать вероятность $p(\text{Маш1}|\text{Мон3})$ того, что машина находится за дверью 1 после того, как Монти открыл дверь 3:

$$p\ (\text{Маш1}|\text{Мон3}) = \frac{\text{р}(\text{Маш1}) \cdot \text{р}(\text{Мон3}|\text{Маш1})}{\text{р}(\text{Маш1}) \cdot \text{р}(\text{Мон3}|\text{Маш1}) + \text{р}(\text{Маш2}) \cdot \text{р}(\text{Мон3}|\text{Маш2}) + \text{р}(\text{Маш3}) \cdot \text{р}(\text{Мон3}|\text{Маш3})}$$

$$p\ (\text{Маш1}|\text{Мон3}) = \frac{\frac{1}{3} \cdot \frac{1}{2}}{\frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot 1 + 1/3 \cdot 0} = 1/3$$

$$p\ (\text{Маш2}|\text{Мон3}) = \frac{\frac{1}{3} \cdot 1}{\frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot 1 + \frac{1}{3} \cdot 0} = 2/3$$

То есть вероятность того, что машина стоит за дверью 1, остается неизменной, а вероятность того, что машина находится за дверью 2, увеличивается до 2/3. Вероятности р(Машина1), р(Машина2) и

p(Машина3) называются априорными вероятностями, а p(Машина1|Монти3) называется апостериорной вероятностью. Условная вероятность p(Монти3|Машина1|) того, что Монти откроет дверь 3, если машина стоит за дверью 1, равняется 1/2, потому что Монти может выбирать между дверью 2 и дверью 3, и предполагается, что этот выбор происходит случайным образом. Условная вероятность p(Монти3|Машина2) того, что Монти откроет дверь 3, если машина стоит за дверью 2, равна 1, так как Монти не имеет выбора, потому что он не может открыть дверь 1. Наконец, p(Монти3|Машина3) равна нулю, потому что Монти не может показать машину участнику. Такое количество объяснений и вычислений показывает, почему люди часто оказываются сбитыми с толку, когда начинают размышлять в терминах условных вероятностей.

## Маммографическое обследование

Рассмотрим два способа предоставления информации о результатах маммографического тестирования, которые я использовал на лекциях с врачами. Первый, запутавший многих людей, использует понятие условной вероятности. Объяснения таких условных вероятностей, как чувствительность и показатель ложных положительных результатов, можно найти на рис. 4. Тестирование может иметь 4 последствия: 1) результат положительный, пациент болен; 2) результат положительный, пациент не болен; 3) результат отрицательный, пациент болен и 4) результат отрицательный, пациент не болен. Вероятности возникновения этих 4 результатов получили название: а) чувствительности; б) доли ложных положительных результатов; в) доли ложных отрицательных результатов и г) специфичности обнаружения (вероятность правильного необнаружения или доля истинных отрицательных результатов). Например, чувствительность обнаружения – вероятность получения положительного результата, если человек действительно болен, – обозначается как $p\ (положительный\ результат/болезнь)$. Такая вероятность называется условной, потому что она представляет собой не просто вероятность события A, но вероятность события A при условии наступления события B.

| Результат теста | Болезнь | |
|---|---|---|
| | Да | Нет |
| Положительный | а) чувствительность | б) доля ложных положительных результатов |
| Отрицательный | в) доля ложных отрицательных результатов | г) специфичность |

Рис. 4. Возможное распределение результатов тестирования

Чтобы понять смысл условных вероятностей, нам нужно выполнить сложные вычисления, которые многим из нас покажутся трудными (о чем свидетельствует грустное лицо в левой части рис. 5). Эта формула получила название формулы Байеса по имени английского священника Томаса Байеса (около 1702–1761), которому приписывается ее открытие.

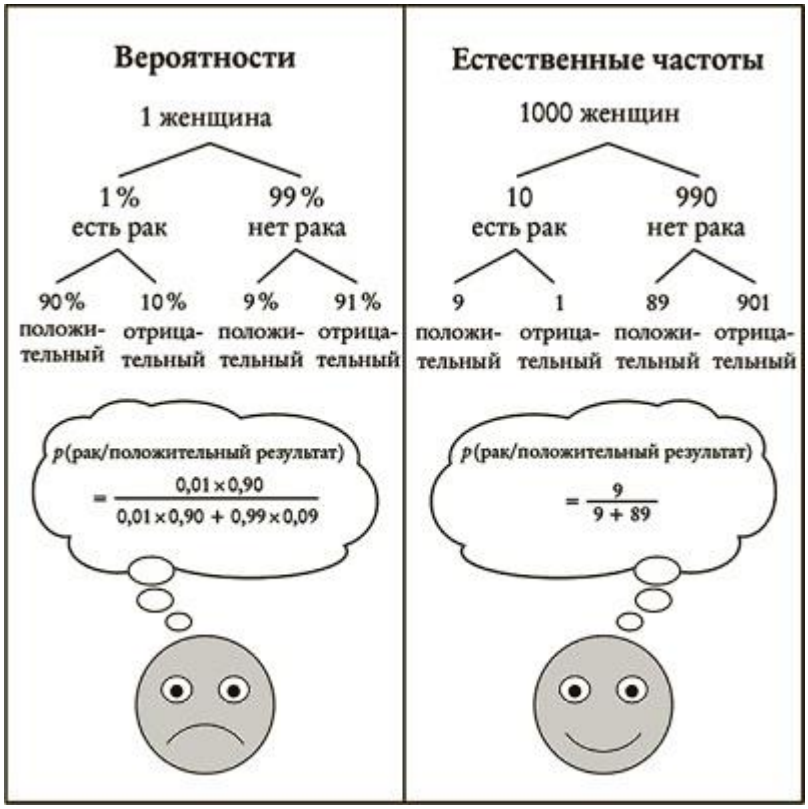


Рис. 5. Вероятность заболевания раком груди в случае положительного результата маммографического обследования

Как мы уже видели, значительно доступнее передать ту же информацию, использовав значения естественных частот. Для представления задачи в терминах естественных частот вы количество людей (в данном случае 1 тыс. женщин) разделяете на две группы – те, кто соответствуют, и те, кто не соответствуют условию (наличие рака груди), далее эти группы разбиваются на подгруппы в зависимости от новой информации (результатов тестирования). Четыре числа в нижней части левого дерева представляют собой значения условных вероятностей (см. рис. 5). В отличие от условных вероятностей (или относительных частот) общая сумма естественных частот всегда соответствует числу в верхней части дерева. Их польза состоит в том, что они выполняют за вас большую часть расчетов, облегчая понимание результатов (о чем говорит улыбающееся лицо на рис. 5).

## Формула Байеса

Формула рассчитывает вероятность гипотез с учетом новых данных. Для простого случая бинарной гипотезы (H и не H, например, есть рак или нет рака) и данных D (например, положительный результат тестирования) формула выглядит так:

$$p(H|D) = \frac{p(H) \cdot p(D|H)}{p(H) \cdot p(D|H) + p(\text{не } H) \cdot p(D|\text{не } H)}$$

где p(D|H) – апостериорная вероятность, p(H) – априорная вероятность, p(D|H) – вероятность D при условии H, p(D|не H) это вероятность D при условии не H.

Многим трудно понимать эту формулу. Но вот полезный совет. Интересно, что расчет p(D|H) становится более интуитивно понятным, когда исходные величины представляются в виде естественных частот, а не вероятностей. Для естественных частот правило выглядит так:

$$p(H|D) = \frac{a}{a+b}$$

где *a* – это число D случаев H, а *b* – это число D случаев не H.

## Литература по формуле Байеса

Условная вероятность. Теорема Байеса

Наивный байесовский классификатор документов в Excel

Идеи Байеса для менеджеров (в конце заметки есть довольно обширный дополнительный список литературы по теме)

Формула Байеса

Дж. Хей. Введение в методы байесовского статистического вывода

Моррис. Наука об управлении. Байесовский подход