ПРИКЛАДНАЯ СТАТИСТИКА

О карасях, рыбалке и бабушках[*]

When the facts change, I change my mind. What do you do, sir?

John Maynard Keynes

Рассмотрим простой пример. Пусть в озере живут караси и щуки. Петя, живущий в деревне по соседству, выловил в нём карася, щуку и ещё одного карася, а после серьёзно задумался о том с какой вероятностью, p, он таскает карасей из озера. Петя предполагает, что в озере настолько много рыбы, что вылов одного карася несильно меняет вероятность поймать нового карася, т.е. наблюдения $y_1 = 1, y_2 = 0, y_3 = 1$ независимы и одинаково распределены.

О том какие у бабушек бывают распределения

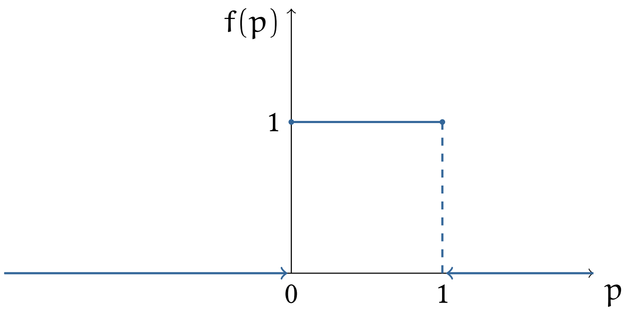
Если бы Петя был частотным статистиком, то он бы воспользовался методом максимального правдоподобия или методом наименьших квадратов и нашёл бы оценку требуемой вероятности. Тем не менее, в родной деревне Пети широко практикуется байесовское воспитание, в связи с чем ему не хотелось бы пользоваться стандартными методами.

Идея! Мы ничего не знаем о параметре p. Давайте опишем наше незнание с помощью какой-то **априорной функции распределения**. Важно помнить, что на данные при этом смотреть нельзя. Наши априорные ожидания никак не должны быть с ними связаны. В случае Пети, он сначала должен задать распределение p, а уже после идти таскать рыб.

[*]Эта pdf-ка, по факту, представляет из себя кусочек недописанной виньетки по Байесовским методам: https://github.com/FUlyankin/book_about_bayes

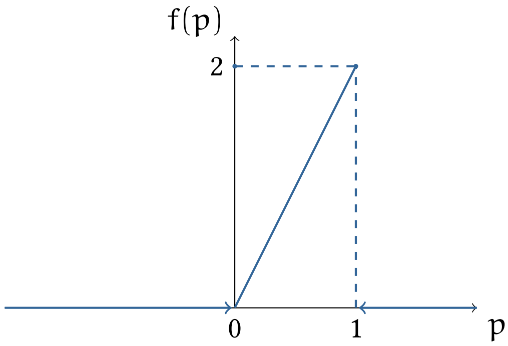
Например, если мы вообще ничего не знаем о том, что происходит в пруду, то логично взять в качестве априорного распределения равномерное, $p \sim \mathcal{U}[0;1]$. Тем самым мы не только скажем, что настолько ничего не знаем о параметре p, что допускаем абсолютно любое значение этого параметра, но и одновременно с этим откинем все невозможные значения, ограничив p отрезком от 0 до 1.

$$f(p) = \begin{cases} 1 & , p \in [0;1] \\ 0 & , \text{иначе} \end{cases}$$



В то же самое время, если у нас есть любящая порыбачить (а заодно и внука) бабушка, которая говорит, что за свою жизнь выловила из озера карасей в несколько раз больше, чем щук, то вполне логично поверить ей и предположить, что у параметра p будет распределение с плотностью

$$f(p) = \begin{cases} 2p & , p \in [0;1] \\ 0 & , \text{иначе.} \end{cases}$$



Тогда в своих априорных предположениях мы учтём многолетний опыт бабушки, а вместе с ним большое число случайных выборок из местного прудика, которые мы не видели. Если бабушка не врёт, и в пруду ничего с тех пор не поменялось, дополнительная информация поможет нам получить более точные оценки. Однако, если бабушка Пети никогда не ловила рыбу (или это вообще не его бабушка, хотя она и утверждает обратное), то принимать её априорное мнение о рыбе на веру ни в коем случае нельзя. Вы должны верить в априорное распределение и должны быть готовы сделать на него денежную ставку.

Давайте посмотрим, что у нас получится при разных априорных мнениях. Пусть $p \sim \mathcal{U}[0;1]$. Найдём апостериорную плотность распределения параметра p. Воспользуемся формулой Байеса:

$$f(p \mid y_1, y_2, y_3) = \frac{f(p, y_1, y_2, y_3)}{f(y_1, y_2, y_3)} = \frac{f(y_1, y_2, y_3 \mid p) \cdot f(p)}{f(y_1, y_2, y_3)}.$$

В знаменателе полученной дроби стоит значение совместной плотности распределения трёх случайных величин в точке y_1, y_2, y_3. Это какая-то константа. Пренебрежём ей для лёгкости расчётов. Чуть позже мы восстановим её назад. С помощью значка $\propto$ будем записывать равенство с точностью до константы

$$\frac{f(y_1, y_2, y_3 \mid p) \cdot f(p)}{f(y_1, y_2, y_3)} \propto f(y_1, y_2, y_3 \mid p) \cdot f(p).$$

Вспоминаем о том, что собранные нами наблюдения независимы и получаем

$$\begin{aligned} f(y_1, y_2, y_3 \mid p) \cdot f(p) &= f(y_1 \mid p) \cdot f(y_2 \mid p) \cdot f(y_3 \mid p) \cdot f(p) = \\ &= \mathbb{P}(y_1 = 1 \mid p) \cdot \mathbb{P}(y_2 = 0 \mid p) \cdot \mathbb{P}(y_3 = 1 \mid p) \cdot f(p) = p \cdot (1-p) \cdot p \cdot 1. \end{aligned}$$

Выходит, что апостериорная плотность распределения параметра p должна иметь вид

$$f(p \mid y_1, y_2, y_3) = \text{const} \cdot p^2 \cdot (1-p).$$

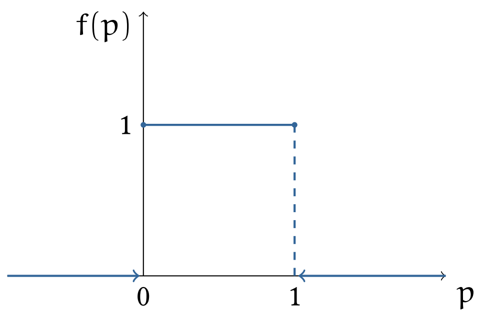
Осталось восстановить нормировочную константу. Вспоминаем, что интеграл по области определения апостериорной плотности распределения должен быть равен единице

$$\text{const} \cdot \int_0^1 p^2 \cdot (1-p)\, dp = 1 \quad \Rightarrow \quad \text{const} = 12$$

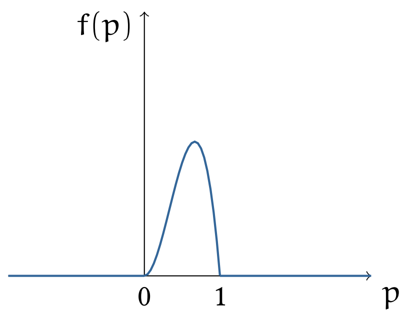
Итак, ваши авации! Апостериорное распределение параметра p:

$$f(p \mid y_1, y_2, y_3) = \begin{cases} 12 \cdot p^2 \cdot (1-p) & p \in [0;1] \\ 0 & \text{иначе} \end{cases}$$

Априорное распределение:



Апостериорное распределение:



В априорном мнении Петя не знал где находится p и все точки для него были одинаково предпочтительны. Апостериорное мнение говорит, что вероятность поймать карася гораздо ближе к единице, чем к нулю. Проделаем те же самые рассуждения, но уже учитывая априорное мнение бабушки.

По аналогии получаем

$$f(p \mid y_1, y_2, y_3) = \mathrm{const} \cdot p^2 \cdot (1-p) \cdot 2p = \mathrm{const} \cdot p^3 \cdot (1-p).$$

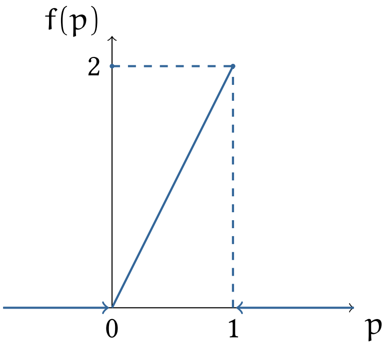
Восстанавливаем константу:

$$\mathrm{const} \cdot \int_0^1 p^3 \cdot (1-p)\, dp = 1 \quad \Rightarrow \quad \mathrm{const} = 20.$$

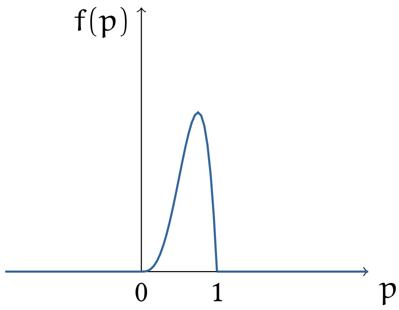
Снова получаем апостерирную функцию плотности

$$f(p \mid y_1, y_2, y_3) = \begin{cases} 20 \cdot p^3 \cdot (1-p) & p \in [0; 1] \\ 0 & \text{иначе.} \end{cases}$$

Априорное распределение:



Апостериорное распределение:



Если учесть и мнение бабушки и нашу выборку, то получится, что шансы того, что карасей в озере мало, минимальны.

Сравним между собой априорную вероятность $\mathbb{P}(p > 0.5)$ и апостериорную вероятность того, что $\mathbb{P}(p > 0.5 \mid y_1, \ldots, y_3)$, а также априорное и апостериорное математические ожидания, $\mathbb{E}(p)$ и $\mathbb{E}(p \mid y_1, \ldots, y_3)$.

Равномерное распределение:	Распределение бабушки:
$\mathbb{P}(p > 0.5) = \int_{0.5}^1 1\, dp = 0.5$	$\mathbb{P}(p > 0.5) = \int_{0.5}^1 2p\, dp = 0.75$
$\mathbb{P}(p > 0.5 \mid y) \approx 0.68$	$\mathbb{P}(p > 0.5 \mid y) = 0.81$
$\mathbb{E}(p) = \int_0^1 p \cdot 1\, dp = 0.5$	$\mathbb{E}(p) = \int_0^1 p \cdot 2p\, dp \approx 0.66$
$\mathbb{E}(p \mid y) = \int_0^1 12 \cdot p^3 \cdot (1-p)\, dp = 0.6$	$\mathbb{E}(p \mid y) = \int_0^1 20 \cdot p^4 \cdot (1-p)\, dp \approx 0.66$

Видим, что в первой ситуации вероятность того, что карасей больше чем щук, при учёте наблюдений увеличивается. Ровно как и доля карасей. Во втором случае, грубо говоря, мои наблюдения подтверждают мнение бабушки и математическое ожидание не изменяется. По той же причине вероятность того, что карасей больше чем щук увеличивается ещё сильнее.

Кстати говоря, иногда возникают ситуации, в которых апостериорный результат не зависит от того, во что мы верим. Это говорит о том, что у нас очень много данных и взаимосвязь в них прослеживается достаточно чётко.

Ещё раз, ещё раз:

- априорное распределение выбирается до сбора данных;
- с помощью априорного распределения мы пытаемся описать своё незнание;
- оно отбрасывает заведомо неверные значения параметра;
- вы должны быть готовы сделать денежную ставку на выбранное вами априорное распределение;
- на выходе мы получаем целое апостериорное распределение, с помощью которого можем отвечать на разные вопросы.

О точечных оценках

Мы получаем на выходе гораздо больше, чем просто точечную оценку. В конечном итоге вся информация о параметре p содержится в его апостериорном распределении, с помощью которого можно отвечать на любые вопросы, касающиеся этого параметра.

Тем не менее, если от нас требуют точечную оценку, в качестве неё мы могли бы использовать, математическое ожидание, моду или медиану. Конкретный выбор зависит от того как именно нас накажу за то, если мы ошибёмся. Выбор точечной β_F зависит от выбранной функции потерь.

Так, например, если мы угадали параметр, то в награду получаем половину царства и принцессу, а если не угадали, то нам отрубают голову, выгоднее всего для нас назвать самое вероятное значение параметра, то есть моду апостериорного распределения.

Если функция потерь квадратичная, $(\beta_F - \beta)^2$, у нас отнимают площадь царства (и, возможно, площадь принцессы) пропорциональную квадрату отклонения спрогнозированного нами значения от настоящего, то выгоднее всего назвать в качестве оценки математическое ожидание апостериорного распределения.

Если функция потерь абсолютная, $|\beta_F - \beta|$, то в качестве оценки выгодна медиана. Выбор функции потерь, в свою очередь, зависит от поставленной перед нами задачи.

Сконцентрируемся. Закроем глаза и попытаемся отыскать в чертогах разума определения медианы и моды. Медиана — это квантиль уровня 0.5. Иными словами это такое значение случайной величины, что

$$\mathbb{P}(p < Med) = \mathbb{P}(p > Med) = 0.5.$$

Найдём её!

$$\mathbb{P}(p > \mathrm{Med}) = 0.5 \quad \Rightarrow \quad \int_0^{\mathrm{Med}} 12 \cdot p^2 \cdot (1-p)\, dp = 0.5$$

Взятие этого интеграла приведёт нас к уравнению четвёртой степени. Нам подойдёт решение $\mathrm{Med}(p) \approx 0.61$. Скорее всего, слова «уравнение четвёртой степени» оставили у впечатлительного читателя не очень хороший осадок. Компьютеры умеют избегать таких сложностей.

Модой непрерывной случайной величины называется такое её значение, при котором плотность распределения достигает локального максимума. Вполне логично, что $\mathrm{Mod}(p) = \frac{2}{3}$:

$$(12 \cdot p^2 \cdot (1-p))' = p \cdot (2 - 3 \cdot p) = 0 \quad \Rightarrow \quad p = \frac{2}{3} \vee p = 0.$$

Таким образом мы получили целых три точечные оценки: 0.6, 0.61 и 0.66. Как это не странно, они расположены довольно близко друг к другу. По мере увеличения количества наблюдений, пик апостериорного распределения будет становится всё острее, а точечные оценки будут становиться всё ближе.

Стоит отметить, что иногда в качестве точечной оценки выбирают какой-то квантиль апостериорного распределения. Когда боятся завысить прогноз, берут квантиль меньше медианы. Например, можно взять 30% квантиль. Когда боятся занизить прогноз, берут кввантиль выше медианы. Например, можно взять 70% квантиль. В такой ситуации мы имеем дело с квантильной функцией потерь, которая по-разному штрафует перепрогноз и недопрогноз.

О доверительных и байесовских интервалах

В частотном подходе мы часто делали интервальные оценки, строили доверительные (confidence) интервалы. В байесовском подходе также можно делать интервальные оценки, а именно, строить **байесовские (bayesian или credible) интервалы.**

Между доверительным и байесовским интервалом есть тонкая разница. Если мы построили 95% доверительный интервал, то говорить, что истинное значение параметра p попадает в этот интервал с вероятностью 0.95 неправильно. В случае доверительного интервала случайными величинами являются его границы.

Правильно сказать, что интервал накрывает истинное значение параметра с вероятностью 95%, и он может как содержать его, так и не содержать, но метод построения обеспечивает вероятность накрытия в 95%. Это связано с тем, что мы работаем при построении интервала не с истинным значением параметра p, а с его оценкой $\hat{p}$. Для байесовского интервала, действительно, вероятность попадания параметра p в него равна 0.95. В случае байесовского подхода мы строим предиктивный интервал.

Обычно, нам хотелось получить самые короткие интервалы. Почему самые короткие? Если Петя говорит, что с вероятностью 0.95 температура завтра будет лежать в интервале от 2 до 5 градусов, а Вася говорит, что от 3 до 10 градусов, ошибаться они будут одинаково, в 5% случаев, однако точность прогноза будет выше у Пети. Самый короткий байесовский интервал называется **HPD (highest probability density interval)**. Конечно же, можно строить интервалы для любых вероятностей, а не только для 0.95.

Для нашего случая, чтобы найти HPD, необходимо решить следующую задачу:

$$\begin{cases} b - a \longrightarrow \min\limits_{a,b} \\ \int_a^b 12 \cdot p^2 \cdot (1-p)\, dp = 0.95. \end{cases}$$

Можно взять интеграл, получить ограничение $4b^3 - 3b^4 - 4a^3 + 3a^4 = 0.95$, не забыть, что $0 \leqslant a, b \leqslant 1$, выписать лагранджиан и получить, что $a \approx 0.23$, $b \approx 0.96$. При этом, значение плотности апостериорного распределения в точке a совпадёт для нашего случая с её значением в точке b. Если у непрерывной случайной величины одна мода, тогда для самого короткого интервала $f(a) = f(b)$. Можно воспользоваться жтим и найти доверительный интервал для нашей задачи

$$\begin{cases} b - a \longrightarrow \min\limits_{a,b} \\ f(a) = f(b). \end{cases}$$

Те читатели, которые не выпали из повествования после слов «уравнение четвёртой степени», сейчас, скорее всего, тоже потеряли интерес к чтению. Однако спешу обрадовать, на практике все вычислительные сложности на себя берёт компьютер, и здесь все эти примеры находятся лишь для того, чтобы показать, где именно возникают вычислительные сложности.

О прогнозах

Когда мы строим какую-то модель, мы хотим на выходе получить прогноз. В данном случае нам было бы безумно интересно получить ответ на вопрос, какая рыба будет выловлена в озере следующей. Логично, что если у нас есть апостериорное распределение параметра p, то прогнозом будет какое-то распределение для нового значения y. Наш прогноз не будет точечным. Дело осталось за малым, преобразовать $f(p \mid y)$ в $\mathbb{P}(y_4 = \text{карась} \mid y)$. Сделаем это несколькими способами.

Способ первый, безынтегральный: мы знаем, что повторное математическое ожидание убирает условие, то есть

$$\mathbb{E}(Z) = \mathbb{E}(\mathbb{E}(Z \mid W)).$$

Если случайная величина Z принимает значения 0 и 1, тогда

$$\mathbb{P}(Z = 1) = \mathbb{E}(Z) = \mathbb{E}(\mathbb{P}(Z = 1 \mid W)).$$

Более того, если есть какое-то дополнительное условие A, тогда выполнится

$$\mathbb{P}(Z = 1 \mid A) = \mathbb{E}(\mathbb{P}(Z = 1 \mid W, A) \mid A).$$

Чтобы осознать это, будем индексом под математическим ожиданием указывать относительно какого распределения мы ищем это математическое ожидание. В первой ситуации мы искали математическое ожидание относительно $\mathbb{P}$, значит

$$\mathbb{E}_{\mathbb{P}}(Z) = \mathbb{E}_{\mathbb{P}}(\mathbb{E}_{\mathbb{P}}(Z|W)).$$

Если мы рассмотрим $\mathbb{P}(Z = 1 \mid A)$, то мы, сказав что наступило событие A, наложим на изначальное пространство элементарных исходов какое-то ограничение и перейдём к новой вероятностной мере $\mathbb{P}_A$, для которой также выполняется

$$\mathbb{E}_{\mathbb{P}_A}(Z) = \mathbb{E}_{\mathbb{P}_A}(\mathbb{E}_{\mathbb{P}_A}(Z \mid W))$$

.

Но что такое $\mathbb{P}_A$? Это ничто иное, как условная вероятность некоторого события, $\mathbb{P}(\ \ldots \mid A)$. Делаем везде замену и получаем, что

$$\mathbb{E}_{\mathbb{P}}(Z \mid A) = \mathbb{E}_{\mathbb{P}}(\mathbb{E}_{\mathbb{P}}(Z \mid W, A) \mid A).$$

Вернёмся к задаче и применим к ней этот факт:

$$\begin{aligned}
\mathbb{P}(y_4 = \text{карась} \mid y_1, y_2, y_3) = \\
&= \mathbb{E}(\mathbb{P}(y_4 = \text{карась} | p, y_1, y_2, y_3) \mid y_1, y_2, y_3) = \\
&\qquad = \mathbb{E}(p \mid y_1, y_2, y_3) = 0.6.
\end{aligned}$$

Такой хитрый способ найти прогноз не является универсальным. Поэтому посмотрим на интегралы, которые помогают сделать это в общем случае.

Способ второй, хитро-интегральный: распишем искомую вероятность по формуле условной вероятности.

$$\mathbb{P}(y_4 = \text{карась} \mid y_1, y_2, y_3) =$$
$$= \frac{\mathbb{P}(y_1 = \text{карась}, y_2 = \text{щука}, y_3 = \text{карась}, y_4 = \text{карась})}{\mathbb{P}(y_1 = \text{карась}, y_2 = \text{щука}, y_3 = \text{карась})} =^*$$

Найти ни верхнюю вероятность ни нижнюю в силу того, что y_1, y_2, y_3, y_4 и p являются случайными величинами, мы не можем. Более того, эти случайные величины зависимы. Случайная величина p влияет на реализацию каждой из этих трёх случайных величин.

Заметим, что $y_1|p$, $y_2|p$, $y_3|p$, $y_4|p$ независимые случайные величины, а $\mathbb{P}(y_1 = \text{карась}) = \mathbb{E}(y_1 = \text{карась} \mid p)$. Воспользуемся этим:

$$^* = \frac{\mathbb{E}(\mathbb{P}(y_1 = 1, y_2 = 0, y_3 = 1, y_4 = 1 \mid p))}{\mathbb{E}(\mathbb{P}(y_1 = 1, y_2 = 0, y_3 = 1 \mid p))} = \frac{\mathbb{E}(p^3(1-p))}{\mathbb{E}(p^2(1-p))} =$$
$$= \frac{\mathbb{E}(p^3) - \mathbb{E}(p^4)}{\mathbb{E}(p^2) - \mathbb{E}(p^3)} = \frac{\frac{1}{4} - \frac{1}{5}}{\frac{1}{3} - \frac{1}{4}} = \frac{12}{20} = 0.6.$$

Конечно же для поиска всех математических ожиданий вида $\mathbb{E}(p^k)$ пришлось брать интегралы.

Способ третий, интегрально-влобовый: поговорим о прогнозировании чуть более подробно, в общем, непрерывном случае. Из совместной плотности распределения $f(x, y)$ можно получить частную плотность $f(x)$, выинтригрировав совместную плотность по переменной y, а именно

$$f(x) = \int f(x, y)\, dy = \int f(x \mid y) f(x)\, dy.$$

Эта формула является аналогом формулы для поиска полной вероятности. Мы перебираем континуальное количество гипотез для переменной Y и находим плотность для X. Будем рассуждать для общего случая. Пусть у нас есть объясняемая переменная y и объясняющая x. Что мы сделали? Мы сделали байесовский вывод и получили апостериорную плотность для параметра,

$$f(\beta \mid x, y) \propto f(y \mid x, \beta) \cdot f(\beta \mid x).$$

Теперь мы хотим перейти от известной нам апостериорной плотности для параметра β к плотности для нового значения y_{new}, $f(y_{new} \mid x_{new}, x, y)$. Выинтегрируем из уже известных нам плотностей лишние части и получим требуемое

$$f(y_{new} \mid x_{new}, x, y) = \int f(y_{new}, \beta \mid x_{new}, x, y) \, d\beta =$$
$$= \int f(y_{new} \mid x_{new}, x, y, \beta) f(\beta \mid x, y) \, d\beta.$$

Под знаком интеграла находится произвведение нашей модели и апостериорной плотности распределения. Их обе мы знаем. Для случая карасей и щук получаем

$$f(y \mid p) = \int f(y \mid p) f(p \mid y) \, dp = \int p \cdot f(p \mid y) \, dp = \mathbb{E}(p \mid y) = 0.6$$

Таким образом, получаем требуемое распределение. Вероятность того, что выловлен карась, равна 0.6. Делая всё это, мы снова сталкиваемся с вычислительными сложностями. Нужно брать интегралы. Повсюду куча интегралов. Решая упражнение с распределением Бернулли и тремя наблюдениями, мы уже накопили кучу вычислительных проблем. Позже мы немного поговорим о том, как компьютер побеждает эти пробелмы с помощью сэмплирования.