Глава 1

Что такое Баесовский подход и кому он сдался

> When the facts change, I change my mind. What do you do, sir?
>
> John Maynard Keynes

В предыдущей главе мы попытались на простых примерах вникнуть в байесовскую философию и понять что же в ней такого особенного. Никакой глубокой математики мы при этом не использовали. Пора исправить это небольшое упущение и вникнуть в проблему на новом уровне.

1.1 Новый шаг вглубь байесовской норы

До того, как мы открыли эту книгу, нам часто приходилось оценивать различные параметры. Для этого в нашем арсенале находилось довольно большое количество методов. Мы использовали в повседневной жизни метод максимального правдоподобия, метод моментов, самые разные методы наименьших квадратов[1] и даже различные более сложные методы вроде обобщённого

[1]Было бы интересно узнать как много МНК знает читатель. Одному из авторов на ум прямо сейчас приходит около 10 разных методов.

метода моментов (GMM). Все эти методы относятся к частотной статистике.

Используя их, мы предполагали, что:

1. у нас есть какой-то детерминированный неизвестный параметр β, который нужно оценить;
2. у нас есть данные $y_1, \dots, y_n$, связанные с этим параметром;
3. у нас есть модель, которая описывает как наблюдение y_i зависит от параметра β;
4. у нас есть методы, которые позволяют получить оценку $\hat{\beta}(y_1, \dots, y_n)$;
5. полученная нами оценка является случайной величиной, так как она является функцией от наблюдений, то есть **статистикой**;
6. у нас есть ряд теорем (ЦПТ, дельта-метод и другие), которые говорят нам о том как распределена случайная величина $\hat{\beta}$, как для неё можно оценить дисперсию, $\widehat{\mathrm{Var}}(\hat{\beta})$, и построить доверительный интервал;
7. на основе этих теорем мы можем проверять свои гипотезы, связанные с параметром β и получать крутые (или не очень) результаты.

Грубо говоря, каждый раз, исследуя различные явления, мы шли по такой, отработанной несколькими поколениями статистиков методологии и получали некоторый результат.

При всём этом нам безумно сильно, в силу человеческой любознательной природы, хотелось бы получить ответ, например, на вопрос: «Какова вероятность того, что β больше трёх, $P(\beta > 3 \mid y_1, \dots, y_n)$?», но в силу того что β является неизвестной константой, получить ответ на этот довольно естественный вопрос мы не могли.

Для того, чтобы выкрутиться из сложившейся ситуации, мы формулировали гипотезу $H_0 : \beta = 3$ против альтернативной гипотезы $H_1 : \beta > 3$.

За кадром мы формулировали несколько теорем, порождающих критерий для проверки этой гипотезы и некоторые условия, при которых его можно было бы использовать. Статистика для проверки гипотезы, при её верности имела

какое-то распределение. Меткое попадание наблюдаемого значения в критическую область говорило нам о том, что β значимо отличается от трёх, но при этом оно ничего не могло рассказать о существенности этого различия. Понятия статистической значимости и смысловой существенности в этой неестественной, дробящей студенческий мозг при изучении, процедуре смешались в единое целое.

У многих опытных читателей, скорее всего, в голове промелькнула мысль: «Как так? Я же могу оценить вероятность $\hat{P}(\hat{\beta} > 3)$!». Да, можете, но это оценка вероятности для оценки. А нам хотелось бы работать с истинным значением β.

Кроме того, в частотном подходе мы ставим перед собой только одну гипотезу и производим её проверку. Все остальные существующие гипотезы мы отбрасываем. Обычно это оправдывается методом максимального правдоподобия, который оценивает параметр, стараясь максимизировать вероятность появления выборки. Мы не отвергаем гипотезу H_0, если значение правдоподобия при замене параметра $\hat{\beta}_{MLE}$ на тройку изменяется незначительно.

Байесовцы, в свою очередь, говорят, что никогда точно неизвестно, какая из гипотез верна, и поэтому нельзя просто выбирать одну гипотезу из огромной кучи и тестировать её. Нужно вычислить апостериорную вероятность каждой возможной гипотезы и при прогнозировании учесть всё. Сумма вероятностей всех возможных гипотез равна единице, поэтому, если какая-то гипотеза становится вероятнее, вероятность других уменьшается. Байесовский подход вынуждает таскать за собой множество гипотез вместо одной. И для человека это довольно тяжело. Это было одной из причин того, почему байесовский подход не был раньше популярен. Сегодня благодаря мощным компьютерам таскать за собой целые распределения стало менее проблематично и интерес к байесовскому подходу начал расти.

Внимательный читатель, падкий на красивые идеи, уже усвоил, что философия байесовского подхода позволяет взглянуть на мир под другим углом. Давайте не будем разделять между собой $\hat{\beta}$ и β, а всё наше незнание о параметре β мы будем формулировать в виде **априорного закона** распределения. Параметр β будет случайной величиной, а не константой.

После того, как мы сформировали наше мнение, соберём кучу наблюдений и с помощью формулы Байеса пересчитаем априорное распределение параметра β в **апостериорное**

$$f(\beta \mid y) = \frac{f(y \mid \beta) \cdot f(\beta)}{f(y)}.$$

На выходе мы получаем целое распределение $f(\beta \mid y)$. Это гораздо больше, чем точечная оценка параметра. Используя его можно получить ответы на любые, интересующие нас вопросы, в том числе измерить вероятность того, что β принимает значение больше трёх. Эта вероятность и будет отражать существенность эффекта.

Важно отметить, что байесовский подход — это другой подход к оцениванию параметров, а не другая модель. В его рамках можно изучать любые классические модели. Изменения касаются способа моделирования неизвестных параметров.

Изобразим всё вышесказанное в виде красивой таблички под номером 1.1 и расположим её на странице 5. Надеемся, что это удовлетворит страсть нашего читателя к структурированию новой информации.

Ещё раз, ещё раз! Мы не знаем параметр β. Степень своего незнания мы представляем в виде априорного закона распределения. После мы собираем выборку и с помощью обычной формулы условной вероятности делаем пересчёт. Вот и вся любовь. Попробуем заняться этой, новой для нас любовью, на каком-нибудь простом примере.

1.2 О карасях, рыбалке и бабушках

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

Частотный подход раньше его называли классическим	Байесовский подход скоро его будут называть классическим
• β — неизвестный параметр, константа • Данные: $y_1, \dots, y_n$ • Модель: описывает связь y_i с β • Методы: ML, OLS, GM и т.п. $\Rightarrow \hat{\beta}(y_1, \dots, y_n)$ • Теоремы: ЦПТ, дельта-метод и т.п. $\Rightarrow \widehat{\mathrm{Var}}(\hat{\beta}), \quad \beta \in [\hat{\beta}_L; \hat{\beta}_H]$ • Прогнозы и гипотезы	• Нет разделения между β и $\hat{\beta}$ • Априорное распределение: $f(\beta)$ выражает моё незнание о β • Данные: $y_1, \dots, y_n$ • Модель: описывает связь y_i с β • Апостериорное распределение: $f(\beta \mid y_1 \dots y_n)$ получаем байесовским пересчётом • В качестве точечной оценки для β можно взять апостериорную медиану, апостериорное математическое ожидание, апостериорную моду или любой другой апостериорный квантиль • Можно построить байесовский интервал (credible or baessian interval), $P(\beta \in [\beta_L; \beta_H]) = 0.95$ • Можно получить распределение $\hat{y}_{n+1}$

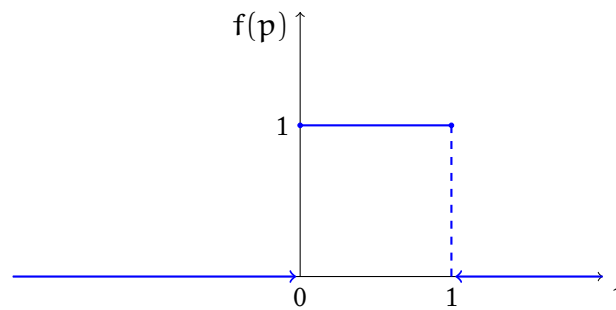
Таблица 1.1. Красивая таблица, которая призвана удовлетворить страсть нашего вдумчивого читателя к структурированию новой информации.

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.

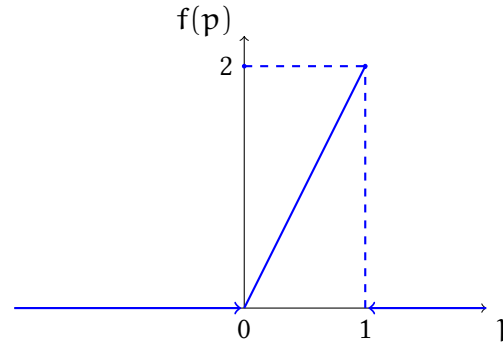
Например, если мы вообще ничего не знаем о том, что происходит в пруду, то логично взять в качестве априорного распределения равномерное, $p \sim \mathcal{U}[0;1]$. Тем самым мы не только скажем, что настолько ничего не знаем о параметре p, что допускаем абсолютно любое значение этого параметра, но и 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 два раза 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. Как уже отмечалось выше, вы должны быть готовы сделать на своё априорное мнение 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 (что довольно наивно)[2] и получаем

$$f(y_1, y_2, y_3 \mid p) \cdot f(p) = f(y_1 \mid p) \cdot f(y_2 \mid p) \cdot f(y_3 \mid p) \cdot f(p) =$$
$$= P(y_1 = 1 \mid p) \cdot P(y_2 = 0 \mid p) \cdot P(y_3 = 1 \mid p) \cdot f(p) = p \cdot (1 - p) \cdot p \cdot 1.$$

Выходит, что апостериорная плотность распределения параметра p должна иметь вид

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

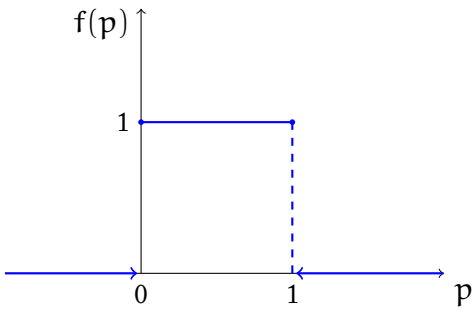
Осталось восстановить нормировочную константу. Вспоминаем, что интеграл по области определения апостериорной плотности распределения должен быть равен единице

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

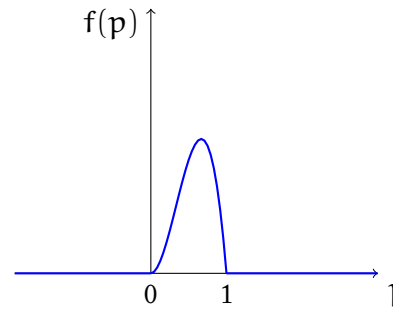
Итак, ваши авации! Апостериорное распределение параметра p:

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

Априорное распределение:



Апостериорное распределение:



[2]Упорный читатель поймёт эту шутку в следующей главе

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

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

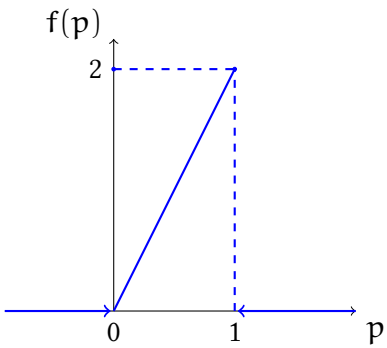
Восстанавливаем константу:

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

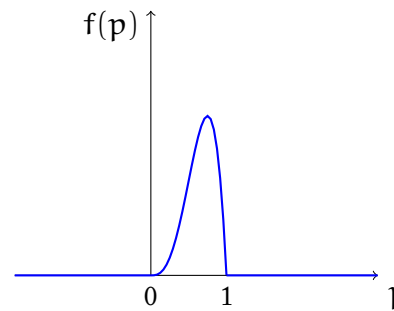
Снова получаем апостеринрую функцию плотности

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

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 $P(p > 0.5)$ и апостериорную вероятность того, что $P(p > 0.5 \mid y_1, \dots, y_3)$, а также априорное и апостериорное математические ожидания, $E(p)$ и $E(p \mid y_1, \dots, y_3)$.

Равномерное распределение:	Распределение бабушки:
$P(p > 0.5) = \int_{0.5}^{1} 1 \, dp = 0.5$	$P(p > 0.5) = \int_{0.5}^{1} 2p \, dp = 0.75$
$P(p > 0.5 \mid y) \approx 0.68$	$P(p > 0.5 \mid y) = 0.81$
$E(p) = \int_{0}^{1} p \cdot 1 \, dp = 0.5$	$E(p) = \int_{0}^{1} p \cdot 2p \, dp \approx 0.66$
$E(p \mid y) = \int_{0}^{1} 12 \cdot p^3 \cdot (1-p) \, dp = 0.6$	$E(p \mid y) = \int_{0}^{1} 20 \cdot p^4 \cdot (1-p) \, dp \approx 0.66$

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 β_F зависит от выбранной функции потерь. Вспомните, что происходило с красавицей в одном из чудес первой главы. Она выбирала в зависимости от системы наград и наказаний.

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.

На самом деле, байесовский подход позволяет анализировать различные функции потерь, которые мы можем использовать для обучения моделей, и понимать, что именно мы получаем в качестве прогноза на выходе. Иногда можно даже обнаружить, что функция потерь была нами выбрана не очень удачно. О таких вещах мы обязательно поговорим немного позже, а сейчас для полноты картины найдём для равномерного априорного распределения моду и медиану.

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

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

Найдём её!

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

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. Про такие сложности и про то как умело на практике их избегают компьютеры мы поговорим в следующей главе.

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.

О доверительных и байесовских интервалах

Едем дальше. 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. Доверительные интервалы обладают странным свойством. Если мы построили 95% доверительный интервал, то говорить, что истинное значение параметра p попадает в этот интервал с вероятностью 0.95 неправильно. Этот 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.

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.

Отдельно стоит сказать, что при определённых условиях значение функции плотности будет одинаковой на правом и левом конце интервала. Чтобы выяснить когда происходит такая интересная штука, читателю предлагается решить упражнение 11.

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. Зная факт из упражнения 11, можно было бы воспользоваться тем, что $f(a) = f(b)$ и найти доверительный интервал из 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», легко могли выпасть сейчас. 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 проблемы.

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 $f(b \mid y)$ в $P(y_4 = \text{карась} \mid y)$. Сделаем это несколькими способами.

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

$$E(Z) = E(E(Z \mid W)).$$

Если случайная величина Z принимает значения 0 и 1, тогда

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

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

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

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 P, значит

$$E_P(Z) = E_P(E_P(Z|W)).$$

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

$$E_{P_A}(Z) = E_{P_A}(E_{P_A}(Z \mid W))$$

.

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

$$E_P(Z \mid A) = E_P(E_P(Z \mid W, A) \mid A).$$

Вернёмся к задаче и применим к ней этот интересный факт:

$$P(y_4 = \text{карась} \mid y_1, y_2, y_3) = \\ = E(P(y_4 = \text{карась}|p, y_1, y_2, y_3) \mid y_1, y_2, y_3) = \\ = E(p \mid y_1, y_2, y_3) = 0.6.$$

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.

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

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. Выше, мы вскользь вспомнили о чуде, связанном со Спящей Красавицей и её потерями. Сейчас пришло время вспомнить про чудо условной независимости.

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

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

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

Способ третий, интегрально-влобовый: поговорим о прогнозировании чуть более подробно, в общем, непрерывном случае. В первой главе, когда мы

обсуждали чудеса условной вероятности и вспоминали формулы, мы выяснили, что из совместной плотности распределения $f(x, y)$ можно получить частную плотность $f(x)$, выинтригировав совместную плотность по переменной y, а именно

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

Мы говорили, что эта формула является аналогом формулы для поиска полной вероятности. Мы перебираем континуальное количество гипотез для переменной Y и находим плотность для X. Раз уж мы начали делать флэшбэки к первой главе, давайте заодно вспомним упражнение 9. Если недобросовестный читатель проигнорировал его, то самое время вернуться назад и разобрать.

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

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

Это же так? Я боюсь ошибиться в этой части с условиями.

Получаем интеграл произведения двух известных нам плотностей. Для случая карасей и щук получаем

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

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. Тому как компьютер борется с этими проблемами мы посвятим несколько глав.

Сейчас мы предлагаем читателю закрепить всё то, о чём мы говорили выше и попробовать решить парочку упражнений. В упражнении 1 читателю предлагается сделать пару ставок на априорные распределения. В упражнении 2 предлагается сделать байесовский вывод с одним наблюдением и геометрическим распределением. По механике оно повторяет всё то, о чём говорилось выше, однако в нём нет ни лагранджианов, ни чётвертых степеней.

1.3 Распределение Бернулли в общем случае

Вы уже решили самостоятельно предложенные выше упражнения? Если нет, бегом решать! Если да, то вы большой молодец и заслуживаете новых интересностей! В этом разделе речь пойдёт про общий случай для распределения Бернулли. Давайте представим себе следущую ситуацию.

Испанские конкистадоры высадились в Южной Америке. Вместе с собой они привезли кучу интересных прогрессивных европейских вундервафель. Они предлагают индейцам сыграть в следующую игру. Испанцы загадывают одну из сторон монетки. Эрнан Кортес подкидывает её. Если выпадает сторона, загаданная испанцами, то индейцы покидают свои земли. Если выпадает сторона, загаданная индейцами, европейцы отдают им все свои вундервафли и уходят с миром. Старейшины местного племени собрались на совет, на котором каждый из них хотел бы высказать своё мнение о выпадении орла в виде априорного распределения.

- Первый старейшина вообще не знает что такое вероятность.
- Второй старейшина не знает что такое монетка, но знает что такое вероятность.

- Третий старейшина однажды видел монетку, он даже вертел её в своих руках и говорит, что, если она сделана хорошо (а скорее всего это так), то будет выпадать с вероятностью $\frac{1}{2}$.

- Четвёртый старейшина уверен, что монетка у испанцев односторонняя. На ней есть только та сторона, которую выберут испанцы. Какую именно сторону они выберут, он не знает.

- Пятый старейшина склонен полагать, что монетка неправильная и та сторона, которую выберут испанцы,выпадает чаще. Какую именно сторону они выберут, он не знает.

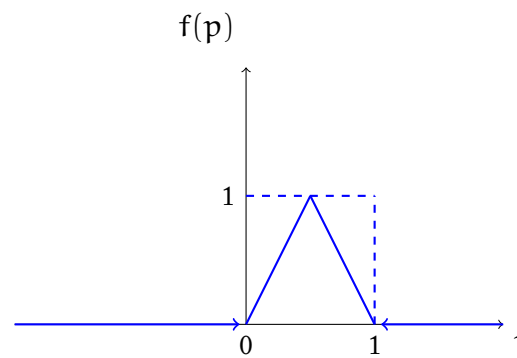
Попробуйте самостоятельно придумать априорные распределения, а уже после читайте решение.

Попробуем сформулировать априорные ожидания старейшин на привычном для нас языке. Поскольку первый старейшина не знает что такое вероятность, он может закодировать своё незнание несобственным распределением $\mathcal{U}[-\infty; +\infty]$.

Второй старейшина знает что такое вероятность, он знает, что она лежит на отрезке от нуля до единицы. Он своё незнание может выразить как $\mathcal{U}[0; 1]$.

Пик априорного распределения третьего старейшины должен приходиться на $\frac{1}{2}$. Этого можно было бы достичь, используя треугольное распределение.

$$f(\beta) = \begin{cases} 4p & , p \in [0; \frac{1}{2}] \\ 4(1-p) & , p \in [\frac{1}{2}; 1] \\ 0 & , \text{иначе} \end{cases}$$



Другим выходом из этой ситуации может стать бэта-распределение, с которым читатель, скорее всего, знакомится впервые. Не беда! Вот его плотность:

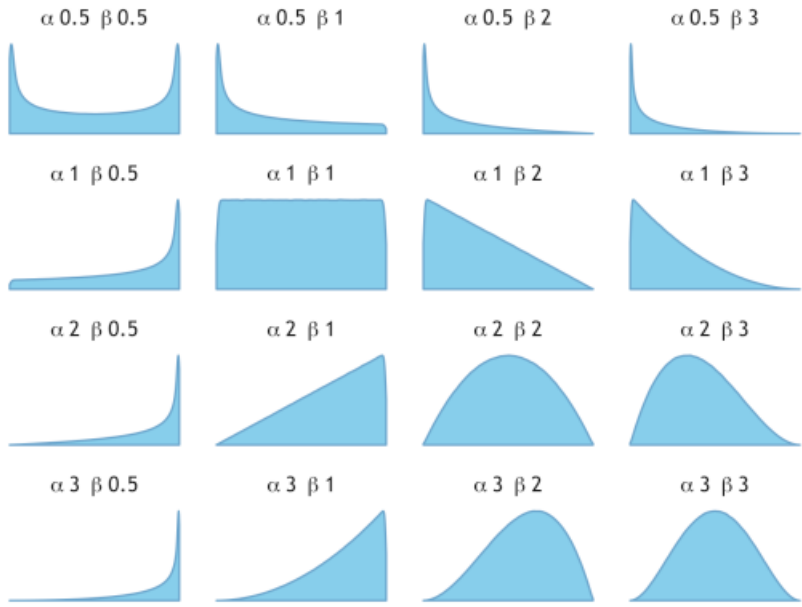


Рис. 1.1. Разные формы бэта-распределения

$$f(p) = \begin{cases} \frac{1}{B(\alpha,\beta)} \cdot p^{\alpha-1} \cdot (1-p)^{\beta-1}, & \text{ежели } p \in [0;1] \\ 0, & \text{иначе.} \end{cases}$$

В данном случае, $B(\alpha, \beta)$ известная из курса по матану бэта-функция. Именно в честь неё распределение так называется. Случайная величина, имеющая такое распределение, принимает свои значения на отрезке от нуля до единицы. Благодаря варьированию параметров α и β мы можем получать совершенно разные формы для априорной плотности. Примеры плотностей можно увидеть на рисунке 1.1 на странице 19.

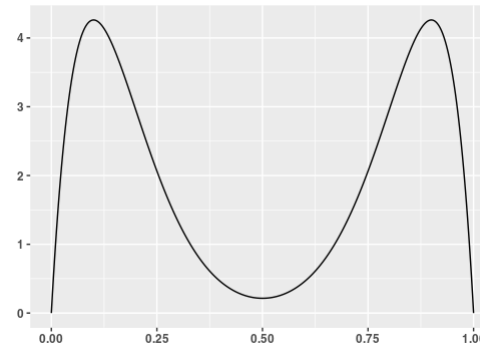
Третий старейшина может выбрать распределение с пиком в точке 0.5. Например, ему подойдёт распределение с $\alpha = \beta = 2$. При разной степени уверенности, он может выбрать разную остроту пика, как следует отрегулировав параметры. Например, для $\alpha = \beta = 3$ на хвосты приходится меньше вероятностной массы. Старейшина сильнее уверен в правильности монетки.

Мнение четвёртого старейшины можно описать с помощью U- образного распределения с $\alpha = \beta = 0.5$. В этом случае основная вероятностная матрица концентрируется у нуля (на монетке две решки) и у единицы (на монетке два орла). Понизив значения параметров, мы сделаем распределение ещё более резким и уберём ещё больше вероятностной массы из центра.

Мнение пятого старейшины можно выразить с помощью бимодального распределения. Например, с пиками в 0.3 и 0.7. U-образное в данном случае не подойдёт, так как оно сигнализирует именно либо о нуле, либо о единице. Для двугорбого распределения можно воспользоваться смесью двух бэта-распределений и сложить плотности, например, $B(2, 10)$ и $B(10, 2)$. Обратите внимание, сложить не случайные величины, а именно плотности. Мы хотим получить смесь из двух распределений. После складывания нужно не забыть отнормировать площадь всего этого дела к единице.

Прогнав следующий код в R можно посмотреть как будет выглядеть сумма плотностей двух бэта-распределений.

```
library(ggplot2)
x <- seq(0,1,by = 0.001)
y1 = dbeta(x,10,2)
y2 = dbeta(x,2,10)
qplot(x,y1+y2,geom='line')
```



Теперь, зная, как задаются различные априорные мнения, мы можем перейти к процедуре вероятностного вывода и посмотреть как будет выглядеть байесовская оценка для p в общем случае. Пусть индейцы поиграли с испанцами в предложенную игру на что-нибудь незначительное и собрали выборку $y_1, \ldots, y_n$. Априорно предполагается, что параметр p имеет бэта-распределение, $B(\alpha, \beta)$. Найдём апостериорную плотность.

Так как $y_1 + \ldots + y_n = n \cdot \bar{y}$, то

$$f(p \mid y_1, \dots y_n) \propto f(y_1, \dots, y_n \mid p) \cdot f(p) = \\ = p^{n\bar{y}}(1-p)^{n-n\bar{y}} \cdot p^{\alpha-1} \cdot (1-p)^{\beta-1} = \\ = p^{n\bar{y}+\alpha-1} \cdot (1-p)^{n-n\bar{y}+\beta-1}.$$

Уже видно, что на выходе снова получилось бэта-распределение, но с новыми параметрами. Восстановим недостающую константу

$$\text{const} \cdot \int_0^1 p^{n\bar{y}+\alpha-1} \cdot (1-p)^{n-n\bar{y}+\beta-1} = 1 \Rightarrow \text{const} = \frac{1}{B(n\bar{y}+\alpha, n-n\bar{y}+\beta)}$$

Параметр p имеет апостериорное бэта-распределение с параметрами $n\bar{y} + \alpha$ и $n - n\bar{y} + \beta$. К параметру α прибавляется количество выпавших орлов. К параметру β прибавляется количество выпавших решек. Априорные значения параметров α и β интерпретируются, как наша вера в правильность монеты. Чем сильнее дисбаланс между ними, тем сильнее скошена наша вера. Чем больше они, тем меньше дисперсия нашей веры.

Попробуем использовать в качестве точечной оценки математическое ожидание апостериорного распределения. Для этого его нужно посчитать.

$$E(p \mid y_1, \dots y_n) = \\ = \int_0^1 p \cdot \frac{1}{B(n\bar{y}+\alpha, n-n\bar{y}+\beta)} \cdot p^{n\bar{y}+\alpha-1}(1-p)^{n-n\bar{y}+\beta-1} dp = \\ = \int_0^1 \frac{1}{B(n\bar{y}+\alpha, n-n\bar{y}+\beta)} \cdot p^{n\bar{y}+\alpha+1-1}(1-p)^{n-n\bar{y}+\beta-1} dp = \\ = \frac{B(n\bar{y}+\alpha+1, n-n\bar{y}+\beta)}{B(n\bar{y}+\alpha, n-n\bar{y}+\beta)} = \frac{n\bar{y}+\alpha}{n+\alpha+\beta} \cdot \frac{B(n\bar{y}+\alpha, n-n\bar{y}+\beta)}{B(n\bar{y}+\alpha, n-n\bar{y}+\beta)} = \\ = \frac{n\bar{y}+\alpha}{n+\alpha+\beta}$$

Выглядит внушительно. Даже пришлось немного вспомнить свойства бэта-функции.[3] Обратите внимание, что до этого дня никакой другой метод оцени-

[3] $B(q+1, p) = \frac{q}{p+q} \cdot B(q, p)$

вания не давал нам такую красивую точечную оценку. Все эти вычисления с интегралами лишний раз подчёркивают, что даже в самом простом случае нам нужно обращаться к такому мощному инструменту, как бэта-функции. Очень редко попадаются случаи, когда апостериорную плотность можно найти в символьном виде. Чаще всего интегралы вообще не берутся.

При отсутствии наблюдений получаем оценку $\frac{\alpha}{\alpha+\beta}$, совпадающую с априорным математическим ожиданием. Если, наоборот, наблюдений очень много, вклад априорного мнения становится всё менее значительным и оценка приближается к среднему.

Если при этом монетка, действительно, является односторонней, мы никогда не получим в качестве точечной оценки единицу. При каждом новом подбрасывании апостериорная оценка будет всё ближе к единице, но никогда не достигнет её.

В каком-то смысле байесовский способ оценивания предполагает встроенную регуляризацию[4]. Если бы мы использовали на одном наблюдении метод максимального правдоподобия, мы бы получили $p = 1$ и переобучились бы под выборку, сделав неправильные выводы о природе исследуемого явления. Байесовский вывод дал бы близкую к единице оценку, но не равную ей, немного застраховав нас от запоминания выборки. Об этом мы более подробно поговорим позже. В качестве упражнения дотошному читателю предлагается найти апостериорные оценки для каждого старейшины и сравнить их между собой.

1.4 Маша, медведи и нормальное распределение

До этого мы рассматривали ситуацию, когда наблюдения у нас дискретны, а распределение параметра непрерывно. Давайте посмотрим на ситуацию, когда непрерывны оба распределения.

Маша прячется от Медведей в точке m на числовой прямой. Есть несколько Медведей, каждый из которых обнюхивает всю числовую прямую в поисках Маши. Медведю номер i кажется, что Машей сильней всего пахнет в точке y_i.

[4]Под регуляризацией понимается накладывание дополнительных ограничение на параметры модели, которые могли бы спасти нас от переобучения, то есть от излишне сильного вылизывания обучающих данных. Об этом мы начнём по-серьёзному говорить в 5 главе и продолжим в 6.

Естественно, Медведи могут ошибаться, например, у них может быть заложен нос, поэтому $y_i \mid m \sim \mathcal{N}(m, 2^2)$. При фиксированном m величины y_i независимы. Известно, что $y_1 = 0.5$, $y_2 = -1$. Априорно известно, что место, где спряталась Маша имеет нормальное распределение, $m \sim \mathcal{N}(1, 4^2)$. Нам нужно:

1. Найти апостериорную плотность распределения параметра m.
2. Найти апостериорные моду, медиану и математическое ожидание.
3. Найти $P(m > 1 \mid y_1, y_2)$.
4. Найти $f(y_3 \mid y_1, y_2)$ и $E(y_3 \mid y_1, y_2)$.

Целеустремлённый читатель должен сначала самостоятельно попытаться решить эту задачу, а уже после продолжить читать текст.

Посмотрим немного подробнее на наше априорное мнение о том, где сидит Маша, $m \sim \mathcal{N}(1, 4^2)$. Значение 1 в данном случае — наше лучшее предположение о том, где она может находиться, а 2^2, в свою очередь, это наша степень доверия к этому предположению. Чем меньшее значение дисперсии мы берём в нашем априорном мнении, тем больше наше доверие к нему.

Делай раз! Апостериорная плотность Маши:

$$f(m \mid y_1, y_2) \propto f(y_1, y_2 \mid m) \cdot f(m) = \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{(0.5-m)^2}{2\cdot 4}\right) \cdot \\ \cdot \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{(-1-m)^2}{2\cdot 4}\right) \cdot \frac{1}{4\sqrt{2\pi}} \exp\left(-\frac{(m-1)^2}{2\cdot 16}\right)$$

Воспользуемся магической силой уже привычного нам значка $\propto$ и для простоты расчётов пренебрежём кучей констант

$$f(m \mid y_1, y_2) \propto \exp\left(-\frac{(0.5-m)^2}{2\cdot 4}\right) \cdot \exp\left(-\frac{(-1-m)^2}{2\cdot 4}\right) \cdot \exp\left(-\frac{(m-1)^2}{2\cdot 16}\right)$$

Сольём всё,что находится под знаком экспоненты в единое целое и попробуем упростить

$$\frac{(0.5-m)^2}{2\cdot 4}+\frac{(-1-m)^2}{2\cdot 4}+\frac{(m-1)^2}{2\cdot 16}=$$
$$=\frac{4(m-0.5)^2+4(m+1)^2+(m-1)^2}{32}=\frac{9m^2+2m+6}{32}$$

Используем двойную магию. С одной стороны пренебрегаем константой, с другой создаём новую для того, чтобы создать полный квадрат. Привыкайте пользоваться этой магией для своего удобства. Не забываем перекинуть в знаменатель лишнюю девятку

$$\exp\left(-\frac{9m^2+2m+6}{32}\right)\propto\exp\left(-\frac{9m^2+2m}{32}\right)=$$
$$=\exp\left(-\frac{m^2+\frac{2}{9}m}{\frac{32}{9}}\right)=\exp\left(-\frac{m^2+2\cdot\frac{1}{9}m+\frac{1}{81}-\frac{1}{81}}{\frac{32}{9}}\right)\propto$$
$$\propto\exp\left(-\frac{m^2+2\frac{1}{9}m+\frac{1}{81}}{\frac{32}{9}}\right)=\exp\left(-\frac{(m+\frac{1}{9})^2}{2\cdot(4/3)^2}\right)$$

Параметр m имеет нормальное апостериорное распределение

$$m\mid y_1,y_2\sim\mathcal{N}(-1/9,(4/3)^2).$$

Обратите внимания, что после того как Медведи попытались вынюхать где находится Маша, самое вероятное её положение изменилось, а дисперсия её положения уменьшилась.

Новая информация сместила априорную плотность влево и вытянула её вверх, в силу того, что Медведи вынюхали похожие вещи.

Делай два! Мода и медиана для нормального распределения совпадают с математическим ожиданием. Какую бы функцию ошибки мы не выбрали, при таком априорном мнении мы натолкнёмся на один и тот же результат.

Делай три! Обратите внимание, что до запуска Медведей, $P(m>1)=0.5$. После запуска, эта вероятность уменьшится,так как распределение очень сильно съедет влево.

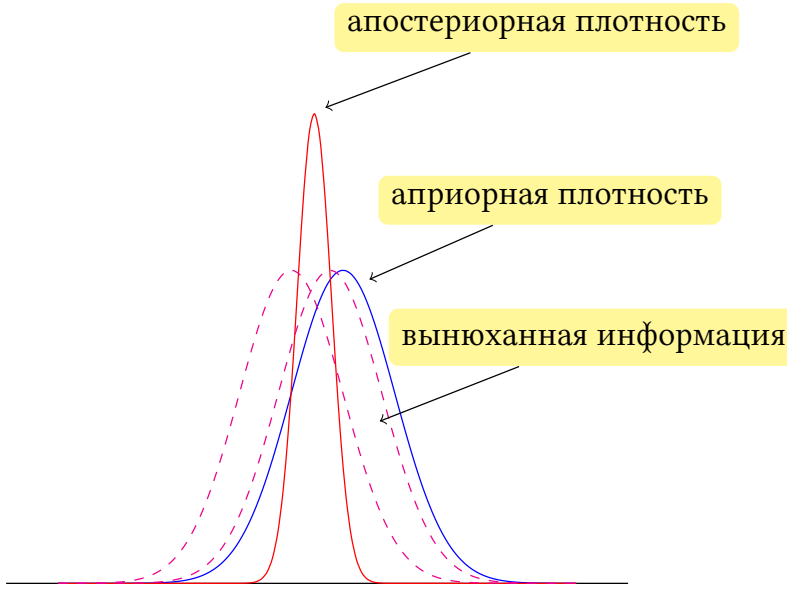


Рис. 1.2. Информация о Маше

$$P(m > 1 \mid y_1, y_2) = 1 - P(m \leqslant 1 \mid y_1, y_2) =$$
$$= 1 - P\left(\frac{m + {}^1/_9}{{}^4/_3} \leqslant \frac{1 + {}^1/_9}{{}^4/_3} \mid y_1, y_2\right) = 1 - \Phi\left(\frac{10}{12}\right) \approx 0.2.$$

Делай четыре! Найдём $f(y_3 \mid y_1, y_2)$ и $E(y_3 \mid y_1, y_2)$. Будем делать это под слоганом: «Каждой Маше по три Медведя!»:

Я не уверен, что то, что ниже правильно. Можно ли попроще?

$$f(y_3 \mid y_1, y_2) = \int_{-\infty}^{+\infty} f(y_3, m \mid y_1, y_2)\, dm =$$
$$= \int_{-\infty}^{+\infty} f(y_3 \mid y_1, y_2, m) \cdot f(m \mid y_1, y_2)\, dm.$$

Чтобы найти плотности распределение y_3, мы должны провести свёртку по двум нормальным распределениям

$$\int_{-\infty}^{+\infty} \mathcal{N}(m, 4) \cdot \mathcal{N}\left(-\frac{1}{9}, \frac{16}{9}\right) \, dm.$$

Если взять этот интеграл (а мы оставляем это дотошному читателю в виде упражнения), можно получить, что

$$y_3 \mid y_1, y_2 \sim \mathcal{N}\left(-\frac{1}{9}, \frac{52}{9}\right).$$

Теперь, когда нюхательные способности третьего Машиного Медведя предсказаны, вы можете попробовать проделать всё то же самое самое, предположив, что вам вообще ничего неизвестно и $m \sim \mathcal{U}(-\infty; +\infty)$. Именно это вам предлагается сделать в упражнении 3. Стоит отметить, что результат у вас, при этом, получится похожим на случай нормального априорного распределения с большой дисперсией. Почему это происходит именно так, мы постараемся выяснить ровно через одну главу, когда будем обсуждать энтропию.

После вы можете попытаться решить упражнение 4. Это точно такой же байесовский вывод, но для нормального распределения в общем случае. Многие формулы в этом выводе получаются довольно громоздкими. Поэтому, если вы запутаетесь по ходу его самостоятельного решения, не отчаивайтесь и загляните в наше решение. В упражнении 5 вас ждёт несколько прикольных асимптотических выводов, связанных с нормальными априорными распределениями.

1.5 О максимальном правдоподобии, серьёзности и бастардах

Экспериментальный раздел, в котором я явно перегнул с аналогиями.

Обычно, в серьёзных книгах по математике сначала формулируется теорема, а после идёт её доказательство. Если бы каждая серия Игры Престолов была бы устроена, как теорема, то в её начале томный грубый закадровый голос произносил бы длиный перечень персонажей, которые должны умереть в

течение серии, а уже после показывалось бы как именно они умирают. В конце бы голос говорил: «Все видели? А? Они умерли! Я же говорил!».

Почему в математике нельзя обойтись без спойлеров? Мир бы тогда был совершенно другим! В каждой книге вслед за безупречной пеленой логических рассуждений следовал бы какой-нибудь безумный факт, сносящий крышу читателя. При этом, накануне лекции недруги молодого студента, собирающегося её посетить, жадно штудировали бы книги в поисках спойлеров, а наутро, врывались бы в его комнату в общежитии и выдавали бы список из утверждений и теорем. Студент же, в свою очередь, пытался бы с зажатыми ладонями ушами добежать скорее до лекционной аудитории, чтобы увидеть оригинальные рассуждения лектора, а в конце лекции обомлеть от удивления.

В таком мире всё было бы совершенно иначе. Оценка метода максимального правдоподобия является частным случаем байесовской оценки. В качестве априорного распределения нужно взять равномерное, а в качестве точечной оценки моду апостериорного распределения. Неприятно было сейчас нарваться на такой спойлер, правда?

Чтож, теперь, зная кто умрёт в этом эпизоде, давайте заметим следущую штуку:

$$\underbrace{f(p \mid y_1, \dots, y_n)}_{\text{апостериорная плотность}} \propto \underbrace{f(y_1, \dots, y_n \mid p)}_{\text{функция правдоподобия}} \cdot \underbrace{f(p)}_{\text{априорная плотность}}$$

Если $p \sim U[0; 1]$, тогда $f(p) \propto const$, значит

$$f(p \mid y_1, \dots y_n) \propto f(y_1 \dots y_n \mid p).$$

Таким образом максимум функции правдоподобия совпадает с максимумом апостериорной плотности. В свою очередь точка, где плотность достигает максимума, называется модой. И правда, получилось, что

$$\hat{p}^{ML} = \text{Mod}(p \mid y_1, \dots y_n).$$

Например, для случая карасей

$$\hat{p}^{ML} = \bar{y} = \frac{2}{3} = \text{Mod}\, f(p \mid y_1, y_2, y_3).$$

Забавно, но если мы делаем допущение, что все гипотезы априори одинаково вероятны, байесовский подход сведётся к принципу максимального правдоподобия. Результаты, полученные с помощью частотных методов являются частным случаем байесовского подхода.

А что, если изначально гипотезы не одинаково правдоподобны априори? Получается, что в такой ситуации использование метода максимального правдоподобия приводит нас к неверным выводам? Вспомните упражнение про золотые монеты, которое вы, конечно же сделали после изучения истории про бабушку и карасей.

В этом упражнении оценка максимального правдоподобия сказала нам, что вероятность вытащить из шляпы серебряную монету равна единице. И это крайне неточно. В шляпе, из которой тянулись монетки, было несколько золотых экземпляров. Если увеличить число испытаний, то оценка станет надёжнее. Как известно, оценки максимального правдоподобия состоятельны, но, тем не менее, непонятно насколько огромным должен быть размер выборки, чтобы охватить все существующие ситуации.

Байесовский подход позволяет в каком-то смысле избежать такой проблемы. Он говорит, что вероятность высокого значения p довольно велика, но тем не менее, не отбрасывает возможность самых маленьких значений этого параметра. Все видели? Да? А я же говорил[5]!

Ещё раз, ещё раз! Метод максимального правдоподобия — частный случай байсовского вывода. На самом деле, довольно часто будет получаться, что какая-то вероятностная модель порождает что-то уже давно нам знакомое, настолько же знакомое, как и метод максимального правдоподобия.

[5] Авторы бы хотели отметить, что негативно относятся к спойлерам и принести свои извинения осерчавшему на них читателю. Один из авторов хотел бы передать привет одному из своих преподавателей, который в его студенчестве проспойлерил ему многие вещи из Игры Престолов. В том числе, во время зачёта по программированию, с которого автор ну никак не мог убежать с закрытыми ушами.

Давайте для полноты картины выясним, как соотносятся между собой мода апостериорного распределения и ML-оценка при произвольном априорном распределении. Попробуйте накидать своё решение на бумажке, а после подглядите в следующую строчку со спойлерами:

$$\begin{aligned}\operatorname{Mod}(\beta \mid y) = \arg\max_{\beta} f(\beta \mid y) = \\ = \arg\max_{\beta} f(y \mid \beta) \cdot f(\beta) = \arg\max_{\beta} (\ln f(y \mid \beta) + \ln f(\beta)).\end{aligned}$$

Видно, что мода апостериорного распределения, в общем случае, это коррекция ML оценки. Этот факт снова наталкивает нас на мысли о регуляризации, которые будут как следует развиты в шестой главе.

1.6 Про сопряжённые распределения и сложности

В задачах про Машу и Медведей априорное распределение было нормальным. После применения формулы Байеса мы снова получали на выходе нормальное распределение, но уже с другими параметрами. Новые параметры

при этом зависели от старых параметров и собранных наблюдений. Таким образом, чтобы получить из априорного распределения апостериорное, в данном случае не надо применять никаких сложных формул и не нужно никаких больших компьютерных мощностей. Аналогично получалось в случае бэта-распределения для монеток. Удобно.

А что если в качестве априорных распределений всегда брать такие, чтобы после домножения на правдоподобие, апостериорное распределение принадлежало тому же самому классу распределений, но с другими параметрами? Тогда бы все байесовские алгоритмы свелись бы к банальному пересчёту и всё. Конечно же мы ограничим себя в выборе априорных распределений, зато у нас не будет никаких вычислительных проблем.

Именно так размышляли байесовцы более чем тридцать лет назад. Они решили пожертвовать своей свободой выбора в пользу простоты вычислений и нашли кучу распределений, которые ведут себя подобным образом и даже придумали для них название.

Определение. Семейство распределений $f(\beta \mid \Theta)$ называется **сопряжённым по отношению к наблюдаемой выборке**, если апостериорное распределение $f(\beta \mid y_1, \dots, y_n)$ тоже принадлежит этому семейству.

Любопытный читатель, скорее всего, захочет узнать немножечко побольше о таких распределениях. Сделать это он может, например в статейке Айвазяна [**?**].

Такие распределения моментально освобождали исследователя от необходимости брать сложные интегралы, решать уравнения высоких степеней, искать условные экстремумы и делать другие страшные вещи. Всё это было достаточно сделать только лишь один раз. Наши Деды позаботились о потомках, провели все эти вычисления и передали нам в наследие формулы пересчёта[6].

Тем не менее, это обернулось для байесовского подхода лишь дополнительной критикой. В результате, в течение 20 века байесовские модели получили слабое распространение.

В 1990-е годы произошли первые изменения. Мир вспомнил про наивный байесовский классификатор (про него мы подробно поговорим в следующей главе). Оказалось, что он иногда работает на порядок лучше, чем другие, более

[6]На википедии можно найти целую таблицу с формулами пересчёта: https://en.wikipedia.org/wiki/Conjugate_prior

сложные модели. Долгое время самые простые версии таких классификаторов с очень высокой точностью защищали пользователей интернета от спама.

Этот алгоритм называется наивным из-за того, что он допускает независимость там, где её, по гамбургскому счёту, нет. Например, в случае фильтра для спама вводится предпосылка, что вхождение слов в текст — независимые события, что конечно же не так. Однако это работает. Видимо, для успешного выявления спама в подробности сообщения вникать не обязательно: достаточно просто уловить суть, посмотрев какие слова оно содержит. Немного позже мы обязательно обучим такой алгоритм.

В дополнение ко всему прочему, люди придумали MCMC, алгоритмы Монте-Карло по схеме Марковской цепи. Благодаря мощным компьютерам и MCMC байесовцы снова обрели свободу. Теперь, выбирая какое-то сложное априорное распределение, они не бояться случайно нарваться на убойные вычисления. Компьютер за них генерирует любые апостериорные распределения и предоставляет исследователю возможность ответить с помощью них на свои вопросы.

Постепенно выяснилось, что байесовские методы в состоянии элегантно решать разные самые сложные задачи. Например, совсем недавно байесовские методы скрестили с нейронными сетями. Это скрещивание открыло перед людьми колоссальные возможности.

Одной из особенностей нейронных сетей является их самоуверенность. Если взять кучу разных картинок и случайным образом расставить на них метки, а после на этой разметке обучить свёрточную нейросеть, она запомнит какой картинке соответствует какая метка и будет на тренировочной выборке показывать превосходные результаты. При этом, на тестовой выборке, её ответы ничем не будут отличаться от случайного угадывания. Сеть никак не обучилась, не нашла никаких закономерностей и никак не сообщила об этом человеку. Она просто запомнила картинки.

После, для каждой новой картинки, сеть будет выдавать какой-то прогноз. Точно также она выдаст прогноз, если мы отдадим ей какую-то совершенно новую картинку, которую она никогда раньше не видела, вместо того, чтобы сказать нам, что такую картинку она видит впервые.

Более того, сеть можно обмануть. Представим, что мы отдали ей котика а и она распознала его. В качестве ответа на нашу картинку, сеть выдала метку «котейка». Возьмём ту же самую картинку с котиком, откроем её в каком-

нибудь графическом редакторе и слегка подёргаем пиксели на лапке. Снова отдадим картинку нейросетке. Нейросеть спокойно может заявить нам, что это вовсе не котик, а самолёт.

В 2015-2016 годы были предприняты первые попытки скрестить нейросети с байесовским подходом. Судя по всему, такое скрещивание позволяет решить проблему самоуверенности сетей. Байесовские сети сообщают человеку насколько они уверенны в своих прогнозах. На нейросети в этой книге мы смотреть не будем. Тем не менее, кое-какие довольно новые модели постараемся пощупать.

Сейчас мы предлагаем читателю прорешать оставшиеся упражнения. Навыки байесовского вывода пригодятся нам в следующих главах. Как говорится, Деды вычисляли!

1.7 Ещё задачи

Упражнение 1. Предположим, что вы оказались в одной из ситуаций, приведённых ниже. Как вы будете действовать?

1. Пусть мы вообще ничего не знаем о коэффициентах в модели линейной регрессии. Как мы можем замоделировать это незнание?

   Другой вопрос: какое распределение можно использовать в качестве априорного, если мы знаем, что коэффициент β принимает значение на отрезке $[\beta_d; \beta_u]$?

2. Пусть нам необходимо оценить какой-то коэффициент в модели линейной регрессии, β. Нам точно известно, что он положительный, $\beta \in [0; +\infty)$, и, скорее всего, не принимает высокие значения. Какое распределение для параметра β можно было бы взять в качестве априорного?

3. Снова коэффициент, снова в регрессии. Теперь он точно лежит на отрезке от 0 до 42. Наши априорные мысли говорят, что он скорее всего ближе к 42, чем к нулю. Как задать такое априорное распределение?

4. Джордж Мартин пишет романы из цикла «Песнь льда и пламени». Пусть случайная величина Y — это число убитых в книге персонажей.

Так как мы имеем дело со случайной величиной — счётчиком, её можно было бы замоделировать с помощью распределения Пуассона с параметром λ, характеризующим частоту, с которой умирают персонажи. Какое априорное распределение можно использовать для параметра λ?

5. И снова регрессия! Только теперь коэффициента два, и мы уверены, что чем больше первый, тем меньше второй. Более того, бабушка говорит, что первый должен быть в районе тройки. Про второй никто ничего не знает. Что делать?

Тут задача про смеси, которую надо переделать

На самом деле, в описанных выше ситуациях, можно вести себя по-разному. Априорная информация сформулирована довольно размыто. Перед нами стоит задача математически точно сформулировать это размытое мнение. Можно сделать это многими разными способами. Например, авторы готовы сделать ставки на следующие распределения.

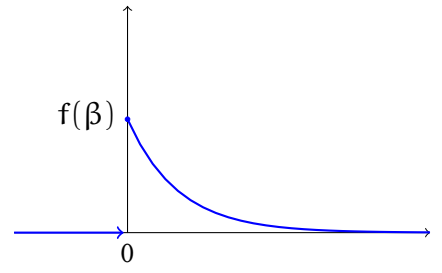
1. Если мы совсем ничего не знаем о параметре β, то нужно брать максимально размытое распределение. Например, нормальное распределение с большой дисперсией, $\mathcal{N}(0, 10^4)$. Если нам известно, что параметр β принимает значение на отрезке $[\beta_d; \beta_u]$, то в качестве априорного можно взять равномерное на этом отрезке распределение. Более того, в первом случае, когда $\beta \in \mathbb{R}$, в качестве априорного распределения можно взять $\mathcal{U}(\infty; +\infty)$. Конечно, настоящего равномерного распределения на неограниченном интервале не бывает, так как

$$\int_{-\infty}^{+\infty} 1 \, dx = \infty \neq 1.$$

Такое распределение называется **несобственным**. Оно говорит, что мы ничего не знаем о параметре и, что удивительно, не доставляет никаких технических неудобств. Байесовский вывод даст апостериорное распределение, пригодное для дальнейшего использования.

2. В качестве априорного распределения для коэфициента можно рассмотреть экспоненциальное, $\beta \sim Exp(-\beta)$. Область определения будет ограничена только положительными значениями. Большие значения коэффициент будет принимать с низкими вероятностями (подумайте о том может ли это побороть мультиколинеарность).

$$f(\beta) = \begin{cases} \exp(-\beta) & , \beta \in [0; +\infty) \\ 0 & , \text{иначе} \end{cases}$$



Также можно выразить своё мнение с помощью любого другого стандартного распределения, обладающего необходимыми для нас свойствами. Например, подойдёт распрелеление хи-квадрат или любое другое гамма-распределение.

3. Давайте возьмём какое-нибудь известное распределение, например экспоненциальное, и перешкалируем его к отрезку $[0; 42]$. Случайная величина $\beta \sim Exp(\lambda)$ принимает значение на $[0; +\infty)$. Тогда Случайная величина, имеющая распределение $\frac{42}{1+\alpha \cdot Exp(\lambda)}$ будет принимать значение как раз на требуемом отрезке! Изменяя параметр α можно следить за тем в какой части отрезка концентрируется подавляющая масса распределения. С помощью подобных трюков, можно модернизировать различные стандартные распределения и использовать их в качестве априорных.

4. Параметр λ характеризует интенсивность, с которой умирают персонажи. Мы наслышаны о том, что Джордж Мартин довольно кровожадный писатель, склонный к убийствам. Тем не менее, количество персонажей, которые погибли в течение книги не может быть слишком большим. Отсюда следует, что параметр λ можно замоделировать распределением, сосредотачивающим основную вероятностную массу в самом

начале числовой оси. Для таких целей вполне подойдёт любое гамма-распределение. Например, экспоненциальное или хи-квадрат. Параметры такого распределения, конечно же, нужно выбирать в зависимости от того насколько сильно мы верим в кровожадность Джорджа Мартина.

Кроме того, можно было предположить, что $f(\lambda) \propto \frac{1}{\lambda}$. Это распределение несобственное, но после байесовского вывода мы получим собственное апостериорное распределение.

5. Коэффициенты взаимосвязаны. Априорное распределение будет двумерным! Например, нормальным со следующими параметрами:

$$\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} \sim \left(\begin{pmatrix}3\\ 0\end{pmatrix} \; ; \; \begin{pmatrix}10^2 & -0.2\\ -0.2 & 10^5\end{pmatrix}\right)$$

Упражнение 2. В шапке лежит куча серебрянных и золотых монет. Пусть p — доля серебряных монет. Априорно Марианне кажется, что в шляпе серебряных монет намного больше, чем золотых. Она описывает свои ожидания плотностью:

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

Марианна тянет из шапки с возвратом монетки до тех пор, пока не вытащит серебряную. Пусть Марианна вытащила серебряную монету с первой попытки. Найдите:

1. Апостериорное распределение $f(p \mid y)$;
2. $P(p > 0.5 \mid y)$, $E(p \mid y)$, $\text{Med}(p \mid y)$, $\text{Mod}(p \mid y)$;
3. $P(y_{new} = 1 \mid y)$;
4. Самый короткий байесовский предиктивный интервал.

Начнём с апостерионого распределения:

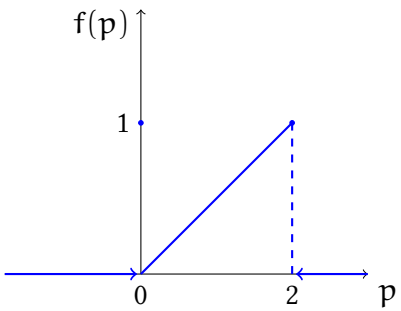
$$f(p \mid y) = \propto f(y \mid p) \cdot f(p) = p \cdot 2 \cdot p.$$

Осталось восстановить нормировачную константу:

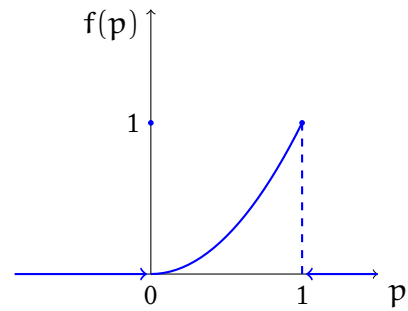
$$f(p \mid y) = const \cdot p^2$$

$$const \cdot \int_0^1 p^2 dp = \frac{c}{3} = 1 \quad \Rightarrow \quad const = 3$$

Априорное распределение:



Апостериорное распределение:



$$P(p > 0.5 \mid y) = \int_{0.5}^1 3p^2 dp = \frac{7}{8} \qquad E(p \mid y) = \int_0^1 3p^3 dp = \frac{3}{4}$$

$$P(p > Med) = 0.5 \quad \Rightarrow \quad \int_0^{Med} 3p^2 dp = 0.5 \quad \Rightarrow \quad Med = \frac{1}{\sqrt[3]{2}} \approx 0.8$$

$$P(y_{new} = 1 \mid y) = E(P(y_{new} = 1 \mid p, y) \mid y) = E(p \mid y) = \frac{3}{4}.$$

Апостериорная плотность достигает своего максимума в точке 1. $\Rightarrow$ $Mod(p \mid y) = 1$.

Построим самый короткий байесовский предиктивный интервал (HPD). Так как апостериорная плотность возрастает, самый короткий интервал будет примыкать к правой границе её области определения.

$$\int_q^1 3p^2 dp = 0.95 \quad \Rightarrow 1 - q^3 = 0.95 \quad \Rightarrow q = \sqrt[3]{0.05}$$

Самый короткий байесовский предиктивный интервал это $[\sqrt[3]{0.05}; 1]$.

Упражнение 3. Та же самая задача, про ту же самую Машу, что и в главе, но у нас новое априорное распределение, $m \sim \mathcal{U}(-\infty; +\infty)$. Сравните полученный результат с тем, что получалось в предыдущей задаче. А ещё найдите моду. Верно ли, что она совпадает с оценкой максимального правдоподобия? Почему?

Получим апостериорную плотность

$$f(m \mid y_1, y_2) \propto \exp\left(-\frac{(0.5-m)^2}{2\cdot 4}\right)\cdot \exp\left(-\frac{(-1-m)^2}{2\cdot 4}\right) \propto$$
$$\propto \exp\left(-\frac{(m+0.25)^2}{2\cdot 2}\right).$$

Нарисуем табличку!

Априорное распределение:	Апостериорное распределение:
$\mathcal{N}(1, 4^2)$	$\mathcal{N}(-\frac{1}{9}, 1.77)$
$\mathcal{U}(-\infty; +\infty)$	$\mathcal{N}(-\frac{1}{4}, 2)$

В случае, когда мы не ввели никакой априорной информации о параметре m, мы получили большую дисперсию, а также более сильно смещённое влево математическое ожидание. В случае, когда мы ввели априорное распределение, мы, предоставив дополнительную информацию, уменьшили дисперсию. Более того, мы не позволили вынюханным наблюдениям сдвинуть апостериорное математическое ожидание влево слишком сильно. Кстати говоря, модой будет -0.25. Она действительно совпадает с оценкой максимального правдоподобия, $\bar{y}$.

Упражнение 4. И снова Маша спряталась от Медведей в точке m на числовой прямой. Есть несколько Медведей, каждый из которых пытается вынюхать где же находится Маша. Медведю номер i кажется, что Машей сильнее всего пахнет в точке y_i. Всего у нас n Медведей. Медведи могут ошибаться, например, у них может быть заложен нос, поэтому $y_i \mid m \sim \mathcal{N}(m, \sigma^2)$. При фиксированном

m величины y_i независимы. Априорно известно, что место, где спряталась Маша имеет нормальное распределение, $m \sim \mathcal{N}(\mu, \tau^2)$. Найдите апостериорное распределение параметра m.

На самом деле нам нужно проделать ровно то же самое, что и для двух наблюдений. На первом шаге нужно избавиться от всех констант. На втором шаге нужно сгруппировать квадрат. Априорная плотность распределения и правдоподобие (совместная плотность распределения данных) имеют вид:

$$f(m) = \frac{1}{\sqrt{2\pi\tau^2}} \cdot \exp\left(-\frac{(m-\mu)^2}{2\tau^2}\right)$$
$$f(y \mid m, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - m)^2\right)$$

Найдём апостериорное распределение:

$$f(m \mid y, \sigma^2) \propto f(y \mid m, \sigma^2) \cdot f(m) =$$
$$= \frac{1}{(2\pi)^{(n+1)/2} \cdot \sigma^n \cdot \tau} \cdot \exp\left(-\frac{(m-\mu)^2}{2\tau^2} - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n}(y_i - m)^2\right) \propto^*$$

Обратим внимание на то, что в нашей формуле находится куча констант, которыми можно пренебречь, а потом восстановить по аналогии с тем, как мы это делали раньше. Множитель перед экспонентой — константа. Отбросим его. Раскроем скобки внутри экспоненты.

$$^* \propto \exp\left(-\frac{m^2 - 2m\mu}{2\tau^2} - \frac{\mu^2}{2\tau^2} - \frac{nm^2 - 2m\sum_{i=1}^{n} y_i}{2\sigma^2} - \frac{\sum_{i=1}^{n} y_i^2}{2\sigma^2}\right) \propto$$
$$\propto \exp\left(-\frac{m^2 - 2m\mu}{2\tau^2} - \frac{nm^2 - 2m\sum_{i=1}^{n} y_i}{2\sigma^2}\right)$$

Обратим внимание на то, что второе и четвёртое слагаемые внутри экспоненты — это константы, они не содержат m. Мы можем вытащить их в отдельный множитель и вынести за основную экспоненту. Пожертвуем ими. В

оставшихся двух слагаемых везде есть случайность в виде нашего m. Сделаем несколько очень громоздких преобразований. Сначала приведём дроби к одному и тому же знаменателю, а затем сгруппируем скобочки.

$$\frac{m^2 - 2m\mu}{2\tau^2} + \frac{nm^2 - 2m\sum_{i=1}^{n} y_i}{2\sigma^2} = \\ = \frac{(\sigma^2/n)\cdot(m^2 - 2\mu m)}{2(\sigma^2/n)\tau^2} + \frac{\tau^2(m^2 - 2m\bar{y})}{2(\sigma^2/n)\tau^2} = \\ = \frac{[\sigma^2/n + \tau^2]\cdot\left[m^2 - \frac{2\mu m(\sigma^2/n)}{\sigma^2/n+\tau^2} - \frac{2m\bar{y}\tau^2}{\sigma^2/n+\tau^2}\right]}{2(\sigma^2/n)\tau^2}$$

Переобозначим всё, что не зависит от m как константу:

$$\tilde{\mu} = \frac{\mu\cdot(\sigma^2/n)}{(\sigma^2/n) + \tau^2} + \frac{\bar{y}\cdot\tau^2}{(\sigma^2/n) + \tau^2}.$$

Выражение выше станет более компактным:

$$\frac{[\sigma^2/n + \tau^2]\cdot[m^2 - 2\tilde{\mu}\cdot m]}{2(\sigma^2/n)\tau^2}.$$

Переобозначим ещё одну константу:

$$\tilde{\tau}^2 = \frac{\tau^2\sigma^2/n}{(\sigma^2/n) + \tau^2}.$$

Всё, что находится в этой дроби снова не зависит от m. В итоге получаем, что

$$f(m \mid y, \sigma^2) \propto \exp\left(-\frac{m^2 - 2\tilde{\mu}m}{2\tilde{\tau}^2}\right) \propto \\ \propto \exp\left(-\frac{m^2 - 2\tilde{\mu}m + \tilde{\mu}^2}{2\tilde{\tau}^2}\right) = \exp\left(-\frac{(m-\tilde{\mu})^2}{2\tilde{\tau}^2}\right).$$

Обратите внимание, что мы дополнили числитель недостающей константой для полноты квадрата. Мы можем себе это позволить, так как с самого начала относимся ко всем константам по-наплевательски и надеемся восстановить

их в самом конце. В итоге видим, что апостериорное распределение снова будет нормальным с параметрами $\tilde{\mu}$ и $\tilde{\tau}^2$. Решив эту задачу, мы получили две формулы для пересчёта параметров из априорного нормально распределения в апостериорное.

Упражнение 5. В предыдущей задаче[7] мы получили формулы пересчёта для параметров нормального распределения:

$$\tilde{\mu} = \frac{\mu \cdot (\sigma^2/n)}{(\sigma^2/n) + \tau^2} + \frac{\bar{y} \cdot \tau^2}{(\sigma^2/n) + \tau^2} \qquad \tilde{\tau}^2 = \frac{\tau^2 \sigma^2/n}{(\sigma^2/n) + \tau^2}$$

Получается, что апостериорное среднее это взвешенное среднее априорного распределения и среднего по выборке, $\tilde{\mu} = w \cdot \mu + (1 - w) \cdot \bar{y}$. Найдите:

1. Предел апостериорного среднего при $\tau \to \infty$, а после при $\tau \to 0$.
2. Предел апостериорной дисперсии при $\tau \to \infty$, а потом при $\tau \to 0$.
3. Предел апостериорного среднего при $n \to \infty$, а также предел апостериорной дисперсии при $n \to \infty$
4. Сравните апостериорную дисперсию с априорной и с дисперсией $\bar{y}$.
5. Проинтерпретируйте все эти результаты. Найти-то каждый может.

Обратите внимание, что речь в этой задаче идёт про нормальное распределение. Другие распределения ведут себя похожим образом, но не так явно. Итак, пределы : 3

[7] Обычно когда в книгах пишут что-то в стиле "Подставив формулу (13) в формулу (42) мы получили формулу (73). Воспользуемся ей и получим формулу (101) возникает нужда листать страницы до формулы (13), потом до формулы (42) и так далее, и это выбешивает. Почему нельзя напечатать эти формулы ещё раз? В этой задаче мы делаем отсылку к формулам пересчёта, полученным в предыдущей задаче. И, обратите внимание, печатаем эти формулы ещё раз, чтобы читателю было удобно и он не бросал нашу книгу. Всё для комфорта любознательного читателя!

$$\lim_{\tau\to\infty}\frac{\mu\cdot(\sigma^2/n)}{(\sigma^2/n)+\tau^2}+\frac{\bar{y}\cdot\tau^2}{(\sigma^2/n)+\tau^2}=0+\lim_{\tau\to\infty}\frac{\bar{y}}{\sigma^2/\tau^2 n+1}=\bar{y}$$
$$\lim_{\tau\to 0}\frac{\mu\cdot(\sigma^2/n)}{(\sigma^2/n)+\tau^2}+\frac{\bar{y}\cdot\tau^2}{(\sigma^2/n)+\tau^2}=\mu+\lim_{\tau\to 0}\frac{\bar{y}}{\sigma^2/\tau^2 n}=\mu$$

Если наша неуверенность в априорном мнении, τ, очень велика, то априорная информация теряет своё значение и вес перед ней зануляется. Все выводы в таком случае мы делаем по выборке. И наоборот, если наша неуверенность в априорном мнении, τ, очень низка, то знакомство с данными никак не может изменить наши представления о значении m. Вес перед частью, отвечающей за выборку, зануляется.

$$\lim_{\tau\to\infty}\frac{\tau^2\sigma^2/n}{(\sigma^2/n)+\tau^2}=\lim_{\tau\to\infty}\frac{\sigma^2/n}{\sigma^2/\tau^2 n+1}=\frac{\sigma^2}{n}$$
$$\lim_{\tau\to 0}\frac{\tau^2\sigma^2/n}{(\sigma^2/n)+\tau^2}=\lim_{\tau\to 0}\frac{\sigma^2/n}{\sigma^2/\tau^2 n+1}=0$$

Если наша априорная неуверенность τ очень велика, то она теряет своё значение. Апостериорная неуверенность совпадает с дисперсией $\bar{y}$, а оценка с $\bar{y}$. И наоборот, если наша априорная неуверенность слишком мала, то наша апостериорная оценка совпадёт с μ, а дисперсия окажется нулевой.

$$\lim_{n\to\infty}\frac{\mu\cdot(\sigma^2/n)}{(\sigma^2/n)+\tau^2}+\frac{\bar{y}\cdot\tau^2}{(\sigma^2/n)+\tau^2}=\lim_{n\to\infty}\frac{\mu}{1+\frac{\tau^2}{\sigma^2/n}}+\bar{y}=\bar{y}$$
$$\lim_{n\to\infty}\frac{\tau^2\sigma^2/n}{(\sigma^2/n)+\tau^2}=\lim_{n\to\infty}\frac{\tau^2}{1+\dfrac{\tau^2}{\sigma^2/n}}=0$$

Если данных много, то они подавляют любое априорное мнение. Более того, по мере накопления данных, доверие к апостериорному распределению увеличивается и апостериорная дисперсия $\tilde{\tau}$ уменьшается.

Отдельно стоит обратить внимание на то, что апостериорная дисперсия меньше как дисперсии априорного распределения, так и дисперсии $\bar{y}$:

$$\frac{\tau^2\sigma^2/n}{(\sigma^2/n)+\tau^2} < \tau^2 \qquad \frac{\tau^2\sigma^2/n}{(\sigma^2/n)+\tau^2} < \frac{\sigma^2}{n}.$$

При этом

$$\frac{1}{\tilde{\tau}^2} = \frac{(\sigma^2/n)+\tau^2}{\tau^2\sigma^2/n} = \frac{1}{\tau^2} + \frac{1}{\sigma^2/n}.$$

«Точность» апостериорного распределения равна сумме «выборочной точности» и «априорной точности». Получается, что объединение выборочной и априорной информации приводит к увеличению «точности».

Упражнение 6. Вася сказал Кате, что подтягивается 7 раз, чтобы она впечатлилась и пошла с ним на свидание, но она не поверила и послала своего младшего брата Витю проверить сколько раз подтягивается Вася. Любой парень знает, что число подтягиваний это довольно случайная штука. Васины подтягивания распределены следующим образом:

β	4	5	6	7	8
$P(\beta = k)$	$\frac{1}{16}$	$\frac{2}{8}$	$\frac{3}{4}$	$\frac{2}{8}$	$\frac{1}{16}$

Эта табличка и есть априорное мнение о том, насколько сильный из Васьки подтягун. К сожалению Витя очень невнимательный и может ошибиться в подсчётах на одно подтягивание. К счастью, он очень ответственный и знает насколько часто он ошибается:

ε	-1	0	1
$P(\varepsilon = k)$	$\frac{3}{8}$	$\frac{2}{8}$	$\frac{3}{8}$

Итоговое число подтягиваний получается по формуле $Y = \beta + \varepsilon$.

1. Витя насчитал, что Вася подтянулся 6 раз. Найдите апостериорное распределение Васиных подтягиваний после подхода.

2. Из-за невнимательности Вити было решено на следующий день замерить результаты ещё раз. Апостериорное распределение первого дня используется в качестве априорного. Витя насчитал 5 подтягиваний. Найдите новое апостериорное распределение.

3. Предположим, что подтягуны делали подсчёты иначе. Сначала они пронаблюдали 6 и 5 подтягиваний, а после сделали байесовский пересчёт. Как будут отличаться результаты от случая последовательного байесовского вывода?

4. Удастся ли Ваське завоевать сердце Катьки?

1. Параметр имеет дискретное распределение. Данные также распределены дискретно, Нам нужно получить на выходе апостериорное распределение количества подтягиваний. Подготовим почву для применения формулы Байеса и найдём нормировочную константу по формуле полной вероятности:

$$P(y = 6) = \frac{3}{8} \cdot \frac{2}{8} + \frac{2}{8} \cdot \frac{3}{4} + \frac{3}{8} \cdot \frac{2}{8} = \frac{3}{8}$$

Теперь нужно по формуле Байсе переоценить вероятности всех априорных гипотез.

$$P(\beta = k \mid Y = 6) = \frac{P(y = 6 \mid \beta = k) \cdot P(\beta = k)}{P(y = 6)}$$

В конечном итоге получим апостериорное распределение:

β	4	5	6	7	8
$P(\beta = k)$	0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	0

Забавно, но вероятности того, что Вася подтянулся 5 или 7 раз не изменились, а априорная информация с хвостов перетекла в 6 подтягиваний.

2. Посмотрим, что случится при повторном применении формулы Байеса.

$$P(y = 5) = \frac{3}{8} \cdot \frac{1}{2} + \frac{2}{8} \cdot \frac{1}{4} + \frac{3}{8} \cdot 0 = \frac{1}{4}$$

β	4	5	6	7	8
$P(\beta = k)$	0	$\frac{1}{4}$	$\frac{3}{4}$	0	0

Вероятность ещё сильнее перетекла к 6.

3. Посмотрим, что будет, если делать пересчёт сразу же.

$$P(\beta = k \mid y_1 = 6, y_2 = 5) = \frac{P(y = 6 \mid \beta = k) \cdot P(y = 5 \mid \beta = k) \cdot P(\beta = k)}{P(y = 6, y = 5)}$$

В этот раз не будем искать нормировочную константу влоб. Восстановим её в самом конце.

$$\begin{aligned} P(\beta = 6 \mid y_1 = 6, y_2 = 5) \propto \\ \propto P(y = 6 \mid \beta = 6) \cdot P(y = 5 \mid \beta = 6) \cdot P(\beta = 6) = \\ = \left(\frac{3}{4} \cdot \frac{2}{8}\right) \cdot \left(\frac{2}{8} \cdot \frac{3}{8}\right) \cdot \frac{2}{8} = \frac{9}{2048} \end{aligned}$$

По аналогии найдём все остальные числа, пропорциональные вероятностям:

β	4	5	6	7	8
$P(\beta = k)$	0	$\frac{9}{2048}$	$\frac{12}{2048}$	0	0

Восстанавливаем нормировочную константу из уравнения:

$$\text{const} \cdot \left(\frac{9}{2048} + \frac{12}{2048}\right) = 1 \Rightarrow \text{const} = \frac{2048}{12}$$

И получаем, что итоговое распределение ничем не отличается от апостериорного распределения из предыдущего пункта. Повторим здесь ещё раз важную мысль: выход одних байесовских моделей можно использовать в качестве входа для других. При первом выводе мы сохранили в апостериорном распределении всю информацию о наших параметрах. При втором мы её дополнили. Это эквивалентно одному большому байесовскому выводу. Это свойство оказывается полезным при масштабировании байесовских моделей на большие объемы данных.

4. Судя по всему, не получится :(

Упражнение 7. В ресторане ПушкинЪ работают два повара. Первый божественно готовит окрошку, второй ризотто. Саша пришёл в ресторан, чтобы отобедать. Время готовки блюда Y имеет экспоненциальное распределение с параметром α. Повара отличаются друг от друга скоростью готовки. Для первого $\alpha = 0.1$. Для второго $\alpha = 0.2$. Работают они в ресторане день через день. Чья смена сегодня Саша не знает. К счастью, он встретил на выходе свою подругу Наташу и успел спросить у неё как долго она ждала блюдо. Выяснилось, что 10 минут. Саша хочет отведать самое вкусное блюдо текущего повара.

1. Какой из поваров готовит быстрее?

2. Какое блюдо Саше нужно заказать?

Давайте вспомним природу экспоненциального распределения. Параметр α в нём отвечает за то, насколько тяжёлый у плотности хвост. Более того, мы помним, что $E(Y) = \frac{1}{\alpha}$. Для первого повара среднее значение времени готовки составит 10 минут, для второго 20 минут.

Повара работают день через день, значит логично было бы взять в качестве априорного распределения Саши равномерное:

α	0.1	0.2
$P(\alpha = k)$	0.5	0.5

Распределение дискретное. Значит апостериорное распределение также будет дискретным. Наши данные, при этом, порождаются непрерывным распределением. Наташа сказала, что ей заказ несли 10 минут, значит функция правдоподобия будет выглядеть следующим образом:

$$f(y \mid \alpha) = \alpha e^{-\alpha \cdot 10}$$

Подготовимся к пересчёту вероятностей:

$$P(\alpha = k \mid y) = \frac{f(y \mid \alpha = k) \cdot P(\alpha = k)}{f(y)}$$
$$f(y) = P(\alpha = 0.1) \cdot f(y \mid \alpha = 0.1) + P(\alpha = 0.2) \cdot f(y \mid \alpha = 0.2)$$
$$f(y) = 0.5 \cdot 0.1e^{-0.1y} + 0.5 \cdot 0.2e^{-0.2y}$$

Сделаем пересчёт вероятностей:

$$P(\alpha = 0.1 \mid y) \propto f(y \mid \alpha = 0.1) \cdot P(\alpha = 0.1) = 0.1e^{-1} \cdot 0.5$$
$$P(\alpha = 0.2 \mid y) \propto f(y \mid \alpha = 0.2) \cdot P(\alpha = 0.2) = 0.2e^{-2} \cdot 0.5$$

Последний штрих, поделим вероятности на нормировочную константу $f(y)$, которую мы восстановили выше:

α	0.1	0.2
$P(\alpha = k)$	$\frac{0.1e^{-1}}{0.1e^{-1}+0.2e^{-2}}$	$\frac{0.2e^{-2}}{0.1e^{-1}+0.2e^{-2}}$

Или, что то же самое:

α	0.1	0.2
$P(\alpha = k)$	0.58	0.42

Напрашивается вывод: Саше нужно заказывать окрошку.

Упражнение 8. Кровожадный Джордж Мартин пишет романы из цикла «Песнь льда и пламени». Кирилл совсем недавно начал знакомиться с книгами известного писателя и пока что прочитал всего-навсего n книг, в каждой из которых погибло $y_1, \dots, y_n$ персонажей. Персонажи гибнут с некоторой интенсивностью λ, то есть $y_i \mid \lambda \sim \text{Poiss}(\lambda)$. При фиксированной интенсивности смерти персонажей в рамках одной книги не зависят от того, что написано в другой. Априорно кровожадность Джорджа Мартина описывается как $\lambda \propto \frac{1}{\lambda}$.

1. Найдите апостериорную условную функцию плотности λ с учётом прочитанных книг.

2. Найдите математическое ожидание и моду для апостериорного распределения.

3. Найдите дисперсию для апостериорного распределения. Как ведёт себя дисперсия апостериорного распределения при увеличении выборки?

4. Пусть в качестве точечной оценки Кирилл выбрал моду. Зная решение предыдущего пункта, вы бы порекомендовали ему это делать? Може быть в качестве точечной оценки лучше было бы взять математическое ожидание?

Hint: для решении задачи придётся вспомнить о том, как работают гамма-функции. Не забывайте, что $\Gamma(z) = \int_0^{+\infty} e^{-t} \cdot t^{z-1}\,dt$, а также, что $\Gamma(z+1) = z \cdot \Gamma(z)$. Также полезным может оказаться тот факт, что $\sum y_i = n\bar{y}$.

Продолжаем оттачивать технику байесовского вывода на новых упражнениях!

$$f(\lambda \mid y_1, \dots, y_n) \propto f(y_1, \dots, y_n \mid \lambda) \cdot f(\lambda) = \frac{e^{-n\cdot\lambda} \cdot \lambda^{n\cdot\bar{y}}}{y_1! \cdot \dots \cdot y_n!} \cdot \frac{1}{\lambda} \propto e^{-n\cdot\lambda} \cdot \lambda^{n\bar{y}-1}$$

Для того, чтобы дальше можно было бы использовать гамма-функции сделаем небольшой финт ушами:

$$e^{-n\cdot\lambda} \cdot \lambda^{n\bar{y}-1} = e^{-n\cdot\lambda} \cdot (n\cdot\lambda)^{n\bar{y}-1} \cdot \frac{1}{n^{n\bar{y}-1}} \propto e^{-n\cdot\lambda} \cdot (n\cdot\lambda)^{n\bar{y}-1}$$

Восстанавливаем нормировочную константу.

$$\text{const} \cdot \int_0^{+\infty} e^{-n\cdot\lambda} \cdot (n\cdot\lambda)^{n\bar{y}-1}\,d\lambda = \\ = \frac{\text{const}}{n} \int_0^{+\infty} e^{-n\cdot\lambda} \cdot (n\cdot\lambda)^{n\bar{y}-1}\,d(n\lambda) = \frac{\Gamma(n\bar{y})}{n} \cdot \text{const} = 1$$

Отсюда получаем апостериорное гамма-распределение:

$$f(\lambda \mid y_1, \dots y_n) = \frac{n}{\Gamma(n\bar{y})} e^{-n\cdot\lambda} \cdot (n\cdot\lambda)^{n\bar{y}-1}.$$

Найдём математическое ожидание:

$$E(\lambda \mid y_1, \dots, y_n) = \int_0^{+\infty} \lambda \cdot \frac{n}{\Gamma(n\bar{y})} e^{-n\cdot\lambda} \cdot (n\cdot\lambda)^{n\bar{y}-1}\, d\lambda = \\ = \frac{1}{n\Gamma(n\bar{y})} \int_0^{+\infty} e^{-n\cdot\lambda} \cdot (n\cdot\lambda)^{n\bar{y}+1-1}\, d(n\lambda) = \frac{\Gamma(n\bar{y}+1)}{n\Gamma(n\bar{y})} = \bar{y}.$$

Найдём моду:

$$e^{-n\cdot\lambda} \cdot (n\cdot\lambda)^{n\bar{y}-1} \longrightarrow \max_{\lambda}$$

$$-n\cdot e^{-n\cdot\lambda} \cdot (\lambda n)^{n\bar{y}-1} + e^{-n\cdot\lambda} n(n\bar{y}-1)(\lambda n)^{n\bar{y}-2} = \\ = e^{-n\cdot\lambda}(\lambda n)^{n\bar{y}-2}(-n\lambda n + n(n\bar{y}-1)) = 0$$

$$\mathrm{Mod}(\lambda \mid y_1, \dots y_n) = \bar{y} - \frac{1}{n}.$$

Пришёл черёд дисперсии:

$$E(\lambda^2 \mid y_1, \dots, y_n) = \int_0^{+\infty} \lambda^2 \cdot \frac{n}{\Gamma(n\bar{y})} e^{-n\cdot\lambda} \cdot (n\cdot\lambda)^{n\bar{y}-1}\, d\lambda = \\ = \frac{1}{n^2\Gamma(n\bar{y})} \int_0^{+\infty} e^{-n\cdot\lambda} \cdot (n\cdot\lambda)^{n\bar{y}+2-1}\, d(n\lambda) = \frac{\Gamma(n\bar{y}+2)}{n^2\Gamma(n\bar{y})}.$$

Раскрываем гамма-функцию:

$$\frac{\Gamma(n\bar{y}+2)}{n^2\Gamma(n\bar{y})} = \frac{(n\bar{y}+1)\cdot\Gamma(n\bar{y}+1)}{n^2\Gamma(n\bar{y})} = \\ = \frac{(n\bar{y}+1)\cdot n\bar{y}\cdot\Gamma(n\bar{y})}{n^2\Gamma(n\bar{y})} = \frac{(n\bar{y}+1)\cdot n\bar{y}}{n^2} = \bar{y}^2 + \frac{\bar{y}}{n}.$$

Последний штрих:

$$\mathrm{Var}(\lambda \mid y_1, \dots, y_n) = \mathrm{E}(\lambda^2 \mid y_1, \dots, y_n) - (\mathrm{E}(\lambda \mid y_1, \dots, y_n))^2 = \frac{\bar{y}}{n}.$$

Как это не странно, при увеличении выборки, дисперсия апостериорного распределения уменьшается. Наши представления о парметре λ становятся всё более и более конкретными. При неограниченном увеличении выборки апостериорное распределение становится вырожденным. Дисперсия случайной величины сходится к нулю

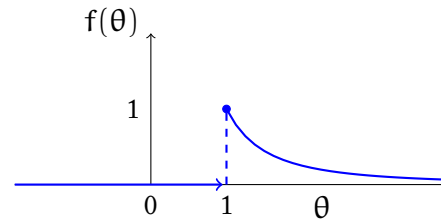
$$\operatorname*{plim}_{n\to\infty} \frac{\bar{y}}{n} = \operatorname*{plim}_{n\to\infty} \frac{\mathrm{E}(y_i \mid \lambda)}{n} = \operatorname*{plim}_{n\to\infty} \frac{\lambda}{n} = 0.$$

С вероятностью единица (почти наверное) случайная величина λ принимает своё истиное значение.

На малых выборках мода апостериорного распределения будет давать смещённую точечную оценку. В конечном счёте при увеличении выборки обе оценки сходятся к истиному значению параметра. Наше апостериорное распределение описывает то насколько сильно мы не знаем что происходит с параметром λ после анализа выборки. При бесконечно большой выборке, мы знаем о параметре λ абсолютно всё.

Упражнение 9. В городе N есть одна железная дорога. Эконометресса Анна Каренина каждый день в случайный момент времени приходит на вокзал города N и ждёт поезд. Время, которое Аня прождала поезд, она фиксирует в своём блокноте. Известно, что поезда ходят с интервалом в θ минут, так что время ожидания поезда пассажиром можно считать случайной величиной, имеющей равномерное распределение, $y_i \sim \mathcal{U}[0; \theta]$. В блокноте Ани уже есть записи $y_1, \dots, y_n$. Аня думает, что задержка между поездами, θ не может быть очень большой. Более того, Ане кажется, что параметр θ имеет распределение Парето с плотностью:

$$f(\theta) = \begin{cases} \frac{1}{\theta^2} & , \theta \geqslant 1 \\ 0 & , \text{иначе} \end{cases}$$



Это означает, что Аня не верит в большие задержки. При этом, её вера убывает по степенному закону[8].

1. Существует ли в данной задаче оценка максимального правдоподобия?
2. Найдите апостериорное распределение для параметра θ.
3. Найдите математическое ожидание и моду апостериорного распределения. Используйте все три точечные оценки для ответа на следующие вопросы.
4. Аня пришла на вокзал и ждала поезд 5 минут. Как Аня оценивает время ожидания поезда? На второй день Аня прождала поезд 2 минуты. Как изменились Анины оценки? На третий день Аня ждала поезд 10 минут. Что произошло с её оценками? Какая из точечных оценок наиболее адекватно, на ваш взгляд, себя ведёт? Почему? Что при поступлении каждого нового наблюдения происходит с апостериорным распределением Ани?
5. Нужно ли Ане срочно бросить заниматься дурью на вокзале и уйти домой, в поместье?

Найдём оценку максимального правдоподобия.

$$f(y_1, \dots, y_n \mid \theta) = \begin{cases} \frac{1}{\theta^n}, 0 \leqslant y_{max} \leqslant \theta \\ 0, \text{ иначе} \end{cases} \quad \to \max_\theta .$$

Взятие производной по θ даст нам уравнение, которое нельзя решить относительно θ. Мы имеем дело с нерегулярным случаем. Из-за того, что область

[8] Нассим Талеб в своей книге «Чёрный Лебедь» очень любит распределение Парето.

определения случайной величины зависит от значения параметра, не выполняются условия регулярности и оценку максимального правдоподобия, в привычном для нас смысле, найти нельзя. Однако, мы можем оценить параметр на основе ограничений из выборки. Логично, что $y_1 < \theta, \dots, y_n < \theta$. Так как функция $f(y \mid \theta)$ убывает по θ, возьмём самое маленькое возможное значение этого параметра и получим, что $\theta^{ML} = y_{max}$.

Найдём для нашего параметра апостериорное распределение. Будем помнить о том, что область определения зависит от значения параметра.

$$f(\theta \mid y) \propto f(y \mid \theta) \cdot f(\theta) = \begin{cases} \frac{1}{\theta^n} \cdot \frac{1}{\theta^2}, \theta \geqslant \max(y_{max}, 1) \\ 0, \text{ иначе} \end{cases}$$

Восстанавливаем константу:

$$\text{const} \cdot \int_{\max(y_{max},1)}^{+\infty} \frac{1}{\theta^{n+2}} \, d\theta = 1 \Rightarrow \text{const} = (n+1) \cdot (\max(y_{max}, 1))^{n+1}$$

На выходе получаем распределение Парето. Выглядит оно не очень удобно. Судя по форме этого распределения, его максимум достигается в самой левой точке. Значит мода будет находиться в точке $\max(y_{max}, 1)$. Честное взятие интеграла даст нам математическое ожидание.

Наблюдения	Математическое ожидание	Мода
$y_1, \dots, y_n$	$(1 + \frac{1}{n}) \cdot \max(y_{max}, 1)$	$\max(y_{max}, 1)$
$y_1 = 5$	10	5
$y_1 = 5, y_2 = 2$	7.5	5
$y_1 = 5, y_2 = 2, y_3 = 10$	13.3	10

Надо как-то переделать упражнение, чтобы появилась мораль

Упражнение 10. Несколько упражнений на сопряжённые распределения и формулы пересчёта.

1. $Y \sim Bern(p)$, $p \sim B(\alpha, \beta)$. Покажите, что апостериорное распределение также будет B-распределением и выведите формулы пересчёта для параметров апостериорного распределения $\tilde{\alpha}$ и $\tilde{\beta}$.

2. $Y \sim \mathrm{Binom}(p, n), p \sim B(\alpha, \beta)$. Покажите, что апостериорное распределение также будет B-распределением и выведите формулы пересчёта для параметров апостериорного распределения $\tilde{\alpha}$ и $\tilde{\beta}$.

3. $Y \sim \mathrm{Geom}(p)$, $p \sim B(\alpha, \beta)$. Покажите, что апостериорное распределение также будет B-распределением и выведите формулы пересчёта для параметров апостериорного распределения $\tilde{\alpha}$ и $\tilde{\beta}$.

4. $Y \sim \mathrm{Poiss}(\lambda)$, $\lambda \sim \Gamma(s, r)$. Покажите, что апостериорное распределение также будет Γ-распределением и выведите формулы пересчёта для параметров апостериорного распределения $\tilde{s}$ и $\tilde{r}$.

5. $Y \sim \mathrm{Exp}(\alpha)$, $\alpha \sim \Gamma(s, r)$. Покажите, что апостериорное распределение также будет Γ-распределением и выведите формулы пересчёта для параметров апостериорного распределения $\tilde{s}$ и $\tilde{r}$.

Написать решения (или послать читателя в Айвазяна, что выгоднее)

Упражнение 11. Пусть $f(x)$ — плотность распределения случайной величины X. Исследователь Вова зафиксировал уровень значимости α и нашёл самый коротки доверительный интервал $[a; b]$. Известно, что у случайной величины X только одна мода. Правда ли, что для этого интервала будет выполнено равенство $f(a) = f(b)$?

Условие про одну моду лишнее или нет? Все условия учёл?

Разберёмся с этой задачкой двумя способами. Первый способ решения будет графическим. Пусть $f(b) > f(a)$, тогда если мы сдвинем значение b вправо на ε, мы захватим розовый кусочек площади.

Сделать картинку

Для того, чтобы компенсировать вновь приобретённую площадь, значение a также нужно сдвинуть вправо на значение δ. Из-за того, что $f(b) > f(a)$, мы должны будем сдвинуть a на большую величину, чем ε, чтобы компенсировать вновь появившуюся площадь. Таким образом, из-за того, что $\delta > \varepsilon$, при каждом новом сдвиге a и b будут сближаться друг с другом до тех пор, пока значения $f(a)$ и $f(b)$ не сравняются и мы не получим самый короткий доверительный интервал.

Нормально описание или кривовато? Как исправить?

Второй способ решения будет алгебраическим. Он будет чуть более сложным. Для того, чтобы найти самый короткий доверительный интервал, мы должны решить задачку вида:

$$\begin{cases} b - a \longrightarrow \min\limits_{a,b} \\ \int\limits_a^b f(x)\,dx = 1 - \alpha \end{cases}$$

Вспомним, что $\int_a^b f(x)\,dx = F(b) - F(a)$ и получим, что $b = F^{-1}(\alpha + F(a))$. Подставим полученное ограничение в минимизируемую функцию

$$F^{-1}(\alpha + F(a)) - a \to \min_a.$$

Осталась техническая часть. Берём производную обратной функции и приравниваем её к нулю. При этом, не забываем, что $y'_x = \frac{1}{x'_y}$.

За x в нашей формуле обозначается $\alpha + F(a)$, за y обозначается $F^{-1}(\alpha + F(a))$. Значит обратной функцией будет $F(F^{-1}(\alpha + F(a)))$. Получаем следующее уравнение

$$\frac{1}{F'(F^{-1}(\underbrace{\alpha + F(a)}_{F(b)}))} \cdot f(a) - 1 = 0.$$

В итоге получаем, что

$$\frac{1}{F'(b)} \cdot f(a) = 1 \quad \Rightarrow \quad f(a) = f(b).$$

Обратите внимание, что при бимодальном распределении самый короткий доверительный интервал может оказаться рваным.

Картинка