







Байесовские методы. Лекция 7. Иерархические модели.

Целищев М.А.

МГУ им. М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математической статистики

весна 2021

Список литературы

-  J. Kruschke. Doing Bayesian Data Analysis, Second Edition. A Tutorial with R, JAGS, and Stan. Academic Press, 2014.
-  R. McElreath. Statistical Rethinking. A Bayesian Course with Examples in R and Stan, Second Edition. Chapman and Hall CRC, 2020.
-  O. Martin. Bayesian Analysis with Python. Introduction to Statistical Modeling and Probabilistic Programming using PyMC3 and ArviZ, Second Edition. Packt, 2018.
-  K. P. Murphy. Machine Learning: A Probabilistic Perspective. The MIT Press, 2012.
-  A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, D. Rubin. Bayesian Data Analysis, Third Edition. CRC Press, 2013.
-  S. Brooks, A. Gelman, G. L. Jones and X.-L. Meng. Handbook of Markov Chain Monte Carlo. Chapman & Hall/CRC, 2011.

Модель Бернулли

Вернёмся к тому, с чего начали курс лекций. Пусть в некоторой стране скончалось d человек из n заражённых вирусом. Нужно оценить летальность вируса.

Классическая статистика: оценка макс. правдоподобия $\theta_{\text{ML}} = \frac{d}{n}$.

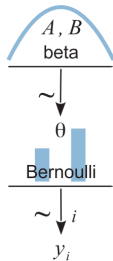
Байесова статистика:

Правдоподобие: $(y_i | \theta) \sim \text{Be}(\theta), \quad i = 1, \dots, n$

Априорное: $\theta \sim \text{Beta}(\alpha, \beta)$

Тогда

$$\begin{aligned} (\theta | Y) &\sim \text{Beta}(\alpha + \sum_i y_i, \beta + n - \sum_i y_i) = \\ &= \text{Beta}(\alpha + d, \beta + n - d). \end{aligned}$$



Если же по каким-то причинам Beta-распределение не подходит в качестве априорного, то можно взять и любое другое, и пользоваться сэмплами из апостериорного распределения, известного с точностью до нормировочной константы.

А что, если известны данные для двух стран? n_1, n_2 — количество инфицированных, d_1, d_2 — количество жертв. Как оценить летальность вируса?

- Можно рассматривать каждую страну по отдельности.

Минус: если у одной из стран малый размер выборки, то оценка летальности для неё получится менее точной

- А можно объединить данные: считать, что из $n_1 + n_2$ заражённых скончалось $d_1 + d_2$ человек и, тем самым, свести задачу к предыдущему слайду.

Плюс: получили большую по размеру выборку

Минус: есть основания полагать, что у разных стран летальность может быть разная (из-за разных условий и разных «методик подсчёта» числа инфицированных и числа умерших)

А можно ли сделать нечто среднее из этих двух подходов?

Байесовская модель для нескольких классов

Пусть y_{ij} — индикатор того, что i -ый инфицированный в j -ой стране скончался, $i = 1, \dots, n_j$, $j = 1, 2$.

Опишем модель данных (правдоподобие) по-байесовски:

$$(y_{ij}|\theta_j) \sim \text{Be}(\theta_j), \quad i = 1, \dots, n_j, \quad j = 1, 2.$$

Не хватает априорного распределения вектора (θ_1, θ_2) .

- Если брать θ_1 и θ_2 независимыми с некоторыми маргинальными распределениями $p(\theta_1)$ и $p(\theta_2)$, то есть $p(\theta_1, \theta_2) = p(\theta_1)p(\theta_2)$, то приходим к апостериорному

$$p(\theta_1, \theta_2|Y) \propto p(Y|\theta_1, \theta_2)p(\theta_1, \theta_2) = p(Y_1|\theta_1)p(\theta_1) \cdot p(Y_2|\theta_2)p(\theta_2),$$

где Y_j — вектор индикаторов летальных исходов j -ой страны (которые для разных стран *условно независимы при фиксированных* θ_1, θ_2).

Итого получается, что апостериорное распределение факторизуется по странам, то есть задача решается для каждой страны по отдельности.

Байесовская модель для нескольких классов

Пусть y_{ij} — индикатор того, что i -ый инфицированный в j -ой стране скончался, $i = 1, \dots, n_j$, $j = 1, 2$.

Опишем модель данных (правдоподобие) по-байесовски:

$$(y_{ij}|\theta_j) \sim \text{Be}(\theta_j), \quad i = 1, \dots, n_j, \quad j = 1, 2.$$

Не хватает априорного распределения вектора (θ_1, θ_2) .

- Если же считать $\theta_1 = \theta_2 = \theta$ и дать ему некоторое маргинальное распределение $p(\theta)$, то приходим к апостериорному

$$p(\theta|Y) \propto p(Y|\theta)p(\theta) = p(Y_1|\theta)p(Y_2|\theta)p(\theta).$$

Итого получается, что данные по странам Y_j объединяются, то есть оценивается «общая летальность».

Байесовская модель для нескольких классов

Пусть y_{ij} — индикатор того, что i -ый инфицированный в j -ой стране скончался, $i = 1, \dots, n_j$, $j = 1, 2$.

Опишем модель данных (правдоподобие) по-байесовски:

$$(y_{ij}|\theta_j) \sim \text{Be}(\theta_j), \quad i = 1, \dots, n_j, \quad j = 1, 2.$$

Не хватает априорного распределения вектора (θ_1, θ_2) .

- А что, если не впадать в крайности независимости/точного совпадения параметров θ_1 и θ_2 ?

Какое же априорное распределение $p(\theta_1, \theta_2)$ использовать?

Как минимум, имеет смысл просить, чтобы маргинальные распределения $p(\theta_1)$ и $p(\theta_2)$ совпадали.

Но, как мы помним, совместное распределение $p(\theta_1, \theta_2)$ не определяется однозначно своими маргинальными распределениями $p(\theta_1)$ и $p(\theta_2)$.

Оказывается, что хорошим вариантом будет брать вектор (θ_1, θ_2) *симметрично зависимым*.

Definition

Случайные величины ξ_1, \dots, ξ_d *симметрично зависимы* (или *перестановочны*), если распределение случайного вектора (ξ_1, \dots, ξ_d) остаётся неизменным при произвольной перестановке его элементов, то есть

$$(\xi_{i_1}, \dots, \xi_{i_d}) \stackrel{d}{=} (\xi_1, \dots, \xi_d)$$

для любой перестановки i_1, \dots, i_d чисел $1, \dots, d$.

Замечание. Ясно, что если $\xi_1, \xi_2, \dots, \xi_d$ независимы и одинаково распределены (i.i.d.), то они симметрично зависимы.

Также очевидно, что если ξ_1, \dots, ξ_d симметрично зависимы, то их маргинальные распределения совпадают, то есть

$$p_{\xi_1}(x) = \dots = p_{\xi_d}(x) \quad \forall x \in \mathbb{R}.$$

Однако же симметрично зависимые случайные величины совсем не обязаны быть независимыми!

Definition

Случайные величины ξ_1, \dots, ξ_d *симметрично зависимы* (или *перестановочны*), если распределение случайного вектора (ξ_1, \dots, ξ_d) остаётся неизменным при произвольной перестановке его элементов, то есть

$$(\xi_{i_1}, \dots, \xi_{i_d}) \stackrel{d}{=} (\xi_1, \dots, \xi_d)$$

для любой перестановки i_1, \dots, i_d чисел $1, \dots, d$.

Примеры симметрично зависимых случайных величин:

- Пусть в урне n белых и m чёрных шаров. Шары тянут один за другим без возвращения, и пусть X_i — индикатор того, что i -ый шар белый. Тогда X_1, X_2, \dots, X_{n+m} , конечно же, зависимы, но зависимы симметрично, поскольку

$$\mathbf{P}(X_1 = x_1, \dots, X_{n+m} = x_{n+m}) = \frac{n! m!}{(n+m)!} = \frac{1}{C_{n+m}^n}$$

при $x_i \in \{0, 1\}$ и $\sum x_i = n$.

Definition

Случайные величины ξ_1, \dots, ξ_d *симметрично зависимы* (или *перестановочны*), если распределение случайного вектора (ξ_1, \dots, ξ_d) остаётся неизменным при произвольной перестановке его элементов, то есть

$$(\xi_{i_1}, \dots, \xi_{i_d}) \stackrel{d}{=} (\xi_1, \dots, \xi_d)$$

для любой перестановки i_1, \dots, i_d чисел $1, \dots, d$.

Примеры симметрично зависимых случайных величин:

- Пусть $X \sim \mathcal{N}(\mathbf{a}, \Sigma)$ с вектором средних $\mathbf{a} = (a, \dots, a) \in \mathbb{R}^d$ и ковариационной матрицей $\Sigma \in \mathbb{R}^{d \times d}$, у которой совпадают все диагональные элементы $\sigma_{ii} = \sigma^2$, а также совпадают все недиагональные элементы $\sigma_{ij} = \rho$, $i \neq j$.

Definition

Случайные величины ξ_1, \dots, ξ_d *симметрично зависимы* (или *перестановочны*), если распределение случайного вектора (ξ_1, \dots, ξ_d) остаётся неизменным при произвольной перестановке его элементов, то есть

$$(\xi_{i_1}, \dots, \xi_{i_d}) \stackrel{d}{=} (\xi_1, \dots, \xi_d)$$

для любой перестановки i_1, \dots, i_d чисел $1, \dots, d$.

Пример с.в., не являющихся симметрично зависимыми:

- Ежедневные цены закрытия контрактов по нефти (или любых других финансовых инструментов). Для них порядок наблюдений имеет значение: низкие значения сегодня дают большую уверенность в том, что завтрашние значения будут не сильно выше.

Условная независимость

Definition

Говорят, что случайные величины X_1, \dots, X_d с совместной плотностью $p(x_1, \dots, x_d)$ *условно независимы при фиксированном ϕ* , если

$$p(x_1, \dots, x_d | \phi) = \prod_{i=1}^n p(x_i | \phi).$$

Совместная плотность X_1, \dots, X_d при этом записывается в виде:

$$p(x_1, \dots, x_d) = \int p(x_1, \dots, x_d | \phi) p(\phi) d\phi = \int \left(\prod_{i=1}^d p(x_i | \phi) \right) p(\phi) d\phi,$$

где $p(\phi)$ — плотность ϕ . Иными словами, условно независимые случайные величины — это смесь распределений с независимыми компонентами со *смешивающим распределением* $p(\phi)$.

Вообще-то понятие условной независимости уже встречалось в курсе: условно независимыми были исходы подбрасывания монетки, при фиксированной вероятности выпадения решки.

Lemma

Если X_1, \dots, X_d условно независимы при фиксированном ϕ и все условные маргинальные распределения $p(x_i|\phi)$ совпадают, то X_1, \dots, X_d симметрично зависимы.

Док-во. Для любой перестановки i_1, \dots, i_d чисел $1, \dots, d$ имеем

$$p(x_1, \dots, x_d) = \int \left(\prod_{i=1}^d p(x_i|\phi) \right) p(\phi) d\phi = p(x_{i_1}, \dots, x_{i_d}). \quad \square$$

Для бесконечных последовательностей с.в. справедлив и обратный результат:

Theorem (де Финетти, 1931, в оригинале для бернуллиевских с.в.)

Пусть X_1, X_2, \dots — последовательность симметрично зависимых с.в. Тогда найдётся такая с.в. ϕ , что X_1, X_2, \dots условно независимы при фиксированном ϕ , то есть $p(x_1, \dots, x_d) = \int \left(\prod_{i=1}^d p(x_i|\phi) \right) p(\phi) d\phi \quad \forall d \in \mathbb{N}$.

ДЗ(*): показать, что для конечных последовательностей с.в. этот результат не является справедливым.

Так что с вирусом?

Вернёмся к задаче оценки летальности вируса для двух стран.

Правдоподобие:

$$(y_{ij}|\theta_j) \sim \text{Be}(\theta_j), \quad i = 1, \dots, n_j, \quad j = 1, 2.$$

Так как всё же задать априорное распределение вектора (θ_1, θ_2) ? Поскольку нет априорных предпочтений между странами, то лучший вариант — симметричная зависимость, а по теореме де Финетти это (почти) тоже самое, что условная независимость при фиксированном ϕ . Кто такой ϕ ?

Параметры θ_1 и θ_2 при таком подходе рассматриваются как i.i.d. элементы выборки из некоторого распределения $p(\theta|\phi)$ *популяции* возможных летальностей для разных стран, а ϕ — неизвестный *гиперпараметр* этого распределения, который, как всегда в байесовском подходе, имеет своё априорное распределение $p(\phi)$.

Иерархическая модель

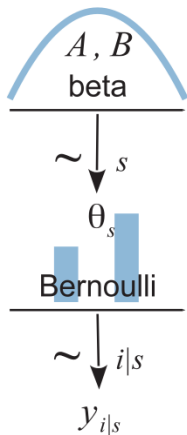
Каким взять $p(\theta_j|\phi)$ и что такое ϕ ?

Поскольку θ_j (летальность вируса в j -ой стране) принимает значения из $[0, 1]$, то разумным выбором будет бета-распределение: $\theta_j|A, B \sim \text{Beta}(A, B)$, и тогда $\phi = (A, B)$ — гиперпараметры.

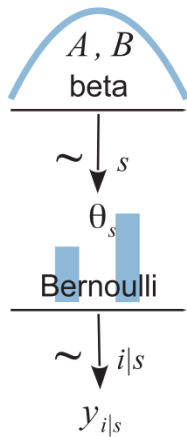
Тогда *иерархическая* модель будет выглядеть так:

$$\begin{cases} y_{ij}|\theta_j \sim \text{Be}(\theta_j), & i = 1, \dots, n_j, \quad j = 1, 2, \\ \theta_j|A, B \sim \text{Beta}(A, B), & j = 1, 2, \\ A, B \sim p(A, B) \end{cases}$$

При этом все y_{ij} считаются условно независимыми при фиксированных θ_1, θ_2 , а θ_1, θ_2 условно независимы при фиксированных A, B .



Иерархическая модель



$$\begin{cases} y_{ij}|\theta_j \sim \text{Be}(\theta_j), & i = 1, \dots, n_j, \quad j = 1, 2, \\ \theta_j|A, B \sim \text{Beta}(A, B), & j = 1, 2, \\ A, B \sim p(A, B) \end{cases}$$

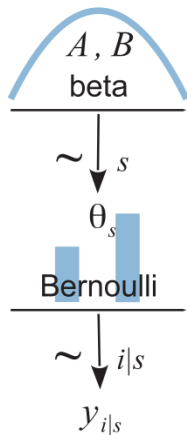
Задача байесовского вывода — найти апостериорное распределение параметров при известных наблюдениях Y , то есть

$$\begin{aligned} p(\theta_1, \theta_2, A, B | Y) &\propto p(Y | \theta_1, \theta_2, A, B) p(\theta_1, \theta_2, A, B) = \\ &= p(Y_1|\theta_1)p(Y_2|\theta_2) \cdot p(\theta_1|A, B)p(\theta_2|A, B) \cdot p(A, B). \end{aligned}$$

Если A, B брать известными (т.е. с вырожденным распределением), то апостериорное распределение факторизуется (отдельно на каждую страну).

Разумнее взять их невырожденными. А с каким распределением?

Иерархическая модель



$$\begin{cases} y_{ij}|\theta_j \sim \text{Be}(\theta_j), & i = 1, \dots, n_j, \quad j = 1, 2, \\ \theta_j|A, B \sim \text{Beta}(A, B), & j = 1, 2, \\ A, B \sim p(A, B) \end{cases}$$

$$p(\theta_1, \theta_2, A, B | Y) \propto \\ \propto p(Y_1|\theta_1)p(Y_2|\theta_2) \cdot p(\theta_1|A, B) \cdot p(\theta_2|A, B)p(A, B).$$

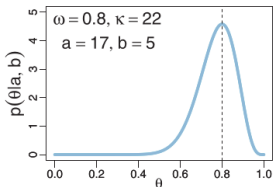
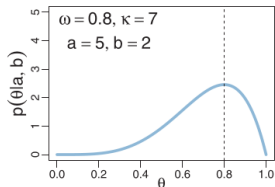
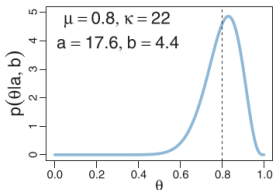
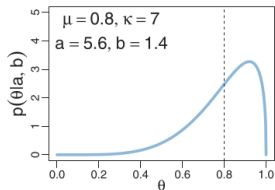
Поскольку $A, B > 0$, то можно для них брать любое распределение, сосредоточенное на $(0, +\infty)$, например, гамма-распределение Γ , или обратное-гамма распределение InvGamma .

А совместное для A и B каким брать априорное распределение? Они независимы? Почему?

Перепараметризация

Вспомним, что при значениях $A > B \geq 1$ распределение $\text{Beta}(A, B)$ с плотностью $p(\theta|A, B) \propto \theta^{A-1}(1-\theta)^{B-1}$ «сосредоточено правее» на $[0, 1]$, и чем больше A и B , тем это распределение «уже» (больше напоминает вырожденное).

Лучше использовать иную параметризацию бета-распределения, т.к. физический смысл гиперпараметров A и B «затуманен».



Проверить (дз!):

$$\text{Среднее: } \mu = \frac{A}{A+B}$$

$$\text{Мода: } \omega = \frac{A-1}{A+B-2}$$

Концентрация:

$$\kappa := A+B$$

Дисперсия:

$$\sigma^2 = \frac{\mu(1-\mu)}{\kappa+1}$$

Тогда давайте возьмём в качестве двух параметров вместо A и B такие, которые бы имели физический смысл, например характеризовали точку максимума (моду) $\omega = \frac{A-1}{A+B-2}$ и степень концентрации $\kappa = A + B$ параметров θ_j .

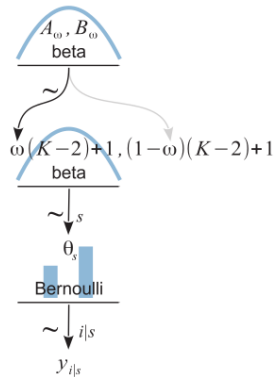
При этом (дз!) $A = \omega(\kappa - 2) + 1$, $B = (1 - \omega)(\kappa - 2) + 1$.

Можно интерпретировать $\omega \in (0, 1)$ как точку, вокруг которой априорно сосредоточены летальности разных стран, а $\kappa > 2$ — как априорную степень концентрации летальностей вокруг ω .

Можно взять ω и κ априорно независимыми и даже считать $\kappa = K$ фиксированным, а $\omega \sim \text{Beta}(A_\omega, B_\omega)$.

Вся модель:

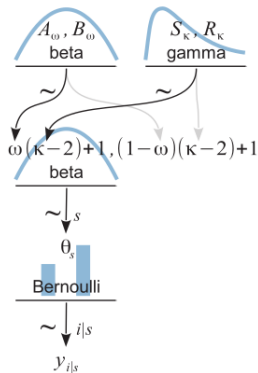
$$\begin{cases} y_{ij} | \theta_j \sim \text{Be}(\theta_j), & i = 1, \dots, n_j, \quad j = 1, 2, \\ \theta_j | \omega, \kappa \sim \text{Beta}(\omega(\kappa - 2) + 1, (1 - \omega)(\kappa - 2) + 1), & j = 1, 2, \\ \omega \sim \text{Beta}(A_\omega, B_\omega) \\ \kappa = K \end{cases}$$



Нужно больше иерархии

А можно взять невырожденное априорное распределение и на гиперпараметр $\kappa > 2$:

$$\begin{cases} y_{ij} | \theta_j \sim \text{Be}(\theta_j), & i = 1, \dots, n_j, \quad j = 1, 2, \\ \theta_j | \omega, \kappa \sim \text{Beta}(\omega(\kappa - 2) + 1, (1 - \omega)(\kappa - 2) + 1), & j = 1, 2, \\ \omega \sim \text{Beta}(A_\omega, B_\omega) \\ \kappa - 2 \sim \Gamma(S_\kappa, R_\kappa) \end{cases}$$



Дальше можно «навесить» априорное распределение и на гипер-гиперпараметры $A_\omega, B_\omega, S_\kappa, R_\kappa$, но зачем? Если не получается объяснить зачем, то и не нужно. В таком случае следует брать их постоянными, но так, чтобы распределения на верхнем уровне иерархии были не слишком «узкими» (если априорно нет оснований в этом сомневаться).

Зачем всё это?

Итак, можно посчитать апостериорное распределение всех параметров (ну или хотя бы семплировать из него):

$$p(\theta_1, \theta_2, \omega, \kappa | Y) \propto \text{Likelihood} \times \text{Prior} = \left(\prod_j p(Y_j | \theta_j) \right) \cdot \left(\prod_j p(\theta_j | \omega, \kappa) \right) p(\omega) p(\kappa)$$

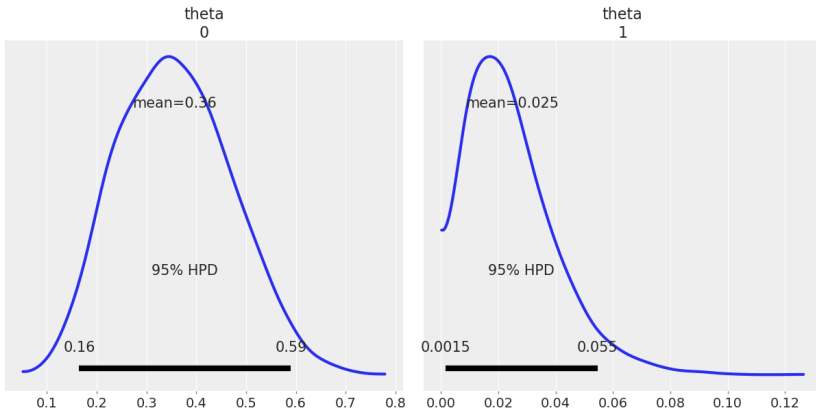
А зачем всё это нужно? Чем это поможет?

Да много чем! Можно исследовать:

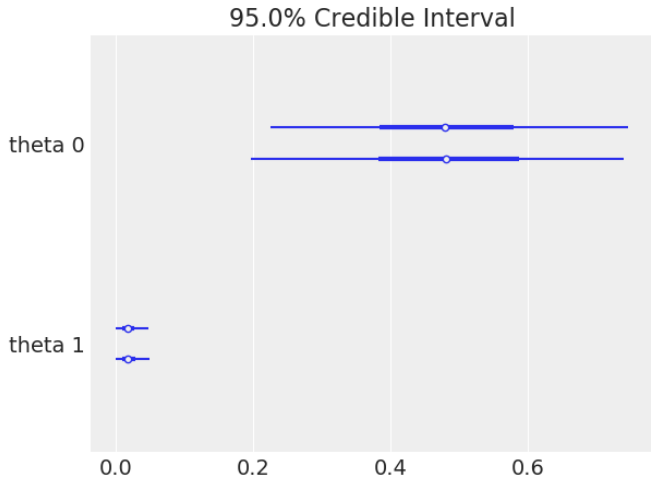
- апостериорное распределение летальности θ_j для каждой страны (которое зависит от всего датасета Y), в том числе делать прогнозы для новых наблюдений из этой страны,
- разницу между летальностями для разных стран,
- апостериорное распределение гиперпараметров (ω, κ) и прогнозировать летальность для стран, у которых нет наблюдений.

Апостериорные распределения θ_j

Нагенерировав достаточно семплов из $\theta_j|Y$, можно построить соответствующие гистограммы и сгладить их:

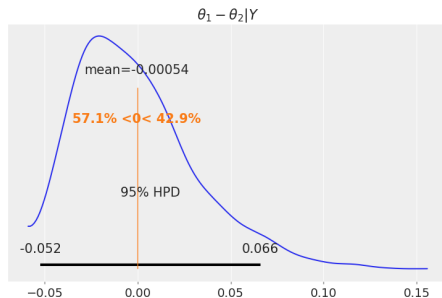
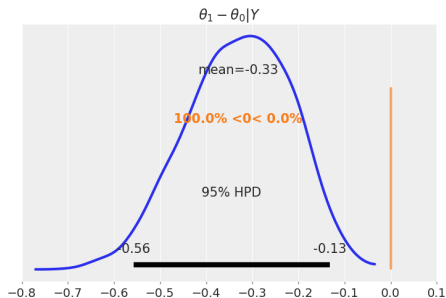


Апостериорные распределения θ_j



Апостериорное распределение $\theta_1 - \theta_2$

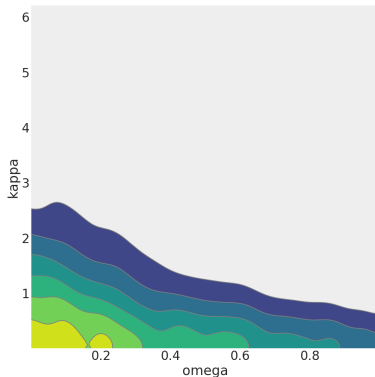
А можно тем же образом исследовать апостериорное распределение разности летальностей для двух стран:



Как это сделать? Очень просто: вычесть нагенерированные семплы для θ_2 из семплов θ_1 , это и будет выборка из $\theta_1 - \theta_2$.

Апостериорное распределение гиперпараметров

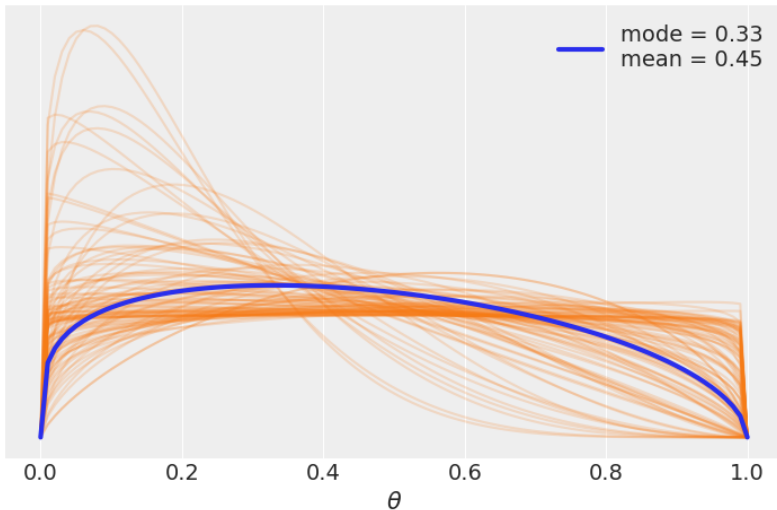
Можно оценить апостериорное распределение $\omega, \kappa|Y$:



Зачем оно может понадобиться? С его помощью можно генерировать летальности для стран, которых ещё нет в датасете Y .

Апостериорное распределение гиперпараметров

Возможные реализации $p(\theta|\omega, \kappa)$ для насемплированных ω, κ :

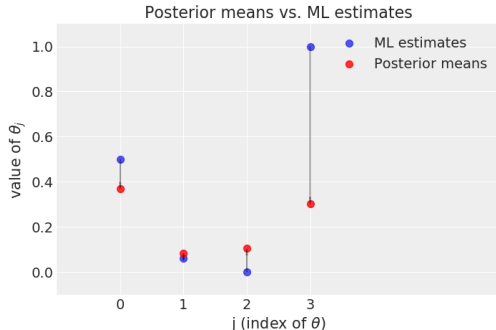


Shrinkage

Можно сравнить среднее апостериорных распределений $\theta_j|Y$ с оценками максимального правдоподобия $\theta_j^{\text{ML}} = \frac{d_j}{n_j}$.

Если гиперпараметры невырождены, то среднее апостериорных распределений будет ближе к «общему центру», чем оценки ML.

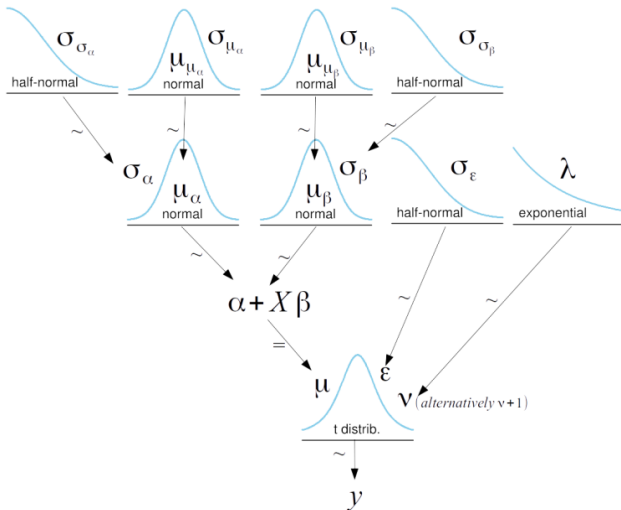
Этот эффект называется сжатием (*shrinkage*) и демонстрирует, что байесовы оценки используют информацию всего датасета для оценки летальности одной группы.



При этом чем более «сжаты» априорные распределения гиперпараметров, тем сильнее наблюдается этот эффект.

Иерархические модели

Иерархические модели применяют и для линейной регрессии:



Здесь:

y — наблюдения

X — предикторы