Московский государственный университет имени М.В. Ломоносова

Факультет вычислительной математики и кибернетики

Отчёт по первому практическому заданию: Байесовские рассуждения

Номер варианта: **[3]**

Каратыщев Дмитрий Иванович, 417 группа

Дата выполнения работы:
2 октября 2024 г.

Содержание

1 Постановка задачи

Рассматривается модель посещаемости студентами ВУЗа нескольких лекций по курсу. Аудитория данного курса состоит из студентов профильного факультета, а также студентов других факультетов. Через a обозначено количество студентов, поступивших на профильный факультет, а через b – количество студентов других факультетов. Студенты профильного факультета посещают лекцию с некоторой вероятностью p_1, а студенты остальных факультетов – с вероятностью p_2. Через c обозначено количество студентов, посетивших данную лекцию. На лекции по курсу ведётся запись студентов. При этом каждый студент записывается сам, а также, быть может, записывает своего товарища, которого на лекции на самом деле нет. Студент записывает своего товарища с некоторой вероятностью p_3. Через d обозначено общее количество записавшихся на данной лекции. Величины a и b имеют дискретное равномерное распределение на интервалах $[a_{min}, a_{max}]$ и $[b_{min}, b_{max}]$ соответственно.

Имеются две вероятностные модели, которые необходимо исследовать. Первая модель выглядит следующим образом:

$$p(a, b, c_1, \ldots, c_N, d_1, \ldots, d_N) = p(a)p(b)\prod_{n=1}^{N} p(d_n|c_n)p(c_n|a, b),$$

$$d_n|c_n \sim c_n + Bin(c_n, p_3),$$

$$c_n|a, b \sim Bin(a, p_1) + Bin(b, p_2),$$

$$a \sim \mathcal{U}[a_{min}, a_{min}]$$

$$b \sim \mathcal{U}[b_{min}, b_{min}]$$

Вторая модель является упрощением первой: меняется распределение на $c_n|a, b$:

$$p(a, b, c_1, \ldots, c_N, d_1, \ldots, d_N) = p(a)p(b)\prod_{n=1}^{N} p(d_n|c_n)p(c_n|a, b),$$

$$d_n|c_n \sim c_n + Bin(c_n, p_3),$$

$$c_n|a, b \sim Pois(ap_1 + bp_2)$$

$$a \sim \mathcal{U}[a_{min}, a_{min}]$$

$$b \sim \mathcal{U}[b_{min}, b_{min}]$$

В 3-м варианте параметры поставленной задачи выглядят так: $a_{min} = 75,\ a_{max} = 90,\ b_{min} = 500,\ b_{max} = 600,\ p_1 = 0.1,\ p_2 = 0.01,\ p_3 = 0.3,\ N = 50$. При этом рекомендуется брать следующие диапазоны значений: для c_n - $[0, a_{max} + b_{max}]$, а для d_n - это $[0, 2(a_{max} + b_{max})]$

Исследование подразумевает рассмотрение следующих вопросов:

1. Вывод формул для всех необходимых распределений аналитически;
2. Нахождение математического ожидания и дисперсии априорных распределений $p(a),\ p(b),\ p(c_n),\ p(d_n)$;
3. Реализация генератора выборки $d_1, \ldots, d_N$ из модели при заданных значениях параметров a и b;
4. Анализ того, как происходит уточнение прогноза для величины b по мере прихода новой информации. Для этого необходимо построить графики и найти мат. ожидание и дисперсию для распределений $p(b), p(b|d_1), \ldots, p(b|d_1, \ldots, d_N)$, где выборка $d_1, \ldots, d_N$
   - 4.1. Сгенерирована из модели при параметрах a и b, равных математическим ожиданиям своих априорных распределений, округлённых до ближайшего целого;
   - 4.2. $d_1 = \ldots = d_N$, где $\forall i \in \overline{1, N}\ d_i$ равно математическому ожиданию своего априорного распределения, округлённого до ближайшего целого.

   Также необходимо провести аналогичный эксперимент, если дополнительно известно значение a и сравнить результаты с первым экспериментом.

5. Проведение временных замеров по оценке всех необходимых распределений $p(c_n),\ p(d_n),\ p(b|d_1), \ldots, p(b|d_1, \ldots, d_N),\ p(b|a),\ p(b|a, d_1), \ldots, p(b|a, d_1, \ldots, d_N)$;

6. Сравнение двух моделей с использованием результатов всех предыдущих пунктов. Демонстрация того, где максимально проявляется разница между ними, через конкретные примеры (не обязательно из экспериментов выше). Объяснение причин подобного результата.

2 Теоретическая часть

В данном разделе приведены теоретические выкладки, необходимые для данного практического задания, а также для программной реализации.

2.1 Задание 1. Аналитический вывод формул для распределений

1.

$$a \sim Unif[a_{min}, a_{max}]$$

$$p_a(x) = \frac{I[a_{min} \le x \le a_{max}]}{a_{max} - a_{min} + 1}$$

2.

$$b \sim Unif[b_{min}, b_{max}]$$

$$p_b(x) = \frac{I[b_{min} \le x \le b_{max}]}{b_{max} - b_{min} + 1}$$

3. Рассмотрим третью модель, в которой рассматривается следующий вид c_n

$$c_n|a, b \sim Bin(a, p_1) + Bin(b, p_2)$$

Используем свёртку двух биномиальных распределений ($p(x) = \sum_k p_{first}(k)p_{second}(x-k)$) с учётом того, что $k \ge 0,\ k \le a,\ x - k \le b$, а также умножаем полученный результат на индикаторную функцию для обозначения границ значений c_n

$$p_{c_n|a,b}(x) = \sum_{k=max(0,x-b)}^{min(a,x)} \binom{a}{k} \binom{b}{x-k} p_1^k p_2^{x-k} (1-p_1)^{a-k} (1-p_2)^{b-(x-k)} I[0 \le x \le a+b]$$

4. Теперь запишем аналогичное распределение, но уже для четвёртой модели.

$$c_n|a, b \sim Pois(ap_1 + bp_2)$$

$$p_{c_n|a,b}(k) = \frac{e^{-(ap_1+bp_2)}(ap_1 + bp_2)^k}{k!} I[0 \le k \le a+b]$$

5. Найдём $p_{c_n}(x)$ через формулу полной вероятности.

$$p_{c_n}(x) = \sum_a \sum_b p(c_n = x \mid a, b)p(a)p(b)I[0 \le x \le a_{max} + b_{max}]$$

$$p_{c_n}(x) = (a_{max} - a_{min} + 1)^{-1}(b_{max} - b_{min} + 1)^{-1} \sum_a \sum_b p(c_n = x \mid a, b)I[0 \le x \le a_{max} + b_{max}]$$

6. Аналогичным образом распишем вид распределений $p_{c_n|b}$ и $p_{c_n|a}$.

$$p_{c_n|b}(x) = (a_{max} - a_{min} + 1)^{-1} \sum_a p(c_n = x|a, b)I[0 \le x \le a_{max} + b]$$

$$p_{c_n|a}(x) = (b_{max} - b_{min} + 1)^{-1} \sum_b p(c_n = x|a, b)I[0 \le x \le a + b_{max}]$$

7. При определении следующего распределения учитываем, что $d_n \in [0, 2(a_{max} + b_{max})]$

$$d_n|c_n \sim c_n + Bin(c_n, p_3)$$

$$p_{d_n|c_n}(x) = \binom{c_n}{x - c_n} p_3^{x-c_n}(1-p_3)^{2c_n - x} I[c_n \le x \le 2c_n]$$

8. Для определения $p_{d_n}(x)$ воспользуемся формулой полной вероятности. Все использующиеся в следующей формуле распределения были получены выше.

$$p_{d_n}(x) = \sum_{k=0}^{a_{max}+b_{max}} p(d_n = x \mid c_n = k) p(c_n = k) I[0 \le x \le 2(a_{max} + b_{max})]$$

9. Также нам могут понадобиться распределения $p_{d_n|a}$, $p_{d_n|b}$ и $p_{d_n|a,b}$. Используя результат для p_{d_n}, сделаем вывод о виде данных распределений.

$$p_{d_n|a,b}(x) = \sum_{k=0}^{a+b} p(d_n = x \mid c_n = k) p(c_n = k|a,b) I[0 \le x \le 2(a+b)]$$

$$p_{d_n|a}(x) = \sum_{b=0}^{b_{max}} p_{d_n|a,b}(x) p(b) [0 \le x \le 2(a + b_{max})]$$

$$p_{d_n|b}(x) = \sum_{a=0}^{a_{max}} p_{d_n|a,b}(x) p(a) [0 \le x \le 2(a_{max} + b)]$$

10. Теперь рассмотрим подсчёт распределений $p(b|d_1),\ p(b|d_1, d_2),\ \ldots,\ p(b|d_1, d_2, \ldots, d_N)$

$$p(b|d_1) = [\textit{ф-ла Байеса}] = \frac{p(d_1|b)p(b)}{p(d_1)} = \frac{\sum_{a=0}^{a_{max}} p(d_1|a,b)p(a)p(b)}{p(d_1)}$$

$$p(b|d_1, d_2, \ldots d_n) = [\textit{ф-ла Байеса}] = \frac{p(d_1, d_2, \ldots d_n|b)p(b)}{p(d_1, d_2, \ldots d_n)} = \frac{\sum_{a=0}^{a_{max}} p(d_1, d_2, \ldots d_n|a,b)p(a)p(b)}{p(d_1, d_2, \ldots d_n)} =$$

$$= [d_i|a,b \textit{ независимы}] = \frac{p(b)\sum_{a=0}^{a_{max}} \prod_{i=1}^{n} p(d_i|a,b)p(a)}{\sum_{b=0}^{b_{max}} \sum_{a=0}^{a_{max}} \prod_{i=1}^{n} p(d_i|a,b)p(a)p(b)} = [p(a), p(b) \textit{ не зависят от } a, b] =$$

$$= \frac{\sum_{a=0}^{a_{max}} \prod_{i=1}^{n} p(d_i|a,b)}{\sum_{b=0}^{b_{max}} \sum_{a=0}^{a_{max}} \prod_{i=1}^{n} p(d_i|a,b)}$$

11. Теперь получим $p(b|a), p(b|a, d_1),\ p(b|a, d_1, d_2),\ \ldots,\ p(b|a, d_1, d_2, \ldots, d_N)$

В силу независимости a и b имеем $p(b|a) = p(b)$. Запишем оставшиеся распределения.

$$p(b|a, d_1) = [\textit{ф-ла Байеса}] = \frac{p(d_1|a,b)p(b)}{p(d_1|a)}$$

$$p(b|a, d_1, d_2, \ldots d_n) = [\textit{ф-ла Байеса}] = \frac{p(d_1, d_2, \ldots d_n|a,b)p(b)}{p(d_1, d_2, \ldots d_n|a)} = [d_i|a,b \textit{ независимы}] =$$

$$\frac{\prod_{i=1}^{n} p(d_i|a,b)p(b)}{\sum_{b=0}^{b_{max}} p(d_1, d_2, \ldots d_n|a,b)p(b)} = \frac{\prod_{i=1}^{n} p(d_i|a,b)}{\sum_{b=0}^{b_{max}} \prod_{i=1}^{n} p(d_i|a,b)}$$

2.2 Задание 2. Нахождение математического ожидания и дисперсии априорных распределений $p(a), p(b), p(c_n), p(d_n)$

1. Начнём с равномерного распределения для случайных величин a и b. Используя табличные данные, сразу же имеем следующие формулы.

$$\mathbb{E}a = \frac{a_{min} + a_{max}}{2},\ \mathbb{E}b = \frac{b_{min} + b_{max}}{2}$$

$$\mathbb{D}a = \frac{(a_{max} - a_{min} + 1)^2 - 1}{12},\ \mathbb{D}b = \frac{(b_{max} - b_{min} + 1)^2 - 1}{12}$$

2. Теперь найдём необходимые статистики для c_n в 3-й модели.
$\mathbb{E}c_n = \mathbb{E}_{a,b}\mathbb{E}[c_n|a,b] = \mathbb{E}_{a,b}\mathbb{E}[Bin(a,p_1) + Bin(b,p_1)] = \mathbb{E}_{a,b}[\mathbb{E}(Bin(a,p_1)) + \mathbb{E}(Bin(b,p_2))] =$
$= \mathbb{E}_{a,b}[ap_1 + bp_2] = p_1\mathbb{E}a + p_2\mathbb{E}b = p_1\frac{a_{min} + a_{max}}{2} + p_2\frac{b_{min} + b_{max}}{2}$
$\mathbb{D}c_n = \mathbb{D}_{a,b}\mathbb{D}[c_n|a,b] = \mathbb{D}_{a,b}\mathbb{D}[Bin(a,p_1) + Bin(b,p_1)] = \mathbb{D}_{a,b}[\mathbb{D}(Bin(a,p_1)) + \mathbb{D}(Bin(b,p_2))] = =$
$\mathbb{D}_{a,b}[ap_1(1-p_1)+bp_2(1-p_2)] = p_1(1-p_1)\mathbb{D}a+p_2(1-p_2)\mathbb{D}b = \frac{p_1(1-p_1)((a_{max} - a_{min} + 1)^2 - 1)}{12} +$
$\frac{p_2(1-p_2)((b_{max} - b_{min} + 1)^2 - 1)}{12}$

3. Рассмотрим подсчёт статистик для c_n в 4-й модели.
$\mathbb{E}c_n = \mathbb{E}_{a,b}\mathbb{E}[c_n|a,b] = \mathbb{E}_{a,b}\mathbb{E}[Pois(ap_1 + bp_1)] = \mathbb{E}_{a,b}[ap_1 + bp_2] = p_1\mathbb{E}a + p_2\mathbb{E}b = p_1\frac{a_{min} + a_{max}}{2} +$
$p_2\frac{b_{min} + b_{max}}{2}$
$\mathbb{D}c_n = \mathbb{D}_{a,b}\mathbb{D}[c_n|a,b] = \mathbb{D}_{a,b}\mathbb{D}[Pois(ap_1 + bp_1)] = \mathbb{D}_{a,b}[ap_1 + bp_2] = p_1^2\mathbb{D}a + p_2^2\mathbb{D}b$

4. Перейдём к подсчёту статистик для d_n

$$\mathbb{E}d_n = \mathbb{E}_{c_n}\mathbb{E}[d_n|c_n] = \mathbb{E}_{c_n}[\mathbb{E}c_n + \mathbb{E}[Bin(c_n,p_3)]] = \mathbb{E}c_n + \mathbb{E}_{c_n}[c_np_3] = (1+p_3)\mathbb{E}c_n$$

$$\mathbb{D}d_n = \mathbb{D}_{c_n}\mathbb{D}[d_n|c_n] = \mathbb{D}_{c_n}[\mathbb{D}c_n + \mathbb{D}[Bin(c_n,p_3)]] = \mathbb{D}[p_3(1-p_3)c_n] = p_3^2(1-p_3)^2\mathbb{D}c_n$$

5. Теперь распишем подсчёт математического ожидания и дисперсии $b|d_1,\dots,d_n$ и $b|a,d_1,\dots,d_n$. В силу того, что все эти распределения являются дискретными, запишем статистики по определению.

$$\mathbb{E}[b|d_1,\dots,d_n] = \sum_{b=b_{min}}^{b_{max}} b\cdot p(b|d_1,\dots,d_n)$$

$$\mathbb{E}[b|a,d_1,\dots,d_n] = \sum_{b=b_{min}}^{b_{max}} b\cdot p(b|a,d_1,\dots,d_n)$$

$$\mathbb{E}[b^2|d_1,\dots,d_n] = \sum_{b=b_{min}}^{b_{max}} b^2\cdot p(b|d_1,\dots,d_n)$$

$$\mathbb{E}[b^2|a,d_1,\dots,d_n] = \sum_{b=b_{min}}^{b_{max}} b^2\cdot p(b|a,d_1,\dots,d_n)$$

$$\mathbb{D}[b|d_1,\dots,d_n] = \mathbb{E}[b^2|d_1,\dots,d_n] + (\mathbb{E}[b|d_1,\dots,d_n])^2$$
$$\mathbb{D}[b|a,d_1,\dots,d_n] = \mathbb{E}[b^2|a,d_1,\dots,d_n] + (\mathbb{E}[b|a,d_1,\dots,d_n])^2$$

3 Практическая часть

3.1 Задание 1. Нахождение математического ожидания и дисперсии априорных распределений $p(a), p(b), p(c_n), p(d_n)$

Подсчёт требуемых статистик осуществлялся через взвешенное среднее библиотеки `numpy`. Результатом стали следующие значения:

Распределение	Математическое ожидание		Дисперсия	
	3 модель	4 модель	3 модель	4 модель
$p(a)$	82.5	82.5	21.25	21.25
$p(b)$	550.0	550.0	850.0	850.0
$p(c_n)$	13.75	13.75	**10.3275**	**0.2975**
$p(d_n)$	17.875	17.875	**0.4554**	**0.0131**

Таблица 1: Статистики априорных распределений

Видно, что различие между моделями заключается лишь в дисперсии для распределений $p(c_n)$ и $p(d_n)$.

3.2 Задание 2. Уточнение апостериорных распределений

Первый эксперимент заключался в семплировании выборки $d_1, \ldots, d_N$ с параметрами a, b, равными своим математическим ожиданиям, округлённым до ближайшего целого числа. После семплирования была проведена оценка математического ожидания и дисперсии каждого апостериорного распределения $p(b|d_1), \ldots, p(b|d_1, \ldots, d_N)$ как без учёта параметра a, так и вместе с ним. Результаты приведены на общем графике.

Рис. 1: Первый эксперимент

1. Наблюдается тенденция повышения среднего значения и понижения дисперсии для величины b
2. Графики достаточно зашумлены в силу случайного семплирования d_i
3. Уточнение знания об a сильно влияет на среднее и дисперсию апостериорного распределения. Видно, что модели без данного параметра слабее уточняют распределение на b в силу более

высокой дисперсии и того, что, несмотря на одинаковую тенденцию к повышению, график математического ожидания у данных моделей ниже, чем у моделей с уточнённым параметром a.

4. Замена суммы биномиальных распределений пуассоновским хорошо заметна - это видно по резкому спаду среднего значения на первых, примерно, 10 значений N, а также по более высокой дисперсии, нежели у 3 модели с биномиальными распределениями на c_n. Это, в целом, легко объяснимо, так как замена 3 модели на 4-ю отражается именно в изменении дисперсии на c_n и d_n, что можно было видеть в сводке статистик для априорных распределений выше

5. Лучшей себя показывает 3-я модель с уточнённым параметром a

Второй эксперимент заключался в выборе среднего значения d_i как сэмплов для всей выборки, то есть $d_1 = \ldots = d_N = mean(d)$. Далее были проведены аналогичные первому эксперименту действия по подсчёту среднего и дисперсии для каждого из апостериорных распределений. Результаты можно видеть на следующем графике:

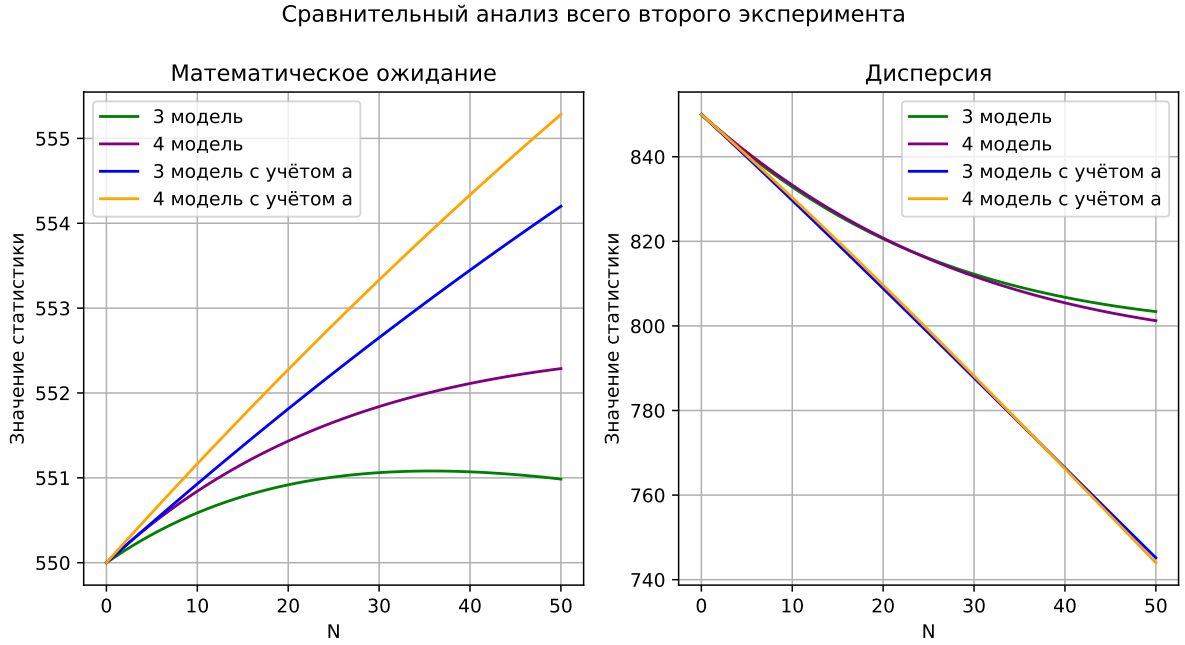

Рис. 2: Первый эксперимент

1. Уточнение знания о d_i и a увеличивает математическое ожидание количества студентов непрофильных факультетов (b) и уменьшает дисперсию данной величины, что делает каждое следующее апостериорное распределение более определённым, с меньшей энтропией.

2. Знание a - числа студентов профильного факультета - преобразует график почти в линейный. При этом без данной величины графики словно выходят на плато с большим значением N

3. В отличие от случайного семплирования d_i, данные графики гладкие, незашумлённые.

4. Уточнение апостериорного распределения действительно отражает то, насколько хорошо мы знаем рассматриваемую модель.

5. Сильной разницы между 3 и 4 моделями здесь не наблюдается

3.3 Задание 3. Измерение времени для оценки необходимых распределений

Третья модель оказалась быстрее в плане подсчёта распределений. Для основных априорных распределений результаты следующие:

$p(c_n)$		$p(d_n)$	
3 модель	4 модель	3 модель	4 модель
0.092	0.213	**0.258**	0.347

Таблица 2: Время оценки априорных распределений $p(c_n), p(d_n)$

Для апостериорных распределений был построен следующий график:

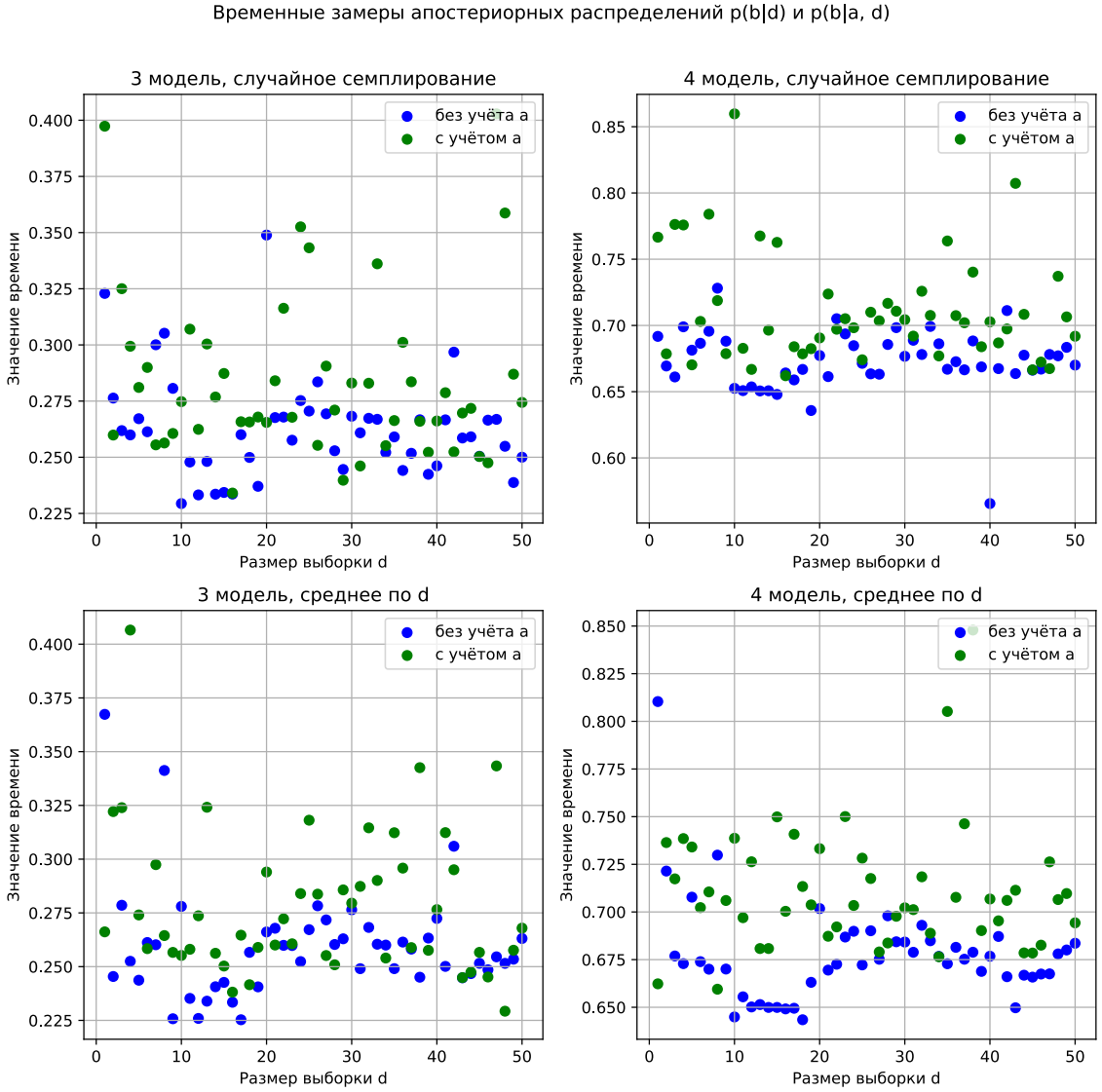

Рис. 3: Временные замеры, апостериорные распределения

Легко видеть, что 3 модель быстрее 4 во всех экспериментах. При этом важно заметить, что время оценки распределения увеличивается, если известен параметр a

4 Выводы

Были проанализированы 2 предложенные модели посещения студентами лекций. В рамках задания приведены теоретические выкладки, необходимые для реализации модели.

По результатам экспериментов видно, что 3 модель, несмотря на свой более «сложный» вид в силу суммы биномиальных распределений, выглядит более предпочтительной. Подсчёт $p(c_n)$ в виде свёртки оказывается даже быстрее, нежели подсчёт того же распределения для 4-й модели.

По графикам оценки времени апостериорных распределений, третья модель тоже выигрывает почти в два раза.

Случайное семплирование значений d_i показывает, что 3 модель лучше подходит под тенденцию увеличения математического ожидания, а также снижения дисперсии оценки b. В случае с взятием среднего по d_i видно, что обе модели не сильно отличаются друг от друга, а больше влияет знание о величине a.

Также важно отметить, что 3-я модель более подходит под действительность, так как бесконечное число студентов - это всего лишь абстракция, которую в реальной жизни наблюдать не получится.

В целом результат того, что 3-я модель лучше 4-й, легко объясним неудачной аппроксимацией Пуасоновским распределением, ибо наша модель и здравый смысл не подразумевают бесконечного количества студентов двух факультетов. Касательно времени вычисления - вероятно, операция свёртки реализована более оптимальным способом, что делает её более быстрой для подсчёта распределения $p(c_n)$