Problem 1. Вывести формулы для всех необходимых далее распределений аналитически.

$$p(a) = \frac{1}{a_{max} - a_{min} + 1}$$

$$p(b) = \frac{1}{b_{max} - b_{min} + 1}$$

В случае модели номер 3 имеем:

$$p(c_1 = c|a,b) = \sum_{i=0}^{\min(a,c)} p(X_1 = i|a) \cdot p(X_2 = c - i|b) = \sum_{i=0}^{\min(a,c)} \binom{a}{i} p_1^i (1-p_1)^{a-i} \cdot \binom{b}{c-i} p_2^{c-i} (1-p)^{b-c+i}$$

В случае модели номер 4 имеем:

$$p(c_1 = c|a,b) = \frac{(ap_1 + bp_2)^c \exp(-(ap_1 + bp_2))}{c!}$$

Идём далее:

$$p(d_1 = d|c_1) = p(X = d - c_1|c_1) = \binom{c_1}{d - c_1} p_3^{d-c_1} (1-p)^{2c_1 - d}$$

С базовыми разобрались, теперь переходим к тем, что в заданиях:

$$p(c_n) = p(c_1) = \sum_{a,b} p(a,b,c_1) = \sum_{a,b} p(c_1|a,b) \cdot p(a) \cdot p(b)$$

$$p(d_n) = p(d_1) = \sum_{c_1} p(c_1, d_1) = \sum_{c_1} p(d_1|c_1) \cdot p(c_1)$$

$$p(b|d_1,...,d_n) = \frac{p(b,d_1,...,d_n)}{p(d_1,...,d_n)} = \frac{\sum_{a,c_1,...,c_n} p(a,b,c_1,...,c_n,d_1,...,d_n)}{\sum_{a,b,c_1,...,c_n} p(a,b,c_1,...,c_n,d_1,...,d_n)} = \frac{\sum_{a,c_1,...,c_n} p(a)p(b)\prod_i p(d_i|c_i) \cdot p(c_i|a,b)}{\sum_{a,b,c_1,...,c_n} p(a)p(b)\prod_i p(d_i|c_i) \cdot p(c_i|a,b)}$$

$$p(b|a,d_1,...,d_n) = \frac{p(b,a,d_1,...,d_n)}{p(a,d_1,...,d_n)} = \frac{\sum_{c_1,...,c_n} p(a,b,c_1,...,c_n,d_1,...,d_n)}{\sum_{b,c_1,...,c_n} p(a,b,c_1,...,c_n,d_1,...,d_n)} = \frac{\sum_{c_1,...,c_n} p(a)p(b)\prod_i p(d_i|c_i) \cdot p(c_i|a,b)}{\sum_{b,c_1,...,c_n} p(a)p(b)\prod_i p(d_i|c_i) \cdot p(c_i|a,b)}$$

Problem 2.
Найти математические ожидания и дисперсии априорных распределений $p(a), p(b), p(c_n), p(d_n)$.

Для первых двух распределений есть готовые формулы, так что пользуемся ими:

$$\mathbb{E}[a] = \frac{a_{min} + a_{max}}{2} = 82.5$$

$$\mathbb{E}[b] = \frac{b_{min} + b_{max}}{2} = 550$$

$$\mathbb{V}ar[a] = \frac{(a_{max} - a_{min} + 1)^2 - 1}{12} = 21.25$$

$$\mathbb{V}ar[b] = \frac{(b_{max} - b_{min} + 1)^2 - 1}{12} = 850$$

С остальными чуть интереснее: мы знаем условные распределения, а значит можно пользоваться формулами полного матожидания и дисперсии (считаем для модели 3 и 4 соответственно):

$$\mathbb{E}[c_n] = \mathbb{E}[\mathbb{E}[c_n|a,b]] = \mathbb{E}[\mathbb{E}[\text{Bin}(a,p_1)] + \mathbb{E}[\text{Bin}(b,p_2)]] = \mathbb{E}[ap_1 + bp_2] = p_1\mathbb{E}[a] + p_2\mathbb{E}[b] = 13.75$$

$$\mathbb{V}ar[c_n] = \mathbb{E}[\mathbb{V}ar[c_n|a,b]] + \mathbb{V}ar[\mathbb{E}[c_n|a,b]] = \mathbb{E}[ap_1(1-p_1) + bp_2(1-p_2)] + \mathbb{V}ar[ap_1 + bp_2] =$$
$$= p_1(1-p_1)\mathbb{E}[a] + p_2(1-p_2)\mathbb{E}[b] + p_1^2\mathbb{V}ar[a] + p_2^2\mathbb{V}ar[b] = 13.17$$

$$\mathbb{E}[c_n] = \mathbb{E}[\mathbb{E}[c_n|a,b]] = \mathbb{E}[\mathbb{E}[\text{Poiss}(ap_1 + bp_2)]] = \mathbb{E}[ap_1 + bp_2] = p_1\mathbb{E}[a] + p_2\mathbb{E}[b] = 13.75$$

$$\mathbb{V}ar[c_n] = \mathbb{E}[\mathbb{V}ar[c_n|a,b]] + \mathbb{V}ar[\mathbb{E}[c_n|a,b]] = \mathbb{E}[ap_1 + bp_2] + \mathbb{V}ar[ap_1 + bp_2] = \mathbb{E}[ap_1 + bp_2] + p_1^2\mathbb{V}ar[a] + p_2^2\mathbb{V}ar[b] = 14.05$$

$$\mathbb{E}[d_n] = \mathbb{E}[\mathbb{E}[d_n|c_n]] = \mathbb{E}[\mathbb{E}[c_n + \text{Bin}(c_n, p_3)]] = \mathbb{E}[c_n + c_n p_3] = (1 + p_3)\mathbb{E}[c_n] = 17.875$$

$$\mathbb{V}ar[d_n] = \mathbb{E}[\mathbb{V}ar[d_n|c_n]] + \mathbb{V}ar[\mathbb{E}[d_n|c_n]] = \mathbb{E}[c_n p_3(1-p_3)] + \mathbb{V}ar[c_n(1+p_3)] = p_3(1-p_3)\mathbb{E}[c_n] + (1+p_3)^2\mathbb{V}ar[c_n] =$$

$$= 25.14 \text{ для модели 3}$$

$$= 26.63 \text{ для модели 4}$$

		mean	varience
0	p(a)	82.50	21.25
1	p(b)	549.90	850.00
2	p(c)3	13.80	13.16
3	p(c)4	13.80	14.04
4	p(d)3	17.87	25.14
5	p(d)4	17.87	26.62

Рис. 1: Видим, что полученные значения совпадают.

Уже сейчас можно заметить, что дисперсия у модели 4 (которая с Пуассоном) чуть больше, но мы вернёмся к этому чуть позже.

Problem 3. Реализовать генератор выборки $d_1, ..., d_N$ из модели при заданных значениях параметров a, b.

Кратко проговорим идею (реализация в коде имеется). Сначала сэмплируем c по заданным a, b: если это модель 3, то делаем сумму двух биномиальных распределений, если 4, то из Пуассона; далее к полученной c прибавляем ещё одно биномиальное.

Problem 4. Пронаблюдать, как происходит уточнение прогноза для величины b по мере прихода новой косвенной информации. Для этого построить графики и найти мат.ожидание и дисперсию для распределений $p(b), p(b|d_1), ..., p(b|d_1, ..., d_N)$,где выборка $d1, ..., d_N$ 1) сгенерирована из модели при параметрах a, b, равных мат.ожиданиям своих априорных распределений, округленных до ближайшего целого и 2) $d_1 = ... = d_N$, где d_n равно мат.ожиданию своего априорного распределения, округленного до ближайшего целого. Провести аналогичный эксперимент, если дополнительно известно значение a. Сравнить результаты двух экспериментов.

Смотря на график можно сказать следующее:

1. Дисперсия модели 3 меньше, чем у модели 4;
2. У дисперсии есть нисходящий тренд (то есть в целом она становится меньше). В первом случае это происходит не монотонно, потому что d не одинаковые. Но каждое новое значение позволяет оценить b с меньшим разбросом.

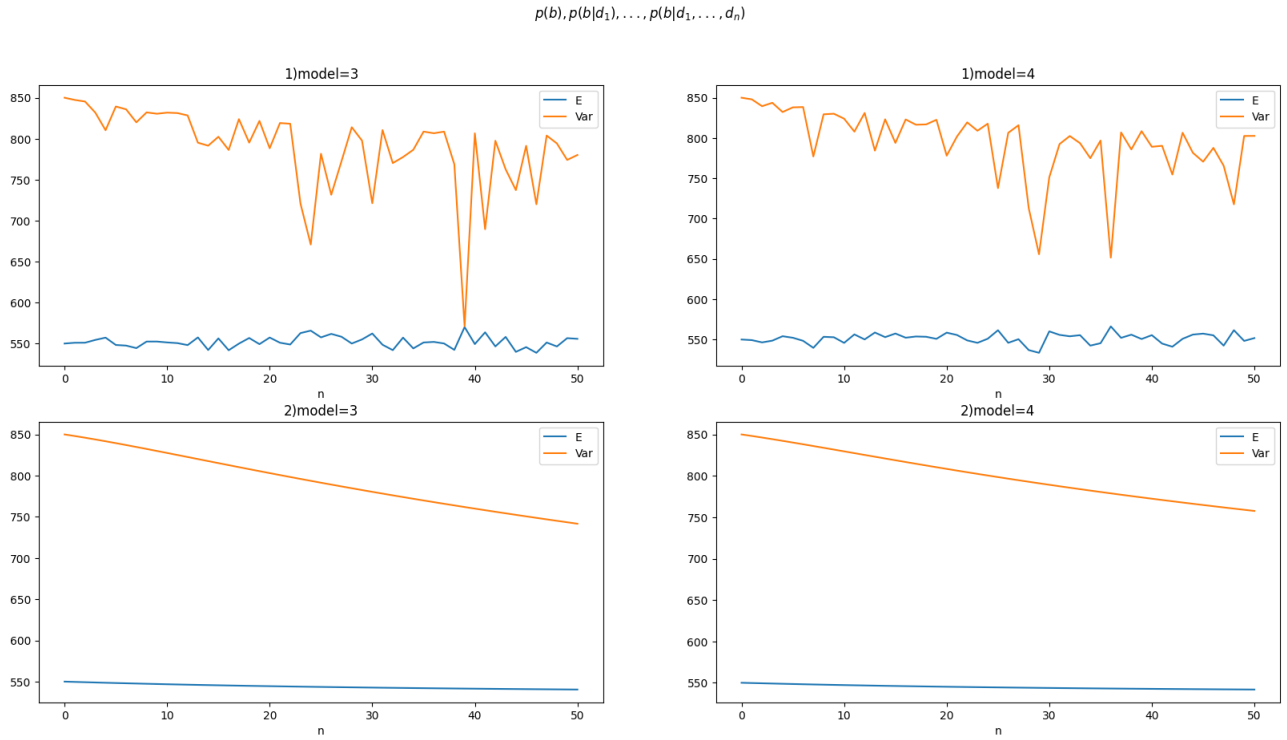

Рис. 2: График средних и дисперсий для b при условии d_n.

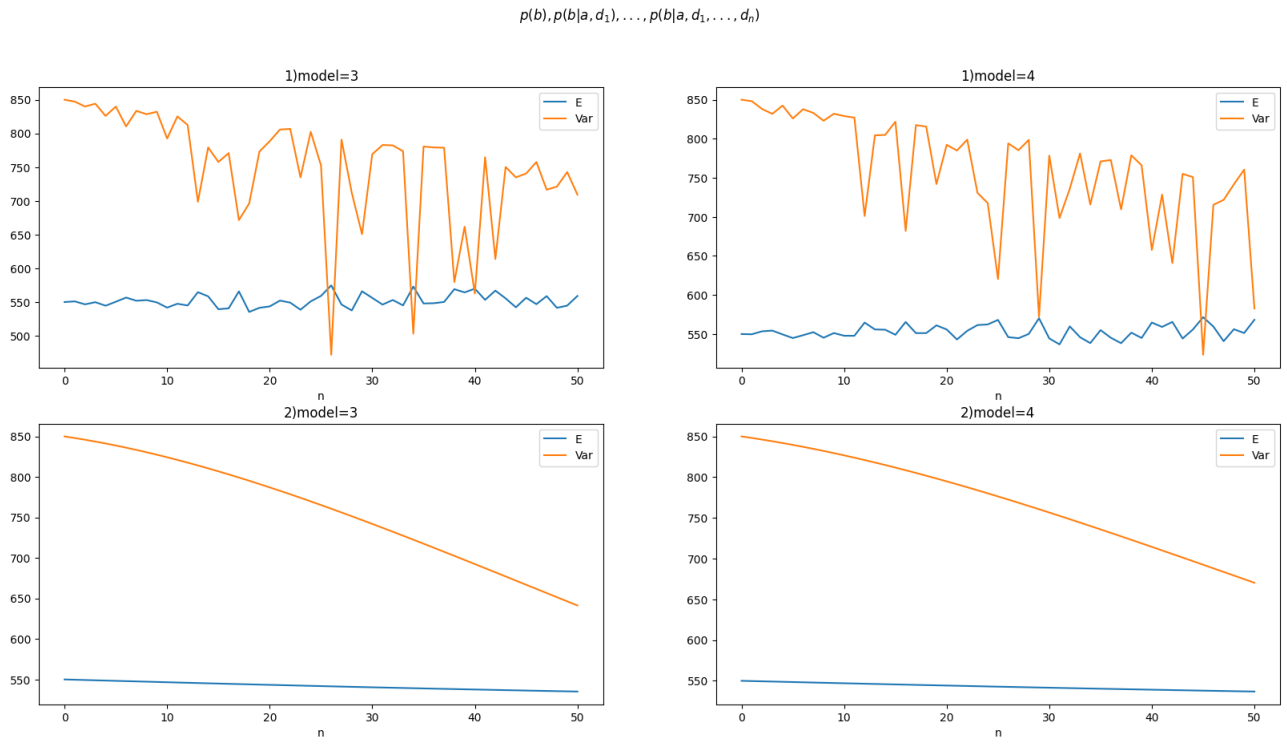

Рис. 3: График средних и дисперсий для b при условии a и d_n.

3. Матожидание во втором случае также уменьшается. Физический смысл такой: видя, что каждую неделю на лекции записывается одинаковое число человек, мы можем слегка понижать оценку количества записавшихся студентов.

   Теперь перейдём к графику, где добавляется условие на a:

   Наблюдения тут все те же, только добавляется ещё одно: наличие условия a ещё сильнее снижает дисперсию. Опять же тут нет ничего необычного: если мы понимаем, сколько студентов с профиля записались на курс, то и делать оценку на количество непрофильных студентов можем лучше.

Problem 5. Провести временные замеры по оценке всех необходимых распределений $p(c_n), p(d_n), p(b|d_1, ..., d_n), p(b|a, d_1, ..., d_n)$.

Время, приведённое в таблице, это среднее 100 запусков.

		model3(ms)	model4(ms)
0	p(c)	23	18
1	p(d)	43	38
2	p(b|d1, ..., dn)	49	45
3	p(b|a, d1, ..., dn)	39	36

Рис. 4: Время запусков необходимых распределений.

Видно, что посчитать распределение при условии a сильно проще, потому что не надо делать по нему суммирование. Также время для модели 4 чуть меньше, чем для модели номер 3. Происходит это, как мне кажется, потому что считать биноминальное распределение сложнее из-за перестановок.

Problem 6. Используя результаты всех предыдущих пунктов, сравнить две модели. Показать, где максимально проявляется разница между ними (привести конкретный пример, не обязательно из экспериментов выше). Объяснить причины подобного результата.

Перед нам встаёт классический trade-off: у нас есть модель 3, которая имеет меньшую дисперсию, но при этом сложносчитаемое распределение, и модель 4, у которой дисперсия выше, но и считается всё быстрее. Как я писал выше, на графике с матожиданием и дисперсией, дисперсия у модели 3 заметно ниже, при это время подсчёта отличается на совсем немного. Так что я бы отдал предпочтение модели 3.