Домашнее задание 1

Королев Кирилл

Вариант 1

Рассматриваются модели 1 и 2 с параметрами $a_{min} = 75, a_{max} = 90, b_{min} = 500, b_{max} = 600, p_1 = 0.1, p_2 = 0.01, p_3 = 0.3$.

Задание 1

Вывести формулы для всех необходимых далее распределений аналитически.

Распределение $p(a)$

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

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

Распределение $p(b)$

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

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

Для первой модели $p(c \mid a, b)$

$$X \sim Bin(a, p_1),\ Y \sim Bin(b, p_2),\ c \mid a, b \sim X + Y$$

$$p(c = k \mid a, b) = \sum_{j=0}^{\min(a,k)} p(X = j)\, p(Y = k - j) = \sum_{j=0}^{\min(a,k)} \binom{a}{j} p_1^j (1 - p_1)^{a-j} \binom{b}{k-j} p_2^{k-j} (1 - p_2)^{b+j-k}$$

Для второй модели $p(c \mid a, b)$

$$c \mid a, b \sim Poiss(ap_1 + bp_2)$$

$$p(c = k \mid a, b) = \frac{(ap_1 + bp_2)^k}{k!} e^{-(ap_1 + bp_2)}$$

Распределение $p(d \mid c)$

$$X \sim Bin(c, p_3),\ d \mid c \sim c + X$$

$$p(d = k \mid c) = p(c + X = k \mid c) = p(X = k - c \mid c) = \binom{c}{k-c} p_3^{k-c} (1 - p_3)^{2c-k}$$

Распределение $p(c \mid a)$

$$p(c \mid a) = \frac{p(a,c)}{p(a)} = \sum_b \frac{p(c,a,b)}{p(a)} = \sum_b p(c \mid a,b)p(b)$$

Распределение $p(c \mid b)$

$$p(c \mid b) = \frac{p(b,c)}{p(b)} = \sum_a \frac{p(c,a,b)}{p(b)} = \sum_a p(c \mid a,b)p(a)$$

Распределение $p(c \mid d)$

$$p(c \mid d) = \frac{p(d \mid c)p(c)}{\sum_{c'} p(d \mid c')p(c')}$$

Распределение $p(c)$

$$p(c) = \sum_{a,b} p(a,b,c) = \sum_{a,b} p(c \mid a,b)p(a)p(b)$$

Распределение $p(d)$

$$p(d) = \sum_c p(c,d) = \sum_c p(d \mid c)p(c)$$

Распределение $p(c \mid a,b,d)$

$$p(c \mid a,b,d) = \frac{p(a,b,c,d)}{p(a,b,d)} = \frac{p(a,b,c,d)}{\sum_{c'} p(a,b,c',d)} = \frac{p(d \mid c)p(c \mid a,b)p(a)p(b)}{\sum_{c'} p(d \mid c')p(c' \mid a,b)p(a)p(b)} = \frac{p(d \mid c)p(c \mid a,b)}{\sum_{c'} p(d \mid c')p(c' \mid a,b)}$$

Задание 2

Найти математические ожидания и дисперсии априорных распределений $p(a), p(b), p(c), p(d)$.

Матожидания и дисперсии дискретных равномерных распределений p(a) и p(b) считаются по известным формулам.

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

$$\mathbb{D}a = \frac{(a_{max} - a_{min} + 1)^2 - 1}{12} = \mathbf{21.25}$$

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

$$\mathbb{D}b = \frac{(b_{max} - b_{min} + 1)^2 - 1}{12} = \mathbf{850}$$

Посчитаем ожидания и дисперсии распределений $p(c)$ и $p(d)$. Зная условные распределения $p(c \mid a,b)$ и $p(d \mid c)$ можем воспользоваться формулой полного матожидания.

$$\mathbb{E}[c \mid a,b] = \mathbb{E}[Bin(a,p_1) + Bin(b,p_2) \mid a,b] = ap_1 + bp_2$$

$$\mathbb{E}c = \mathbb{E}_{a,b}\left[\mathbb{E}[c \mid a,b]\right] = \frac{a_{min} + a_{max}}{2}p_1 + \frac{b_{min} + b_{max}}{2}p_2 = \mathbf{13.75}$$

Посчитаем следующую условную дисперсию

$$\mathbb{D}[c \mid a, b] = \mathbb{D}[Bin(a, p_1) + Bin(b, p_2) \mid a, b] = ap_1(1 - p_1) + bp_2(1 - p_2)$$

Тогда по формуле для условной дисперсии получаем

$$\mathbb{D}c = \mathbb{E}[\mathbb{D}[c \mid a, b]] + \mathbb{D}[\mathbb{E}[c \mid a, b]] = \mathbb{E}[ap_1(1 - p_1) + bp_2(1 - p_2)] + \mathbb{D}[ap_1 + bp_2] =$$

$$= \frac{a_{min} + a_{max}}{2} p_1(1-p_1) + \frac{b_{min} + b_{max}}{2} p_2(1-p_2) + p_1^2 \frac{(a_{max} - a_{min} + 1)^2 - 1}{12} + p_2^2 \frac{(b_{max} - b_{min} + 1)^2 - 1}{12} = \mathbf{13.1675}$$

Для второй модели делаем аналогично.

$$\mathbb{E}[c \mid a, b] = \mathbb{E}[Poiss(ap_1 + bp_2) \mid a, b] = ap_1 + bp_2$$

$$\mathbb{E}c = \mathbb{E}_{a,b}\left[\mathbb{E}[c \mid a, b]\right] = \frac{a_{min} + a_{max}}{2} p_1 + \frac{b_{min} + b_{max}}{2} p_2 = \mathbf{13.75}$$

$$\mathbb{D}[c \mid a, b] = \mathbb{D}[Poiss(ap_1 + bp_2) \mid a, b] = ap_1 + bp_2$$

$$\mathbb{D}c = \mathbb{E}[\mathbb{D}[c \mid a, b]] + \mathbb{D}[\mathbb{E}[c \mid a, b]] = \mathbb{E}[ap_1 + bp_2] + \mathbb{D}[ap_1 + bp_2] =$$

$$= \frac{a_{min} + a_{max}}{2} p_1 + \frac{b_{min} + b_{max}}{2} p_2 + p_1^2 \frac{(a_{max} - a_{min} + 1)^2 - 1}{12} + p_2^2 \frac{(b_{max} - b_{min} + 1)^2 - 1}{12} = \mathbf{14.0475}$$

Ожидание распределения $p(d)$ одинаково для обоих моделей, так как $\mathbb{E}c$ совпадает.

$$\mathbb{E}d = \mathbb{E}[\mathbb{E}[d \mid c]] = \mathbb{E}[\mathbb{E}[c + Bin(c, p_3) \mid c]] = \mathbb{E}[c + cp_3] = \mathbf{17.875}$$

Дисперсия для распределения $p(d)$.

$$\mathbb{D}d = \mathbb{E}[\mathbb{D}[d \mid c]] + \mathbb{D}[\mathbb{E}[d \mid c]] = \mathbb{E}[cp_3(1 - p_3)] + \mathbb{D}[c(1 + p_3)] = p_3(1 - p_3)\mathbb{E}c + (1 + p_3)^2\mathbb{D}c$$

Таким образом, для первой модели $\mathbb{D}c \approx \mathbf{25.141}$, а для второй $\mathbb{D}c \approx \mathbf{26.628}$.

	expectation	variance
pa	82.500	21.250
pb	550.000	850.000
pc1	13.750	13.168
pc2	13.750	14.048
pd1	17.875	25.141
pd2	17.875	26.628

Ожидание и дисперсия априорных распределений, округленные до 3-х знаков после запятой

Ожидание распределения $p(d)$ одинаково для обоих моделей, так как $\mathbb{E}c$ совпадает.

$$\mathbb{E}d = \mathbb{E}[\mathbb{E}[d \mid c]] = \mathbb{E}[\mathbb{E}[c + Bin(c, p_3) \mid c]] = \mathbb{E}[c + cp_3] = \mathbf{17.875}$$

Дисперсия для распределения $p(d)$.

$$\mathbb{D}d = \mathbb{E}[\mathbb{D}[d \mid c]] + \mathbb{D}[\mathbb{E}[d \mid c]] = \mathbb{E}[cp_3(1 - p_3)] + \mathbb{D}[c(1 + p_3)] =$$

$$= p_3(1 - p_3)\mathbb{E}c + (1 + p_3)^2\mathbb{D}c$$

Таким образом, для первой модели $\mathbb{D}c \approx \mathbf{25.141}$, а для второй $\mathbb{D}c \approx \mathbf{26.628}$.

Задание 3

Пронаблюдать, как происходит уточнение прогноза для величины c по мере прихода новой косвенной информации. Для этого построить графики и найти мат.ожидание и дисперсию для распределений $p(c), p(c|a), p(c|b), p(c|d), p(c|a, b), p(c|a, b, d)$ при параметрах a, b, d, равных мат.ожиданиям своих априорных распределений, округленных до ближайшего целого.

Для обеих моделей получаем

$$\mathbb{E}c = p_1\mathbb{E}a + p_2\mathbb{E}b = 13.75$$

В качестве a берем $[\mathbb{E}a] = 82$ и для обеих моделей получаем

$$\mathbb{E}[c \mid a] = \mathbb{E}_b[\mathbb{E}[c \mid a, b]] = \mathbb{E}_b[ap_1 + bp_2] = ap_1 + p_2\mathbb{E}b = 13.7$$

В качестве b берем $[\mathbb{E}b] = 550$ и для обеих моделей получаем

$$\mathbb{E}[c \mid b] = \mathbb{E}_a[\mathbb{E}[c \mid a, b]] = \mathbb{E}_a[ap_1 + bp_2] = p_1\mathbb{E}a + bp_2 = 13.75$$

Для первой модели ожидание $\mathbb{E}[c \mid d]$ посчитаем численно

$$\mathbb{E}[c \mid d] \approx 13.896$$

Для второй модели ожидание $\mathbb{E}[c \mid d]$ тоже посчитаем численно

$$\mathbb{E}[c \mid d] \approx 13.894$$

Знаем из предыдущего пункта, что $\mathbb{E}[c \mid a, b] = ap_1 + bp_2$, таким образом, $\mathbb{E}[c \mid a = [\mathbb{E}a], b = [\mathbb{E}b]] = 13.7$ для обеих моделей.

Для первой модели ожидание $\mathbb{E}[c \mid a, b, d]$ посчитаем численно

$$\mathbb{E}[c \mid a, b, d] \approx 13.891$$

Для второй модели ожидание $\mathbb{E}[c \mid a, b, d]$ тоже посчитаем численно

$$\mathbb{E}[c \mid a, b, d] \approx 13.889$$

$\mathbb{D}c$ уже считали, в частности, для первой модели $\mathbb{D}c = 13.1675$, а для второй $\mathbb{D}c = 14.0475$.

В качестве a берем $[\mathbb{E}a] = 82$ и для первой модели получаем

$$\mathbb{D}[c \mid a] = \mathbb{E}_b[\mathbb{D}[c \mid a, b]] + \mathbb{D}_b[\mathbb{E}[c \mid a, b]] = \mathbb{E}_b[ap_1(1 - p_1) + bp_2(1 - p_2)] + \mathbb{D}_b[ap_1 + bp_2] =$$
$$= p_1(1 - p_1)a + p_2(1 - p_2)\mathbb{E}b + p_2^2\mathbb{D}b = 12.91$$

Для второй модели получаем

$$\mathbb{D}[c \mid a] = \mathbb{E}_b[\mathbb{D}[c \mid a, b]] + \mathbb{D}_b[\mathbb{E}[c \mid a, b]] = \mathbb{E}_b[ap_1 + bp_2] + \mathbb{D}_b[ap_1 + bp_2] =$$
$$= p_1a + p_2\mathbb{E}b + p_2^2\mathbb{D}b = 13.785$$

В качестве b берем $[\mathbb{E}b] = 550$ и для первой модели получаем

$$\mathbb{D}[c \mid b] = \mathbb{E}_a[\mathbb{D}[c \mid a, b]] + \mathbb{D}_a[\mathbb{E}[c \mid a, b]] = \mathbb{E}_a[ap_1(1 - p_1) + bp_2(1 - p_2)] + \mathbb{D}_a[ap_1 + bp_2] =$$
$$= p_1(1 - p_1)\mathbb{E}a + p_2(1 - p_2)b + p_1^2\mathbb{D}a = 13.0825$$

Для второй модели получаем

$$\mathbb{D}[c \mid b] = \mathbb{E}_a[\mathbb{D}[c \mid a, b]] + \mathbb{D}_a[\mathbb{E}[c \mid a, b]] = \mathbb{E}_a[ap_1 + bp_2] + \mathbb{D}_a[ap_1 + bp_2] =$$
$$= p_1\mathbb{E}a + p_2b + p_1^2\mathbb{D}a = 13.9625$$

	expectation	variance
pc	13.750	13.168
pc_a	13.700	12.910
pc_b	13.750	13.083
pc_d	13.896	1.534
pc_ab	13.700	12.825
pc_abd	13.891	1.529
pc2	13.750	14.048
pc_a2	13.700	13.785
pc_b2	13.750	13.963
pc_d2	13.894	1.544
pc_ab2	13.700	13.700
pc_abd2	13.889	1.540

Ожидание и дисперсия условных распределений, округленные до 3-х знаков после запятой

Для первой модели дисперсию $\mathbb{D}[c \mid d]$ посчитаем численно

$$\mathbb{D}[c \mid d] \approx 1.534$$

Для второй модели дисперсию $\mathbb{D}[c \mid d]$ тоже посчитаем численно

$$\mathbb{D}[c \mid d] \approx 1.544$$

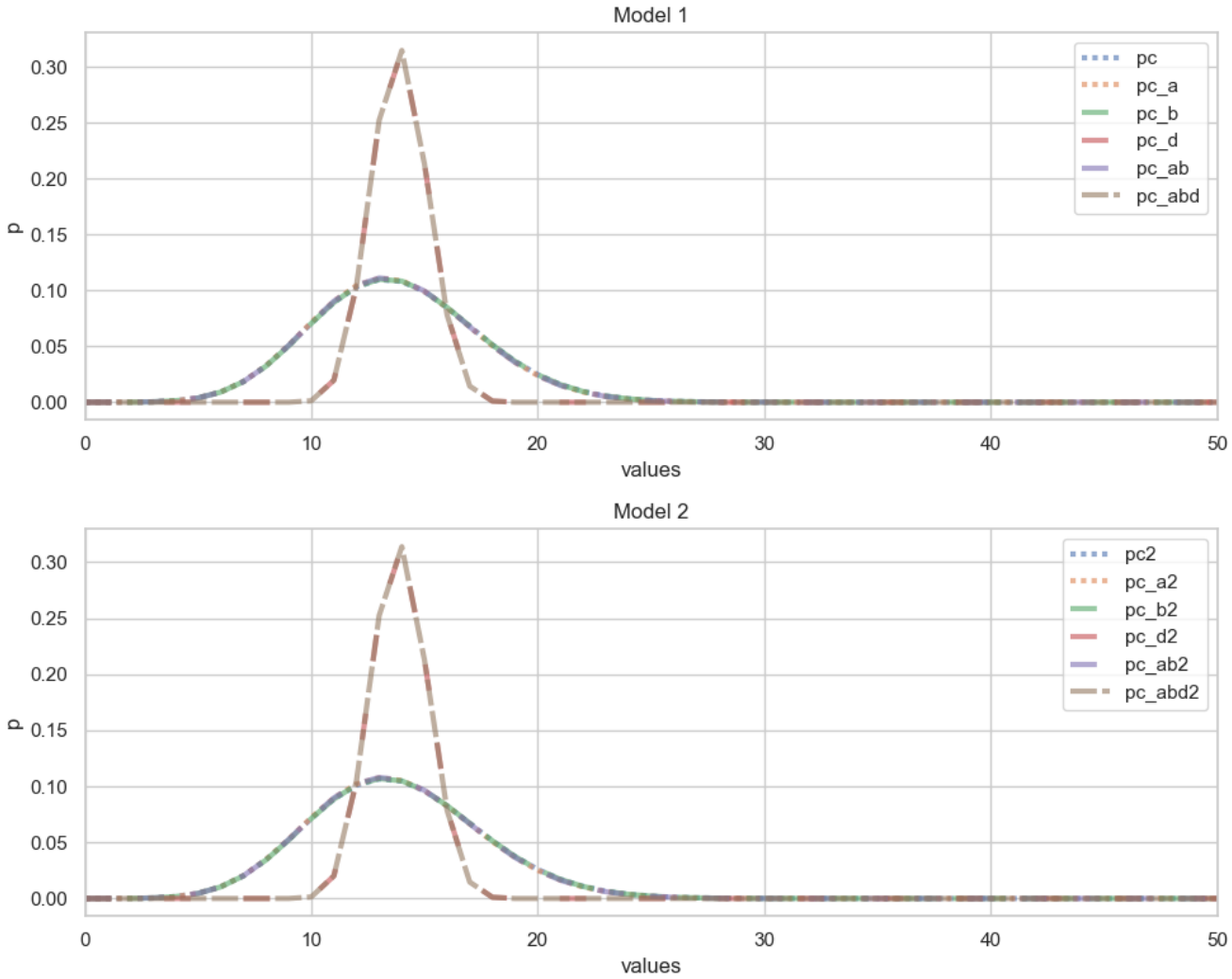
Знаем из предыдущего пункта, что $\mathbb{D}[c \mid a, b] = ap_1(1 - p_1) + bp_2(1 - p_2)$, для первой модели, таким образом, $\mathbb{D}[c \mid a = [\mathbb{E}a], b = [\mathbb{E}b]] = 12.825$. Для второй модели $\mathbb{D}[c \mid a, b] = ap_1 + bp_2$ и $\mathbb{D}[c \mid a = [\mathbb{E}a], b = [\mathbb{E}b]] = 13.7$.

Для первой модели дисперсию $\mathbb{D}[c \mid a, b, d]$ посчитаем численно

$$\mathbb{D}[c \mid a, b, d] \approx 1.529$$

Для второй модели дисперсию $\mathbb{D}[c \mid a, b, d]$ тоже посчитаем численно

$$\mathbb{D}[c \mid a, b, d] \approx 1.54$$

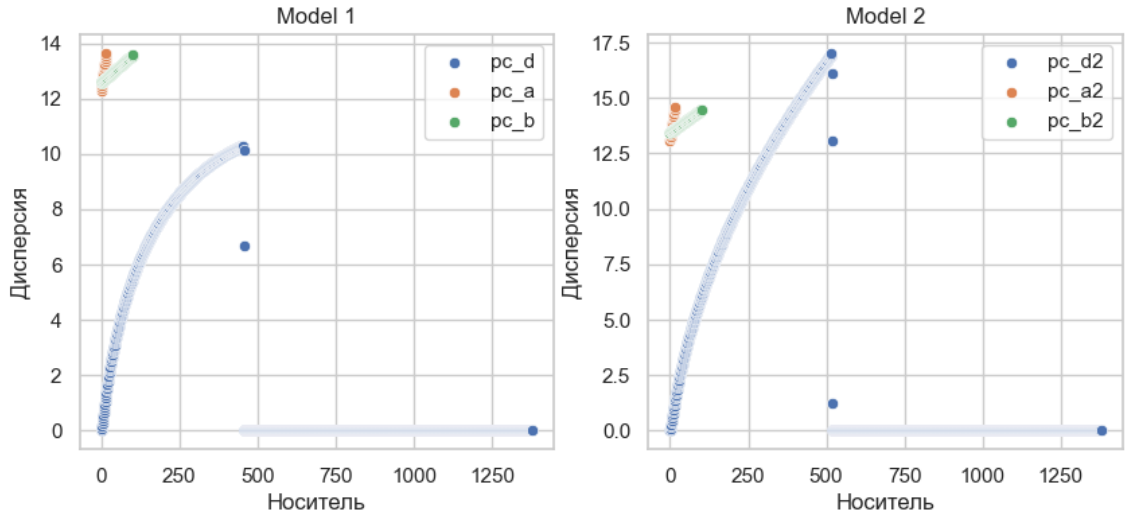


Графики условных распределений для первой и второй модели

Видим, что обуславливаясь на d, мы наибольшим образом уточняем прогноз c, распределение становится более вырожденным, дисперсия значительно уменьшается. Обуславливание на a и b дает мало информации, распределение почти не меняется по сравнению с априорным.

Задание 4

Определить, какая из величин a, b, d вносит наибольший вклад в уточнение прогноза для величины c (в смысле дисперсии распределения). Для этого проверить верно ли, что $D[c|d] < D[c|b]$ и $D[c|d] < D[c|a]$ для любых допустимых значений a, b, d. Найти множество точек (a, b) таких, что $D[c|b] < D[c|a]$. Являются ли множества $\{(a, b)|D[c|b] < D[c|a]\}$ и $\{(a, b)|D[c|b]D[c|a]\}$ линейно разделимыми?



Сравнение дисперсий

Модель 1

Действительно, после численного подсчета дисперсий, оказалось верно, что $\mathbb{D}[c \mid d] < \mathbb{D}[c \mid a]$ и $\mathbb{D}[c \mid d] < \mathbb{D}[c \mid b]$ для всех допустимых значений a, b, d.

$$\mathbb{D}[c \mid a] = \mathbb{E}_b[\mathbb{D}[c \mid a, b]] + \mathbb{D}_b[\mathbb{E}[c \mid a, b]] = \mathbb{E}_b[ap_1(1-p_1) + bp_2(1-p_2)] + \mathbb{D}_b[ap_1 + bp_2] = p_1(1-p_1)a + p_2(1-p_2)\mathbb{E}b + p_2^2\mathbb{D}b$$

$$\mathbb{D}[c \mid b] = \mathbb{E}_a[\mathbb{D}[c \mid a, b]] + \mathbb{D}_a[\mathbb{E}[c \mid a, b]] = \mathbb{E}_a[ap_1(1-p_1) + bp_2(1-p_2)] + \mathbb{D}_a[ap_1 + bp_2] = p_1(1-p_1)\mathbb{E}a + p_2(1-p_2)b + p_1^2\mathbb{D}a$$

Если $\mathbb{D}[c \mid b] - \mathbb{D}[c \mid a] = 0$ задает уравнение прямой, то множества будут линейно разделимыми. Действительно, получаем прямую

$$b = \frac{p_1(1-p_1)a + p_2(1-p_2)\mathbb{E}b + p_2^2\mathbb{D}b - p_1(1-p_1)\mathbb{E}a - p_1^2\mathbb{D}a}{p_2(1-p_2)}$$

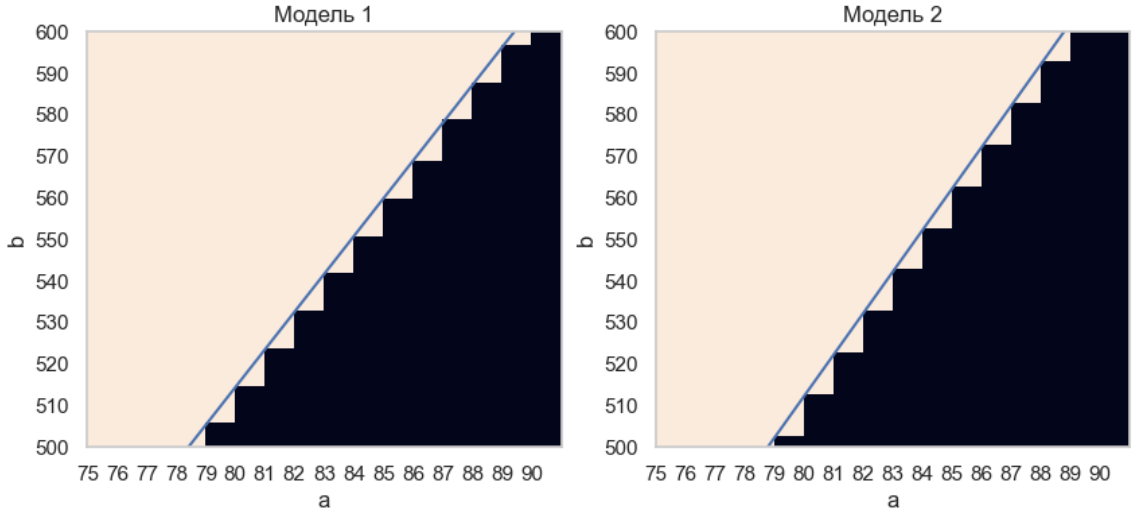
Модель 2

Для второй модели наблюдаем, что неравенство не выполняется и дисперсия для $\mathbb{D}[c \mid d]$ увеличивается.

$$\mathbb{D}[c \mid a] = \mathbb{E}_b[\mathbb{D}[c \mid a, b]] + \mathbb{D}_b[\mathbb{E}[c \mid a, b]] = \mathbb{E}_b[ap_1 + bp_2] + \mathbb{D}_b[ap_1 + bp_2] = p_1a + p_2\mathbb{E}b + p_2^2\mathbb{D}b$$

$$\mathbb{D}[c \mid b] = \mathbb{E}_a[\mathbb{D}[c \mid a, b]] + \mathbb{D}_a[\mathbb{E}[c \mid a, b]] = \mathbb{E}_a[ap_1 + bp_2] + \mathbb{D}_a[ap_1 + bp_2] = p_1\mathbb{E}a + p_2b + p_1^2\mathbb{D}a$$

Аналогично, $\mathbb{D}[c \mid b] - \mathbb{D}[c \mid a] = 0$ задает уравнение прямой, следовательно, множества линейно разделимы.



Слева множество точек (a, b), где $\mathbb{D}[c|b] < \mathbb{D}[c|a]$, справа, где неравенство обратное

Задание 5

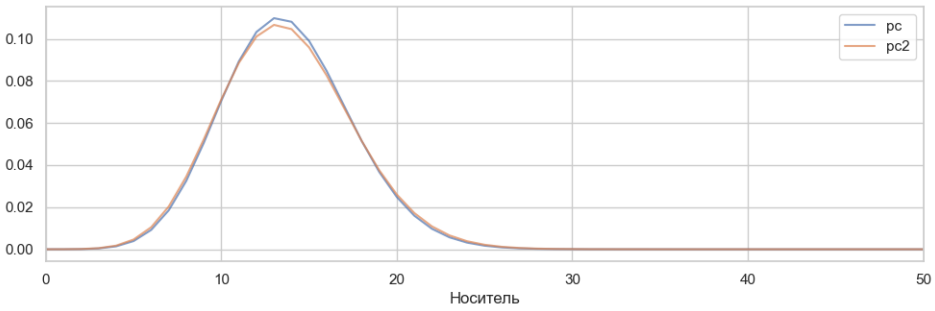
В качестве условий подавался весь носитель.

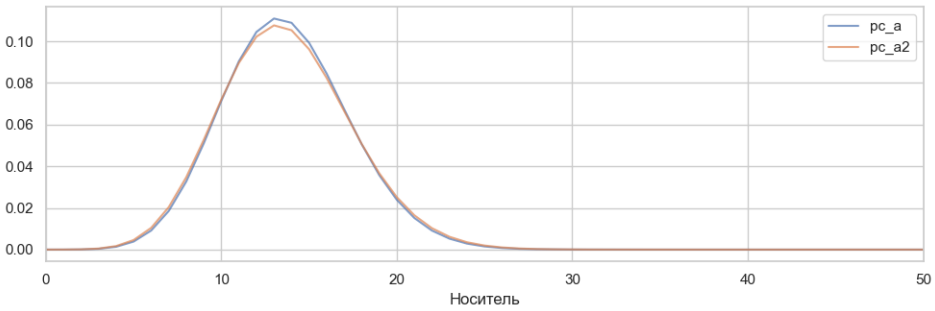
$p_1(c)$	$p_2(c)$	$p_1(c\|a)$	$p_2(c\|a)$	$p_1(c\|b)$	$p_2(c\|b)$	$p_1(c\|d)$	$p_2(c\|d)$
665 ms	71.4 ms	650 ms	61.7 ms	623 ms	61.2 ms	673 ms	120 ms

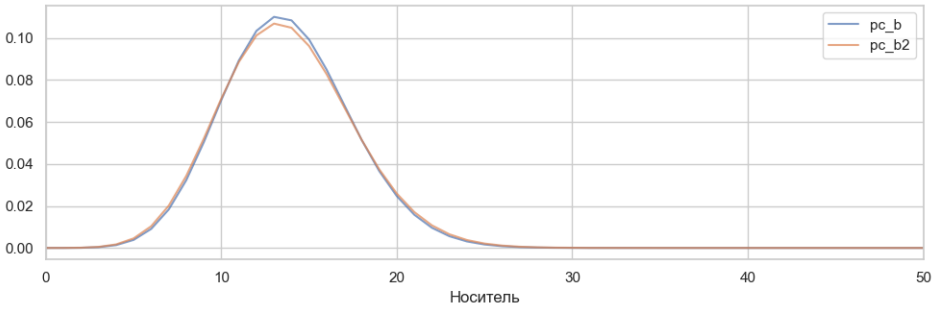
$p_1(c\|a,b)$	$p_2(c\|a,b)$	$p_1(c\|a,b,d)$	$p_2(c\|a,b,d)$	$p_1(d)$	$p_2(d)$
628 ms	68 ms	21.2 s	20.3 s	669 ms	116 ms

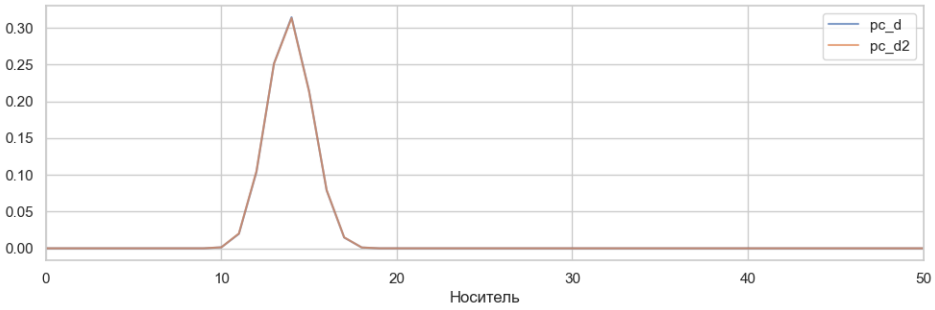
Задание 6

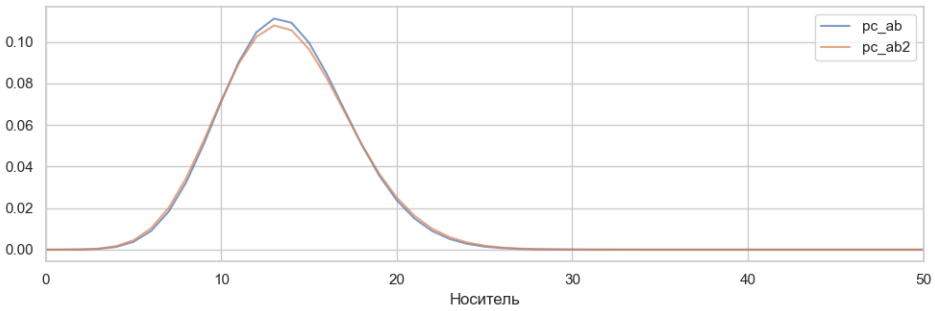
Основное отличие в моделях заключается в распределении $p(c \mid a, b)$. Хоть распределение Пуассона сохраняет ожидание, тем не менее дисперсия увеличивается. Вместо исходной модели с трудоемким вычислением распределения суммы биномиальных распределений, мы заменили ее на более удобную для вычислений аппроксимацию, но пожертвовали дисперсией. Распределения почти совпадают, но во второй модели получаются более "размазанными"(см. графики ниже). Также во второй модели не выполняется логичная гипотеза о том, что d вносит наибольший вклад в уточнении прогноза c в смысле меньшей дисперсии.

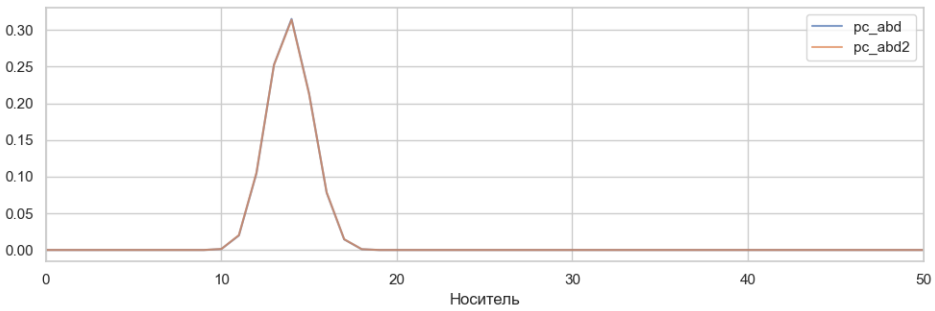












Сравнения распределений двух моделей