Практическая работа №1

Котов Артем, МОиАД ВШЭ СПб

2 ноября 2021 г.

Содержание

Параметры:

- $a \in [75, 90]$
- $b \in [500, 600]$
- $c \in [0, 690]$
- $d \in [1380]$
- $p_1 = 0.1$
- $p_2 = 0.01$
- $p_3 = 0.3$

Обычно, в суммах не указаны нижние и верхние пределы, что означает, что суммирование ведется по всем возможным значениям соответствующих величин.

Task 1

Условие: Вычисляем $p(a), p(b), p(c), p(d), p(c|a), p(c|b), p(c|d), p(c|ab), p(c|abd)$

Решение.

Замечание. $U[a,b] \sim \frac{1}{b-a+1}[a \leqslant x \leqslant b]$ $Bin(n,p) \sim \binom{n}{x} p^x (1-p)^{n-x}$

$$p(a) = \frac{1}{16}[75 \leqslant a \leqslant 90]$$
$$p(b) = \frac{1}{101}[500 \leqslant b \leqslant 600]$$

Для вычисления $p(c)$ нам потребуется $p(c|ab)$ (для второй модели в целом, аналогично, разве что свертка упростится):

$$p(c|ab) \sim \mathcal{B}(a, p_1) + \mathcal{B}(b, p_2) \sim \sum_i^c \binom{a}{i} p_1^i (1-p_1)^{a-i} \binom{b}{c-i} p_2^{c-i} (1-p_2)^{b-c+i},$$

тогда

$$p(c) = \sum_{ab} p(c|ab)p(a)p(b) = \frac{1}{1616} \sum_{a=a_{\min}}^{a_{\max}} \sum_{b=b_{\min}}^{b_{\max}} \sum_{i=0}^{c} \binom{a}{i} \binom{b}{c-i} p_1^i p_2^{c-i} (1-p_1)^{a-i} (1-p_2)^{b-c+i}.$$

Но это не удобно программировать, лучше оставить в виде сверток двух биномиальных:

$$p(c|ab) = \sum_{x=0}^{c} p(\mathcal{B}(a, p_1) = x) p(\mathcal{B}(b, p_2) = c - x)$$

$$p(c|a) = \sum_b p(c|ab)p(b) = \frac{1}{101}\sum_{b=500}^{600} p(c|ab)$$
$$p(c|b) = \sum_a p(c|ab)p(a) = \frac{1}{16}\sum_{a=75}^{90} p(c|ab)$$
$$p(c) = \sum_{ab} p(c|ab)\underbrace{p(a)p(b)}_{p(ab)}$$

Теперь с d:

$$p(d|c) = p(c + \mathcal{B}(c, p_3) = d) = p(\mathcal{B}(c, p_3) = d - c)$$
$$p(d) = \sum_c p(d|c)p(c)$$
$$p(c|d) = \frac{p(d|c)p(c)}{p(d)}$$
$$p(c|abd) = \frac{P(abcd)}{p(abd)} = \frac{p(d|c)p(c|ab)p(a)p(b)}{\sum_c p(d|c)p(c|ab)p(a)p(b)}$$

■

Task 2

Условие: Найти математические ожидания и дисперсии априорных распределений $p(a), p(b), p(c), p($

Решение.

Вообще все можем посчитать просто так:

$$\mathbb{E}x = \sum_x xp(x)$$
$$\mathbb{D}x = \sum_x x^2p(x) - (\mathbb{E}x)^2$$

Для a, b можем посчитать ручками:

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

Таблица 1: Мат. ожидания и дисперсии $p(a), p(b), p(c)$ и $p(d)$, округленные до 2-ух знаков после запятой.

№ Модели		$p(a)$	$p(b)$	$p(c)$	$p(d)$
1	$\mathbb{E}$	82.50	550.00	13.75	17.88
	$\mathbb{D}$	21.25	850.00	13.17	25.14
2	$\mathbb{E}$	82.50	550.00	13.75	17.87
	$\mathbb{D}$	21.25	850.00	14.05	26.63

Из таблицы видно, что мат. ожидания почти не чувствуют разницу в моделях, а вот дисперсии для c и d несколько отличаются. ■

Task 3

Условие: Пронаблюдать, как происходит уточнение прогноза для величины c по мере прихода новой косвенной информации.

Решение.

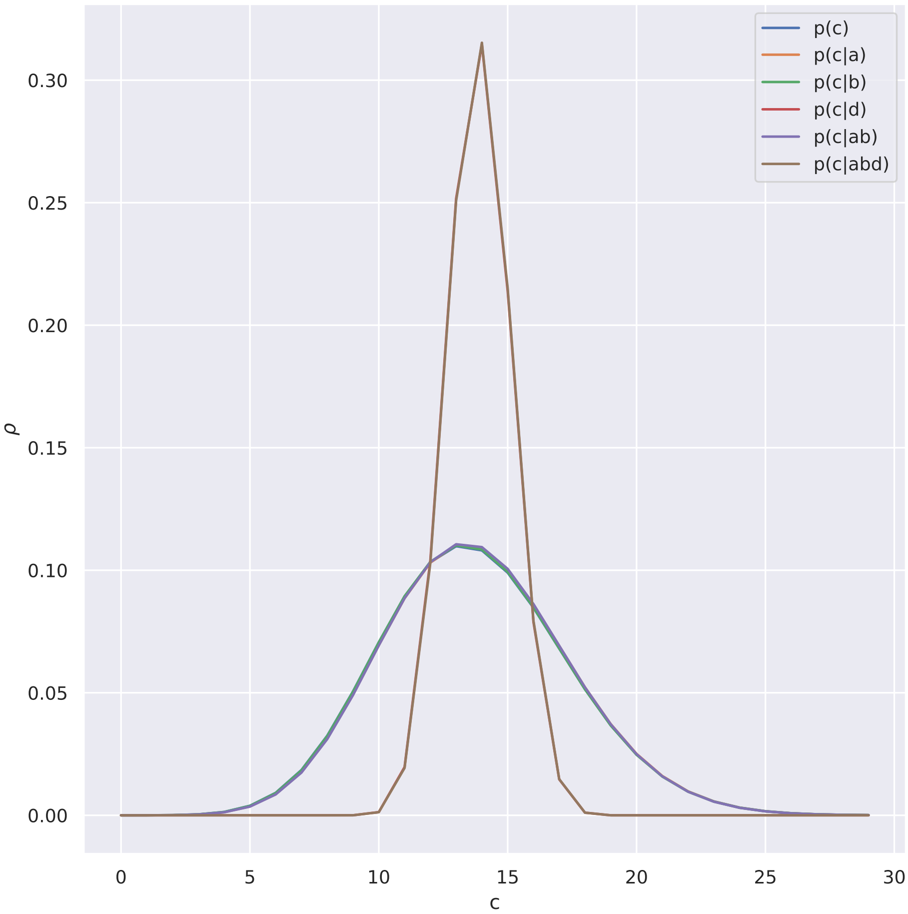

Рис. 1: График плотности распределений $p(c), p(c|a), p(c|b), p(c|d), p(c|ab), p(c|abd)$ для первой модели, где $c \in [0, 30]$, a, b, d равны своим математическим ожиданиям.

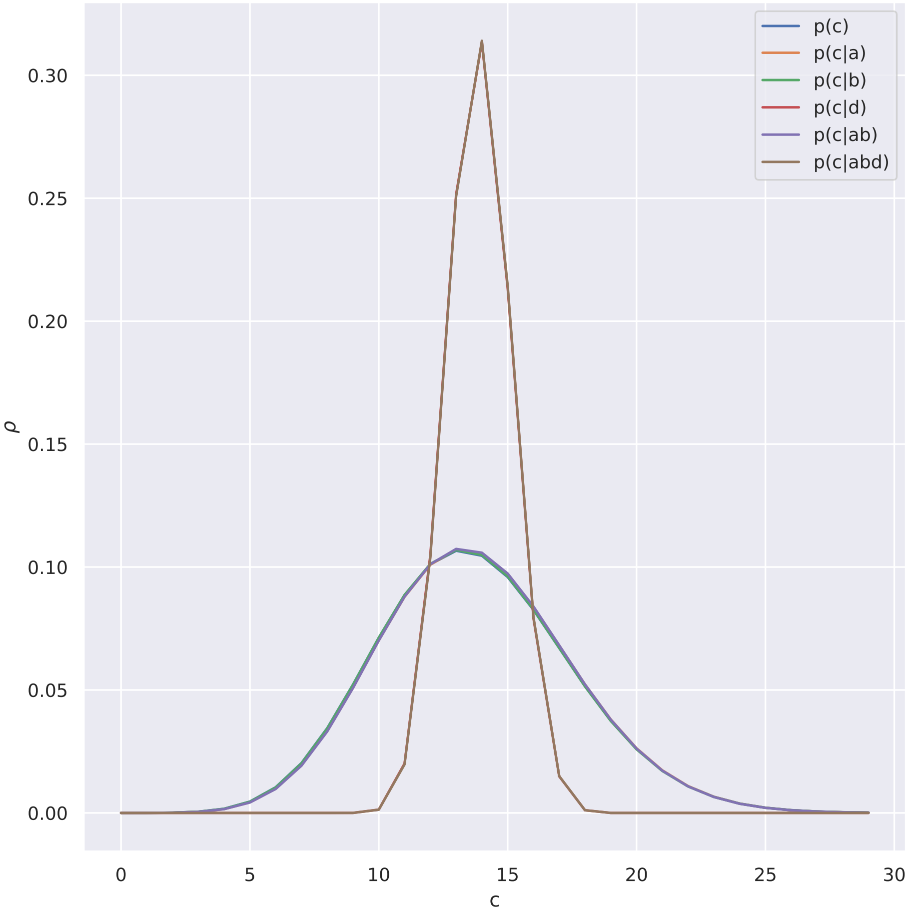

Рис. 2: График плотности распределений $p(c), p(c|a), p(c|b), p(c|d), p(c|ab), p(c|abd)$ для второй модели, где $c \in [0, 30]$, a, b, d равны своим математическим ожиданиям.

На графиках видно, что многие распределения совпали, так, например, $p(c|d)$ и $p(c|abd)$ крайне похожи для обоих моделей. Также видно, что знание о d существенно уменьшает дисперсию величины c, при это дополнительная информация о a, b уже существенно не влияет на распределение.

Таблица 2: Мат. ожидания и дисперсии $p(c), p(c|a), p(c|b), p(c|d), p(c|ab), p(c|abd)$, округленные до 2-ух знаков после запятой.

№ Модели		$p(c)$	$p(c\|a)$	$p(c\|b)$	$p(c\|d)$	$p(c\|ab)$	$p(c\|abd)$
1	$\mathbb{E}$	13.75	13.80	13.75	13.90	13.80	13.90
	$\mathbb{D}$	13.17	13.00	13.08	1.53	12.92	1.53
2	$\mathbb{E}$	13.75	13.80	13.75	13.89	13.80	13.90
	$\mathbb{D}$	14.05	13.88	13.96	1.54	13.80	1.54

Для этих распределений, в целом, аналогично, мат ожидания не чувствуют разницу в моделях, а дисперсии у второй модели систематически больше. ■

Task 4

Условие: Определить, какая из величин a, b, d вносит наибольший вклад в уточнение прогноза для величины c (в смысле дисперсии распределения).

Проведенный численный эксперимент показал, что для первой модели условия $\mathbb{D}[c|d] < \mathbb{D}[c|b]$ и $\mathbb{D}[c|d] < \mathbb{D}[c|a]$ выполняются для любых допустимых значений $a \in [75, 90]$, $b \in [500, 600]$ и $d \in [0, 1380]$. Однако для второй модели это оказывается неверным, к сожалению, аналитически показать это строго пока не удалось, т.е. может быть так, что это просто численная ошибка, но я, скорее, склоняюсь к тому, что это свойства модели.

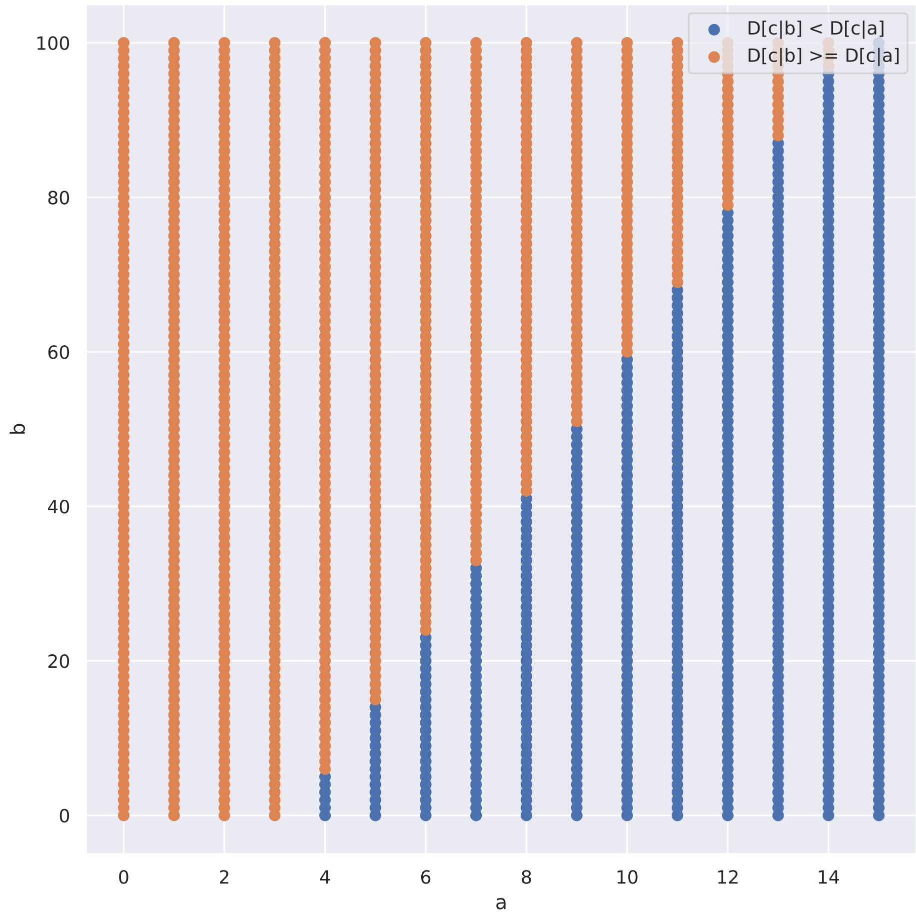

Рис. 3: График множества точек $(a, b) : \mathbb{D}[c|b] < D[c|a]$ (синий) и $(a, b) : \mathbb{D}[c|b] \geqslant D[c|a]$ (оранжевый) для первой модели.

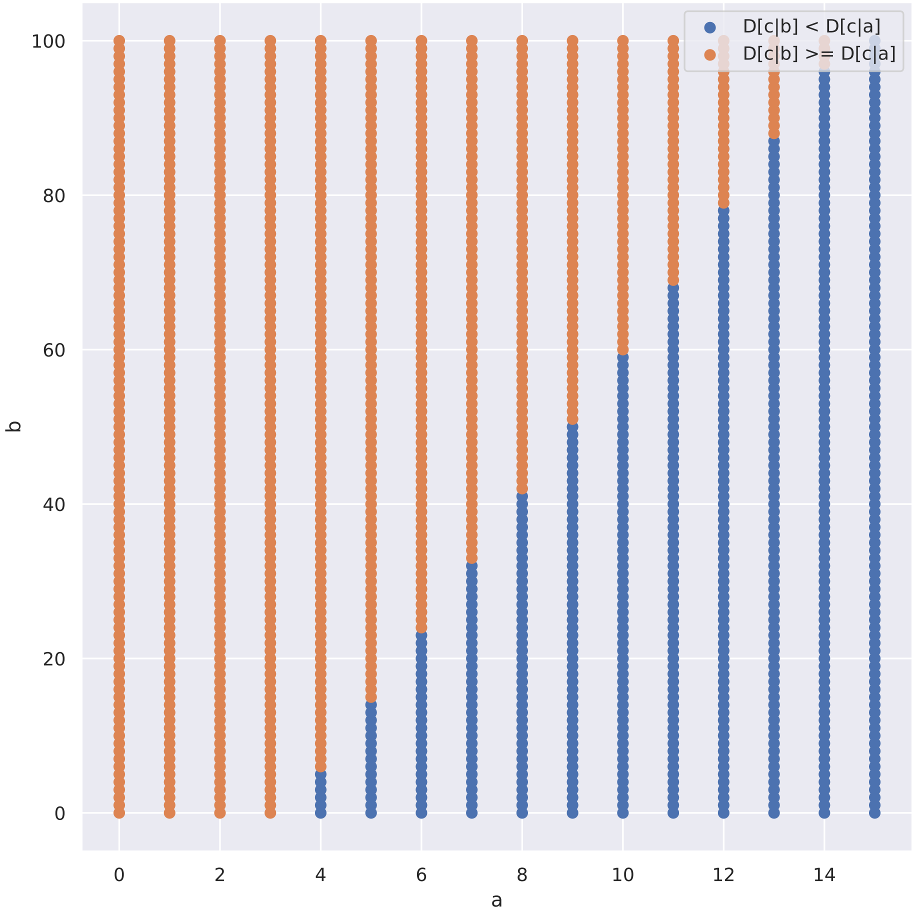

Рис. 4: График множества точек $\{(a, b) : \mathbb{D}[c|b] < D[c|a]\}$ (синий) и $\{(a, b) : \mathbb{D}[c|b] \geqslant D[c|a]\}$ (оранжевый) для второй модели.

В целом, из графиков можно сделать вывод, что эти множества таки линейно разделимы для обоих моделей.

Task 5

Условие: Провести временные замеры по оценке всех необходимых распределений $p(c), p(c|a), p(c|b)$

Решение.

Таблица 3: Временные замеры [мс] расчетов $p(c), p(c|a), p(c|b), p(c|d), p(c|ab), p(c|abd), p(d)$. Расчеты проведены векторно сразу для всех допустимых значений a, b, c, d.

№ Модели	$p(c)$	$p(c\|a)$	$p(c\|b)$	$p(c\|d)$	$p(c\|ab)$	$p(c\|abd)$	$p(d)$
1	64 ± 7	86 ± 22	95 ± 19	173 ± 11	89 ± 14	4630 ± 26	161 ± 6
2	83 ± 20	91 ± 18	111 ± 42	198 ± 47	63 ± 20	4890 ± 170	141 ± 4

■

Task 6

По большей степени можно выделить 2 существенных отличия:

1) Дисперсии величин $c, d, c|a, c|b, c|ab$ и $c|abd$ у второй модели систематически больше, чем у первой

2) Нарушаются условия $\mathbb{D}[c|d] < \mathbb{D}[c|b]$ и $\mathbb{D}[c|d] < \mathbb{D}[c|a]$ для второй модели, т.е. существуют такие a, b и d, что эти неравенства невыполнены.