

# Практическая работа №1

Котов Артем, МОиАД ВШЭ СПб

3 ноября 2021 г.

## Содержание

|        |   |
|--------|---|
| Task 1 | 2 |
| Task 2 | 3 |
| Task 3 | 5 |
| Task 4 | 7 |
| Task 5 | 9 |
| Task 6 | 9 |

Параметры:

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

Обозначения:

- Обычно, в суммах не указаны нижние и верхние пределы, что означает, что суммирование ведется по всем возможным значениям соответствующих величин.
- $\mathcal{B}$  — биномиальное распределение.
- $\mathcal{P}$  — Пуассоновское распределение.

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

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

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

$$p_{\text{model1}}(c|ab) \sim \mathcal{B}(a, p_1) + \mathcal{B}(b, p_2) \sim \sum_i^c \binom{a}{i} p_1^i (1-p_1)^{a-i} \binom{b}{c-i} p_2^{c-i} (1-p_2)^{b-c+i}$$
$$p_{\text{model2}}(c|ab) \sim \mathcal{P}(ap_1 + bp_2)$$

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  $p(a), p(b), p(c), p(d)$ .

*Решение.*

Вообще все можем посчитать просто так:

$$\mathbb{E}x = \sum_x xp(x)$$

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

Для  $a, b$  можем посчитать ручками:

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

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

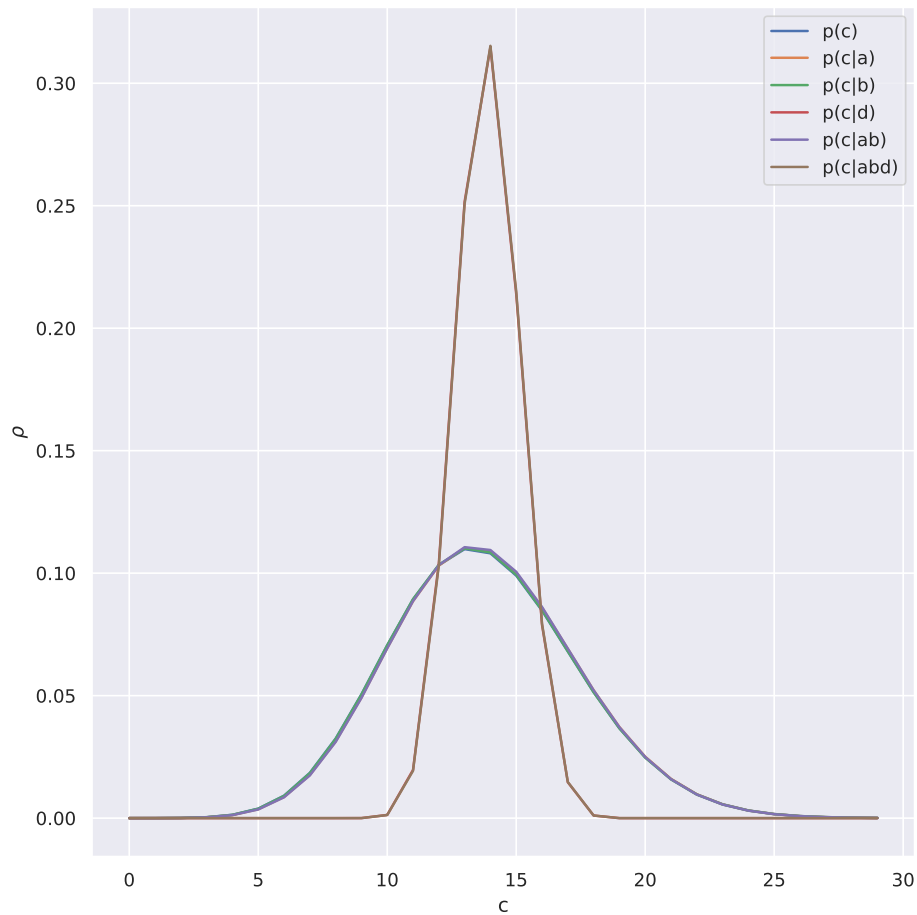
| № Модели |   | $p(a)$ | $p(b)$ | $p(c)$ | $p(d)$ |
|----------|---|--------|--------|--------|--------|
| 1        | Е | 82.50  | 550.00 | 13.75  | 17.88  |
|          | Д | 21.25  | 850.00 | 13.17  | 25.14  |
| 2        | Е | 82.50  | 550.00 | 13.75  | 17.87  |
|          | Д | 21.25  | 850.00 | 14.05  | 26.63  |

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

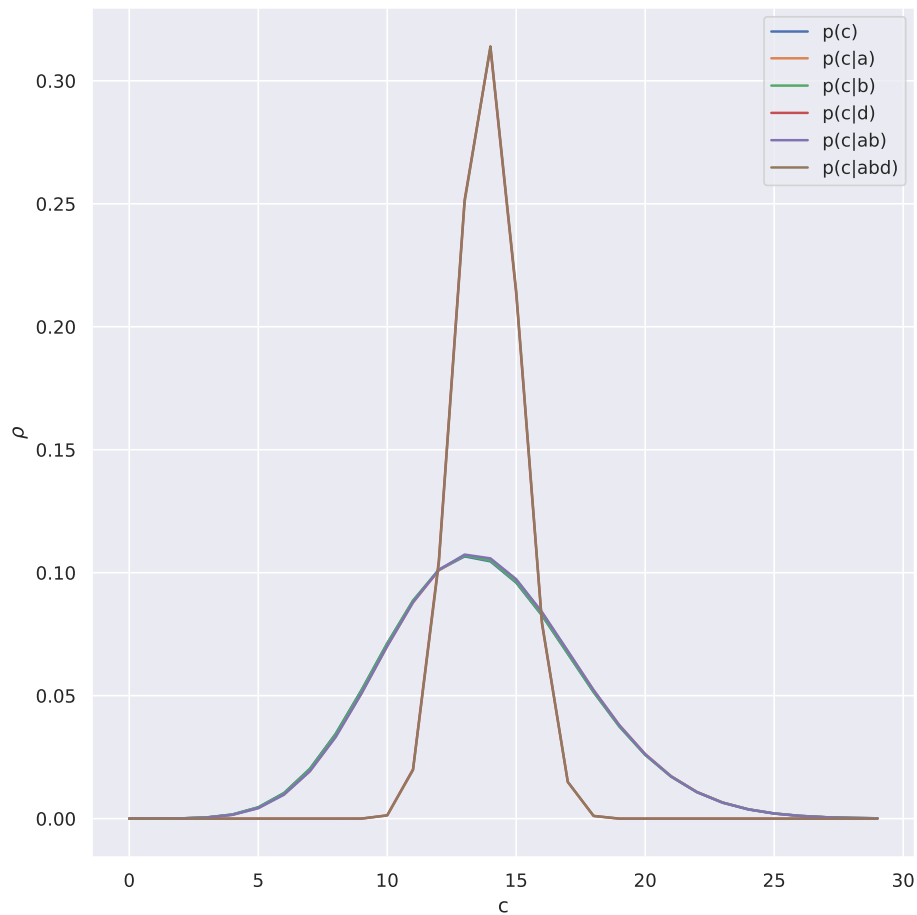
## Task 3

**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 обеих 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.

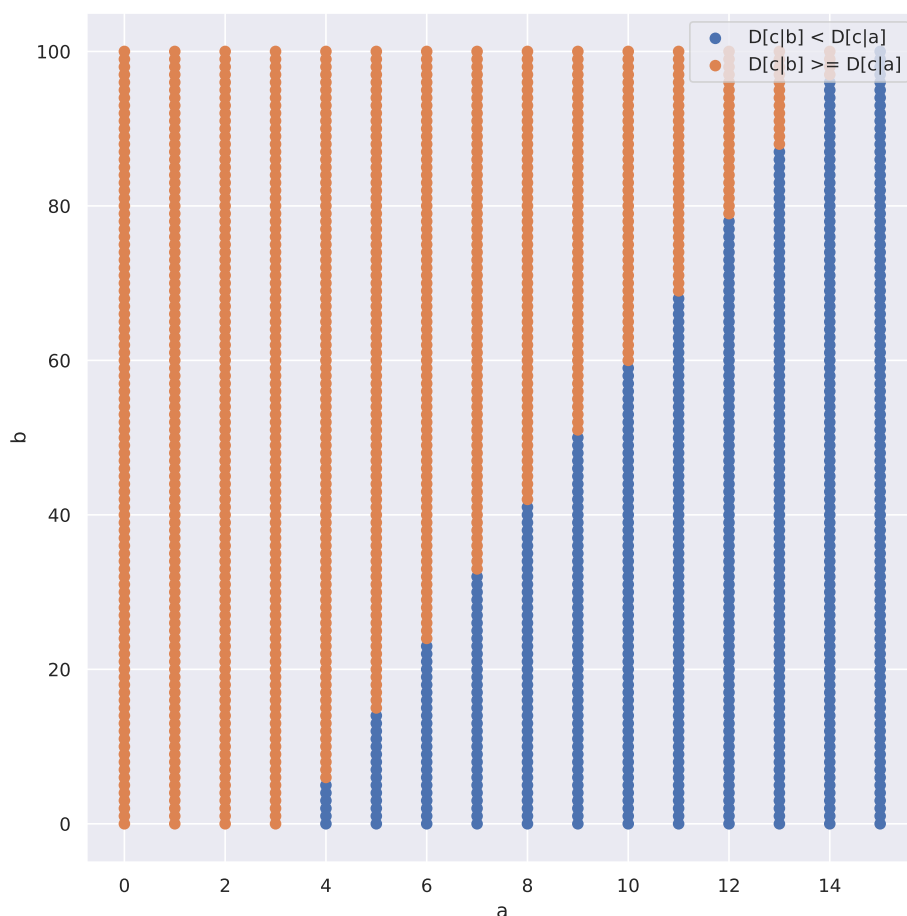
| № Модели |   | $p(c)$ | $p(c a)$ | $p(c b)$ | $p(c d)$ | $p(c ab)$ | $p(c abd)$ |
|----------|---|--------|----------|----------|----------|-----------|------------|
| 1        | Е | 13.75  | 13.80    | 13.75    | 13.90    | 13.80     | 13.90      |
|          | Д | 13.17  | 13.00    | 13.08    | 1.53     | 12.92     | 1.53       |
| 2        | Е | 13.75  | 13.80    | 13.75    | 13.89    | 13.80     | 13.90      |
|          | Д | 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. ■

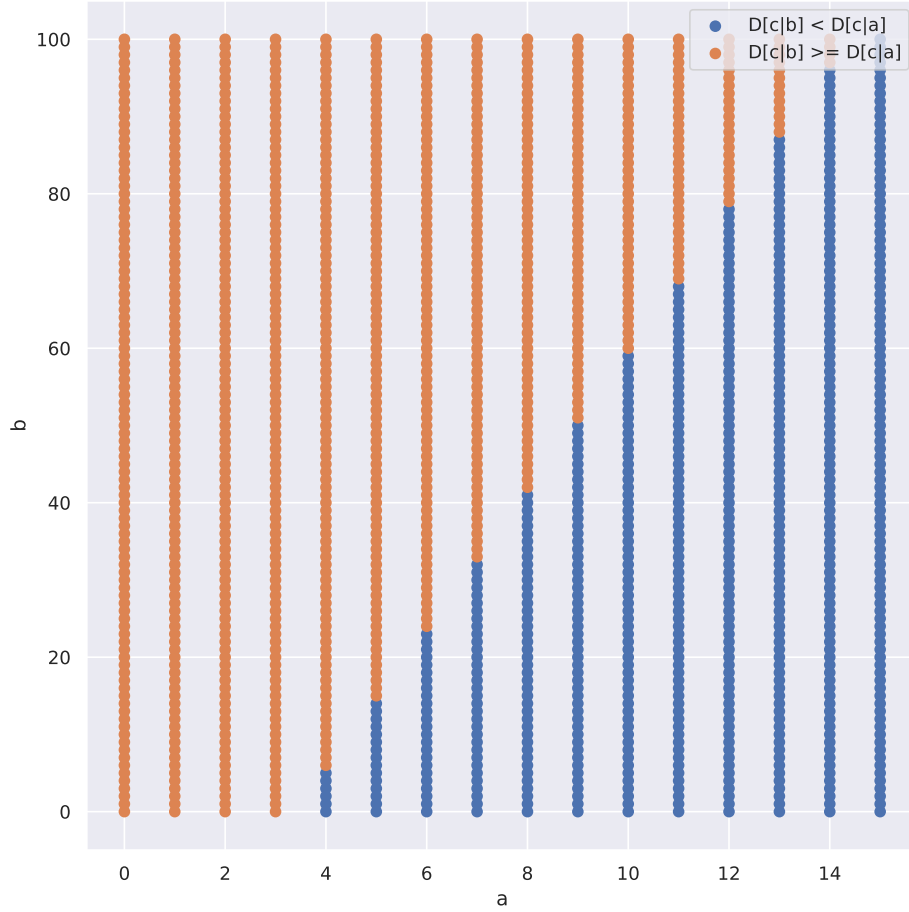
## Task 4

**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  $(a, b) : \mathbb{D}[c|b] \geq D[c|a]$  (оранжевый) для первой модели.



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

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

А теперь попробуем привести показательство: рассмотрим  $p(c|ab)$  и заметим, что  $a|b = a$ , т.к.  $a$  и  $b$  независимые.

Введем  $y = c|b$  и  $x = a|b (= a)$ , тогда  $c|ab = y|x$

**Замечание.** Не важно, в каком порядке обуславливать на  $a$  и  $b$ : сначала  $a$ , а потом  $b$  или наоборот, т.к.  $a$  и  $b$  независимые.

По формуле полной вариации:

$$\begin{aligned}
 \mathbb{D}y &= \mathbb{E}(\mathbb{D}[y|x]) + \mathbb{D}(\mathbb{E}[y|x]) \\
 \mathbb{D}[y|x] &= \mathbb{D}[c|ab] = \mathbb{D}[\mathcal{B}(a, p_1) + \mathcal{B}(b, p_2)] = ap_1(1 - p_1) + bp_2(1 - p_2) \\
 \mathbb{E}_a \mathbb{D}[y|x] &= \mathbb{E}a \cdot p_1(1 - p_1) + \underbrace{b}_{\text{т.к. } b \text{ не зависит от } a} p_2(1 - p_2) \\
 \mathbb{E}[y|x] &= \mathbb{E}[\mathcal{B}(a, p_1) + \mathcal{B}(b, p_2)] = ap_1 + bp_2 \\
 \mathbb{D}_a(\mathbb{E}[y|x]) &= \mathbb{D}ap_1^2,
 \end{aligned}$$



тогда получаем, что

$$\mathbb{D}[c|b] = \mathbb{E}a \cdot p_1(1 - p_1) + bp_2(1 - p_2) + \mathbb{D}a \cdot p_1^2.$$

Аналогично

$$\mathbb{D}[c|a] = \mathbb{E}b \cdot p_2(1 - p_2) + ap_1(1 - p_1) + \mathbb{D}b \cdot p_2^2.$$

Рассмотрим множество точек, где  $\mathbb{D}[c|a] = \mathbb{D}[c|b]$ , т.е.

$$bp_2(1 - p_2) = C + ap_1(1 - p_1),$$

где  $C$  – некая константа, определяемая через  $p_1, p_2, \mathbb{E}b, \mathbb{E}a$ . Заметим, что это уравнение есть уравнение прямой в плоскости  $(a, b)$ :  $b = ka + c$ , т.е. множество точек, где  $\mathbb{D}[c|a] = \mathbb{D}[c|b]$  есть прямая в плоскости  $(a, b)$ , т.е. множества действительно линейно делимы.

Для модели Пуассона все аналогично, разве что для  $c|ab$  мы получим  $\mathcal{P}(ap_1 + bp_2)$ , что в итоге даст:

$$\mathbb{D}[c|a] = \mathbb{E}ap_1 + bp_2 + \mathbb{D}ap_1$$

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

что приведет снова к уравнению прямой в плоскости  $(a, b)$ .

## 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        | $64 \pm 7$  | $86 \pm 22$ | $95 \pm 19$  | $173 \pm 11$ | $89 \pm 14$ | $4630 \pm 26$  | $161 \pm 6$ |
| 2        | $83 \pm 20$ | $91 \pm 18$ | $111 \pm 42$ | $198 \pm 47$ | $63 \pm 20$ | $4890 \pm 170$ | $141 \pm 4$ |



## 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.