Московский государственный университет им. М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования

Задание №1.
Вероятностные модели посещаемости курса

Автор: Арбузова Дарья
Группа: 417

Москва, 2014

Содержание

1 Постановка задачи

Рассмотрим модель посещаемости студентами одного курса лекции. Пусть аудитория данного курса состоит из студентов профильной кафедры, а также студентов других кафедр. Обозначим

- a — количество студентов, распределившихся на профильную кафедру;
- b — количество студентов других кафедр на курсе;
- c — количество студентов на данной лекции;
- d — общее количество записавшихся на данной лекции;
- p_1 — вероятность посещения лекции студентом профильной кафедры;
- p_2 — вероятность посещения лекции студентом любой из остальных кафедр;
- p_3 — вероятность, с которой студент записывает своего товарища.

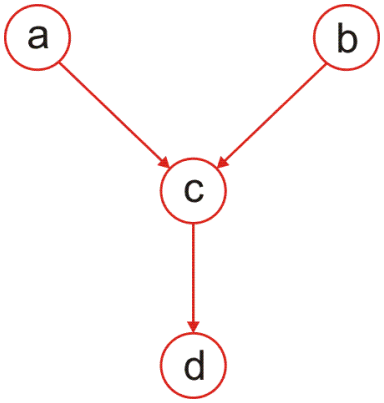


Рис. 1: Графическая модель

В модели 1:

- $a \sim R[a_{min}; a_{max}]$
- $b \sim R[b_{min}; b_{max}]$
- $c|a, b \sim B(a, p_1) + B(b, p_2)$
- $d|c \sim c + B(c, p_3)$

В модели 2:

- $a \sim R[a_{min}; a_{max}]$
- $b \sim R[b_{min}; b_{max}]$
- $c|a, b \sim Poiss(ap_1 + bp_2)$
- $d|c \sim c + B(c, p_3)$

Значения параметров: $a_{min} = 15, a_{max} = 30, b_{min} = 250, b_{max} = 350, p_1 = 0.5, p_2 = 0.05, p_3 = 0.5$.

В данном задании необходимо исследовать поведение моделей при различных параметрах, а также оценить плотность некоторых распределений с приходом новой информации.

2 Вывод формул для расчёта распределений

Проведём вывод необходимых распределений для рассматриваемых моделей, пользуясь фактами из теории вероятностей:

1. $a \sim R[a_{min}; a_{max}]$

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

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

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

2. $b \sim R[b_{min}; b_{max}]$ аналогично:

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

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

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

3. $b|a$

$$p(b|a) = \frac{p(a,b)}{p(a)} = \frac{\sum\limits_{c=0}^{a+b}\sum\limits_{d=c}^{2c} p(a,b,c,d)}{p(a)} = p(b) \cdot \sum_{c=0}^{a+b} p(c|a,b) \cdot \sum_{d=c}^{2c} p(d|c) = p(b)$$

Величины a и b независимы в рамках рассматриваемых моделей.

4. (a) Модель 1: $c|a,b \sim B(a,p_1) + B(b,p_2)$
Пусть $c = x + y$, где $x \sim B(a,p_1), y \sim B(b,p_2)$, тогда

$$p(c|a,b) = \sum_{k=0}^{c} p(x=k;a,p_1) \cdot p(y=c-k;b,p_2) =$$

$$= \sum_{k=0}^{c} C_a^k p_1^k (1-p_1)^{a-k} \cdot C_b^{c-k} p_2^{c-k} (1-p_2)^{b+k-c}$$

(b) Модель 2: $c|a,b \sim Poiss(ap_1 + bp_2)$

$$p(c|a,b) = e^{\lambda}\frac{\lambda^c}{c!},$$

где $\lambda = ap_1 + bp_2$.

5. $d|c \sim c + B(c,p_3)$

$$p(d|c) = C_c^{d-c} p_3^{d-c} (1-p_3)^{2c-d}$$

6. $c|a$

$$p(c|a) = \frac{p(a,c)}{p(a)} = \frac{\sum\limits_{b=b_{min}}^{b_{max}}\sum\limits_{d=0}^{2(a+b)} p(a,b,c,d)}{p(a)} =$$

$$= p(b) \cdot \sum_{b=b_{min}}^{b_{max}} p(c|a,b) \cdot \sum_{d=0}^{2(a+b)} p(d|c) = p(b) \cdot \sum_{b=b_{min}}^{b_{max}} p(c|a,b)$$

7. $c|b$ аналогично:

$$p(c|b) = p(a) \cdot \sum_{a=a_{min}}^{a_{max}} p(c|a,b)$$

8. c

$$p(c) = \sum_{a=a_{min}}^{a_{max}}\sum_{b=b_{min}}^{b_{max}}\sum_{d=0}^{2(a+b)} p(a,b,c,d) =$$

$$= p(a) \cdot p(b) \cdot \sum_{a=a_{min}}^{a_{max}}\sum_{b=b_{min}}^{b_{max}} p(c|a,b) \cdot \sum_{d=0}^{2(a+b)} p(d|c) = p(a) \cdot p(b) \cdot \sum_{a=a_{min}}^{a_{max}}\sum_{b=b_{min}}^{b_{max}} p(c|a,b)$$

9. d

$$p(d) = \sum_{c=0}^{a_{max}+b_{max}} p(d|c)p(c)$$

3 Априорные распределения

Требуется рассчитать математические ожидания и дисперсии априорных распределений a, b, c и d.

Пусть для некоторый случайной величины x известно её распределение, тогда

$$\mathbb{E}[x] = \sum_{x=x_{min}}^{x_{max}} xp(x)$$

$$\mathbb{D}[x] = \sum_{x=x_{min}}^{x_{max}} x^2 p(x) - (\mathbb{E}[x])^2$$

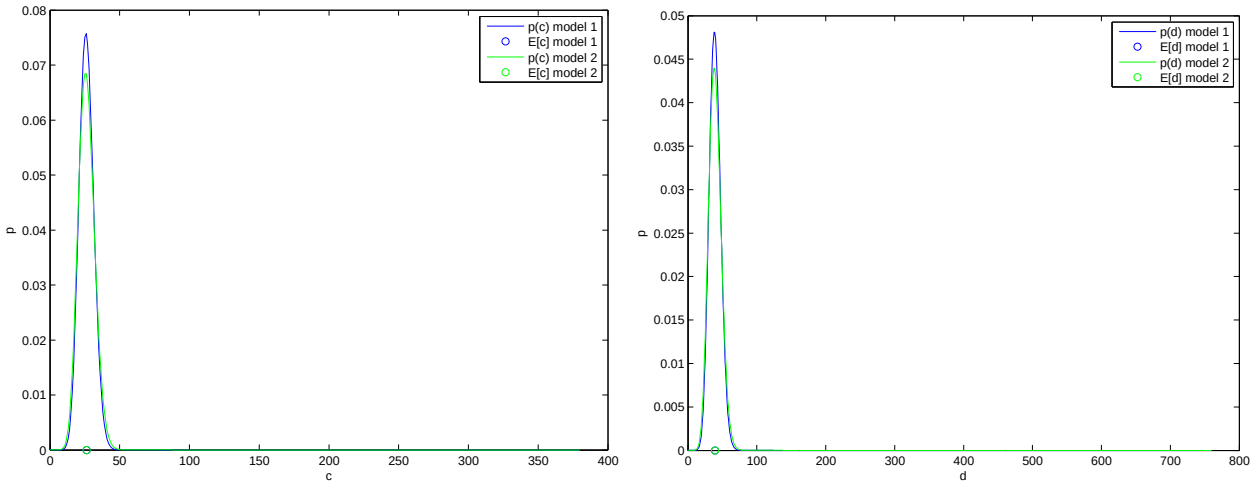
В пункте 2 было показано, как получить априорные распределения, имея $p(c|a,b)$ и $p(d|c)$.

Результаты приведены в таблице 1 (прочерк означает одинаковое поведение в обеих моделях):

Величина	Модель 1		Модель 2	
	$\mathbb{E}$	$\mathbb{D}$	$\mathbb{E}$	$\mathbb{D}$
a	22.5	21.25	-	-
b	300	850	-	-
c	26.25	27.3125	-	33.6875
d	39.375	68.0156	-	82.3594

Таблица 1: Априорные распределения

Приведём вид исследованных распределений на графиках ниже (см. рис. 2):



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

Рис. 2: Распределения $p(c)$ и $p(d)$ для моделей 1 и 2

4 Прогноз величины b

Требуется пронаблюдать, как происходит уточнение прогноза для величины b с приходом новой информации.

Рассмотрим распределения $b, b|a, b|a, d$. Как было показано выше в пункте 2, b и $b|a$ распределены одинаково, поэтому будем сравнивать только b с $b|a, d$.

На рисунке 3 показано распределение соответствующих величин для моделей 1 и 2, а также указаны их матожидания.

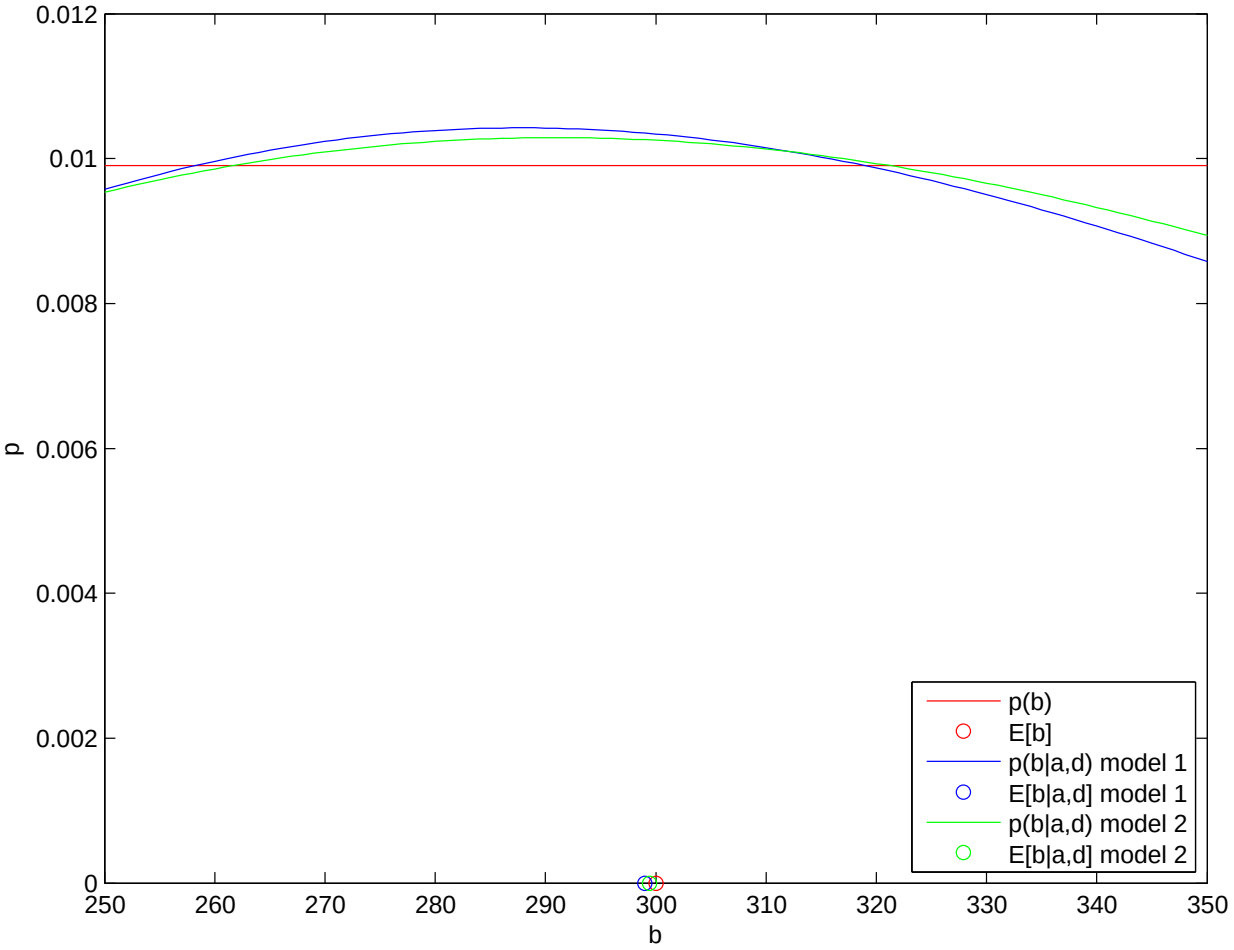
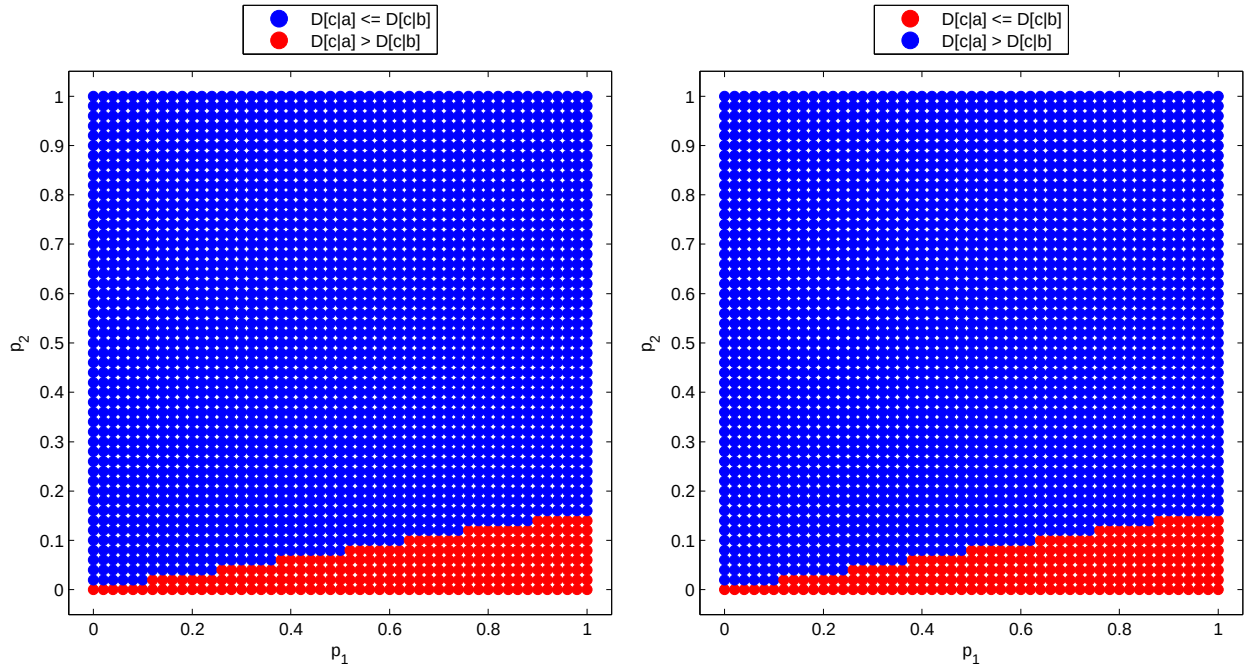


Рис. 3: Уточнение прогноза для величины b с приходом новой информации для моделей 1 и 2

5 Влияние параметров p_1 и p_2

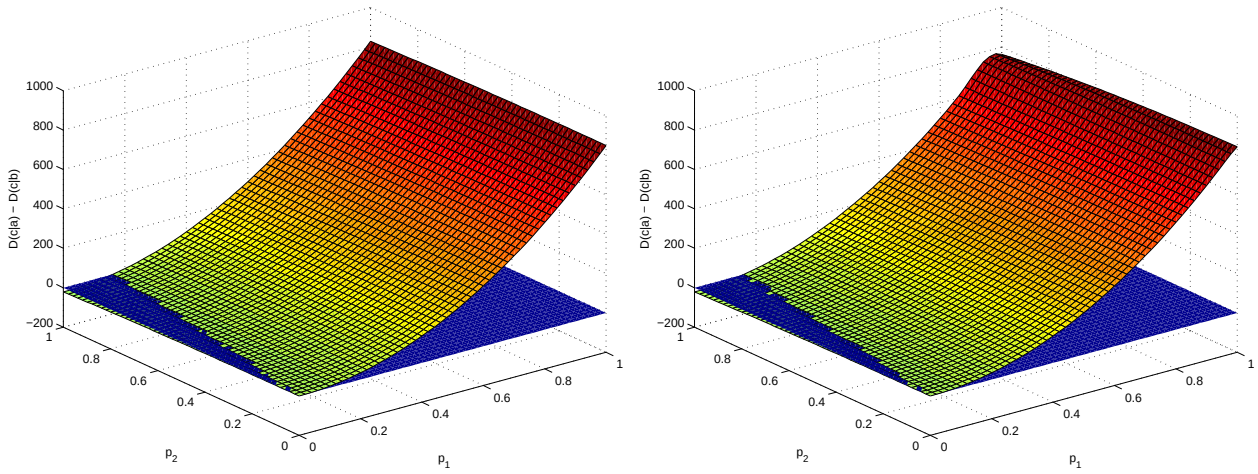
Требуется исследовать, каким образом единичный квадрат разбивается на области $\{(p_1, p_2)|\mathbb{D}[c|a] \leqslant \mathbb{D}[c|b]\}$ и $\{(p_1, p_2)|\mathbb{D}[c|a] > \mathbb{D}[c|b]\}$. Результаты для моделей 1 и 2 представлены на рисунках 4a и 4b соответственно.



(a) Модель 1 (b) Модель 2

Рис. 4: Соотношение между $\mathbb{D}[c|a]$ и $\mathbb{D}[c|b]$ в зависимости от p_1 и p_2 для моделей 1 и 2

Для большей наглядности рассмотрим трёхмерные графики зависимости $\mathbb{D}[c|a] - \mathbb{D}[c|b]$ от параметров p_1 и p_2 (рисунки 5a и 5b для моделей 1 и 2 соответственно). Синим обозначена плоскость, соответствующая равенству дисперсий.



(a) Модель 1 (b) Модель 2

Рис. 5: Зависимость $\mathbb{D}[c|a] - \mathbb{D}[c|b]$ от p_1 и p_2 для моделей 1 и 2

Очевидно, что области $\{(p_1, p_2)|\mathbb{D}[c|a] \leqslant \mathbb{D}[c|b]\}$ и $\{(p_1, p_2)|\mathbb{D}[c|a] > \mathbb{D}[c|b]\}$ в модели 2 не являются линейно разделимыми. Результаты модели 1 не являются достаточно наглядными и требуют большего анализа.

Найдём 3 точки на границе областей и проверим, лежат ли они на одной прямой. Для значений $p_1 = 0.05, 0.5$ и 0.95 вычислим соответствующие p_2 с помощью бинарного поиска с критерием останова $|\mathbb{D}[c|a] - \mathbb{D}[c|b]| < 10^{-7}$. В таблице 2 представлены координаты точек и значение разности дисперсий в них.

Точка	p_1	p_2	$\mathbb{D}[c\|a] - \mathbb{D}[c\|b], 10^{-8}$
z_1	0.05	0.0095100594	4.46
z_2	0.5	0.0799816155	0.12
z_3	0.95	0.1503011680	-7.21

Таблица 2: Граничные точки

Составим 2 вектора, $v_1 = z_2 - z_1$ и $v_2 = z_3 - z_2$, и проверим их на коллинеарность. Получим $v_1 = (0.45, 0.07047), v_2 = (0.45, 0.07032)$. Поскольку координаты граничных точек вычислены с высокой степенью точности, то различие координат векторов в 4-м знаке даёт повод утверждать, что они неколлинеарны, следовательно, кривая, разделяющая области, не является прямой.

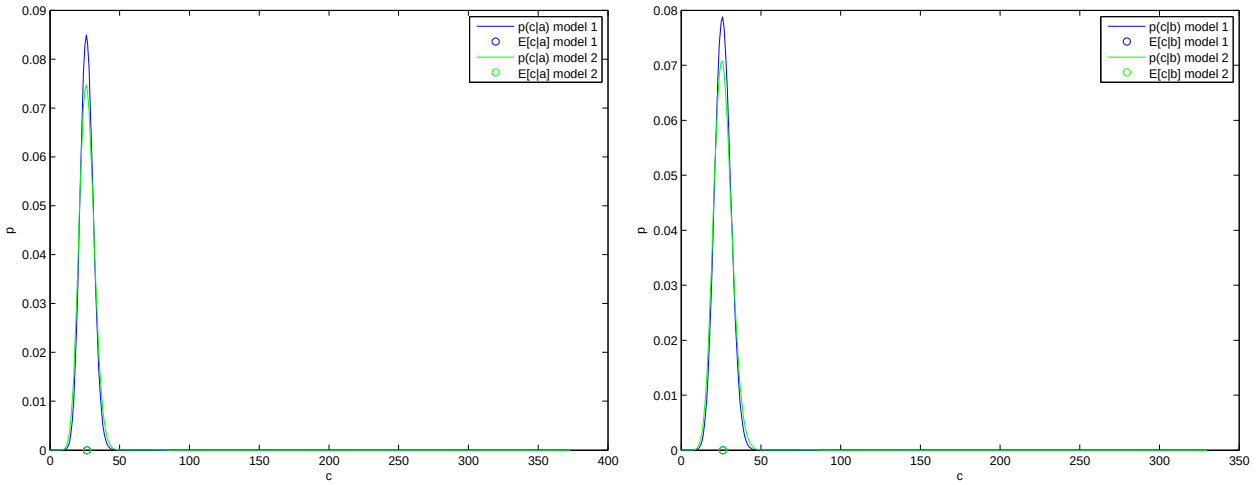
6 Временные замеры

Требуется рассчитать распределения $p(c), p(c|a), p(c|b), p(b|a), p(b|a,\ d), p(d)$. Значения параметров a, b и d положены равными их матожиданию ($a = 23, b = 300, d = 39$).

Замеры производятся для полного набора выходных аргументов функций. Результаты представлены в таблице 3.

Распределение	Время, с	
	Модель 1	Модель 2
$p(c)$	0.183	0.022
$p(c\|a)$	0.107	0.013
$p(c\|b)$	0.038	0.007
$p(b\|a)$	0.001	-
$p(b\|a,\ d)$	0.106	0.028
$p(d)$	0.259	0.101

Таблица 3: Временные замеры



(a) $p(c|a)$ (b) $p(c|b)$

Рис. 6: Распределения $p(c|a)$ и $p(c|b)$ для моделей 1 и 2

7 Сравнение моделей 1 и 2

При больших значениях параметров p_1 и p_2 и малом числе испытаний (параметрах a, b) пуассоновское распределение недостаточно хорошо приближает сумму биномиальных. В этом можно убедиться, положив значения параметров $a_{min} = 1, a_{max} = 5, b_{min} = 10, b_{max} = 15, p_1 = 0.9$, см. рис. 7:

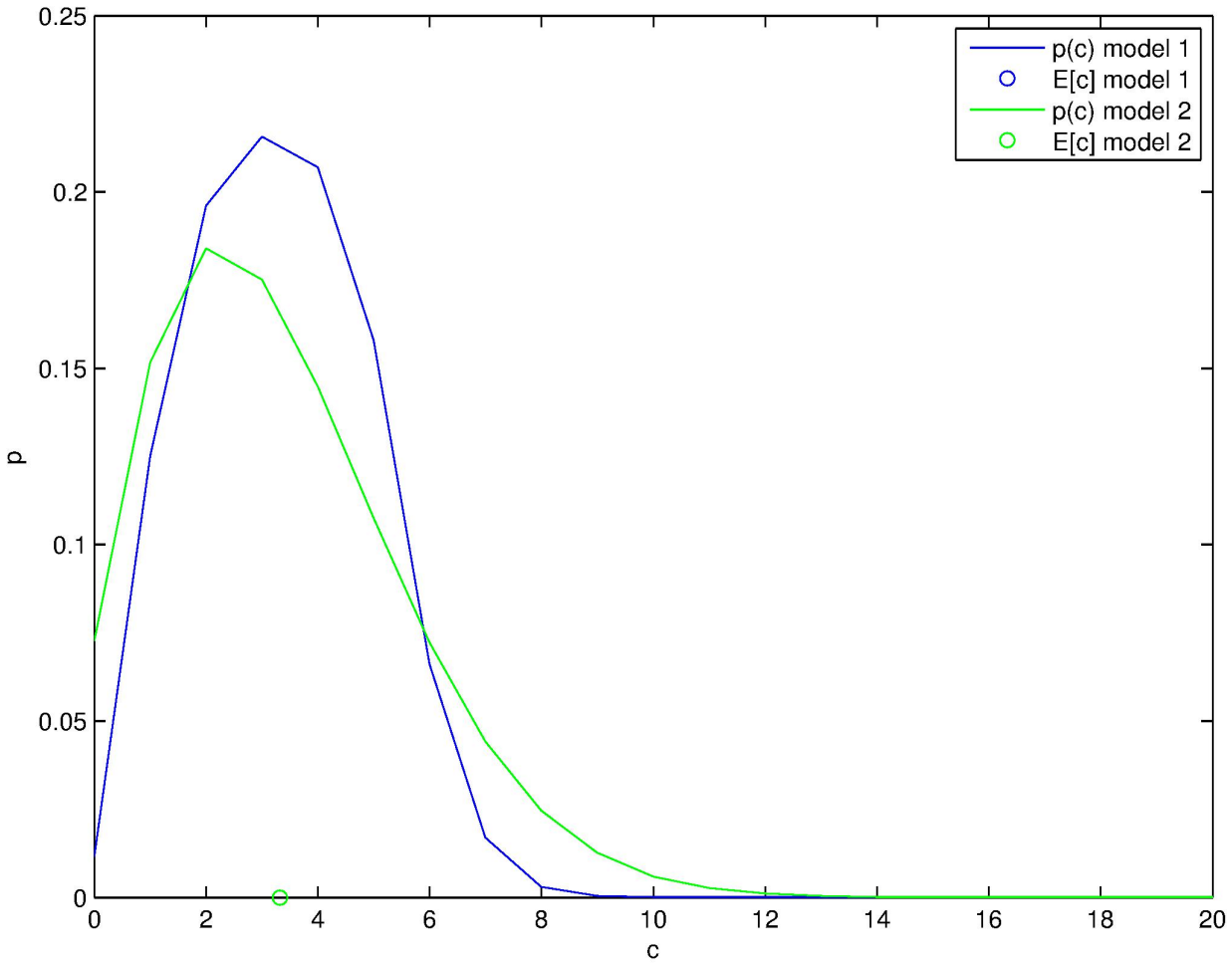


Рис. 7: Распределение $p(c)$ для моделей 1 и 2