

# Домашнее задание 1

Юрасов Никита Андреевич

Обновлено 31 октября 2019 г.

## Содержание

<b>1</b>	<b>Описание основных характеристик</b>	<b>2</b>
1.1	Распределение Пуассона . . . . .	2
1.2	Распределение Эрланга . . . . .	4
<b>2</b>	<b>Поиск примеров событий</b>	<b>6</b>
2.1	Распределение Пуассона . . . . .	6
2.2	Распределение Эрланга . . . . .	8
<b>3</b>	<b>Моделирование</b>	<b>10</b>
3.1	Распределение Пуассона . . . . .	10
3.2	Распределение Эрланга . . . . .	11

Все графики, которые в дальнейшем будут вставлены в эту работу, были сконструированы с помощью библиотеки `matplotlib` в Jupyter Notebook, который будет приложен вместе с работой (ДЗ1.ipynb).

Все формулировки определений были взяты из книги А. В. Иванова «Теория Вероятностей (Краткий курс)»

Для выполнения работы были выбраны два распределения:

1. Дискретное распределение: распределение Пуассона
2. Непрерывное распределение: распределение Эрланга

## 1 Описание основных характеристик

### 1.1 Распределение Пуассона

Пусть случайная величина задается законом распределением:

$$f(x) = \frac{\mu^x}{x!} e^{-\mu}, \quad \mu > 0, x \in \mathbb{N} \quad (1)$$

Функция распределения:

$$F_\xi(x) = P(\xi < x) = \sum_{k=0}^{x-1} \frac{\mu^k}{k!} e^{-\mu}, \quad \mu > 0, x \in \mathbb{N} \quad (2)$$

### Математическое ожидание

**Определение 1.1.** Математическое ожидание неотрицательной дискретной случайной величины называется:

$$E\xi = \sum_i a_i p_i = \sum_i a_i P(\xi = a_i)$$

Тогда, из этого определения можно вывести математическое ожидание распределения Пуассона (1):

$$E\xi = \sum_{k=0}^{\infty} k \frac{\mu^k}{k!} e^{-\mu} = e^{-\mu} \sum_{k=1}^{\infty} \frac{\mu^k}{(k-1)!} = \mu e^{-\mu} \sum_{k=1}^{\infty} \frac{\mu^{k-1}}{(k-1)!} = \mu e^{-\mu} \sum_{m=0}^{\infty} \frac{\mu^m}{m!} = \mu e^{-\mu} e^{\mu} = \mu$$

### Дисперсия

**Определение 1.2.** Дисперсия – это такое число, которое выражается вторым центральным моментом

$$D\xi = E(\xi - E\xi)^2 \quad (3)$$

Чтобы посчитать дисперсию случайной величины  $\xi$  проще всего воспользоваться выводом факториальных моментов.

**Определение 1.3.** Факториальный момент – это число  $E\xi^{[m]} = E\xi(\xi - 1) \dots (\xi - m + 1)$

Тогда:

$$E\xi(\xi - 1) = \sum_{k=0}^{\infty} k(k-1) \frac{\mu^k}{k!} e^{-\mu} = \mu^2 e^{-\mu} \sum_{k=2}^{\infty} \frac{\mu^{k-2}}{(k-2)!} = \mu^2 e^{-\mu} e^{\mu} = \mu^2$$

Если  $E\xi^2 = E\xi(\xi - 1) + E\xi = \mu^2 + \mu$ , а  $D\xi = E\xi^2 - (E\xi)^2$ , то  $D\xi = \mu^2 + \mu - \mu^2 = \mu$

В итоге мы получили для распределения Пуассона  $E\xi = \mu, D\xi = \mu$

## Производящая и характеристическая функции

**Определение 1.4.** Производящей функцией случайной величины  $\xi$  называется функция

$$\varphi_{\xi}(z) = Ez^{\xi} = \sum_{k=0}^{\infty} z^k p_k, \quad z \in \mathbb{C}, \quad |z| \leq 1$$

**Определение 1.5.** Характеристической функцией произвольной случайной величины  $\xi$  называется функция действительного аргумента при дискретном распределении:

$$g(t) = \sum_{k=0}^{\infty} e^{itk} P(\xi = k)$$

Производящая функция для распределения Пуассона:

$$\varphi_{\xi}(x) = \sum_{k=0}^{\infty} \frac{\mu^k}{k!} e^{-\mu} z^k = e^{-\mu} \sum_{k=0}^{\infty} \frac{(\mu z)^k}{k!} = e^{-\mu} e^{\mu z} = e^{\mu(z-1)}$$

Характеристическая функция:

$$g(t) = \sum_{k=0}^{\infty} e^{itk} \frac{\mu^k}{k!} e^{-\mu} = e^{-\mu} \sum_{k=0}^{\infty} \frac{(\mu e^{it})^k}{k!} = e^{-\mu} e^{\mu e^{it}} = e^{\mu(e^{it}-1)}$$

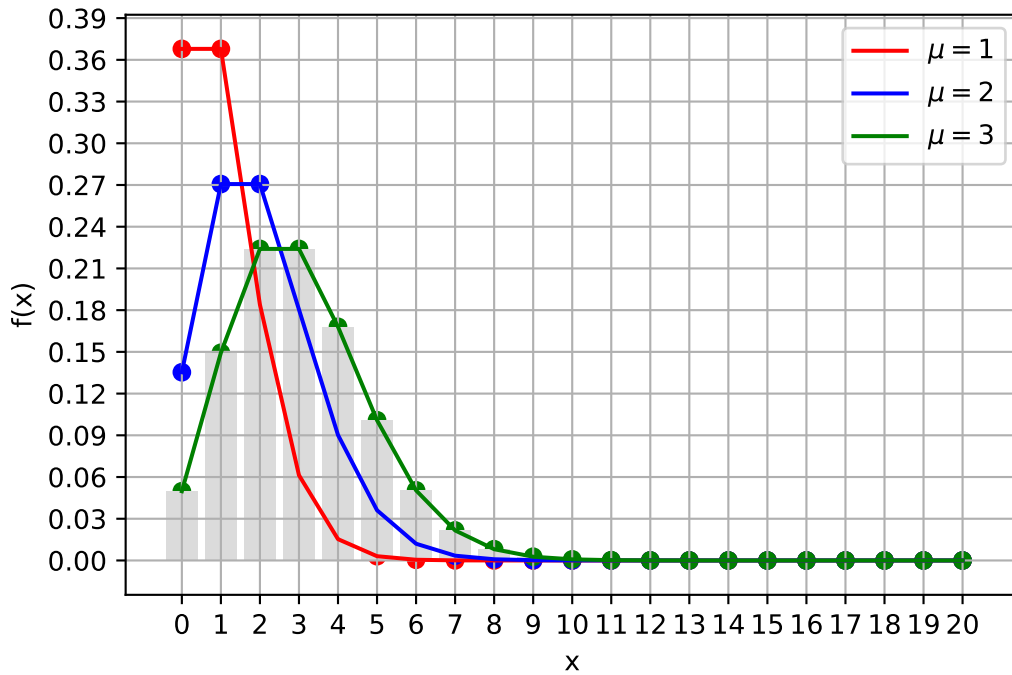


Рис. 1: Распределение Пуассона

## 1.2 Распределение Эрланга

Пусть случайная величина задается распределением:

$$f(x) = \frac{\lambda^m}{(m-1)!} x^{m-1} e^{-\lambda x}, \quad x, \lambda \in \mathbb{R}, \quad x, \lambda > 0, m \in \mathbb{N} \quad (4)$$

Найдем функцию распределения по определению:

$$F(x) = P(\xi \leq x) = \int_{-\infty}^x f(x) dx = \frac{\lambda^m}{\Gamma(m)} \int_0^x x^{m-1} e^{-\lambda x} dx$$

Либо же, функцию распределения можно выразить через закон Пуассона:

$$F(x) = 1 - \sum_{k=0}^{n-1} \frac{(\lambda x)^k}{k!} e^{-\lambda x}$$

### Математическое ожидание

**Определение 1.6.** Математическое ожидание произвольной непрерывной случайной величины  $\xi(\omega)$  называется интеграл Лебега от нее по мере  $P$ :  $E\xi(\omega) = \int_{\Omega} \xi(\omega) P(d\omega)$ .

При абсолютной сходимости:  $E\xi(\omega) = \int_{-\infty}^{+\infty} x f(x) dx$ .

Согласно определению 1.6 можно получить математическое ожидание распределения Эрланга:

$$\begin{aligned} E\xi &= \int_0^{+\infty} x \frac{\lambda^m}{(m-1)!} x^{m-1} e^{-\lambda x} = \boxed{\text{делаем замену } z = \lambda x} = \int_0^{+\infty} \frac{z}{\lambda} \frac{\lambda^m}{(m-1)!} \frac{z^{m-1}}{\lambda^{m-1}} e^{-z} \frac{dz}{\lambda} = \\ &= \boxed{\Gamma(n) = (n-1)!} = \frac{1}{\lambda \Gamma(m)} \int_0^{+\infty} z^m e^{-z} dz = \frac{\Gamma(m+1)}{\lambda \Gamma(m)} = \frac{m!}{\lambda(m-1)!} = \frac{m}{\lambda} \end{aligned}$$

### Дисперсия

**Определение 1.7.** Дисперсией случайной величины  $\xi$ , имеющей  $E\xi$ , называется число:

$$D\xi = E(\xi - E\xi)^2 = E\xi^2 - (E\xi)^2$$

По определению 1.7 найдем дисперсию распределения Эрланга:

$$\begin{aligned} E\xi^2 &= \int_0^{+\infty} x^2 \frac{\lambda^m}{(m-1)!} x^{m-1} e^{-\lambda x} dx = \boxed{\text{сделаем замену } z = \lambda x} = \int_0^{+\infty} \frac{z^{m+1} \lambda^m e^{-z} dz}{\lambda^{m+1} (m-1)! \lambda} = \\ &= \frac{1}{\lambda^2 \Gamma(m)} \int_0^{+\infty} z^{m+1} e^{-z} dz = \frac{\Gamma(m+2)}{\lambda^2 \Gamma(m)} = \frac{(m+1)!}{\lambda^2 (m-1)!} = \frac{(m+1)m}{\lambda^2} \end{aligned}$$

Тогда:  $D\xi = E\xi^2 - (E\xi)^2 = \frac{(m+1)m}{\lambda^2} - \left(\frac{m}{\lambda}\right)^2 = \frac{m}{\lambda^2}$

## Характеристическая функция

**Определение 1.8.** Характеристической функцией непрерывной случайной величины называют функцию от действительного аргумента

$$g(t) = g_{\xi}(t) = Ee^{it\xi} = \int_{\mathbb{R}} e^{itx} dF(x) = \int_{-\infty}^{+\infty} e^{itx} f(x) dx$$

Найдем характеристическую функцию распределения Эрланга по определению:

$$\begin{aligned} g(t) &= Ee^{it\xi} = \int_0^{+\infty} e^{itx} \frac{\lambda^m x^{m-1} e^{-\lambda x}}{(m-1)!} dx = \frac{\lambda^m}{\Gamma(m)} \int_0^{+\infty} x^{m-1} e^{-x(\lambda - it)} dx = \boxed{\text{замена } y = x(\lambda - it)} = \\ &= \frac{\lambda^m}{\Gamma(m)} \int_0^{+\infty} \frac{y^{m-1} e^{-y}}{(\lambda - it)^{m-1}} \frac{dy}{\lambda - it} = \frac{\lambda^m}{\Gamma(m)(\lambda - it)} \int_0^{+\infty} y^{m-1} e^{-y} dy = \frac{\lambda^m \Gamma(m)}{\Gamma(m)(\lambda - it)^m} = \frac{\lambda^m}{(\lambda - it)^m} = \\ &= \left( \frac{\lambda}{\lambda - it} \right)^m = \left( \frac{\lambda - it}{\lambda} \right)^{-m} = \left( 1 - \frac{it}{\lambda} \right)^{-m} \end{aligned}$$

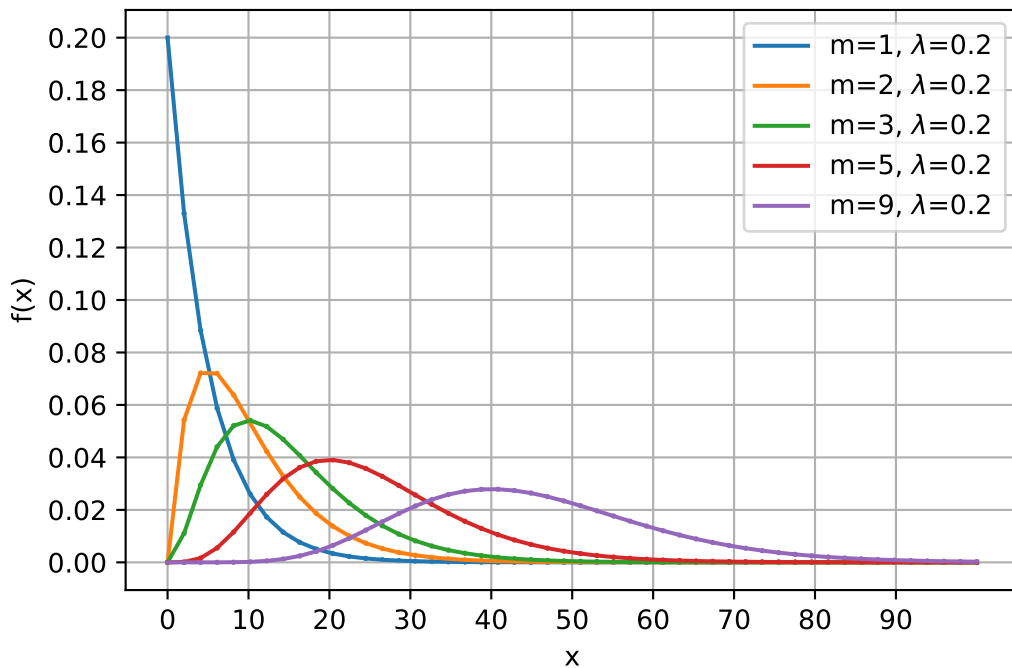


Рис. 2: Плотность распределения Эрланга

## 2 Поиск примеров событий

### 2.1 Распределение Пуассона

#### Типичная интерпретация

Распределение Пуассона описывает вероятность наступления  $k$  независимых событий за какое-то фиксированное время  $t$  с интенсивностью  $\lambda$ .

#### Нетипичная интерпретация

Распределение Пуассона играет большую роль в ставках на спорт [3]. Зная определенные коэффициенты, связанные с командами, можно рассчитать вероятность количества забитых голов той или иной командой и даже предсказать ставки, которые выставит букмекер.

Можно сказать, что сумма голов в матче подчиняется распределению Пуассона, а раз сумма распределена по Пуассону, то и каждый гол в свою очередь распределен по этому закону.

Пусть, для определенности, в один промежуток времени можно забить всего один гол, а каждый забитый гол является независимым событием.

Опишем некоторые вспомогательные значения:

- Среднее количество забитых мячей дома (на выезде) в сезоне по всем командам:

$$\mu_{in}, \mu_{out} = \left[ \frac{\text{КОЛ-ВО ГОЛОВ В СЕЗОНЕ}}{\text{КОЛ-ВО МАТЧЕЙ}} \right]$$

- Усредненный и взвешенный показатель забитых голов дома и на выезде:

$$\mathcal{A}_{in} = \frac{\sum \text{забитых дома}}{19/\mu_{in}}$$

$$\mathcal{A}_{out} = \frac{\sum \text{забитых на выезде}}{19/\mu_{out}}$$

- Усредненный и взвешенный показатель пропущенных голов дома и на выезде:

$$\mathcal{D}_{in} = \frac{\sum \text{пропущенные дома}}{19/\mu_{out}}$$

$$\mathcal{D}_{out} = \frac{\sum \text{пропущенные на выезде}}{19/\mu_{in}}$$

**NB!** Во всех значениях, описанных выше, важно учитывать **где** команда играет и **где** забивает голы.

Пусть у нас есть будущая игра  $X$  против  $Y$ . Тогда, чтобы рассчитать примерное показатель забитых голов в матче, воспользуемся формулами для расчета параметров каждой команды:

$$\lambda_X = \mathcal{A}_{in_X} \times \mathcal{D}_{out_Y} \times \mu_{in}$$

$$\lambda_Y = \mathcal{A}_{out_Y} \times \mathcal{D}_{in_X} \times \mu_{out}$$

Получив параметр  $\lambda$  для каждой команды, можно найти вероятность количества голов в предстоящем противостоянии по формулам:

$$P_X(k) = \frac{(\lambda_X)^k}{k!} e^{-\lambda_X} - \text{вероятность } k \text{ голов, забитых командой } X$$

$$P_Y(k) = \frac{(\lambda_Y)^k}{k!} e^{-\lambda_Y} - \text{вероятность } k \text{ голов, забитых командой } Y$$

Как нетрудно заметить, формулы для вычисления количества голов есть закон распределения Пуассона с параметром  $\lambda_i$ .

### Известные соотношения

1. Биноминальное распределение  $\longrightarrow$  распределение Пуассона

$$\begin{aligned} P(\xi = k) &= \lim_{n \rightarrow \infty} C_n^k p^k (1-p)^{n-k} = \boxed{p = \frac{\mu}{n}, \lim_{n \rightarrow \infty} p = 0} = \lim_{n \rightarrow \infty} C_n^k \frac{\mu^k}{n^k} \left(1 - \frac{\mu}{n}\right)^{n-k} = \\ &= \lim_{n \rightarrow \infty} \frac{n!}{(n-k)!k!} \left(\frac{\mu^k}{n^k}\right) \left(1 - \frac{\mu}{n}\right)^n \left(1 - \frac{\mu}{n}\right)^{-k} = \boxed{\lim_{n \rightarrow \infty} \left(1 - \frac{\mu}{n}\right)^n = e^{-\mu}, \lim_{n \rightarrow \infty} \left(1 - \frac{\mu}{n}\right)^{-k} = 1}, \\ \text{также } \lim_{n \rightarrow \infty} \frac{n!}{(n-k)!n^k} &= \boxed{\text{фиксируем } k} = \lim_{n \rightarrow \infty} \frac{n}{n} \cdot \frac{n-1}{n} \cdot \frac{n-2}{n} \dots \frac{n-k+1}{n} = 1. \end{aligned}$$

В итоге:  $P(\xi = k) = \frac{\mu^k}{k!} e^{-\mu}$

2. Распределение Пуассона  $\longleftrightarrow$  показательное распределение

$$\begin{aligned} P(\xi = k) &= \frac{(\mu t)^k}{k!} e^{-\mu t} \\ P(\xi = 0) &= e^{-\mu t} \\ P(\xi > 0) &= e^{-\mu t} \\ P(\xi \leq 0) &= 1 - e^{-\mu t} \end{aligned}$$

А это функция показательного распределения

3. Распределение Пуассона  $\longrightarrow$  Нормальное (Гаусса) распределение

При увеличении параметра  $\lambda$  распределение Пуассона стремится к нормальному распределению (распределению Гаусса) с параметрами  $\sigma = \sqrt{\lambda}$  и сдвигом  $\lambda$ . Для вывода нужно воспользоваться формулой Стирлинга:

$$n! = n^n e^{-n} \sqrt{2\pi n} \cdot e^{-\frac{\theta}{12n}(1+O(1/n))}, \quad n \rightarrow \infty, 0 < \theta < 1$$

Разложение в ряд Тейлора  $\ln\left(\frac{\lambda}{k}\right)^k$  в окрестности точки  $k = \lambda$ :

$$\ln\left(\frac{\lambda}{k}\right)^k = -(k - \lambda) - \frac{(k - \lambda)^2}{2\lambda} + O(k - \lambda)^3$$

$$\sqrt{k} \approx \sqrt{\lambda}$$

Тогда:

$$P(\xi = k) = \frac{1}{\sqrt{2\pi\lambda}} \cdot e^{-\frac{(k-\lambda)^2}{2\lambda}}$$

## 2.2 Распределение Эрланга

### Типичная интерпретация

Распределение Эрланга является одним из основных распределений математической статистики для получения случайных величин. Так как параметра  $m$  – целое число, распределение Эрланга описывает время, необходимое, для появления ровно  $m$  событий при условии, что они независимы и появляются с постоянной интенсивностью  $\lambda$ .

Распределение Эрланга (частный случай гамма-распределения) можно увидеть в диагностировании рака на разных временных этапах жизни человека. Считается, что на появление рака влияют канцерогенные факторы (одного общепринятого списка не существует, но для определенности будем считать, что таков список существует и он конечен), в последовательно проявлении которых возникает рак 1-й степени. Список видов рака обширен, но в дальнейшем будут рассмотрены 20 самых часто появляющихся видов. Если рассматривать появление любого вида рака в любой момент времени жизни человека (от 0 до 100 лет), то можно построить график, который очень схож с плотностью распределения Эрланга: «вероятность» появления рака в раннем возрасте и глубокой старости близка к 0, а в середине жизни (35-55 лет) достигает своего пика. В дальнейшей работе будет рассмотрена самая обширная база данных США по статистке рака с 1999 по 2015 года, а так же проиллюстрирована аппроксимация распределения Эрланга (с разными параметрами для каждого вида рака) к реальным данным.

### Нетипичная интерпретация

Применение распределения Эрланга можно найти в организации промышленных перевозок [2].

Пусть у нас есть какой-то поток поездов, который ездит по одному маршруту и прибывает на одну и ту же станцию. Тогда разобьем весь отрезок времени, в который приходят поезда, на интервалы точками  $t_0, t_1, t_2, \dots, t_k$  и будем их считать в какой-то условно выбранной величине  $T$ . Тогда можно получить отрезки, в которые поезда прибывают:  $(t_0, t_1), (t_1, t_2), \dots, (t_{k-1}, t_k)$ . Определим числовые характеристики для полученных отрезков:

$$M(t) = \sum_{i=1}^k P_i \bar{t}_i - \text{математическое ожидание;}$$

$$D(t) = \alpha_2 - [M(t)]^2 - \text{дисперсия;}$$

$$\delta(t) = \sqrt{D(t)} - \text{среднеквадратичное отклонение,}$$

где  $P_i$  вычисляется следующим образом: найдем количество поездов  $m_i$ , которые прибыли в данный отрезок времени  $(t_{i-1}, t_i)$ , и поделим на общее количество поездов  $N$  ( $P_i = \frac{m_i}{N}$ ), а  $\bar{t}_i$  – есть отрезок  $(t_{i-1}, t_i)$ .

Согласно полученным числовым характеристикам можно найти параметр распределения Эрланга по потоку поездов  $l$  и среднечасовую интенсивность  $\lambda$ :

$$l = \frac{[M(t)]^2}{D(t)}$$



$$\lambda = \frac{1}{M(t)}$$

Установлено, что для предприятий с внешним прибытием и объемом меньшим 10 млн тонн в год распределение межпоездных интервалов описывается распределением Эрланга 1-го порядка, а свыше 10 млн т. - законом Эрланга 2-го порядка.

### **Известные соотношения**

1. Распределение Эрланга  $\longleftarrow$  экспоненциальное распределение

$$\Gamma(1, \lambda) \equiv \text{Exp}(1/\lambda)$$

2. Распределение Эрланга  $\longleftarrow \chi^2$  распределение

$$\Gamma(m/2, 2) \equiv \chi^2(m)$$

3. Распределение Эрланга  $\longrightarrow$  Бета-распределение

Так как распределение Эрланга является частным случаем гамма-распределения, то:

$$\text{Пусть } \xi \sim \Gamma(\alpha, 1), \eta \sim \Gamma(\beta, 1), \alpha, \beta \in \mathbb{N}$$

Тогда:

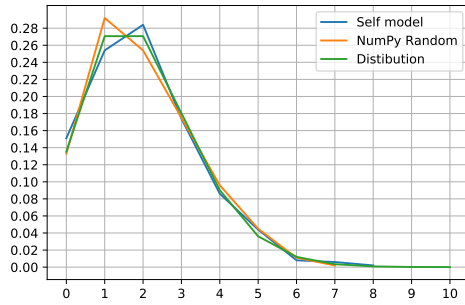
$$\frac{\xi}{\xi + \eta} \sim \text{Be}(\alpha, \beta),$$

где  $\text{Be}(\alpha, \beta)$ — есть бета-распределение.

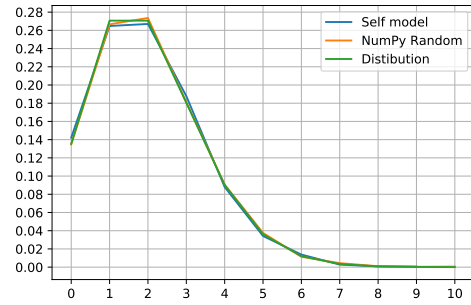
## 3 Моделирование

### 3.1 Распределение Пуассона

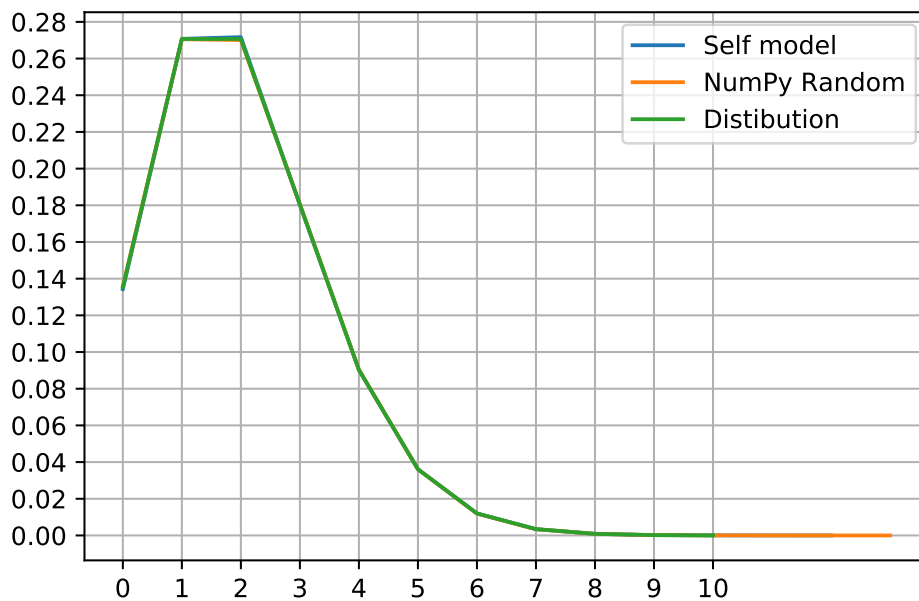
Моделирование случайной величины Пуассона выполнено методом генерации равномерно распределенных случайных величин на отрезке  $(0, 1)$  (используя функцию библиотеки NumPy: `numpy.random.uniform(low=0.0, high=1.0, size=None)`) [4].



а) 1000 точек



б) 5000 точек



в) 1000000 точек

Рис. 3: Моделирование случайной величины Пуассона в сравнении с библиотечной функцией и функцией распределения

### Оценка времени

```
In [5]: %%timeit
data = create_data((transform_numbers(random_poisson(2, 10000))))

56 ms ± 3.55 ms per loop (mean ± std. dev. of 7 runs, 10 loops each)

In [6]: %%timeit
data2 = create_data((transform_numbers(np.random.poisson(2, 10000))))

4.77 ms ± 48.6 µs per loop (mean ± std. dev. of 7 runs, 100 loops each)
```

Рис. 4: Оценка времени моделирования

## 3.2 Распределение Эрланга

Для моделирования случайно величины Эрланга воспользуемся методом обратной функции[1]. Из факта, что сумма экспоненциальных случайных величин распределена по закону Эрланга, можно смоделировать случайную величину Эрланга.

Функция экспоненциального распределения:

$$F(x) = 1 - e^{-\lambda x}$$

Пусть  $R \sim U(0, 1)$ . Тогда:

$$R = \int_0^x f(x) dx = \int_0^x \lambda e^{-\lambda x} = 1 - e^{-\lambda x} = F(x)$$

Отсюда:

$$x = -\frac{1}{\lambda} \ln(1 - R)$$

Так как случайная величина  $(1 - R)$  распределена так же как и  $R$ , то:

$$x = -\frac{1}{\lambda} \ln(R)$$

Пусть  $\xi_i \sim Pois(\lambda)$ ,  $i = \overline{1, k}$ . Тогда  $\sum_{i=1}^k \xi_i \sim Erlang(\lambda, k)$

Что касается библиотечного моделирования: в библиотеке **numpy**, в модуле **random** есть функция **gamma(shape, scale, size=None)**, которая может моделировать случайные величины, но по другому виду распределения (взято с [сайта NumPy](#)):

$$p(x) = x^{k-1} \frac{e^{-x/\theta}}{\theta^k \Gamma(k)}$$

Отличие составляет параметр  $\theta$  или, в случае этой работы,  $\lambda$ , который должен быть обратным ( $\theta^{-1}$ ).

Ниже представлены 3 рисунка моделирования случайной величины Эрланга при 3 разных количествах точек

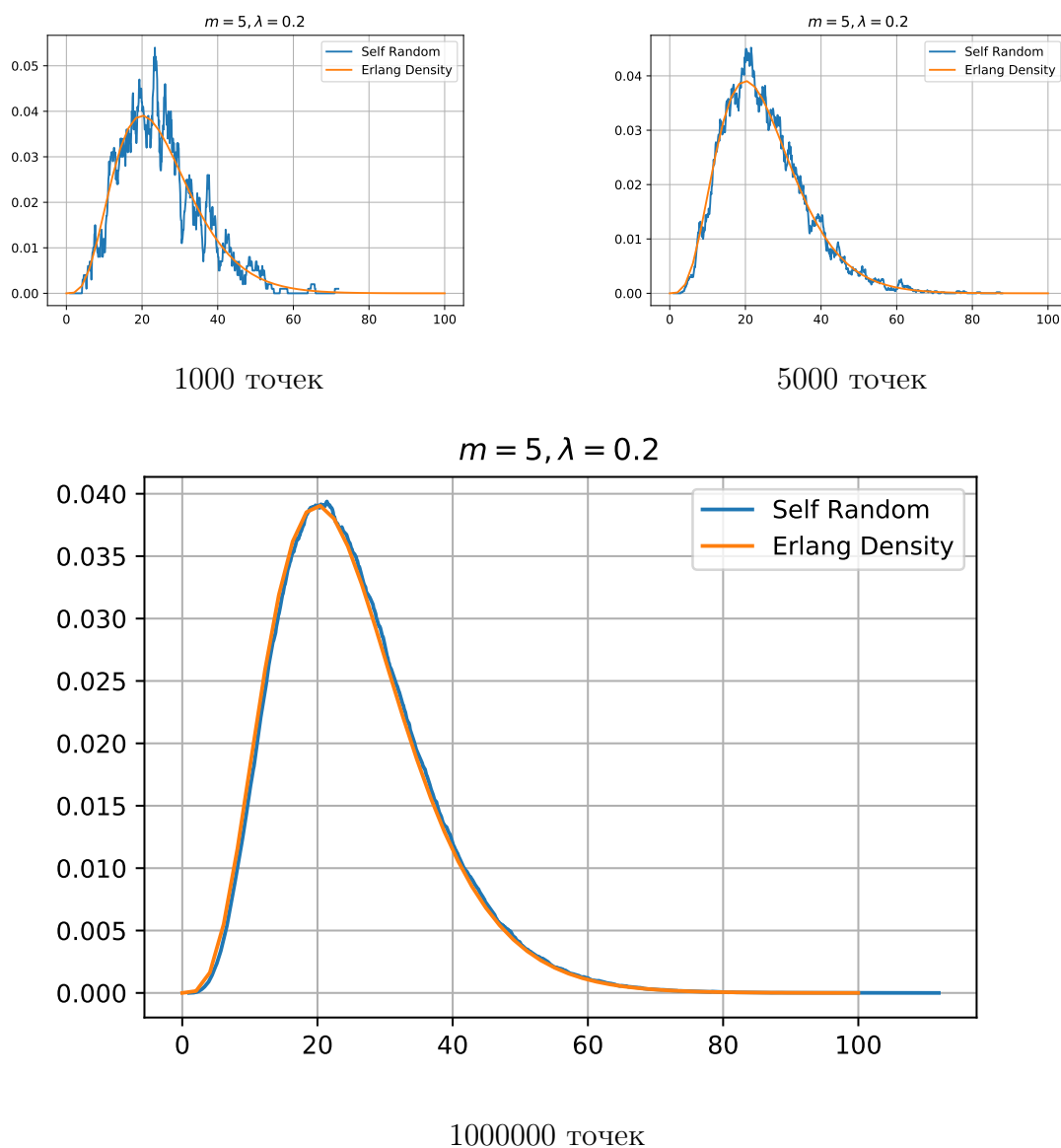


Рис. 5: Моделирование случайной величины Эрланга в сравнении с функцией распределения с параметрами  $m = 5, \lambda = 0.2$

**Оценка времени** Ниже будет приведен код, который с помощью `magic commands` в Jupyter notebook вычисляет работу времени отдельного куска кода. В двух ячейках проиллюстрированы оценки времени для функции распределения и самостоятельного моделирования.

Как видно из рисунка 6, реализация библиотечной функции быстрее в примерно 330 раз.

## Оценка времени

```
In [12]: %%timeit
random_erlang(5,0.2,100000)

1.65 s ± 149 ms per loop (mean ± std. dev. of 7 runs, 1 loop each)
```

```
In [13]: %%timeit
np.random.gamma(5,0.2,100000)

5.12 ms ± 17.5 µs per loop (mean ± std. dev. of 7 runs, 100 loops each)
```

Рис. 6: Скриншот с оценкой времени

## Список литературы

- [1] Вадзинский Р.Н. Справочник по вероятностным распределениям. — СПб.: Наука, 2001, 295 с.
- [2] Акулиничев В. М. Организация перевозок на промышленном транспорте. — МСК: Высшая Школа, 1983, 53 с.
- [3] <https://dashee87.github.io/football/python/predicting-football-results-with-statistical-modelling/>
- [4] Ивченко Г. И., Медведев Ю. И. Введение в математическую статистику. — М.: Издательство ЛКИ, 2010, 65 с.