Отчет по лабораторной работе № 1

**Анализ одномерных случайных величин**

Выполнено:

Барбаков Илья Олегович,

Группа J4150

Санкт-Петербург

2022

**Оглавление**

[1. Обоснование выбора данных 1](#_Toc117510010)

[2. Построение непараметрической оценки PDF в виде гистограммы и использование функции плотности ядра 1](#_Toc117510011)

[3. Оценка порядковой статистики и её представление в виде "ящика с усами" ………………….. 3](#_Toc117510015)

[4. Подбор теоретических распределений, которые лучше всего отражают эмпирические данные 4](#_Toc117510018)

[5. Оценка параметров распределения случайной величины с помощью метода максимального правдоподобия и LS-методов 5](#_Toc117510019)

[6. Проверка эмпирических и теоретических распределений с помощью квантильного графика 11](#_Toc117510031)

[7. Статистические тесты 13](#_Toc117510034)

[8. Сэмплирование случайной величины 14](#_Toc117510038)

[Исходный код 20](#_Toc117510048)

# Обоснование выбора данных

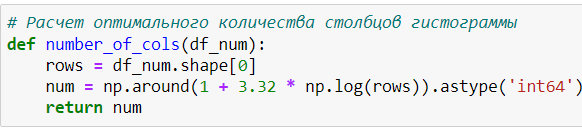
Для задачи анализа одномерных случайных величин выбран датасет с данными действий клиентов телефонной компании, а также с целевой переменной оттока каждого клиента. Выборка имеет большое количество записей, что обеспечит достаточный уровень репрезентативности данных.

Задачей работы является анализ нескольких декретных случайных величин, а также непрерывных. В качестве дискретной величины было выбрано 2 параметра: количество международных звонков, а также количество звонков в службу поддержки клиентов. Эти показатели распределятся с оптимальным количеством классов (20 и 8), и также большинство этих классов представлены как минимум 10 значениями в каждом, что позволит успешно подобрать закон распределения.

Для анализа непрерывных случайных величин были выбраны следующие показатели: плата за вечерние и ночные звонки. По первичному анализу данные распределены близко к нормальному и не имеют несколько пиков, благодаря чему можно будет подобрать известный закон распределения.

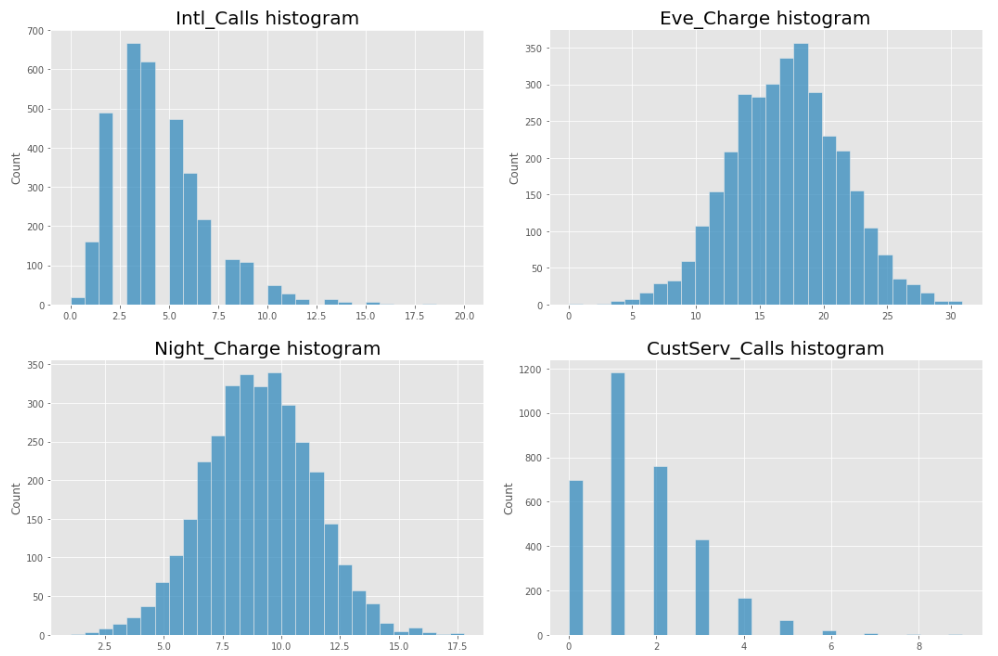
# Построение непараметрической оценки PDF в виде гистограммы и использование функции плотности ядра

Гистограмма – непараметрический способ оценки распределений, позволяющий первично оценить где находятся данные. Выборка может быть разной и поэтому чтобы отразить на графике гладкость и унимодальность необходимо подобрать количество столбцов диаграммы. Напишем для этого функцию:



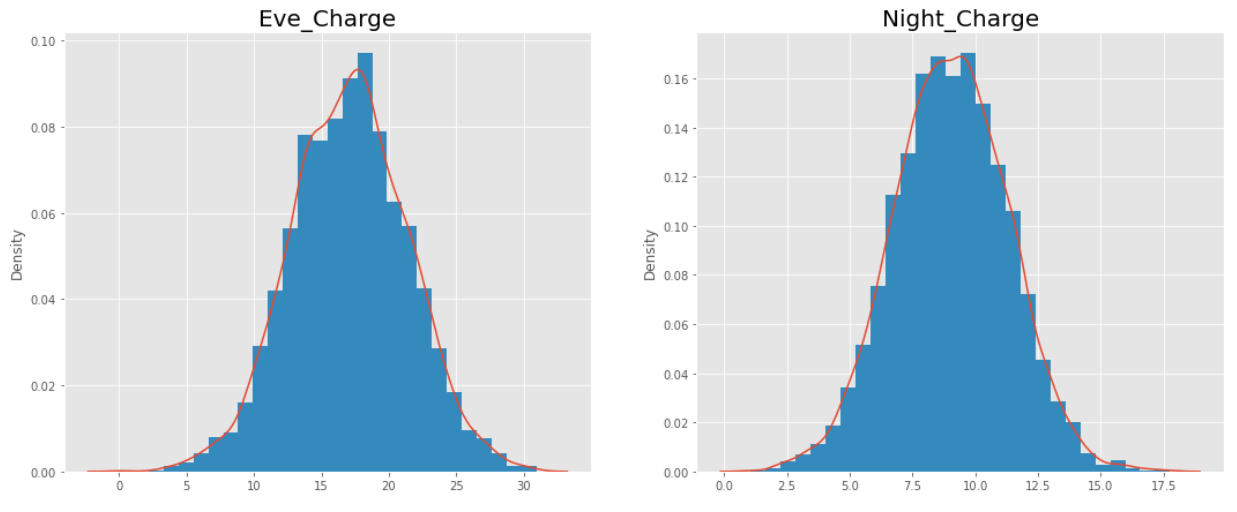
## Рисунок 1 - Функция расчета количества столбцов гистограммы

Теперь строим гистограммы для всех случайных величин:



## Рисунок 2 - Гистограммы случайных величин

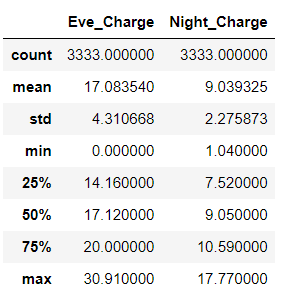
Для получения гладких оценок плотности распределения непрерывной случайной величины, применяем метод ядерного сглаживания, на основе Гауссовой функции. Для этого применяем метод sns.kdeplot, который выводит график. Параметр «bw\_adjust» - ширину окна подбираем эмпирически.



## Рисунок 3 - Ядерное сглаживание

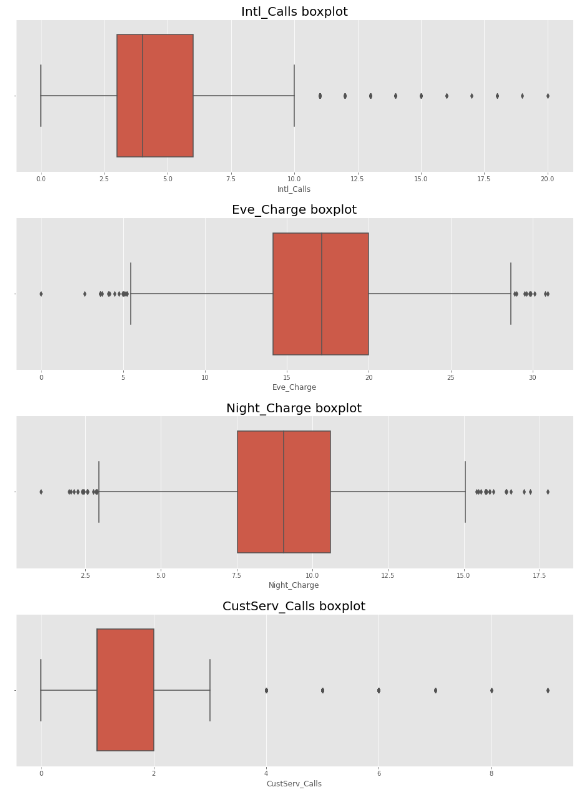
# Оценка порядковой статистики и её представление в виде "ящика с усами"

Выводим порядковую и описательную статистику для непрерывных случайных величин. Она включает в себя значения среднего, максимального, минимального, стандартного отклонения, а также квантилей.



## Рисунок 4 – Статистика

Выводим график в виде «ящика с усами» - диаграмму размаха, которая наглядно показывает медиану, нижний и верхний [квартили](https://ru.wikipedia.org/wiki/%D0%9A%D0%B2%D0%B0%D0%BD%D1%82%D0%B8%D0%BB%D1%8C#%D0%9C%D0%B5%D0%B4%D0%B8%D0%B0%D0%BD%D0%B0_%D0%B8_%D0%BA%D0%B2%D0%B0%D1%80%D1%82%D0%B8%D0%BB%D0%B8), минимальное и максимальное значение выборки и [выбросы](https://ru.wikipedia.org/wiki/%D0%92%D1%8B%D0%B1%D1%80%D0%BE%D1%81_(%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0)). Реализуем вывод графиков с помощью метода библиотеки seaborn – boxplot.



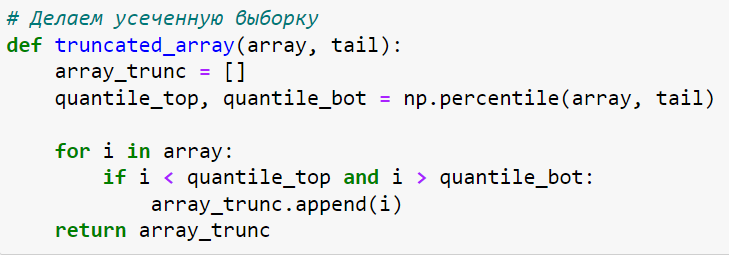
## Рисунок 5 - Ящики с усами

# Подбор теоретических распределений, которые лучше всего отражают эмпирические данные

После проведения первичного анализа, были выбраны следующие теоретические распределения случайной непрерывной величины: нормальное, логарифмически нормальное, гамма, бета и альфа. Для дискретной величины: биномиальное и распределение Пуассона.

# Оценка параметров распределения случайной величины с помощью метода максимального правдоподобия и LS-методов

Изучая распределения, по графику квантилей было заметно, что присутствуют выбросы на хвостах выборки. Такие аномальные значения не отображают логику закона распределения основных данных и могут искажать оценку. Создадим функцию, реализующую создание усеченной выборки данных.



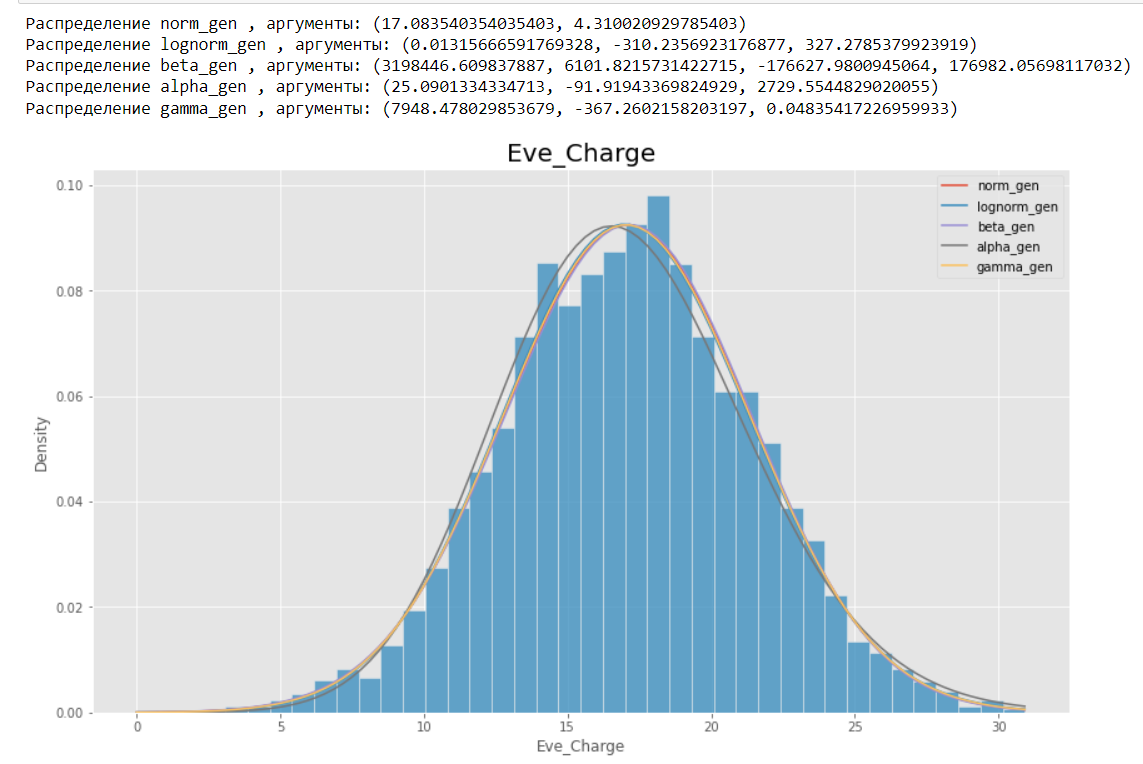
## Рисунок 6 - Усеченная выборка

Подбираем параметры с помощью метода максимального правдоподобия для каждого закона распределения. Суть данного метода заключается в оценке параметров распределения, путем максимизации функции правдоподобия. В python данный метод для непрерывных случайных величин реализован с помощью функции «scipy.stats.rv\_continuous.fit».

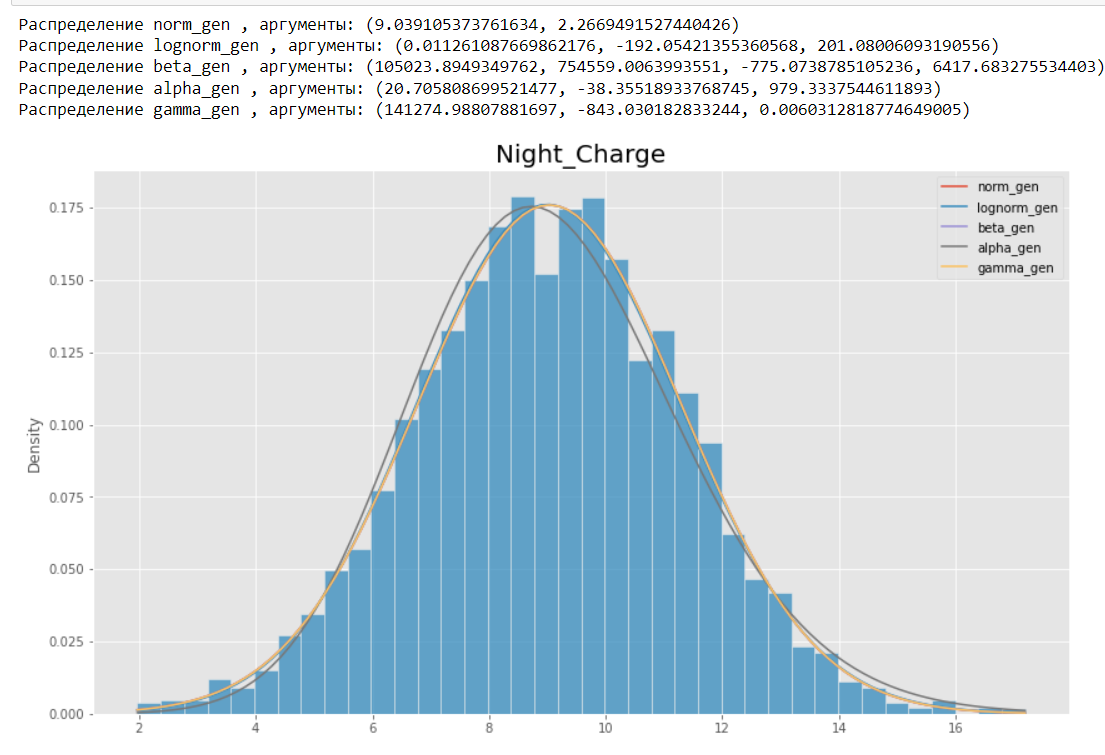


## Рисунок 7 - Метод максимального правдоподобия для непрерывных СВ

Выводим графики теоретических распределений для непрерывных случайных величин.

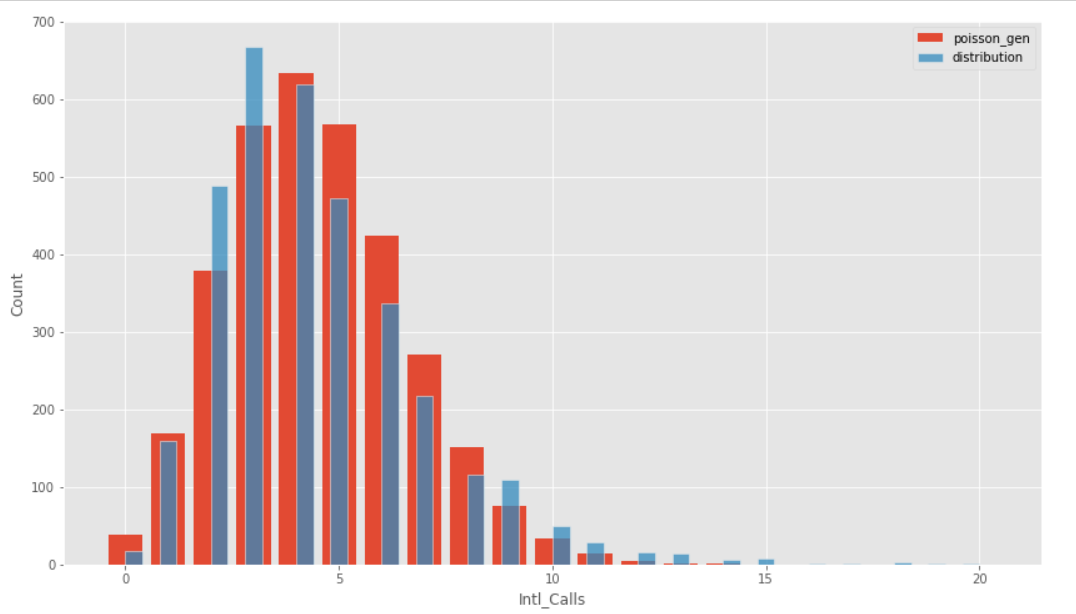


## Рисунок 8 – Метод MLE для «Eve\_Charge»

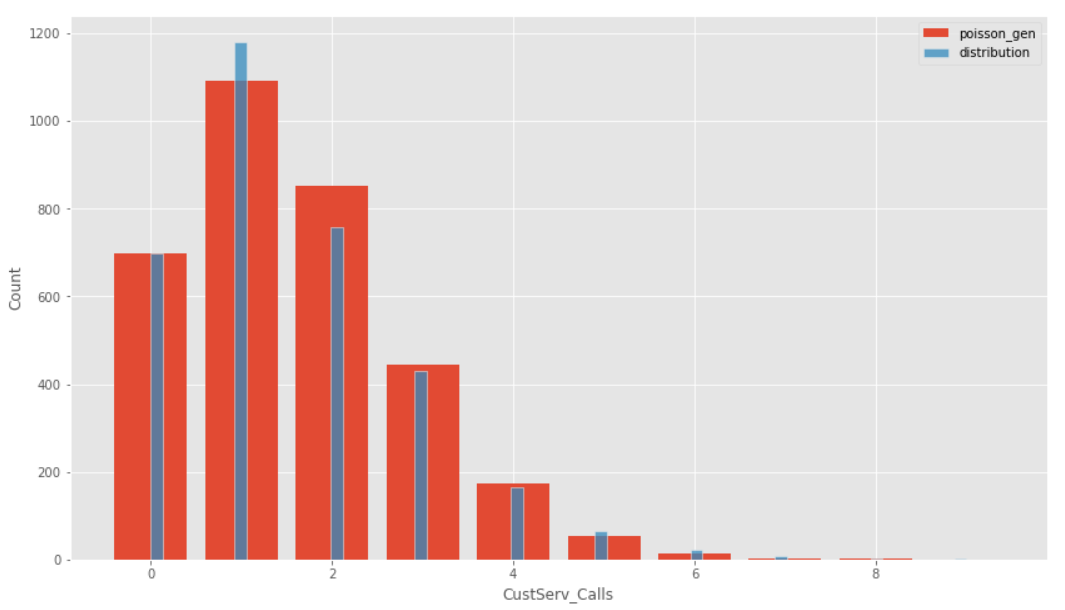


## Рисунок 9 – Метод MLE для «Night\_Charge»

Оценка параметров распределений для дискретных случайных величин представлена в пакете scipy методом «scipy.stats.fit». Выводим графики теоретического Пуассоновского распределения для показателей «Intl\_Calls» и «CustServ\_Calls».

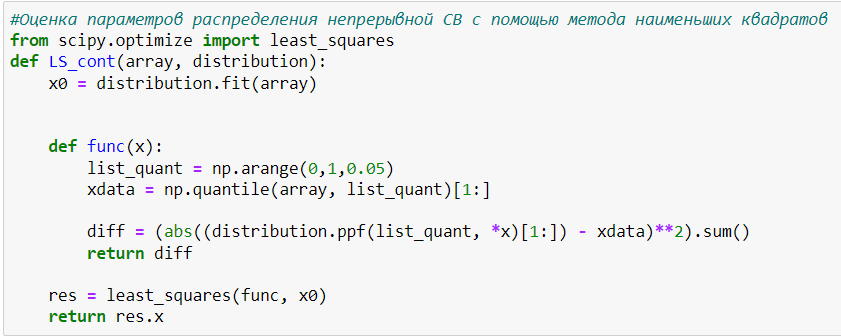


## Рисунок 10 - Метод MLE для «Intl\_Calls»



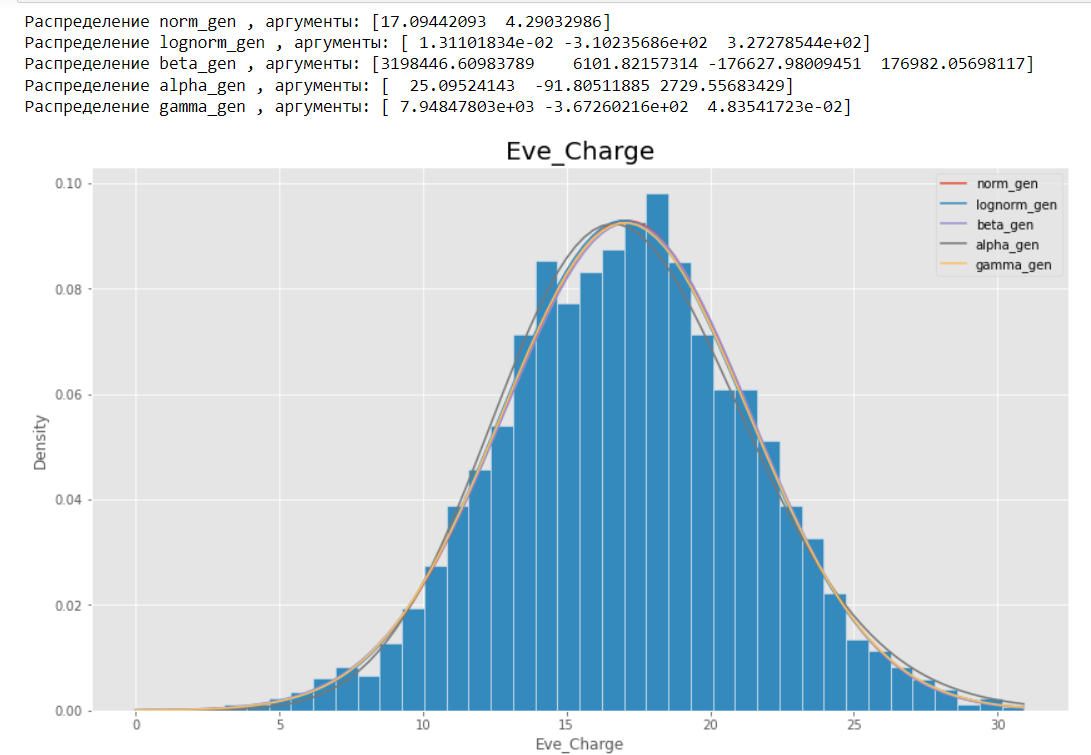
## Рисунок 11 - Метод MLE для «CustServ\_Calls»

Оценивать параметры теоретического распределения также возможно с помощью метода минимальных квадратов, суть которого заключается в минимизации суммы квадратов ошибки по определенным квантилям выборки.

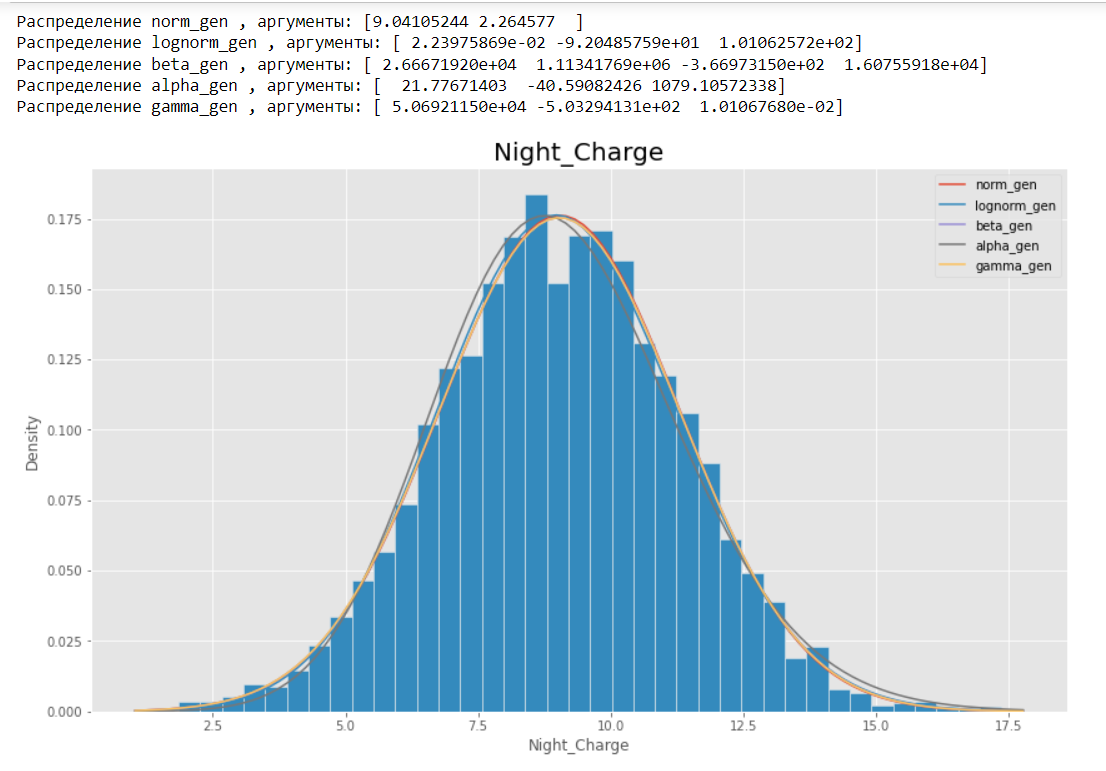


## Рисунок 12 - Метод наименьших квадратов

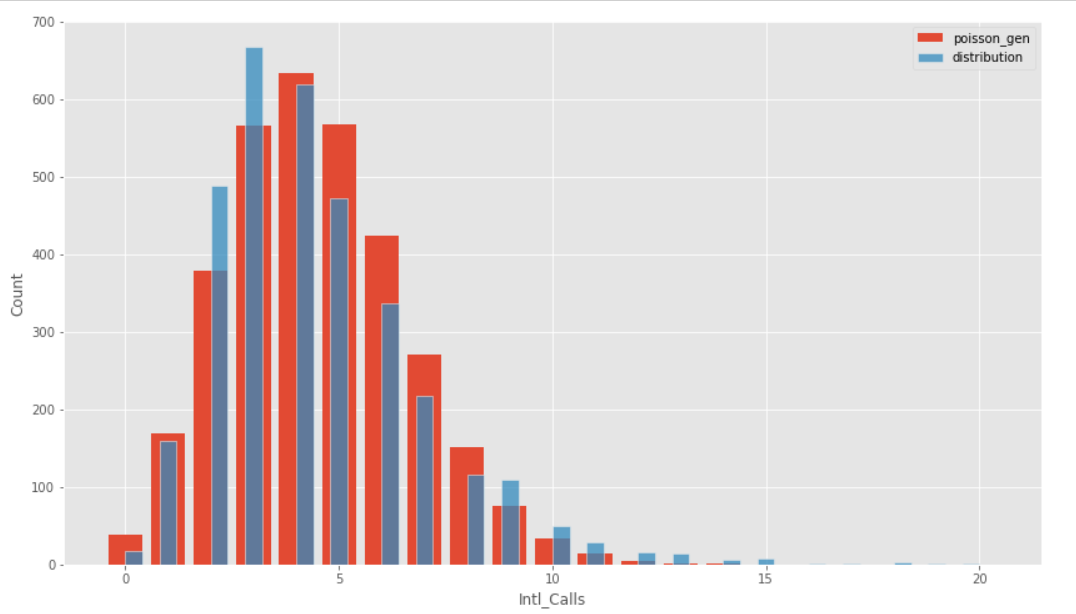
Выведем графики теоретических распределений по параметрам, подобранным с помощью МНК.



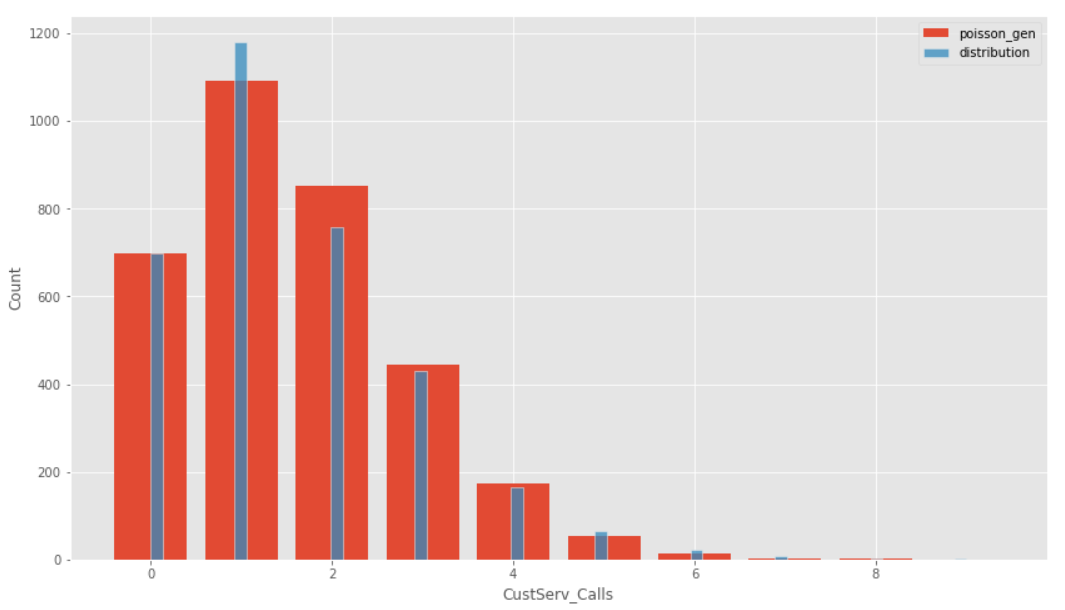
## Рисунок 13 - Метод МНК для «Eve\_Charge»



## Рисунок 14 - Метод МНК для «Night\_Charge»



## Рисунок 15 - Метод МНК для «Intl\_Calls»

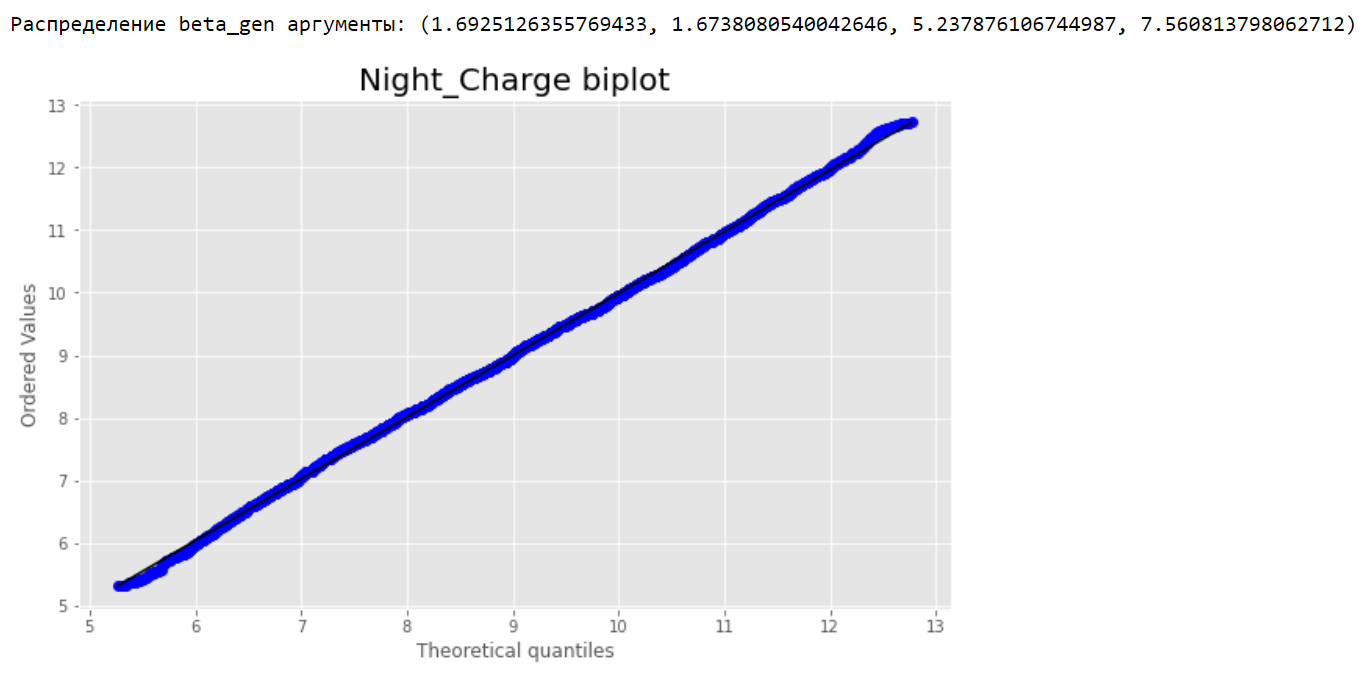


## Рисунок 16 - Метод МНК для «CustServ\_Calls»

С первого взгляда практически все законы довольно хорошо описывают распределения, поэтому необходимо провести дальнейшие проверки.

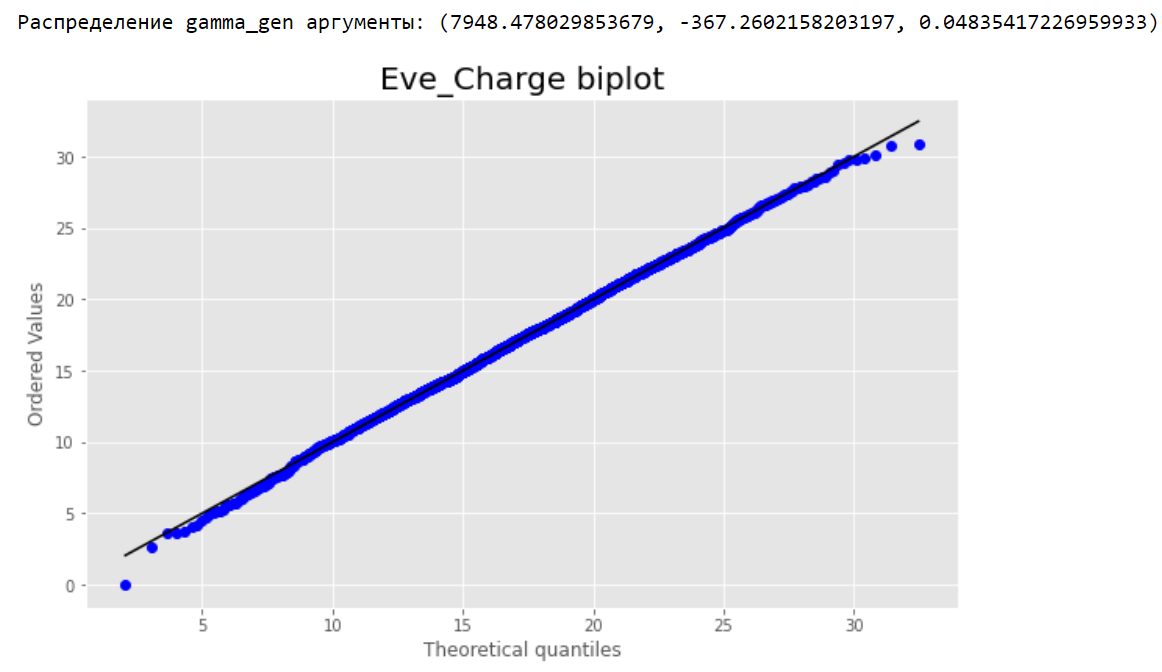
# Проверка эмпирических и теоретических распределений с помощью квантильного графика

Проведем оценку подбора распределения с помощью квантильного биплота. Чем ближе точки значений квантилей лежат на главной диагонали, тем лучше подобрано теоретическое распределение. Для параметра «Night\_Charge» наилучшим распределением выступает beta-распределение.



## Рисунок 17 - Квантильный биплот «Night\_Charge»

Для показателя «Eve\_Charge» оптимальным распределением выступает gamma-распределение.

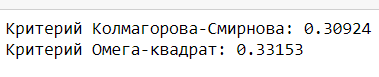


## Рисунок 18 - Квантильный биплот «Eve\_Charge»

Точки значений теоретических квантилей лежат практически на главной диагонали, что порождает вывод о том, что подобран оптимальный закон распределения.

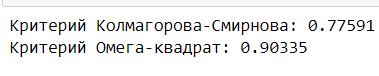
# Статистические тесты

Оценим критерии согласия - критерии для сравнения эмпирического распределения и теоретического. Для непрерывных случайных величин применим критерии Колмогорова-Смирнова и Омега-квадрат. Для статистических тестов примем уровень достоверности равным 95%. Таким образом, результат для «Night\_Charge»:



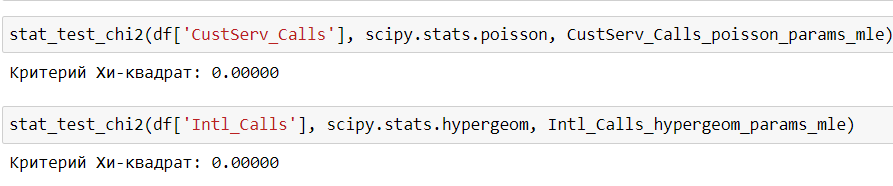
## Рисунок 19 - Статистический тест «Night\_Charge»

P-значение превышает 0.05, таким образом мы принимаем нулевую гипотезу, о том что данные распределены по beta закону. Далее применим эти же критерии для «Eve\_Charge». Значение p также превышает 0.05, что говорит нам о том, что данные распределены по gamma закону.



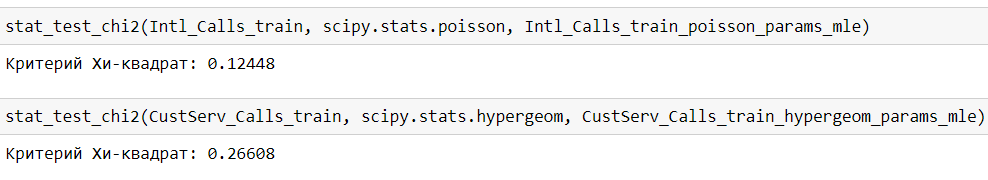
## Рисунок 20 - Статистический тест «Eve\_Charge»

Для дискретных случайных величин воспользуемся тестом Хи-квадрат, который проверяет нулевую гипотезу о том, что категориальные данные имеют заданные частоты.



## Рисунок 21 - Тест Хи-квадрат

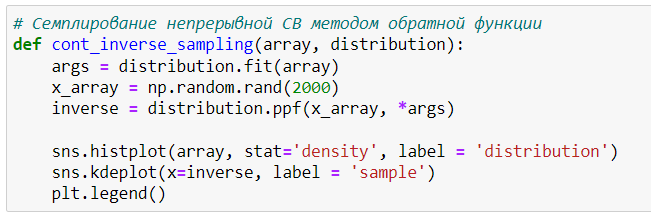
Результаты теста для всего датасета получаются нулевыми, поэтому попробуем сократить объем выборки и опять сравнить данные с помощью критерия Хи-квадрат. Теперь теоретическое распределение подобрано хорошо.



## Рисунок 22 - Рисунок 23 - Тест Хи-квадрат для усеченной выборки

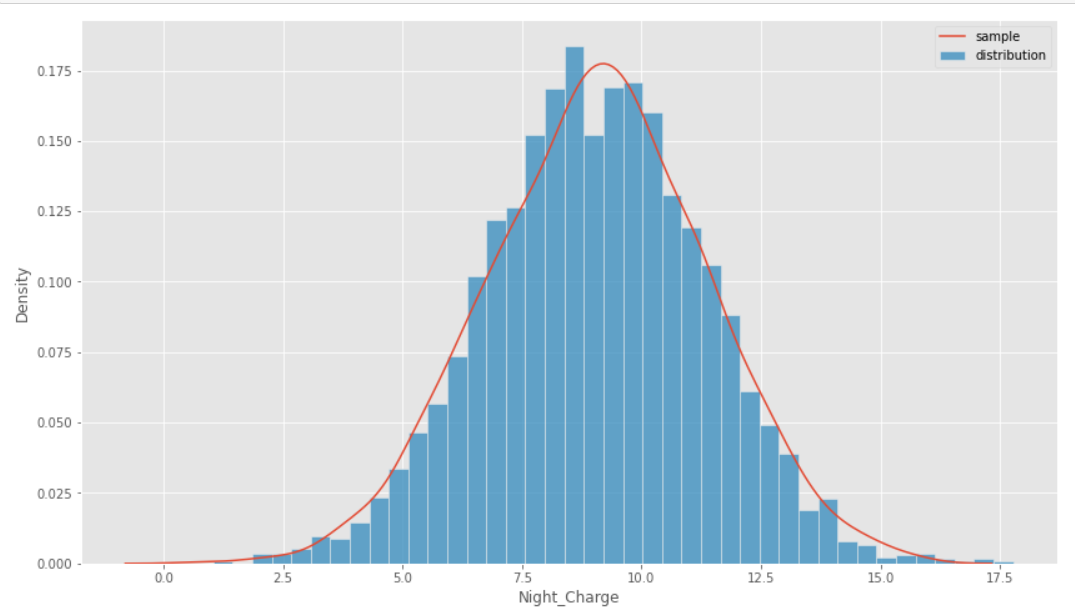
# Сэмплирование случайной величины

Далее необходимо произвести сэмплирование случайной величины, основываясь на уже оценённых распределениях. Первым способом выступает сэмплирование обратной функцией. Его смысл заключается в нахождении обратной функции подобранного распределения. Далее генерация равномерной случайной величины в диапазоне от 0 до 1 и далее применение этих данных в уже найденную функцию.

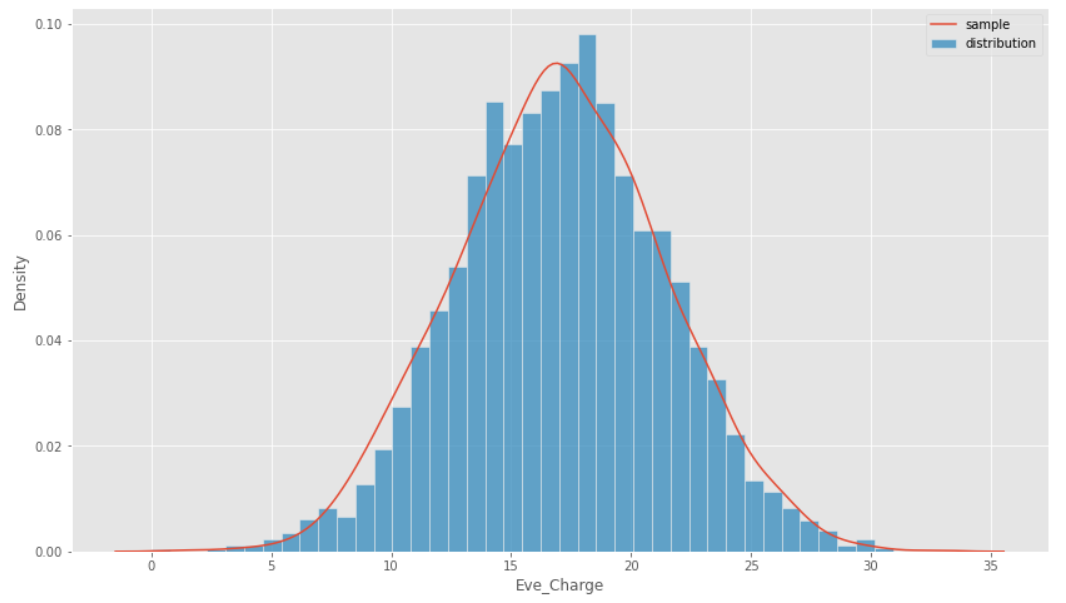


## Рисунок 24 - Сэмплирование методом обратной функции

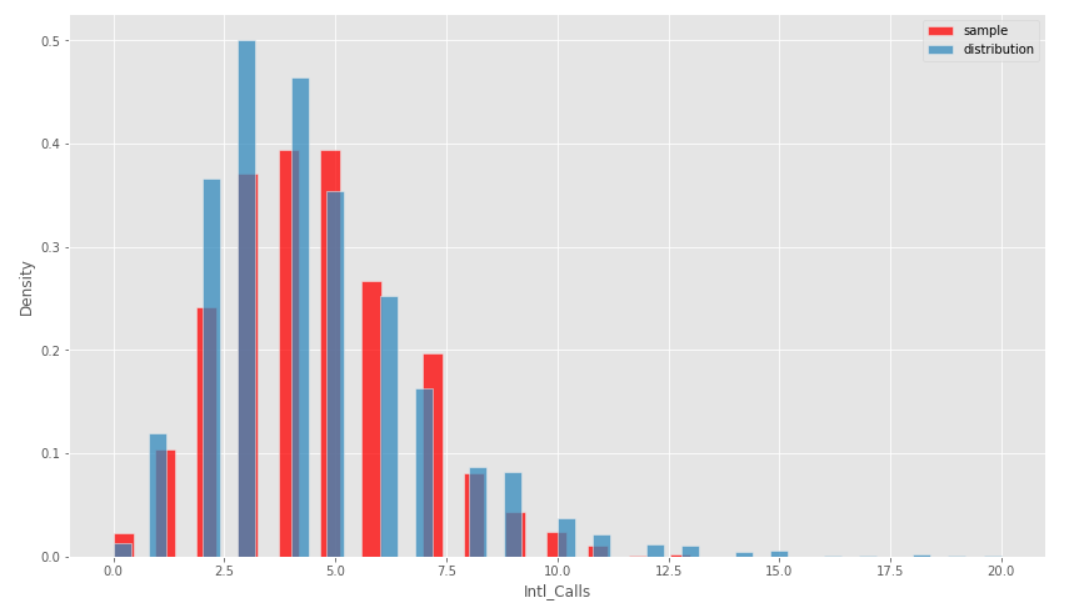
Выведем на график результат сэмплирование, чтобы оценить приближенность выборки к выбранному распределению.



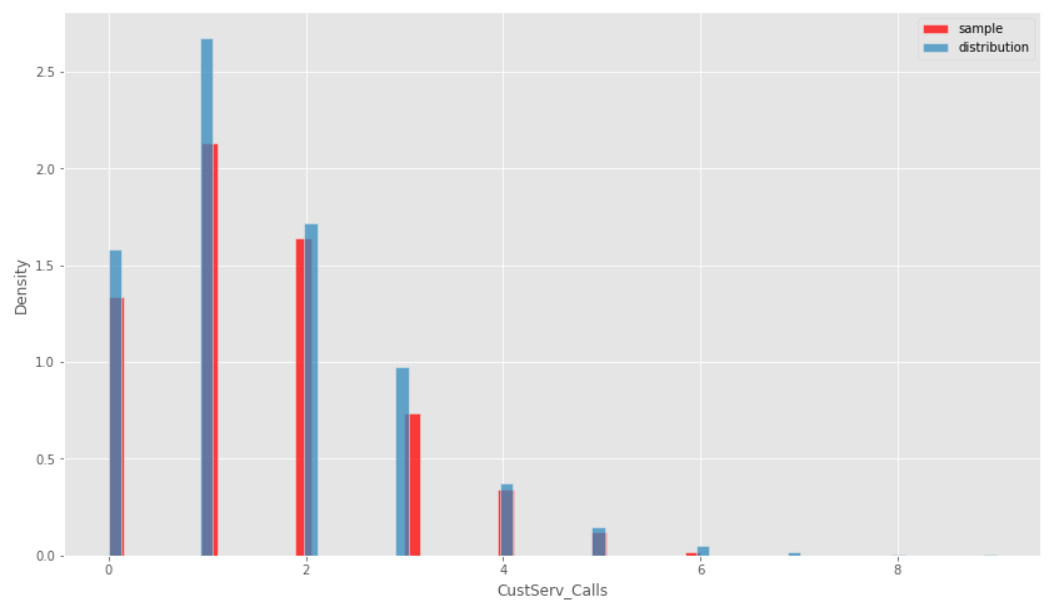
## Рисунок 25 - Сэмплирование обратной функцией для «Night\_Charge»



## Рисунок 26 - Сэмплирование обратной функцией для «Eve\_Charge»



## Рисунок 27 - Сэмплирование обратной функцией для «Intl\_Calls»

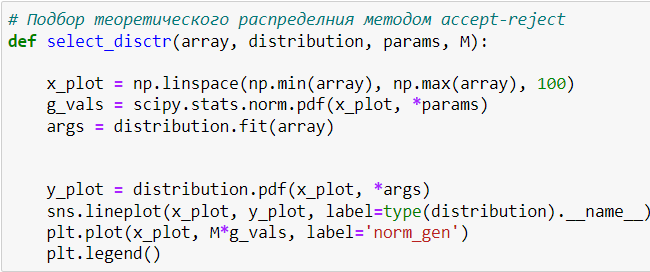


## Рисунок 28 - Сэмплирование обратной функцией для «CustServ\_Calls»

Исходя из графиков, сэмплированная выборка подчиняется выбранным законам распределения и может быть использована в дальнейших процедурах.

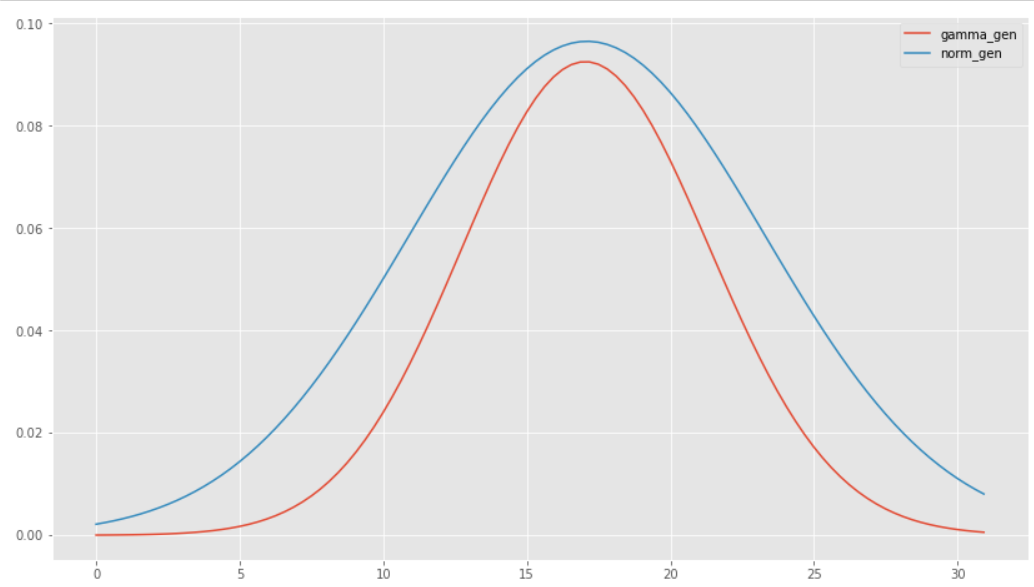
Вторым способом сэмплирование случайной величины является accept-reject метод, суть которого заключается в выборе теоретического распределение, которое по графику накрывает выборку. Далее мы сэмплируем значения из уже известного распределения и выбираем только те значения, которые попадают и в исходную выборку.

Первым шагом подберем теоретическое распределение, которое оптимально накроет наши данные.

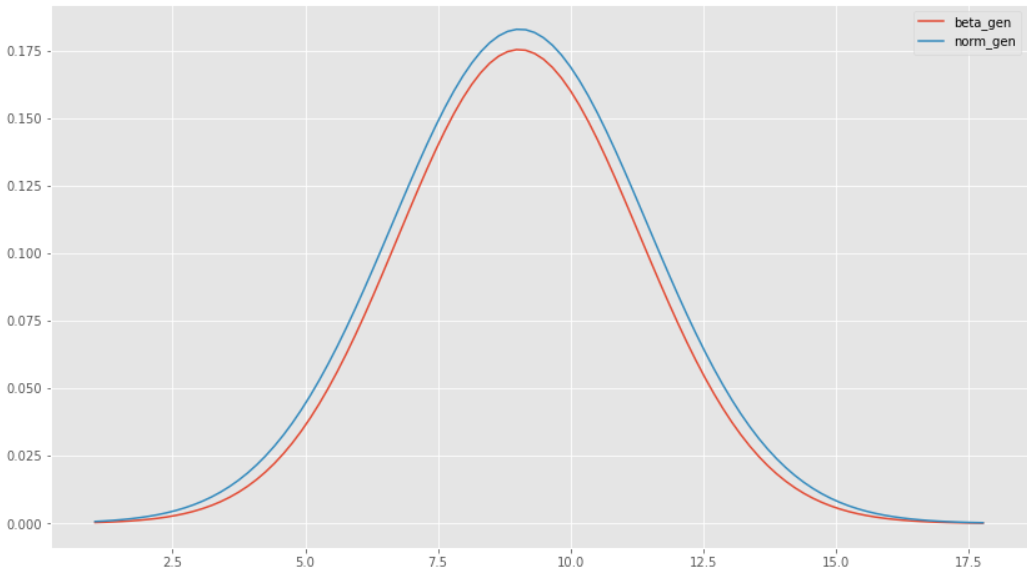


## Рисунок 29 - Подбор теоретического распределения методом accept-reject

В качестве теоретического распределения будем использовать нормальное, так как оно наиболее оптимально подходит по графику. Подбираем параметры нормального распределения так, чтобы оно накрывало исходное распределение, но зазор между ними был минимальным.

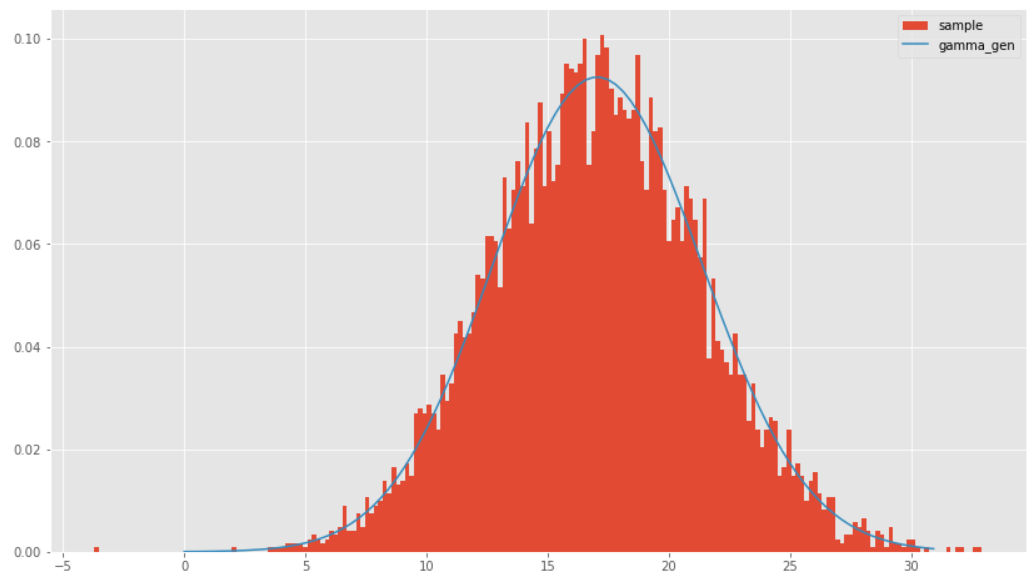


## Рисунок 30 - Подбор теоретического распределения методом accept-reject для «Eve\_Charge»



## Рисунок 31 - Подбор теоретического распределения методом accept-reject для «Night\_Charge»

Далее реализуем accept-reject сэмплирование.



## Рисунок 32 - Accept-reject сэмплирование для «Eve\_Charge»

## C:\Users\Test\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Снимок экрана (84).png

## Рисунок 33 - Accept-reject сэмплирование для «Night\_Charge»

Сэмплированная выборка попадает под изначальное распределение случайной величины, таким образом, параметры подобраны верно и сэмплирование проведено успешно.

## 

# Исходный код

Ссылка на репозиторий:

<https://github.com/barbak11/MOMO/tree/main/Lab1_one-dimensional%20random%20variable>