Отчет по лабораторной работе № 1

Анализ одномерных случайных величин

Выполнено:

Карпов П.А.

J4150

Санкт-Петербург

2022

Оглавление

[1. Обоснование выбора датасета 3](#_Toc117474373)

[2. Построение непараметрической оценки PDF в виде гистограммы и использование функции плотности ядра 3](#_Toc117474374)

[3. Оценка порядковой статистики 4](#_Toc117474375)

[4. Подбор теоретических распределений, которые лучше всего отражают эмпирические данные. 5](#_Toc117474376)

[Моделирование непрерывных СВ: 5](#_Toc117474377)

[Моделирование смеси распределений Гаусса 7](#_Toc117474378)

[5. Сэмплирование 10](#_Toc117474379)

# 1. Обоснование выбора датасета

Исходный датасет состоит представляет собой таблицу 240 строк × 6 колонок и содержит сведения о поставках оборудования в компании “Газпромнефть - Снабжение”

Название колонок и тип случайной величины представлены в таблице 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Прогнозная дата поставки** | **Дата ФП по последней входящей поставке** | **План.сумма с НДС 18%** | **Курс2** | **Задержка поставки** |
| **…** | **…** | **…** | **…** | **…** |

Таблица1. Название колонок датасета

Были исследованы следующие признаки:

1. План.сумма с НДС 18%
2. Курс2
3. Задержка поставки

Все признаки представлены непрерывными СВ.

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

Ядерная оценка PDF для непрерывных СВ:

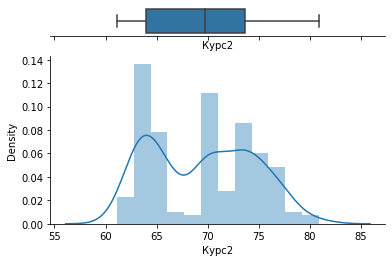


Рисунок 1. Ядерная оценка плотности Курс2.

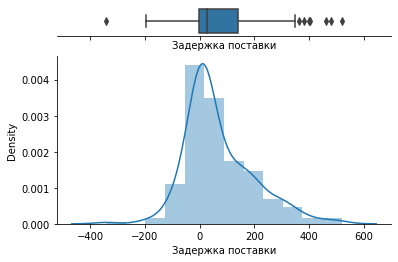


Рисунок 2. Ядерная оценка плотности Задержка поставки

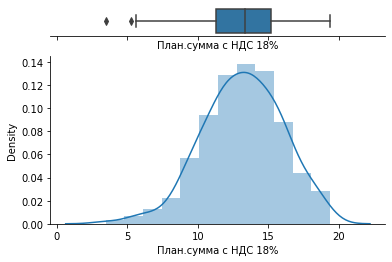


Рисунок 3. Ядерная оценка плотности План.сумма с НДС 18%

# 3. Оценка порядковой статистики

Выводим порядковую и описательную статистику для СВ:



Рисунок 4. Статистика по непрерывным СВ

# 4. Подбор теоретических распределений, которые лучше всего отражают эмпирические данные.

## Моделирование непрерывных СВ:

Для моделирования непрерывных случайных величин *“Задержка поставки”, “Курс2”* и *“План.сумма с НДС 18%”* было проведено сравнение всех возможных непрерывных распределений из библиотеки *scipy.stats* Наиболее подходящее под данные распределение выбиралось согласно максимально возможному значению *pvalue* теста Колмогорова-Смирнова. Для лучшего распределения также было рассчитано значение *pvalue* критерия Крамера-Мизеса.

Распределения подбирались с помощью метода максимального правдоподобия (МП) и с помощью метода наименьших квадратов (МНК).

Таким образом, для непрерывных распределений получим:

Сравнение реального и эмпирического распределения для колонки *"Задержка поставки"* представлено на рисунке 5.

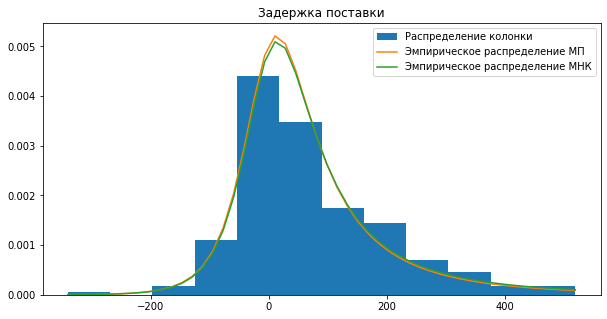


Рисунок 5. Сравнение реального и эмпирического распределения для колонки "Задержка поставки"

Наиболее подходящим распределением оказалось распределение *norminvgauss* со значением *pvalue = 0.29 и* *0.22* для критерия Колмогорова-Смирнова и Крамера-Мизеса соответственно. Параметры распределения представлены в таблице 2.

|  |  |  |
| --- | --- | --- |
|  | α | β |
| ML | 0.796 | 0.540 |
| LSM | 0.785 | 0.559 |

Таблица 2. Параметры распределения norminvgauss

На рисунке 6 представлен q-q биплот для св *“Задержка поставки”,* распределённой согласно norminvgauss

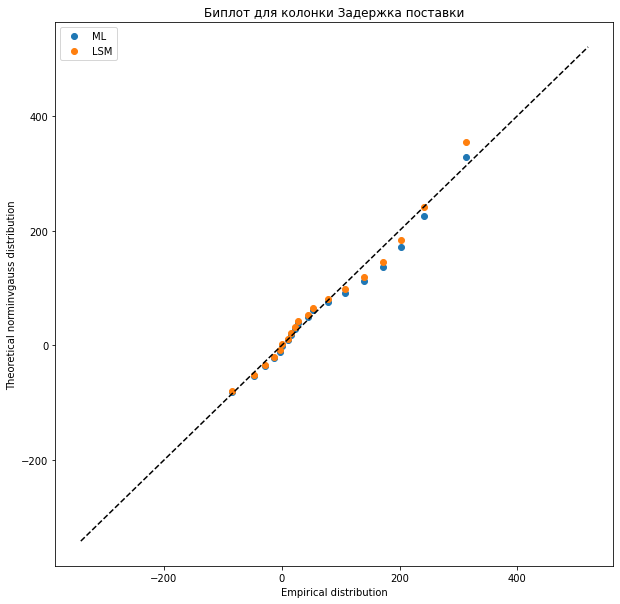


Рисунок 6. Квантильный биплот для св “Задержка поставки”, распределённой согласно norminvgauss

При моделировании непрерывной св *“План.сумма с НДС 18%”*, лучшим подобранным распределением оказалось *genextreme* со значением *pvalue = 0.99 и 0.98* для критериев Колмогорова-Смирнова и Крамера-Мизеса соответственно. Его график представлен на рисунке 7.

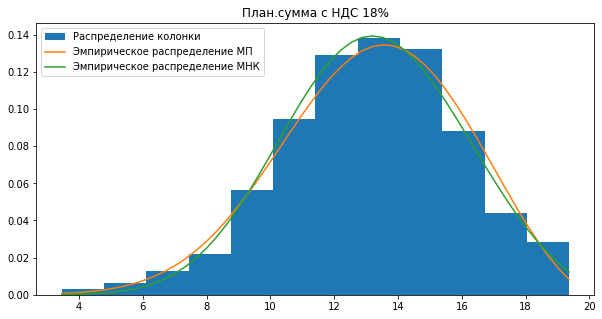


Рисунок 7. Сравнение реального и эмпирического распределения для колонки " План.сумма с НДС 18%"

Параметры распределения представлены в таблице 3.

|  |  |  |
| --- | --- | --- |
|  | *c* | *s* |
| ML | 2.691 | 3.676 |
| LSM | 2.564 | 3.704 |

Таблица 3. Параметры распределения genextreme

На рисунке 8 представлен q-q биплот для св *“План.сумма с НДС 18%”*, распределённой согласно *genextreme.*

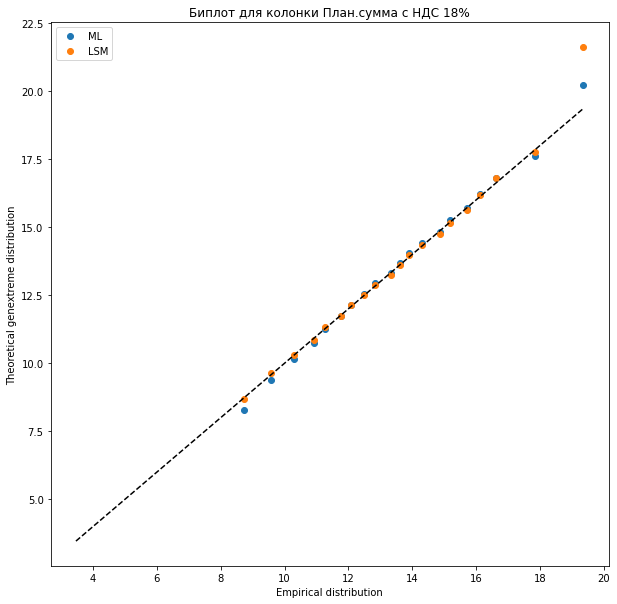
**

Рисунок 8. Квантильный биплот для св “План.сумма с НДС 18%”, распределённой согласно powerlognorm

При моделировании непрерывной св *“Курс2”*, лучшим подобранным распределением оказалось *bradford* со значением *pvalue = 0.01 и 0.02* для критерия Колмогорова-Смирнова и Крамера-Мизеса соответственно. Его график представлен на рисунке 9.

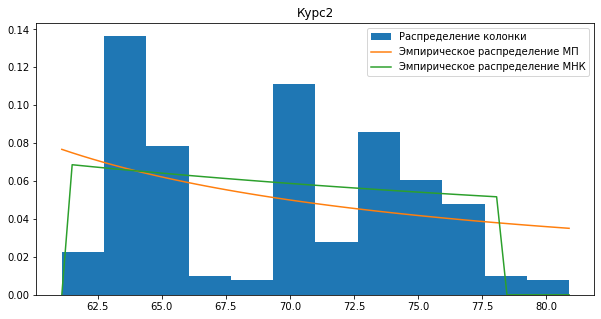


Рисунок 9. Сравнение реального и эмпирического распределения для колонки "Курс2"

Поскольку значение *pvalue = 0.01 и 0.02 < 0.05*, гипотезу о том, что св *“Курс2”* распределена согласно распределению *bradford* следует отвергнуть.

Из рисунка 9 видно, что распределение СВ мультимодально, а значит стоит попробовать подобрать смесь гауссовых распределений для её описания.

## Моделирование смеси распределений Гаусса

Для моделирования распределения колонки *“Курс2”* была подобрана смесь 2-х распределений Гаусса, представленная на рисунке 10.

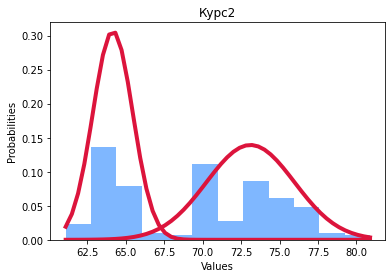


Рисунок 10. Смесь распределений Гаусса, описывающих св "Курс2"

Параметры компонент гауссовой смеси представлены на рисунке 11.

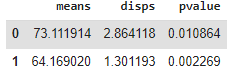


Рисунок 11. Параметры компонент гауссовой смеси (means – мат.ожидание, disps – дисперсия, pvalue – значение теста Колмогорова Смирнова)

Значения *pvalue* у обеих компонент *< 0.05,* поэтому гипотезу о нормальности распределения выделенных кластеров следует отвергнуть.

Тем не менее, каждый из 2-х полученных кластеров можно отдельно проверить на всех непрерывных распределениях из библиотеки *scipy.stats*

Лучшим распределением кластера под номером 0 оказалось *gengamma* со значением *pvalue = 0.17 и 0.26* для критериев Колмогорова-Смирнова и Крамера-Мизеса соответственно.

Его график представлен на рисунке 12.

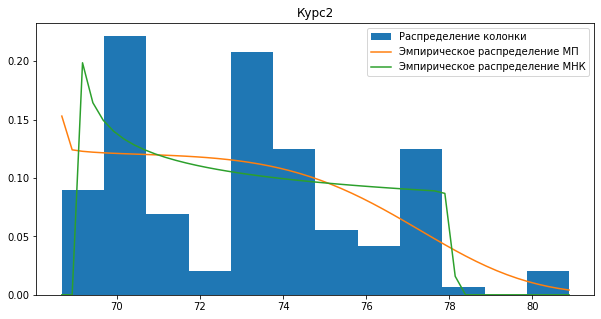


Рисунок 12. Сравнение реального и эмпирического распределения кластера 0 колонки "Курс2"

На рисунке 13 представлен q-q биплот для кластера 0 колонки *"Курс2"*

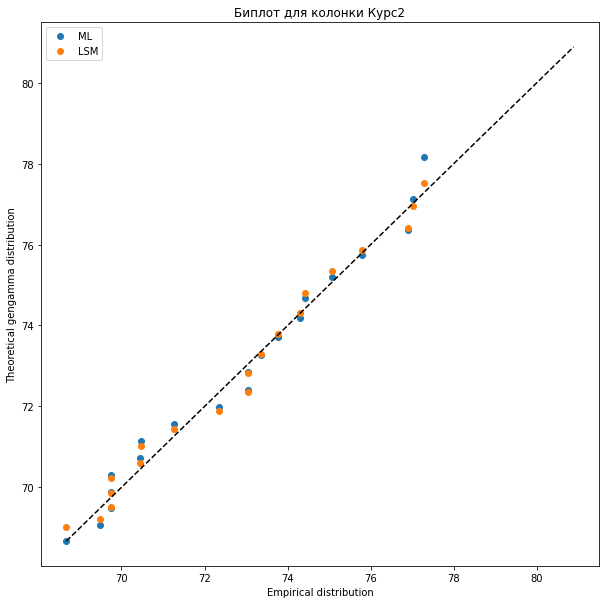


Рисунок 13. Квантильный биплот для кластера 0 колонки "Курс2", распределённого согласно gengamma

Лучшим распределением кластера под номером 1 оказалось *moyal* со значением *pvalue = 0.10*

Его график представлен на рисунке 14.

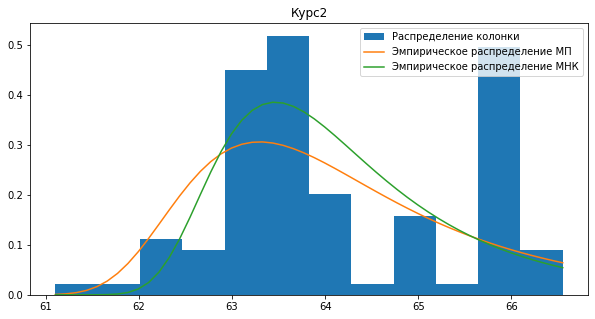


Рисунок 14. Сравнение реального и эмпирического распределения кластера 1 колонки "Курс2"

На рисунке 15 представлен q-q биплот для кластера 1 колонки *"Курс2"*

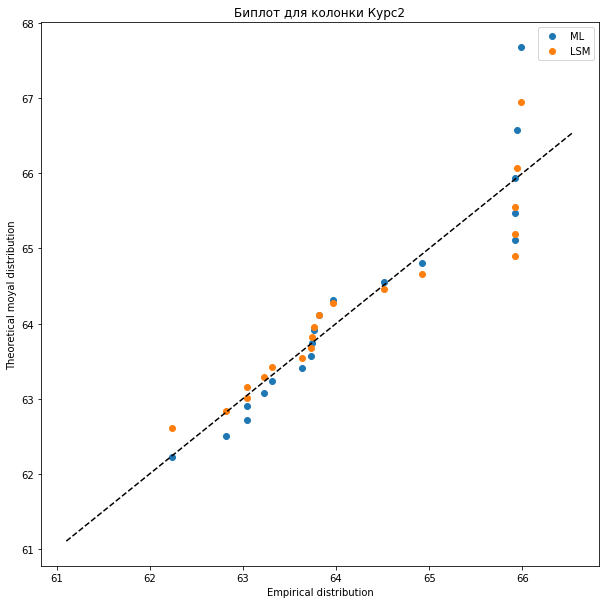


Рисунок 15. Квантильный биплот для кластера 1 колонки "Курс2", распределённого согласно moyal

Таким образом, непрерывную СВ *“Курс2”* можно представить как смесь из распределений *gengamma и moyal*.

Распределение СВ *“Курс2”* представлено на рисунке 16.

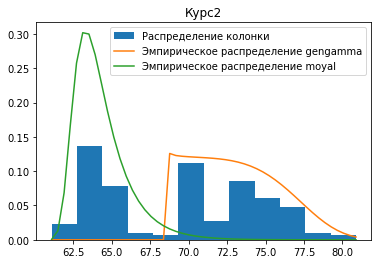


Рисунок 16. Сравнение реального и эмпирического распределения для колонки "Курс2"

# 5. Сэмплирование

Для сэмплирования из подобранных распределения были выбраны 2 метода:

1. Метод обратной функции
2. Ratio-Of-Uniforms

При сэмплировании из распределения методом обратной функции необходимо определить кумулятивную функцию распределения вероятностей (CDF) распределения и вычислить значения переменных от равномерно распределенных значений вероятностей. Сэмплирование производиться с помощью метода PPF из библиотеки *scipy.* Метод возвращает значения СВ при заданном значении вероятности.

Метод Ratio-Of-Uniforms является одним из видов метода accept reject сэмплинга. Для сэмплирования подбирается равномерное распределение, полностью накрывающее исследуемое распределение. Пример представлен на рисунке 17.

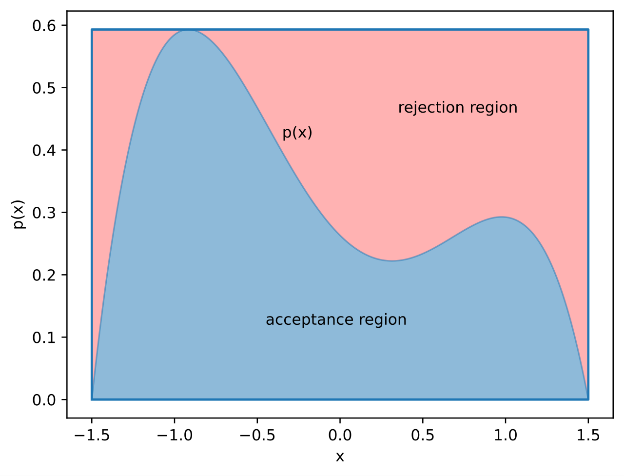
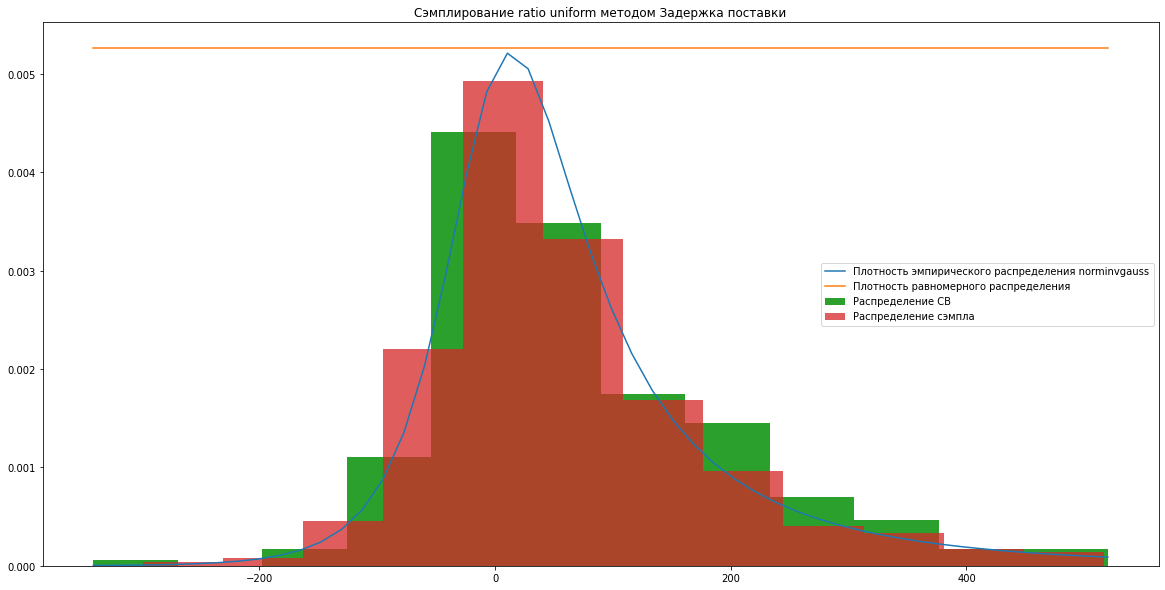
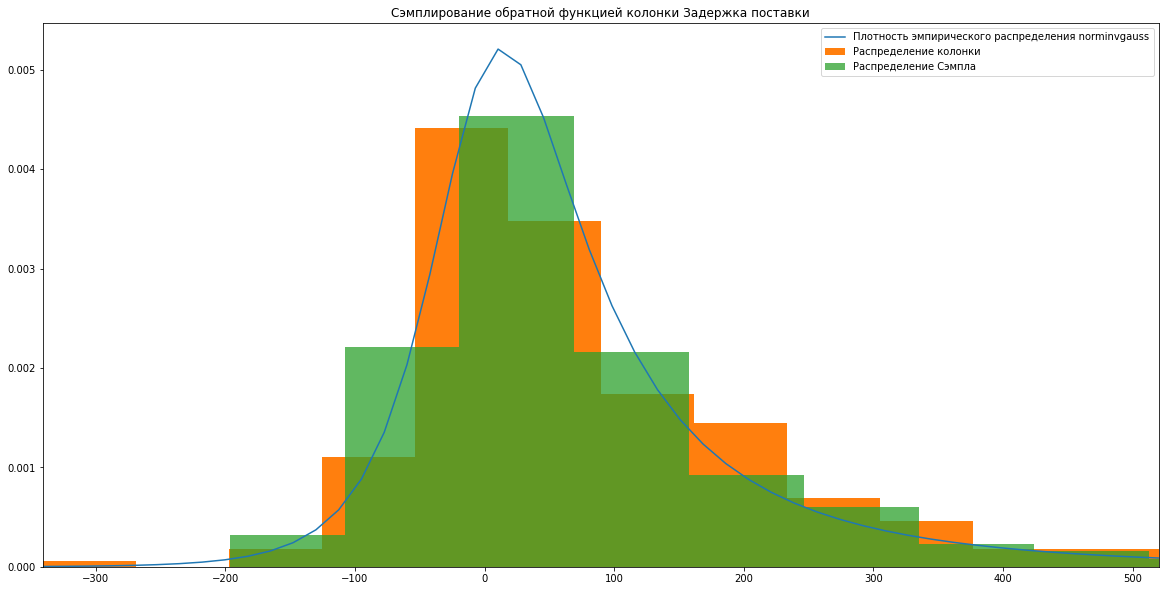


Рисунок 17. Ratio-Of-Uniforms method example

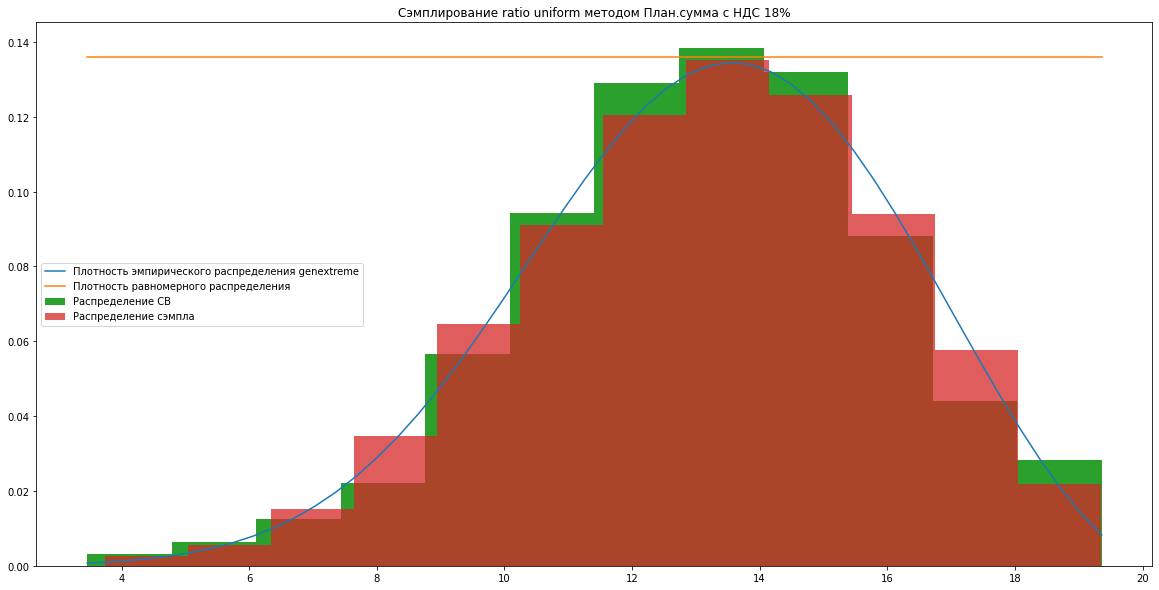
Если сэмплированное значение попало в *acceptance region*, то его значение принимается, иначе – отвергается и в выборку сэмплов не попадает.

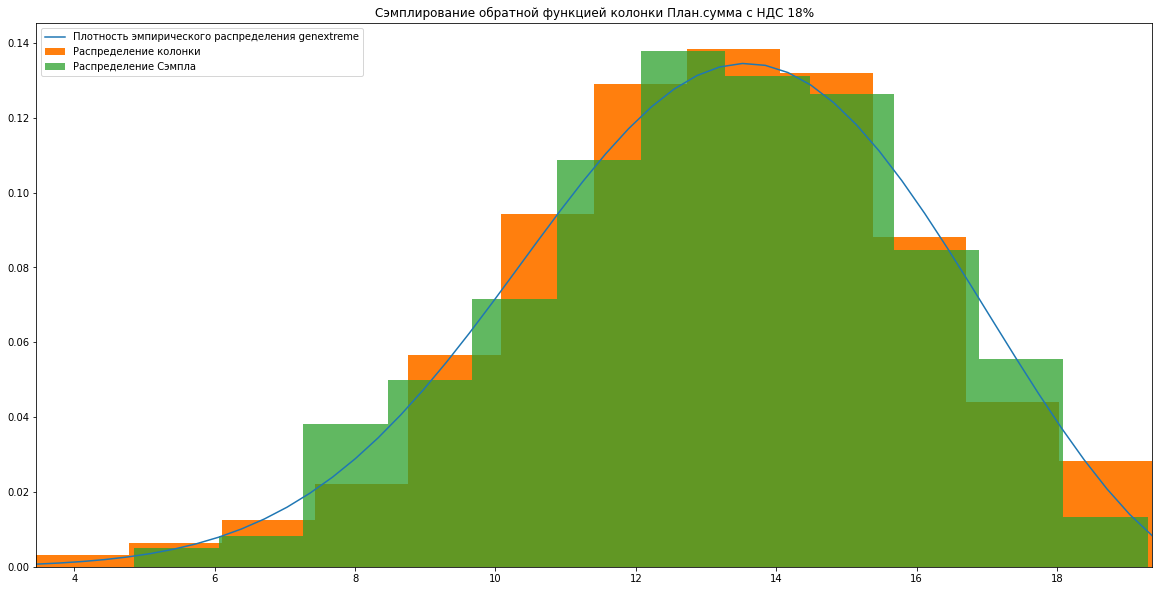
Результаты сэмплирования 2-мя методами представлены на рисунках:



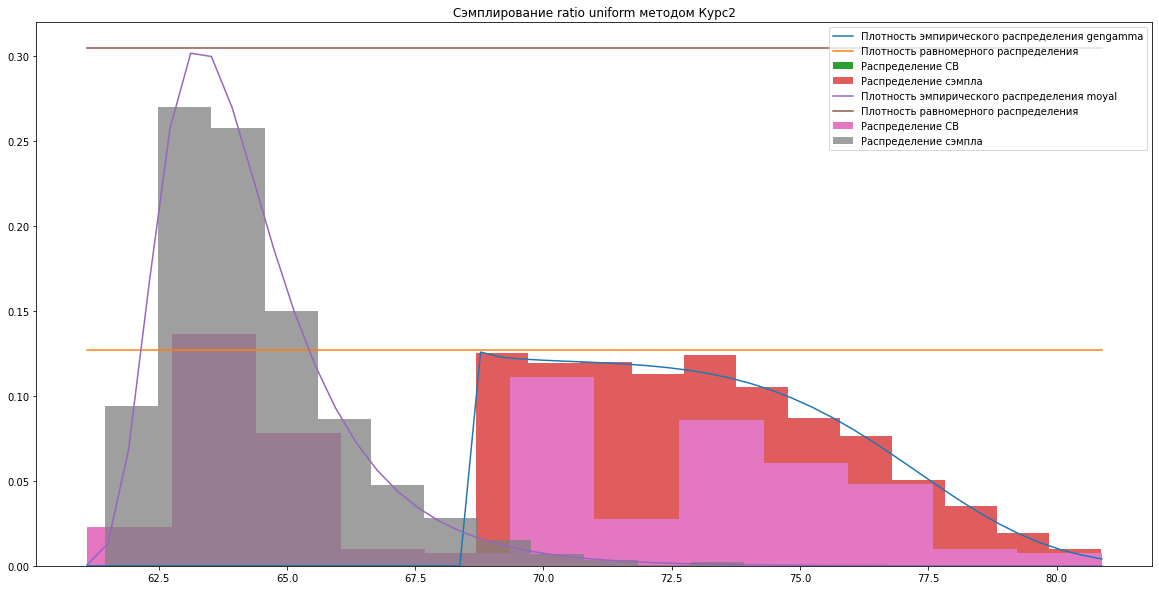


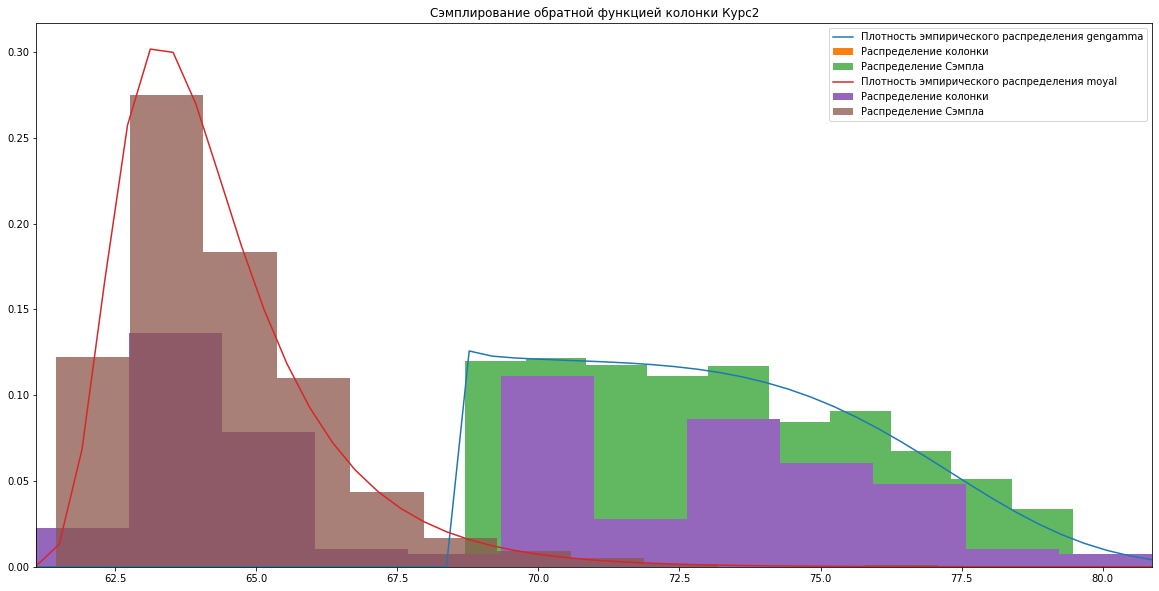
*Рисунки 18-19. Сэмплирование колонки “Задержка поставки”*





*Рисунки 20-21. Сэмплирование колонки “План.сумма с НДС 18%”*





*Рисунки 22-23. Сэмплирование колонки “Курс2”*

# Исходный код

Ссылка на GitHub: <https://github.com/PabloKarpacho/Lab1>