Отчет по лабораторной работе № 1

Анализ одномерных случайных величин

Выполнено:

Леженков Г.Г.

J4151

Санкт-Петербург

2022

Оглавление

[1. Обоснование выбора данных: 2](#_Toc118304618)

[2. Порядковая статистика: 2](#_Toc118304619)

[3. Построение непараметрической оценки PDF в виде гистограммы и использование функции плотности ядра: 3](#_Toc118304620)

[4. Подбор наилучших распределений, отражающих данные: 4](#_Toc118304621)

[5. Сэмплирование 10](#_Toc118304622)

# **Обоснование выбора данных:**

Датасет имеет размер 6 столбцов x 370 строк. В нем содержится информация о диагностических данных женщин-пациенток из национального института диабета, болезней органов пищеварения и почек.

Название колонок и случайные величины предоставлены в таблице ниже:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Глюкоза  **Glucose** | Кровяноедавление  **BloodPressure** | Инсулин  **Insulin** | Индексмассытела  **BMI** | Функцияродословнойдиабета  **Dp\_func** | Результат  **Outcome** |
| … | … | … | … | … | … |

Для исследования были выбраны следующие непрерывные СВ:

1. **Glucose**
2. **BloodPressure**
3. **Insulin**
4. **BMI**
5. **Dp\_func**

Величины Outcome имеет тип Булево.

# **Порядковая статистика:**

Выводим описательную и порядковую статистику для непрерывных СВ:

Изображение выглядит как стол

Автоматически созданное описание

Рис. 1 Статистика по непрерывным СВ

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

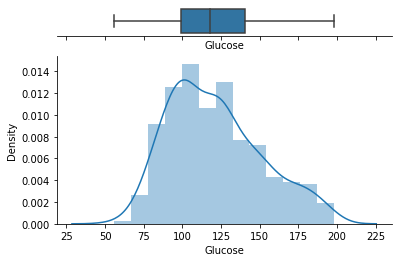


Рис 2 Ядерная оценка плотности столбца «Glucose»

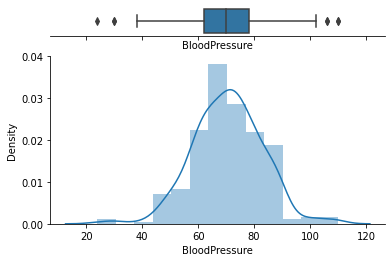


Рис 3 Ядерная оценка плотности столбца «BloodPressure»

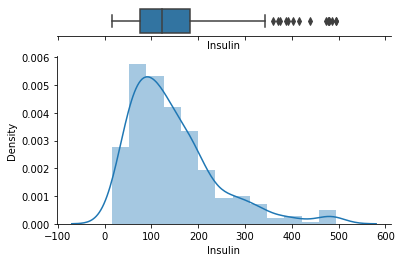


Рис 4 Ядерная оценка плотности столбца «Insulin»

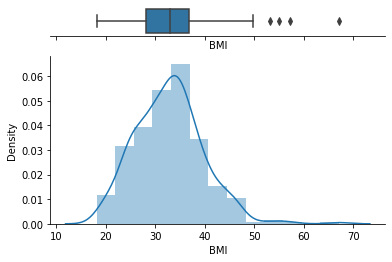


Рис 5 Ядерная оценка плотности столбца «BMI»

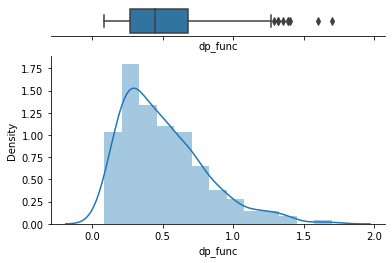


Рис 6 Ядерная оценка плотности столбца «dp\_func»

# **Подбор наилучших распределений, отражающих данные:**

Для получения наилучшего распределения для непрерывных СВ произведено сравнение всех непрерывных распределений из библиотеки *scipy.stats* посредством поиска максимального значения Pvalue теста Колмогорова-Смирнова. Также после выбора и построения распределений было подсчитано значение Pvalue критерия Крамера-Мизеса-Смирнова.

Распределения выбирались на основе 2 методов: метод наименьших квадратов (МНК) и метода максимального правдоподобия (МП).

Таким образом получаем:

1. **Glucose**

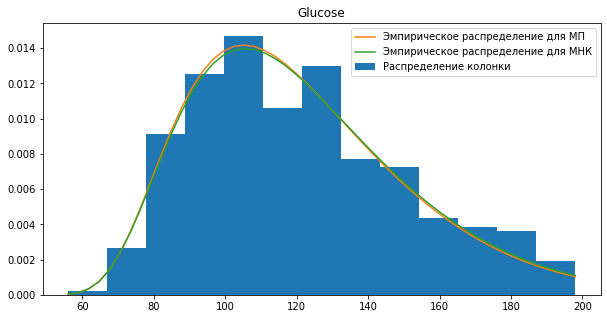


Рис 7 сравнение реального и эмпирического распределения для столбца Glucose

Наиболее подходящим распределением получилось kstwobign со значениями для критерия Колмогорова-Смирнова и критерия Крамера-Мизеса *pvalue = 0.730, 0.762* соответственно.

Параметры представлены в таблице ниже:

|  |  |  |
| --- | --- | --- |
| **метод** | **α** | **β** |
| ML | 17.6 | 118.9 |
| LSM | 16.8 | 120.6 |

Табл. 1 параметры для распределения *kstwobign*

Также был построен q-q биплот для данной СВ, распределенной по *kstwobign*

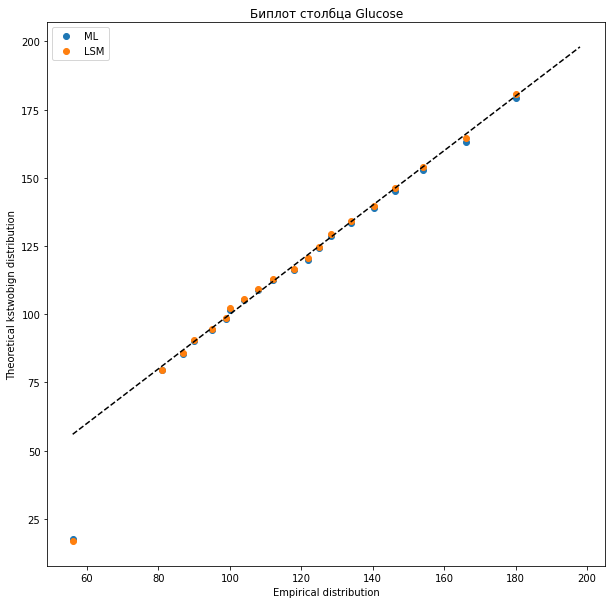


Рис 8 q-q биплот для СВ «Glucose», распределенной по *kstwobign*

1. **BloodPressure**

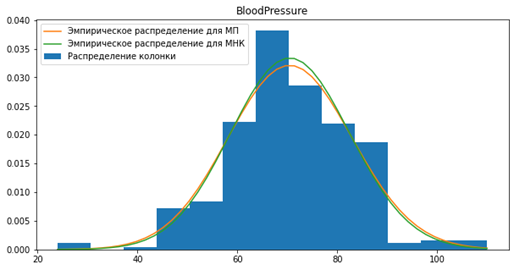


Рис 9 сравнение реального и эмпирического распределения для столбца *BloodPressure*

Наиболее подходящим распределением получилось *fatiguelife* со значениями для критерия Колмогорова-Смирнова и критерия Крамера-Мизеса *pvalue = 0.428, 0.516* соответственно.

Параметры представлены в таблице ниже:

|  |  |  |
| --- | --- | --- |
| **метод** | **α** | **β** |
| ML | 0.01 | -1111 |
| LSM | 0.01 | -1111 |

Табл. 2 параметры для распределения *fatiguelife*

Также был построен q-q биплот для данной СВ, распределенной по *fatiguelife*

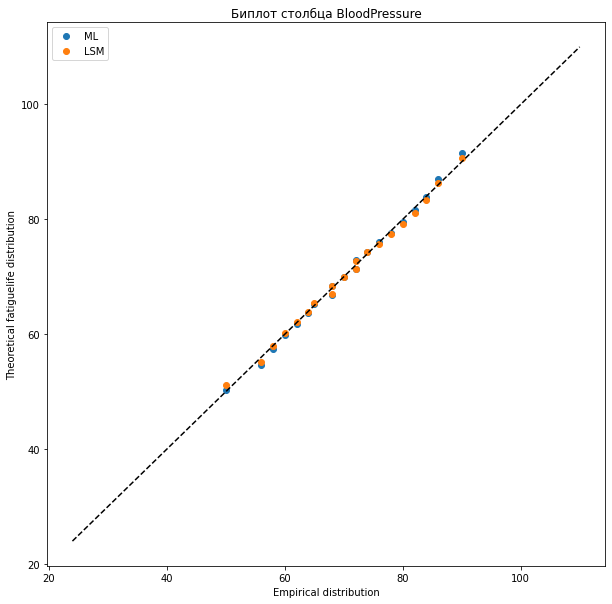


Рис 10 q-q биплот для СВ «BloodPressure», распределенной по *fatiguelife*

**3)    Insulin**

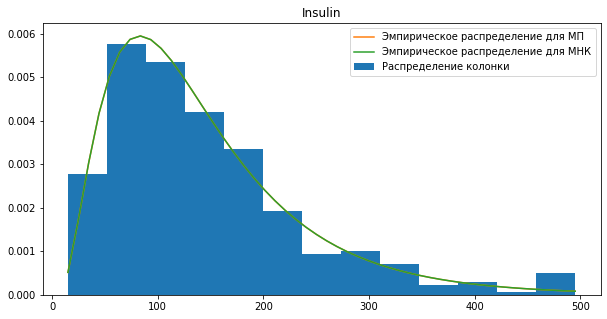


Рис 11 сравнение реального и эмпирического распределения для столбца Insulin

Наиболее подходящим распределением получилось *exponweib* со значениями для критерия Колмогорова-Смирнова и критерия Крамера-Мизеса *pvalue = 0.998, 0.995* соответственно.

Параметры представлены в таблице ниже:

|  |  |  |
| --- | --- | --- |
| **метод** | **α** | **β** |
| ML | 3.809 | 0.870 |
| LSM | 3.809 | 0.870 |

Табл. 3 параметры для распределения *exponweib*

Также был построен q-q биплот для данной СВ, распределенной по *exponweib*

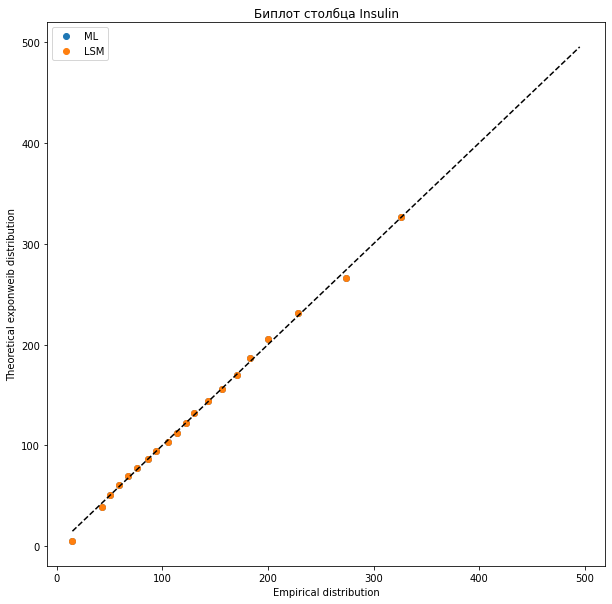


Рис 12 q-q биплот для СВ «Insulin», распределенной по *exponweib*

**4)    BMI**

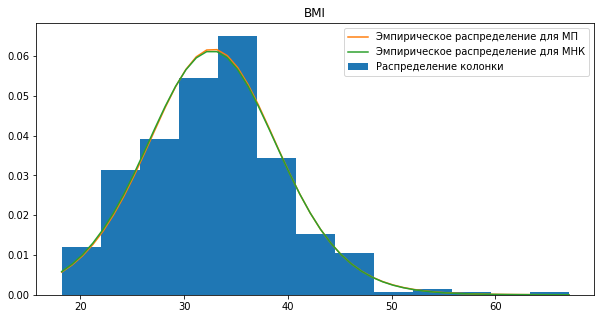


Рис 13 сравнение реального и эмпирического распределения для столбца BMI

Наиболее подходящим распределением получилось t со значениями для критерия Колмогорова-Смирнова и критерия Крамера-Мизеса *pvalue = 0.813, 0.802* соответственно.

Параметры представлены в таблице ниже:

|  |  |  |
| --- | --- | --- |
| **метод** | **α** | **β** |
| ML | 13.44 | 32.73 |
| LSM | 15.39 | 32.67 |

Табл. 4 параметры для распределения *t*

Также был построен q-q биплот для данной СВ, распределенной по *t*

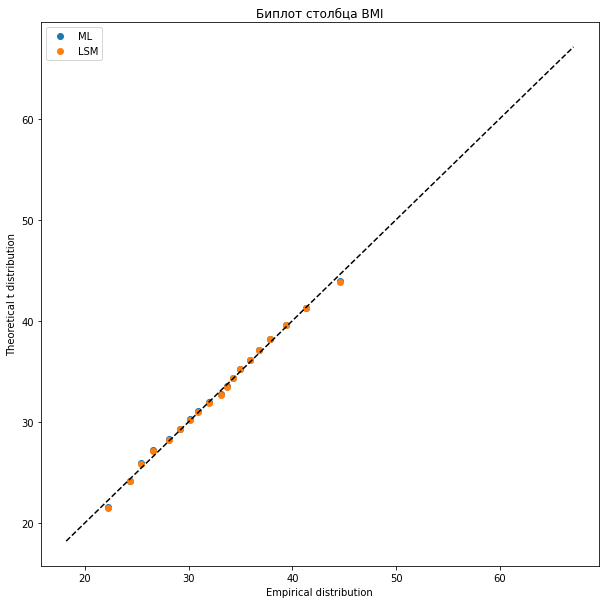


Рис 14 q-q биплот для СВ «BMI», распределенной по *t*

**5)    dp\_func**

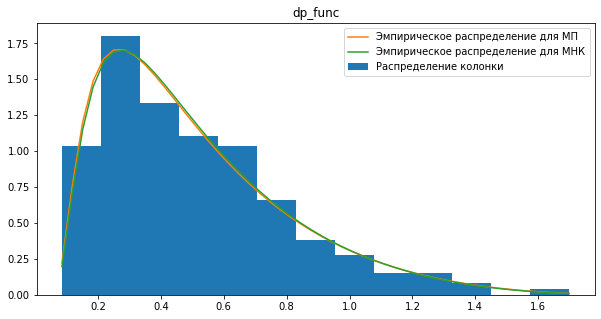


Рис 15 сравнение реального и эмпирического распределения для столбца dp\_func

Наиболее подходящим распределением получилось johnsonsb со значениями для критерия Колмогорова-Смирнова и критерия Крамера-Мизеса *pvalue = 0.994, 0.943* соответственно.

Параметры представлены в таблице ниже:

|  |  |  |
| --- | --- | --- |
| **метод** | **α** | **β** |
| ML | 1.56 | 1.074 |
| LSM | 1.56 | 1.088 |

Табл. 5 параметры для распределения *johnsonsb*

Также был построен q-q биплот для данной СВ, распределенной по *johnsonsb*

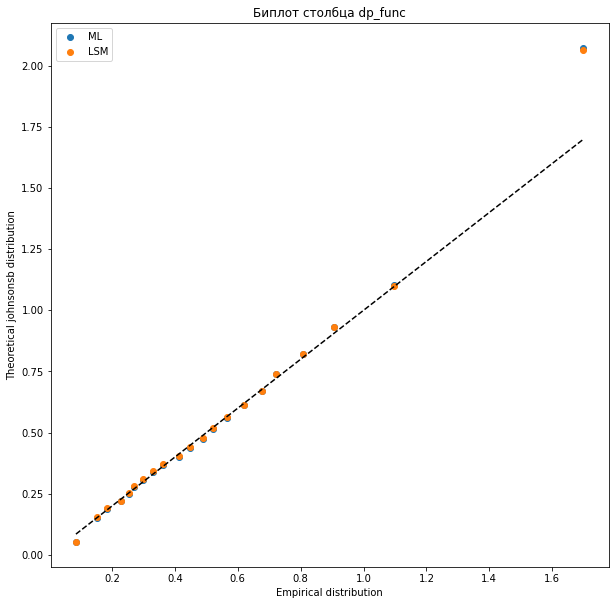


Рис 16 q-q биплот для СВ «dp\_func», распределенной по johnsonsb

# **Сэмплирование**

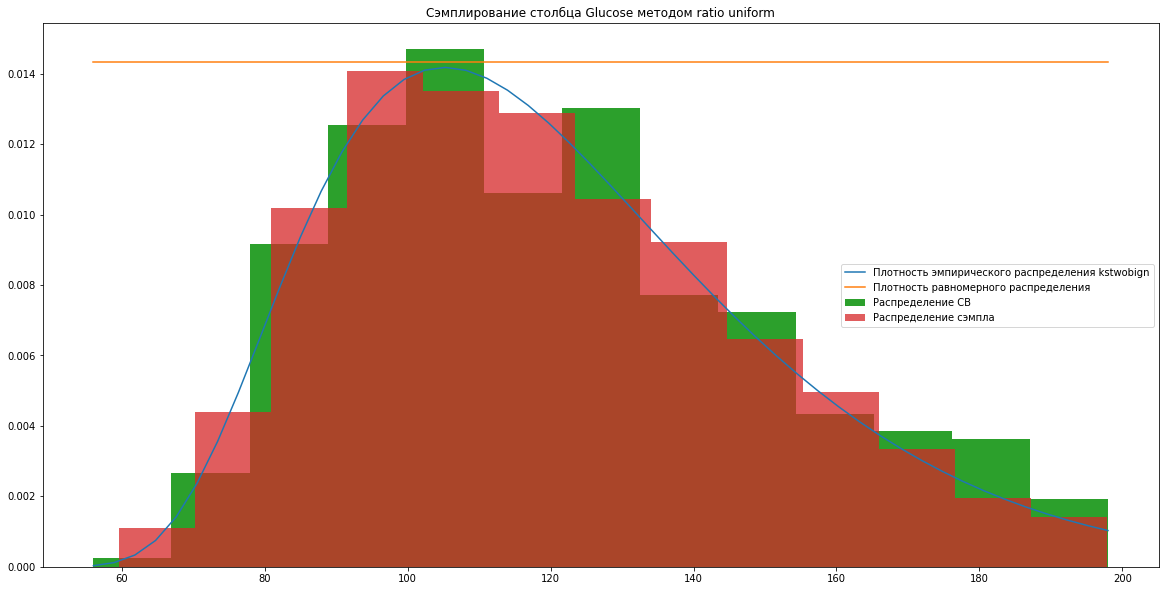
Для сэмплирования были выбраны 2 метода:

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

Метод Ratio-Of-Uniforms является подметодом accept reject сэмплинга. Подбирается равномерное распределение, покрывающее текущее распределение.

Если сэмплирование попадает в область acceptance region, то оно принимается, иначе отвергается и, соответсвенно не попадает в выборку сэмплов.

При сэмплировании методом обратной функцией в начале необходимо определиться с кумулятивной функцией распределения вероятности – CDF. Далее вычисляются значения переменных от равномерно распределенных значений вероятностей. Сэмплирование производиться с помощью метода PPF из библиотеки *scipy.* Метод возвращает значения СВ при заданном значении вероятности.



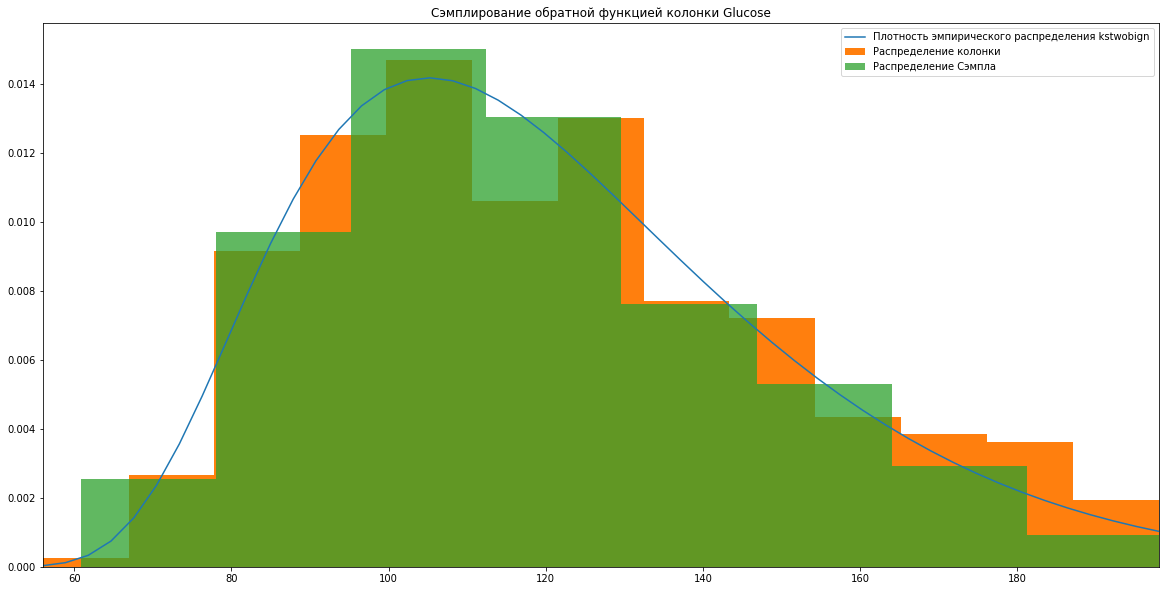


Рис 17-18 Сэмплирование столбца Glucose

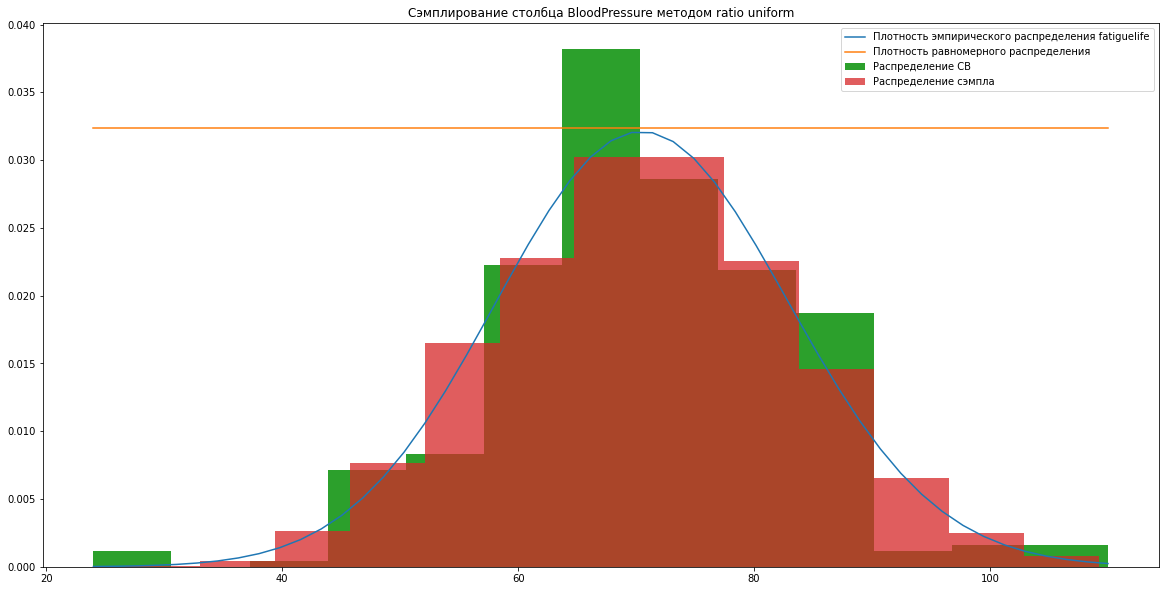
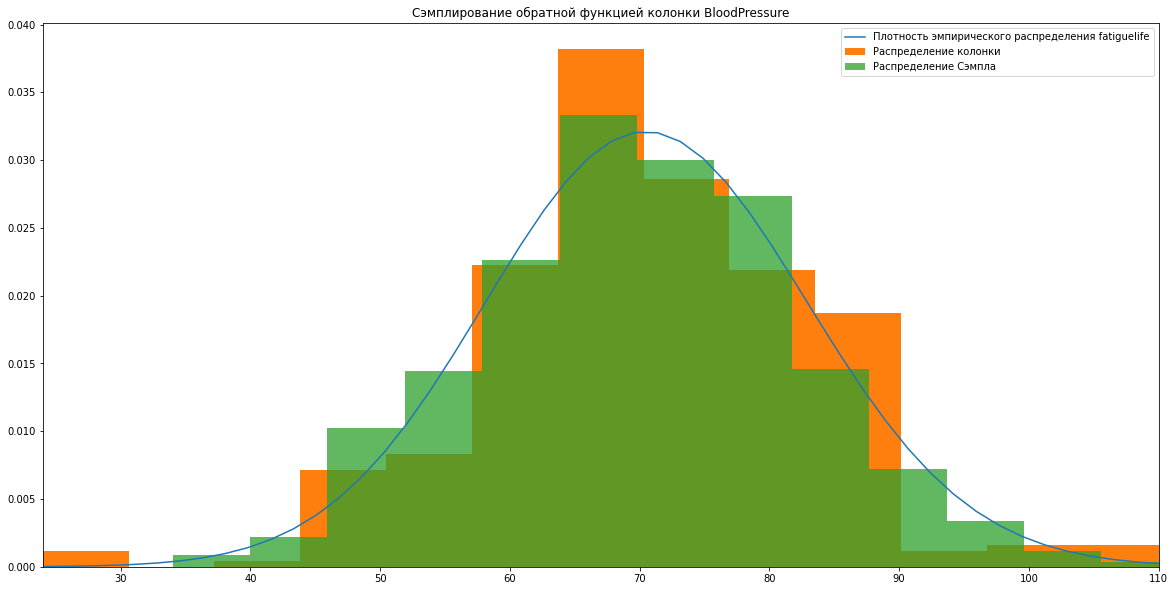
 

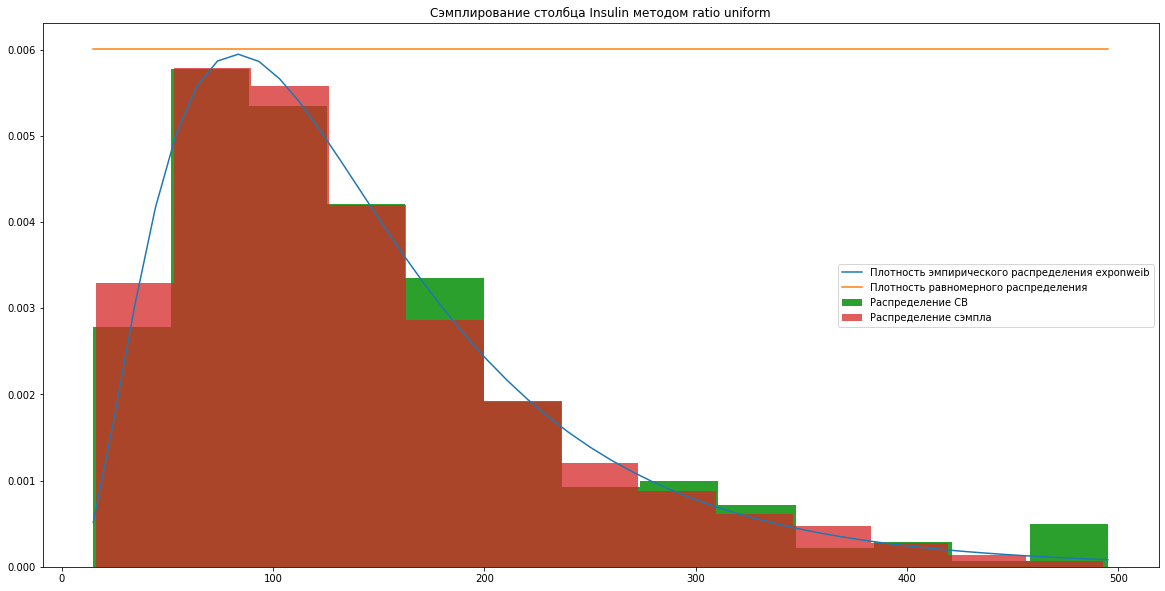
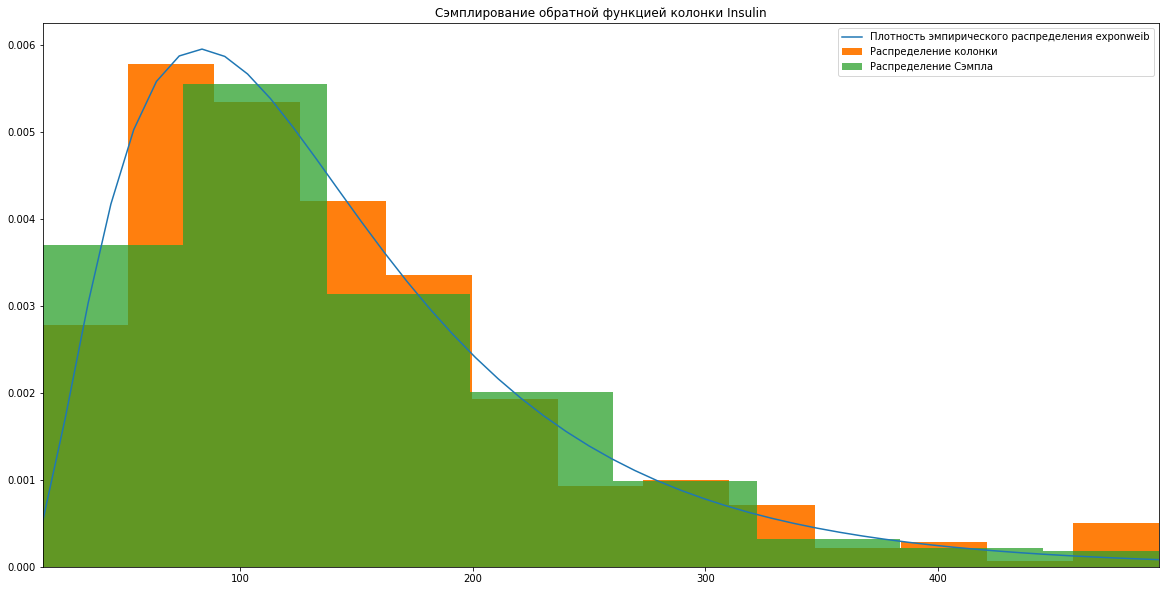
Рис 19-20 Сэмплирование столбца BloodPressure  

Рис 21-22 Сэмплирование столбца Insulin

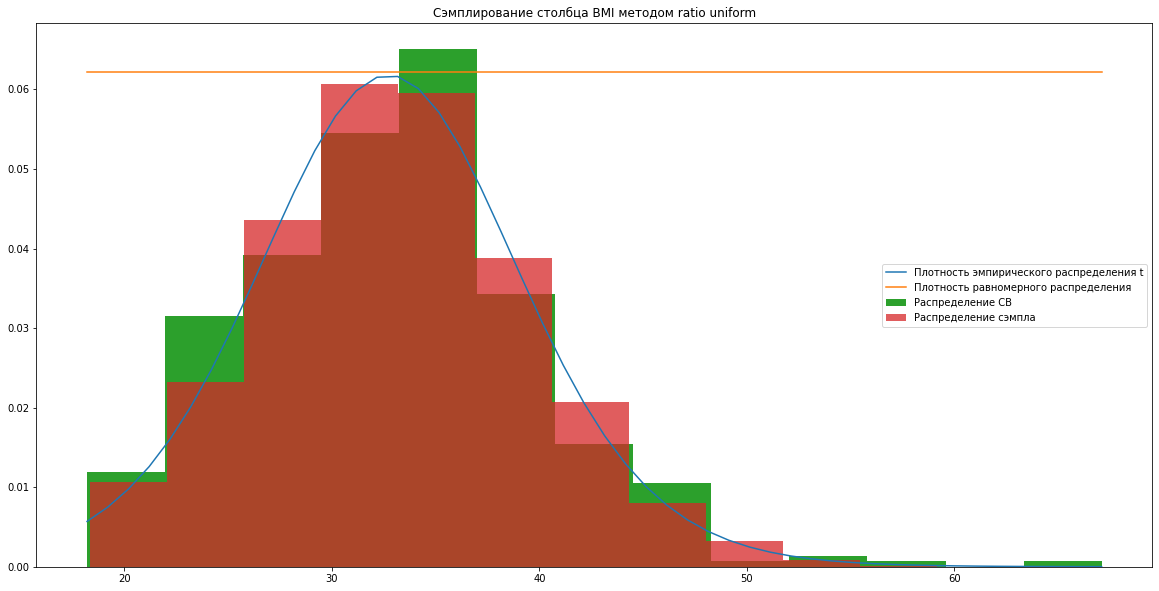
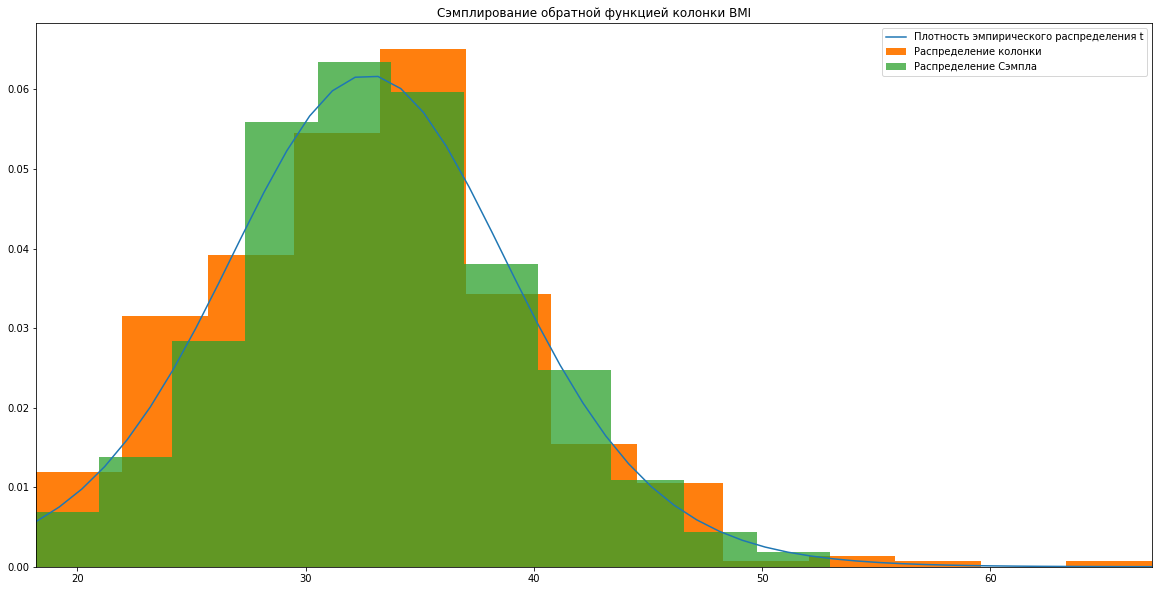
 

Рис 23-24 Сэмплирование столбца BMI

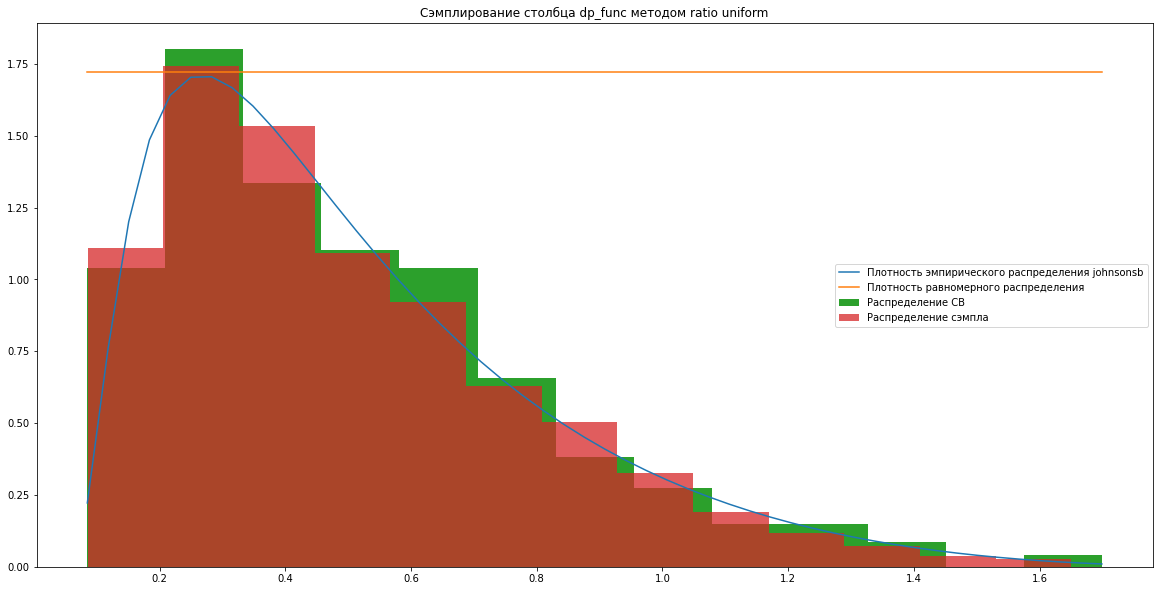
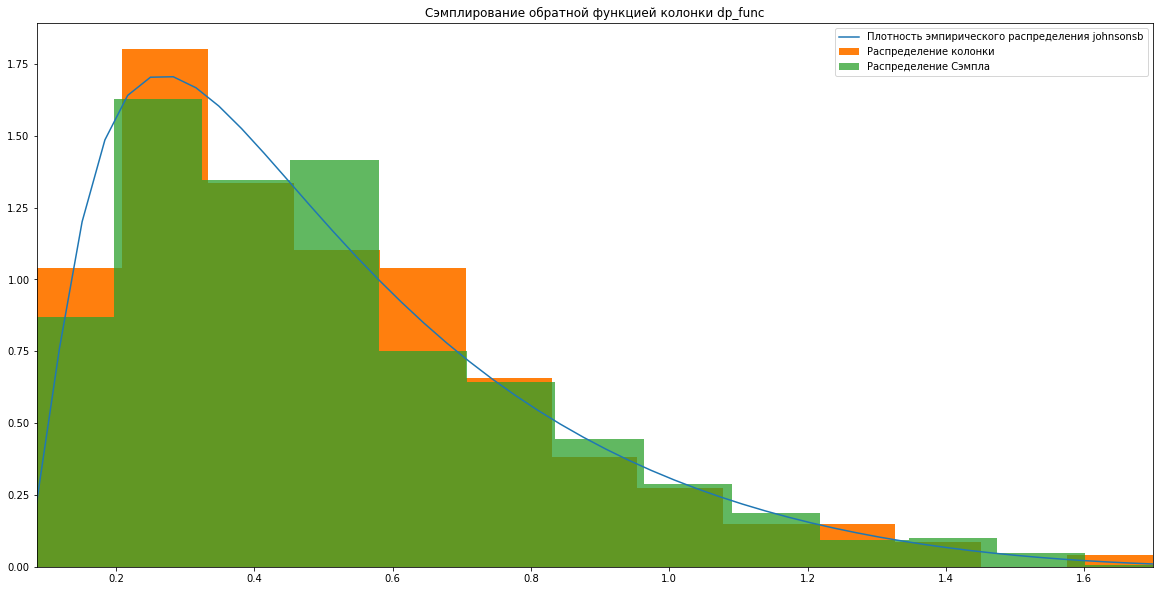
 

Рис 25-26 Сэмплирование столбца dp\_func

**Исходный код:** <https://github.com/Lezhenkov/Lab1>