Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»

Факультет программной инженерии и компьютерной техники

Расчетно-графическая работа №1 по дисциплине «Математическая статистика»

Вариант 1, Закон Бернулли

Выполнили**:** Вавилина Екатерина, Медведева Даниэла
Группа: P3230
Преподаватель: Лукина Марина Владимировна

Санкт-Петербург
2025

Цель работы

Цель данной работы- исследование закона распределения Бернулли, оценка его параметров методом моментов и методом максимального правдоподобия, а также анализ выборки, распределенной по равномерному закону.

1. Генерация выборки

Возьмем вероятность выпадения 1 равной 0.
Таким образом будем симулировать эксперимент с подбрасыванием монетки, где выпадение 1 равно выпадению орла, а выпадение 0 равно выпадению решки. Другие варианты исключены.

Код для генерации выборки из 30 значений(это делается с помощью функции np.random.binomial, где мы задаем кол-во испытаний и вероятность успеха):

```
import numpy as np

# Исходная выборка (Бернулли)
np.random.seed(10)
sample = np.random.binomial(1, 0.5, 30)
```

Получившиеся значения:
[1 0 1 1 0 0 0 1 0 0 1 1 0 1 1 1 1 0 1 1 1 0 0 1 0 0 1 1 1 1]

2. Оценка параметра распределения по методу моментов

Метод моментов - это способ оценить параметры распределения, используя *математическое ожидание и дисперсию.*
Метод моментов основан на том, что параметры распределения связаны с его моментами, такими как *математическое ожидание и дисперсия.* Если мы знаем, как выражается математическое ожидание через параметр pp, то можем найти его оценку, просто заменив ожидание на *выборочное среднее*.

1. Запись формулы для вычисления математического ожидания через параметр распределения

Для случайной величины X, распределенной по *закону Бернулли* с параметром p, математическое ожидание определяется как:

$$E(X) = p = 0.5$$

где p — вероятность успеха (выпадения единицы).

2. Выражение параметра распределения через математическое ожидание

Так как $E(X) = p$, то отсюда сразу следует, что:

$$p = E(X) = 0.5$$

3. Вычисление оценки параметра, подставив в формулу оценку математического ожидания

Метод моментов заменяет математическое ожидание на его выборочную оценку — выборочное среднее:

$$E(X) = \frac{1}{n}\sum_{i=1}^{n} \quad x_i = \frac{1}{30}(1 + 0 + 1 + 1 + \cdots + 1 + 1 + 1 + 1) = \frac{18}{30} = 0.6$$

Тогда: $p = E(X) = 0.6$

4. Запись формулы для вычисления дисперсии через параметр распределения

Для распределения Бернулли дисперсия вычисляется по формуле:

и для p= 0.5 дисперсия будет:

$$D(X) = p(1 - p) = 0.25$$

5. Выражение параметра распределения через дисперсию

Мы знаем, что дисперсия для распределения Бернулли равна p(1−p)p(1−p). Подставляем выборочную дисперсию в эту формулу:

$$D(X) = p - p^2 \text{ или } p^2 - p + D(X) = 0$$

Это квадратное уравнение относительно p, решая его, получаем:

$$p = \frac{1 \pm \sqrt{1 - 4D(X)}}{2} = \frac{1 \pm \sqrt{1 - 4 \cdot 0{,}25}}{2} = \frac{1 \pm \sqrt{1 - 1}}{2} = 0{,}5$$

Нас интересует результат в интервале [0;1].

6. Вычисление оценки параметра, подставив в формулу оценку дисперсии

Выборочная дисперсия рассчитывается по формуле:

$$D_{несм}(X) = \frac{1}{n - 1}\sum_{i=1}^{n} \quad (x_i - E(X))^2 \approx 0.2483$$

$$p = \frac{1 \pm \sqrt{1 - 4 \cdot 0.2483}}{2} = \frac{1 \pm 0.0824}{2}$$

Тогда есть 2 возможных корня:

$p = 0.4588$ и $p = 0.5412$

Из двух корней мы выбираем тот, который ближе к выборочному среднему, т.е. p=0.5412(оно ближе к 0.6)

7. Сравнение полученных оценок между собой и с точным значением параметра, заложенном при моделировании

В результате оценки мы получили $E(X) = 0.6$ вместо ожидаемого 0.5 и $D_{несм}(X) = 0.2483$ вместо ожидаемого 0.25.

Это связано с относительно малым размером выборки. *Чем больше размер выборки, тем точнее будут результаты (об этом говорит закон больших чисел).*

Метод моментов через математическое ожидание дал p = 0.6. Оценка может быть завышена или занижена в зависимости от конкретной выборки (в данном случае завышена). При увеличении размера выборки оценка становится точнее.

Метод моментов через исправленную дисперсию дал два корня $p = 0.4588$ и $p = 0.5412$. Истинное значение 0.5 лежит между ними, что логично, так как уравнение $D(X) = p(1 - p) = 0.25$ является квадратным.

Какой корень выбрать? По смыслу правдоподобнее взять корень, ближайший к оценке по мат. ожиданию. Здесь $p = 0.5412$ ближе к 0.6, поэтому его предпочтительнее использовать.

Все полученные значения отличаются от того, что было задано при генерации выборки (0.5). Самыми близкими значениями являются $p = 0.4588$ (заниженное значение) и $p = 0.5412$ (завышенное значение). *Если бы у нас вероятность изначально была отлична от 0.5, ближайшим был бы только 1 из 2 корней.*

8*. Код программы (для проверки расчетов)

Файл lab_1_task_2.py на гитхабе https://github.com/SnowLullaby/Matstat

3. Оценка параметра распределения по методу максимального правдоподобия

1. Построим функцию максимального правдоподобия на основании формулы вероятности

Вероятность закона Бернулли:

$$P(X_i = 1) = 0.5 \text{ и } P(X_i = 0) = 0.5$$

Тогда функция правдоподобия для всех испытаний запишется в виде:

$f(p) = p^{sum\ 1} \cdot p^{n-sum\ 1}$, где n - размер выборки (30), а sum 1 - количество 1 в выборке.

Подставим сумму в выборке (sum 1 = 18)

Получаем:

$$f(p) = p^{18} \cdot (1-p)^{30-18} = p^{18} \cdot (1-p)^{12}$$

2. Построим логарифмическую функцию максимального правдоподобия

Берем натуральный логарифм функции правдоподобия:

$$L(p) = ln(p^{18}(1-p)^{12}) = 18ln(p) + 12ln(1-p)$$

3. Найдем оценку параметра максимизируя функцию правдоподобия

Для поиска максимума продифференцируем по p:

$$\frac{\delta}{\delta p} L(p) = \frac{18}{p} - \frac{12}{1-p}$$

Приравниваем производную к нулю и решаем уравнение:

$$\frac{18}{p} - \frac{12}{1-p} = 0$$

$$\frac{18}{p} = \frac{12}{1-p}$$

$$18 - 18p = 12p$$

$$18 = 30p \rightarrow\ p = \frac{18}{30} = 0.6$$

4. Сравним оценки параметра, полученные разными методами и с точным значением параметра, заложенном при моделировании

Оценка метода максимального правдоподобия совпала с оценкой метода моментов через среднее (0.6).

Метод моментов по исправленной дисперсии дал две оценки (0.5412 и 0.4588), которые ближе к истинному значению (0.5).

Разница между истинным значением (0.5) и оценкой (0.6) обусловлена случайностью конкретной выборки, и при увеличении числа испытаний (по закону больших чисел) эта разница уменьшается, стремясь к истинному значению p.

4. Первичная обработка выборки (НСВ)

Распределение по равномерному закону

0.83	0.56	1.58	0.81	0.89	0.19	–1.77	–0.89
–3.99	–0.40	0.48	–0.74	–3.43	–3.47	–4.55	–1.51
1.29	–3.09	–0.76	0.80	–4.77	0.75	–1.13	–3.03
–3.60	–3.65	–3.67	–0.13	–3.22	–3.56	0.75	0.99
–1.12	–4.42	–4.60	–4.35	0.36	–3.80	0.18	–3.43
0.39	–3.85	1.84	–4.63				

1. Построим группированную выборку

Выберем наименьший элемент и округлим его в меньшую сторону до одной цифры после запятой наим x

Наименьшее число: -4.77, округляя получим -4.8

Выберем наибольший элемент и округлим его в большую сторону до одной цифры после запятой наиб x

Наибольшее число: 1.84, округляя получим 1.9

Выберем шаг разбиения $\frac{x_{макс}-x_{мин}}{k} = h$***, где k число интервалов*** $k = log_2 N + 1$

$k = log_2 44 + 1 = 7$ (округляем в большую сторону)

$$h = \frac{1.9 + 4.8}{7} \approx 0.96$$

Построим группированную выборку

Группированная выборка с частотами:

Интервал	Середина интервала	Частота	Относительная частота
[-4.8, -3.8)	-4.32	8	0.181818

[-3.8, -2.9)	-3.36	11	0.250000
[-2.9, -1.9)	-2.41	0	0.000000
[-1.9, -1.0)	-1.45	4	0.090909
[-1.0, -0.0)	-0.49	5	0.113636
[-0.0, 0.9)	0.46	12	0.272727
[0.9, 1.9)	1.42	4	0.090909

Частота считается по количеству чисел, попавших в интервал.

Относительная частота равна частному частоты и 44 (т.е. размера выборки)

2. По группированной выборке построим оценку функции распределения и ее график

Для построения оценки функции распределения найдем середину каждого бакета (т.е. получившегося интервала). Затем для каждой середины будем накапливать относительные частоты появления чисел, меньше данного.

Эмпирическая функция распределения:

Интервал ЭФР

x <= -4.32 0.000000

-4.32 < x <= -3.36 0.181818

-3.36 < x <= -2.41 0.431818

-2.41 < x <= -1.45 0.431818

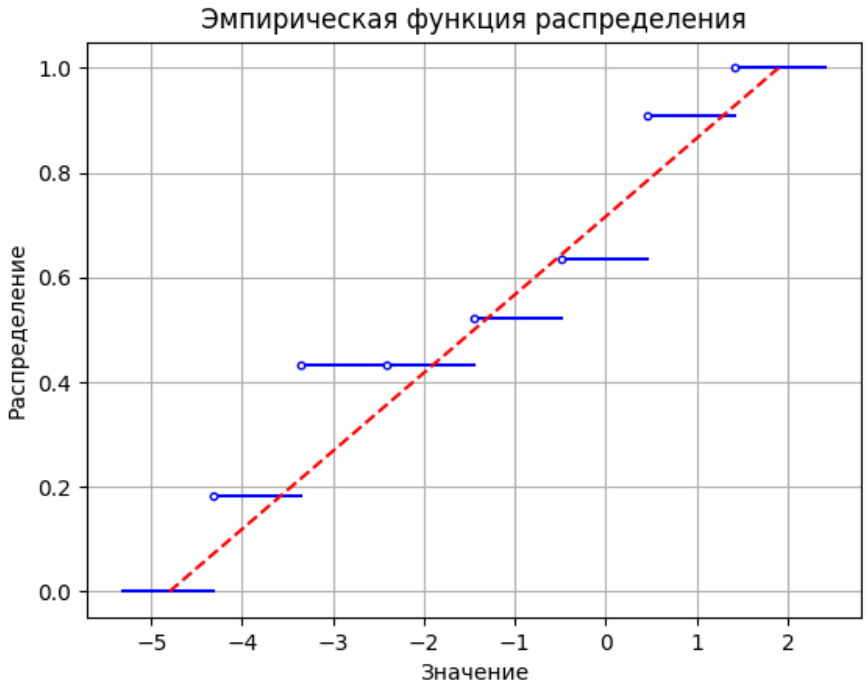
-1.45 < x <= -0.49 0.522727

-0.49 < x <= 0.46 0.636364

0.46 < x <= 1.42 0.909091

1.42 < x 1.000000

Нанесем полученные данные на график:

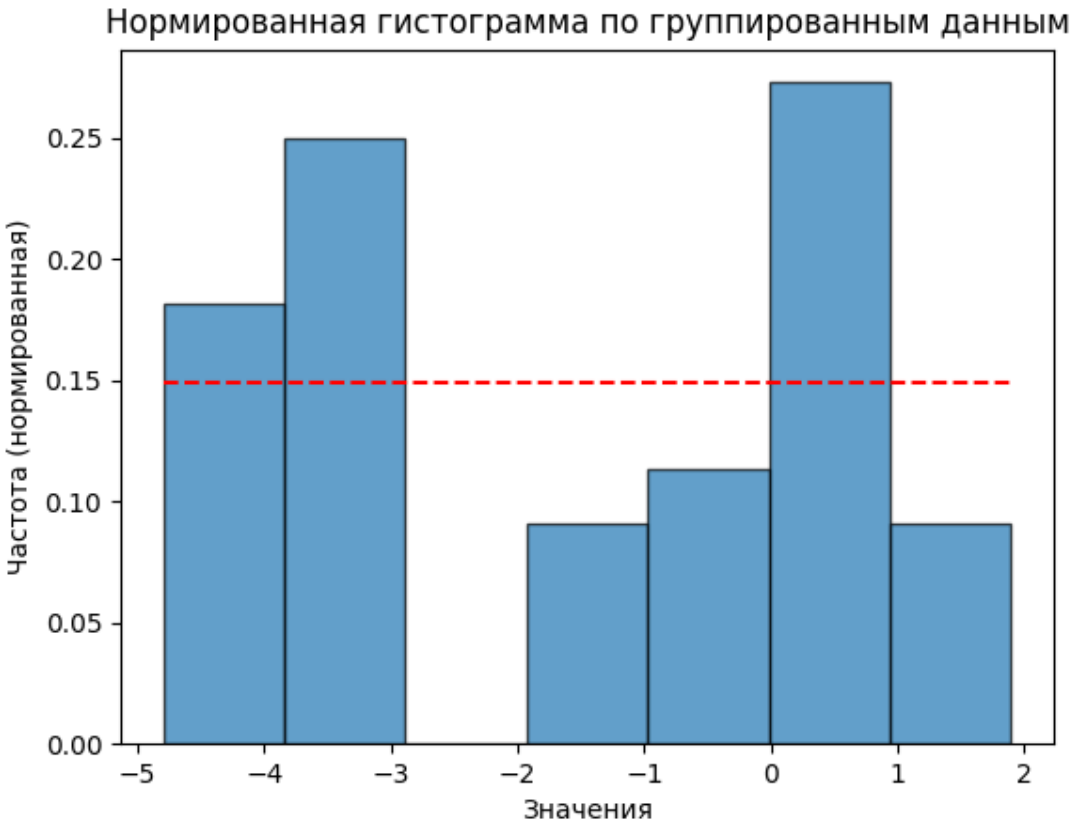


Синим - наша функция распределения.

Красным пунктиром - теоретическая функция равномерного распределения. Где a - совпадает с найденным ранее минимальным значением, а b - с максимальным.

3. Построим гистограмму

По полученным в пункте 1 данным построим гистограмму. Красным обозначено теоретическое поведение равномерно распределенной функции (при a и b равны ранее найденным минимуму и максимуму соответственно).



4. Вычислим оценки математического ожидания, дисперсии и среднеквадратического отклонения, используя формулы для группированной выборки

$$\widehat{m} = \sum_{i=1}^{k} x_i^{cp} * \widehat{p}_i \text{ и } \sigma^2 = \sum_{i=1}^{k} (x_i^{cp} - \widehat{m})^2 * \widehat{p}_i$$

Пользуясь таблицей, полученной в пункте 1 задания 4, подставим значения в функцию и посчитаем полученное значение. Полученные значения:

$\widehat{m} = -1.5588$

$$\widehat{D} = \sigma^2 = 4.2565$$

$$\sigma = 2.0631$$

5. По графику и гистограмме оценим, насколько наша выборка соответствует указанному в задании закону распределения

По гистограмме видно, что у выборки присутствуют 2 пика - во 2 интервале и в 6 интервале. При этом превышение в этих областях значительное относительно теоретического распределения. Более того, 3 бакет вообще не содержит значений, что не свойственно равномерному распределению

По графику ЭФР тоже видно, что полученные отрезки вероятностей плохо ложатся на теоретическую прямую.

Исходя из сказанного выше мы предполагаем, что наша выборка не соответствует указанному (равномерному) распределению.

6*. Код программы (для построений и расчетов)

Файл lab_1_task_4.py на гитхабе https://github.com/SnowLullaby/Matstat

5. Оценка параметров распределения по методу моментов

1. Используя формулу плотности вероятности соответствующего закона распределения, получим формулы, выражающие математическое ожидание и дисперсию СВ через параметры распределения

$$\widehat{m} = \int_{-\infty}^{+\infty} x \cdot f(x) = \int_{a}^{b} x \cdot \frac{1}{b-a} dx = \frac{x^2}{2} \cdot \frac{1}{b-a} |_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{(b-a)(b+a)}{2(b-a)} = \frac{b+a}{2}$$

$$\widehat{D} = \int_{-\infty}^{+\infty} x^2 \cdot f(x)dx - M^2(x) = \frac{x^3}{3} \cdot \frac{1}{b-a} |_a^b - \frac{(b+a)^2}{4} = \frac{b^3 - a^3}{3(b-a)} - \frac{(b+a)^2}{4} = \frac{4(b^2+ab+a^2)}{12} - \frac{3(b+a)^2}{12}$$

$$= \frac{4b^2 + 4ab + 4a^2 - 3b^2 - 6ab - 3a^2}{12} = \frac{b^2 - 2ab + a^2}{12} = \frac{(b-a)^2}{12}$$

2. Выразим параметры распределения через математическое ожидание и дисперсию

$$a = 2\widehat{m} - b$$

$$\widehat{D} = \frac{(b - (2\widehat{m} - b))^2}{12} = \frac{4(b - \widehat{m})^2}{12} = \frac{(b - \widehat{m})^2}{3}$$

$$b - \widehat{m} = \sqrt{3\widehat{D}}$$

$b = \widehat{m} + \sqrt{3\widehat{D}}$ и тогда $a = \widehat{m} - \sqrt{3\widehat{D}}$

3. Вычислим оценки параметров, подставив в формулы оценки математического ожидания и дисперсии

$$a = -1.5588 - \sqrt{3 \cdot 4.2565} = -5.1322$$

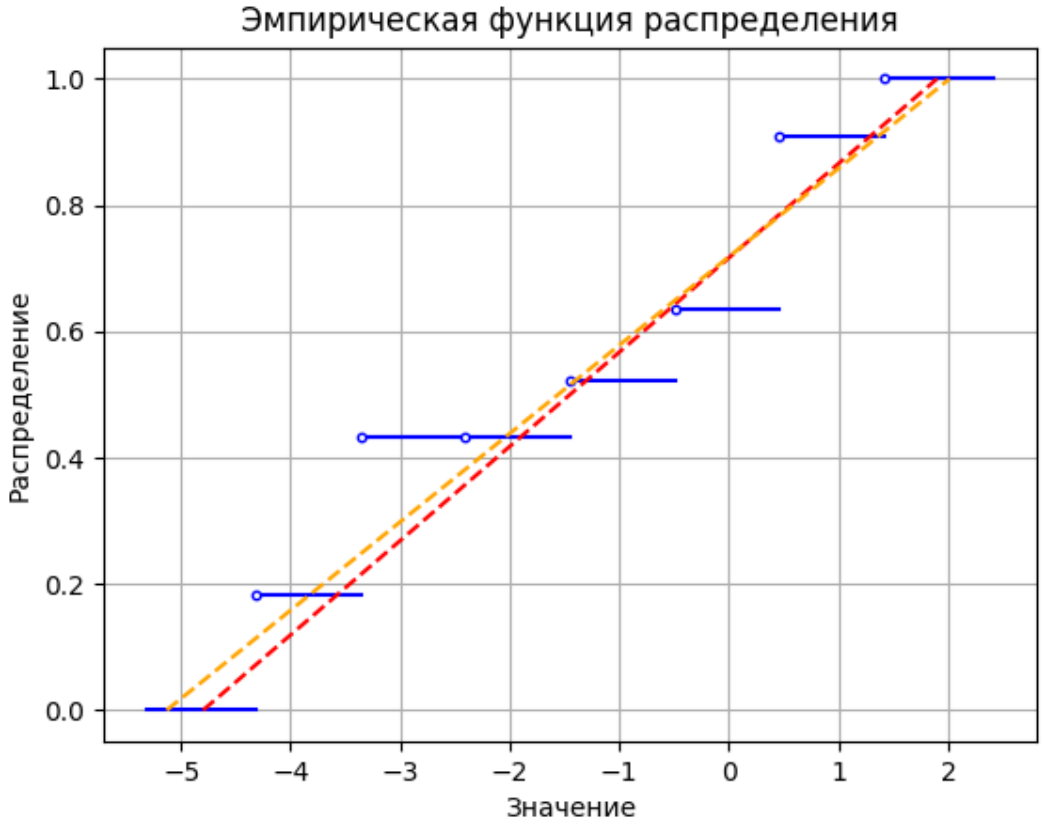
$$b = -1.5588 + \sqrt{3 \cdot 4.2565} = 2.0146$$

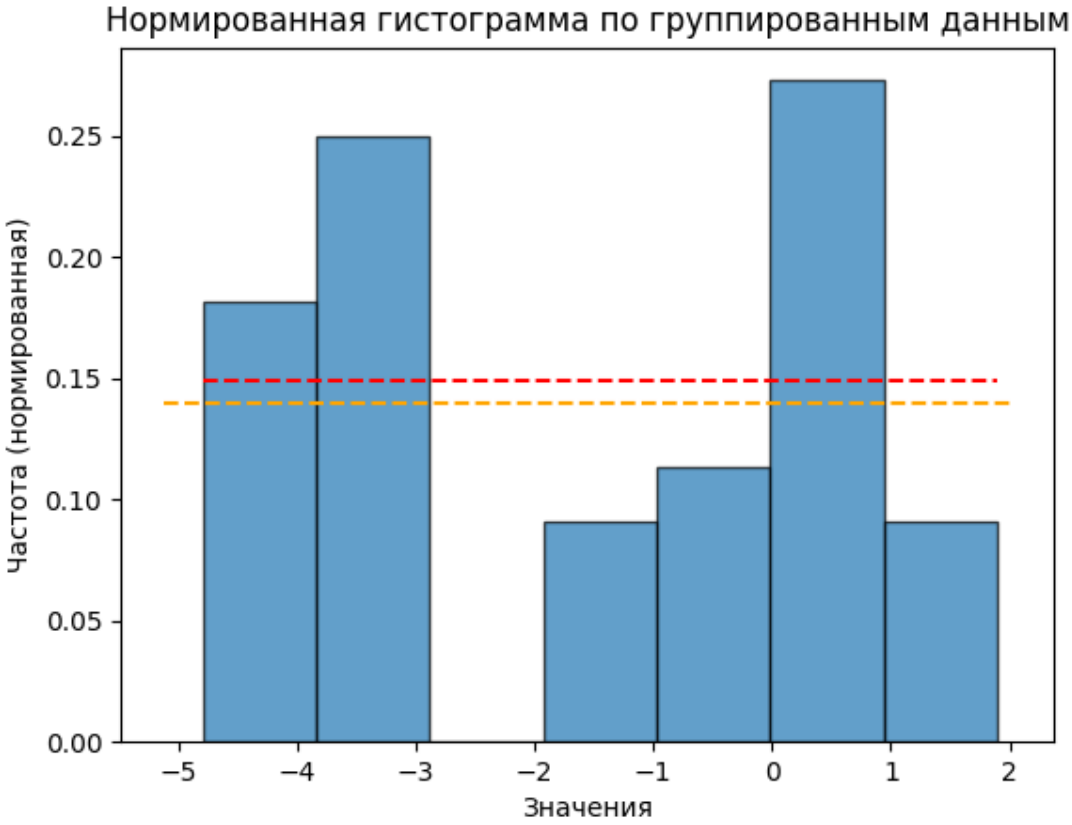
4. Построим графики плотности вероятности и функции распределения с полученными параметрами. В этих же осях координат изобразить ранее полученные оценку функции распределения и гистограмму

Для построения графиков (в том числе для пункта 4.2 и 4.3) воспользуемся следующим способом:

1. Построим множество x принадлежащих интервалу [a, b]
2. Для каждого x будем считать значение следующим образом (в зависимости от типа графика):
    a. Если мы строим ЭФР, то нам нужны кумулятивные вероятности. Они будут считаться как $\int_{-\infty}^{x} \frac{1}{b-a} dx = \int_{a}^{x} \frac{1}{b-a} dx = \frac{x-a}{b-a}$
    b. Если мы строим гистограмму - вероятности будут равны $\frac{1}{b-a}$

На графиках ниже оранжевым указаны графики плотности вероятности и ЭФР с рассчитанными ранее параметрами. Красный - график, основанный на данных выборки.





6. Оценка параметров распределения по методу максимального правдоподобия

1. Построим функцию максимального правдоподобия на основании плотности вероятности заданного закона распределения

$f(a,b) = \prod_{i=1}^{44} \frac{1}{b-a} = \frac{1}{(b-a)^{44}}$ при условии, что все 44 значения лежат внутри [a, b].

2. Построим логарифмическую функцию максимального правдоподобия

$$L(a,b) = ln(\frac{1}{(b-a)^{44}}) = -44 \cdot ln(b-a)$$

3. Найдем оценки параметров максимизируя функцию правдоподобия

$$\frac{\partial}{\partial a} = -44\frac{1}{b-a} = \frac{-44}{b-a}$$

$$\frac{\partial}{\partial b} = 44\frac{1}{b-a} = \frac{44}{b-a}$$

Эти функции никогда не равны 0, значит у нас нет точек локального минимума, локального максимума или седловых точек. Следовательно нам надо определить поведение функции и разбирать значения по границам допустимых a и b.

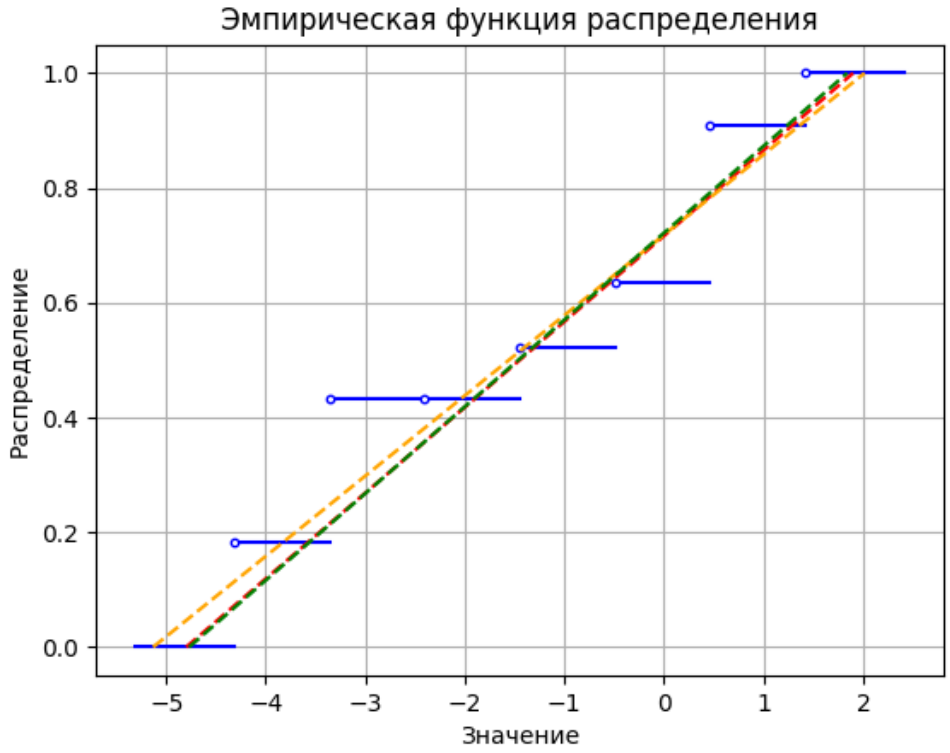
$-44 \cdot ln(b-a)$ возрастает, когда a максимально приближается к b (т.к. тогда логарифм стремится к минус бесконечности, а умножение на -44 делает это выражение стремящимся к бесконечности).

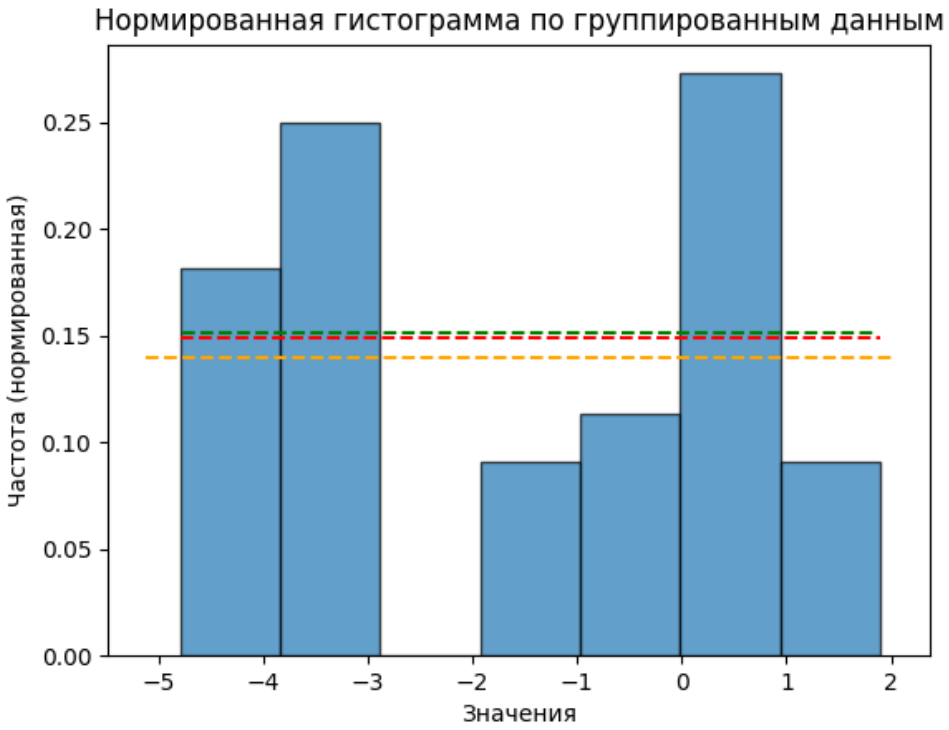
Тогда нам надо найти максимально приближенные друг к другу a и b при условии ограничений $a \leq -4.77$ и $b \geq 1.84$

Максимально близкими являются $a = -4.77$ и $b = 1.84$

4. Построим графики плотности вероятности и функции распределения с полученными параметрами. В этих же осях координат изобразить ранее полученные оценку функции распределения и гистограмму (и все построенные до этого графики)

Строить будем аналогично тому, как строили в пункте 5.4. На графике новую функцию обозначим зеленым





Вывод

В ходе работы мы исследовали распределение Бернулли и равномерное распределение. Оценка параметра распределения

Бернулли с помощью метода моментов и максимального правдоподобия дала схожие результаты ($p \approx 0.6$), что немного отклоняется от истинного значения 0.5. Это связано с малым размером выборки. В случае с равномерным распределением выборка не полностью соответствовала теоретическому распределению, что также объясняется ограниченным количеством данных. Мы пришли к выводу, что для более точных оценок необходимо увеличивать размер выборки.