**Национальный исследовательский университет**
**Высшая школа экономики**
**Московский институт электроники и математики**

Департамент прикладной математики
кафедра компьютерной безопасности

# Долгосрочное домашнее задание

## по матетматической статистике

**Дискретное распределение:** *логарифмическое распределение*
**Непрерывное распределение:** *треугольное распределение*

Выполнил
Ишкинин Д.С.

Проверил
Чухно А.Б.

Москва 2023

# Оглавление


**1 Характеристики вероятностных распределений 4**

1.1 Дискретное распределение . . . . . 4

1.1.1 Основные характеристики распределения . . . . . 4

1.1.2 Примеры событий, которые могут быть описаны выбранными случайными величинами . . . . . 5

1.1.3 Описание способа моделирования выбранных случайных величин . . . . . 6

1.2 Непрерывное распределение . . . . . 7

1.2.1 Основные характеристик распределения . . . . . 7

1.2.2 Примеры событий, которые могут быть описаны выбранными случайными величинами . . . . . 9

1.2.3 Описание способа моделирования выбранных случайных величин . . . . . 10

**2 Основные понятия математической статистики 12**

2.1 Дискретное распределение . . . . . 12

2.1.1 Генерация выборок выбранных случайных величин . . . 12

2.1.2 Построение эмпирической функции распределения . . . . 12

2.1.3 Построение гистограммы и полигона частот . . . . . 17

2.1.4 Вычисление выборочных моментов . . . . . 21

2.2 Непрерывное распределение . . . . . 23

2.2.1 Генерация выборок выбранных случайных величин . . . 23

2.2.2 Построение эмпирической функции распределения . . . . 23

2.2.3 Построение гистограммы и полигона частот . . . . . 27

2.2.4 Вычисление выборочных моментов . . . . . 32

**3 Построение точечных оценок параметра распределения 33**

3.1 Дискретное распределение . . . . . 33

3.1.1 Получение оценок методом моментов и методом максимального правдоподобия . . . . . 33

3.1.2 Поиск оптимальных оценок . . . . . 34

3.1.3 Работа с данными . . . . . 35

3.2 Непрерывное распределение . . . . . 37

3.2.1 Получение оценок методом моментов и методом максимального правдоподобия . . . . . 37

3.2.2 Поиск оптимальных оценок . . . . . 39

3.2.3 Работа с данными . . . . . 39

**4 Проверка статистических гипотез 43**
4.1 Дискретное распределение . . . . . 43
4.1.1 Проверка гипотезы о виде распределения . . . . . 43
4.1.2 Задание для данных, описываемых распределением . . . 53
4.2 Непрерывное распределение . . . . . 54
4.2.1 Проверка гипотезы о виде распределения . . . . . 55

**5 Различение статистических гипотез 69**
5.1 Теория . . . . . 69
5.2 Описание критерия отношения правдоподобия . . . . . 70
5.3 Вычисление функции отношения правдоподобия . . . . . 70
5.4 Вычисление критической области . . . . . 70
5.5 Вычисление минимального количества материала . . . . . 71

# Домашнее задание 1.

---

# Характеристики вероятностных распределений

### 1. Дискретное распределение

$\xi \sim Log(\theta)$, $P(\xi = x) = -\ln(1-\theta)^{-1} \cdot \theta^x \cdot x^{-1}$, $x \in \mathbb{N}$, $\theta = \frac{1}{13}$

#### 1.1.1. Основные характеристики распределения

▷ *Функция распределения*

$\forall x < 1$ $F_\xi(x) = P(\xi \leq x) = 0$, т. к. $\xi$ принимает только натуральные значения

$$\forall x \geq 1 \; F_\xi(x) = P(\xi \leq x) = \sum_{k=1}^{\lfloor x \rfloor} P(\xi = k) = -\sum_{k=1}^{\lfloor x \rfloor} \ln(1-\theta)^{-1} \cdot \theta^k \cdot k^{-1} =$$

$$= -\frac{1}{\ln(1-\theta)} \sum_{k=1}^{\lfloor x \rfloor} \frac{\theta^k}{k}$$

$$\text{Тогда } F_\xi(x) = \begin{cases} 0, & \text{если } x < 1 \\ \frac{1}{\ln \frac{13}{12}} \sum\limits_{k=1}^{\lfloor x \rfloor} \frac{1}{k \cdot 13^k}, & \text{если } x \geq 1 \end{cases}$$

▷ *Математическое ожидание*

Для дискретного распределения справедлива следующая формула:

$$M\xi = \sum_{i \geq 1} x_i P(\xi = x_i)$$

Тогда:

$$M\xi = \sum_{i \geq 1} iP(\xi = i) = -\sum_{i \geq 1} i \frac{\theta^i}{\ln(1-\theta) \cdot i} = -\frac{1}{\ln(1-\theta)} \sum_{i \geq 1} \theta^i = -\frac{1}{\ln(1-\theta)} \frac{\theta}{1-\theta} =$$

$$= -\frac{1}{\ln \frac{12}{13}} \frac{1}{12} = \frac{1}{12 \ln \frac{13}{12}}$$

▷ *Дисперсия*

Для нахождения дисперсии воспользуемся следующими формулами:

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

$$M\xi^2 = \sum_{i \geq 1} x_i^2 P(\xi = x_i)$$

$M\xi^2 = \sum\limits_{i\geq 1} i^2 P(\xi = i) = -\sum\limits_{i\geq 1} i^2 \frac{\theta^i}{\ln(1-\theta)\cdot i} = -\frac{1}{\ln(1-\theta)} \sum\limits_{i\geq 1} i \cdot \theta^i =$

$= -\frac{1}{\ln(1-\theta)} \sum\limits_{i\geq 0} (i+1) \cdot \theta^{i+1} = -\frac{\theta}{\ln(1-\theta)} \sum\limits_{i\geq 0} (i+1) \cdot \theta^i = -\frac{\theta}{\ln(1-\theta)} \frac{1}{(1-\theta)^2}$

Тогда: $D\xi = -\frac{\theta}{\ln(1-\theta)} \frac{1}{(1-\theta)^2} - \frac{1}{\ln^2(1-\theta)} \frac{\theta^2}{(1-\theta)^2} = -\theta \frac{\ln(1-\theta)+\theta}{(1-\theta)^2 \ln^2(1-\theta)} = -\frac{1}{13} \frac{\ln \frac{12}{13} + \frac{1}{13}}{\frac{12^2}{13^2} \ln^2 \frac{12}{13}} =$

$= -\frac{13 \ln \frac{12}{13} + 1}{144 \ln^2 \frac{12}{13}}$

▷ *Квантиль уровня* $\gamma$

Квантиль распределения уровня $\gamma$ случайной величины $\xi$ - значение $x_\gamma$ случайной величины $\xi$, для которого выполняется следующее равенство:

$$P(\xi \leq x_\gamma) = \gamma$$

Возможные значения $\gamma$:

1. $\gamma = 0 \Rightarrow x_\gamma < 1$. Но значение случайной величины, принимающей натуральные значения, не может быть меньше 1. Т. е. $\nexists x_\gamma : F_\xi(x_\gamma) = 0$
2. $\gamma \in (0,1) \Rightarrow F_\xi(x_\gamma) = \frac{1}{\ln \frac{13}{12}} \sum\limits_{k=1}^{\lfloor x_\gamma \rfloor} \frac{1}{k \cdot 13^k} = \gamma \Rightarrow \sum\limits_{k=1}^{x_\gamma} \frac{1}{k \cdot 13^k} = \ln \frac{13}{12} \gamma$
   Решив уравнение, найдем искомый $x_\gamma$.

### 1.1.2. Примеры событий, которые могут быть описаны выбранными случайными величинами

▷ **Пример интерпретации распределения:**

– Логарифмическое распределение довольно популярно в страховании для моделирования частоты возникновения претензий. Оно используется например, для описания: количества товаров, приобретенных потребителем за определенный период; количества видов птиц и растений в определенной местности; и количества паразитов на одного хозяина.

– Данное распределение также используется в микробиологии для моделирования количества организмов в кластере.

– Логарифмическое распределение используется для описания разнообразия выборки, то есть того, сколько элементов данного типа содержится в выборке элементов. Например, оно позволяет описать количество особей определенного вида в выборке комаров.

– Физик Фрэнк Бенфорд собрал данные, чтобы посмотреть, будут ли естественные и социологические наборы данных подчиняться логарифмическому закону. Он часто находил соответствие, работая с такими датасетами, как площади рек, статистика американской лиги бейсбола, атомные веса элементов, показатели смертности.

▷ **Известные соотношения между распределениями:**

Пусть выборка $X = (X_1, ..., X_n)$ из распределения $\mathcal{L}(\xi)$, где $\xi \sim Log(\theta)$. Пусть случайная величина $p \sim Pois(\lambda)$. Тогда $T = \sum\limits_{i=1}^{p} X_i$ имеет отрицательное биномиальное распределение: $T \sim \overline{Bi}(1-\theta, -\frac{\lambda}{\ln(1-\theta)})$.

Для доказательства данного факта нам потребуются следующие производящие функции:

1. $g_p(s) = Ms^p = \sum\limits_{r=0}^{\infty} P(p=r)s^r = \sum\limits_{r=0}^{\infty} \frac{\lambda^r}{r!}e^{-\lambda}s^r = e^{-\lambda}\sum\limits_{r=0}^{\infty}\frac{(\lambda s)^r}{r!} =$
$= e^{-\lambda}e^{\lambda s} = e^{\lambda(s-1)}$
2. $g_\xi(s) = Ms^\xi = \sum\limits_{r=1}^{\infty} P(\xi=r)s^r = \sum\limits_{r=1}^{\infty}\frac{(-1)\theta^r}{\ln(1-\theta)\cdot r}s^r = \frac{1}{\ln(1-\theta)}\sum\limits_{r=1}^{\infty}\frac{(-1)(\theta s)^r}{r} =$
$= \frac{1}{\ln(1-\theta)}\sum\limits_{r=1}^{\infty}\frac{(-1)^{r-1}(-\theta s)^r}{r} = \frac{\ln(1-\theta s)}{\ln(1-\theta)}$

Пользуясь соотношением: $g_T(s) = exp[\lambda(g_\xi(s)-1)]$, получим:

$g_T(s) = exp[\lambda(\frac{\ln(1-\theta s)}{\ln(1-\theta)}-1)] = exp[\frac{\lambda}{\ln(1-\theta)}\ln\frac{1-\theta s}{1-\theta}] = (\frac{1-\theta s}{1-\theta})^{\frac{\lambda}{\ln(1-\theta)}} = (\frac{1-\theta}{1-\theta s})^{-\frac{\lambda}{\ln(1-\theta)}}$

Учитывая, что для $\overline{Bi}(n,p)$ производящая функция $g(s) = (\frac{p}{1-(1-p)s})^n$, получаем, что $g_T(s)$ есть производящая для $\overline{Bi}(-\frac{\lambda}{\ln(1-\theta)}, 1-\theta)$.

Таким образом, мы доказали, что пуассоновская сумма случайных величин, распределенных логарифмически с параметром $\theta$, имеет отрицательное биномиальное распределение с параметрами $1-\theta$ и $-\frac{\lambda}{\ln(1-\theta)}$.

### 1.1.3. Описание способа моделирования выбранных случайных величин

Пусть имеется источник непрерывных случайных величин, распределенных равномерно на отрезке $[0,1]$. Получим по выборке $\mathcal{U} = (\mathcal{U}_1, ..., \mathcal{U}_n)$ из распределения $\mathcal{L}(\mathcal{R}[0,1])$ выборку $X = (X_1, ..., X_n)$ из распределения $\mathcal{L}(\xi)$, где $\xi \sim Log(\theta)$.

Обозначим $S_m = \sum\limits_{i=1}^{m} p_i$, где $p_i = P(\xi = i)$. Тогда учитывая, что $\mathcal{U}_i$ равномерно распределена на отрезке $[0,1]$, получим:

$$P(S_{m-1} \le \mathcal{U}_i < S_m) = P(S_{m-1} \le \mathcal{U}_i < S_{m-1} + p_m) = p_m \ \forall m \ge 2$$

Тогда номер $m_0$, соответствующий значению случайной величины $X_i$, может быть определен из неравенства:

$$S_{m_0-1} \le \mathcal{U}_i < S_{m_0}$$

Случайная величина $X_i$ определяется только через $\mathcal{U}_i$. Тогда случайные величины $X_1, ..., X_n$ независимы и одинаково распределены.

Алгоритмически $m_0$ можно вычислить следующим образом: для каждого значения $\mathcal{U}_i$ $(i \in \overline{1,n})$ будем последовательно вычитать $p_j$, где $j \in \mathbb{N}$, пока разность будет неотрицательной. Номер $m_0$, после которого разность стала отрицательной, и будет значением случайной величины $X_i$.

## 2. Непрерывное распределение

Плотность распределения: $f_\xi(x) = \begin{cases} \frac{2x}{\theta}, & \text{если } x \in [0,\theta] \\ \frac{2(1-x)}{1-\theta}, & \text{если } x \in (\theta,1] \\ 0, & \text{иначе} \end{cases}; \ \theta = 0.6$

### 1.2.1. Основные характеристик распределения

▷ *Функция распределения* $F_\xi(x) = \int\limits_{-\infty}^{x} f_\xi(t)dt$

$\forall x < 0 \ F_\xi(x) = \int\limits_{-\infty}^{x} 0dt = 0$

$\forall x \in [0,\theta] \ F_\xi(x) = \int\limits_{-\infty}^{0} 0dt + \int\limits_{0}^{x} \frac{2t}{\theta}dt = \frac{1}{\theta}t^2|_0^x = \frac{x^2}{\theta}$

$\forall x \in (\theta,1] \ F_\xi(x) = \int\limits_{-\infty}^{0} 0dt + \int\limits_{0}^{\theta} \frac{2t}{\theta}dt + \int\limits_{\theta}^{x} \frac{2(1-t)}{1-\theta}dt = \frac{1}{\theta}t^2|_0^\theta + \frac{2}{1-\theta}\int\limits_{\theta}^{x}(1-t)dt =$
$= \theta + \frac{1}{1-\theta}(2t - t^2)|_\theta^x = \frac{\theta(1-\theta)+2x-x^2-2\theta+\theta^2}{1-\theta} = \frac{-\theta-(x^2-2x)}{1-\theta} = \frac{-\theta-(x-1)^2+1}{1-\theta} =$
$= 1 - \frac{(1-x)^2}{1-\theta}$

$\forall x > 1 \ F_\xi(x) = \int\limits_{-\infty}^{0} 0dt + \int\limits_{0}^{\theta} \frac{2t}{\theta}dt + \int\limits_{\theta}^{1} \frac{2(1-t)}{1-\theta}dt + \int\limits_{1}^{x} 0dt = \frac{1}{\theta}t^2|_0^\theta + \frac{1}{1-\theta}(2t-t^2)|_\theta^1 =$
$= \theta + \frac{1}{1-\theta}(2 - 1 - 2\theta + \theta^2) = \frac{\theta-\theta^2+1-2\theta+\theta^2}{1-\theta} = \frac{1-\theta}{1-\theta} = 1$

Тогда $F_\xi(x) = \begin{cases} 0, & \text{если } x < 0 \\ \frac{x^2}{\theta}, & \text{если } x \in [0,\theta] \\ 1 - \frac{(1-x)^2}{1-\theta}, & \text{если } x \in (\theta,1] \\ 1, & \text{если } x > 1 \end{cases} = \begin{cases} 0, & \text{если } x < 0 \\ \frac{5}{3}x^2, & \text{если } x \in [0,0.6] \\ 1 - \frac{5}{2}(1-x)^2, & \text{если } x \in (0.6,1] \\ 1, & \text{если } x > 1 \end{cases}$

▷ *Математическое ожидание*

$$M\xi = \int\limits_{\mathbb{R}} x f_\xi(x)dx = \int\limits_{-\infty}^{0} x\cdot 0\cdot dx + \int\limits_{0}^{\theta} x\frac{2x}{\theta}dx + \int\limits_{\theta}^{1} x\frac{2(1-x)}{1-\theta}dx + \int\limits_{1}^{+\infty} x\cdot 0\cdot dx =$$
$$= \frac{2}{\theta}\frac{x^3}{3}|_0^\theta + \frac{2}{1-\theta}(\frac{x^2}{2} - \frac{x^3}{3})|_\theta^1 = \frac{2}{3}\theta^2 + \frac{1}{1-\theta}(1 - \frac{2}{3} - \theta^2 + \frac{2}{3}\theta^3) = \frac{2\theta^2(1-\theta)+1-3\theta^2+2\theta^3}{3(1-\theta)} =$$
$$= \frac{-\theta^2+1}{3(1-\theta)} = \frac{\theta+1}{3} = \frac{8}{15}$$

▷ *Дисперсия*

Для нахождения дисперсии воспользуемся следующими формулами:

$$D\xi = M\xi^2 - (M\xi)^2$$
$$Mg(\xi) = \int\limits_{\mathbb{R}} g(x) f_\xi(x)dx$$

$$M\xi^2 = \int\limits_{\mathbb{R}} x^2 f_\xi(x)dx = \int\limits_{-\infty}^{0} x^2\cdot 0\cdot dx + \int\limits_{0}^{\theta} x^2\frac{2x}{\theta}dx + \int\limits_{\theta}^{1} x^2\frac{2(1-x)}{1-\theta}dx + \int\limits_{1}^{+\infty} x^2\cdot 0\cdot dx =$$
$$= \frac{2}{\theta}\frac{x^4}{4}|_0^\theta + \frac{2}{1-\theta}(\frac{x^3}{3} - \frac{x^4}{4})|_\theta^1 = \frac{\theta^3}{2} + \frac{2}{1-\theta}(\frac{1}{3} - \frac{1}{4} - \frac{\theta^3}{3} + \frac{\theta^4}{4}) = \frac{3\theta^3(1-\theta)+1-4\theta^3+3\theta^4}{6(1-\theta)} =$$
$$= \frac{-\theta^3+1}{6(1-\theta)} = \frac{1+\theta+\theta^2}{6}$$

Тогда: $D\xi = \frac{1+\theta+\theta^2}{6} - \frac{(\theta+1)^2}{9} = \frac{3+3\theta+3\theta^2-2\theta^2-4\theta-2}{18} = \frac{\theta^2-\theta+1}{18} = \frac{9/25-3/5+1}{18} =$
$= \frac{9-15+25}{450} = \frac{19}{450}$

▷ *Квантиль уровня $\gamma$*

Квантиль распределения уровня $\gamma$ случайной величины $\xi$ - значение $x_\gamma$ случайной величины $\xi$, для которого выполняется следующее равенство:

$$P(\xi \le x_\gamma) = \gamma$$

Возможные значения $\gamma$:

1. $\gamma = 0 \Rightarrow x_\gamma \le 0$
2. $\gamma = 1 \Rightarrow x_\gamma \ge 1$
3. $\gamma \in (0,1) \Rightarrow F_\xi(x_\gamma) = \begin{cases} \frac{5}{3}x_\gamma^2, & x_\gamma \in (0, 0.6] \\ 1 - \frac{5}{2}(1-x_\gamma)^2, & x_\gamma \in (0.6, 1) \end{cases} = \gamma,$

$$\begin{cases} x_\gamma^2 = \frac{3}{5}\gamma, & \gamma \in (0, 0.6] \\ \frac{5}{2}(1-x_\gamma)^2 = 1-\gamma, & \gamma \in (0.6, 1) \end{cases}$$

При извлечении корней помним о том, где лежит $x_\gamma$:

$$\begin{cases} x_\gamma = \sqrt{\frac{3}{5}\gamma}, & \gamma \in (0, 0.6] \\ 1 - x_\gamma = \sqrt{\frac{2}{5}(1-\gamma)}, & \gamma \in (0.6, 1) \end{cases}, \begin{cases} x_\gamma = \sqrt{\frac{3}{5}\gamma}, & \gamma \in (0, 0.6] \\ x_\gamma = 1 - \sqrt{\frac{2}{5}(1-\gamma)}, & \gamma \in (0.6, 1) \end{cases}$$

### 1.2.2. Примеры событий, которые могут быть описаны выбранными случайными величинами

▷ **Пример интерпретации распределения:**

– Треугольное распределение обычно используется для описания совокупности с ограниченными выборочными данными, и особенно, когда взаимосвязь между переменными известна, но данных мало (возможно, из-за высокой стоимости сбора). Поэтому данное распределение было названо распределением "недостатка знаний".

– Треугольное распределение часто используется при принятии бизнес-решений, особенно при моделировании. Как правило, когда о распределении результата известно не так много (например, только его наименьшее и наибольшее значения), и если известен наиболее вероятный результат, то он может быть смоделирован с помощью треугольного распределения.

– Данное распределение также широко используется в управлении проектами для моделирования событий, которые происходят в интервале, определяемом минимальным и максимальным значением.

– Треугольное распределение применяется при вероятностном описании процессов аварийности и травматизма в человеко-машинных системах, вероятностном моделировании ошибок в линейной модели наблюдений, прогнозировании, экономической оценки риска.

▷ **Известные соотношения между распределениями:**

– Пусть случайная величина $\xi$ имеет треугольное распределение с параметром $\theta$. Тогда случайная величина $\eta = 1 - \xi$ распределена по треугольному закону с параметром $1 - \theta$:

*Доказательство.* $F_\eta(x) = P(1 - \xi \le x) = P(\xi \ge 1 - x) =$

$$= 1 - F_\xi(1-x) = 1 - \begin{cases} 0, & \text{если } 1-x < 0 \\ \frac{(1-x)^2}{\theta}, & \text{если } 1-x \in [0,\theta] \\ 1 - \frac{x^2}{1-\theta}, & \text{если } 1-x \in (\theta, 1] \\ 1, & \text{если } 1-x > 1 \end{cases} =$$

$$= \begin{cases} 1, & \text{если } x > 1 \\ 1 - \frac{(1-x)^2}{\theta}, & \text{если } x \in [1-\theta, 1] \\ \frac{x^2}{1-\theta}, & \text{если } x \in [0, 1-\theta) \\ 0, & \text{если } x < 0 \end{cases} = F_\xi(1-\theta) \qquad \square$$

– Пусть случайная величина $\mathcal{U} \sim \mathcal{R}[0,1]$. Пусть $F(x)$ - функция распределения для треугольной случайной величины.

$$F^{-1}(x) = \begin{cases} \sqrt{\theta x}, & \text{если } x \in [0, \theta] \\ 1 - \sqrt{(1-\theta)(1-x)}, & \text{если } x \in (\theta, 1] \end{cases}$$

Тогда случайная величина $F^{-1}(\mathcal{U})$ имеет треугольное распрделение с параметром $\theta$. Подробный вывод $F^{-1}(x)$ и доказательство данного факта приведено в следующем разделе.

### 1.2.3. Описание способа моделирования выбранных случайных величин

Пусть имеется источник непрерывных случайных величин, распределенных равномерно на отрезке $[0, 1]$. Получим по выборке $\mathcal{U} = (\mathcal{U}_1, ..., \mathcal{U}_n)$ из распределения $\mathcal{L}(\mathcal{R}[0, 1])$ выборку $X = (X_1, ..., X_n)$ из треугольного распределения $\mathcal{L}(\xi)$.

Заметим, что функция треугольного распределения $F(x)$ непрерывная, строго возрастает на отрезке $[0, 1]$ и выражается в элементарных функциях. Тогда для случайной величины $\eta = F^{-1}(\mathcal{U})$ верно:

$$P(\eta \le x) = P(F^{-1}(\mathcal{U}) \le x) = P(\mathcal{U} \le F(x)) = F(x)$$

Т. е. случайная величина $\eta = F^{-1}(\mathcal{U})$ имеет функцию распределения $F(x)$. Вычислив $F^{-1}(x)$, получим выборку $(F^{-1}(\mathcal{U}_1), ..., F^{-1}(\mathcal{U}_n))$ из треугольного распределения.

$$F(x) = \begin{cases} 0, & \text{если } x < 1 \\ \frac{x^2}{\theta}, & \text{если } x \in [0, \theta] \ (*) \\ 1 - \frac{(1-x)^2}{1-\theta}, & \text{если } x \in (\theta, 1] \ (**) \\ 1, & \text{если } x > 1 \end{cases}$$

▷ Найдем функцию, обратную к $F(x)$. Для этого заметим, что для обратной функции область определения - область значений прямой функции (отрезок $[0, 1]$).

$F(\theta) = \frac{\theta^2}{\theta} = \theta \Rightarrow$ для $F^{-1}(x)$ пересечение графиков обратных функций будет в точке с абсциссой $\theta$.

▷ Найдем функцию, обратную к $(*)$: $y = \frac{x^2}{\theta} \Rightarrow x = \pm\sqrt{\theta y}$

Делая переобозначение, получим: $y = \pm\sqrt{\theta x}$.

Заметим, что для $(*)$ рассматривался отрезок $[0, \theta]$, т. е. правая ветвь параболы. Тогда, обратная функция для $(*)$: $y = \sqrt{\theta x}$.

▷ Найдем функцию, обратную к $(**)$:

$y = 1 - \frac{(1-x)^2}{1-\theta} \Rightarrow (x-1)^2 = (1-y)(1-\theta) \Rightarrow x - 1 = \pm\sqrt{(1-\theta)(1-y)} \Rightarrow x = 1 \pm \sqrt{(1-\theta)(1-y)}$

Делая переобозначение, получим: $y = 1 \pm \sqrt{(1-\theta)(1-x)}$.

Заметим, что для $(**)$ рассматривался полуинтервал $(\theta, 1]$, т. е. левая ветвь параболы. Тогда, обратная функция для $(**)$: $y = 1 - \sqrt{(1-\theta)(1-x)}$.

Тогда: $F^{-1}(x) = \begin{cases} \sqrt{\theta x}, & \text{если } x \in [0, \theta] \\ 1 - \sqrt{(1-\theta)(1-x)}, & \text{если } x \in (\theta, 1] \end{cases} =$

$$= \begin{cases} \sqrt{0.6x}, & \text{если } x \in [0, 0.6] \\ 1 - \sqrt{0.4(1-x)}, & \text{если } x \in (0.6, 1] \end{cases}$$

Таким образом, по реализациям выборки равномерно распределенных случайных величин сможем смоделировать через обратную функцию реализации выборки случайных величин с треугольным распределением.

# Домашнее задание 2.

## Основные понятия математической статистики

### 1. Дискретное распределение

Реализацию необходимых алгоритмов моделирования случайных величин, а также построение графиков можно посмотреть здесь.

### 2.1.1. Генерация выборок выбранных случайных величин

Для генерации выборок логарифмической случайной величины используется описанный мною ранее способ моделирования.

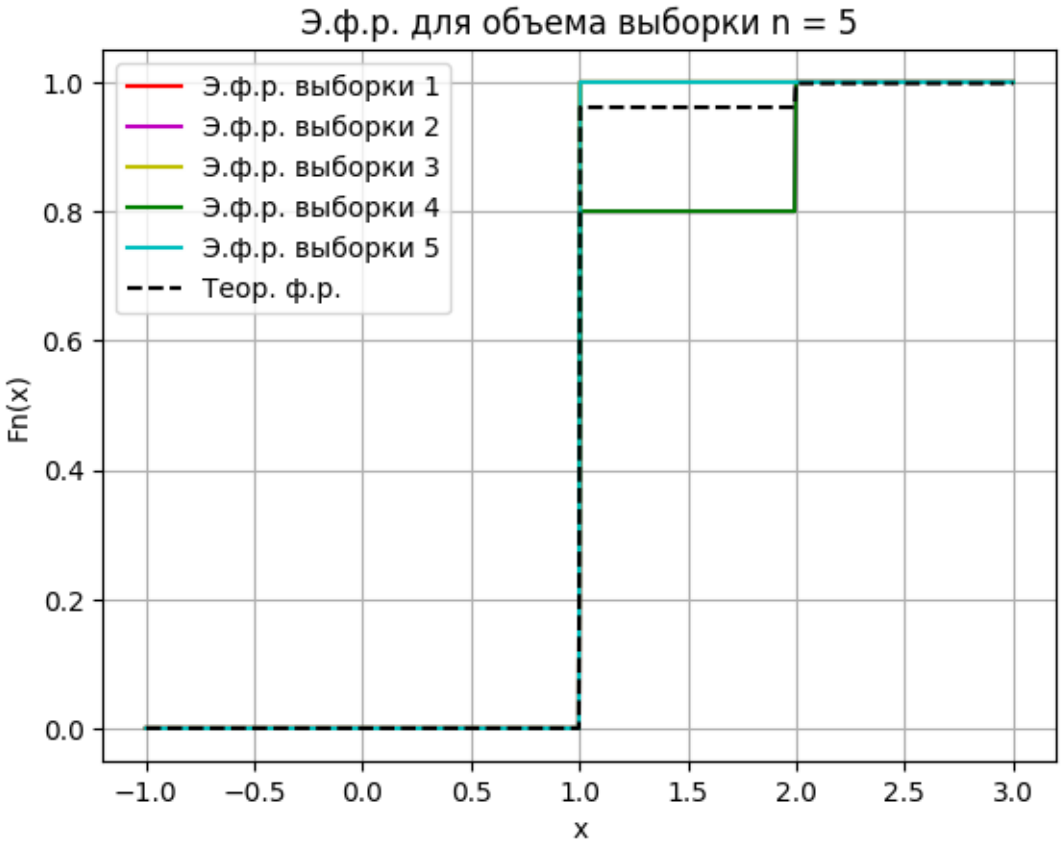
Объёмы сгенерированных выборок: $n = [5, 10, 100, 200, 400, 600, 800, 1000]$.

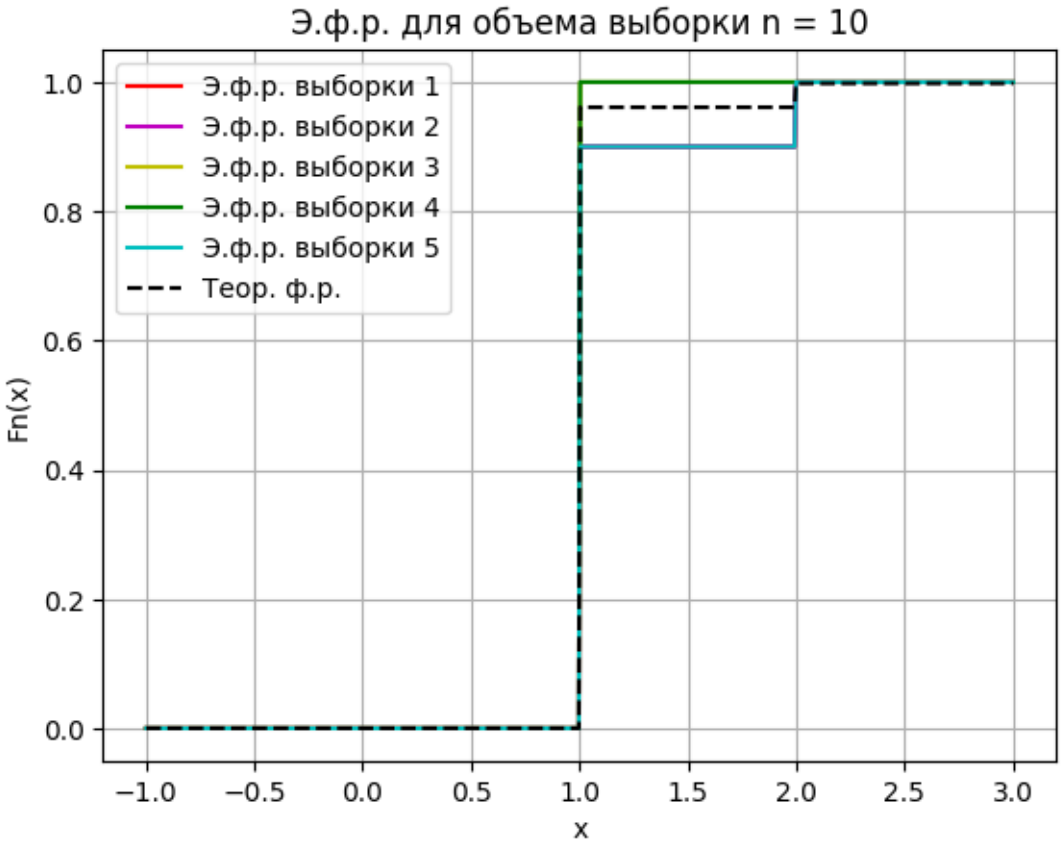
### 2.1.2. Построение эмпирической функции распределения

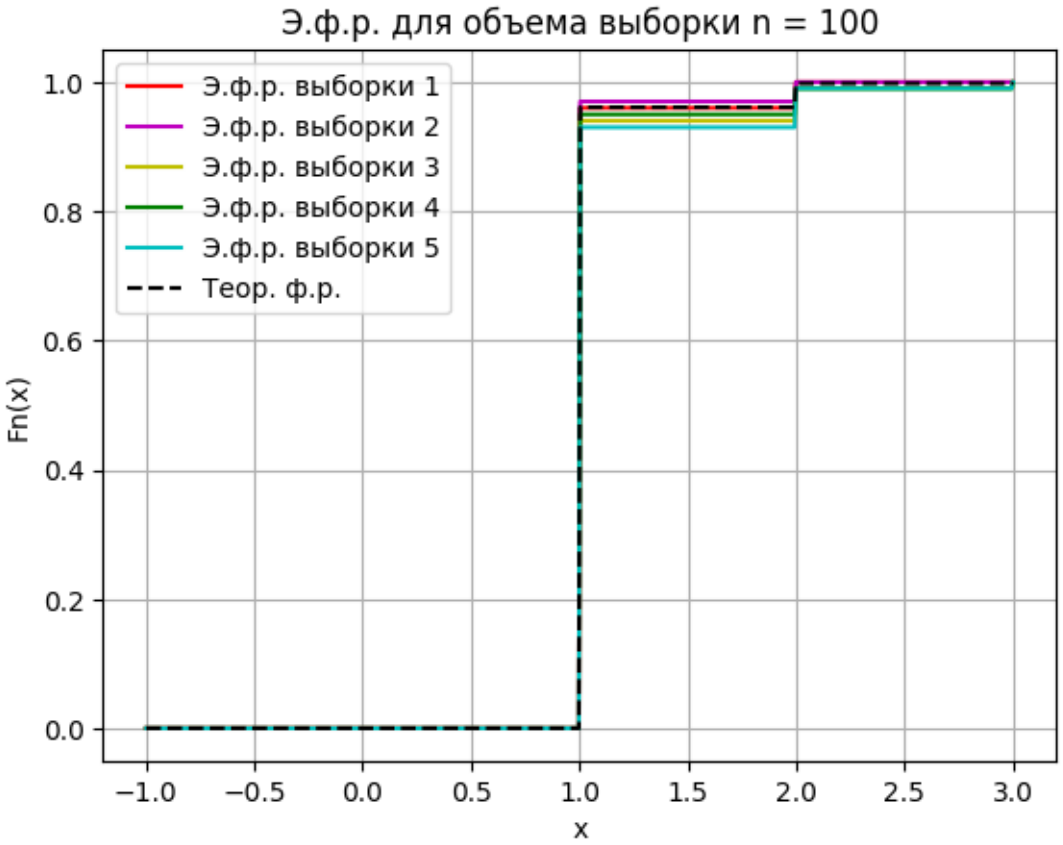
Эмпирическая функция распределения: $\mathscr{F}_n(t) = \frac{1}{n}\sum\limits_{i=1}^{n} I(x_i \leq t)$

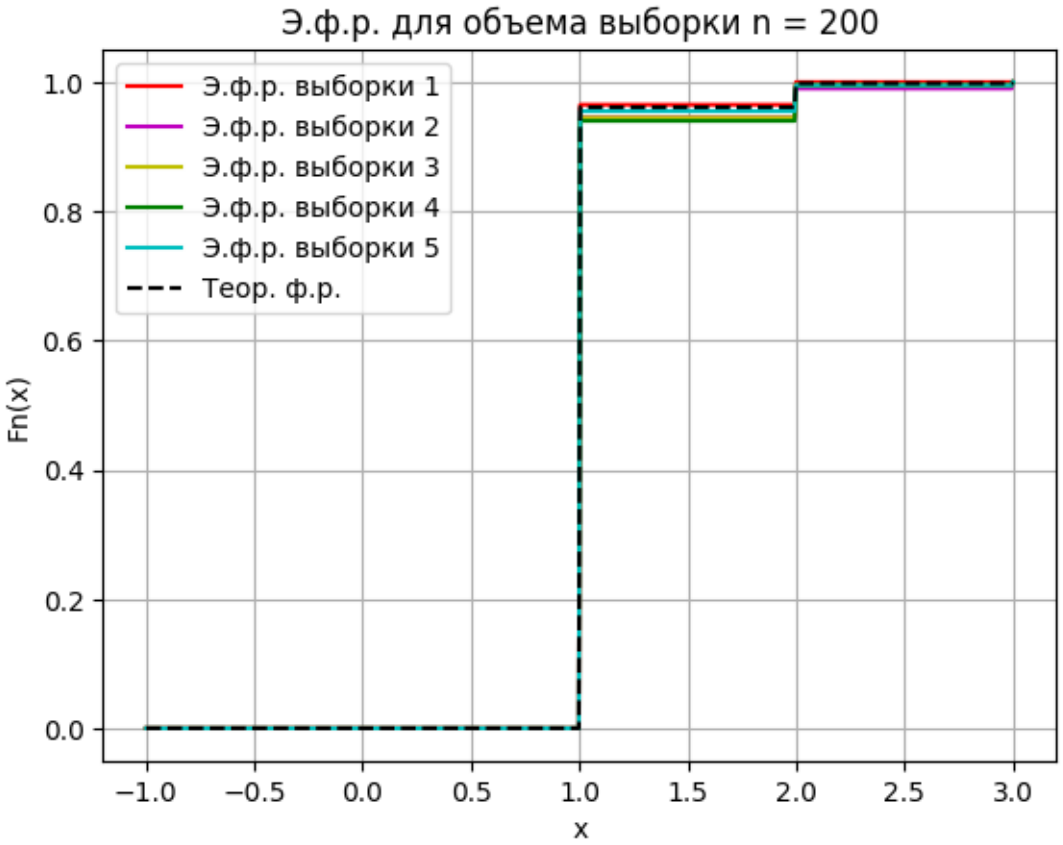
Замечу, что параметр $\theta = \frac{1}{13} \Rightarrow P(1) > 0.96 \Rightarrow$ согласно предъявленному мною алгоритму моделирования логарифмических случайных величин подавляющее большинство равномерных случайных величин получат значение 1 в логарифмическом распределении $\Rightarrow$ логично, что график имеет “немного”скачков

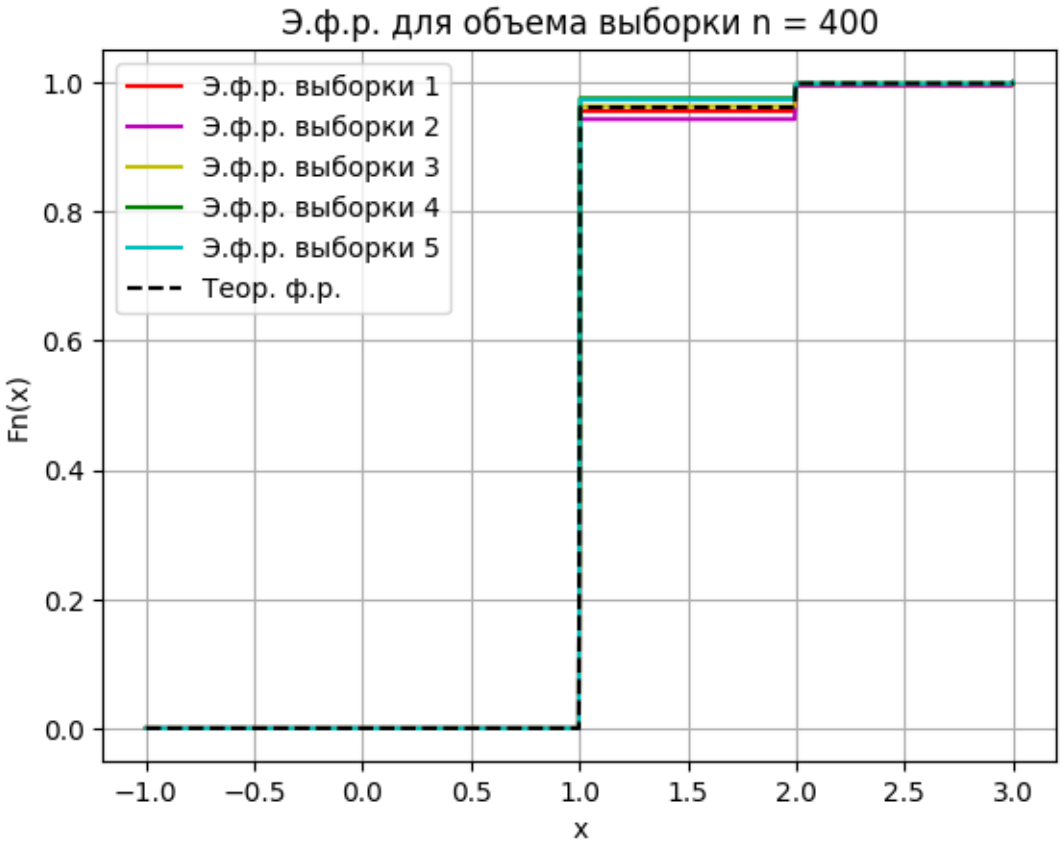
Графики эмпирических функций распределения для каждого из объемов выборки (на каждом графике изображена э. ф. р. для 5 выборок данного объема, а также функция распределения случайной величины):

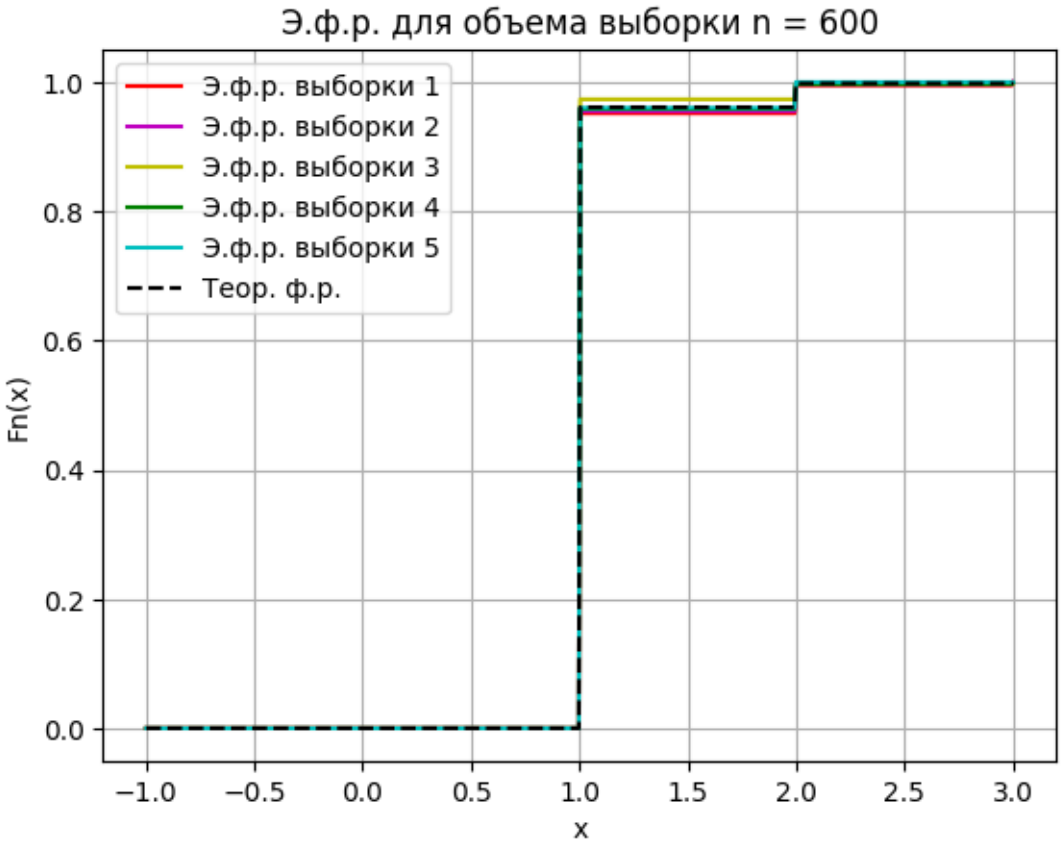

Э.ф.р. для объема выборки n = 5
Э.ф.р. выборки 1
Э.ф.р. выборки 2
Э.ф.р. выборки 3
Э.ф.р. выборки 4
Э.ф.р. выборки 5
Теор. ф.р.
Fn(x)
x
1.0
0.8
0.6
0.4
0.2
0.0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0

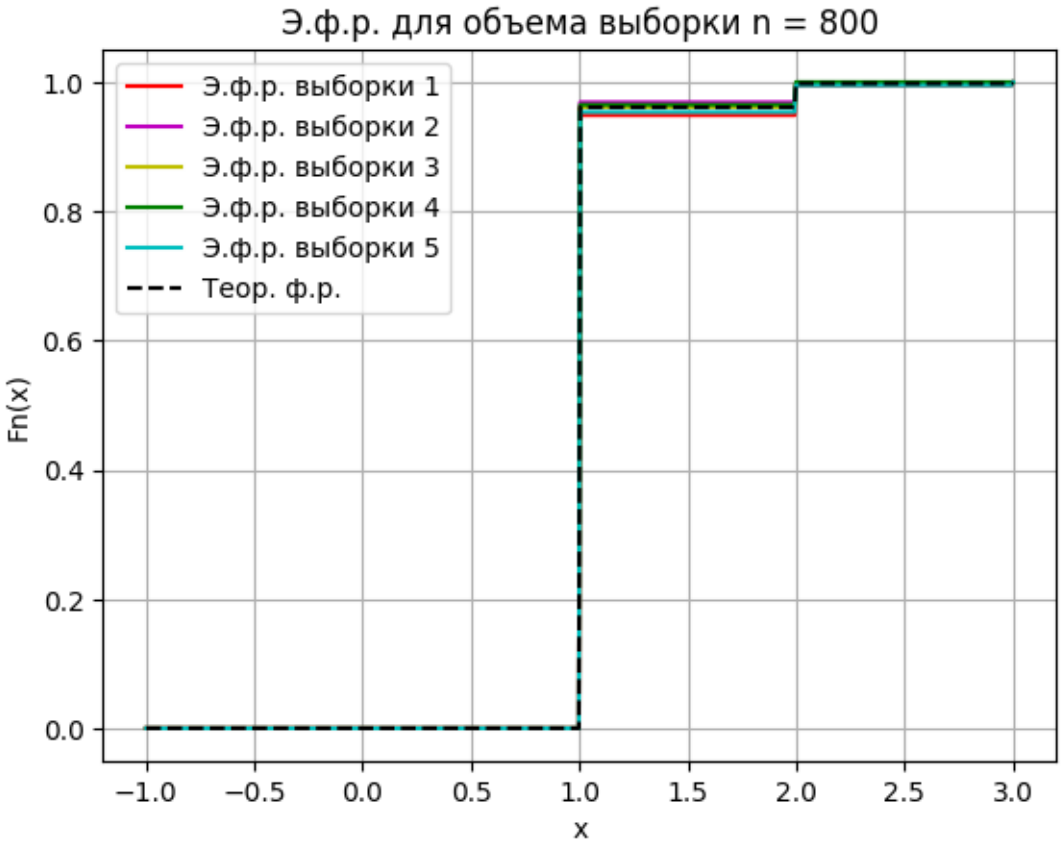

Э.ф.р. для объема выборки n = 10
Э.ф.р. выборки 1
Э.ф.р. выборки 2
Э.ф.р. выборки 3
Э.ф.р. выборки 4
Э.ф.р. выборки 5
Теор. ф.р.
Fn(x)
x
1.0
0.8
0.6
0.4
0.2
0.0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0

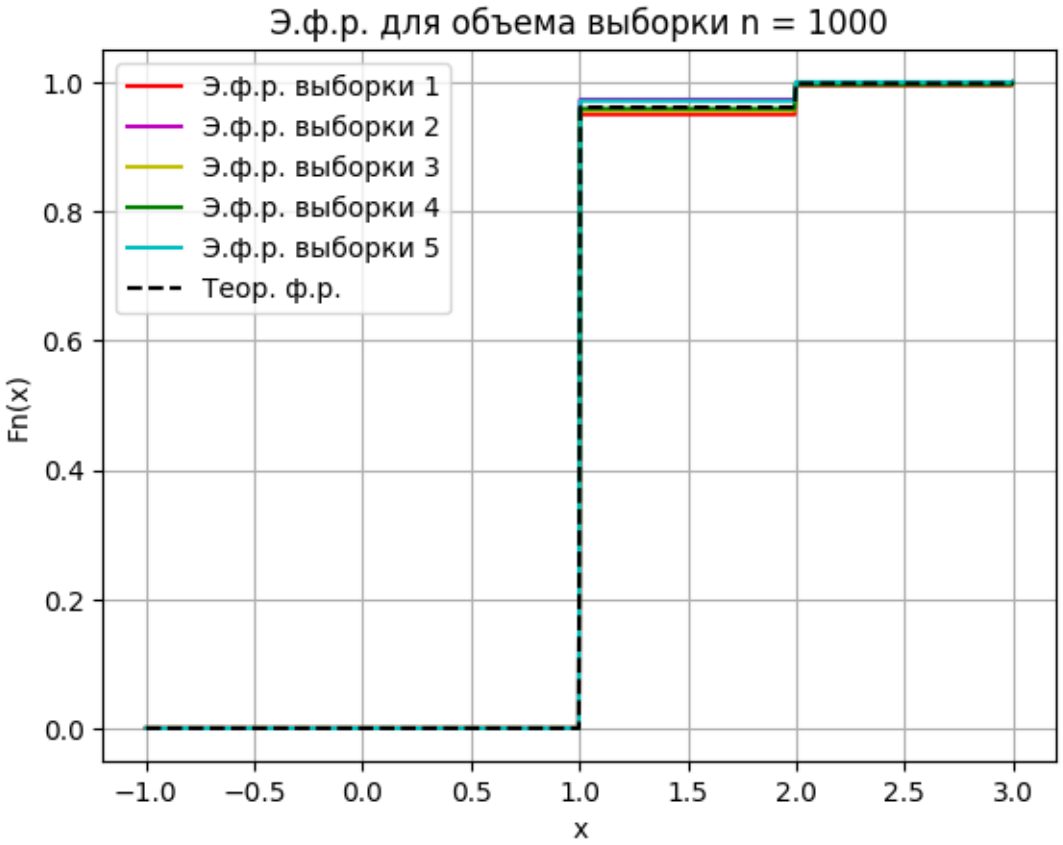

Э.ф.р. для объема выборки n = 100
Э.ф.р. выборки 1
Э.ф.р. выборки 2
Э.ф.р. выборки 3
Э.ф.р. выборки 4
Э.ф.р. выборки 5
Теор. ф.р.
Fn(x)
x
1.0
0.8
0.6
0.4
0.2
0.0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0


Э.ф.р. для объема выборки n = 200
Э.ф.р. выборки 1
Э.ф.р. выборки 2
Э.ф.р. выборки 3
Э.ф.р. выборки 4
Э.ф.р. выборки 5
Теор. ф.р.
Fn(x)
x
1.0
0.8
0.6
0.4
0.2
0.0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0


Э.ф.р. для объема выборки n = 400
Э.ф.р. выборки 1
Э.ф.р. выборки 2
Э.ф.р. выборки 3
Э.ф.р. выборки 4
Э.ф.р. выборки 5
Теор. ф.р.
Fn(x)
x
1.0
0.8
0.6
0.4
0.2
0.0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0


Э.ф.р. для объема выборки n = 600
Э.ф.р. выборки 1
Э.ф.р. выборки 2
Э.ф.р. выборки 3
Э.ф.р. выборки 4
Э.ф.р. выборки 5
Теор. ф.р.
Fn(x)
x
1.0
0.8
0.6
0.4
0.2
0.0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0



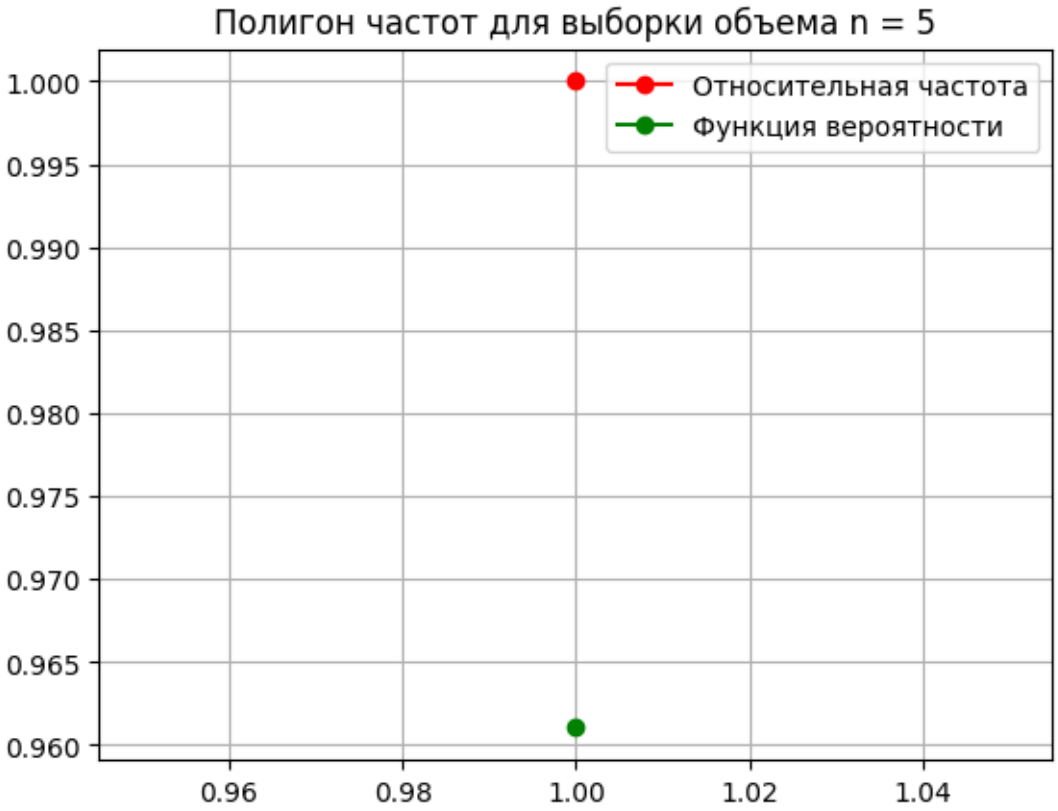

Вычисленные значения

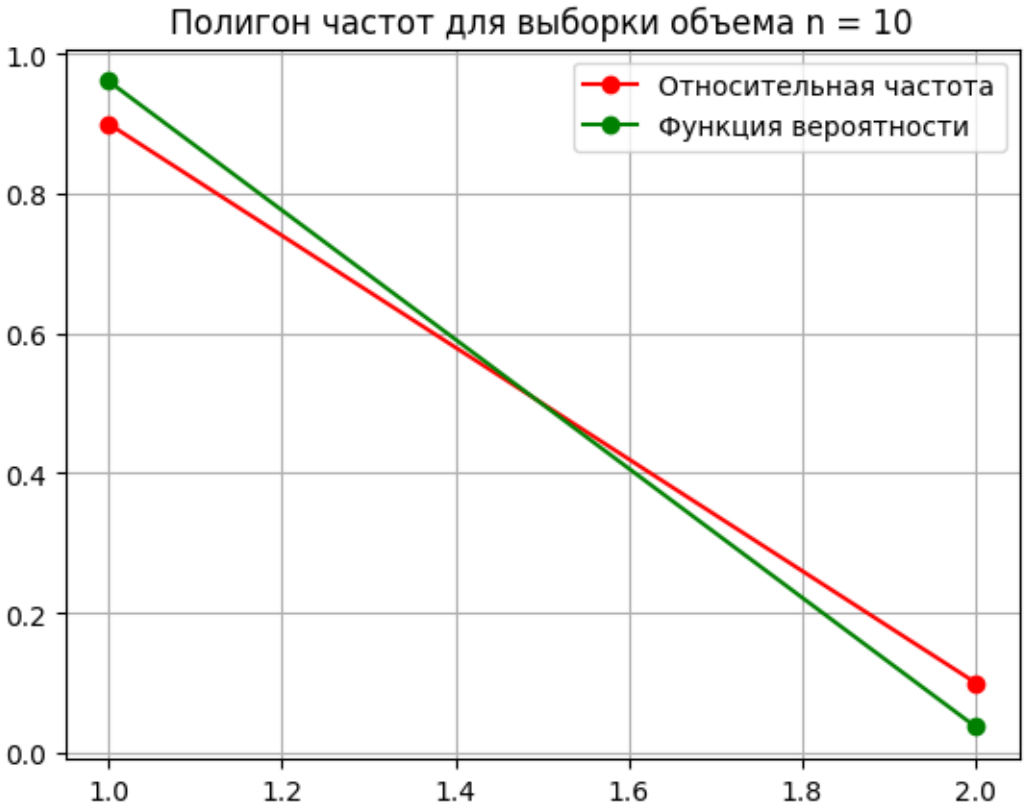
$$D_{m,n} = \sqrt{\frac{nm}{n+m}} \sup_{x \in \mathbb{R}} |\mathscr{F}_n(x) - \mathscr{F}_m(x)|$$

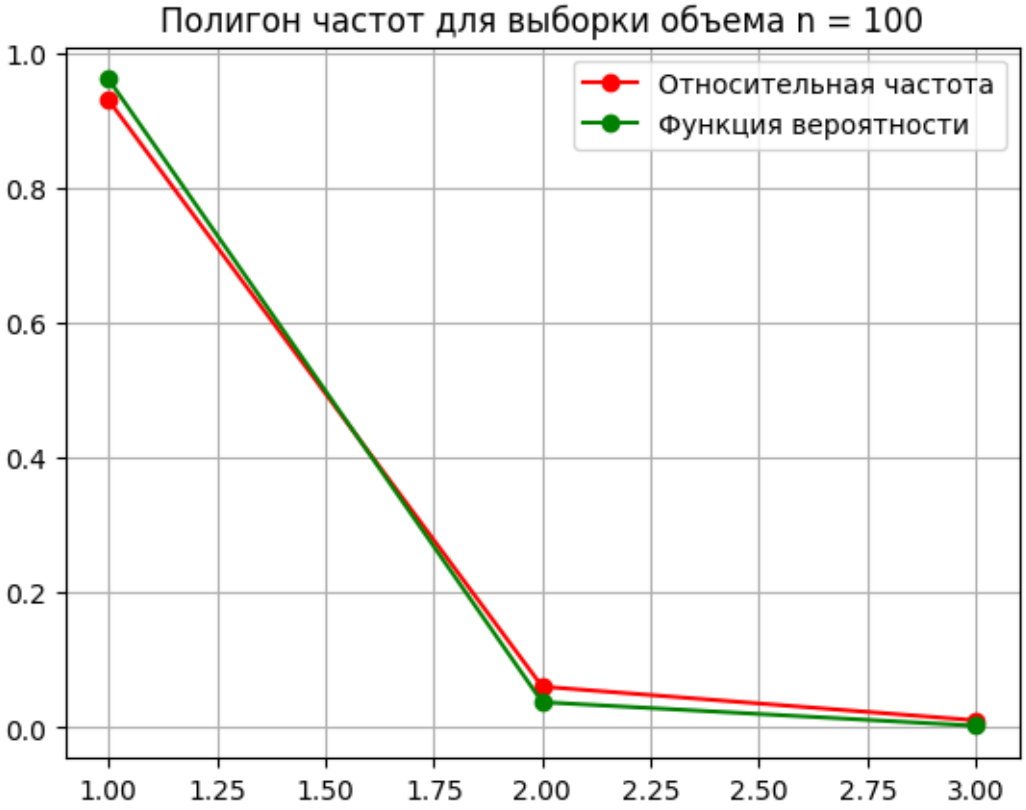
для каждой пары эмпирических функций распеределения находятся здесь ($D_{m,n}$ расчитывались для всехвозможных пар с учетом пятёрки выборок для каждого объёма).

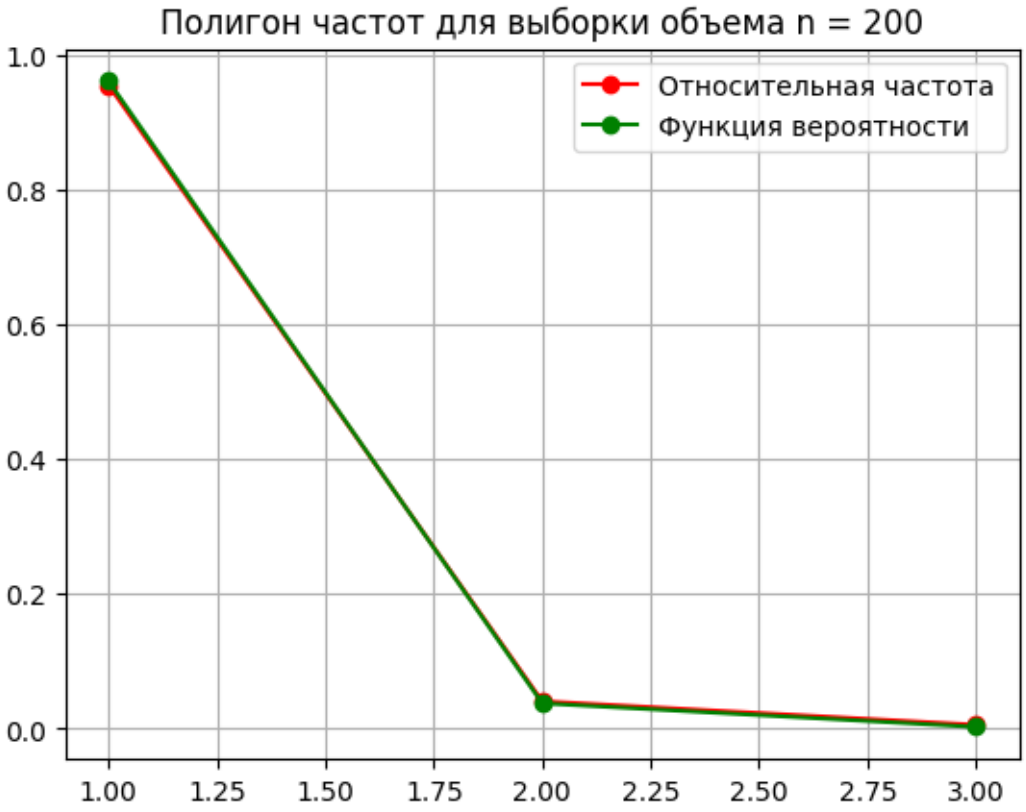
### 2.1.3. Построение гистограммы и полигона частот

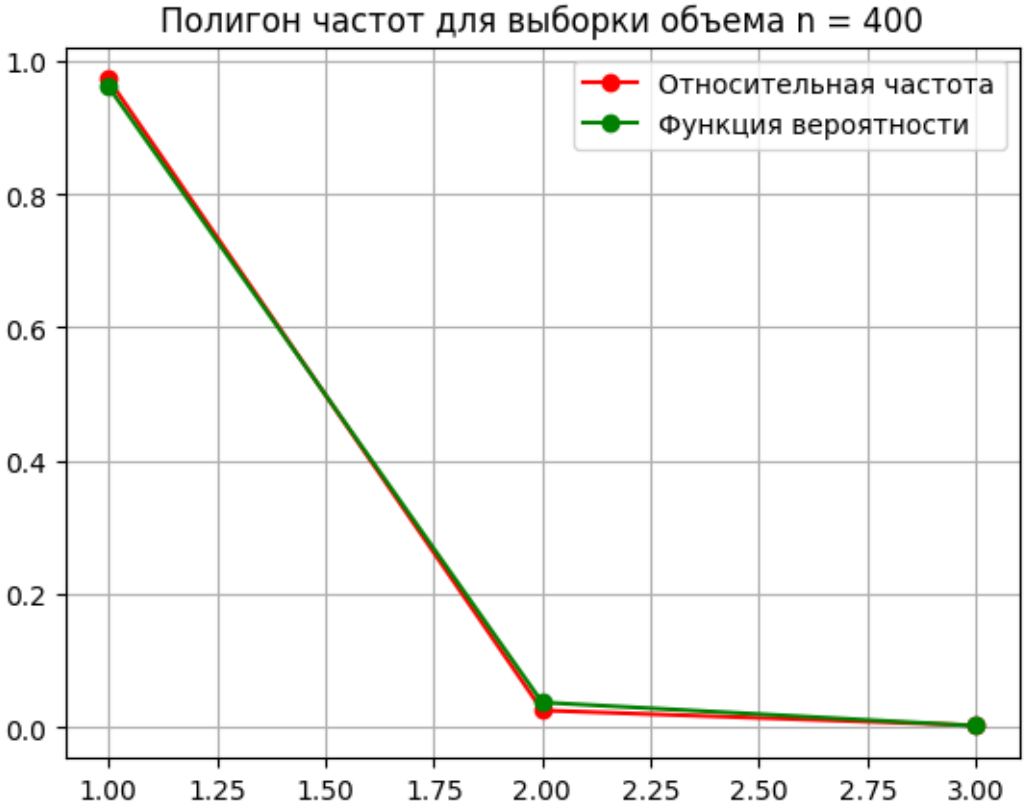
Построим полигон относительных частот для каждого объёма выборки.

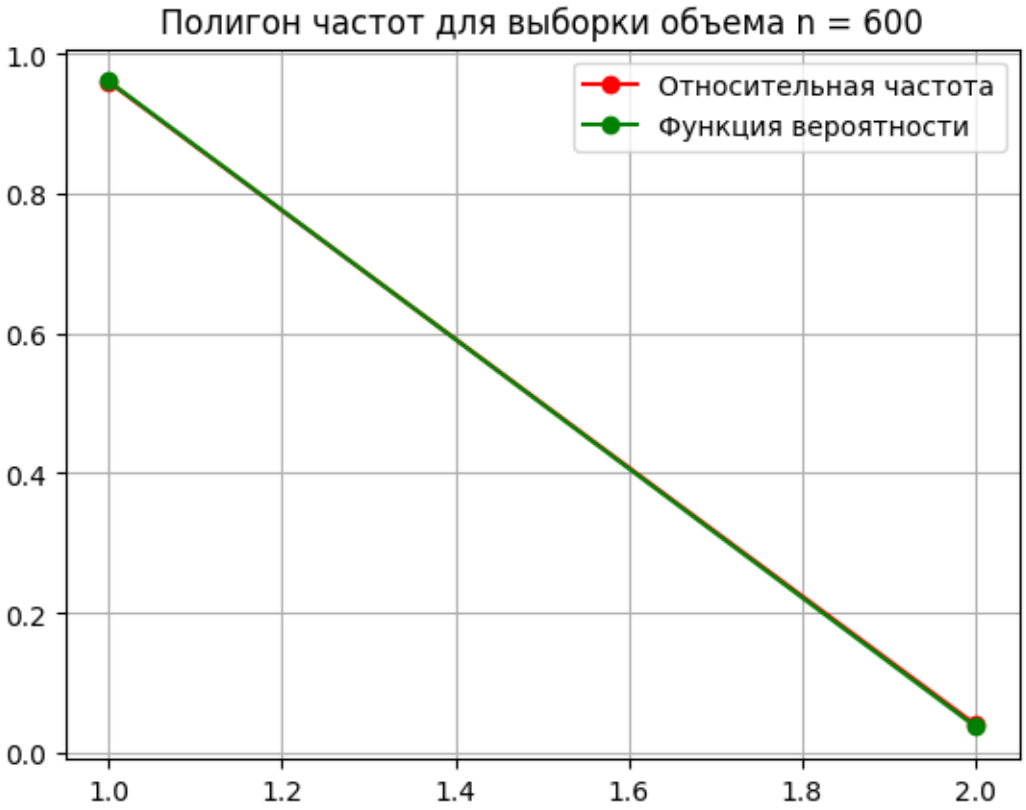


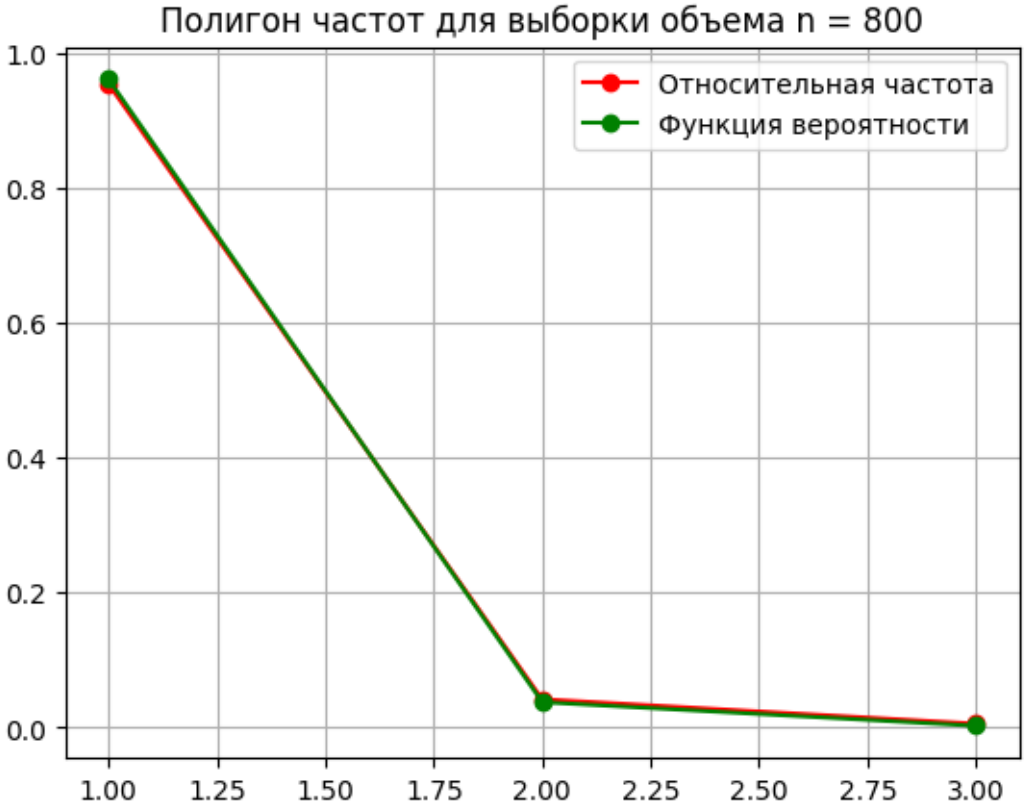

Полигон частот для выборки объема n = 10
Относительная частота
Функция вероятности
1.0
0.8
0.6
0.4
0.2
0.0
1.0
1.2
1.4
1.6
1.8
2.0

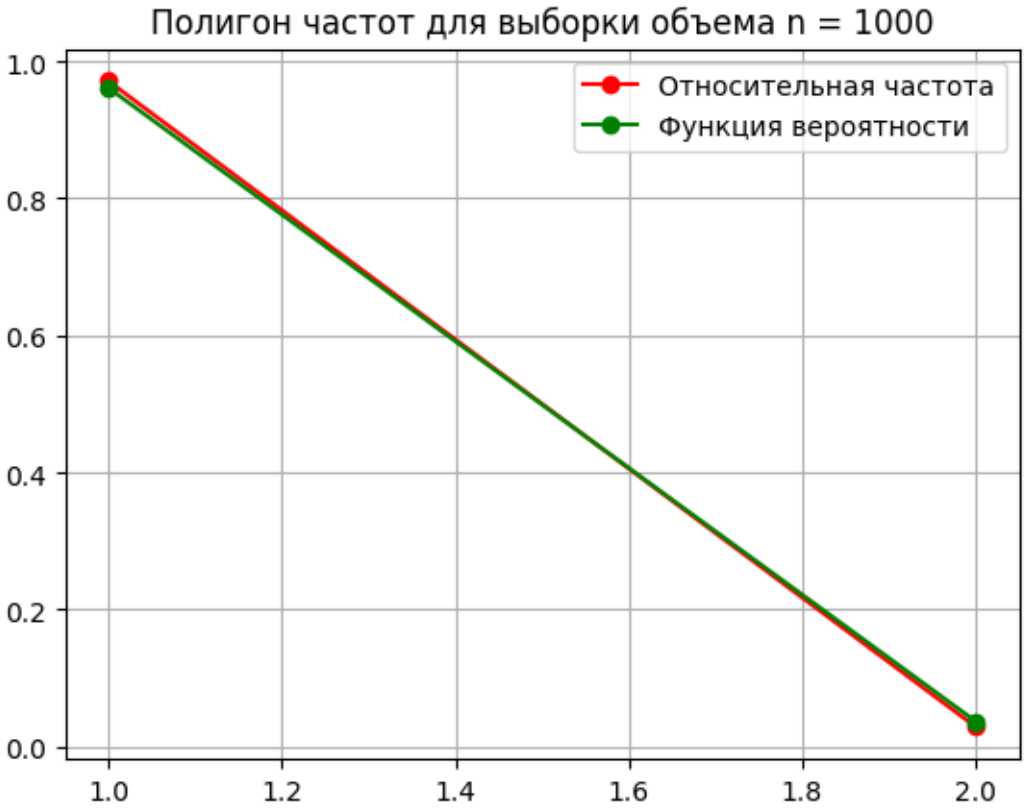

Полигон частот для выборки объема n = 100
Относительная частота
Функция вероятности
1.0
0.8
0.6
0.4
0.2
0.0
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00


Полигон частот для выборки объема n = 200
Относительная частота
Функция вероятности
1.0
0.8
0.6
0.4
0.2
0.0
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00


Полигон частот для выборки объема n = 400
Относительная частота
Функция вероятности
1.0
0.8
0.6
0.4
0.2
0.0
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00


Полигон частот для выборки объема n = 600
Относительная частота
Функция вероятности
1.0
0.8
0.6
0.4
0.2
0.0
1.0
1.2
1.4
1.6
1.8
2.0


Полигон частот для выборки объема n = 800
Относительная частота
Функция вероятности
1.0
0.8
0.6
0.4
0.2
0.0
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00




Видим, что графики относительных частот при увеличении объёма выборки всё лучше описывают графики функции вероятности.

Пусть логарифмическая случайная величина $\xi$ приняла значение $x_i$.

$\nu_i = \sum\limits_{j=1}^{n} I(X_j = x_i)$ - число элементов выборки, принявших значение $x_i$. Тогда согласно закону больших чисел относительная частота $\nu_i^* = \frac{\nu_i}{n}$ при $n \to \infty$ сходится по вероятности к $M[I(X_j = x_i)] = P(\xi = x_i)$. Такая картина и наблюдается на графиках: с увеличением $n$ относительные частоты сближаются с вероятностями.

Результат на графиках также логичен из следующей теоремы:

*Теорема.* Относительная частота произвольного события является оптимальной оценкой для вероятности этого события.

### 2.1.4. Вычисление выборочных моментов

Значения выборочного среднего $\overline{X} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$ и выборочной дисперсии $\overline{S}^2 = \frac{1}{n}\sum\limits_{i=1}^{n}(X_i - \overline{X})^2$ для каждой сгенерированной находятся здесь.

Свойства оценок:

- ▷ Выборочное среднее - несмещённая оценка для математического ожидания случайной величины:

$$M\overline{X} = M[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} MX_i = \frac{n}{n}M\xi = M\xi$$

▷ Выборочное среднее - состоятельная оценка для математического ожидания случайной величины:

$D\overline{X} = D[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n^2}\sum_{i=1}^{n} DX_i = \frac{n}{n^2}D\xi = \frac{D\xi}{n} \to 0$ при $n \to \infty$

По неравенству Чебышева: $P(|\overline{X} - M\overline{X}| \geq \epsilon) \leq \frac{D\overline{X}}{\epsilon^2} \to 0$ при $n \to \infty$

Тогда $\overline{X} \xrightarrow{P} M\xi$.

▷ Выборочная дисперсия - смещённая оценка для дисперсии случайной величины:

$M\overline{S}^2 = \frac{n-1}{n}D\xi \neq D\xi$ (формула долго выводилась в лекциях)

▷ Выборочная дисперсия - состоятельная оценка для дисперсии случайной величины:

$D\overline{S}^2 = \frac{(n-1)^2}{n^3}(\mu_4 - \frac{n-3}{n-1}\mu_2^2)$, где $\mu_k = M(\xi - M\xi)^k$

По неравенству Чебышева: $P(|\overline{S}^2 - M\overline{S}^2| \geq \epsilon) \leq \frac{D\overline{S}^2}{\epsilon^2} \iff$
$P(|\overline{S}^2 - \frac{n-1}{n}D\xi| \geq \epsilon) \leq \frac{1}{\epsilon^2}\frac{(n-1)^2}{n^3}(\mu_4 - \frac{n-3}{n-1}\mu_2^2) \to 0$ при $n \to \infty$

Тогда при $n \to \infty$ получим, что: $\overline{S}^2 \xrightarrow{P} D\xi$.

Таким образом, при $n \to \infty$ выборочное среднее сходится по вероятности к математическому ожиданию случайной величины, а выборочная дисперсия - к дисперсии случайной величины. Покажем это в нашем примере.

Истинные значения математического ожидания и дисперсии:
$M\xi = \frac{1}{12\ln\frac{13}{12}} \approx 1.04111$; $D\xi = -\frac{13\ln\frac{13}{12}+1}{144\ln^2\frac{12}{13}} \approx 0.04396$

Анализируя полученные результаты и сравнивая экспериментальные значения с истинными значениями математического ожидания и дисперсии, видим, что эксперементальные значения близки к истинным. В этом можно убедиться, посмотрев на разницу между ними здесь. Действительно, разница мала.

## 2. Непрерывное распределение

Реализацию необходимых алгоритмов моделирования случайных величин, а также построение графиков можно посмотреть здесь.

### 2.2.1. Генерация выборок выбранных случайных величин

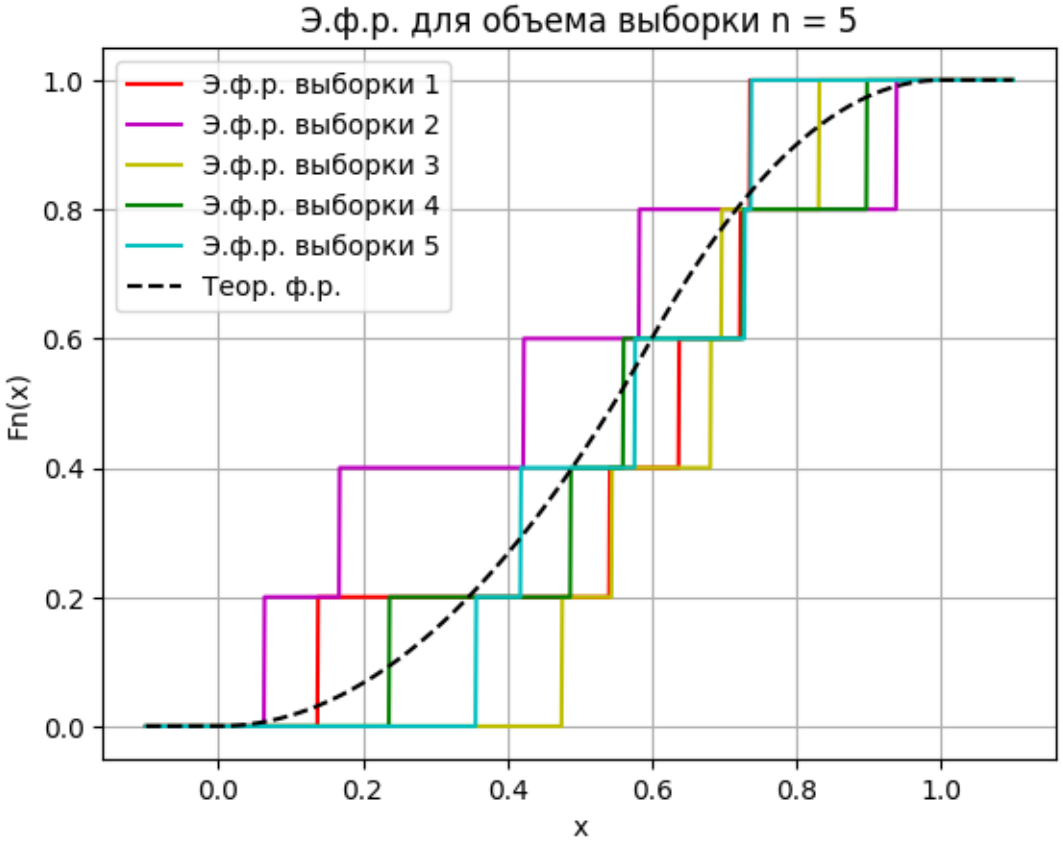
Для генерации выборок логарифмической случайной величины используется описанный мною ранее способ моделирования.

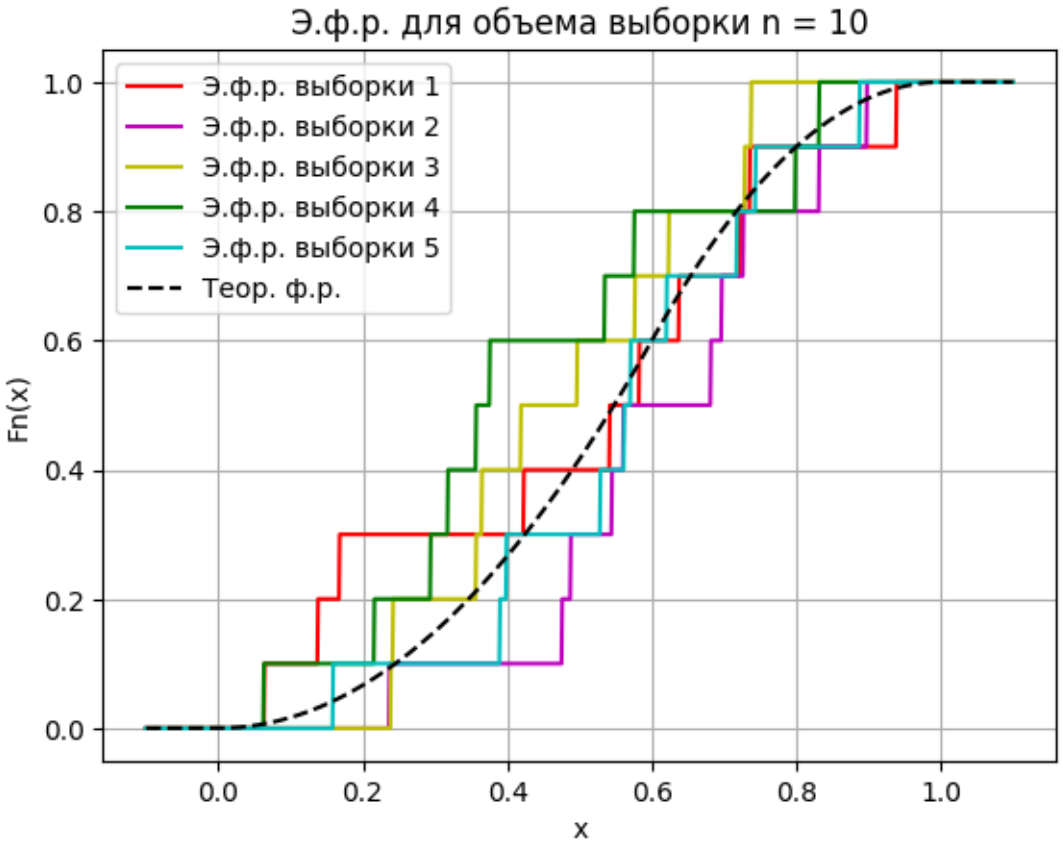
Объёмы сгенерированных выборок: $n = [5, 10, 100, 200, 400, 600, 800, 1000]$.

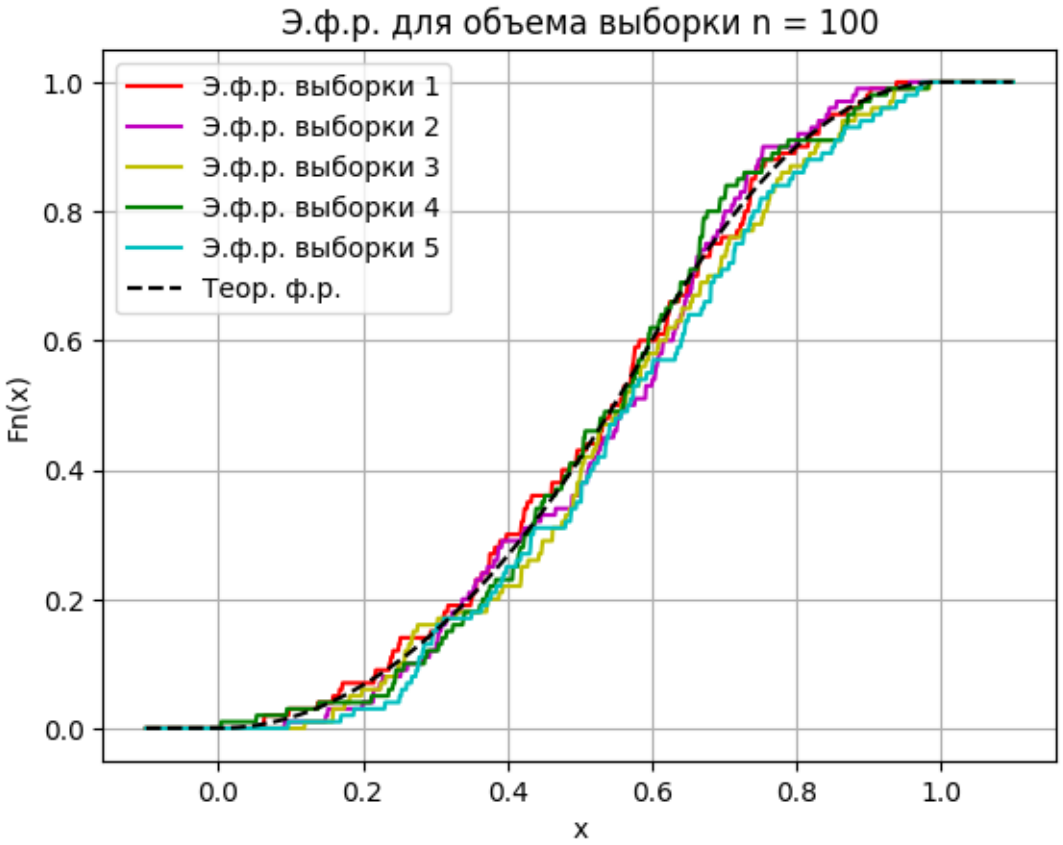
### 2.2.2. Построение эмпирической функции распределения

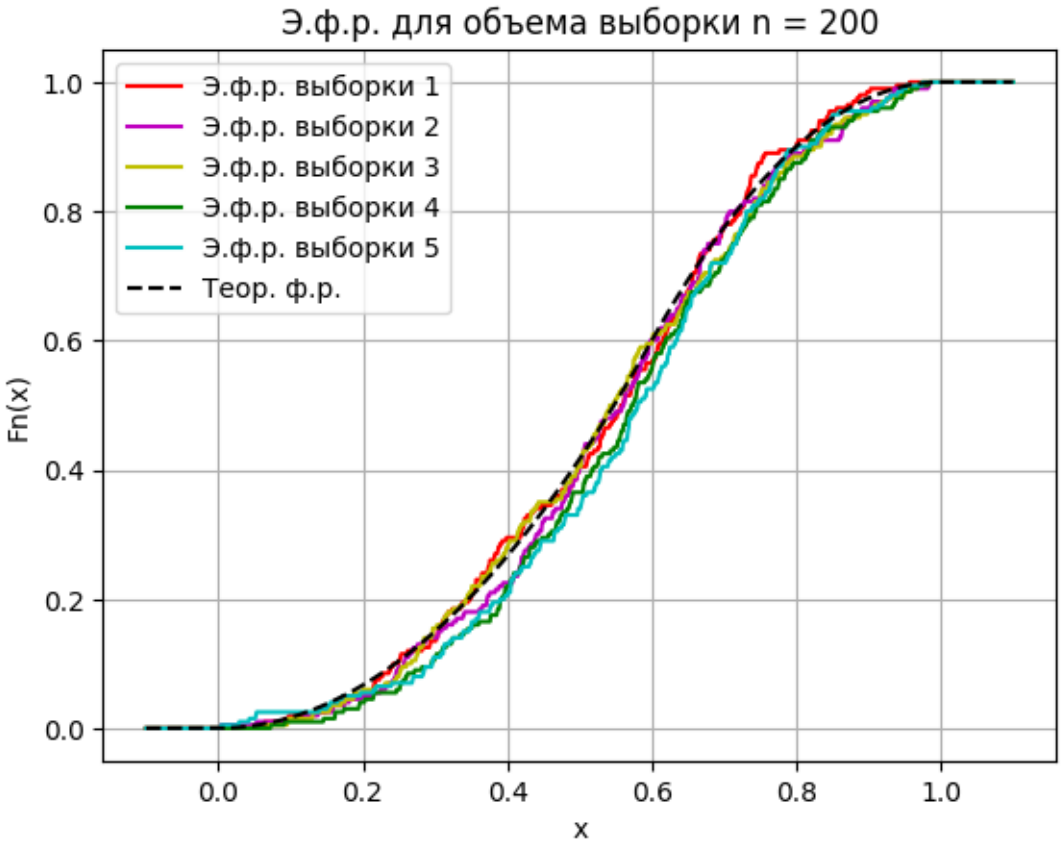
Эмпирическая функция распределения: $\mathscr{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq t)$

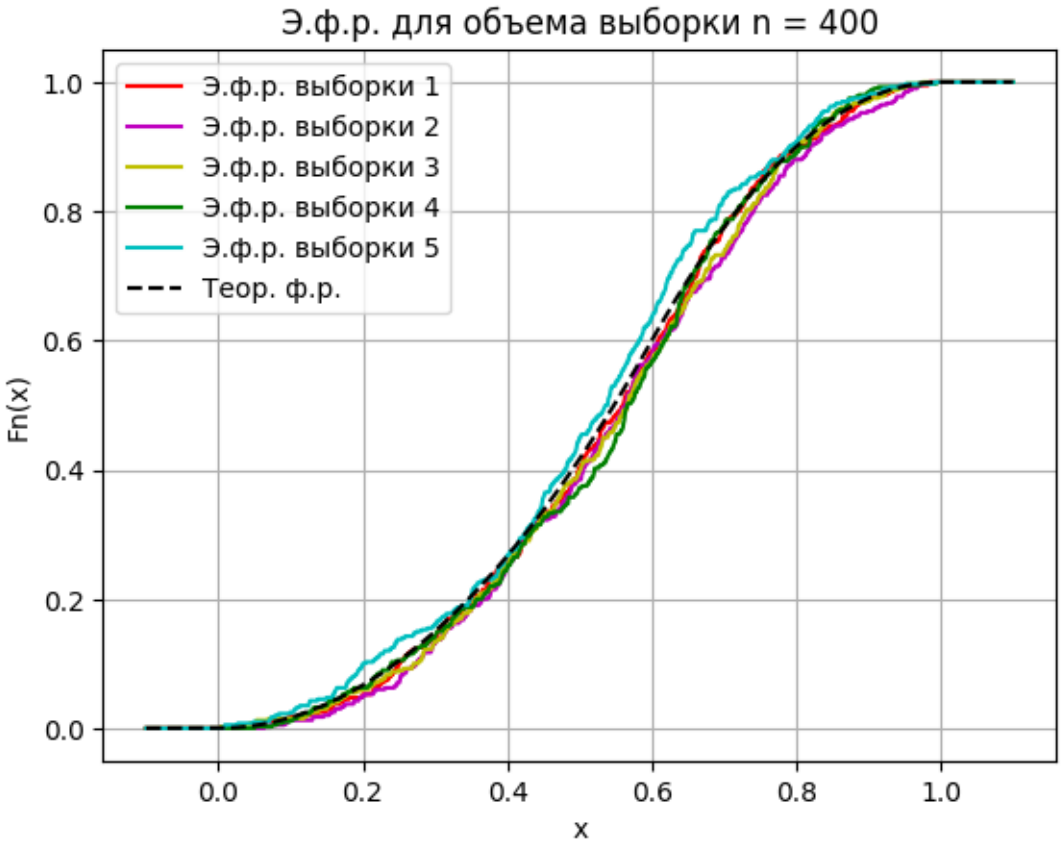
Графики эмпирических функций распределения для каждого из объемов выборки (на каждом графике изображена э. ф. р. для 5 выборок данного объема, а также функция распределения случайной величины):

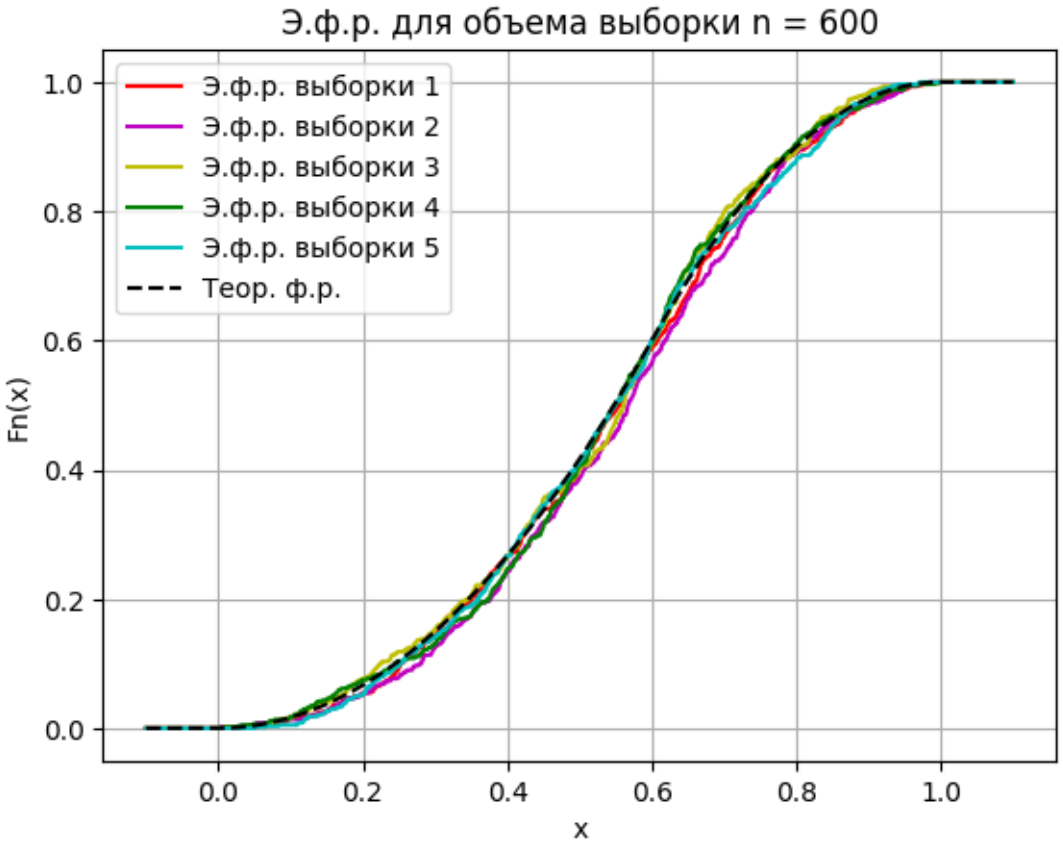


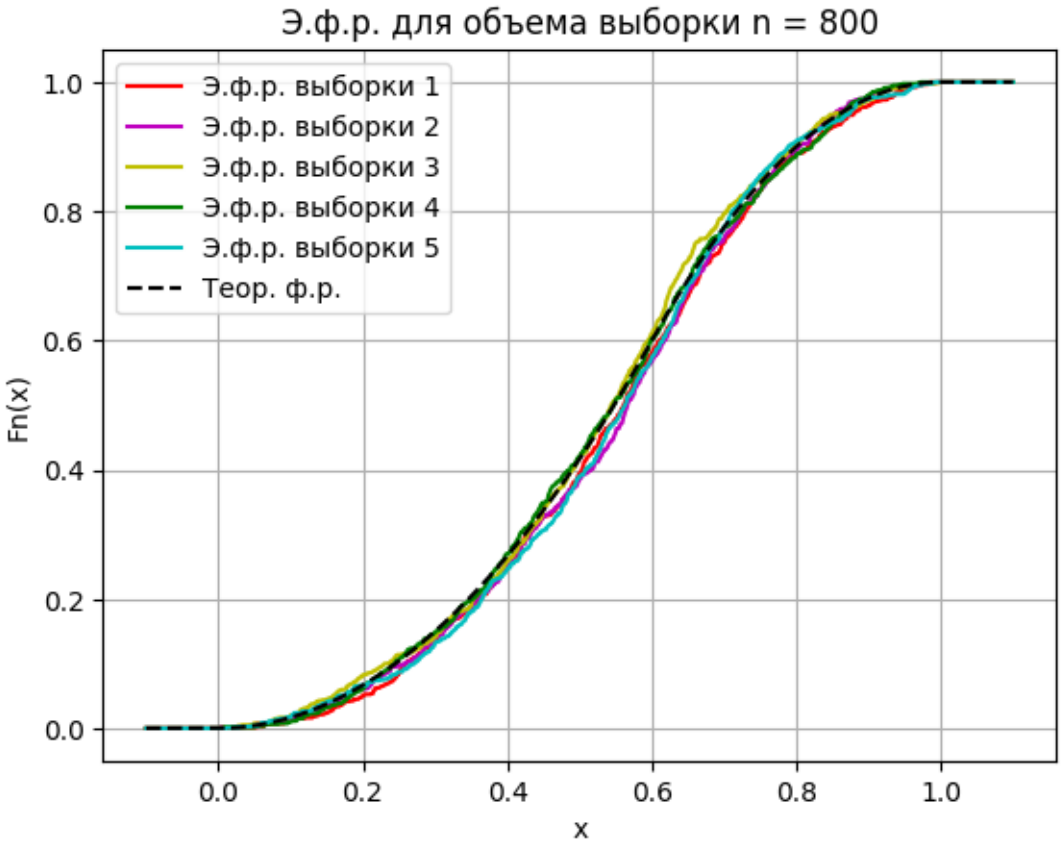

Э.ф.р. для объема выборки n = 10
Э.ф.р. выборки 1
Э.ф.р. выборки 2
Э.ф.р. выборки 3
Э.ф.р. выборки 4
Э.ф.р. выборки 5
Теор. ф.р.
Fn(x)
x
0.0
0.2
0.4
0.6
0.8
1.0

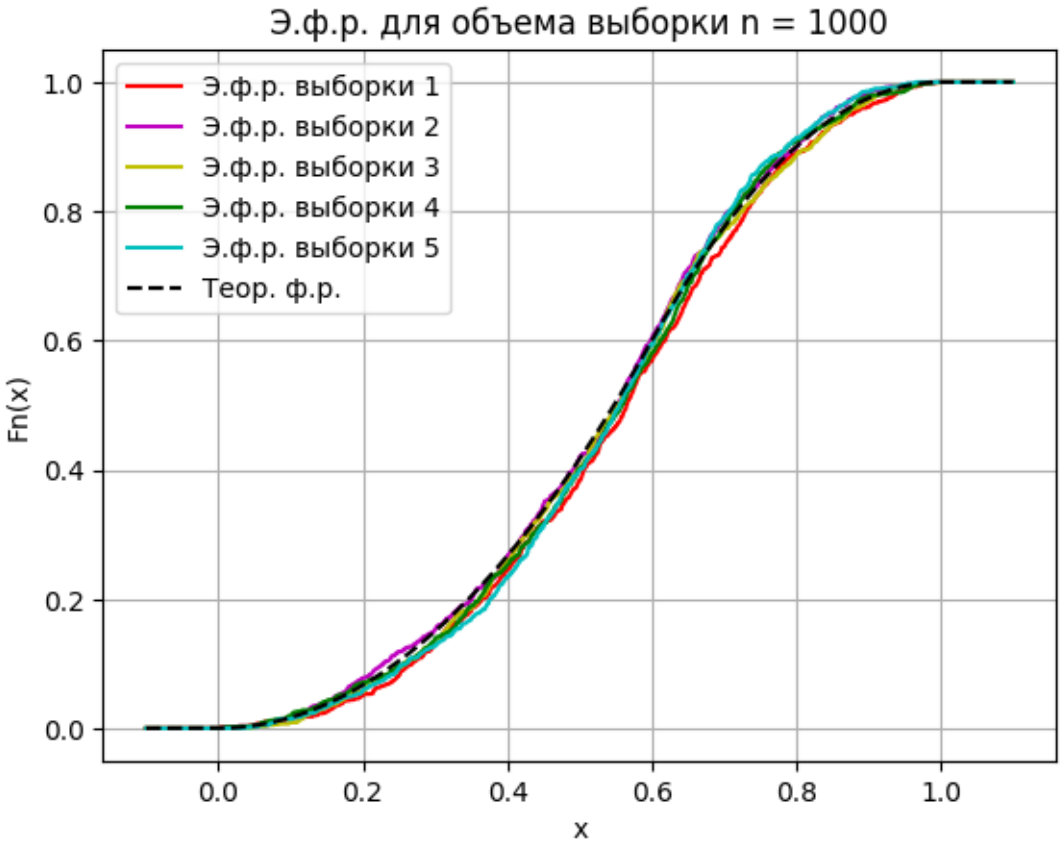

Э.ф.р. для объема выборки n = 100
Э.ф.р. выборки 1
Э.ф.р. выборки 2
Э.ф.р. выборки 3
Э.ф.р. выборки 4
Э.ф.р. выборки 5
Теор. ф.р.
Fn(x)
x
0.0
0.2
0.4
0.6
0.8
1.0


Э.ф.р. для объема выборки n = 200
Э.ф.р. выборки 1
Э.ф.р. выборки 2
Э.ф.р. выборки 3
Э.ф.р. выборки 4
Э.ф.р. выборки 5
Теор. ф.р.
Fn(x)
x
0.0
0.2
0.4
0.6
0.8
1.0


Э.ф.р. для объема выборки n = 400
Э.ф.р. выборки 1
Э.ф.р. выборки 2
Э.ф.р. выборки 3
Э.ф.р. выборки 4
Э.ф.р. выборки 5
Теор. ф.р.
Fn(x)
x
0.0
0.2
0.4
0.6
0.8
1.0


Э.ф.р. для объема выборки n = 600
Э.ф.р. выборки 1
Э.ф.р. выборки 2
Э.ф.р. выборки 3
Э.ф.р. выборки 4
Э.ф.р. выборки 5
Теор. ф.р.
Fn(x)
x
0.0
0.2
0.4
0.6
0.8
1.0


Э.ф.р. для объема выборки n = 800
Э.ф.р. выборки 1
Э.ф.р. выборки 2
Э.ф.р. выборки 3
Э.ф.р. выборки 4
Э.ф.р. выборки 5
Теор. ф.р.
Fn(x)
x
0.0
0.2
0.4
0.6
0.8
1.0



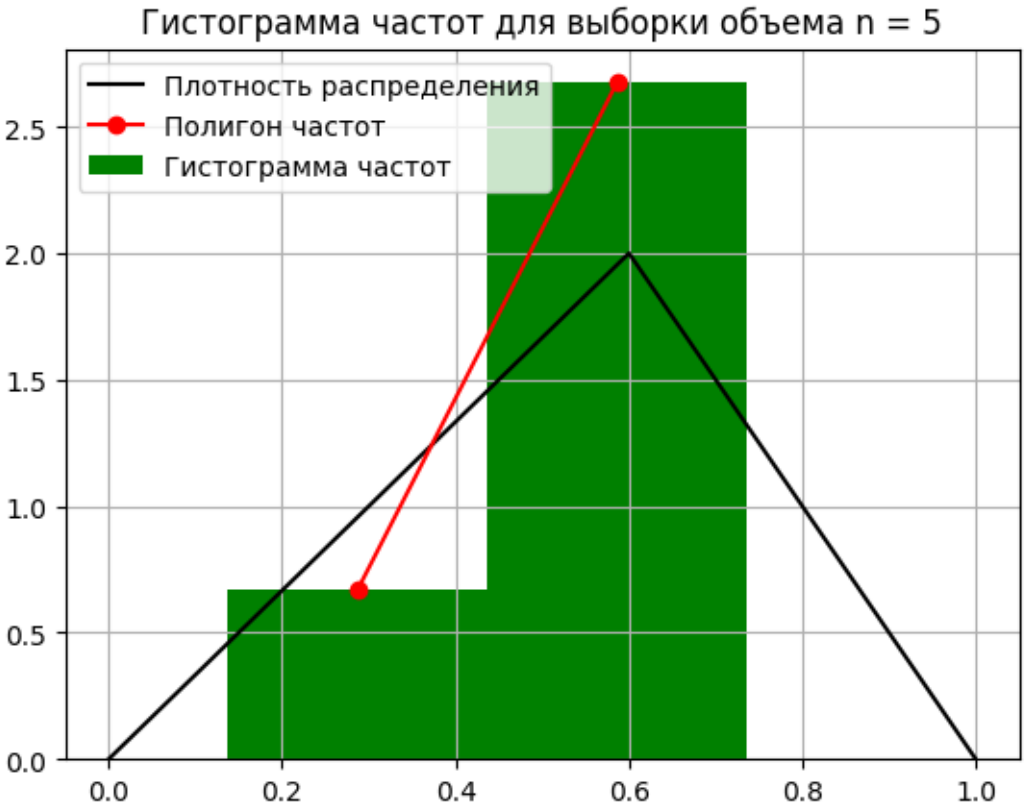

Вычисленные значения

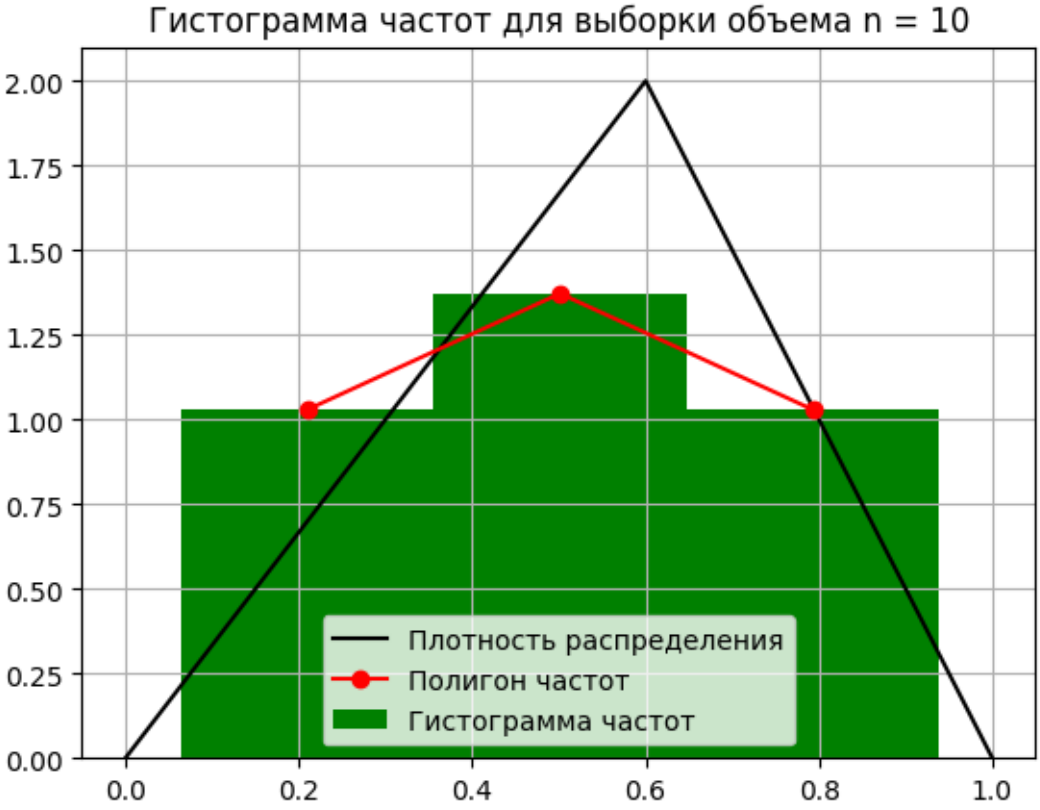
$$D_{m,n} = \sqrt{\frac{nm}{n+m}} \sup_{x \in \mathbb{R}} |\mathscr{F}_n(x) - \mathscr{F}_m(x)|$$

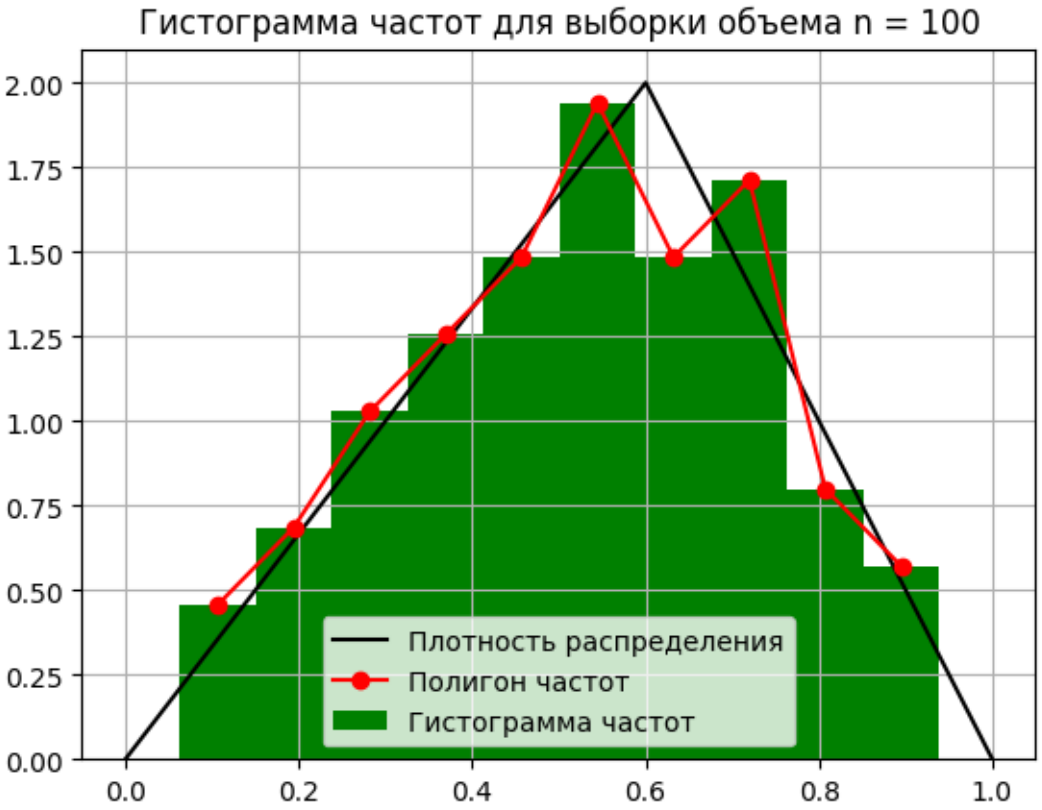
для каждой пары эмпирических функций распеределения находятся здесь ($D_{m,n}$ расчитывались для всехвозможных пар с учетом пятёрки выборок для каждого объёма).

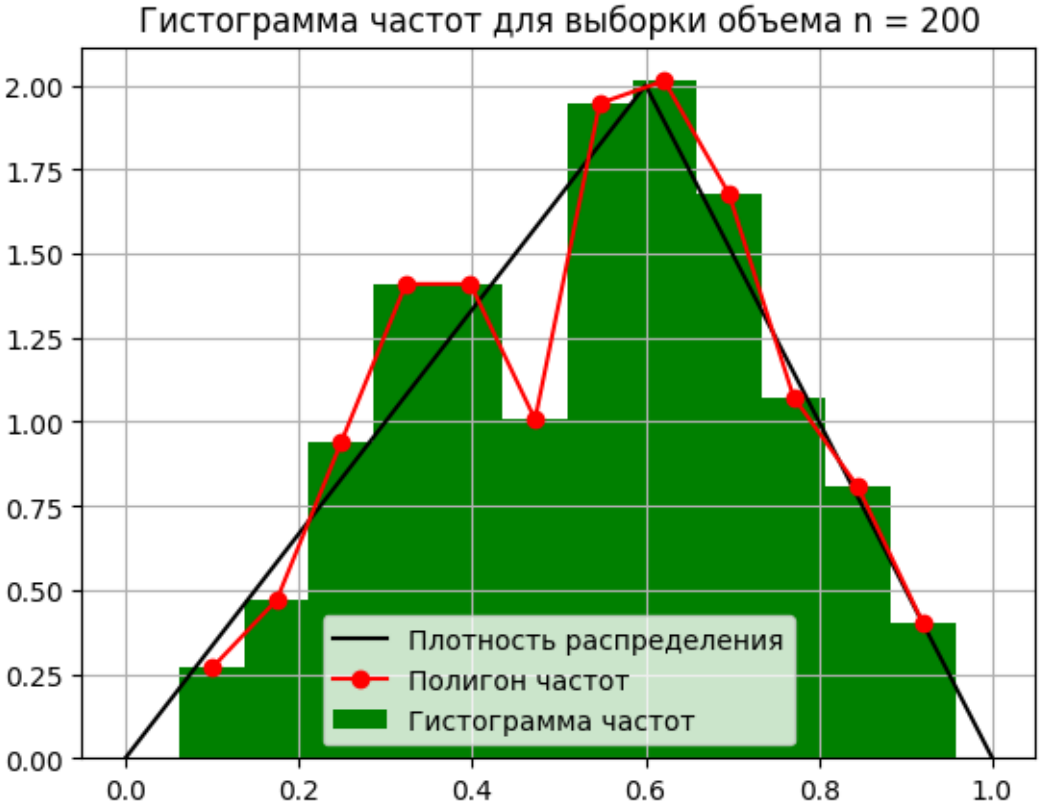
### 2.2.3. Построение гистограммы и полигона частот

Построим гистограмму и полигон частот для треугольного распределения

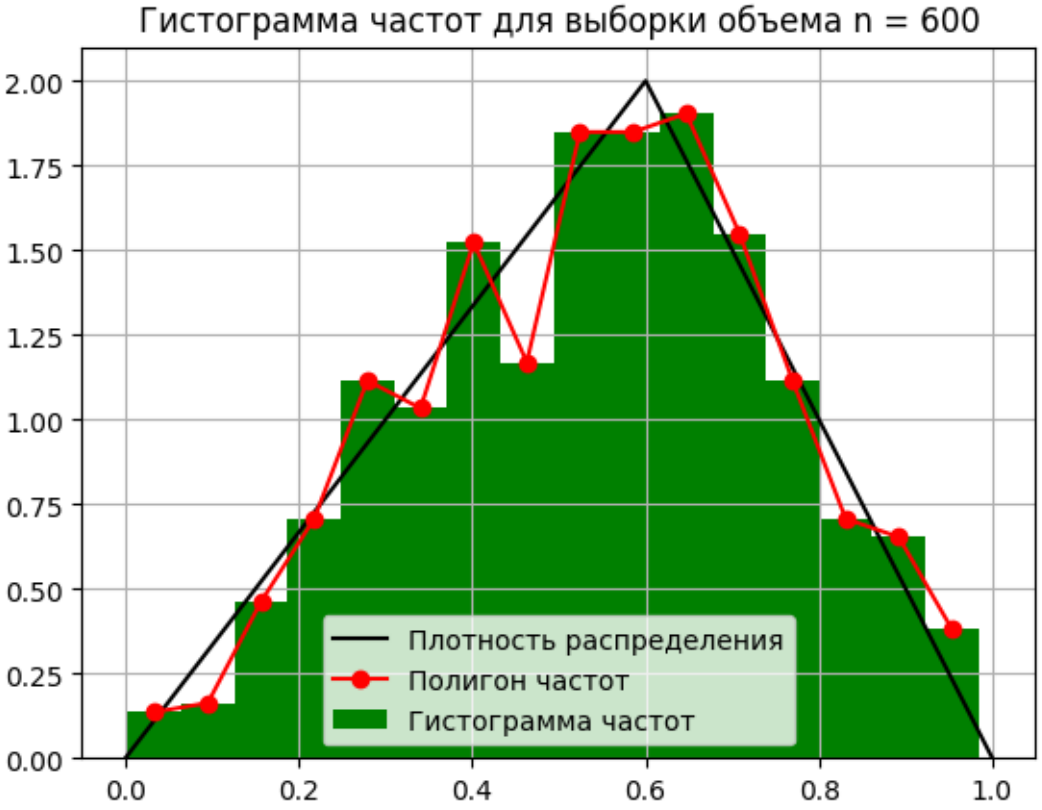

Гистограмма частот для выборки объема n = 5
Плотность распределения
Полигон частот
Гистограмма частот
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0


Гистограмма частот для выборки объема n = 10
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.8
1.0
Плотность распределения
Полигон частот
Гистограмма частот


Гистограмма частот для выборки объема n = 100
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.8
1.0
Плотность распределения
Полигон частот
Гистограмма частот


Гистограмма частот для выборки объема n = 200
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.8
1.0
Плотность распределения
Полигон частот
Гистограмма частот


Гистограмма частот для выборки объема n = 400
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.8
1.0
Плотность распределения
Полигон частот
Гистограмма частот


Гистограмма частот для выборки объема n = 600
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.8
1.0
Плотность распределения
Полигон частот
Гистограмма частот



Видим, что построенные графики гистограм и полигонов частот при увеличении объёма выборки всё лучше описывают графики плотности распределения.

Разбивая область возможных значений случайной величины $\xi$ на непересекающиеся интервалы $\Delta_1, ..., \Delta_N$, находим $\nu_i = \sum\limits_{j=1}^{n} I(X_j \in \Delta_i)$ - число наблю-

дений, попавших в интервал $\Delta_i$ и строим $\hat{f}_n(x) = \frac{\nu_i}{n|\Delta_i|}$, где $x \in \Delta_i \ \forall i = \overline{1,n}$ , а $|\Delta_i|$ - длина соответствующего интервала.

Пусть $\Delta_i = [a, b)$. Согласно закону больших чисел относительная частота $\frac{\nu_i}{n}$ при $n \to \infty$ сходится по вероятности к

$$M[I(X_j \in \Delta_i)] = P(\xi \in \Delta_i) = P(\xi \in [a,b)) = P(\xi < b) - P(\xi < a) =$$
$$= F(b) - F(a) = \int\limits_{-\infty}^{b} f(x)dx - \int\limits_{-\infty}^{a} f(x)dx = \int\limits_{a}^{b} f(x)dx = \int\limits_{\Delta_i} f(x)dx =$$
$$\overset{\text{Th о среднем}}{=} f(a_i)|\Delta_i|, \text{ где } a_i \in \Delta_i$$

Значит, при больших $n$ и достаточно мелком разбиении $\hat{f}_n(x) \approx f(a_i)$, где $x \in \Delta_i$, т. е. гистограмма $\hat{f}_n(x)$ будет приближать плотность $f(x)$, что подтверждается и на нашем примере.

Результат на графиках также логичен из следующей теоремы:

*Теорема.* Относительная частота произвольного события является оптимальной оценкой для вероятности этого события.

В нашем случае теорема говорит о том, что при большом объёме выборки относительная частота $\frac{\nu_i}{n} \approx P(\xi \in \Delta_i)$. Тогда выполнив деление на длину интервала, получим что $\hat{f}_n(x) \approx f(x)$.

### 2.2.4. Вычисление выборочных моментов

Значения выборочного среднего $\overline{X} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$ и выборочной дисперсии $\overline{S}^2 = \frac{1}{n}\sum\limits_{i=1}^{n}(X_i - \overline{X})^2$ для каждой сгенерированной находятся здесь.

Свойства оценок (выводы следующих свойств были ранее):

- ▷ Выборочное среднее - несмещённая и состоятельная оценка для математического ожидания случайной величины.
- ▷ Выборочная дисперсия - смещённая и состоятельная оценка для дисперсии случайной величины.

При $n \to \infty$ выборочное среднее сходится по вероятности к математическому ожиданию случайной величины, а выборочная дисперсия - к дисперсии случайной величины. Покажем это в нашем примере.

Истинные значения математического ожидания и дисперсии:
$M\xi = \frac{8}{15} \approx 0.53333$; $D\xi = \frac{19}{450} \approx 0.04222$

Анализируя полученные результаты и сравнивая экспериментальные значения с истинными значениями математического ожидания и дисперсии, видим, что эксперементальные значения близки к истинным. В этом можно убедиться, посмотрев на разницу между ними здесь. Действительно, разница мала.

# Домашнее задание 3.

## Построение точечных оценок параметра распределения

### 1. Дискретное распределение

Значения полученных оценок для сгенерированных выборок получаются программно здесь.

#### 3.1.1. Получение оценок методом моментов и методом максимального правдоподобия

▷ *Метод моментов*

Заметим, что $\alpha_1 = M\xi = -\frac{1}{\ln(1-\theta)}\frac{\theta}{1-\theta} < \infty$, а также что $\theta \in (0,1)$ - одномерный неизвестный параметр. Тогда рассмотрим систему с одним неизвестным $\theta$:

$$\begin{cases} \hat{\alpha_1} = \alpha_1 = -\frac{1}{\ln(1-\theta)}\frac{\theta}{1-\theta} \end{cases}$$

Не можем получить явное выражение для оценки $\hat{\theta}$, построенной по методу моментов, через $\hat{\alpha_1}$, но для неё верно: $\frac{\hat{\theta}}{ln(1-\hat{\theta})(1-\hat{\theta})} = -\hat{\alpha_1}$. Сравним для сгенерированных выборок значения $-\hat{\alpha_1}$ с левой частью равенства. Полученные значения находятся здесь.

▷ *Метод максимального правдоподобия*

По определению: оценка максимального правдоподобия - построенное по реализации выборки $\overline{x}$ значение

$$\hat{\theta}_{\text{м.м.п.}} = \arg\max_{\theta\in(0,1)} L(\overline{x};\theta)$$

Знаем, что: $P(\xi = x) = -\ln(1-\theta)^{-1} \cdot \theta^x \cdot x^{-1}$. Исходя из этого, найдем функцию правдоподобия:

$$L(\overline{x};\theta) = \prod_{i=1}^{n} P(\xi = x_i) = \frac{(-1)^n}{\ln^n(1-\theta)} \prod_{i=1}^{n} \frac{\theta^{x_i}}{x_i} = \frac{(-1)^n}{\ln^n(1-\theta)} \frac{\theta^{\sum\limits_{i=1}^{n} x_i}}{\prod\limits_{i=1}^{n} x_i}$$

Заметим, что

$$\frac{\partial}{\partial\theta}\frac{1}{\ln^n(1-\theta)} = -\frac{n}{\ln^{n+1}(1-\theta)}\frac{1}{1-\theta}(-1) = \frac{n}{(1-\theta)\ln^{n+1}(1-\theta)}$$

$$\frac{\partial}{\partial\theta}\frac{\theta^{\sum x_i}}{\prod x_i} = \frac{1}{\prod x_i}\sum x_i\theta^{\sum x_i-1}$$

Тогда для $\hat{\theta}_{\text{м.м.п.}}$ будет верно:

$$\frac{\partial L(\overline{x};\theta)}{\partial\theta} = -\left(\frac{n}{(1-\theta)\ln^{n+1}(1-\theta)}\cdot\frac{\theta^{\sum x_i}}{\prod x_i} + \frac{1}{\ln^n(1-\theta)}\cdot\frac{1}{\prod x_i}\sum x_i\theta^{\sum x_i-1}\right) = 0$$

$$\frac{n}{(1-\theta)\ln(1-\theta)}\theta^{\sum x_i} + \sum x_i\theta^{\sum x_i-1} = 0$$

$$\frac{\theta}{(1-\theta)\ln(1-\theta)} = -\frac{1}{n}\sum_{i=1}^{n} x_i$$

Т. е. для оценки $\hat{\theta}_{\text{м.м.п.}}$ получили выражение, аналогичное выражению для $\hat{\theta}_{\text{м.м.}}$: $\frac{\hat{\theta}}{ln(1-\hat{\theta})(1-\hat{\theta})} = -\hat{\alpha_1}$. Полученные значения находятся всё ещё здесь.

### 3.1.2. Поиск оптимальных оценок

Преобразуем функцию вероятности:

$$P(\xi = x) = -\frac{1}{\ln(1-\theta)}\frac{\theta^x}{x} = \exp\{-\ln(-\ln(1-\theta)) + x\ln\theta - \ln x\} =$$
$$= \exp\{\ln\theta\cdot x - \ln(-\ln(1-\theta)) - \ln x\}$$

Положим: $A(\theta) = ln\theta; B(x) = x; C(\theta) = -\ln(-\ln(1-\theta)); D(x) = -\ln x \Rightarrow$ логарифмическое параметрическое семейство - экспоненциальное.

Заметим, что $\forall x \in \mathbb{N}\ \sqrt{P_\theta(\xi = x)} = \sqrt{-\frac{1}{\ln(1-\theta)}\frac{\theta^x}{x}}$ - непрерывная и дифференцируемая функция по $\theta$ (помним, что $\theta \in (0,1)$). А также

$$\frac{\partial P_\theta(\xi = x)}{\partial\theta} = \sqrt{\frac{\theta^x}{x}}\frac{1}{2\sqrt{-\frac{1}{\ln(1-\theta)}}}\frac{-1}{(1-\theta)\ln^2(1-\theta)} = -\sqrt{\frac{\theta^x}{x}}\frac{\sqrt{-\ln(1-\theta)}}{2(1-\theta)\ln^2(1-\theta)}$$

- непрерывная функция по $\theta$. Таким образом, $\forall x \in \mathbb{N}\ \sqrt{P_\theta(\xi = x)}$ непрерывно дифференцируема во всех точках $\theta \in (0,1)$. Значит, можем утверждать о регулярности семейства.

Найдем функцию вклада выборки:

$$V(X;\theta) = \frac{\partial \ln L(X;\theta)}{\partial\theta} = \sum_{i=1}^{n} \frac{\partial \ln P_\theta(X_i)}{\partial\theta} = \sum_{i=1}^{n} \frac{\partial}{\partial\theta}(A(\theta)B(X_i) + C(\theta) + D(X_i)) =$$

$$= A'(\theta)\sum_{i=1}^{n} B(X_i) + n \cdot C'(\theta) = n \cdot A'(\theta)(\frac{1}{n}\sum_{i=1}^{n} B(X_i) + \frac{C'(\theta)}{A'(\theta)})$$

$$\Rightarrow \frac{1}{n \cdot A'(\theta)} V(X;\theta) = \frac{1}{n}\sum_{i=1}^{n} B(X_i) - (-\frac{C'(\theta)}{A'(\theta)})$$

$C'(\theta) = -\frac{1}{-\ln(1-\theta)} \cdot \frac{-1}{1-\theta} \cdot (-1) = \frac{1}{(1-\theta)\ln(1-\theta)}$
$A'(\theta) = \frac{1}{\theta}$

$a(\theta) = \frac{1}{n \cdot A'(\theta)}$; $T(X) = \frac{1}{n}\sum\limits_{i=1}^{n} B(X_i) = \frac{1}{n}\sum\limits_{i=1}^{n} X_i = \overline{X}$;
$\tau(\theta) = -\frac{C'(\theta)}{A'(\theta)} = -\frac{\theta}{(1-\theta)\ln(1-\theta)}$

Воспользуемся вычисленными ранее $M\overline{X}$ и $M\xi$: $M\overline{X} = M\xi = -\frac{\theta}{(1-\theta)\ln(1-\theta)}$
$\Rightarrow \overline{X}$ - несмещенная оценка параметра $\tau(\theta)$.

Тогда, пользуясь следствием из неравенства Рао-Крамера, получим что $D_\theta\overline{X} = \frac{[\tau'(\theta)]^2}{i_n(\theta)}$ (дисперсия других несмещённых оценок не меньше). Тогда $\overline{X}$ - оптимальная оценка параметра $\tau(\theta) = -\frac{\theta}{(1-\theta)\ln(1-\theta)}$. Заметим, что разрешить данное уравнение относительно $\theta$ и получить оптимальную оценку для $\theta$ имеющимися способами не представляется возможным.

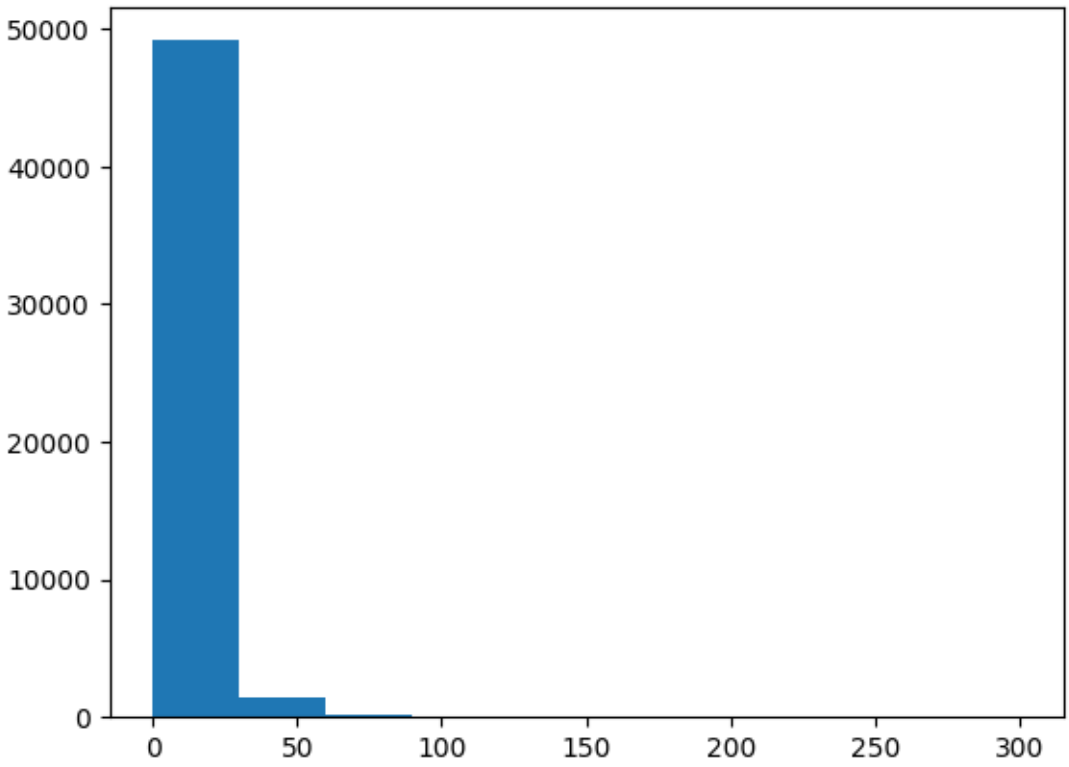
Приведем значения полученной оценки для сгенерированных выборок. Полученные значения находятся здесь.

### 3.1.3. Работа с данными

Как мы знаем, логарифмическое распределение используется для описания разнообразия выборки, то есть того, сколько элементов данного типа содержится в выборке элементов.

Выбранный мною датасет "Game Recommendations on Steam" описывает более 38 миллионов предварительно обработанных пользовательских рекомендаций (отзывов) из магазина Steam, а также содержит таблицу игр с информацией о рейтингах, ценах в долларах США, дате выхода и т.д.

Рассмотрим столбец price_final, содержащий цены на игры после скидок (указаны в долларах США).




▷ *Значение выборочного среднего и выборочной дисперсии*

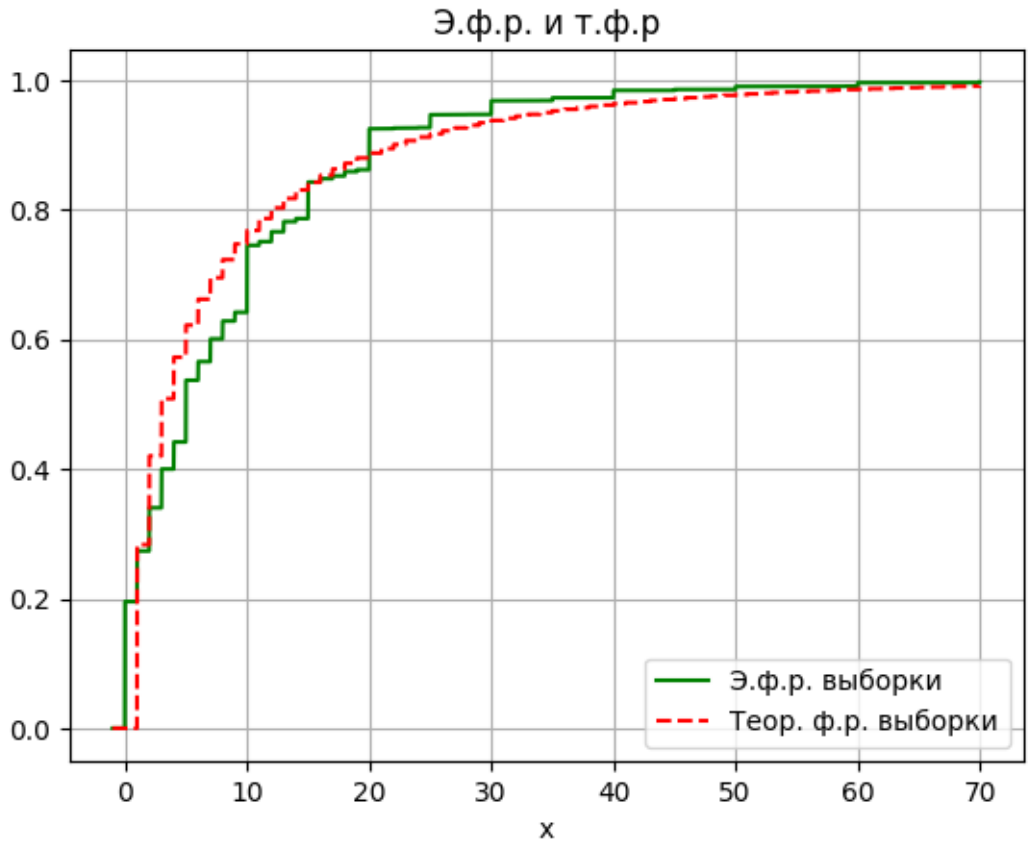
$\overline{X} = 8.61442$; $\overline{S}^2 = 132.75303$

▷ *Значения оценок*

Для оценки $\hat{\theta}$, построенной и по методу моментов, и по методу максимального правдоподобия, было вычислено: $\frac{\hat{\theta}}{ln(1-\hat{\theta})(1-\hat{\theta})} = -\hat{\alpha_1}$.

Для наших данных: $-\hat{\alpha_1} = -8.61442$. Численное решение уравнения относительно $\hat{\theta}$ даёт решение $\hat{\theta} = 0.96713$

Посмотрим на графики эмпирической и теоретической функции распределения:




Таким образом, графики эмпирической и теоретической функции распределения близки, но не совпадают. Значение оптимальной оценки параметра проверять нет смысла, т. к. уравнение относительно $\hat{\theta}$ будет идентичным.

## 2. Непрерывное распределение

Значения полученных оценок для сгенерированных выборок получаются программно здесь.

### 3.2.1. Получение оценок методом моментов и методом максимального правдоподобия

▷ *Метод моментов*

Заметим, что $\alpha_1 = M\xi = \frac{\theta+1}{3} < \infty$, а также что $\theta \in (0, 1)$ - одномерный неизвестный параметр. Тогда рассмотрим систему с одним неизвестным $\theta$:

$$\left\{\hat{\alpha_1} = \alpha_1 = \frac{\theta+1}{3}\right.$$

Тогда оценка, построенной по методу моментов $\hat{\theta} = 3\hat{\alpha_1} - 1$.

Приведем значения полученной оценки для сгенерированных выборок. Полученные значения находятся здесь.

▷ *Метод максимального правдоподобия*

Плотность распределения: $f_\xi(x) = \begin{cases} \frac{2x}{\theta}, & \text{если } x \in [0, \theta] \\ \frac{2(1-x)}{1-\theta}, & \text{если } x \in (\theta, 1]; \\ 0, & \text{иначе} \end{cases}$

Функция правдоподобия:

$$L(\overline{x}; \theta) = \prod_{i=1}^{n} f(x_i) = 2^n \prod_{i=1}^{n} \left( \frac{x_i}{\theta} I(0 \le x_i \le \theta) + \frac{1 - x_i}{1 - \theta} I(\theta < x_i \le 1) \right)$$

Стандартные методы вычисления оценки максимального правдоподобия (приравнивание производной функции правдоподобия к нулю или поиск точки с максимальным значением функции правдоподобия на графике) не помогают её вычислить. Согласно трудам Эдди Оливера (1972), не хватает методов математического анализа для поиска максимума.

1. Если $x_{(j)} < \theta < x_{(j+1)}$, то:

$$L(\overline{x}; \theta) = 2^n \prod_{i=1}^{j} \frac{x_i}{\theta} \prod_{i=j+1}^{n} \frac{1 - x_i}{1 - \theta} = 2^n \theta^{-j} (1 - \theta)^{-(n-j)} \prod_{i=1}^{j} x_i \prod_{i=j+1}^{n} (1 - x_i)$$

$$\Rightarrow \frac{\partial \ln L}{\partial \theta} = -\frac{j}{\theta} + \frac{n - j}{1 - \theta} \Rightarrow \frac{\partial^2 \ln L}{\partial \theta} = \frac{j}{\theta^2} + \frac{n - j}{(1 - \theta)^2} > 0$$

   Из положительности второй производной понятно, что любая существующая стационарная точка будет минимумом, а не максимумом.

2. Если $0 < \theta < x_{(1)}$, то: $L(\overline{x}; \theta) = 2^n (1 - \theta)^{-n} \prod_{i=1}^{n} (1 - x_i)$ - строго возрастающая функция от $\theta$

3. Если $x_{(n)} < \theta < 1$, то: $L(\overline{x}; \theta) = 2^n \theta^{-n} \prod_{i=1}^{n} x_i$ - строго убывающая функция от $\theta$

Вывод из рассуждений выше: $\hat{\theta}_{\text{м.м.п.}} = x_{(i)}$, где $i \in \overline{1, n}$.

Соответственно, $\frac{\partial \ln L}{\partial \theta}$ должна быть положительной слева от $x_{(j)}$, а справа - отрицательной:

$$\lim_{\theta \to x_{(j)}-} \left( -\frac{j-1}{\theta} + \frac{n-j+1}{1-\theta} \right) > 0$$

$$\lim_{\theta \to x_{(j)}+} \left( -\frac{j}{\theta} + \frac{n-j}{1-\theta} \right) < 0$$

$$-\frac{j}{x_{(j)}} + \frac{n-j}{1-x_{(j)}} < 0 < -\frac{j-1}{x_{(j)}} + \frac{n-j+1}{1-x_{(j)}} \Rightarrow$$

$$-j + jx_{(j)} + nx_{(j)} - jx_{(j)} < 0 < -j + jx_{(j)} + 1 - x_{(j)} + nx_{(j)} - jx_{(j)} + x_{(j)} \Rightarrow$$
$$-j + nx_{(j)} < 0 < -(j-1) + nx_{(j)} \Rightarrow -\frac{j}{n} < -x_{(j)} < -\frac{j-1}{n} \Rightarrow \frac{j-1}{n} < x_{(j)} < \frac{j}{n}$$

Таким образом:

$$\hat{\theta}_{\text{м.м.п.}} = \arg\max_{\theta \in (0,1)} L(\overline{x}; \theta) = \arg\max_{\theta \in \Theta} L(\overline{x}; \theta), \text{ где } \Theta = \{x_{(i)} : \frac{i-1}{n} < x_{(i)} < \frac{i}{n}, i = \overline{1,n}\}$$

Приведем значения полученной оценки для сгенерированных выборок. Полученные значения находятся здесь.

### 3.2.2. Поиск оптимальных оценок

Функция правдоподобия:

$$L(\overline{x}; \theta) = \prod_{i=1}^{n} f(x_i) = 2^n \prod_{i=1}^{n} \left( \frac{x_i}{\theta} I(0 \le x_i \le \theta) + \frac{1-x_i}{1-\theta} I(\theta < x_i \le 1) \right)$$

Возьмем $g(T(\overline{x}); \theta) = L(\overline{x}; \theta)$ и $h(\overline{x}) = 1 \Rightarrow T(X) = (X_{(1)}, X_{(n)})$.

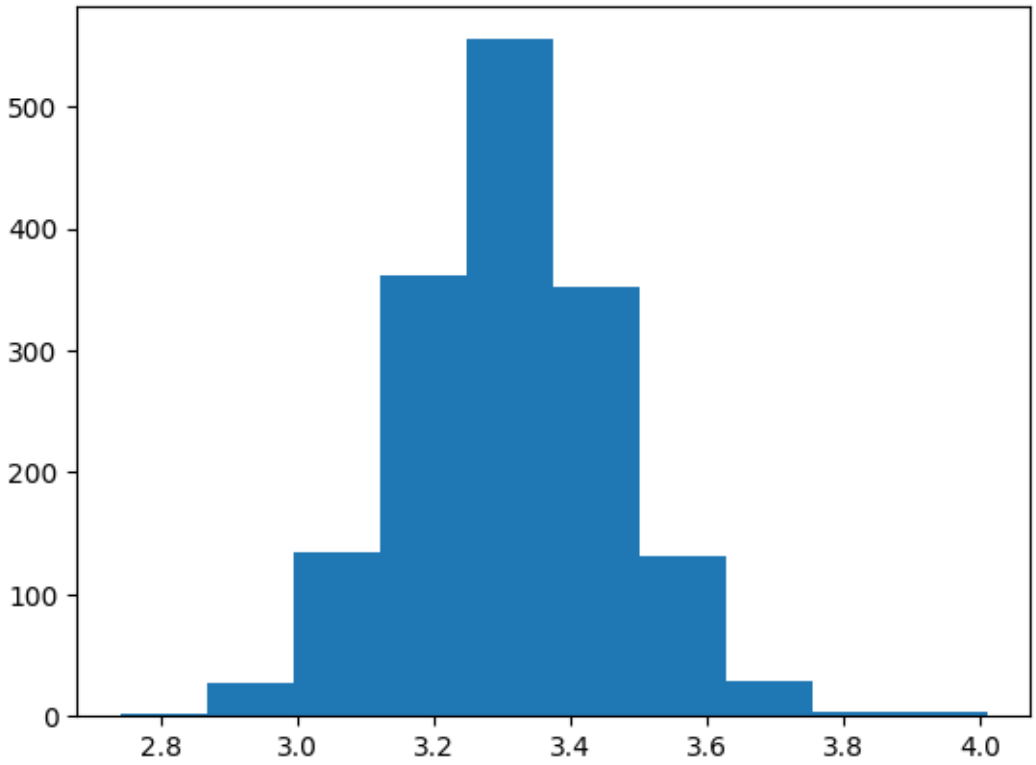
По критерию факторизации статистика $T(X)$ - достаточная. Заметим, что мы имеем дело с многомерной достаточной статистикой для скалярного параметра $\theta$, которая является неполной. Значит, не можем применить теорему о том, что произвольная функция от полной достаточной статистики является оптимальной оценкой своего математического ожидания.

### 3.2.3. Работа с данными

Как мы знаем, треугольное распределение обычно используется для описания совокупности с ограниченными выборочными данными и особенно, когда данных мало (возможно, из-за высокой стоимости сбора). В данных, приведенных далее, около 1600 объектов.

Выбранный мною датасет "Red Wine Quality" описывает варианты красного португальского вина "Винью Верде".

Рассмотрим столбец pH, который описывает, насколько кислым является вино по шкале от 0 (очень кислое) до 14 (базовое), причем большинство вин находятся в диапазоне от 3-4 по этой шкале.




▷ *Значение выборочного среднего и выборочной дисперсии*

$\overline{X} = 3.31111$; $\overline{S}^2 = 0.02382$

Найдем математического ожидание случайной величины с треугольным распределением:

Плотность распределения: $f_\xi(x) = \begin{cases} \frac{2(x-a)}{(b-a)(\theta-a)}, & \text{если } x \in [a, \theta] \\ \frac{2(b-x)}{(b-a)(b-\theta)}, & \text{если } x \in (\theta, b]; \\ 0, & \text{иначе} \end{cases}$

$$M\xi = \int\limits_{\mathbb{R}} x f_\xi(x)dx = \int\limits_a^\theta x\frac{2(x-a)}{(b-a)(\theta-a)}dx + \int\limits_\theta^b x\frac{2(b-x)}{(b-a)(b-\theta)}dx =$$

$$= \frac{2}{b-a}\left(\frac{1}{\theta-a}\int\limits_a^\theta (x^2-ax)dx + \frac{1}{b-\theta}\int\limits_\theta^b (bx-x^2)dx\right) =$$

$$= \frac{2}{b-a}\left(\frac{1}{\theta-a}(\frac{x^3}{3}-\frac{ax^2}{2})|_a^\theta + \frac{1}{b-\theta}(\frac{bx^2}{2}-\frac{x^3}{3})|_\theta^b)\right) =$$

$$= \frac{2}{b-a}\left(\frac{1}{\theta-a}(\frac{\theta^3}{3}-\frac{a\theta^2}{2}-\frac{a^3}{3}+\frac{a^3}{2}) + \frac{1}{b-\theta}(\frac{b^3}{2}-\frac{b^3}{3}-\frac{b\theta^2}{2}+\frac{\theta^3}{3})\right) =$$

$$= \frac{1}{3(b-a)}\left(\frac{2\theta^3-3a\theta^2+a^3}{\theta-a} + \frac{b^3-3b\theta^2+2\theta^3}{b-\theta}\right) =$$

$$= \frac{1}{3(b-a)}\left((2\theta^2-a\theta-a^2)-(2\theta^2-b\theta-b^2)\right) = \frac{-a\theta-a^2+b\theta+b^2}{3(b-a)} =$$

$$= \frac{a+b+\theta}{3}$$

▷ *Значения оценок*

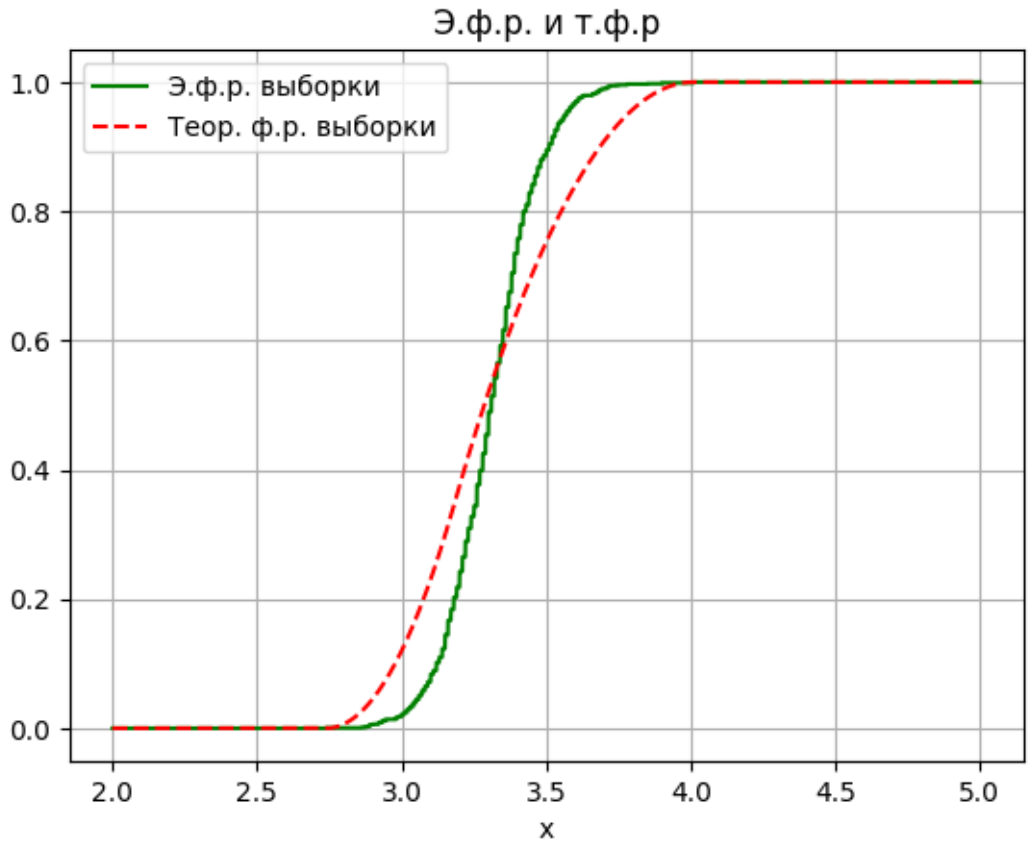
Заметим, что $\alpha_1 = M\xi = \frac{a+b+\theta}{3} < \infty$, а также что $\theta \in (0,1)$ - одномерный неизвестный параметр. Тогда рассмотрим систему с одним неизвестным $\theta$:

$$\left\{\hat{\alpha_1} = \alpha_1 = \frac{a+b+\theta}{3}\right.$$

Тогда оценка, построенная по методу моментов: $\hat{\theta} = 3\hat{\alpha_1} - a - b$.

Для наших данных: $a = 2.74; b = 4.01; \hat{\alpha_1} = 3.31111 \Rightarrow \hat{\theta} = 3.18334$.

Посмотрим на графики эмпирической и теоретической функции распределения:




Таким образом, графики эмпирической и теоретической функции распределения близки, но не совпадают. Значение оптимальной оценки параметра по вышеупомянутым причинам проверить не можем.

# Домашнее задание 4.

---

# Проверка статистических гипотез

### 1. Дискретное распределение

Программу, реализующую применение критериев, можно посмотреть здесь.

### 4.1.1. Проверка гипотезы о виде распределения

### Критерий согласия Колмогорова (Смирнова)

$X = (X_1, ..., X_n)$ - выборка из распределения $\mathcal{L}(\xi)$ с неизвестной функцией распределения $F_\xi(x)$. Простая гипотеза $H_0$: $F_\xi(x) = F(x)$, где $F(x)$ - функция распределения логарифмической случайной величины с параметром $\theta = \frac{1}{13}$.

Заметим, что $F(x)$ имеет точки разрыва $\Rightarrow$ вместо теоремы Колмогорова будем использовать следующее утверждение:

Пусть выборка $Y = (Y_1, ..., Y_n)$ из $\mathcal{L}(\mathcal{R}[0,1])$, выборка $X = (X_1, ..., X_n)$ из $\mathcal{L}(\xi)$: $F_\xi(x)$ имеет точки разрыва. Построим следующую случайную величину:

$$U_i = F(X_i-) + Y_i \cdot [F(X_i) - F(X_i-)],$$

где $F(X_i-) = \lim\limits_{z \to 0} F(X_i - z)$. Тогда с. в. $U_i \sim \mathcal{R}[0,1]$.

Благодаря данному утверждению получим случайные величины, распределенные равномерно на отрезке $[0,1]$, и будем применять критерий уже для непрерывного распределения. Заметим, что для $\mathcal{R}[0,1]$ $F(x) = x$, $x \in [0,1]$. Тогда статистика, соответствующая рассматриваемому критерию:
$D_n = \sup\limits_{0 \le u \le 1} |\hat{F}_n(u) - u|$. Необходимые рассуждения содержатся далее.

Используемые уровни значимости $\alpha$ и соответствующие им значения $\lambda_\alpha$:

| $\alpha$ | $\lambda_\alpha$ |
|---|---|
| 0.01 | 1.62762 |
| 0.05 | 1.3581 |
| 0.1 | 1.22385 |

Применим критерий к каждой сгенерированной выборки (✓ - принимаем гипотезу $H_0$, $\times$ - отвергаем её):

| $n$ | $i$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.1$ |
|---|---|---|---|---|
| 5 | 0 | ✓ | ✓ | ✓ |
| 5 | 1 | ✓ | ✓ | ✓ |
| 5 | 2 | ✓ | ✓ | ✓ |
| 5 | 3 | ✓ | ✓ | ✓ |
| 5 | 4 | ✓ | ✓ | ✓ |
| 10 | 0 | ✓ | ✓ | ✓ |
| 10 | 1 | ✓ | ✓ | ✓ |
| 10 | 2 | ✓ | ✓ | ✓ |
| 10 | 3 | ✓ | ✓ | ✓ |
| 10 | 4 | ✓ | ✓ | ✓ |
| 100 | 0 | ✓ | ✓ | ✓ |
| 100 | 1 | ✓ | ✓ | ✓ |
| 100 | 2 | ✓ | ✓ | ✓ |
| 100 | 3 | ✓ | ✓ | ✓ |
| 100 | 4 | ✓ | ✓ | × |
| 200 | 0 | ✓ | ✓ | ✓ |
| 200 | 1 | ✓ | ✓ | ✓ |
| 200 | 2 | ✓ | ✓ | ✓ |
| 200 | 3 | ✓ | ✓ | ✓ |
| 200 | 4 | ✓ | ✓ | ✓ |
| 400 | 0 | ✓ | ✓ | ✓ |
| 400 | 1 | ✓ | ✓ | ✓ |
| 400 | 2 | ✓ | ✓ | ✓ |
| 400 | 3 | ✓ | ✓ | ✓ |
| 400 | 4 | ✓ | ✓ | ✓ |
| 600 | 0 | ✓ | ✓ | ✓ |
| 600 | 1 | ✓ | ✓ | ✓ |
| 600 | 2 | ✓ | ✓ | ✓ |
| 600 | 3 | ✓ | ✓ | ✓ |
| 600 | 4 | ✓ | ✓ | ✓ |
| 800 | 0 | ✓ | ✓ | ✓ |
| 800 | 1 | ✓ | ✓ | ✓ |
| 800 | 2 | ✓ | ✓ | ✓ |
| 800 | 3 | ✓ | ✓ | ✓ |
| 800 | 4 | ✓ | ✓ | ✓ |
| 1000 | 0 | ✓ | ✓ | ✓ |
| 1000 | 1 | ✓ | ✓ | ✓ |
| 1000 | 2 | ✓ | ✓ | ✓ |
| 1000 | 3 | ✓ | ✓ | ✓ |
| 1000 | 4 | ✓ | ✓ | ✓ |

Таким образом, почти для каждой сгенерированной выборки при разных уровнях значимости принимается нулевая гипотеза.

**Критерий согласия хи-квадрат**

Пусть выборка $X = (X_1, ..., X_n)$ из распределения $\mathcal{L}(\xi)$, где $\xi$ - дискретная случайная величина, принимающая значения $1, ..., N$ с вероятностями $p_1, ..., p_N$ $(p_1 + ... + p_N = 1)$.

Введем случайную величину $\nu_k^{(n)} = \sum\limits_{i=1}^{n} Ind(X_i = k)$ - частота встречаемости значения $k$ ($k = \overline{1, N}$ и $\nu_1^{(n)} + ... + \nu_N^{(n)} = n$).

Определим случайный вектор частот: $\overline{\nu}^{(n)} = (\nu_1^{(n)}, ..., \nu_N^{(n)})$, который имеет полиномиальное распределение: $P(\overline{\nu}^{(n)} = (m_1, ..., m_N)) = \frac{n!}{m_1! \cdot ... \cdot m_N!} p_1^{m_1} \cdot ... \cdot p_N^{m_N}$.

Рассмотрим статистику: $X_N^2 = \sum\limits_{i=1}^{N} \frac{(\nu_i^{(n)} - np_i)^2}{np_i} = \sum\limits_{i=1}^{N} (\frac{(\nu_i^{(n)})^2}{np_i} - 2\nu_i^{(n)} + np_i) =$

$= \sum\limits_{i=1}^{N} \frac{(\nu_i^{(n)})^2}{np_i} - 2\sum\limits_{i=1}^{N} \nu_i^{(n)} + n\sum\limits_{i=1}^{N} p_i = \sum\limits_{i=1}^{N} \frac{(\nu_i^{(n)})^2}{np_i} - 2n + n = \sum\limits_{i=1}^{N} \frac{(\nu_i^{(n)})^2}{np_i} - n.$

$X_N^2$ - статистика Пирсона или статистика хи-квадрат.

Заметим, что относительная частота события $\{\xi = i\}$ $\frac{\nu_i^{(n)}}{n}$ является состоятельной оценкой вероятности $p_i$ этого события. Тогда если гипотеза $H_0$ справедлива, то при больших $n$ разности $|\frac{\nu_i^{(n)}}{n} - p_i|$ должны быть малы $\Rightarrow$ значение статистики $X_N^2$ не должно быть слишком большим.

Тогда критическая область определяется следующим образом:

$$\mathfrak{X}_{1,\alpha} = \{X_N^2 > t_\alpha\}, \text{ где}$$
критическая граница $t_\alpha$ при заданном уровне значимости $\alpha$
выбирается из условия $P(X_N^2 > t_\alpha | H_0) = \alpha$

Выбрать $t_\alpha$ поможет следующая теорема.

**Определение.** *Пусть случайные величины $\xi_1, ..., \xi_n$ независимы и $\xi_i \sim N(0, 1)$. Тогда с. в. $\chi_n^2 = \sum\limits_{i=1}^{n} \xi_i^2$ имеет распределение хи-квадрат с n степенями свободы.*

**Теорема.** *Пусть случайный вектор частот $\overline{\nu}^{(n)} = (\nu_1^{(n)}, ..., \nu_N^{(n)})$ имеет полиномиальное распределение с параметрами n и $(p_1, ..., p_N)$.*
*Тогда при $n \to \infty$ распределение $X_N^2$ (статистики Пирсона) при справедливости гипотезы $H_0$ сходится к распределению $\chi_{N-1}^2$ (хи-квадрат с N-1 степенью свободы).*

На практике критерий хи-квадрат можно использовать для расчетов с хорошим приближением при $n \geq 50$ и $\nu_j^{(n)} \geq 5$ $\forall j \in \overline{1, N}$.

Если эти условия выполнены, то:

$$P(X_N^2 > t_\alpha | H_0) \approx 1 - F_{N-1}(t_\alpha),$$
где $F_{N-1}(t)$ - функция распределения $\chi_{N-1}^2$

Следовательно, полагая: $1 - F_{N-1}(t_\alpha) = \alpha$, получаем $t_\alpha = \chi^2_{1-\alpha,N-1}$ - $(1-\alpha)$-квантиль с. в. $\chi^2_{N-1}$.

Таким образом, критерий согласия хи-квадрат:

$$H_0 \text{ отвергается} \iff X^2_N > \chi^2_{1-\alpha,N-1},$$
где $\alpha$ - заданный уровень значимости

С помощью метода группировки наблюдений перейдем к рассмотрению вышеописанной схемы. Т. к. в нашем случае логарифмическая случайная величина $\xi$ принимает натуральные значения, то разобьем ее область значений $\varepsilon$ на $N$ непересекающихся интервалов $\varepsilon_1, ..., \varepsilon_N$, т. е. $\varepsilon = \varepsilon_1 \sqcup ... \sqcup \varepsilon_N$. Теперь будем рассматривать $\nu_k^{(n)} = \sum\limits_{i=1}^{n} Ind(X_i \in \varepsilon_k)$ - число элементов выборки, попавших в интервал $\varepsilon_k$, $k = \overline{1, N}$ $(\nu_1^{(n)} + ... + \nu_N^{(n)} = n)$.

Заметим, что $P(1) = \frac{1}{13 \ln \frac{13}{12}} \approx 0.96 \Rightarrow$ при справедливости гипотезы $H_0$ значений, больших 1, будет гораздо меньше, чем самой 1. Вообще говоря на большее, чем 2 количество интервалов нет смысла разбивать, т. к. в этом случае уже в третьем интервале будет очень мало значений при рассматриваемах объемах выборок.

Будем разбивать $\varepsilon$ на $N = 2$ интервала: $\varepsilon_1 = \{1\}$, $\varepsilon_2 = \{2, 3, 4, ...\}$. Тогда $p_1 = P(1) = \frac{1}{13 \ln \frac{13}{12}}$, $p_2 = 1 - P(1) = 1 - \frac{1}{13 \ln \frac{13}{12}}$.

Также посмотрим на результаты при $N = 3$ ($\varepsilon_1 = \{1\}$, $\varepsilon_2 = \{2\}$, $\varepsilon_3 = \{3, 4, 5, ...\}$, вероятности считаются аналогично), ведь в случае со сложной гипотезой нам будет необходимо значение $N \geq 3$.

Посчитаем $\chi^2_{1-\alpha,N-1}$:

1. $N = 2$:
   - $\triangleright$ $\alpha = 0.01 \Rightarrow \chi^2_{0.99,1} = 6.635$
   - $\triangleright$ $\alpha = 0.05 \Rightarrow \chi^2_{0.95,1} = 3.841$
   - $\triangleright$ $\alpha = 0.1 \Rightarrow \chi^2_{0.9,1} = 2.706$
2. $N = 3$:
   - $\triangleright$ $\alpha = 0.01 \Rightarrow \chi^2_{0.99,2} = 9.21$
   - $\triangleright$ $\alpha = 0.05 \Rightarrow \chi^2_{0.95,2} = 5.991$
   - $\triangleright$ $\alpha = 0.1 \Rightarrow \chi^2_{0.9,2} = 4.605$

Применим критерий к каждой сгенерированной выборки (✓ - принимаем гипотезу $H_0$, $\times$ - отвергаем её):

1. $N = 2$

| $n$ | $i$ | $X_N^2$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.1$ |
|---|---|---|---|---|---|
| 5 | 0 | 0.203 | ✓ | ✓ | ✓ |
| 5 | 1 | 3.461 | ✓ | ✓ | × |
| 5 | 2 | 0.203 | ✓ | ✓ | ✓ |
| 5 | 3 | 3.461 | ✓ | ✓ | × |
| 5 | 4 | 0.203 | ✓ | ✓ | ✓ |
| 10 | 0 | 0.994 | ✓ | ✓ | ✓ |
| 10 | 1 | 0.994 | ✓ | ✓ | ✓ |
| 10 | 2 | 0.406 | ✓ | ✓ | ✓ |
| 10 | 3 | 0.406 | ✓ | ✓ | ✓ |
| 10 | 4 | 0.994 | ✓ | ✓ | ✓ |
| 100 | 0 | 0.003 | ✓ | ✓ | ✓ |
| 100 | 1 | 0.215 | ✓ | ✓ | ✓ |
| 100 | 2 | 1.18 | ✓ | ✓ | ✓ |
| 100 | 3 | 0.325 | ✓ | ✓ | ✓ |
| 100 | 4 | 2.57 | ✓ | ✓ | ✓ |
| 200 | 0 | 0.084 | ✓ | ✓ | ✓ |
| 200 | 1 | 1.371 | ✓ | ✓ | ✓ |
| 200 | 2 | 1.371 | ✓ | ✓ | ✓ |
| 200 | 3 | 2.36 | ✓ | ✓ | ✓ |
| 200 | 4 | 0.194 | ✓ | ✓ | ✓ |
| 400 | 0 | 0.388 | ✓ | ✓ | ✓ |
| 400 | 1 | 3.665 | ✓ | ✓ | × |
| 400 | 2 | 0.023 | ✓ | ✓ | ✓ |
| 400 | 3 | 2.086 | ✓ | ✓ | ✓ |
| 400 | 4 | 1.406 | ✓ | ✓ | ✓ |
| 600 | 0 | 1.403 | ✓ | ✓ | ✓ |
| 600 | 1 | 0.582 | ✓ | ✓ | ✓ |
| 600 | 2 | 2.427 | ✓ | ✓ | ✓ |
| 600 | 3 | 0.116 | ✓ | ✓ | ✓ |
| 600 | 4 | 0.017 | ✓ | ✓ | ✓ |
| 800 | 0 | 3.218 | ✓ | ✓ | × |
| 800 | 1 | 1.274 | ✓ | ✓ | ✓ |
| 800 | 2 | 0.001 | ✓ | ✓ | ✓ |
| 800 | 3 | 0.337 | ✓ | ✓ | ✓ |
| 800 | 4 | 1.131 | ✓ | ✓ | ✓ |
| 1000 | 0 | 3.245 | ✓ | ✓ | × |
| 1000 | 1 | 3.828 | ✓ | ✓ | × |
| 1000 | 2 | 0.674 | ✓ | ✓ | ✓ |
| 1000 | 3 | 0.11 | ✓ | ✓ | ✓ |
| 1000 | 4 | 2.656 | ✓ | ✓ | ✓ |

2. $N = 3$

| $n$ | $i$ | $X_N^2$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.1$ |
|---|---|---|---|---|---|
| 5 | 0 | 0.203 | ✓ | ✓ | ✓ |
| 5 | 1 | 3.741 | ✓ | ✓ | ✓ |
| 5 | 2 | 0.203 | ✓ | ✓ | ✓ |
| 5 | 3 | 3.741 | ✓ | ✓ | ✓ |
| 5 | 4 | 0.203 | ✓ | ✓ | ✓ |
| 10 | 0 | 1.134 | ✓ | ✓ | ✓ |
| 10 | 1 | 1.134 | ✓ | ✓ | ✓ |
| 10 | 2 | 0.406 | ✓ | ✓ | ✓ |
| 10 | 3 | 0.406 | ✓ | ✓ | ✓ |
| 10 | 4 | 1.134 | ✓ | ✓ | ✓ |
| 100 | 0 | 0.226 | ✓ | ✓ | ✓ |
| 100 | 1 | 0.341 | ✓ | ✓ | ✓ |
| 100 | 2 | 3.677 | ✓ | ✓ | ✓ |
| 100 | 3 | 3.209 | ✓ | ✓ | ✓ |
| 100 | 4 | 4.707 | ✓ | ✓ | × |
| 200 | 0 | 0.427 | ✓ | ✓ | ✓ |
| 200 | 1 | 6.745 | ✓ | × | × |
| 200 | 2 | 1.861 | ✓ | ✓ | ✓ |
| 200 | 3 | 2.74 | ✓ | ✓ | ✓ |
| 200 | 4 | 0.945 | ✓ | ✓ | ✓ |
| 400 | 0 | 1.89 | ✓ | ✓ | ✓ |
| 400 | 1 | 4.53 | ✓ | ✓ | ✓ |
| 400 | 2 | 0.09 | ✓ | ✓ | ✓ |
| 400 | 3 | 2.392 | ✓ | ✓ | ✓ |
| 400 | 4 | 1.651 | ✓ | ✓ | ✓ |
| 600 | 0 | 3.376 | ✓ | ✓ | ✓ |
| 600 | 1 | 0.902 | ✓ | ✓ | ✓ |
| 600 | 2 | 2.453 | ✓ | ✓ | ✓ |
| 600 | 3 | 0.556 | ✓ | ✓ | ✓ |
| 600 | 4 | 1.358 | ✓ | ✓ | ✓ |
| 800 | 0 | 5.542 | ✓ | ✓ | × |
| 800 | 1 | 1.604 | ✓ | ✓ | ✓ |
| 800 | 2 | 0.106 | ✓ | ✓ | ✓ |
| 800 | 3 | 1.706 | ✓ | ✓ | ✓ |
| 800 | 4 | 3.992 | ✓ | ✓ | ✓ |
| 1000 | 0 | 6.311 | ✓ | × | × |
| 1000 | 1 | 4.021 | ✓ | ✓ | ✓ |
| 1000 | 2 | 1.522 | ✓ | ✓ | ✓ |
| 1000 | 3 | 1.968 | ✓ | ✓ | ✓ |
| 1000 | 4 | 3.831 | ✓ | ✓ | ✓ |

Таким образом, и при $N = 2$, и при $N = 3$ почти для каждой сгенерированной выборки при разных уровнях значимости принимается нулевая гипотеза.

### Критерий согласия Колмогорова (Смирнова) для сложной гипотезы (в условиях когда неизвестен параметр распределения)

$X = (X_1, ..., X_n)$ - выборка из распределения $\mathcal{L}(\xi)$ с неизвестной функцией распределения $F_\xi(x)$. Сложная гипотеза $H_0$: $F_\xi(x) \in \mathcal{F}_0 = \{Log(\theta), \theta \in (0, 1)\}$.

В случае сложных гипотез распределение $D_n(\theta)$ зависит от вида априорных распределений, от способа получения оценок, размера выборки и от вида параметрического множества. Однако описанную ранее методику проверки гипотезы о виде распределения можно применить и в случае со сложной гипотезой следующим образом.

Как и в случае с простой гипотезой, будем работать с новой выборкой из $\mathcal{L}(\mathcal{R}[0, 1])$. Рассмотрим оценку неизвестного параметра методом максимального правдоподобия, полученную ранее. Для $\hat{\theta}_{\text{м.м.п.}}$ верно: $\frac{\hat{\theta}}{ln(1-\hat{\theta})(1-\hat{\theta})} = -\hat{\alpha_1}$.

Вычислим значения статистики, соответствующей рассматриваемому критерию: $\hat{D}_n = \underset{x \in \mathbb{R}}{sup} |\hat{F}_n(x) - F(x; \hat{\theta}_n)|$, где $\hat{\theta}_n$ - оценка максимального правдоподобия параметра $\theta$. При этом в случае достаточно большой выборки, можно разбить ее на две части: по одной получить оценки на неизвестные параметры, по второй проверить гипотезу о виде распределения.

Используемые уровни значимости $\alpha$ и соответствующие им значения $\lambda_\alpha$:

| $\alpha$ | $\lambda_\alpha$ |
|---|---|
| 0.01 | 1.62762 |
| 0.05 | 1.3581 |
| 0.1 | 1.22385 |

Применим критерий к каждой сгенерированной выборки (✓ - принимаем гипотезу $H_0$, × - отвергаем её):

| $n$ | $i$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.1$ |
|---|---|---|---|---|
| 5 | 0 | ✓ | ✓ | ✓ |
| 5 | 1 | ✓ | ✓ | ✓ |
| 5 | 2 | ✓ | ✓ | ✓ |
| 5 | 3 | ✓ | ✓ | ✓ |
| 5 | 4 | ✓ | ✓ | ✓ |
| 10 | 0 | ✓ | ✓ | ✓ |
| 10 | 1 | ✓ | ✓ | ✓ |
| 10 | 2 | ✓ | ✓ | ✓ |
| 10 | 3 | ✓ | ✓ | ✓ |
| 10 | 4 | ✓ | ✓ | ✓ |
| 100 | 0 | ✓ | ✓ | ✓ |
| 100 | 1 | ✓ | ✓ | ✓ |
| 100 | 2 | ✓ | ✓ | ✓ |
| 100 | 3 | ✓ | ✓ | ✓ |
| 100 | 4 | ✓ | ✓ | ✓ |
| 200 | 0 | ✓ | ✓ | ✓ |
| 200 | 1 | ✓ | ✓ | ✓ |
| 200 | 2 | ✓ | ✓ | × |
| 200 | 3 | ✓ | × | × |
| 200 | 4 | ✓ | ✓ | ✓ |
| 400 | 0 | ✓ | ✓ | ✓ |
| 400 | 1 | ✓ | ✓ | ✓ |
| 400 | 2 | ✓ | ✓ | ✓ |
| 400 | 3 | ✓ | ✓ | ✓ |
| 400 | 4 | ✓ | ✓ | ✓ |
| 600 | 0 | ✓ | ✓ | ✓ |
| 600 | 1 | ✓ | ✓ | ✓ |
| 600 | 2 | ✓ | ✓ | ✓ |
| 600 | 3 | ✓ | ✓ | ✓ |
| 600 | 4 | ✓ | ✓ | ✓ |
| 800 | 0 | ✓ | ✓ | ✓ |
| 800 | 1 | ✓ | ✓ | ✓ |
| 800 | 2 | ✓ | ✓ | ✓ |
| 800 | 3 | ✓ | ✓ | ✓ |
| 800 | 4 | ✓ | ✓ | ✓ |
| 1000 | 0 | ✓ | ✓ | ✓ |
| 1000 | 1 | ✓ | ✓ | × |
| 1000 | 2 | ✓ | ✓ | ✓ |
| 1000 | 3 | ✓ | ✓ | × |
| 1000 | 4 | ✓ | ✓ | ✓ |

Таким образом, почти для каждой сгенерированной выборки при разных уровнях значимости принимается нулевая гипотеза, но заметим, что по сравнению со случаем простой гипотезы, здесь чаще отвергается $H_0$.

**Критерий согласия хи-квадрат для сложной гипотезы (в условиях когда неизвестен параметр распределения)**

Сложная гипотеза для полиномиального распределения имеет следующий вид: $H_0 : \overline{p} = \overline{p}(\theta)$, где $\theta = (\theta_1, ..., \theta_r), \theta \in \Theta, r < N - 1$. При гипотезе $H_0$ вероятности исходов зависят от $\theta$. Соответственно, статистика Пирсона будет зависеть от $\theta$. Благодаря фиксации $\theta = \hat{\theta}_{\text{м.м.п.}}$, можем вычислить статистику Пирсона.

Рассмотрим оценку неизвестного параметра методом максимального правдоподобия, полученную ранее. Для $\hat{\theta}_{\text{м.м.п.}}$ верно: $\frac{\hat{\theta}}{ln(1-\hat{\theta})(1-\hat{\theta})} = -\hat{\alpha_1}$.

По рассмотренной в лекциях теореме:

$$\hat{X}_N^2 = \hat{X}_N^2(\hat{\theta}) = \sum_{i=1}^{N} \frac{(\nu_i^{(n)} - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})} = \sum_{i=1}^{N} \frac{(\nu_i^{(n)})^2}{np_i(\hat{\theta})} - n$$

имеет распределение $\chi^2_{N-1-r}$ (хи-квадрат с $N - 1 - r$ степенями свободы).
Тогда в случае со сложной гипотезой получим:

$$H_0 \text{ отвергается} \iff \hat{X}_N^2 > \chi^2_{1-\alpha,N-1-r},$$
где $\alpha$ - заданный уровень значимости

Остальная логика для применения критерия согласия хи-квадрат остается такой же, как и для случая с простой гипотезой.

Посчитаем $\chi^2_{1-\alpha,N-1-r}$ (в нашем случае $r = 1$, рассматриваемое $N = 3$):

▷ $\alpha = 0.01 \Rightarrow \chi^2_{0.99,1} = 6.635$

▷ $\alpha = 0.05 \Rightarrow \chi^2_{0.95,1} = 3.841$

▷ $\alpha = 0.1 \Rightarrow \chi^2_{0.9,1} = 2.706$

Применим критерий к каждой сгенерированной выборки (✓ - принимаем гипотезу $H_0$, × - отвергаем её):

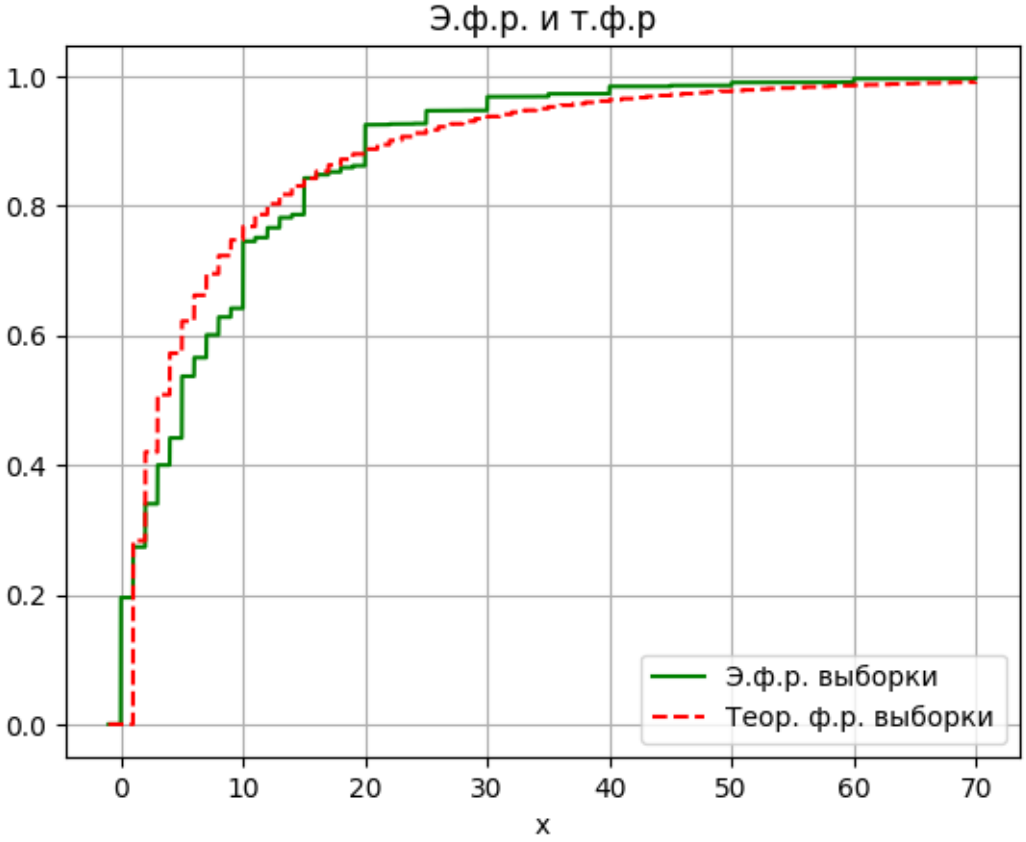
| $n$ | $i$ | $X_N^2$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.1$ |
|---|---|---|---|---|---|
| 5 | 0 | 0.0 | ✓ | ✓ | ✓ |
| 5 | 1 | 0.393 | ✓ | ✓ | ✓ |
| 5 | 2 | 0.0 | ✓ | ✓ | ✓ |
| 5 | 3 | 0.393 | ✓ | ✓ | ✓ |
| 5 | 4 | 0.0 | ✓ | ✓ | ✓ |
| 10 | 0 | 0.166 | ✓ | ✓ | ✓ |
| 10 | 1 | 0.166 | ✓ | ✓ | ✓ |
| 10 | 2 | 0.0 | ✓ | ✓ | ✓ |
| 10 | 3 | 0.0 | ✓ | ✓ | ✓ |
| 10 | 4 | 0.166 | ✓ | ✓ | ✓ |
| 100 | 0 | 0.234 | ✓ | ✓ | ✓ |
| 100 | 1 | 0.129 | ✓ | ✓ | ✓ |
| 100 | 2 | 0.519 | ✓ | ✓ | ✓ |
| 100 | 3 | 1.12 | ✓ | ✓ | ✓ |
| 100 | 4 | 0.187 | ✓ | ✓ | ✓ |
| 200 | 0 | 0.355 | ✓ | ✓ | ✓ |
| 200 | 1 | 1.554 | ✓ | ✓ | ✓ |
| 200 | 2 | 0.05 | ✓ | ✓ | ✓ |
| 200 | 3 | 0.003 | ✓ | ✓ | ✓ |
| 200 | 4 | 0.374 | ✓ | ✓ | ✓ |
| 400 | 0 | 0.748 | ✓ | ✓ | ✓ |
| 400 | 1 | 0.038 | ✓ | ✓ | ✓ |
| 400 | 2 | 0.085 | ✓ | ✓ | ✓ |
| 400 | 3 | 1.108 | ✓ | ✓ | ✓ |
| 400 | 4 | 0.815 | ✓ | ✓ | ✓ |
| 600 | 0 | 0.675 | ✓ | ✓ | ✓ |
| 600 | 1 | 0.095 | ✓ | ✓ | ✓ |
| 600 | 2 | 0.338 | ✓ | ✓ | ✓ |
| 600 | 3 | 0.28 | ✓ | ✓ | ✓ |
| 600 | 4 | 1.407 | ✓ | ✓ | ✓ |
| 800 | 0 | 0.522 | ✓ | ✓ | ✓ |
| 800 | 1 | 0.82 | ✓ | ✓ | ✓ |
| 800 | 2 | 0.12 | ✓ | ✓ | ✓ |
| 800 | 3 | 1.42 | ✓ | ✓ | ✓ |
| 800 | 4 | 1.25 | ✓ | ✓ | ✓ |
| 1000 | 0 | 0.836 | ✓ | ✓ | ✓ |
| 1000 | 1 | 1.002 | ✓ | ✓ | ✓ |
| 1000 | 2 | 1.002 | ✓ | ✓ | ✓ |
| 1000 | 3 | 1.329 | ✓ | ✓ | ✓ |
| 1000 | 4 | 1.202 | ✓ | ✓ | ✓ |

Таким образом, для каждой сгенерированной выборки при разных уровнях значимости принимается нулевая гипотеза.

### 4.1.2. Задание для данных, описываемых распределением

Напомню, что выбранный мною датасет "Game Recommendations on Steam" для 3-ей домашней работы описывает более 38 миллионов предварительно обработанных пользовательских рекомендаций, а также содержит таблицу игр с информацией о ценах в долларах США, дате выхода и т.д. Мы рассматривем столбец price_final, содержащий цены на игры после скидок.

Вспомним графики эмпирической и теоретической функции распределения:




Видим, что есть серьезные отклонения т.ф.р от э.ф.р.

### Критерий согласия Колмогорова (Смирнова) для сложной гипотезы (в условиях когда неизвестен параметр распределения)

Применим критерий к имеющимся данным (✓ - принимаем гипотезу $H_0$, × - отвергаем её):

| $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.1$ |
|---|---|---|
| × | × | × |

Таким образом, при разных уровнях значимости отвергается нулевая гипотеза. Видимо, имеющиеся данные не имеют логарифмическое распределение.

**Критерий согласия хи-квадрат для сложной гипотезы (в условиях когда неизвестен параметр распределения)**

Посчитаем $\chi^2_{1-\alpha,N-1-r}$ (в нашем случае $r = 1$):

1. $N = 3$:
   - ▷ $\alpha = 0.01 \Rightarrow \chi^2_{0.99,1} = 6.635$
   - ▷ $\alpha = 0.05 \Rightarrow \chi^2_{0.95,1} = 3.841$
   - ▷ $\alpha = 0.1 \Rightarrow \chi^2_{0.9,1} = 2.706$
2. $N = 16$ (эвристическая формула Старджесса для определения "оптимального"числа интервалов $N = 3.3 \lg n + 1$):
   - ▷ $\alpha = 0.01 \Rightarrow \chi^2_{0.99,14} = 29.141$
   - ▷ $\alpha = 0.05 \Rightarrow \chi^2_{0.95,14} = 23.685$
   - ▷ $\alpha = 0.1 \Rightarrow \chi^2_{0.9,14} = 21.064$

Применим критерий к имеющимся данным (✓ - принимаем гипотезу $H_0$, $\times$ - отвергаем её):

1. $N = 3$

| $X^2_N$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.1$ |
|---|---|---|---|
| 9043.094 | $\times$ | $\times$ | $\times$ |

2. $N = 16$

| $X^2_N$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.1$ |
|---|---|---|---|
| 68917.796 | $\times$ | $\times$ | $\times$ |

Таким образом, при разных уровнях значимости отвергается нулевая гипотеза. Вывод: имеющиеся данные не соответствуют логарифмическому распределению.

## 2. Непрерывное распределение

Программу, реализующую применение критериев, можно посмотреть здесь.

### 4.2.1. Проверка гипотезы о виде распределения

**Критерий согласия Колмогорова (Смирнова)**

$X = (X_1, ..., X_n)$ - выборка из распределения $\mathcal{L}(\xi)$ с неизвестной функцией распределения $F_\xi(x)$. Простая гипотеза $H_0$: $F_\xi(x) = F(x)$, где $F(x)$ - функция распределения треугольной случайной величины с параметром $\theta = 0.6$.

Статистика Колмогорова: $D_n = D_n(x) = \sup\limits_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)|$ - максимальное отклонение эмпирической ф.р. $\hat{F}_n(x)$ от гипотетической ф.р. $F(x)$. При каждом $x$ $\hat{F}_n(x)$ - оптимальная и состоятельная оценка для $F(x)$, т. е. с увеличением объема выборки $n$ $\hat{F}_n(x)$ сближается с $F(x)$. Значит, при больших n и при справедливости гипотезы $H_0$ значение $D_n$ не должно сильно отклоняться от 0.

Теорема Колмогорова (для непрерывных функций распределения $F(x)$ и для объема выборки $n \geq 20$):

$$P(\sqrt{n}D_n \geq \lambda_\alpha | H_0) = 1 - K(\lambda_\alpha) = \alpha$$

Критерий согласия Колмогорова: если $n \geq 20$, выбран уровень значимости $\alpha$, по которому определяется $\lambda_\alpha$: $K(\lambda_\alpha) = 1 - \alpha$, то

$$H_0 \text{ отвергается} \Longleftrightarrow \sqrt{n}D_n \geq \lambda_\alpha$$

Данному критерию соответствует критическая область:

$$\mathfrak{X}_{1,\alpha} = \{\overline{x} : \sqrt{n}D_n(\overline{x}) \geq \lambda_\alpha\}$$

При этом вместо статистики $S_n = \sqrt{n}D_n$ при малых объемах выборки ($n \leq 20$) используется статистика $S_n = \frac{6nD_n+1}{6\sqrt{n}}$, которая также сходится к распределению Колмогорова, но сходится к нему быстрее.

Т. к. вычисление супремума функции - нетривиальная задача, то будем рассматривать $D_n = max\{D_n^+, D_n^-\}$, где $D_n^+ = \underset{1\leq k\leq n}{max} |\frac{k}{n} - F(X_{(k)})|$,
а $D_n^- = \underset{1\leq k\leq n}{max} |F(X_{(k)}) - \frac{k-1}{n}|$.

Распределение статистики $D_n$ при гипотезе $H_0$ не зависит от вида функции $F(x)$. Тогда имея непрерывную функцию распределения $F(x)$ для треугольной случайной величины и используя таблицы значений функции Колмогорова $K(t)$, рассчитаем критерий для проверки гипотезы относительно $F(x)$.

Используемые уровни значимости $\alpha$ и соответствующие им значения $\lambda_\alpha$:

| $\alpha$ | $\lambda_\alpha$ |
|---|---|
| 0.01 | 1.62762 |
| 0.05 | 1.3581 |
| 0.1 | 1.22385 |

Применим критерий к каждой сгенерированной выборки (✓ - принимаем гипотезу $H_0$, $\times$ - отвергаем её):

| $n$ | $i$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.1$ |
|---|---|---|---|---|
| 5 | 0 | ✓ | ✓ | ✓ |
| 5 | 1 | ✓ | ✓ | ✓ |
| 5 | 2 | ✓ | ✓ | ✓ |
| 5 | 3 | ✓ | ✓ | ✓ |
| 5 | 4 | ✓ | ✓ | ✓ |
| 10 | 0 | ✓ | ✓ | ✓ |
| 10 | 1 | ✓ | ✓ | ✓ |
| 10 | 2 | ✓ | ✓ | ✓ |
| 10 | 3 | ✓ | ✓ | ✓ |
| 10 | 4 | ✓ | ✓ | ✓ |
| 100 | 0 | ✓ | ✓ | ✓ |
| 100 | 1 | ✓ | ✓ | ✓ |
| 100 | 2 | ✓ | ✓ | ✓ |
| 100 | 3 | ✓ | ✓ | ✓ |
| 100 | 4 | ✓ | ✓ | ✓ |
| 200 | 0 | ✓ | ✓ | ✓ |
| 200 | 1 | ✓ | ✓ | ✓ |
| 200 | 2 | ✓ | ✓ | ✓ |
| 200 | 3 | ✓ | ✓ | ✓ |
| 200 | 4 | ✓ | ✓ | $\times$ |
| 400 | 0 | ✓ | ✓ | ✓ |
| 400 | 1 | ✓ | ✓ | ✓ |
| 400 | 2 | ✓ | ✓ | ✓ |
| 400 | 3 | ✓ | ✓ | $\times$ |
| 400 | 4 | ✓ | ✓ | ✓ |
| 600 | 0 | ✓ | ✓ | ✓ |
| 600 | 1 | ✓ | ✓ | $\times$ |
| 600 | 2 | ✓ | ✓ | ✓ |
| 600 | 3 | ✓ | ✓ | ✓ |
| 600 | 4 | ✓ | ✓ | ✓ |
| 800 | 0 | ✓ | ✓ | ✓ |
| 800 | 1 | ✓ | $\times$ | $\times$ |
| 800 | 2 | ✓ | ✓ | ✓ |
| 800 | 3 | ✓ | ✓ | ✓ |
| 800 | 4 | ✓ | ✓ | ✓ |
| 1000 | 0 | ✓ | ✓ | $\times$ |
| 1000 | 1 | ✓ | ✓ | ✓ |
| 1000 | 2 | ✓ | ✓ | ✓ |
| 1000 | 3 | ✓ | ✓ | ✓ |
| 1000 | 4 | ✓ | $\times$ | $\times$ |

Таким образом, почти для каждой сгенерированной выборки при разных уровнях значимости принимается нулевая гипотеза.

**Критерий согласия хи-квадрат**

С помощью вышеописанного метода группировки наблюдений перейдем к рассмотрению дискретной схемы. Разобьем множество значений $\varepsilon$ случайной величины $\xi$ на равновероятные интервалы $\varepsilon_1, ..., \varepsilon_N$: $P(\xi \in \varepsilon_k) = \frac{1}{N}$.

Пусть $N = 3$. Тогда найдем точку $x_\gamma$: $P(\xi \leq x_\gamma) = \frac{1}{3} \Longleftrightarrow$ найдем квантиль уровня $\frac{1}{3}$ с. в. $\xi$. По выведенным ранее формулам $\frac{1}{3}$-квантиль $= 0.45$; $\frac{2}{3}$-квантиль $= 0.63 \Rightarrow \varepsilon_1 = [0, 0.45]$, $\varepsilon_2 = (0.45, 0.63]$, $\varepsilon_3 = (0.63, 1]$. Для $N = 5$ и для $N = 10$ интервалы определяются аналогичным образом.

Посчитаем $\chi^2_{1-\alpha,N-1}$:

1. $N = 3$:
   - $\triangleright$ $\alpha = 0.01 \Rightarrow \chi^2_{0.99,2} = 9.21$
   - $\triangleright$ $\alpha = 0.05 \Rightarrow \chi^2_{0.95,2} = 5.991$
   - $\triangleright$ $\alpha = 0.1 \Rightarrow \chi^2_{0.9,2} = 4.605$
2. $N = 5$:
   - $\triangleright$ $\alpha = 0.01 \Rightarrow \chi^2_{0.99,4} = 13.277$
   - $\triangleright$ $\alpha = 0.05 \Rightarrow \chi^2_{0.95,4} = 9.488$
   - $\triangleright$ $\alpha = 0.1 \Rightarrow \chi^2_{0.9,4} = 7.779$
3. $N = 10$:
   - $\triangleright$ $\alpha = 0.01 \Rightarrow \chi^2_{0.99,9} = 21.666$
   - $\triangleright$ $\alpha = 0.05 \Rightarrow \chi^2_{0.95,9} = 16.919$
   - $\triangleright$ $\alpha = 0.1 \Rightarrow \chi^2_{0.9,9} = 14.684$

Применим критерий к каждой сгенерированной выборки (✓ - принимаем гипотезу $H_0$, $\times$ - отвергаем её):

1. $N = 3$

| $n$ | $i$ | $X_N^2$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.1$ |
|---|---|---|---|---|---|
| 5 | 0 | 1.6 | ✓ | ✓ | ✓ |
| 5 | 1 | 1.6 | ✓ | ✓ | ✓ |
| 5 | 2 | 2.8 | ✓ | ✓ | ✓ |
| 5 | 3 | 0.4 | ✓ | ✓ | ✓ |
| 5 | 4 | 0.4 | ✓ | ✓ | ✓ |
| 10 | 0 | 0.8 | ✓ | ✓ | ✓ |
| 10 | 1 | 2.6 | ✓ | ✓ | ✓ |
| 10 | 2 | 1.4 | ✓ | ✓ | ✓ |
| 10 | 3 | 3.2 | ✓ | ✓ | ✓ |
| 10 | 4 | 0.2 | ✓ | ✓ | ✓ |
| 100 | 0 | 0.56 | ✓ | ✓ | ✓ |
| 100 | 1 | 1.22 | ✓ | ✓ | ✓ |
| 100 | 2 | 1.34 | ✓ | ✓ | ✓ |
| 100 | 3 | 0.14 | ✓ | ✓ | ✓ |
| 100 | 4 | 3.62 | ✓ | ✓ | ✓ |
| 200 | 0 | 1.39 | ✓ | ✓ | ✓ |
| 200 | 1 | 0.37 | ✓ | ✓ | ✓ |
| 200 | 2 | 2.17 | ✓ | ✓ | ✓ |
| 200 | 3 | 2.77 | ✓ | ✓ | ✓ |
| 200 | 4 | 4.12 | ✓ | ✓ | ✓ |
| 400 | 0 | 0.98 | ✓ | ✓ | ✓ |
| 400 | 1 | 3.26 | ✓ | ✓ | ✓ |
| 400 | 2 | 2.405 | ✓ | ✓ | ✓ |
| 400 | 3 | 1.295 | ✓ | ✓ | ✓ |
| 400 | 4 | 6.455 | ✓ | × | × |
| 600 | 0 | 2.59 | ✓ | ✓ | ✓ |
| 600 | 1 | 4.33 | ✓ | ✓ | ✓ |
| 600 | 2 | 0.28 | ✓ | ✓ | ✓ |
| 600 | 3 | 5.67 | ✓ | ✓ | × |
| 600 | 4 | 0.37 | ✓ | ✓ | ✓ |
| 800 | 0 | 3.88 | ✓ | ✓ | ✓ |
| 800 | 1 | 3.423 | ✓ | ✓ | ✓ |
| 800 | 2 | 5.448 | ✓ | ✓ | × |
| 800 | 3 | 0.498 | ✓ | ✓ | ✓ |
| 800 | 4 | 2.92 | ✓ | ✓ | ✓ |
| 1000 | 0 | 6.824 | ✓ | × | × |
| 1000 | 1 | 0.602 | ✓ | ✓ | ✓ |
| 1000 | 2 | 0.854 | ✓ | ✓ | ✓ |
| 1000 | 3 | 1.736 | ✓ | ✓ | ✓ |
| 1000 | 4 | 3.374 | ✓ | ✓ | ✓ |

2. $N = 5$

| $n$ | $i$ | $X_N^2$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.1$ |
|---|---|---|---|---|---|
| 5 | 0 | 2.0 | ✓ | ✓ | ✓ |
| 5 | 1 | 2.0 | ✓ | ✓ | ✓ |
| 5 | 2 | 2.0 | ✓ | ✓ | ✓ |
| 5 | 3 | 2.0 | ✓ | ✓ | ✓ |
| 5 | 4 | 4.0 | ✓ | ✓ | ✓ |
| 10 | 0 | 2.0 | ✓ | ✓ | ✓ |
| 10 | 1 | 1.0 | ✓ | ✓ | ✓ |
| 10 | 2 | 1.0 | ✓ | ✓ | ✓ |
| 10 | 3 | 4.0 | ✓ | ✓ | ✓ |
| 10 | 4 | 2.0 | ✓ | ✓ | ✓ |
| 100 | 0 | 1.9 | ✓ | ✓ | ✓ |
| 100 | 1 | 5.5 | ✓ | ✓ | ✓ |
| 100 | 2 | 2.8 | ✓ | ✓ | ✓ |
| 100 | 3 | 1.7 | ✓ | ✓ | ✓ |
| 100 | 4 | 2.2 | ✓ | ✓ | ✓ |
| 200 | 0 | 1.25 | ✓ | ✓ | ✓ |
| 200 | 1 | 1.05 | ✓ | ✓ | ✓ |
| 200 | 2 | 2.65 | ✓ | ✓ | ✓ |
| 200 | 3 | 4.1 | ✓ | ✓ | ✓ |
| 200 | 4 | 5.85 | ✓ | ✓ | ✓ |
| 400 | 0 | 0.8 | ✓ | ✓ | ✓ |
| 400 | 1 | 4.375 | ✓ | ✓ | ✓ |
| 400 | 2 | 3.925 | ✓ | ✓ | ✓ |
| 400 | 3 | 4.425 | ✓ | ✓ | ✓ |
| 400 | 4 | 4.125 | ✓ | ✓ | ✓ |
| 600 | 0 | 1.2 | ✓ | ✓ | ✓ |
| 600 | 1 | 7.433 | ✓ | ✓ | ✓ |
| 600 | 2 | 4.067 | ✓ | ✓ | ✓ |
| 600 | 3 | 3.467 | ✓ | ✓ | ✓ |
| 600 | 4 | 2.317 | ✓ | ✓ | ✓ |
| 800 | 0 | 3.163 | ✓ | ✓ | ✓ |
| 800 | 1 | 3.862 | ✓ | ✓ | ✓ |
| 800 | 2 | 2.888 | ✓ | ✓ | ✓ |
| 800 | 3 | 1.462 | ✓ | ✓ | ✓ |
| 800 | 4 | 6.438 | ✓ | ✓ | ✓ |
| 1000 | 0 | 6.69 | ✓ | ✓ | ✓ |
| 1000 | 1 | 0.88 | ✓ | ✓ | ✓ |
| 1000 | 2 | 2.44 | ✓ | ✓ | ✓ |
| 1000 | 3 | 4.64 | ✓ | ✓ | ✓ |
| 1000 | 4 | 12.13 | ✓ | × | × |

3. $N = 10$

| $n$ | $i$ | $X_N^2$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.1$ |
|---|---|---|---|---|---|
| 5 | 0 | 9.0 | ✓ | ✓ | ✓ |
| 5 | 1 | 9.0 | ✓ | ✓ | ✓ |
| 5 | 2 | 9.0 | ✓ | ✓ | ✓ |
| 5 | 3 | 5.0 | ✓ | ✓ | ✓ |
| 5 | 4 | 13.0 | ✓ | ✓ | ✓ |
| 10 | 0 | 8.0 | ✓ | ✓ | ✓ |
| 10 | 1 | 6.0 | ✓ | ✓ | ✓ |
| 10 | 2 | 10.0 | ✓ | ✓ | ✓ |
| 10 | 3 | 6.0 | ✓ | ✓ | ✓ |
| 10 | 4 | 6.0 | ✓ | ✓ | ✓ |
| 100 | 0 | 7.4 | ✓ | ✓ | ✓ |
| 100 | 1 | 8.4 | ✓ | ✓ | ✓ |
| 100 | 2 | 5.0 | ✓ | ✓ | ✓ |
| 100 | 3 | 4.2 | ✓ | ✓ | ✓ |
| 100 | 4 | 8.6 | ✓ | ✓ | ✓ |
| 200 | 0 | 6.5 | ✓ | ✓ | ✓ |
| 200 | 1 | 3.5 | ✓ | ✓ | ✓ |
| 200 | 2 | 8.6 | ✓ | ✓ | ✓ |
| 200 | 3 | 10.3 | ✓ | ✓ | ✓ |
| 200 | 4 | 9.5 | ✓ | ✓ | ✓ |
| 400 | 0 | 3.4 | ✓ | ✓ | ✓ |
| 400 | 1 | 12.05 | ✓ | ✓ | ✓ |
| 400 | 2 | 7.6 | ✓ | ✓ | ✓ |
| 400 | 3 | 12.95 | ✓ | ✓ | ✓ |
| 400 | 4 | 16.4 | ✓ | ✓ | × |
| 600 | 0 | 7.3 | ✓ | ✓ | ✓ |
| 600 | 1 | 14.067 | ✓ | ✓ | ✓ |
| 600 | 2 | 17.5 | ✓ | × | × |
| 600 | 3 | 11.1 | ✓ | ✓ | ✓ |
| 600 | 4 | 9.1 | ✓ | ✓ | ✓ |
| 800 | 0 | 10.7 | ✓ | ✓ | ✓ |
| 800 | 1 | 11.025 | ✓ | ✓ | ✓ |
| 800 | 2 | 12.675 | ✓ | ✓ | ✓ |
| 800 | 3 | 4.975 | ✓ | ✓ | ✓ |
| 800 | 4 | 8.025 | ✓ | ✓ | ✓ |
| 1000 | 0 | 14.86 | ✓ | ✓ | × |
| 1000 | 1 | 11.04 | ✓ | ✓ | ✓ |
| 1000 | 2 | 10.24 | ✓ | ✓ | ✓ |
| 1000 | 3 | 5.72 | ✓ | ✓ | ✓ |
| 1000 | 4 | 14.78 | ✓ | ✓ | × |

Таким образом, при $N = 5$ реже всего отвергается нулевая гипотеза - почти для каждой сгенерированной выборки принимается $H_0$.

### Критерий согласия Колмогорова (Смирнова) для сложной гипотезы (в условиях когда неизвестен параметр распределения)

$X = (X_1, ..., X_n)$ - выборка из распределения $\mathcal{L}(\xi)$ с неизвестной функцией распределения $F_\xi(x)$. Сложная гипотеза $H_0$: $F_\xi(x) \in \mathcal{F}_0 = \{Triangle(\theta), \theta \in (0, 1)\}$.

В случае сложных гипотез распределение $D_n(\theta)$ зависит от вида априорных распределений, от способа получения оценок, размера выборки и от вида параметрического множества. Однако описанную ранее методику проверки гипотезы о виде распределения можно применить и в случае со сложной гипотезой следующим образом.

Оценка неизвестного параметра методом максимального правдоподобия, полученная ранее:

$$\hat{\theta}_{\text{м.м.п.}} = \arg\max_{\theta \in (0,1)} L(\overline{x}; \theta) = \arg\max_{\theta \in \Theta} L(\overline{x}; \theta), \text{ где } \Theta = \{x_{(i)} : \frac{i-1}{n} < x_{(i)} < \frac{i}{n}, i = \overline{1, n}\}$$

Вычислим значения статистики, соответствующей рассматриваемому критерию: $\hat{D}_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x; \hat{\theta}_n)|$, где $\hat{\theta}_n$ - оценка максимального правдоподобия параметра $\theta$. При этом в случае достаточно большой выборки, можно разбить ее на две части: по одной получить оценки на неизвестные параметры, по второй проверить гипотезу о виде распределения.

Используемые уровни значимости $\alpha$ и соответствующие им значения $\lambda_\alpha$:

| $\alpha$ | $\lambda_\alpha$ |
|---|---|
| 0.01 | 1.62762 |
| 0.05 | 1.3581 |
| 0.1 | 1.22385 |

Применим критерий к каждой сгенерированной выборки (✓ - принимаем гипотезу $H_0$, × - отвергаем её):

| $n$ | $i$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.1$ |
|---|---|---|---|---|
| 5 | 0 | ✓ | ✓ | ✓ |
| 5 | 1 | ✓ | ✓ | ✓ |
| 5 | 2 | ✓ | ✓ | ✓ |
| 5 | 3 | ✓ | ✓ | ✓ |
| 5 | 4 | ✓ | ✓ | ✓ |
| 10 | 0 | ✓ | ✓ | ✓ |
| 10 | 1 | ✓ | ✓ | ✓ |
| 10 | 2 | ✓ | ✓ | ✓ |
| 10 | 3 | ✓ | ✓ | ✓ |
| 10 | 4 | ✓ | ✓ | ✓ |
| 100 | 0 | ✓ | ✓ | ✓ |
| 100 | 1 | ✓ | ✓ | ✓ |
| 100 | 2 | ✓ | ✓ | ✓ |
| 100 | 3 | ✓ | ✓ | ✓ |
| 100 | 4 | ✓ | ✓ | ✓ |
| 200 | 0 | ✓ | ✓ | ✓ |
| 200 | 1 | ✓ | ✓ | ✓ |
| 200 | 2 | ✓ | ✓ | × |
| 200 | 3 | ✓ | ✓ | ✓ |
| 200 | 4 | ✓ | ✓ | ✓ |
| 400 | 0 | ✓ | ✓ | ✓ |
| 400 | 1 | ✓ | ✓ | ✓ |
| 400 | 2 | ✓ | × | × |
| 400 | 3 | ✓ | ✓ | ✓ |
| 400 | 4 | ✓ | ✓ | ✓ |
| 600 | 0 | ✓ | ✓ | ✓ |
| 600 | 1 | ✓ | ✓ | ✓ |
| 600 | 2 | × | × | × |
| 600 | 3 | ✓ | × | × |
| 600 | 4 | ✓ | ✓ | ✓ |
| 800 | 0 | ✓ | ✓ | ✓ |
| 800 | 1 | ✓ | ✓ | ✓ |
| 800 | 2 | ✓ | × | × |
| 800 | 3 | ✓ | ✓ | ✓ |
| 800 | 4 | ✓ | ✓ | ✓ |
| 1000 | 0 | ✓ | ✓ | ✓ |
| 1000 | 1 | × | × | × |
| 1000 | 2 | ✓ | ✓ | ✓ |
| 1000 | 3 | ✓ | ✓ | × |
| 1000 | 4 | ✓ | ✓ | ✓ |

Таким образом, в большинстве случаев при разных уровнях значимости принимается нулевая гипотеза, но заметим, что по сравнению со случаем простой гипотезы, здесь чаще отвергается $H_0$.

**Критерий согласия хи-квадрат для сложной гипотезы (в условиях когда неизвестен параметр распределения)**

Сложная гипотеза для полиномиального распределения имеет следующий вид: $H_0 : \overline{p} = \overline{p}(\theta)$, где $\theta = (\theta_1, ..., \theta_r), \theta \in \Theta, r < N - 1$. При гипотезе $H_0$ вероятности исходов зависят от $\theta$. Соответственно, статистика Пирсона будет зависеть от $\theta$. Благодаря фиксации $\theta = \hat{\theta}_{\text{м.м.п.}}$, можем вычислить статистику Пирсона.

Рассмотрим оценку неизвестного параметра методом максимального правдоподобия, полученную ранее:

$$\hat{\theta}_{\text{м.м.п.}} = \arg\max_{\theta \in (0,1)} L(\overline{x}; \theta) = \arg\max_{\theta \in \Theta} L(\overline{x}; \theta), \text{ где } \Theta = \{x_{(i)} : \frac{i-1}{n} < x_{(i)} < \frac{i}{n}, i = \overline{1, n}\}$$

По рассмотренной в лекциях теореме:

$$\hat{X}^2_N = \hat{X}^2_N(\hat{\theta}) = \sum_{i=1}^{N} \frac{(\nu_i^{(n)} - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})} = \sum_{i=1}^{N} \frac{(\nu_i^{(n)})^2}{np_i(\hat{\theta})} - n$$

имеет распределение $\chi^2_{N-1-r}$ (хи-квадрат с $N - 1 - r$ степенями свободы).
Тогда в случае со сложной гипотезой получим:

$$H_0 \text{ отвергается} \iff \hat{X}^2_N > \chi^2_{1-\alpha, N-1-r},$$
где $\alpha$ - заданный уровень значимости

Остальная логика для применения критерия согласия хи-квадрат остается такой же, как и для случая с простой гипотезой (использование метода группировки наблюдений и квантилей треугольной случайной величины с $\theta = \hat{\theta}_{\text{м.м.п.}}$ для определения границ интервалов).

Посчитаем $\chi^2_{1-\alpha, N-1-r}$ (в нашем случае $r = 1$):

1. $N = 3$:
   - $\alpha = 0.01 \Rightarrow \chi^2_{0.99,1} = 6.635$
   - $\alpha = 0.05 \Rightarrow \chi^2_{0.95,1} = 3.841$
   - $\alpha = 0.1 \Rightarrow \chi^2_{0.9,1} = 2.706$
2. $N = 5$:
   - $\alpha = 0.01 \Rightarrow \chi^2_{0.99,3} = 11.345$
   - $\alpha = 0.05 \Rightarrow \chi^2_{0.95,3} = 7.815$

- $\alpha = 0.1 \Rightarrow \chi^2_{0.9,3} = 6.251$

3. $N = 10$:

- $\alpha = 0.01 \Rightarrow \chi^2_{0.99,8} = 20.09$
- $\alpha = 0.05 \Rightarrow \chi^2_{0.95,8} = 15.507$
- $\alpha = 0.1 \Rightarrow \chi^2_{0.9,8} = 13.362$

Применим критерий к каждой сгенерированной выборки (✓ - принимаем гипотезу $H_0$, $\times$ - отвергаем её):

1. $N = 3$

| $n$ | $i$ | $X_N^2$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.1$ |
|---|---|---|---|---|---|
| 5 | 0 | 0.4 | ✓ | ✓ | ✓ |
| 5 | 1 | 0.4 | ✓ | ✓ | ✓ |
| 5 | 2 | 1.6 | ✓ | ✓ | ✓ |
| 5 | 3 | 0.4 | ✓ | ✓ | ✓ |
| 5 | 4 | 0.4 | ✓ | ✓ | ✓ |
| 10 | 0 | 0.2 | ✓ | ✓ | ✓ |
| 10 | 1 | 0.2 | ✓ | ✓ | ✓ |
| 10 | 2 | 0.8 | ✓ | ✓ | ✓ |
| 10 | 3 | 0.8 | ✓ | ✓ | ✓ |
| 10 | 4 | 0.2 | ✓ | ✓ | ✓ |
| 100 | 0 | 0.86 | ✓ | ✓ | ✓ |
| 100 | 1 | 0.14 | ✓ | ✓ | ✓ |
| 100 | 2 | 0.26 | ✓ | ✓ | ✓ |
| 100 | 3 | 0.08 | ✓ | ✓ | ✓ |
| 100 | 4 | 0.26 | ✓ | ✓ | ✓ |
| 200 | 0 | 0.19 | ✓ | ✓ | ✓ |
| 200 | 1 | 0.28 | ✓ | ✓ | ✓ |
| 200 | 2 | 2.47 | ✓ | ✓ | ✓ |
| 200 | 3 | 1.33 | ✓ | ✓ | ✓ |
| 200 | 4 | 1.33 | ✓ | ✓ | ✓ |
| 400 | 0 | 0.02 | ✓ | ✓ | ✓ |
| 400 | 1 | 0.365 | ✓ | ✓ | ✓ |
| 400 | 2 | 0.035 | ✓ | ✓ | ✓ |
| 400 | 3 | 1.355 | ✓ | ✓ | ✓ |
| 400 | 4 | 2.765 | ✓ | ✓ | × |
| 600 | 0 | 0.37 | ✓ | ✓ | ✓ |
| 600 | 1 | 0.61 | ✓ | ✓ | ✓ |
| 600 | 2 | 1.17 | ✓ | ✓ | ✓ |
| 600 | 3 | 6.75 | × | × | × |
| 600 | 4 | 0.16 | ✓ | ✓ | ✓ |
| 800 | 0 | 0.227 | ✓ | ✓ | ✓ |
| 800 | 1 | 0.618 | ✓ | ✓ | ✓ |
| 800 | 2 | 4.27 | ✓ | × | × |
| 800 | 3 | 0.543 | ✓ | ✓ | ✓ |
| 800 | 4 | 3.918 | ✓ | × | × |
| 1000 | 0 | 0.854 | ✓ | ✓ | ✓ |
| 1000 | 1 | 0.026 | ✓ | ✓ | ✓ |
| 1000 | 2 | 1.766 | ✓ | ✓ | ✓ |
| 1000 | 3 | 0.728 | ✓ | ✓ | ✓ |
| 1000 | 4 | 3.338 | ✓ | ✓ | × |

2. $N = 5$

| $n$ | $i$ | $X_N^2$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.1$ |
|---|---|---|---|---|---|
| 5 | 0 | 4.0 | ✓ | ✓ | ✓ |
| 5 | 1 | 2.0 | ✓ | ✓ | ✓ |
| 5 | 2 | 2.0 | ✓ | ✓ | ✓ |
| 5 | 3 | 2.0 | ✓ | ✓ | ✓ |
| 5 | 4 | 2.0 | ✓ | ✓ | ✓ |
| 10 | 0 | 1.0 | ✓ | ✓ | ✓ |
| 10 | 1 | 1.0 | ✓ | ✓ | ✓ |
| 10 | 2 | 1.0 | ✓ | ✓ | ✓ |
| 10 | 3 | 1.0 | ✓ | ✓ | ✓ |
| 10 | 4 | 2.0 | ✓ | ✓ | ✓ |
| 100 | 0 | 1.7 | ✓ | ✓ | ✓ |
| 100 | 1 | 5.5 | ✓ | ✓ | ✓ |
| 100 | 2 | 2.3 | ✓ | ✓ | ✓ |
| 100 | 3 | 1.9 | ✓ | ✓ | ✓ |
| 100 | 4 | 0.5 | ✓ | ✓ | ✓ |
| 200 | 0 | 0.35 | ✓ | ✓ | ✓ |
| 200 | 1 | 0.7 | ✓ | ✓ | ✓ |
| 200 | 2 | 5.4 | ✓ | ✓ | ✓ |
| 200 | 3 | 3.25 | ✓ | ✓ | ✓ |
| 200 | 4 | 1.35 | ✓ | ✓ | ✓ |
| 400 | 0 | 0.375 | ✓ | ✓ | ✓ |
| 400 | 1 | 2.725 | ✓ | ✓ | ✓ |
| 400 | 2 | 0.8 | ✓ | ✓ | ✓ |
| 400 | 3 | 3.875 | ✓ | ✓ | ✓ |
| 400 | 4 | 3.725 | ✓ | ✓ | ✓ |
| 600 | 0 | 0.8 | ✓ | ✓ | ✓ |
| 600 | 1 | 1.117 | ✓ | ✓ | ✓ |
| 600 | 2 | 3.917 | ✓ | ✓ | ✓ |
| 600 | 3 | 3.583 | ✓ | ✓ | ✓ |
| 600 | 4 | 2.317 | ✓ | ✓ | ✓ |
| 800 | 0 | 1.025 | ✓ | ✓ | ✓ |
| 800 | 1 | 0.962 | ✓ | ✓ | ✓ |
| 800 | 2 | 2.962 | ✓ | ✓ | ✓ |
| 800 | 3 | 1.462 | ✓ | ✓ | ✓ |
| 800 | 4 | 6.538 | ✓ | ✓ | × |
| 1000 | 0 | 0.67 | ✓ | ✓ | ✓ |
| 1000 | 1 | 0.97 | ✓ | ✓ | ✓ |
| 1000 | 2 | 2.57 | ✓ | ✓ | ✓ |
| 1000 | 3 | 5.98 | ✓ | ✓ | ✓ |
| 1000 | 4 | 12.42 | × | × | × |

3. $N = 10$

| $n$ | $i$ | $X_N^2$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.1$ |
|---|---|---|---|---|---|
| 5 | 0 | 9.0 | ✓ | ✓ | ✓ |
| 5 | 1 | 5.0 | ✓ | ✓ | ✓ |
| 5 | 2 | 9.0 | ✓ | ✓ | ✓ |
| 5 | 3 | 5.0 | ✓ | ✓ | ✓ |
| 5 | 4 | 9.0 | ✓ | ✓ | ✓ |
| 10 | 0 | 6.0 | ✓ | ✓ | ✓ |
| 10 | 1 | 6.0 | ✓ | ✓ | ✓ |
| 10 | 2 | 8.0 | ✓ | ✓ | ✓ |
| 10 | 3 | 4.0 | ✓ | ✓ | ✓ |
| 10 | 4 | 6.0 | ✓ | ✓ | ✓ |
| 100 | 0 | 3.4 | ✓ | ✓ | ✓ |
| 100 | 1 | 7.8 | ✓ | ✓ | ✓ |
| 100 | 2 | 5.8 | ✓ | ✓ | ✓ |
| 100 | 3 | 5.0 | ✓ | ✓ | ✓ |
| 100 | 4 | 6.0 | ✓ | ✓ | ✓ |
| 200 | 0 | 4.2 | ✓ | ✓ | ✓ |
| 200 | 1 | 2.2 | ✓ | ✓ | ✓ |
| 200 | 2 | 11.0 | ✓ | ✓ | ✓ |
| 200 | 3 | 5.6 | ✓ | ✓ | ✓ |
| 200 | 4 | 3.0 | ✓ | ✓ | ✓ |
| 400 | 0 | 2.1 | ✓ | ✓ | ✓ |
| 400 | 1 | 7.75 | ✓ | ✓ | ✓ |
| 400 | 2 | 3.9 | ✓ | ✓ | ✓ |
| 400 | 3 | 10.8 | ✓ | ✓ | ✓ |
| 400 | 4 | 14.15 | ✓ | ✓ | × |
| 600 | 0 | 4.6 | ✓ | ✓ | ✓ |
| 600 | 1 | 3.2 | ✓ | ✓ | ✓ |
| 600 | 2 | 17.1 | ✓ | × | × |
| 600 | 3 | 12.267 | ✓ | ✓ | ✓ |
| 600 | 4 | 7.467 | ✓ | ✓ | ✓ |
| 800 | 0 | 6.225 | ✓ | ✓ | ✓ |
| 800 | 1 | 4.05 | ✓ | ✓ | ✓ |
| 800 | 2 | 12.125 | ✓ | ✓ | ✓ |
| 800 | 3 | 4.35 | ✓ | ✓ | ✓ |
| 800 | 4 | 7.625 | ✓ | ✓ | ✓ |
| 1000 | 0 | 6.46 | ✓ | ✓ | ✓ |
| 1000 | 1 | 12.58 | ✓ | ✓ | ✓ |
| 1000 | 2 | 11.9 | ✓ | ✓ | ✓ |
| 1000 | 3 | 6.28 | ✓ | ✓ | ✓ |
| 1000 | 4 | 15.9 | ✓ | × | × |

Таким образом, получили результат, как и в случае с простой гипотезой - при $N = 5$ реже всего отвергается нулевая гипотеза - почти для каждой сгенерированной выборки принимается $H_0$.

# Домашнее задание 5.

## Различение статистических гипотез

Рассмотрим случай двух простых гипотез $H_0$ и $H_1$. Имеется выборка $X = (X_1, ..., X_n)$ из $\mathcal{L}(\xi)$. Проверим гипотезу $H_0$: $\xi \sim Log(\theta_0)$ против альтернативы $H_1$: $\xi \sim Log(\theta_1)$.

Выбор из двух простых гипотез можно представить в виде параметрической гипотезы. Пусть $\Theta = \{0, 1\}$, $F_\theta(x) = (1-\theta)F_0(x) + \theta F_1(x)$,
$H_0$: $\theta = \theta_0 = 0$, $H_1$: $\theta = \theta_1 = 1$ (здесь $\theta_0$ и $\theta_1$ отличаются от введенных выше, далее будем работать с введенными изначально параметрами).

### 1. Теория

**Определение.** *Пусть $X$ - выборка из неизвестного распределения $F_X \in \mathcal{F}$, где $\mathcal{F}$ - заданное множество априори возможных распределений выборки $X$. Выделим некоторое подмножество $\mathcal{F}_0 \subset \mathcal{F}$, $\mathcal{F}_1 = \mathcal{F} \backslash \mathcal{F}_0$.*

- ▷ *Гипотеза $H_0$ - основная/нулевая гипотеза: $F_X \in \mathcal{F}_0$.*
- ▷ *Гипотеза $H_1$ - альтернативная гипотеза: $F_X \in \mathcal{F}_1$.*

**Определение.** *Если $H_0$ и $H_1$ - простые гипотезы, то*

- ▷ $P(\overline{x} \in \mathfrak{X}_1 | H_0) = P(\mathfrak{X}_1 | H_0) = \alpha$ *- ошибка 1-го рода*
- ▷ $P(\mathfrak{X}_0 | H_1) = \beta$ *- ошибка 2-го рода*

**Определение.** *Функция мощности критерия $W$ - функционал на множестве допустимых распределений $\mathcal{F}$ и выборки $X$: $W(F_X) = W(F_X; \mathfrak{X}_{1,\alpha}) = P(X \in \mathfrak{X}_{1,\alpha} | F_X)$ (вероятность попасть в $\mathfrak{X}_{1,\alpha}$, если $F_X$ - истинное распределение).*

*Вероятности ошибок 1-го и 2-го рода через функцию мощности*

- ▷ $\alpha = W(F_{0,X}) \triangleq P(\mathfrak{X}_{1,\alpha} | F_{0,X}) = P(\mathfrak{X}_{1,\alpha} | H_0)$
- ▷ $\beta = 1 - W(F_{1,X}) \triangleq 1 - P(\mathfrak{X}_{1,\alpha} | F_{1,X}) = 1 - P(\mathfrak{X}_{1,\alpha} | H_1) = P(\mathfrak{X}_{0,\alpha} | H_1)$

*В случае параметрических гипотез функцию мощности критерия можно переписать в следующем виде:*
$W(\theta) = W(\theta; \mathfrak{X}_{1,\alpha}) = P_\theta(X \in \mathfrak{X}_{1,\alpha}) \Longrightarrow \alpha = W(\theta_0; \mathfrak{X}_{1,\alpha})$, $\beta = 1 - W(\theta_1; \mathfrak{X}_{1,\alpha})$.

## 2. Описание критерия отношения правдоподобия

Функция отношения правдоподобия:

$$l(\overline{x}) \triangleq \frac{L(\overline{x};\theta_1)}{L(\overline{x};\theta_0)} = \frac{\prod\limits_{i=1}^{n} f_1(x_i)}{\prod\limits_{i=1}^{n} f_0(x_i)}$$

Критическая область критерия Неймана-Пирсона:
$\mathfrak{X}^*_{1,\alpha} = \{\overline{x} \in \mathfrak{X} : l(\overline{x}) \geq c_\alpha\}$, где $c_\alpha$: ошибка 1 рода равна $\alpha$.
Наиболее мощный критерий с уровнем значимости $\alpha$ - параметрический критерий, минимизирующий ошибку 2 рода при заданной ошибке 1 рода.
По Лемме Неймана-Пирсона: Критическая область $\mathfrak{X}^*_{1,\alpha}$ задает наиболее мощный критерий для гипотезы $H_0$ относительно альтернативы $H_1$ среди всех критериев с уровнем значимости $\alpha$.

## 3. Вычисление функции отношения правдоподобия

Найденная ранее функция правдоподобия:

$$L(\overline{x};\theta) = \frac{(-1)^n}{\ln^n(1-\theta)} \frac{\theta^{\sum\limits_{i=1}^{n} x_i}}{\prod\limits_{i=1}^{n} x_i}$$

Тогда функция отношения правдоподобия:

$$l(\overline{x}) = \frac{\ln^n(1-\theta_0)}{\ln^n(1-\theta_1)} \frac{{\theta_1}^{\sum x_i}}{{\theta_0}^{\sum x_i}} = \frac{\ln^n(1-\theta_0)}{\ln^n(1-\theta_1)} \Big(\frac{\theta_1}{\theta_0}\Big)^{\sum\limits_{i=1}^{n} x_i}$$

## 4. Вычисление критической области

Рассмотрим следующее неравенство:

$$l(\overline{x}) \geq c \Longleftrightarrow \Big(\frac{\theta_1}{\theta_0}\Big)^{\sum\limits_{i=1}^{n} x_i} \geq c \cdot \frac{\ln^n(1-\theta_1)}{\ln^n(1-\theta_0)} \Longleftrightarrow \sum_{i=1}^{n} x_i \cdot \ln\frac{\theta_1}{\theta_0} \geq \ln c + \ln\Big(\frac{\ln^n(1-\theta_1)}{\ln^n(1-\theta_0)}\Big)$$

$$\Longleftrightarrow \sum_{i=1}^{n} x_i \geq \frac{\ln c + \ln\Big(\frac{\ln^n(1-\theta_1)}{\ln^n(1-\theta_0)}\Big)}{\ln\frac{\theta_1}{\theta_0}}$$

Пусть правая часть неравенства равна $t(c) \Longrightarrow P_i(l(\overline{x}) \geq c) = P_i\Big(\sum\limits_{i=1}^{n} x_i \geq t(c)\Big)$.

Рассмотрим асимптотический подход к различению гипотез. Выборка $X = (X_1, ..., X_n)$ из $L(Log(\theta_i)) \Longrightarrow X_1, ..., X_n$ - независимые, одинаково распределенные случайные величины, $MX_1 = -\frac{1}{\ln(1-\theta_i)}\frac{\theta_i}{1-\theta_i} = \mu_i$,
$DX_1 = -\theta_i\frac{\ln(1-\theta_i)+\theta_i}{(1-\theta_i)^2\ln^2(1-\theta_i)} = \sigma_i^2 < \infty \Longrightarrow$ пользуясь ЦПТ (а именно - теоремой Леви), получим, что при достаточно большом значении $n$ ($n \to \infty$):
$\sum\limits_{i=1}^{n} X_i \sim N(n\mu_i, n\sigma_i^2)$

Для определенности будем считать, что $\theta_0 < \theta_1 \Rightarrow \mu_0 < \mu_1$.

Заметим, что если $\xi \sim N(\mu, \sigma^2)$, то $\eta = -\frac{\xi-\mu}{\sigma} \sim N(0,1)$, т. к.

$$f_\xi(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \Longrightarrow F_\eta(x) = P(\eta \le x) = P\Big(\frac{\xi-\mu}{\sigma} \ge -x\Big) = P(\xi \ge \mu - \sigma x) =$$
$$= 1 - F_\xi(\mu - \sigma x) \Longrightarrow f_\eta(x) = F'_\eta(x) = -f_\xi(\mu-\sigma x)(-\sigma) =$$
$$= \sigma\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(-\sigma x)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}}e^{-x^2/2} \Longrightarrow \eta \sim N(0,1)$$

В таком случае с. в. $-\frac{\sum\limits_{i=1}^{n} X_i - n\mu_i}{\sqrt{n}\sigma_i} \sim N(0,1)$. Тогда (обозначим $t_\alpha = t(c_\alpha)$):

$$\alpha = P_0(l(\overline{x}) \ge c_\alpha) = P_0\Big(\sum_{i=1}^{n} x_i \ge t_\alpha\Big) = P_0\Big(\frac{\sum x_i - n\mu_0}{\sqrt{n}\sigma_0} \ge \frac{t_\alpha - n\mu_0}{\sqrt{n}\sigma_0}\Big) =$$
$$= \Phi\Big(-\frac{t_\alpha - n\mu_0}{\sqrt{n}\sigma_0}\Big) = \Phi(-g_\alpha), \text{где } g_\alpha = g(t_\alpha) = g(t(c_\alpha)) = \frac{t_\alpha - n\mu_0}{\sqrt{n}\sigma_0},$$

а $\Phi$ - функция стандартного нормального распределения

Т. к. $\Phi(-g)$ - непрерывная функция, то всегда найдем такое $g_\alpha$. Таким образом, в данном случае критерий Неймана-Пирсона задается критической областью $\mathfrak{X}^*_{1,\alpha} = \{\overline{x} : \frac{\sum x_i - n\mu_0}{\sqrt{n}\sigma_0} \ge g_\alpha\}$, $\Phi(-g_\alpha) = \alpha$.

Из заданного нами $g_\alpha$ следует: $t_\alpha = n\mu_0 + \sqrt{n}\sigma_0 g_\alpha$

$$\beta = P_1(l(\overline{x}) < c_\alpha) = P_1\Big(\sum_{i=1}^{n} x_i < t_\alpha\Big) = P_1\Big(\frac{\sum x_i - n\mu_1}{\sqrt{n}\sigma_1} < \frac{t_\alpha - n\mu_1}{\sqrt{n}\sigma_1}\Big)$$

Правая часть неравенства: $\frac{t_\alpha - n\mu_1}{\sqrt{n}\sigma_1} = \frac{n\mu_0 + \sqrt{n}\sigma_0 g_\alpha - n\mu_1}{\sqrt{n}\sigma_1} = \frac{\mu_0-\mu_1}{\sigma_1}\sqrt{n} + g_\alpha\frac{\sigma_0}{\sigma_1} \Longrightarrow$

$$\beta = P_1\Big(\frac{\sum x_i - n\mu_1}{\sqrt{n}\sigma_1} < \frac{\mu_0-\mu_1}{\sigma_1}\sqrt{n} + g_\alpha\frac{\sigma_0}{\sigma_1}\Big) = \Phi\Big(\frac{\mu_0-\mu_1}{\sigma_1}\sqrt{n} + g_\alpha\frac{\sigma_0}{\sigma_1}\Big)$$

## 5. Вычисление минимального количества материала

Заранее заданы вероятности ошибок $\alpha$ и $\beta$. Определим минимальное число наблюдений $n^* = n^*(\alpha, \beta)$ наблюдений, необходимых для того, чтобы ошибоч-

ные заключения могли быть сделаны с вероятностями, не превосходящими $\alpha$ и $\beta$.

Заметим, что $\beta = \beta(a, n) \to 0$ при $n \to \infty \Rightarrow$ искомое $n^*$ - наименьшее из $n$, для которых $\beta(\alpha, n) \leq \beta$. Обозначим $\zeta_p$ - $p$-квантиль стандартного нормального распределения. Из полученных выражениями для ошибок: $\alpha = \Phi(-g_\alpha)$; $\beta = \Phi\left(\frac{\mu_0-\mu_1}{\sigma_1}\sqrt{n} + g_\alpha\frac{\sigma_0}{\sigma_1}\right)$, следует, что: $\zeta_\alpha = -g_\alpha$; $\zeta_\beta = \frac{\mu_0-\mu_1}{\sigma_1}\sqrt{n} + g_\alpha\frac{\sigma_0}{\sigma_1} \Longrightarrow \zeta_\beta - g_\alpha\frac{\sigma_0}{\sigma_1} = \frac{\mu_0-\mu_1}{\sigma_1}\sqrt{n} \Longrightarrow \sigma_1\zeta_\beta + \sigma_0\zeta_\alpha = (\mu_0-\mu_1)\sqrt{n} \Longrightarrow n = \frac{(\sigma_1\zeta_\beta+\sigma_0\zeta_\alpha)^2}{(\mu_0-\mu_1)^2}$. Так как $n$ должно быть целым числом, то округлим полученный результат вверх. Тогда число наблюдений в критерии Неймана-Пирсона:

$$n^* = \left\lceil \frac{(\sigma_1\zeta_\beta + \sigma_0\zeta_\alpha)^2}{(\mu_0 - \mu_1)^2} \right\rceil$$

Программную иллюстрацию полученных результатов можно посмотреть здесь.