

Правительство Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего профессионального образования
Национальный исследовательский университет
"Высшая школа экономики"
Московский институт электроники и математики Национального
исследовательского университета "Высшая школа экономики"
ОП "Компьютерная безопасность"

ДОМАШНЕЕ ЗАДАНИЕ ПО ДИСЦИПЛИНЕ
"МАТЕМАТИЧЕСКАЯ СТАТИСТИКА"
ВАРИАНТ №21

Работу выполнил:
Смирнов Тимофей Богданович,
студент группы СКБ221

Москва, 2024

Оглавление

I Характеристики вероятностных распределений	5
1 Блок №1. Описание основных характеристик распределения	6
1.1 Дискретное равномерное I	6
1.2 Распределение Лапласа	8
2 Блок №2. Поиск примеров событий, которые могут быть описаны выбранными случайными величинами	12
2.1 Дискретное равномерное I	12
2.2 Распределение Лапласа	14
3 Блок №3. Описание способа моделирования выбранных случайных величин	16
3.1 Дискретное равномерное I	16
3.2 Распределение Лапласа	18
II Основные понятия математической статистики	20
4 Блок №1. Генерация выборок выбранных случайных величин	21
4.1 Дискретное равномерное I	21
4.2 Распределение Лапласа	22
5 Блок №2. Построение эмпирической функции распределения	22
5.1 Дискретное равномерное I	23
5.2 Распределение Лапласа	26
6 Блок №3. Построение гистограммы и полигона частот	29
6.1 Дискретное равномерное I	29
6.2 Распределение Лапласа	33
7 Блок №4. Вычисление выборочных моментов	35
7.1 Дискретное равномерное I	35
7.2 Распределение Лапласа	38

III Построение точечных оценок параметра распределения

41

8 Блок №1. Получение оценок методом моментов и методом максимального правдоподобия	42
8.1 Дискретное равномерное I	42
8.2 Распределение Лапласа	44
9 Блок №2. Поиск оптимальных оценок	47
9.1 Дискретное равномерное I	47
9.2 Распределение Лапласа	49
10 Блок №3. Работа с данными	52
10.1 Дискретное равномерное I	52
10.2 Распределение Лапласа	54
IV Проверка статистических гипотез	56
11 Блок №1. Проверка гипотезы о виде распределения	57
11.1 Критерий согласия Колмогорова	57
11.1.1 Распределение Лапласа	57
11.1.2 Дискретное равномерное I	59
11.2 Критерий согласия хи-квадрат	61
11.2.1 Дискретное равномерное I	61
11.2.2 Распределение Лапласа	63
11.3 Критерий согласия Колмогорова для сложной гипотезы	67
11.3.1 Распределение Лапласа	67
11.3.2 Дискретное равномерное I	69
11.4 Критерий согласия хи-квадрат для сложной гипотезы	70
11.4.1 Дискретное равномерное I	70
11.4.2 Распределение Лапласа	72
12 Блок №2. Проверка гипотезы об однородности выборок	73
12.1 Распределение Лапласа	74

V Различение статистических гипотез	76
13 Блок №1. Вычисление функции отношения правдоподобия	77
13.1 Распределение Лапласа	79
13.2 Дискретное равномерное I	83
VI Ссылки	86

Часть I

Характеристики

вероятностных

распределений

1 Блок №1. Описание основных характеристик распределения

Для каждого из распределений необходимо выписать его основные характеристики:

1. функция распределения $F(x)$,
2. математическое ожидание,
3. дисперсия,
4. Квантиль уровня γ .

1.1 Дискретное равномерное I

Функция распределения случайной величины:

Дискретное равномерное распределение с параметром θ , $\xi \sim R\{1, \dots, \theta\}$. Равномерное распределение на множестве $\{1, \dots, \theta\}$ имеет следующий закон распределения

$$P(\xi = x) = \theta^{-1}, x \in \{1, \dots, \theta\}.$$

Вычислим функцию распределения:

$$F_\xi(x) = \sum_{k=1}^{\lfloor x \rfloor} \theta^{-1} = \frac{\lfloor x \rfloor}{\theta}, 1 \leq x \leq \theta.$$

Функция распределения равномерного закона:

$$F(x) = \begin{cases} 0, & x < 1; \\ \frac{\lfloor x \rfloor}{\theta}, & 1 \leq x \leq \theta; \\ 1, & x > \theta. \end{cases}$$

Математическое ожидание:

$$P(\xi = x) = \theta^{-1}, x \in \{1, \dots, \theta\}.$$

Вычислим математическое ожидание по определению:

$$\begin{aligned} M\xi &= \sum_{x=1}^{\theta} x * P(\xi = x) = \sum_{x=1}^{\theta} x * \theta^{-1} = \theta^{-1} \sum_{x=1}^{\theta} x = \\ &= \theta^{-1} \frac{2 * 1 + 1(\theta - 1)}{2} \theta = \frac{\theta + 1}{2}. \end{aligned}$$

Дисперсия:

$$M\xi = \frac{\theta + 1}{2}.$$

$$M\xi^2 = \sum_{x=1}^{\theta} x^2 * P(\xi = x) = \theta^{-1} * \sum_{x=1}^{\theta} x^2 = \theta^{-1} * \frac{\theta(\theta + 1)(2\theta + 1)}{6} = \frac{(\theta + 1)(2\theta + 1)}{6}.$$

Тогда найдем дисперсию случайной величины ξ :

$$\begin{aligned} D\xi &= M\xi^2 - (M\xi)^2 = \frac{(\theta + 1)(2\theta + 1)}{6} - \frac{(\theta + 1)^2}{4} = \\ &= \frac{8\theta^2 + 12\theta + 4 - 6\theta^2 - 12\theta - 6}{24} = \frac{2\theta^2 - 2}{24} = \frac{\theta^2 - 1}{12}. \end{aligned}$$

Вычисление квантиля:

По определению, вычислим квантиль уровня γ , решив неравенство.

$$F_{\xi}(x) \geq \gamma \iff \frac{x}{\theta} \geq \gamma \iff x \geq \gamma\theta, x \in \{1, \dots, \theta\}.$$

1.2 Распределение Лапласа

Функция распределения случайной величины:

Распределение Лапласа на множестве \mathbb{R} , $x, \mu, \theta \in \mathbb{R}$, $\theta > 0$. Распределение Лапласа на множестве \mathbb{R} задается плотностью распределения вида

$$f(x) = \frac{\theta}{2} * \exp\{-\theta|x - \mu|\};$$

$$f(x) = \frac{\theta}{2} * e^{-\theta|x-\mu|}.$$

Найдем выражение $F(x)$ на множестве \mathbb{R} :

$$F(x) = \int_{-\infty}^x f(t) dt = \int_{-\infty}^x \frac{\theta}{2} \exp\{-\theta|t - \mu|\} dt = \frac{\theta}{2} \int_{-\infty}^x e^{-\theta|t - \mu|} dt \iff$$

$$\iff \begin{cases} F(x) = \frac{\theta}{2} \int_{-\infty}^x \exp^{\theta(t - \mu)} dt \\ t < \mu \\ F(x) = \frac{\theta}{2} \int_{-\infty}^x \exp^{\theta(\mu - t)} dt \\ t \geq \mu \end{cases} \iff$$

$$\iff \begin{cases} F(x) = \frac{\theta}{2e^{\theta\mu}} \int_{-\infty}^x \exp^{\theta t} dt \\ t < \mu \\ F(x) = \frac{\theta}{2} \int_{-\infty}^{\mu} \exp^{\theta(\mu - t)} dt + \frac{\theta}{2} \int_{\mu}^x \exp^{\theta(\mu - t)} dt \\ t \geq \mu \end{cases} \iff$$

$$\iff \begin{cases} F(x) = \left. \frac{\theta}{2\theta e^{\theta\mu}} e^{\theta t} \right|_{-\infty}^x \\ t < \mu \\ F(x) = \lim_{a \rightarrow -\infty} \frac{\theta}{2} \int_a^{\mu} \exp^{\theta(\mu - t)} dt - \frac{e^{\theta\mu}}{2} \int_{\mu}^x \exp^{-\theta t} d(-\theta t) \\ t \geq \mu \end{cases} \iff$$

$$\iff \begin{cases} F(x) = \frac{1}{2e^{\theta\mu}}(e^{\theta x} - e^{\theta(-\infty)}) \\ x < \mu \\ F(x) = \lim_{a \rightarrow -\infty} \left(\frac{1}{2} - \frac{e^{\theta(a-\mu)}}{2} \right) - \frac{e^{\theta\mu}}{2} \left(e^{-\theta x} - e^{-\theta\mu} \right) \\ x \geq \mu \end{cases} \iff$$

$$\iff \begin{cases} F(x) = \frac{1}{2} \exp^{\theta(x-\mu)} \\ x < \mu \\ F(x) = \frac{1}{2} - \frac{e^{\theta(\mu-x)}}{2} + \frac{1}{2} \\ x \geq \mu \end{cases} \iff \begin{cases} F(x) = \frac{1}{2} \exp^{\theta(x-\mu)} \\ x < \mu \\ F(x) = 1 - \frac{1}{2} \exp^{\theta(\mu-x)} \\ x \geq \mu \end{cases}$$

Функция распределения распределения Лапласа:

$$F(x) = \begin{cases} \frac{1}{2} \exp^{\theta(x-\mu)}, & x < \mu; \\ 1 - \frac{1}{2} \exp^{\theta(\mu-x)}, & x \geq \mu. \end{cases}$$

Математическое ожидание:

$$f(x) = \frac{\theta}{2} * e^{-\theta|x-\mu|}.$$

Вычислим математическое ожидание:

$$\begin{aligned} M\xi &= \int_{\mathbb{R}} x * f_{\xi}(x) dx = \int_{-\infty}^{+\infty} x * \frac{\theta}{2} * \exp^{-\theta|x-\mu|} dx = \\ &= \frac{\theta}{2} * \int_{-\infty}^{\mu} x * \exp^{\theta(x-\mu)} dx + \frac{\theta}{2} * \int_{\mu}^{+\infty} x * \exp^{-\theta(x-\mu)} dx = \end{aligned}$$

$$= \begin{cases} u = x \rightarrow du = dx \\ dv = e^{\theta(x-\mu)} \rightarrow v = \frac{e^{\theta(x-\mu)}}{\theta} \\ k = x \rightarrow dk = dx \\ dm = e^{\theta(\mu-x)} \rightarrow m = -\frac{e^{\theta(\mu-x)}}{\theta} \end{cases} = \frac{x e^{\theta(x-\mu)}}{2} \Big|_{-\infty}^{\mu} - \frac{1}{2} \int_{-\infty}^{\mu} \exp^{\theta(x-\mu)} dx -$$

$$\begin{aligned}
-\frac{x e^{\theta(\mu-x)}}{2} \Big|_{\mu}^{+\infty} + \frac{1}{2} \int_{\mu}^{+\infty} \exp^{\theta(\mu-x)} dx &= \frac{\mu e^{\theta(\mu-\mu)}}{2} - 0 - \frac{1}{2\theta} \exp^{\theta(x-\mu)} \Big|_{-\infty}^{\mu} - 0 + \\
&+ \frac{\mu e^{\theta(\mu-\mu)}}{2} - \frac{1}{2\theta} \exp^{\theta(\mu-x)} \Big|_{\mu}^{+\infty} = \mu - \frac{1}{2\theta} + \frac{1}{2\theta} = \mu.
\end{aligned}$$

Дисперсия:

$$\begin{aligned}
M\xi &= \mu. \\
M\xi^2 &= \int_{\mathbb{R}} x^2 * f_{\xi}(x) dx = \int_{-\infty}^{+\infty} x^2 * \frac{\theta}{2} * \exp^{-\theta|x-\mu|} dx = \\
&= \frac{\theta}{2} * \int_{-\infty}^{\mu} x^2 * \exp^{\theta(x-\mu)} dx + \frac{\theta}{2} * \int_{\mu}^{+\infty} x^2 * \exp^{-\theta(x-\mu)} dx = \\
&= \begin{cases} u = x^2 \rightarrow du = 2xdx \\ dv = e^{\theta(x-\mu)} \rightarrow v = \frac{e^{\theta(x-\mu)}}{\theta} \\ k = x^2 \rightarrow dk = 2xdx \\ dm = e^{\theta(\mu-x)} \rightarrow m = -\frac{e^{\theta(\mu-x)}}{\theta} \end{cases} = \frac{x^2 e^{\theta(x-\mu)}}{2} \Big|_{-\infty}^{\mu} - \int_{-\infty}^{\mu} x \exp^{\theta(x-\mu)} dx - \\
&- \frac{x^2 e^{\theta(\mu-x)}}{2} \Big|_{\mu}^{+\infty} + \int_{\mu}^{+\infty} x \exp^{\theta(\mu-x)} dx = \begin{cases} u = x \rightarrow du = dx \\ dv = e^{\theta(x-\mu)} \rightarrow v = \frac{e^{\theta(x-\mu)}}{\theta} \\ k = x \rightarrow dk = dx \\ dm = e^{\theta(\mu-x)} \rightarrow m = -\frac{e^{\theta(\mu-x)}}{\theta} \end{cases} = \\
&= \frac{x^2 e^{\theta(x-\mu)}}{2} \Big|_{-\infty}^{\mu} - \frac{x e^{\theta(x-\mu)}}{2} \Big|_{-\infty}^{\mu} + \int_{-\infty}^{\mu} \theta^{-1} \exp^{\theta(x-\mu)} dx - \frac{x^2 e^{\theta(\mu-x)}}{2} \Big|_{\mu}^{+\infty} - \frac{x e^{\theta(\mu-x)}}{2} \Big|_{\mu}^{+\infty} + \\
&+ \int_{\mu}^{+\infty} \theta^{-1} \exp^{\theta(\mu-x)} dx = \frac{\mu^2}{2} - \frac{\mu}{2} + \frac{1}{\theta^2} e^{\theta(x-\mu)} \Big|_{-\infty}^{\mu} + \frac{\mu^2}{2} + \frac{\mu}{2} - \frac{1}{\theta^2} e^{\theta(\mu-x)} \Big|_{\mu}^{+\infty} = \\
&= \mu^2 + \frac{1}{\theta^2} + \frac{1}{\theta^2} = \mu^2 + \frac{2}{\theta^2}.
\end{aligned}$$

Тогда найдем дисперсию случайной величины ξ :

$$D\xi = M\xi^2 - (M\xi)^2 = \mu^2 + \frac{2}{\theta^2} - \mu^2 = \frac{2}{\theta^2}.$$

Вычисление квантиля:

$$F(x) = \begin{cases} \frac{1}{2} \exp^{\theta(x-\mu)}, & x < \mu; \\ 1 - \frac{1}{2} \exp^{\theta(\mu-x)}, & x \geq \mu. \end{cases}$$

По определению, вычислим квантиль уровня γ , решив уравнения:

$$F(x) = \frac{x-a}{\theta}, \quad x \in [a, a+\theta].$$

- $x < \mu$:

$$\begin{aligned} F_\xi(x) = \gamma &\iff \frac{1}{2} e^{\theta(x-\mu)} = \gamma \iff e^{x\theta} = 2\gamma e^{\theta\mu} \iff e^{\theta x} = e^{\ln 2\gamma e^{\theta\mu}} \iff \\ &\iff x = \frac{\ln 2 + \ln \gamma + \theta\mu \ln e}{\theta} \iff x = \frac{\ln 2\gamma}{\theta} + \mu; \end{aligned}$$

- $x \geq \mu$:

$$\begin{aligned} F_\xi(x) = \gamma &\iff 1 - \frac{1}{2} e^{\theta(\mu-x)} = \gamma \iff e^{-\theta x} = (2 - 2\gamma) e^{-\theta\mu} \iff \\ &\iff e^{-\theta x} = e^{\ln(2 - 2\gamma) e^{-\theta\mu}} \iff x = -\frac{\ln 2 + \ln(1 - \gamma) - \theta\mu}{\theta} \iff \\ &\iff x = -\frac{\ln 2 + \ln(1 - \gamma)}{\theta} + \mu. \end{aligned}$$

Итоговая формула квантиля уровня γ :

$$x = \begin{cases} \mu + \frac{\ln 2\gamma}{\theta}, & x < \mu; \\ \mu - \frac{\ln 2(1 - \gamma)}{\theta}, & x \geq \mu. \end{cases}$$

2 Блок №2. Поиск примеров событий, которые могут быть описаны выбранными случайными величинами

Для каждого из распределений необходимо:

1. привести пример интерпретации распределения – описания события, исходы в котором подчиняются выбранному распределению;
2. известные соотношения между распределениями.

2.1 Дискретное равномерное I

Дискретное равномерное распределение на множестве $\{1, \dots, \theta\}$ - это распределение случайной величины, которая принимает все целочисленные значения из отрезка $[1, \dots, \theta]$ с одинаковыми вероятностями, где $\theta \in R$.

Пример интерпретации распределения:

Рассмотрим **игральный кубик** с шестью гранями. На каждой грани есть точки, количество которых равно номеру грани, на которой они расположены.

- Случайная величина ξ равна числу точек на грани игрального кубика, выпавшего при однократном подбрасывании.
- Каждый исход выпадения любого числа точек из $\{1, \dots, 6\}$ имеет одинаковую вероятность, поскольку кубик является честным.

Формула равномерного распределения $P(\xi = x) = \theta^{-1}, x \in \{1, \dots, \theta\}$ описывает вероятность того, что выпавшее число будет равно x .

Пример: $\theta = 6$ (губик имеет шесть граней). Тогда вероятность выпадения числа $\{1, \dots, 6\}$ равна: $P(\xi = x) = \frac{1}{6}, x \in \{1, \dots, 6\}$.

Этот эксперимент демонстрирует свойства равномерного распределения, бросок кубика с θ гранями является примером события, исходы которого подчиняются дискретному равномерному распределению с параметром θ .

Соотношения, связывающие распределение с другими распределениями:

Геометрическое распределение: Если случайная величина X имеет дискретное равномерное распределение на множестве $\{1, 2, \dots, k\}$, то количество испытаний до первого успеха (где успех — это выпадение конкретного числа) имеет геометрическое распределение с параметром $p = \frac{1}{k}$.

Геометрическое распределение является частным случаем отрицательного биномиального распределения, когда число успехов равно 1. Отрицательное биномиальное распределение описывает количество неудач до m -го успеха в последовательности независимых испытаний Бернулли.

$$P(x) = C_x^{x+m-1} (1-\theta)^{x+1} \theta^m.$$

При $m=1$:

$$P(x) = \theta(1-\theta)^{x-1}.$$

Биномиальное распределение: Если независимые случайные величины X_1, X_2, \dots, X_n , каждая из которых имеет дискретное равномерное распределение на множестве $\{1, 2, \dots, k\}$, то сумма случайных величин $S = X_1 + X_2 + \dots + X_n$ имеет биномиальное распределение с параметрами n и $p = \frac{1}{k}$.

Непрерывное равномерное распределение Разница между дискретным и непрерывным равномерными распределениями заключается в том, что пространство выборки дискретного распределения счётно, тогда как непрерывное распределение неисчислимо.

Дискретное равномерное распределение имеет счётное пространство выборки. Например, если $\theta = 5$, то пространство $\{1, 2, \dots, 5\}$. Каждое значение из этого множества имеет одинаковую вероятность $P(\xi = x) = \theta^{-1}$.

Непрерывное равномерное распределение имеет неисчислимое пространство выборки. Например, если интервал $[a, b]$, то пространство выборки — это все действительные числа в этом интервале. Вероятность попадания в любой подинтервал пропорциональна длине этого подинтервала.

2.2 Распределение Лапласа

Пример интерпретации распределения:

Предположим, что метеорологическая служба прогнозирует температуру на завтра с некоторой средней ошибкой $\mu = 0^\circ\text{C}$ (то есть в среднем прогнозы оказываются точными). Однако возможны отклонения в обе стороны — как завышенные, так и заниженные, и вероятность этих отклонений убывает по мере увеличения их величины. Для моделирования ошибок прогноза используется распределение Лапласа с параметром $\theta = 0.5$.

Плотность вероятности ошибки прогноза температуры x задается формулой:

$$f(x) = \frac{0.5}{2} \exp^{-0.5|x|} = 0.25 * \exp^{-0.5|x|}$$

Вычислим вероятность ошибки прогноза от -2°C до 2°C :

$$\begin{aligned} P(-2 \leq x \leq 2) &= \int_{-2}^2 0.25 * \exp^{-0.5|x|} dx = \int_{-2}^0 0.25 * \exp^{0.5x} dx + \\ &+ \int_0^2 0.25 * \exp^{-0.5x} dx = 0.5 * (1 - 0.3679) + 0.5 * (1 - 0.3679) \approx 0.6321 \end{aligned}$$

Таким образом, вероятность того, что ошибка прогноза будет в пределах от -2°C до 2°C , составляет около 0.6.

Теперь найдем вероятность того, что ошибка прогноза будет больше 5°C :

$$P(|x| > 5) = P(x > 5) + P(x < -5).$$

Распределение симметрично относительно нуля. Найдем вероятность для положительной ошибки:

$$P(x > 5) = \int_5^\infty 0.25 * \exp^{-0.5x} dx \approx 0.5 * 0.0821 \approx 0.04105.$$

$$P(x < -5) = P(x > 5) \approx 0.04105.$$

$$P(|x| > 5) \approx 0.04105 + 0.04105 \approx 0.0821.$$

Таким образом, вероятность того, что ошибка прогноза будет больше 5°C составляет около 0.082.

Этот эксперимент демонстрирует, как распределения Лапласа может использоваться для моделирования ошибок в прогнозировании, где небольшие отклонения вероятны, а вероятность больших ошибок уменьшается экспоненциально.

Соотношения, связывающие распределение с другими распределениями:

Через экспоненциальное распределение: Экспоненциальное распределение можно рассматривать как частный случай распределения хи-квадрат, когда $n = 2$. Это также означает, что экспоненциальное распределение является частным случаем гамма-распределения. Следовательно, экспоненциально распределенная величина может быть интерпретирована как результат суммирования квадратов двух независимых случайных величин, каждая из которых подчиняется стандартному нормальному распределению. Дискретный вариант экспоненциального распределения — это геометрическое распределение.

Плотность вероятности экспоненциально распределения:

$$f(x) = \lambda * \exp^{-\lambda*x}; x \geq 0.$$

В случае экспоненциального распределения если функцию плотности вероятности отразить в область отрицательных значений, то получаем распределение Лапласа. Также в распределении Лапласа вводится значение сдвига относительно нуля для ОХ. Таким образом, распределение Лапласа определено на всем множестве действительных чисел. Экспоненциальное распределение является частным случаем гамма-распределения с параметром формы $k = 1$.

3 Блок №3. Описание способа моделирования выбранных случайных величин

Для каждого из двух распределений (дискретное и непрерывное) необходимо описать способ моделирования выборок с заданными распределениями.

Полагая, что у каждого есть источник непрерывных случайных величин, распределённых равномерно на отрезке $[0, 1]$ (random), необходимо описать и обосновать процедуру получения нужного распределения на основе равномерной выборки.

3.1 Дискретное равномерное I

Для моделирования выборки из дискретного равномерного распределения с параметром θ на основе равномерной выборки на отрезке, используем следующий метод обратной функции. Этот метод позволяет моделировать случайные величины с заданным распределением, используя равномерно распределенные случайные величины:

1. Генерируем случайную величину γ , равномерно распределенную на отрезке $[0, 1)$;
2. Получим дискретную случайную величину ξ .

Используя функцию распределения $F(x) = \frac{\lfloor x \rfloor}{\theta}$; $1 \leq x \leq \theta$; дискретного равномерного распределения, получим случайную величину ξ , принимающую значения из множества $\{1, 2, \dots, \theta\}$ с равной вероятностью:

$$\xi = \lfloor \theta * \gamma \rfloor.$$

При умножении γ на θ , получается равномерная случаная величина, распределенная равномерно на отрезке $[0, \theta)$. И берем целую часть, тем самым переходя во множество $\{0, 1, \dots, \theta - 1\}$. Прибавим единицу, получим последнее отображение : $\{0, 1, \dots, \theta - 1\} \rightarrow \{1, 2, \dots, \theta\}$

Код для моделирования случайной величины:

```
import random

def discrete_model(theta, size=1):
    if not isinstance(theta, int) or theta <= 0:
        raise ValueError("Parametr theta must be positive.")

    discrete_rez = []
    for _ in range(size):
        gamma = random.random()
        xi = int(theta * gamma)+1
        discrete_rez.append(xi)

    return discrete_rez

# Example of use:
theta = 250
sample_num = 100
sample = discrete_model(theta, sample_num)
print(f"Sampling from a discrete uniform distribution with the parameter
      theta = {theta}: {sample}")
```

Output: Выборка 100 элементов из дискретного равномерного распределения с параметром theta = 250:

```
[224, 221, 110, 70, 171, 37, 234, 60, 104, 113, 31, 41, 57, 145, 8, 243, 142,
 55, 79, 97, 75, 130, 159, 186, 203, 223, 33, 7, 231, 68, 101, 181, 118,
 106, 101, 17, 110, 46, 6, 162, 198, 144, 117, 97, 215, 47, 7, 82, 8, 148,
 83, 185, 221, 1, 86, 89, 124, 139, 250, 124, 194, 194, 116, 32, 184, 73,
 214, 210, 42, 70, 150, 114, 226, 28, 114, 9, 227, 237, 190, 114, 49, 87,
 79, 74, 185, 113, 32, 214, 62, 3, 32, 69, 174, 175, 49, 95, 10, 111, 157,
 94]
```

3.2 Распределение Лапласа

Для моделирования выборки из распределения Лапласа на основе равномерно распределенных случайных величин на отрезке $[0, 1]$, будем использовать метод обратных функций.

Описание метода:

1. Генерируем случайную величину γ , равномерно распределенную на отрезке $[0, 1)$;
2. Получим случайную величину ξ с распределением Лапласа.

Используя функцию распределения:

$$F(x) = \begin{cases} \frac{1}{2} \exp^{\theta(x-\mu)}, & x < \mu; \\ 1 - \frac{1}{2} \exp^{\theta(\mu-x)}, & x \geq \mu. \end{cases}$$

, получим случайную величину ξ , принимающую значения из множества \mathbb{R} .

$\xi = F^{-1}(\gamma)$, где $F^{-1}(x)$ - обратная функция распределения. Вычислим ее:

- $\gamma < \frac{1}{2}$:

$$\gamma = \frac{1}{2} \exp^{\theta(x-\mu)} \Rightarrow x = \mu + \frac{\ln 2\gamma}{\theta} \Rightarrow \xi = \mu + \frac{\ln 2\gamma}{\theta}.$$

- $\gamma \geq \frac{1}{2}$:

$$\gamma = 1 - \frac{1}{2} * \exp^{\theta(\mu-x)} \Rightarrow x = \mu - \frac{\ln 2(1-\gamma)}{\theta} \Rightarrow \xi = \mu - \frac{\ln 2(1-\gamma)}{\theta}.$$

Случайная величина γ равномерно распределена на отрезке $[0, 1]$; используя обратную функцию $F^{-1}(\gamma)$, получаем случайную величину ξ , которая имеет распределение Лапласа с параметрами μ и θ .

Код для моделирования случайной величины:

```
import numpy as np

def laplace_model(mu, theta, size=1):
    if theta <= 0:
        raise ValueError("Parametr theta must be positive.")

    gamma = np.random.uniform(low=0, high=1, size=size)

    contin_rez = np.where(
        gamma < 0.5,
        mu + np.log(2 * gamma) / theta, # gamma < 0.5
        mu - np.log(2 * (1 - gamma)) / theta # gamma >= 0.5
    )

    return contin_rez

# Example of use:
mu = 0
theta = 1
sample_num = 30
sample = laplace_model(mu, theta, sample_num)
print(f"Sampling from a continuous Laplace distribution with the parameters
      mu =={mu}, theta={theta}: {sample}")
```

Output: Выборка 30 элементов из непрерывного распределения Лапласа с параметрами mu = 0, theta = 1:

```
[-1.17462875 -0.33336655 0.27555276 0.08434746 0.99524051 0.00647854
 -1.99807001 1.35060294 -0.69527432 -0.28447502 1.12400407 0.81345426
 -0.01500971 1.17469174 1.01651511 0.76347621 -0.93606423 -0.46835001
 -0.60270768 1.2593465 1.68127781 -0.79250059 2.7311924 0.03764289
 -0.26498418 -0.01101862 0.07204468 -4.75664879 1.16907459 -0.08559789]
```

Часть II

Основные понятия математической статистики

4 Блок №1. Генерация выборок выбранных случайных величин

Для каждой из выбранных случайных величин необходимо построить по 5 выборок следующих объемов $n = \{5, 10, 100, 200, 400, 600, 800, 1000\}$.

4.1 Дискретное равномерное I

Одна из построенных выборок с параметром $\theta = 77$ для $n = \{5, 10, 100\}$:

$n = 5$
26
39
36
32
73

$n = 10$
55 22
63 50
51 2
46 77
19 7

$n = 100$									
59	32	37	4	6	43	52	3	55	33
39	3	42	3	45	11	52	59	5	49
33	29	55	14	10	18	45	27	36	73
58	12	2	44	74	23	22	5	53	12
32	76	48	68	43	68	54	39	51	64
14	29	2	64	9	24	74	63	72	39
68	71	20	21	73	13	19	16	31	62
42	66	53	40	52	70	7	48	62	57
60	75	24	28	45	13	33	1	38	16
36	73	12	45	69	32	40	8	59	64

4.2 Распределение Лапласа

Одна из построенных выборок с параметрами $\mu = 22$, $\theta = 7.5$ для $n = \{5, 10\}$:

$n = 5$	$n = 10$	
22.04173534	22.02687505	22.04052231
21.84622421	21.92460163	21.97771776
22.04683823	22.3132322	22.03721312
22.05901924	22.24316246	21.81945995
21.72051219	21.99027478	22.02491901

5 Блок №2. Построение эмпирической функции распределения

Для каждой сгенерированной выборки необходимо построить график эмпирической функции распределения

$$\mathcal{F}_n(t) = \frac{\sum_{i=1}^n I(x_i < t)}{n}.$$

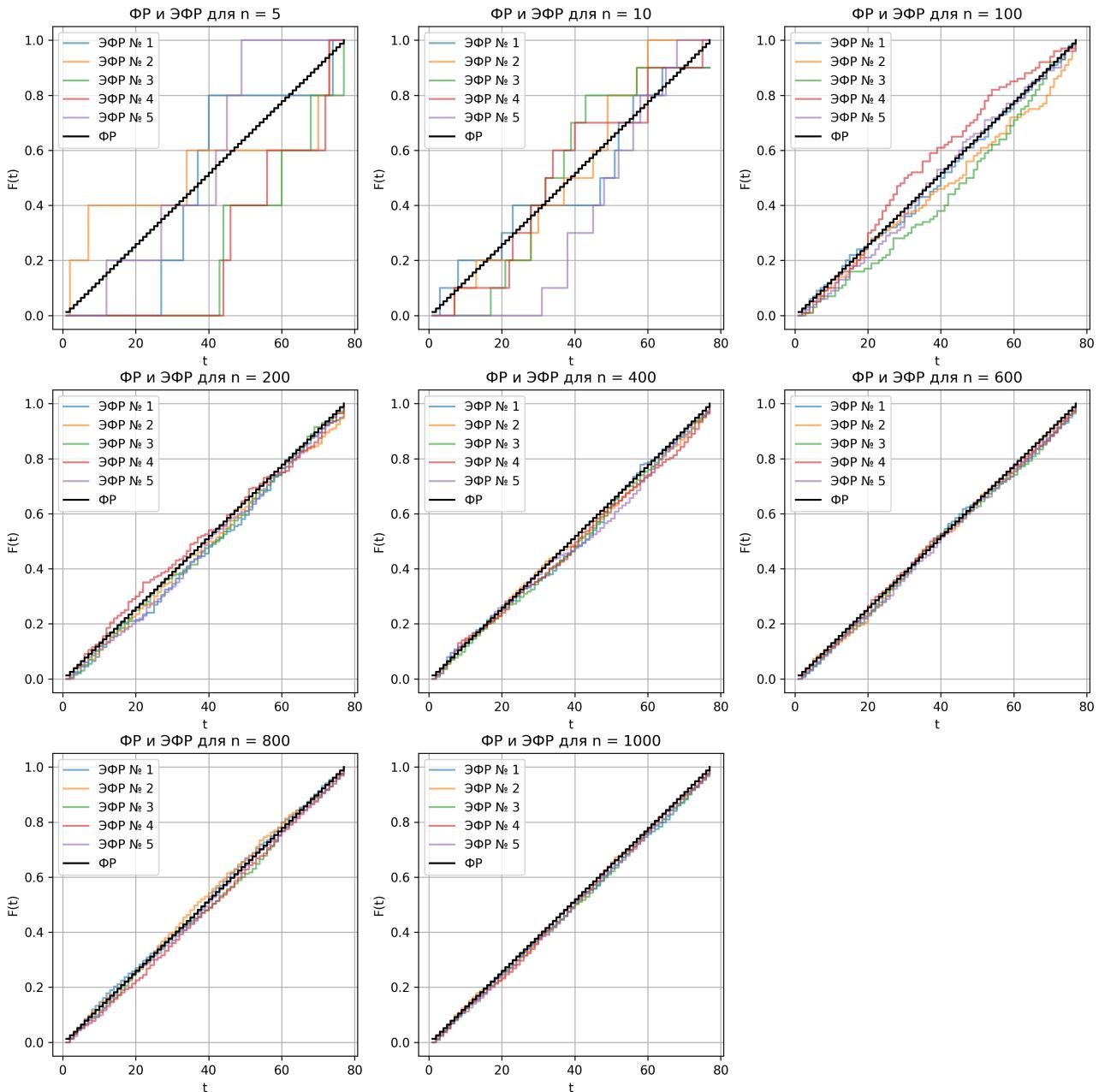
Графики необходимо привести в отчете. На одном графике необходимо отобразить эмпирические функции распределения для каждого из объемов выборки независимо и график функции распределения случайной величины.

Для каждой пары построенных эмпирических $\mathcal{F}_n(x)$ и $\mathcal{F}_m(x)$, $n, m \in \{5, 10, 100, 200, 400, 600, 800, 1000\}$ необходимо вычислить

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

Построим соответствующие графики и вычислим $D_{m,n}$. Исходный код находится по ссылке.

5.1 Дискретное равномерное I



Как видно из графиков, с ростом объема выборки эмпирическая функция распределения приближается к теоретической функции распределения.

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

Вычислим $D_{m,n}$ для каждой пары построенных ЭФР, используя Python. Отобрал изображение только часть полученных данных - возьмем по одной выборке для каждого n . Полные варианты для $D_{m,n}$ находятся по ссылке.

Итоговые значения представим в виде таблицы.

Функция для вычисления эмпирической функции распределения:

```
def F_empirical_cdf(smp, t):
    return np.count_nonzero(smp < t)/len(smp)
```

Функция для вычисления $D_{m,n}$:

```
def compute_Dmn(smp1, smp2):
    n = len(smp1)
    m = len(smp2)
    maxx = max([abs(F_empirical_cdf(smp1, t) - F_empirical_cdf(smp2, t)) for
                t in np.linspace(1, 77, 100)])
    D_mn = np.sqrt((n * m) / (n + m)) * maxx
    return D_mn
```

Создаем таблицу:

```
table = PrettyTable()
table.field_names = ["n/m"] + list(samples.keys())

for n_key, n_sample in samples.items():
    row = [n_key]
    for m_key, m_sample in samples.items():
        D_mn = compute_Dmn(n_sample, m_sample)
        row.append(f"{D_mn:.4f}")
    table.add_row(row)

table
```

Получаем таблицу значений $D_{m,n}$ для каждой пары построенных ЭФР по одной для $n \in \{5, 10, 1000\}$:

n/m	5	10	100	200	400	600	800	1000
5	0.0000	0.7303	0.6983	0.7068	0.7278	0.7348	0.7690	0.7316
10	0.7303	0.0000	0.6332	0.4938	0.5076	0.6795	0.6718	0.5318
100	0.6983	0.6332	0.0000	0.5715	0.4249	0.4783	0.3889	0.3909
200	0.7068	0.4938	0.5715	0.0000	0.6928	0.9390	1.0436	0.7617
400	0.7278	0.5076	0.4249	0.6928	0.0000	1.0586	1.0819	0.7860
600	0.7348	0.6795	0.4783	0.9390	1.0586	0.0000	0.8641	0.9231
800	0.7690	0.6718	0.3889	1.0436	1.0819	0.8641	0.0000	0.9434
1000	0.7316	0.5318	0.3909	0.7617	0.7860	0.9231	0.9434	0.0000

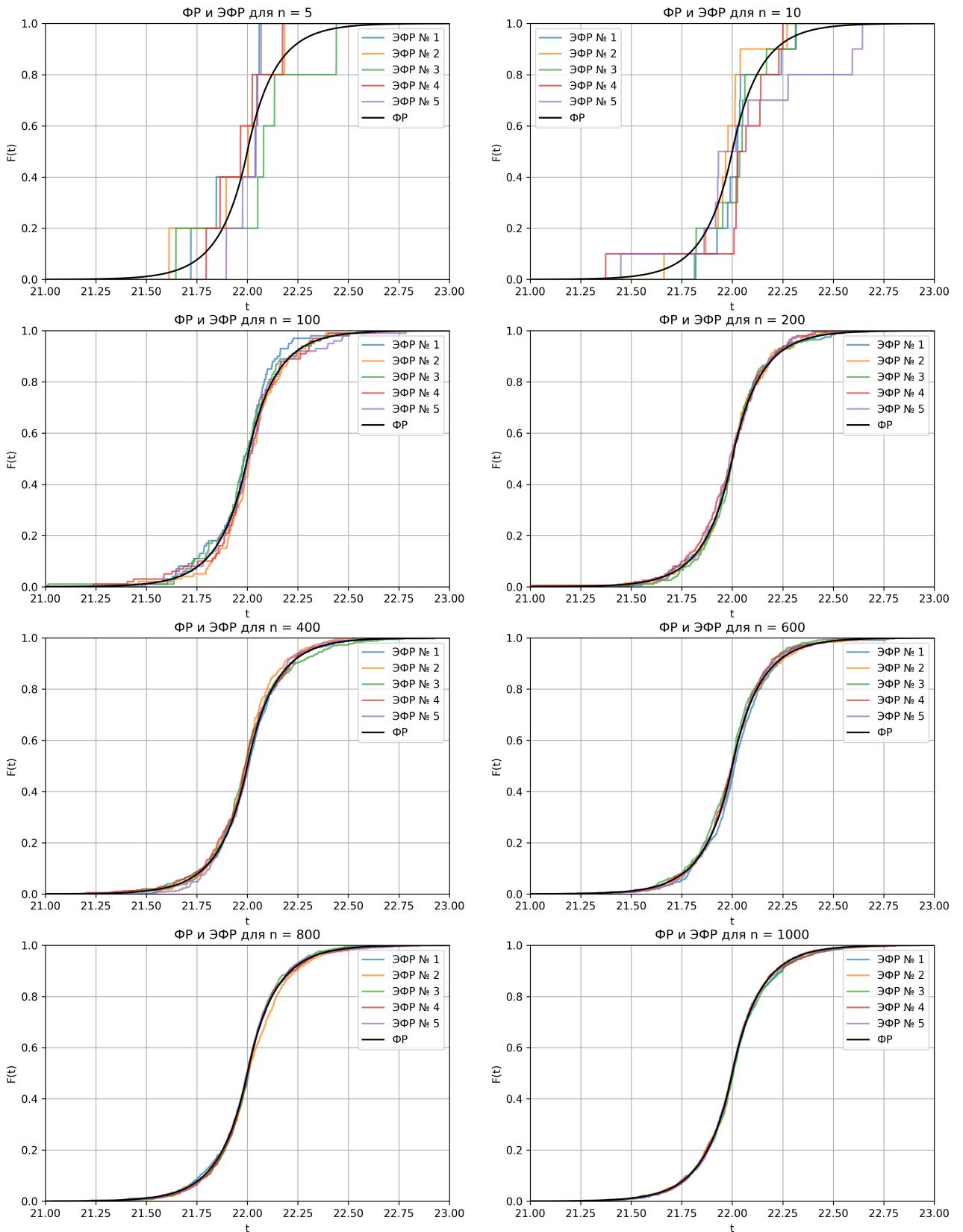
Значения $D_{n,n}$ при $n = 5$:

```

Compute D_(5, 5) for samples 5_1 and 5_1: 0.0
Compute D_(5, 5) for samples 5_1 and 5_2: 0.632455532033676
Compute D_(5, 5) for samples 5_1 and 5_3: 1.264911064067352
Compute D_(5, 5) for samples 5_1 and 5_4: 1.264911064067352
Compute D_(5, 5) for samples 5_1 and 5_5: 0.632455532033676
Compute D_(5, 5) for samples 5_2 and 5_1: 0.632455532033676
Compute D_(5, 5) for samples 5_2 and 5_2: 0.0
Compute D_(5, 5) for samples 5_2 and 5_3: 0.9486832980505138
Compute D_(5, 5) for samples 5_2 and 5_4: 0.9486832980505138
Compute D_(5, 5) for samples 5_2 and 5_5: 0.632455532033676
Compute D_(5, 5) for samples 5_3 and 5_1: 1.264911064067352
Compute D_(5, 5) for samples 5_3 and 5_2: 0.9486832980505138
Compute D_(5, 5) for samples 5_3 and 5_3: 0.0
Compute D_(5, 5) for samples 5_3 and 5_4: 0.31622776601683805
Compute D_(5, 5) for samples 5_3 and 5_5: 0.9486832980505138
Compute D_(5, 5) for samples 5_4 and 5_1: 1.264911064067352
Compute D_(5, 5) for samples 5_4 and 5_2: 0.9486832980505138
Compute D_(5, 5) for samples 5_4 and 5_3: 0.31622776601683805
Compute D_(5, 5) for samples 5_4 and 5_4: 0.0
Compute D_(5, 5) for samples 5_4 and 5_5: 0.948683298050514
Compute D_(5, 5) for samples 5_5 and 5_1: 0.632455532033676
Compute D_(5, 5) for samples 5_5 and 5_2: 0.632455532033676
Compute D_(5, 5) for samples 5_5 and 5_3: 0.9486832980505138
Compute D_(5, 5) for samples 5_5 and 5_4: 0.948683298050514
Compute D_(5, 5) for samples 5_5 and 5_5: 0.0

```

5.2 Распределение Лапласа



Используя те же функции для вычисления ЭФР и $D_{m,n}$, получаем таблицу значений $D_{m,n}$, для каждой пары построенных ЭФР (по одной выборке для каждого n). Полные варианты для $D_{m,n}$ находятся по ссылке.

Функция для вычисления $D_{m,n}$:

```
def compute_Dmn_c(smp1, smp2):
    n = len(smp1)
    m = len(smp2)
    maxx = max([abs(F_empirical_cdf(smp1, t) - F_empirical_cdf(smp2, t)) for
                t in np.linspace(20, 30, 100)])
    D_mn = np.sqrt((n * m) / (n + m)) * maxx
    return D_mn
```

n/m	5	10	100	200	400	600	800	1000
5	0.0000	0.7303	0.5892	0.6957	0.7278	0.7868	0.7189	0.7182
10	0.7303	0.0000	0.8141	0.6481	0.6794	0.6534	0.6207	0.6985
100	0.5892	0.8141	0.0000	0.6124	0.8944	1.0956	0.8132	0.8963
200	0.6957	0.6481	0.6124	0.0000	0.6062	0.9186	0.5692	0.4648
400	0.7278	0.6794	0.8944	0.6062	0.0000	0.8133	0.4899	0.4479
600	0.7868	0.6534	1.0956	0.9186	0.8133	0.0000	0.8024	1.1232
800	0.7189	0.6207	0.8132	0.5692	0.4899	0.8024	0.0000	0.6430
1000	0.7182	0.6985	0.8963	0.4648	0.4479	1.1232	0.6430	0.0000

Значения $D_{n,n}$ при $n = 5$:

```
Compute D_(5, 5) for samples 5_1 and 5_1: 0.0
Compute D_(5, 5) for samples 5_1 and 5_2: 0.316227766016838
Compute D_(5, 5) for samples 5_1 and 5_3: 0.632455532033676
Compute D_(5, 5) for samples 5_1 and 5_4: 0.3162277660168379
Compute D_(5, 5) for samples 5_1 and 5_5: 0.316227766016838
Compute D_(5, 5) for samples 5_2 and 5_1: 0.316227766016838
Compute D_(5, 5) for samples 5_2 and 5_2: 0.0
Compute D_(5, 5) for samples 5_2 and 5_3: 0.6324555320336759
Compute D_(5, 5) for samples 5_2 and 5_4: 0.316227766016838
Compute D_(5, 5) for samples 5_2 and 5_5: 0.316227766016838
Compute D_(5, 5) for samples 5_3 and 5_1: 0.632455532033676
Compute D_(5, 5) for samples 5_3 and 5_2: 0.6324555320336759
Compute D_(5, 5) for samples 5_3 and 5_3: 0.0
Compute D_(5, 5) for samples 5_3 and 5_4: 0.6324555320336759
```

```
Compute D_(5, 5) for samples 5_3 and 5_5: 0.632455532033676
Compute D_(5, 5) for samples 5_4 and 5_1: 0.3162277660168379
Compute D_(5, 5) for samples 5_4 and 5_2: 0.316227766016838
Compute D_(5, 5) for samples 5_4 and 5_3: 0.6324555320336759
Compute D_(5, 5) for samples 5_4 and 5_4: 0.0
Compute D_(5, 5) for samples 5_4 and 5_5: 0.316227766016838
Compute D_(5, 5) for samples 5_5 and 5_1: 0.316227766016838
Compute D_(5, 5) for samples 5_5 and 5_2: 0.316227766016838
Compute D_(5, 5) for samples 5_5 and 5_3: 0.632455532033676
Compute D_(5, 5) for samples 5_5 and 5_4: 0.316227766016838
Compute D_(5, 5) for samples 5_5 and 5_5: 0.0
```

6 Блок №3. Построение гистограммы и полигона частот

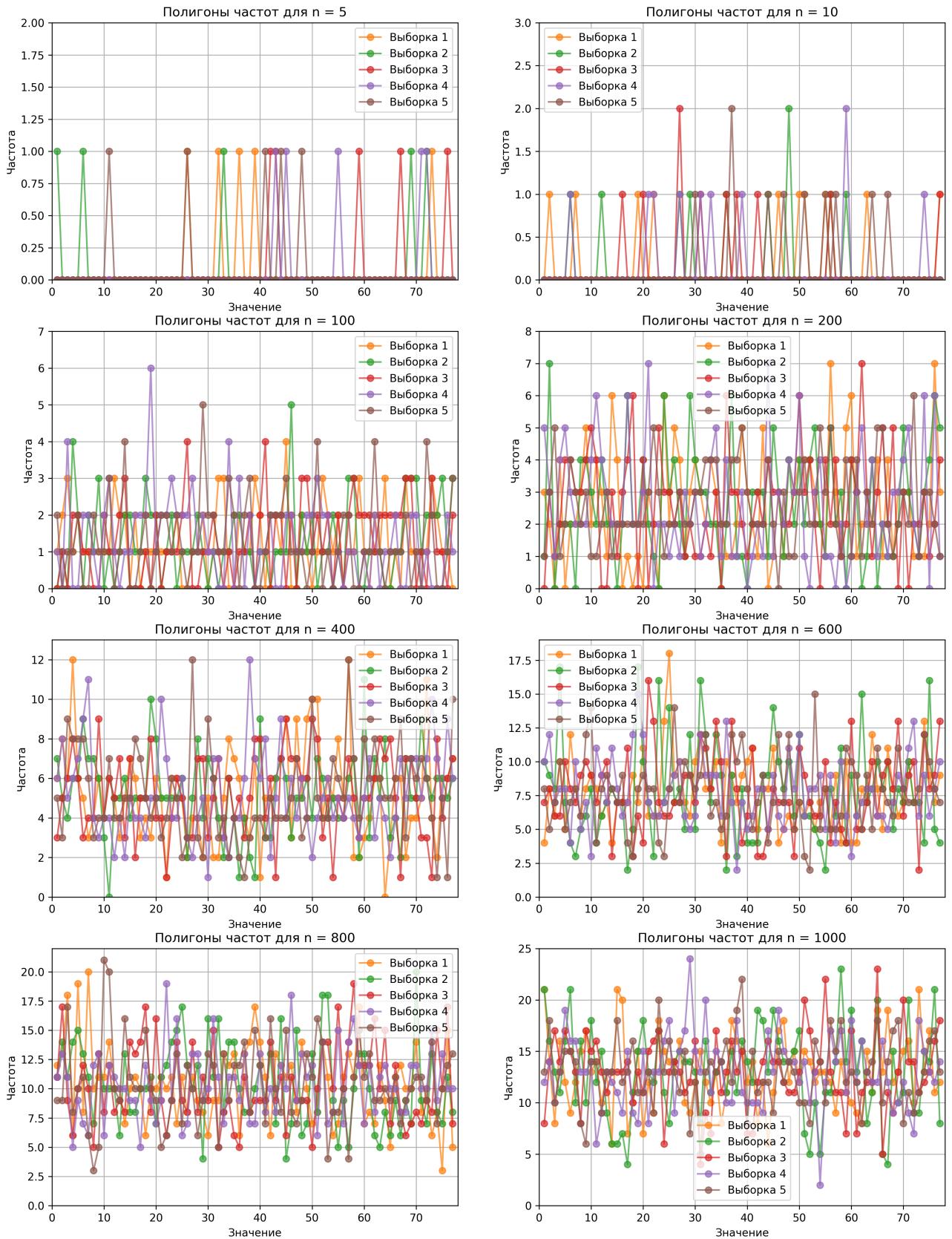
Для каждого распределения и для каждого n необходимо построить и привести в отчете:

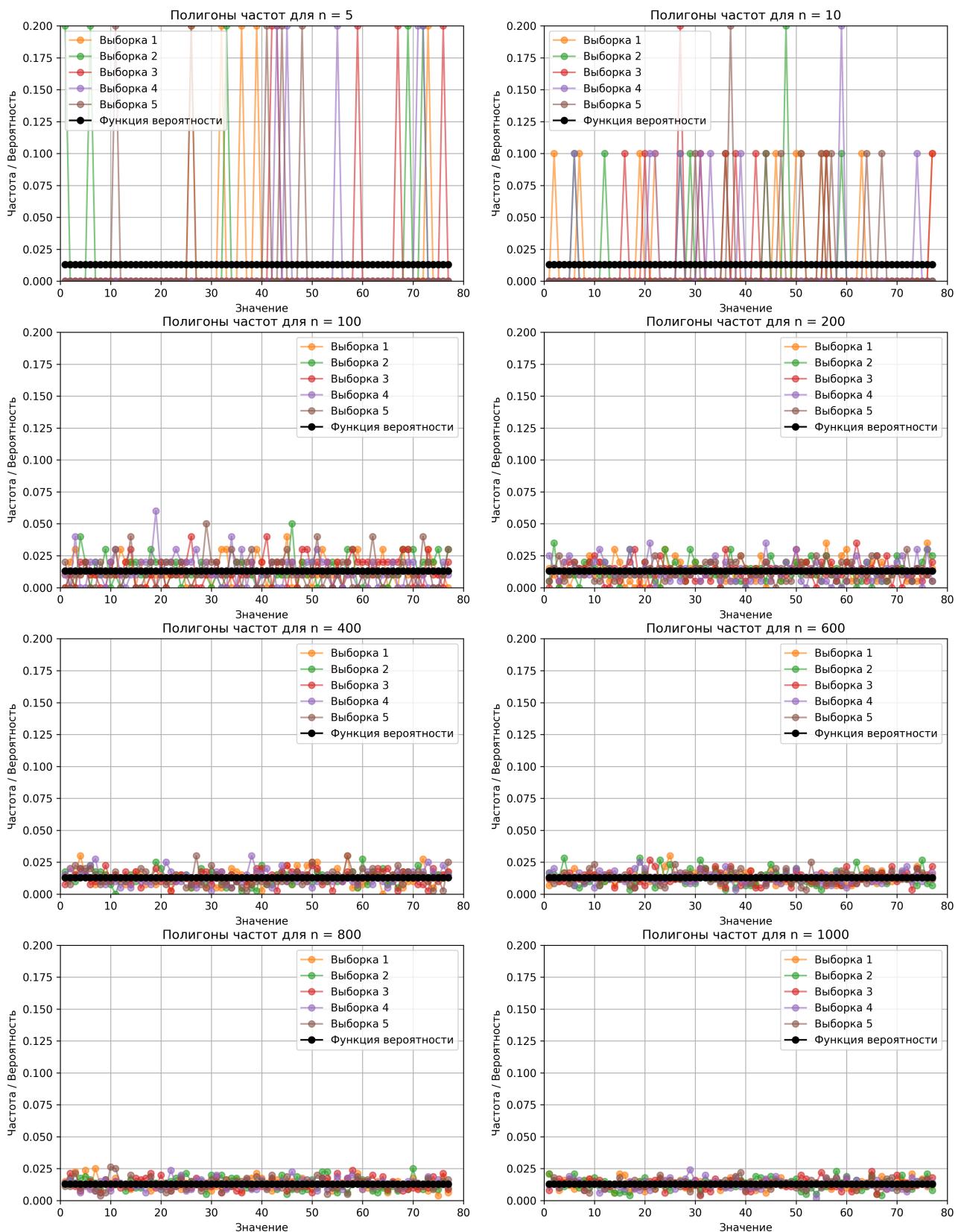
- полигон частот,
- сравнение с плотностью распределения для непрерывных распределений и функцией вероятности для дискретных распределений.

Необходимо пояснить полученные графики. Какие теоремы из курса математической статистики они иллюстрируют?

6.1 Дискретное равномерное I

Построим полигон частот для каждого n . Далее добавим на каждый график функцию вероятности дискретного равномерного распределения. Для выявления приближенности к функции вероятности с ростом n поделим частоту выпадения элемента на количество элементов в выборке и получим следующие графики.

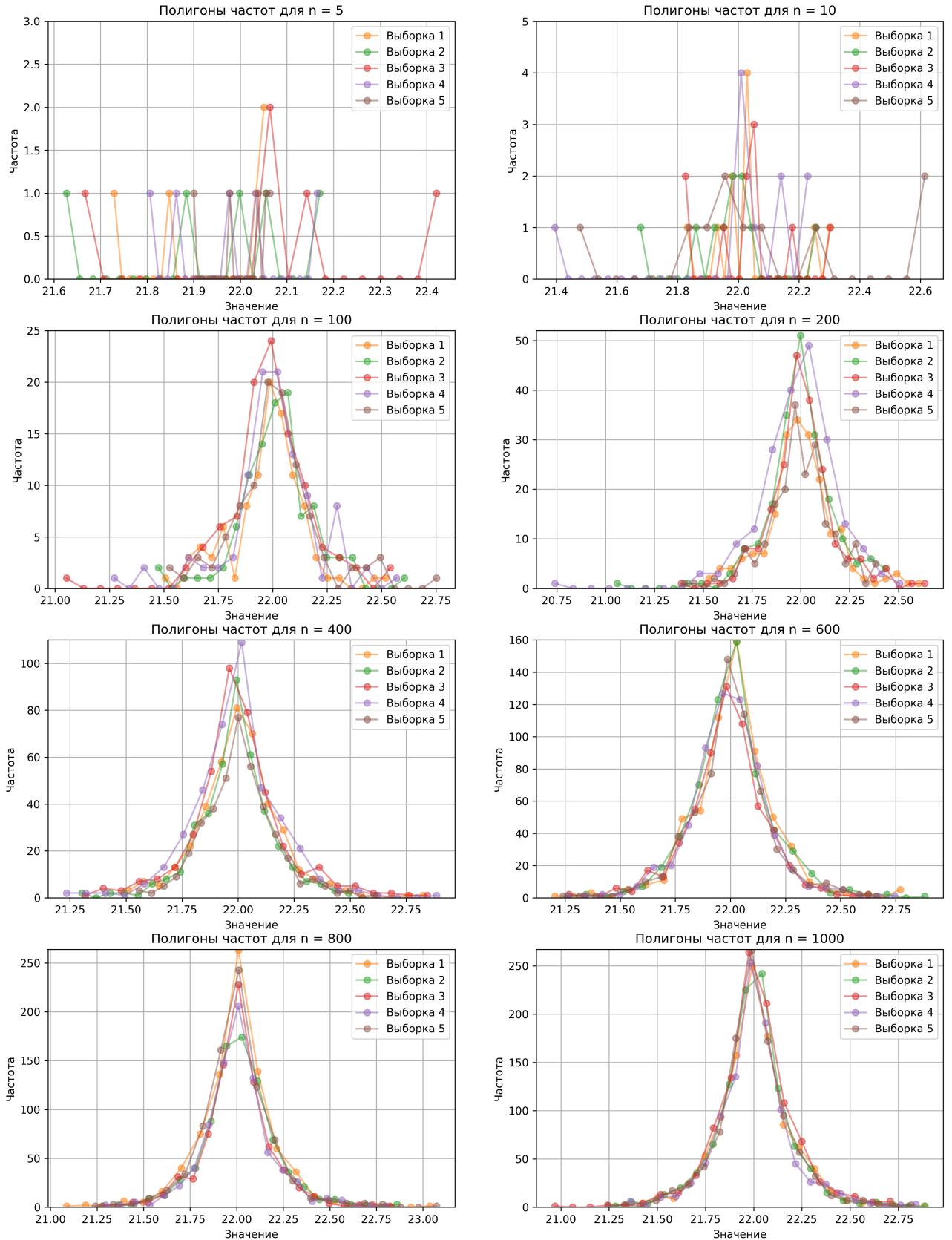


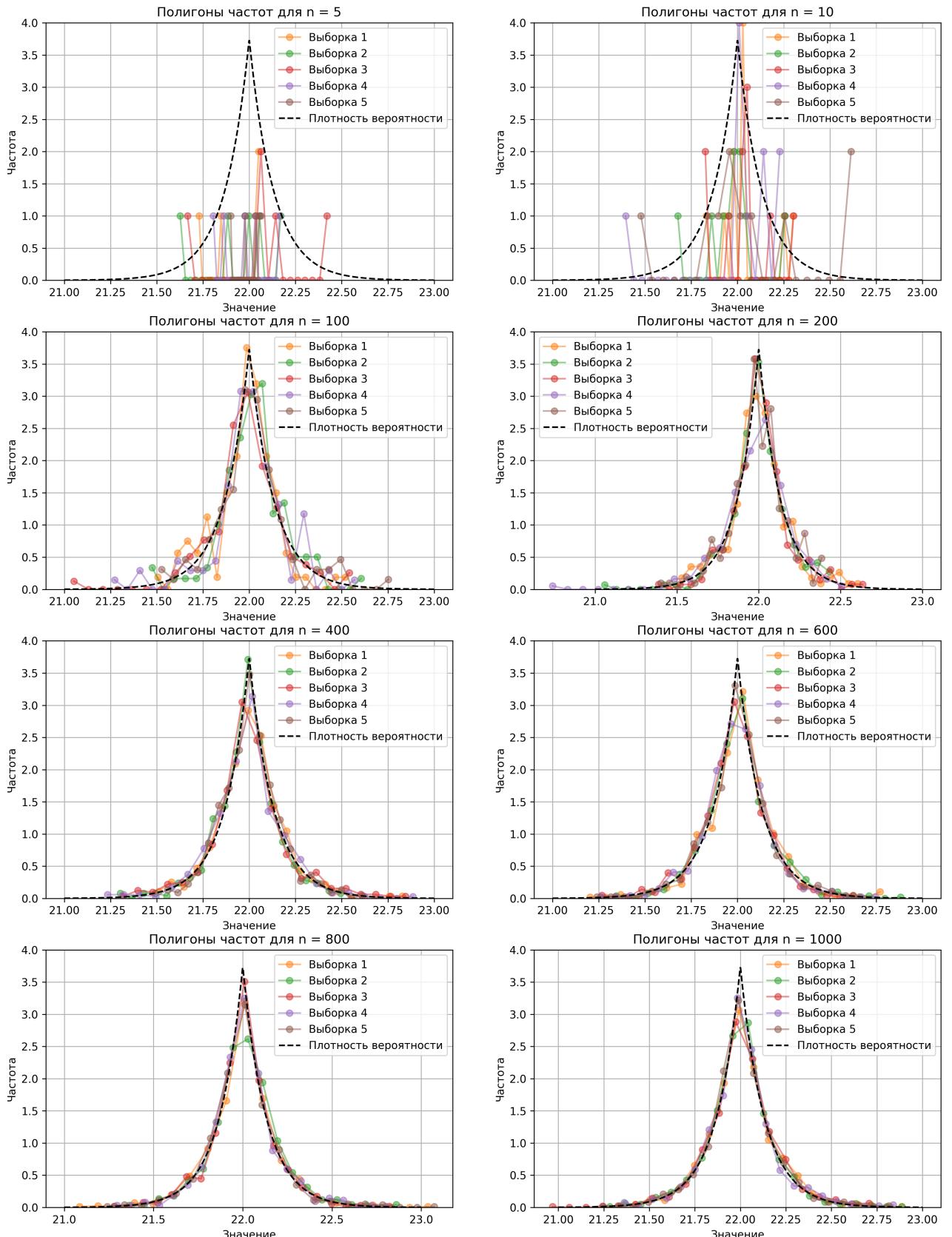


Как видно из графиков, с ростом объема выборки полигоны частот все больше совпадают с функцией вероятности распределения. Смотря на эти графики, можно заметить выполнение Закона больших чисел, т.е. с увеличением объема выборки (числа испытаний) частоты значений стремятся к истинным вероятностям.

ствам. В случае с дискретным равномерным распределением случайная величина принимает θ значений с равными $\frac{1}{\theta}$ вероятностями. Также стоит отметить, что полученные графики для дискретного распределения иллюстрируют теорему Гливенко-Кантелли, так как с ростом объема выборки ЭФР также стремится к теоретическому значению.

6.2 Распределение Лапласа





Как видно из графиков, с ростом объема выборки полигоны частот все больше совпадают с функцией плотности распределения. Смотря на эти графики, можно заметить выполнение Закона больших чисел, т.е. с увеличением объема выборки (числа испытаний) частоты значений стремятся к истинным плотностям веро-

ятности. Аналогично, полученные графики для непрерывного распределения иллюстрируют теорему Гливенко-Кантелли, так как с ростом объема выборки ЭФР также стремится к теоретическому значению и нормированная частота стремится к плотности вероятности.

7 Блок №4. Вычисление выборочных моментов

Для каждого сгенерированной выборки необходимо выписать значение выборочного среднего \bar{X} и выборочной дисперсии \bar{S}^2

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

$$\bar{S}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

Какими свойствами данные оценки обладают?

Также необходимо сравнить значения полученных оценок с истинными значениями математического ожидания и дисперсии.

7.1 Дискретное равномерное I

Истинное значение математического ожидания при $\theta = 77$:

$$M\xi = \frac{\theta + 1}{2} = \frac{77 + 1}{2} = 39.$$

Истинное значение дисперсии при $\theta = 77$:

$$D\xi = \frac{\theta^2 - 1}{12} = \frac{77^2 - 1}{12} = 494.$$

Выборки $n = 5$:

	\bar{X}	M	CF1	S^2	D	CF2
0	41.2	39.0	0.94	271.76	494.0	0.55
1	36.2	39.0	0.93	903.76	494.0	0.17
2	57.4	39.0	0.53	177.04	494.0	0.36
3	57.2	39.0	0.53	152.96	494.0	0.31
4	34.0	39.0	0.87	187.60	494.0	0.38

Выборки $n = 400$:

	\bar{X}	M	CF1	S^2	D	CF2
0	39.02	39.0	1.00	509.06	494.0	0.97
1	39.03	39.0	1.00	528.60	494.0	0.93
2	39.48	39.0	0.99	487.10	494.0	0.99
3	39.66	39.0	0.98	541.31	494.0	0.90
4	39.51	39.0	0.99	531.63	494.0	0.92

Выборки $n = 10$:

	\bar{X}	M	CF1	S^2	D	CF2
0	39.2	39.0	0.99	567.16	494.0	0.85
1	36.5	39.0	0.94	288.45	494.0	0.58
2	37.0	39.0	0.95	295.40	494.0	0.60
3	37.1	39.0	0.95	393.49	494.0	0.80
4	48.9	39.0	0.75	133.09	494.0	0.27

Выборки $n = 600$:

	\bar{X}	M	CF1	S^2	D	CF2
0	38.91	39.0	1.00	490.99	494.0	0.99
1	38.79	39.0	0.99	488.18	494.0	0.99
2	39.19	39.0	1.00	513.56	494.0	0.96
3	38.35	39.0	0.98	509.08	494.0	0.97
4	39.37	39.0	0.99	497.50	494.0	0.99

Выборки $n = 100$:

	\bar{X}	M	CF1	S^2	D	CF2
0	38.72	39.0	0.99	502.80	494.0	0.98
1	41.52	39.0	0.94	559.25	494.0	0.87
2	43.46	39.0	0.89	451.13	494.0	0.91
3	34.27	39.0	0.88	415.46	494.0	0.84
4	38.98	39.0	1.00	456.24	494.0	0.92

Выборки $n = 800$:

	\bar{X}	M	CF1	S^2	D	CF2
0	37.39	39.0	0.96	493.18	494.0	1.00
1	37.29	39.0	0.96	485.98	494.0	0.98
2	39.22	39.0	0.99	501.88	494.0	0.98
3	39.83	39.0	0.98	485.75	494.0	0.98
4	38.84	39.0	1.00	495.26	494.0	1.00

Выборки $n = 200$:

	\bar{X}	M	CF1	S^2	D	CF2
0	40.46	39.0	0.96	478.89	494.0	0.97
1	39.65	39.0	0.98	507.14	494.0	0.97
2	39.45	39.0	0.99	486.31	494.0	0.98
3	37.34	39.0	0.96	547.73	494.0	0.89
4	39.92	39.0	0.98	472.15	494.0	0.96

Выборки $n = 1000$:

	\bar{X}	M	CF1	S^2	D	CF2
0	39.23	39.0	0.99	512.75	494.0	0.96
1	38.74	39.0	0.99	499.26	494.0	0.99
2	39.14	39.0	1.00	511.26	494.0	0.97
3	38.84	39.0	1.00	485.51	494.0	0.98
4	39.23	39.0	0.99	497.73	494.0	0.99

Проверим свойства оценок $T_1(X) = \bar{X}$ и $T_2(X) = \bar{S}^2$:

$$T_1(X) = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i;$$

$$MT_1(X) = \frac{1}{n} \sum_{i=1}^n MX_i = \frac{1}{n} \sum_{i=1}^n \frac{\theta+1}{2} = \frac{\theta+1}{2}.$$

$MT_1(X) = \tau(\theta)$, значит $T_1(X)$ - **несмешенная оценка**.

$$DT_1(X) = D\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \frac{\theta^2 - 1}{12} = \frac{\theta^2 - 1}{12n};$$

$$\lim_{n \rightarrow \infty} DT_1 = \lim_{n \rightarrow \infty} \frac{\theta^2 - 1}{12n} = 0.$$

Так как $\lim_{n \rightarrow \infty} DT_1 = 0$ и T_1 - несмешенная оценка, значит, T_1 - **состоятельная**

оценка.

$$T_2(X) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2;$$

$$\begin{aligned} MT_2(X) &= \frac{1}{n} \sum_{i=1}^n M(X_i - \bar{X})^2 = M\left(\frac{1}{\theta}\left(\sum X_i^2 - 2\bar{X}\sum X_i + n\bar{X}^2\right)\right) = \\ &= M\left(\frac{1}{n}\left(\sum X_i^2 - 2\bar{X} * n\bar{X} + n\bar{X}^2\right)\right) = M\left(\frac{1}{n}\left(\sum X_i^2 - n\bar{X}^2\right)\right) = MX^2 - M\bar{X}^2, \\ \text{так как } MX^2 &= D\bar{X} + (M\bar{X})^2 = \frac{DX}{n} + (M\bar{X})^2, \text{ тогда:} \end{aligned}$$

$$MT_2(X) = MX^2 - \frac{DX}{n} - (MX)^2 = DX - \frac{DX}{n} = \left(1 - \frac{1}{n}\right)DX = \frac{n-1}{n} * \frac{\theta^2 - 1}{12}.$$

Следовательно, оценка T_2 - **смещенная**.

С ростом объема выборки значения выборочного среднего и выборочной дисперсии стремятся к истинным значениям. Выборочное среднее является несмещенной и состоятельной оценкой. Выборочная дисперсия является смещенной и состоятельной оценкой.

* **Состоятельность \bar{S}^2 :**

Оценка \bar{S}^2 является состоятельной, если:

$$\forall \varepsilon \lim_{n \rightarrow \infty} P\left(\left|\bar{S}^2 - D(X)\right| \geq \varepsilon\right) = 0$$

$$E(\bar{S}^2) = \frac{n-1}{n}D(X)$$

$$D(\bar{S}^2) = D\left(\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right) = \frac{1}{n^2} \sum_{i=1}^n D(X_i - \bar{X})^2$$

$$D(X_i - \bar{X})^2 = E(X_i - \bar{X})^2 - (E(X_i - \bar{X}))^2$$

$$E(X_i - \bar{X})^4 \xrightarrow[n \rightarrow \infty]{} \frac{1}{\theta} \sum_{x=1}^{\theta} \left(x - \frac{\theta+1}{2}\right)^4 = \frac{3(\theta^2-1)^2}{20}$$

$$D(X_i - \bar{X})^2 = \frac{3(\theta^2-1)^2}{20} - \left(\frac{\theta^2-1}{12}\right)^2 = \frac{103x^4 - 206x^2 + 103}{720}$$

$$D(\bar{S}^2) = \frac{1}{n^2} \cdot n \cdot \frac{103x^4 - 206x^2 + 103}{720} = \frac{103x^4 - 206x^2 + 103}{720n}$$

Используя неравенство Чебышева, получаем:

$$P \left(\left| \bar{S}^2 - E(\bar{S}^2) \right| \geq \varepsilon \right) \leq \frac{D(\bar{S}^2)}{\varepsilon^2} \xrightarrow{n \rightarrow \infty} 0$$

Следовательно, \bar{S}^2 является **состоятельной** оценкой для $D(X)$

7.2 Распределение Лапласа

Истинное значение математического ожидания при $\mu = 22$ и $\theta = 7,5$:

$$M\xi = \mu = 22.$$

Истинное значение дисперсии при $\mu = 22$ и $\theta = 7,5$:

$$D\xi = \frac{2}{\theta^2} = \frac{2}{7,5^2} \approx 0,036.$$

Выборки $n = 5$:

	\bar{X}	M	CF1	S^2	D	CF2
0	21.94	22.0	1.0	0.02	0.036	0.522
1	21.95	22.0	1.0	0.04	0.036	0.962
2	22.07	22.0	1.0	0.06	0.036	0.194
3	21.97	22.0	1.0	0.02	0.036	0.482
4	22.01	22.0	1.0	0.00	0.036	0.112

Выборки $n = 400$:

	\bar{X}	M	CF1	S^2	D	CF2
0	22.01	22.0	1.0	0.03	0.036	0.930
1	21.99	22.0	1.0	0.03	0.036	0.796
2	22.00	22.0	1.0	0.04	0.036	0.767
3	21.99	22.0	1.0	0.04	0.036	0.901
4	22.01	22.0	1.0	0.03	0.036	0.730

Выборки $n = 10$:

	\bar{X}	M	CF1	S^2	D	CF2
0	22.04	22.0	1.0	0.02	0.036	0.519
1	21.97	22.0	1.0	0.02	0.036	0.581
2	22.03	22.0	1.0	0.02	0.036	0.560
3	22.03	22.0	1.0	0.05	0.036	0.464
4	22.07	22.0	1.0	0.11	0.036	1.215

Выборки $n = 600$:

	\bar{X}	M	CF1	S^2	D	CF2
0	22.01	22.0	1.0	0.03	0.036	0.970
1	22.00	22.0	1.0	0.04	0.036	0.987
2	21.98	22.0	1.0	0.03	0.036	0.903
3	21.99	22.0	1.0	0.03	0.036	0.891
4	22.00	22.0	1.0	0.03	0.036	0.884

Выборки $n = 100$:

	\bar{X}	M	CF1	S^2	D	CF2
0	21.97	22.0	1.0	0.03	0.036	0.817
1	22.02	22.0	1.0	0.03	0.036	0.866
2	21.99	22.0	1.0	0.04	0.036	0.870
3	22.00	22.0	1.0	0.04	0.036	0.854
4	22.01	22.0	1.0	0.04	0.036	0.866

Выборки $n = 800$:

	\bar{X}	M	CF1	S^2	D	CF2
0	22.00	22.0	1.0	0.04	0.036	0.942
1	22.02	22.0	1.0	0.04	0.036	0.947
2	22.00	22.0	1.0	0.03	0.036	0.852
3	22.01	22.0	1.0	0.03	0.036	0.959
4	22.00	22.0	1.0	0.04	0.036	0.958

Выборки $n = 200$:

	\bar{X}	M	CF1	S^2	D	CF2
0	22.00	22.0	1.0	0.03	0.036	0.941
1	22.00	22.0	1.0	0.03	0.036	0.919
2	22.01	22.0	1.0	0.03	0.036	0.827
3	21.98	22.0	1.0	0.04	0.036	0.911
4	21.99	22.0	1.0	0.03	0.036	0.811

Выборки $n = 1000$:

	\bar{X}	M	CF1	S^2	D	CF2
0	22.00	22.0	1.0	0.03	0.036	0.977
1	22.00	22.0	1.0	0.04	0.036	0.989
2	22.01	22.0	1.0	0.04	0.036	0.907
3	22.00	22.0	1.0	0.04	0.036	0.958
4	22.00	22.0	1.0	0.03	0.036	0.969

Проверим свойства оценок $T_1(X) = \bar{X}$ и $T_2(X) = \bar{S}^2$:

$$T_1(X) = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i;$$

$$MT_1(X) = \frac{1}{n} \sum_{i=1}^n MX_i = \frac{1}{n} \sum_{i=1}^n \mu = \mu.$$

$MT_1(X) = \tau(\mu)$, значит $T_1(X)$ - **несмешенная оценка**.

$$DT_1(X) = D\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \frac{2}{\theta^2} = \frac{2}{n\theta^2};$$

$$\lim_{n \rightarrow \infty} DT_1 = \lim_{n \rightarrow \infty} \frac{2}{n\theta^2} = 0.$$

Так как $\lim_{n \rightarrow \infty} DT_1 = 0$ и T_1 - несмешенная оценка, значит, T_1 - **состоятельная**

оценка.

$$T_2(X) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2;$$

$$\begin{aligned} MT_2(X) &= \frac{1}{n} \sum_{i=1}^n M(X_i - \bar{X})^2 = M\left(\frac{1}{\theta}\left(\sum X_i^2 - 2\bar{X}\sum X_i + n\bar{X}^2\right)\right) = \\ &= M\left(\frac{1}{n}\left(\sum X_i^2 - 2\bar{X} * n\bar{X} + n\bar{X}^2\right)\right) = M\left(\frac{1}{n}\left(\sum X_i^2 - n\bar{X}^2\right)\right) = MX^2 - M\bar{X}^2, \end{aligned}$$

так как $M\bar{X}^2 = D\bar{X} + (M\bar{X})^2 = \frac{DX}{n} + (M\bar{X})^2$, тогда:

$$MT_2(X) = MX^2 - \frac{DX}{n} - (MX)^2 = DX - \frac{DX}{n} = \left(1 - \frac{1}{n}\right)DX = \frac{n-1}{n} * \frac{2}{\theta^2}.$$

Следовательно, оценка T_2 - **смещенная**.

С ростом объема выборки значения выборочного среднего и выборочной дисперсии стремятся к истинным значениям. Выборочное среднее является несмещенной и состоятельной оценкой. Выборочная дисперсия является смещенной и состоятельной оценкой.

Часть III

Построение точечных оценок параметра распределения

8 Блок №1. Получение оценок методом моментов и методом максимального правдоподобия

Для каждого из распределений (дискретное и непрерывное) необходимо получить оценки неизвестного параметра методом моментов и методом максимального правдоподобия.

Для каждой выборки, сгенерированной в пункте 2.1, необходимо привести значения полученных оценок.

8.1 Дискретное равномерное I

Дискретное равномерное распределение с параметром θ , $\xi \sim R\{1, \dots, \theta\}$.

Получение оценки методом моментов

Математическое ожидание (первый момент) существует:

$$M\xi = \frac{\theta + 1}{2}.$$

Метод моментов применим:

$$\frac{\theta + 1}{2} = \bar{X} \iff \hat{\theta} = 2\bar{X} - 1$$

Где $\hat{\theta}$ - оценка неизвестного параметра θ , построенная по методу моментов.

Таблица 8.1: Значения оценок ОММ для выборок

	5	10	100	200	400	600	800	1000
1	81.4	77.4	76.44	79.91	77.035	76.8200	73.7724	77.460
2	71.4	72.0	82.04	78.30	77.065	76.5766	73.5776	76.478
3	113.8	73.0	85.92	77.90	77.960	77.3866	77.4300	77.270
4	113.4	73.2	67.54	73.69	78.320	75.7100	78.6500	76.680
5	67.0	96.8	76.96	78.85	78.025	77.7400	76.6850	77.462

Получение оценки методом максимального правдоподобия

Равномерное распределение на множестве $\{1, \dots, \theta\}$ имеет следующий закон распределения

$$P(\xi = x) = \theta^{-1}, x \in \{1, \dots, \theta\}.$$

Вычислим функцию правдоподобия:

$$L(\bar{x}, \theta) = \prod_{i=1}^n P(\xi = x_i) = \prod_{i=1}^n \theta^{-1} * I(x_i \geq 1)I(x_i \leq \theta) = \frac{1}{\theta^n} * I(X_{(1)} \geq 1)I(X_{(n)} \leq \theta)$$

$I(X_{(1)} \geq 1) = 1 \quad \forall \theta \in \mathbb{N}$, тогда функция правдоподобия:

$$L(\bar{x}, \theta) = \begin{cases} \frac{1}{\theta^n}, & \text{если } \theta \geq X_{(n)}; \\ 0, & \text{если } \theta < X_{(n)}; \end{cases}$$

Тогда значение параметра, при котором $L(\bar{x}, \theta)$ примет максимальное значение:

$$\hat{\theta} = X_{(n)}$$

Где $\hat{\theta}$ - оценка неизвестного параметра θ , построенная по методу максимального правдоподобия.

Таблица 8.2: Значения оценок ОМП для выборок

	5	10	100	200	400	600	800	1000
1	73	77	76	77	77	77	77	77
2	72	59	77	77	77	77	77	77
3	76	77	77	77	77	77	77	77
4	72	74	77	77	77	77	77	77
5	48	67	77	77	77	77	77	77

8.2 Распределение Лапласа

Распределение Лапласа на множестве \mathbb{R} , $x, \mu, \theta \in \mathbb{R}$, $\theta > 0$.

Получение оценки методом моментов

Математическое ожидание (первый момент) и дисперсия существуют:

$$M\xi = \mu$$

$$D\xi = \frac{2}{\theta^2}.$$

Метод моментов применим:

$$\begin{cases} M\xi = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i; \\ D\xi = \bar{S} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2; \end{cases} \iff \begin{cases} \mu = \bar{X}; \\ \frac{2}{\theta^2} = \bar{S}; \end{cases} \iff \begin{cases} \hat{\mu} = \bar{X}; \\ \hat{\theta} = \sqrt{\frac{2}{\bar{S}}}. \end{cases}$$

Где $\hat{\theta}$ и $\hat{\mu}$ - оценки неизвестных параметров θ и μ , построенные по методу моментов.

Таблица 8.3: Значения $\hat{\mu}$ оценок ОММ для выборок

	5	10	100	200	400	600	800	1000
1	21.942866	22.039798	21.974956	21.998446	22.008341	22.014678	22.000643	22.003883
2	21.949215	21.970081	22.019188	21.996518	21.986071	22.004593	22.016358	21.998941
3	22.070682	22.029830	21.986599	22.006156	21.998084	21.983399	22.000688	22.008690
4	21.965442	22.027480	21.998890	21.983307	21.989330	21.994507	22.005761	22.003376
5	22.006016	22.069981	22.012103	21.992554	22.005260	22.000285	22.000504	22.002820

Таблица 8.4: Значения $\hat{\theta}$ оценок ОММ для выборок

	5	10	100	200	400	600	800	1000
1	10.377454	10.407509	8.299201	7.730391	7.779032	7.614816	7.291790	7.587478
2	7.360778	9.839216	8.060510	7.822655	8.408152	7.548568	7.308455	7.458549
3	5.580207	10.018233	7.055390	8.247729	6.754404	7.891242	8.124647	7.173391
4	10.800335	6.051400	7.006714	7.185941	7.154069	7.946989	7.658224	7.345550
5	22.411444	4.183131	7.043269	8.328962	8.779717	7.978154	7.346758	7.617792

Получение оценки методом максимального правдоподобия

Распределение Лапласа на множестве \mathbb{R} задается плотностью распределения вида

$$f(x) = \frac{\theta}{2} * \exp\{-\theta|x - \mu|\};$$

Вычислим функцию правдоподобия:

$$L(\bar{x}, \theta) = \prod_{i=1}^n f(x_i) = \prod_{i=1}^n \frac{\theta}{2} * \exp\{-\theta|x_i - \mu|\} = \frac{\theta^n}{2^n} * \prod_{i=1}^n \exp\{-\theta|x_i - \mu|\}$$

Логарифм функции правдоподобия:

$$\ln L(\bar{x}, \theta) = n \ln \frac{\theta}{2} - \theta \sum_{i=1}^n |x_i - \mu|$$

Для нахождения оценки $\hat{\mu}$ максимизируем $\ln L(\mu, \theta)$ по μ . Это равносильно минимизации суммы модулей отклонений:

$$\frac{\partial}{\partial \mu} \ln L(\mu, \theta) = -\theta \sum_{i=1}^n \frac{\partial |X_i - \mu|}{\partial \mu}.$$

Производная от $|X_i - \mu|$:

$$\frac{\partial |X_i - \mu|}{\partial \mu} = \text{sign}(X_i - \mu) = \begin{cases} -1, & \text{если } X_i > \mu, \\ 1, & \text{если } X_i < \mu, \\ 0, & \text{если } X_i = \mu. \end{cases}$$

Тогда:

$$\frac{\partial}{\partial \mu} \ln L(\mu, \theta) = -\theta \sum_{i=1}^n \text{sign}(X_i - \mu).$$

Ставим производную равной нулю:

$$\sum_{i=1}^n \text{sign}(X_i - \mu) = 0.$$

Тогда значение параметра μ , при котором $L(\bar{x}, \theta)$ примет максимальное значение:

$$\hat{\mu} = \text{median}\{X_1, \dots, X_n\}$$

Где $\hat{\mu}$ - оценка неизвестного параметра μ , построенная по методу максимального правдоподобия. Это медиана выборки - значение μ , при котором количества элементов выборки, меньших и больших μ , одинаковы.

Для нахождения оценки $\hat{\theta}$ максимизируем $\ln L(\mu, \theta)$ по θ . Производная:

$$\frac{\partial}{\partial \theta} \ln L(\mu, \theta) = \frac{\partial}{\partial \theta} \left(n \ln \frac{\theta}{2} - \theta \sum_{i=1}^n |X_i - \mu| \right) = \frac{n}{\theta} - \sum_{i=1}^n |X_i - \mu|.$$

$$\frac{n}{\theta} - \sum_{i=1}^n |X_i - \mu| = 0 \iff \hat{\theta} = \frac{n}{\sum_{i=1}^n |X_i - \hat{\mu}|}$$

Где $\hat{\theta}$ - оценка неизвестного параметра θ , построенная по методу максимального правдоподобия. $\hat{\mu}$ - это медиана выборки.

Таблица 8.5: Значения $\hat{\mu}$ оценок ММП для выборок

	5	10	100	200	400	600	800	1000
1	22.041735	22.025897	21.985863	22.000061	22.006376	22.012371	22.006919	22.003322
2	22.004447	21.973577	22.012184	21.997400	21.994274	22.001975	22.004624	22.001386
3	22.080415	22.042556	21.981292	22.001518	21.988611	21.998470	22.004009	22.006326
4	21.966596	22.047166	22.001126	22.002547	21.988571	21.997592	22.004145	21.998165
5	22.038680	21.975081	22.005543	21.986625	22.008490	22.002197	21.996982	21.997025

Таблица 8.6: Значения $\hat{\theta}$ оценок ММП для выборок

	5	10	100	200	400	600	800	1000
1	9.274355	10.822136	8.144349	7.630685	7.569478	7.655102	7.415124	7.384994
2	6.893263	10.646755	8.020382	7.931339	8.280024	7.504690	7.138736	7.422682
3	5.700849	10.025219	7.241377	8.279414	6.848954	7.676734	8.070918	7.209241
4	9.306921	7.246126	7.233615	7.173178	7.182060	7.755750	7.649918	7.415366
5	20.299423	3.959926	7.113093	7.900682	8.317780	7.882622	7.489018	7.567802

9 Блок №2. Поиск оптимальных оценок

Для каждого из распределений (дискретное и непрерывное) необходимо предложить параметрическую функцию $\tau(\theta)$, для которой существует оптимальная оценка.

Для каждой выборки, сгенерированной в пункте 2.1, необходимо привести значения полученных оценок.

9.1 Дискретное равномерное I

Найдем достаточную статистику, используя критерий факторизации:

$$L(\bar{x}, \theta) = \frac{1}{\theta^n} * I(X_{(1)} \geq 1)I(X_{(n)} \leq \theta) = h(\bar{x}) * g(T(\bar{x}), \theta)$$

Значит, $X_{(n)}$ - достаточная статистика.

Проверим ее на полноту по определению:

$$\forall \phi = \phi(X_{(n)}) : M(\phi) = 0 \Rightarrow \phi = 0.$$

$$\begin{aligned} F_{X_{(n)}}(t) &= P(X_{(n)} \leq t) = [F(t)]^n = \frac{\lfloor t \rfloor^n}{\theta^n} \\ \Rightarrow P(X_{(n)} = t) &= F_{X_{(n)}}(t) - F_{X_{(n)}}(t-1) = \frac{\lfloor t \rfloor^n}{\theta^n} - \frac{\lfloor t-1 \rfloor^n}{\theta^n}, t \in [1, \theta] \\ M(\phi(X_{(n)})) &= \sum_{i=1}^{\theta} \phi(i) * P(X_{(n)} = i) = \sum_{i=1}^{\theta} \phi(i) * \left(\frac{i^n}{\theta^n} - \frac{(i-1)^n}{\theta^n} \right) \end{aligned}$$

$$M(\phi(X_{(n)})) = 0:$$

$$\sum_{i=1}^{\theta} \phi(i) * \left(\frac{i^n}{\theta^n} - \frac{(i-1)^n}{\theta^n} \right) = 0$$

$$\sum_{i=1}^{\theta} \phi(i) * (i^n - (i-1)^n) = 0 \Rightarrow \phi(i) * (i^n - (i-1)^n) = 0 \quad \forall i \in \{1, \dots, \theta\}$$

$$(i^n - (i-1)^n) \neq 0 \quad \forall i \in \{1, \dots, \theta\} \Rightarrow \phi(i) = 0 \quad \forall i \in \{1, \dots, \theta\} \quad \forall \theta \in \mathbb{N}$$

Значит, статистика $X_{(n)}$ - достаточная и полная.

Теорема. Если существует полная достаточная статистика, то произвольная функция от нее будет являться оптимальной оценкой своего математического ожидания.

Воспользуемся теоремой, те найдем функцию от нашей полной достаточной статистики $g(X_{(n)})$ такую, что ее математическое ожидание равно нашей параметрической функции $\tau(\theta) = \theta$.

$$M(g(X_{(n)})) = \sum_{i=1}^{\theta} g(i)P(X_{(n)} = i) = \sum_{i=1}^{\theta} g(i) \left(\frac{i^n}{\theta^n} - \frac{(i-1)^n}{\theta^n} \right) = \theta.$$

Заметим, что:

$$\sum_{i=1}^{\theta} P(X_{(n)} = i) = \sum_{i=1}^{\theta} \left(\frac{i^n}{\theta^n} - \frac{(i-1)^n}{\theta^n} \right) = 1 \Rightarrow \sum_{i=1}^{\theta} (i^n - (i-1)^n) = \theta^n.$$

Рассмотрим следующее выражение:

$$\sum_{i=1}^{\theta} (i^{n+1} - (i-1)^{n+1}) = (1 + 2^{n+1} + \dots + \theta^{n+1}) - (0 + 1 + \dots + (\theta-1)^{n+1}) = \theta^{n+1};$$

$$\begin{aligned} & \frac{1}{\theta^n} \sum_{i=1}^{\theta} (i^{n+1} - (i-1)^{n+1}) = \frac{\theta^{n+1}}{\theta^n} = \theta \iff \\ & \iff \frac{1}{\theta^n} \sum_{i=1}^{\theta} (i^{n+1} - (i-1)^{n+1}) \left(\frac{i^n - (i-1)^n}{i^n - (i-1)^n} \right) = \theta \iff \\ & \iff \sum_{i=1}^{\theta} \frac{i^{n+1} - (i-1)^{n+1}}{i^n - (i-1)^n} \left(\frac{i^n}{\theta^n} - \frac{(i-1)^n}{\theta^n} \right) = \theta. \end{aligned}$$

Получили:

$$\begin{aligned} M(g(X_{(n)})) &= \sum_{i=1}^{\theta} g(i) \left(\frac{i^n}{\theta^n} - \frac{(i-1)^n}{\theta^n} \right) = \theta. \\ \Rightarrow g(i) &= \frac{i^{n+1} - (i-1)^{n+1}}{i^n - (i-1)^n} \quad \forall i \in \{1, \dots, \theta\} \end{aligned}$$

Тогда оптимальная оценка θ будет равна:

$$T(X) = g(X_{(n)}) = \frac{X_{(n)}^{n+1} - (X_{(n)} - 1)^{n+1}}{X_{(n)}^n - (X_{(n)} - 1)^n}.$$

Таблица 9.1: Значения оптимальных оценок для выборок

	5	10	100	200	400	600	800	1000
1	87.0055	84.1608	76.3623	77.079	77.0054	77.0004	77.00002	77.000002
2	85.8056	64.3641	77.3709	77.079	77.0054	77.0004	77.00002	77.000002
3	90.6053	84.1608	77.3709	77.079	77.0054	77.0004	77.00002	77.000002
4	85.8056	80.8612	77.3709	77.079	77.0054	77.0004	77.00002	77.000002
5	57.0084	73.1624	77.3709	77.079	77.0054	77.0004	77.00002	77.000002

9.2 Распределение Лапласа

Найдем достаточную статистику, используя критерий факторизации:

$$\begin{aligned} L(\bar{x}, \theta) &= \prod_{i=1}^n f(x_i) = \prod_{i=1}^n \frac{\theta}{2} * \exp\{-\theta|x_i - \mu|\} \\ &= \frac{\theta^n}{2^n} * \exp\{-\theta \sum_{i=1}^n |x_i - \mu|\} = h(\bar{x}) * g(T(\bar{x}), \theta) \end{aligned}$$

Значит, $T(x) = \sum_{i=1}^n |x_i - \mu|$ - достаточная статистика. Далее докажем ее полноту.

Покажем, что распределение Лапласа принадлежит экспоненциальному распределению:

$$f(x) = \frac{\theta}{2} * \exp\{-\theta|x - \mu|\} = \exp\{\ln \frac{\theta}{2} - \theta|x - \mu|\}$$

$$A(\theta) = -\theta, \quad A'(\theta) = -1;$$

$$B(x) = |x - \mu|;$$

$$C(\theta) = \ln \frac{\theta}{2}, \quad C'(\theta) = \frac{1}{\theta};$$

$$D(x) = 0;$$

⇒ распределение Лапласа принадлежит экспоненциальному семейству.

Значит, $|x_i - \mu|$ - полная и достаточная статистика по теореме о полноте экспоненциальных семейств. Следовательно, $T(x) = \sum_{i=1}^n |x_i - \mu|$ - полная статистика.

Мы доказали, что $T(x)$ - полная и достаточная статистика.

Вычислим математическое ожидание статистики $T(x)$:

$$M \left[\sum_{i=1}^n |X_i - \mu| \right] = \sum_{i=1}^n M [|X_i - \mu|] = n \cdot M [|X_1 - \mu|].$$

$$\begin{aligned} M [|X - \mu|] &= \int_{-\infty}^{\infty} |x - \mu| \cdot f(x) dx = \int_{-\infty}^{\infty} |x - \mu| \cdot \frac{\theta}{2} \exp\{-\theta|x - \mu|\} dx = \\ &= \frac{\theta}{2} \left(\int_{-\infty}^{\mu} (\mu - x) \exp\{-\theta(\mu - x)\} dx + \int_{\mu}^{\infty} (x - \mu) \exp\{-\theta(x - \mu)\} dx \right). \end{aligned}$$

Первый интеграл ($x < \mu$)

Заменим $\mu - x = u$, где $x = \mu - u$ и $dx = -du$:

$$\int_{-\infty}^{\mu} (\mu - x) \exp\{-\theta(\mu - x)\} dx = \int_0^{\infty} u \exp\{-\theta u\} du.$$

Второй интеграл ($x > \mu$)

Заменим $x - \mu = u$, где $x = \mu + u$ и $dx = du$:

$$\int_{\mu}^{\infty} (x - \mu) \exp\{-\theta(x - \mu)\} dx = \int_0^{\infty} u \exp\{-\theta u\} du.$$

Оба интеграла идентичны, значит:

$$\begin{aligned} M [|X - \mu|] &= \theta \int_0^{\infty} u \exp\{-\theta u\} du = \theta \left(-\frac{u}{\theta} \exp(-\theta u) \Big|_0^{\infty} + \int_0^{\infty} \frac{1}{\theta} \exp(-\theta u) du \right) = \\ &= \theta \left(0 + \frac{1}{\theta^2} \right) = \frac{1}{\theta}. \end{aligned}$$

Получим:

$$M(T(x)) = M \left[\sum_{i=1}^n |X_i - \mu| \right] = \frac{n}{\theta}.$$

Рассмотрим статистику H как произвольную функцию ϕ от полной доста-
точной статистики:

$$H = \phi(T(x)) = \frac{T(x)}{n} = \frac{\sum_{i=1}^n |x_i - \mu|}{n}.$$

Ее математическое ожидание:

$$M(H) = M\left(\frac{\sum_{i=1}^n |x_i - \mu|}{n}\right) = \frac{1}{n} M\left(\sum_{i=1}^n |x_i - \mu|\right) = \frac{1}{n} M(T(x)) = \frac{1}{\theta}.$$

$$\Rightarrow H = \frac{\sum_{i=1}^n |x_i - \mu|}{n} - \text{оптимальная оценка } \tau(\theta) = \frac{1}{\theta} \text{ по Теореме.}$$

Таблица 9.2: Значения оптимальных оценок для выборок

	5	10	100	200	400	600	800	1000
1	0.107824	0.092403	0.122785	0.131050	0.132110	0.130632	0.134860	0.135410
2	0.145069	0.093925	0.124682	0.126082	0.120773	0.133250	0.140081	0.134722
3	0.175412	0.099748	0.138095	0.120782	0.146008	0.130264	0.123902	0.138711
4	0.107447	0.138005	0.138243	0.139408	0.139236	0.128937	0.130720	0.134855
5	0.049262	0.252530	0.140586	0.126571	0.120224	0.126861	0.133529	0.132139

10 Блок №3. Работа с данными

Для выбранного интерпретации, обоснованной в первой домашней работе найти данные, соответствующие интерпретации. При этом необходимо привести источники данных, а также сами данные (или постоянную ссылку на данные, если они взяты из открытых источников.)

Для полученных данных необходимо проделать такую же работу как и с построенным выборками, а именно:

1. привести значение выборочного среднего и выборочной дисперсии.
2. привести значение предложенной оценки X и (в случае их несовпадения) значение оптимальной оценки.

10.1 Дискретное равномерное I

Дискретное равномерное распределение на множестве $\{1, \dots, \theta\}$ - это распределение случайной величины, которая принимает все целочисленные значения из отрезка $[1, \dots, \theta]$ с одинаковыми вероятностями, где $\theta \in R$.

Мы будем работать с этим датасетом. Он включает в себя данные о победных числах в лотерее. Числа расположены в диапазоне $\{1, \dots, 49\}$.

NUMBER DRAWN 1	NUMBER DRAWN 2	NUMBER DRAWN 3	NUMBER DRAWN 4	NUMBER DRAWN 5	NUMBER DRAWN 6	BONUS NUMBER
3	11	12	14	41	43	13
8	33	36	37	39	41	9
1	6	23	24	27	39	34
3	9	10	13	20	43	34
5	14	21	31	34	47	45
...

Рассмотрим выборочное среднее и выборочную дисперсию нашей выборки:

$$\bar{X} = 25.31;$$

$$\bar{S} = 200.09.$$

Оценки параметра θ :

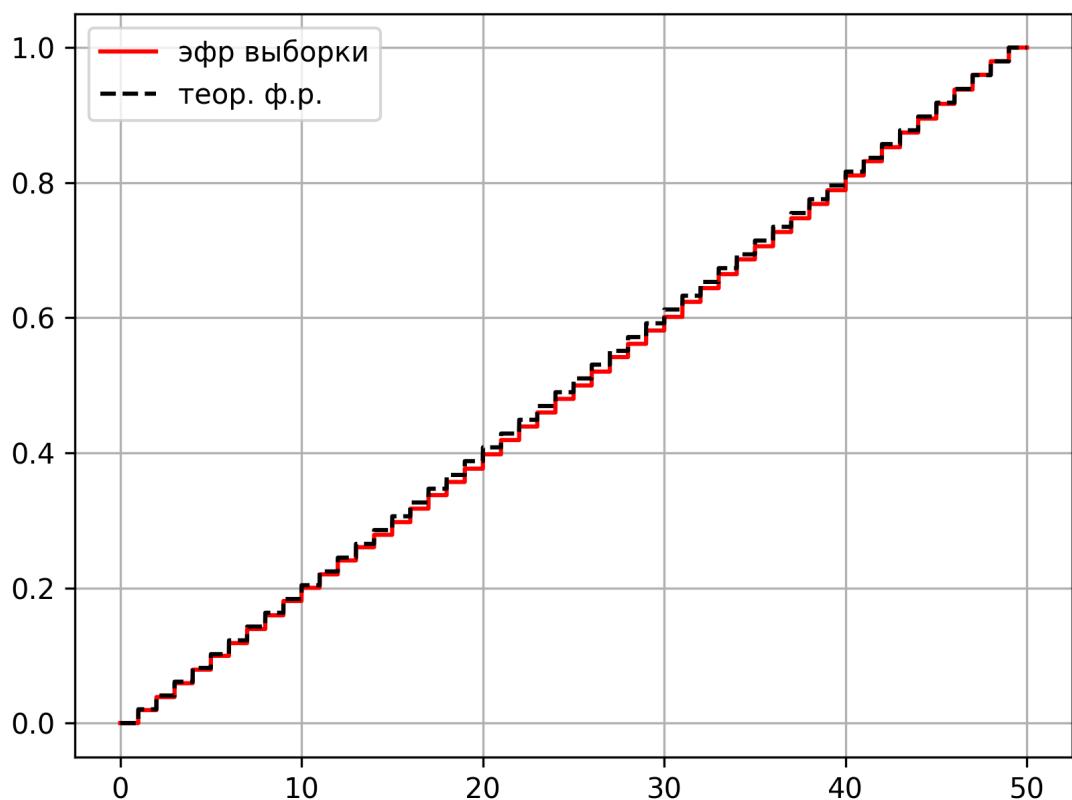
$$\hat{\theta}_{\text{ММ}} = 49.62;$$

$$\hat{\theta}_{\text{ММП}} = 49;$$

$$\hat{\theta}_{\text{опт}} = 49;$$

$$\hat{\theta}_{\text{ММП}} = \hat{\theta}_{\text{опт}} = \theta_{\text{ист}} = 49.$$

Построим графики теоретической ФР и ЭФР для данной выборки:



10.2 Распределение Лапласа

Пример интерпретации распределения:

Предположим, что метеорологическая служба прогнозирует температуру на завтра с некоторой средней ошибкой μ (то есть в среднем прогнозы оказываются точными). Однако возможны отклонения в обе стороны — как завышенные, так и заниженные, и вероятность этих отклонений убывает по мере увеличения их величины. Для моделирования ошибок прогноза используется распределение Лапласа с параметром θ .

Мы будем работать с этим датасетом. Датасет описывает погоду в Сеуле, Южная Корея.

Исходные данные в основном состоят из данных прогноза модели LDAPS на следующий день, максимальной и минимальной температуры на месте в текущий день, а также вспомогательных географических переменных.

Предсказание средней температуры на день и реальное значение температуры в те даты - два основных параметра, которые нам понадобятся.

Date	Present_Tmax	Present_Tmin	pred_tmax	pred_tmin	loss_max	loss_min	loss_mean
2013-07-02	24.4	20.6	24.8	18.7	0.4	-1.9	-0.75
2013-07-03	27.9	17.9	28.1	17.8	0.2	-0.1	0.05
2013-07-04	24.9	21.5	25.2	20.8	0.3	-0.7	-0.20
2013-07-05	27.8	19.9	28.0	19.7	0.2	-0.2	0.00
2013-07-06	27.6	19.6	28.2	19.3	0.6	-0.3	0.15
...

Возьмем разность этих показателей и получим нужную выборку - значения ошибок предсказания температуры. Покажем, что эти данные действительно имеют распределение Лапласа.

Рассмотрим выборочное среднее и выборочную дисперсию нашей выборки:

$$\bar{X} = 0.15;$$

$$\bar{S} = 0.29.$$

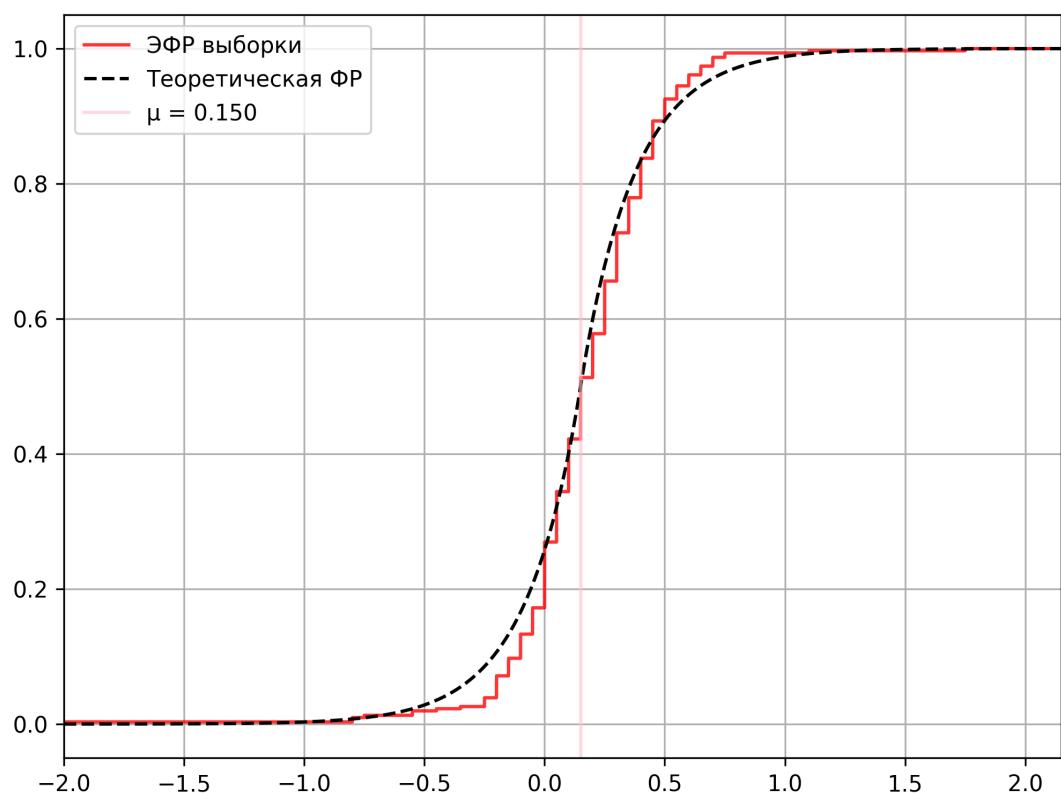
Оценки параметра θ :

$$\hat{\theta}_{\text{ММ}} = 2.63, \quad \hat{\mu}_{\text{ММ}} = 0.15;$$

$$\hat{\theta}_{\text{ММП}} = 4.42, \quad \hat{\mu}_{\text{ММП}} = 0.15;$$

$$\hat{\theta}_{\text{опт}} = 4.42.$$

Построим графики теоретической ФР и ЭФР для данной выборки:



Часть IV

Проверка статистических гипотез

11 Блок №1. Проверка гипотезы о виде распределения

Для каждой сгенерированной выборки необходимо рассмотреть следующие статистики:

- Критерий согласия Колмогорова (Смирнова),
- Критерий согласия хи-квадрат,
- Критерий согласия Колмогорова (Смирнова) для сложной гипотезы (в условиях когда неизвестен параметр распределения),
- Критерий согласия хи-квадрат для сложной гипотезы (в условиях когда неизвестен параметр распределения).

11.1 Критерий согласия Колмогорова

Будут рассмотрены простые гипотезы $\mathcal{F}_0 = \{R\{1, 77\}\}$ и $\mathcal{F}_0 = \{L(22, 7.5)\}$.

11.1.1 Распределение Лапласа

Для применения критерия Колмогорова будем использовать следующую статистику:

$$D_n = D_n(X) = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \quad (4.1)$$

Под $F(x)$ в приведённой формуле подразумевается функция распределения $L(22, 7.5)$ по основной гипотезе. Как нам известно, $\hat{F}_n(x)$ есть оптимальная, несмещенная и состоятельная оценка $F(x)$ в любой точке, и при увеличении объема выборки n значения $\hat{F}_n(x)$ стремятся к значениям $F(x)$. Значит, если верна гипотеза H_0 и n достаточно велико, то значение D_n не должно сильно отклоняться от 0.

По теореме Колмогорова при условии непрерывности функции распределения $F(x)$ и при $n \geq 20$ верно:

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

Из этого следует, что при заданном уровне значимости критерия α критическую область можно задать следующим образом:

$$\mathcal{X}_\alpha = \{\bar{x} : D_n(\bar{x})\sqrt{n} > \lambda_{1-\alpha}\}.$$

Здесь, $\lambda_{1-\alpha}$ - квантиль Колмогоровского распределения уровня $1 - \alpha$, где $\alpha \in (0, 1)$ - уровень значимости.

Отметим, что распределение статистики D_n при гипотезе H_0 не зависит от функции $F(x)$. Таким образом, по таблице значений функции Колмогорова $K(t)$ приведенный критерий может быть рассчитан для любого непрерывного распределения.

Вместо статистики $\sqrt{n}D_n \geq \lambda_{1-\alpha}$ при малых значениях n ($n < 20$) рекомендуется использовать статистику, которая сходится к распределению Колмогорова:

$$S_n = \frac{6nD_n + 1}{6\sqrt{n}}.$$

Таблица для выбранных значений уровней значимости α и соответствующих им квантилей уровня $1 - \alpha$:

α	0,10	0,05	0,01
$\lambda_{1-\alpha}$	1,22	1,36	1,63

Результаты применения критерия Колмогорова:

n = 5			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 10			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H1	H1	H0
5	H0	H0	H0

n = 100			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 200			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 400			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H1	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 600			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H1	H0	H0
2	H0	H0	H0
3	H1	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 800			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H1	H1	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 1000			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

11.1.2 Дискретное равномерное I

Функция $D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)|$ не зависит от вида функции распределения $F(x)$, только в случае, когда $F(x)$ - непрерывная функция. Встает вопрос, что делать если $F(x)$ имеет точки разрыва.

Утверждение. Пусть Y_1, \dots, Y_n — независимые, одинаково распределенные случайные величины, $Y_i \sim \mathcal{R}[0, 1]$, X_1, \dots, X_n — выборка из некоторого распределения, функция которого имеет точки разрыва. Построим следующую случайную величину:

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

где $F(x-) = \lim_{z \rightarrow 0} F(x-z)$. Тогда случайная величина $U_i \sim \mathcal{R}[0, 1]$.

Воспользуемся этим утверждением и получим следующие результаты применения критерия Колмогорова:

n = 5			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H1	H0	H0
4	H1	H0	H0
5	H0	H0	H0

n = 10			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H1	H0	H0

n = 100			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H1	H0	H0
4	H1	H1	H0
5	H0	H0	H0

n = 200			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 400			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 600			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 800			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H1	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 1000			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

11.2 Критерий согласия хи-квадрат

11.2.1 Дискретное равномерное I

Рассмотрим выборку $X = (X_1, \dots, X_n)$ из $\mathcal{L}(\xi)$, где ξ — случайная величина, принимающая значения из $\{1, \dots, N\}$ с вероятностями p_1, \dots, p_N . Определим случайную величину $\nu_k^{(n)}$ как частоту встречаемости значения k в выборке X , $k = \{1, \dots, N\}$, т.е.:

$$\nu_k^{(n)} = \sum_{i=1}^n \mathbb{I}(X_i = k)$$

Рассмотрим вектор частот $\nu^{(n)} = (\nu_1^{(n)}, \dots, \nu_N^{(n)})$. Данный вектор имеет полиномиальное распределение:

$$P(\nu^{(n)} = (m_1, \dots, m_N)) = \frac{n!}{m_1! \cdots m_N!} p_1^{m_1} \cdots p_N^{m_N}$$

Статистику Пирсона (статистику хи-квадрат):

$$\chi^2 = \sum_{i=1}^N \frac{(\nu_i^{(n)} - np_i)^2}{np_i}$$

При увеличении объема выборки n значения $\nu_k^{(n)}$ стремятся к значениям p_k . Значит, если верна гипотеза H_0 и n достаточно велико, то значение X_N^2 не должно сильно отклоняться от 0.

Теорема (Предельное распределение статистики Пирсона). Пусть случайный вектор частот $\nu^{(n)} = (\nu_1^{(n)}, \dots, \nu_N^{(n)})$ имеет полиномиальное распределение с параметрами n и $\vec{p} = (p_1, \dots, p_N)$. Если вектор \vec{p} фиксирован, а $n \rightarrow \infty$, то распределение статистики Пирсона сходится к распределению χ^2 с $N - 1$ степенью свободы.

То есть нам нужно составить статистику, описывающую разброс между p_i и $\frac{\nu_i}{n}$. Также он должна сходиться к кси-квадрат распределению.

Воспользуемся Теоремой и записанным ранее утверждением. Запишем критическую область:

$$P(X_N^2 > t_\alpha \mid H_0) \approx 1 - F_{N-1}(t_\alpha),$$

где $F_{N-1}(t)$ — функция распределения χ^2_{N-1} .

Тогда, полагая $\alpha = 1 - F_{N-1}(t_\alpha)$, получим:

$$t_\alpha = \chi^2_{1-\alpha, N-1},$$

где $\chi^2_{1-\alpha, N-1}$ — квантиль уровня $1 - \alpha$ случайной величины χ^2_{N-1} .

$$\mathcal{X}_\alpha = \left\{ \bar{x} : \sum_{i=1}^N \frac{(\nu_i - np_i)^2}{np_i} > \chi^2_{N-1, 1-\alpha} \right\}.$$

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

Таким образом: гипотеза H_0 отвергается тогда и только тогда, когда $X_n^2 > \chi^2_{1-\alpha, N-1}$, где α — заданный уровень значимости.

Таблица для выбранных значений уровней значимости α и соответствующих им квантилей уровня $1 - \alpha$:

α	0,10	0,05	0,01
$\chi^2_{1-\alpha, N-1}$	92.17	97.35	107.58

Результаты применения критерия хи-квадрат:

n = 5			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 10			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 100			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 200			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 400			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 600			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H1	H1	H1
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 800			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H1	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 1000			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H1	H1	H1
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

11.2.2 Распределение Лапласа

Область значений непрерывной случайной величины ξ разбивается на N непересекающихся интервалов $\Delta_i, i = \{1, N\}$. Далее рассматривается случайная величина η , принимающая значения из $\{1, \dots, N\}$ с вероятностями $P(\xi \in \Delta_1), \dots, P(\xi \in \Delta_N)$ соответственно.

Разобьем множество значений, которые принимает случайная величина из распределения с ненулевыми вероятностями \mathbb{R} на равновероятные отрезки, их количество - N . Будем проверять значения $N \in \{3, 5, 10\}$.

Формула квантиля k_γ уровня γ :

$$k_\gamma = \begin{cases} \mu + \frac{\ln 2\gamma}{\theta}, & \gamma < \frac{1}{2}; \\ \mu - \frac{\ln 2(1-\gamma)}{\theta}, & \gamma \geq \frac{1}{2}. \end{cases}$$

Определим отрезки Δ_i для каждого N :

$$N = 3 \Rightarrow P(\xi \in \Delta_i) = \frac{1}{3}$$

$$\begin{cases} k_{\frac{1}{3}} = 21.95 \\ k_{\frac{2}{3}} = 22.05 \end{cases} \Rightarrow \begin{cases} \Delta_1 = (-\infty, 21.95) \\ \Delta_2 = [21.95, 22.05] \\ \Delta_3 = (22.05, +\infty) \end{cases}$$

$$\mathbf{N} = \mathbf{5} \Rightarrow P(\xi \in \Delta_i) = \frac{1}{5}$$

$$\begin{cases} k_{\frac{1}{5}} = 21.88 \\ k_{\frac{2}{5}} = 21.97 \\ k_{\frac{3}{5}} = 22.03 \\ k_{\frac{4}{5}} = 22.12 \end{cases} \Rightarrow \begin{cases} \Delta_1 = (-\infty, 21.88) \\ \Delta_2 = [21.88, 21.97) \\ \Delta_3 = [21.97, 22.03) \\ \Delta_4 = [22.03, 22.12) \\ \Delta_5 = [22.12, +\infty) \end{cases}$$

$$\mathbf{N} = \mathbf{10} \Rightarrow P(\xi \in \Delta_i) = \frac{1}{10}$$

$$\begin{cases} k_{\frac{1}{10}} = 21.79 \\ k_{\frac{2}{10}} = 21.88 \\ k_{\frac{3}{10}} = 21.93 \\ k_{\frac{4}{10}} = 21.97 \\ k_{\frac{5}{10}} = 22.0 \\ k_{\frac{6}{10}} = 22.03 \\ k_{\frac{7}{10}} = 22.07 \\ k_{\frac{8}{10}} = 22.12 \\ k_{\frac{9}{10}} = 22.21 \end{cases} \Rightarrow \begin{cases} \Delta_1 = (-\infty, 21.79) \\ \Delta_2 = [21.79, 21.88) \\ \Delta_3 = [21.88, 21.93) \\ \Delta_4 = [21.93, 21.97) \\ \Delta_5 = [21.97, 22.0) \\ \Delta_6 = [22.0, 22.03) \\ \Delta_7 = [22.03, 22.07) \\ \Delta_8 = [22.07, 22.12) \\ \Delta_9 = [22.12, 22.21) \\ \Delta_{10} = [22.21, +\infty) \end{cases}$$

Квантили распределения кси-квадрат:

$$\chi^2_{1-0.1,3-1} = 4.61, \quad \chi^2_{1-0.05,3-1} = 5.99, \quad \chi^2_{1-0.01,3-1} = 9.21,$$

$$\chi^2_{1-0.1,5-1} = 7.78, \quad \chi^2_{1-0.05,5-1} = 9.49, \quad \chi^2_{1-0.01,5-1} = 13.28,$$

$$\chi^2_{1-0.1,10-1} = 14.68, \quad \chi^2_{1-0.05,10-1} = 16.92, \quad \chi^2_{1-0.01,10-1} = 21.67.$$

Значения статистик Пирсона для каждой из сгенерированных выборок:
 Таблица для $N = 3$

n	1	2	3	4	5
5	0.4	0.4	5.2	0.4	0.4
10	3.2	3.8	0.8	2.6	2.6
100	0.86	2.06	1.04	0.38	2.18
200	2.29	0.16	0.25	5.89	1.72
400	3.1	2.34	6.64	1.62	0.39
600	7.03	2.71	5.49	5.53	0.09
800	0.52	11.83	0.03	0.57	2.12
1000	6.97	2.28	1.93	0.34	3.37

Таблица для $N = 5$

n	1	2	3	4	5
5	8.0	0.0	4.0	2.0	6.0
10	3.0	2.0	3.0	7.0	2.0
100	2.5	2.0	0.9	1.1	2.2
200	0.5	1.6	3.1	4.35	0.9
400	1.15	6.7	3.92	7.62	1.88
600	6.38	2.73	5.55	2.53	5.12
800	2.95	14.98	3.19	0.81	0.64
1000	1.83	2.66	9.95	6.93	1.31

Таблица для $N = 10$

n	1	2	3	4	5
5	17.0	5.0	5.0	9.0	17.0
10	8.0	4.0	14.0	16.0	8.0
100	12.2	5.2	1.8	4.8	3.2
200	2.1	11.0	7.5	7.4	4.1
400	2.65	11.1	13.8	13.75	16.35
600	10.57	10.3	19.7	7.23	8.2
800	15.55	18.6	5.8	5.6	2.95
1000	15.78	12.88	18.48	13.86	6.5

Результаты применения критерия хи-квадрат для $N \in \{3, 5, 10\}$:

n = 5 :

N = 3			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H1	H0	H0
4	H0	H0	H0
5	H0	H0	H0

N = 5			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H1	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

N = 10			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H1	H1	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H1	H1	H0

n = 10 :

N = 3			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

N = 5			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

N = 10			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 100 :

N = 3			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

N = 5			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

N = 10			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 200 :

N = 3			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H1	H0	H0
5	H0	H0	H0

N = 5			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

N = 10			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H1	H0	H0

n = 400 :

N = 3			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H1	H1	H0
2	H0	H0	H0
3	H1	H0	H0
4	H1	H0	H0
5	H0	H0	H0

N = 5			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

N = 10			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H1	H1	H0
4	H0	H0	H0
5	H0	H0	H0

n = 800 :

N = 3			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H1	H1	H1
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

N = 5			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H1	H1	H1
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

N = 10			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H1	H0	H0
2	H1	H1	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 1000 :

N = 3			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H1	H1	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

N = 5			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H1	H1	H0
4	H0	H0	H0
5	H0	H0	H0

N = 10			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H1	H0	H0
2	H0	H0	H0
3	H1	H1	H0
4	H0	H0	H0
5	H0	H0	H0

11.3 Критерий согласия Колмогорова для сложной гипотезы

11.3.1 Распределение Лапласа

Критерий согласия Колмогорова может быть применен в случае сложной гипотезы в условиях, когда неизвестен параметр распределения, аналогично случаю простой гипотезы. Однако в случае сложных гипотез распределение $D_n(\theta)$, зависит как от вида априорных распределений, так и от способа получения оценок, размера выборки n , вида Θ .

Поэтому статистику D_n необходимо заменить на статистику \widehat{D}_n :

$$\widehat{D}_n = \widehat{D}_n(X) = \sup_{x \in \mathbb{R}} |\widehat{F}_n(x) - F_{\hat{\theta}}(x)|$$

Под $F_{\hat{\theta}}(x)$ в приведенной формуле подразумевается функция распределения $L(\hat{\theta}, 22)$, где оценка $\hat{\theta}$ находится по методу максимального правдоподобия.

Как было найдено ранее, $\hat{\theta} = \frac{n}{\sum_{i=1}^n |X_i - 22|}$.

Также отметим, что в случае достаточного объема выборки ($n \geq 40$) можно разбить ее на 2 равные части, одну из которых использовать для нахождения $\hat{\theta}$, а вторую — для применения критерия.

Результат применения критерия:

n = 5			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H1	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 10			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 100			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H1	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 200			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H1	H0	H0
5	H0	H0	H0

n = 400			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 600			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H1	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 800			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 1000			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

11.3.2 Дискретное равномерное I

Как и в случае простых гипотез для применения критерия согласия Колмогорова воспользуемся случайной величиной U_i , которую будем вычислять согласно следующей формуле:

$$U_i = \lim_{z \rightarrow 0} F_{\hat{\theta}}(X_i - z) + Y_i \left[F_{\hat{\theta}}(X_i) - \lim_{z \rightarrow 0} F_{\hat{\theta}}(X_i - z) \right]$$

Под $F_{\hat{\theta}}(x)$ в приведенной формуле подразумевается функция распределения $R\{1, \hat{\theta}\}$, где оценка $\hat{\theta}$ находится по методу максимального правдоподобия.

Как было найдено ранее, $\hat{\theta} = X_{(n)}$.

Результат применения критерия:

n = 5			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H1	H0	H0
4	H1	H1	H0
5	H0	H0	H0

n = 10			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H1	H1	H0

n = 100			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 200			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 400			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H1	H0	H0

n = 600			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 800				n = 1000			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$		$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0	1	H0	H0	H0
2	H0	H0	H0	2	H0	H0	H0
3	H0	H0	H0	3	H0	H0	H0
4	H0	H0	H0	4	H0	H0	H0
5	H0	H0	H0	5	H0	H0	H0

11.4 Критерий согласия хи-квадрат для сложной гипотезы

11.4.1 Дискретное равномерное I

Критерий согласия хи-квадрат может быть применен для сложной гипотезы в условиях, когда неизвестен параметр распределения. Для этого зафиксируем неизвестный параметр θ при помощи метода максимального правдоподобия.

Ранее было найдено:

$$\hat{\theta} = \hat{\theta}_{\text{м.м.п.}} = X_{(n)}$$

Подход аналогичен случаю с простой гипотезы, но стоит учесть, что статистика Пирсона теперь имеет следующий вид:

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

Также при увеличении объема выборки распределение статистики $\hat{\chi}_N^2$ при условии гипотезы H_0 сходится к распределению χ_{N-2}^2 .

Зададим критическую область фиксированного уровня значимости α :

$$\hat{\mathcal{X}}_\alpha = \left\{ \bar{x} : \sum_{i=1}^N \frac{(\nu_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})} > \chi_{N-2, 1-\alpha}^2 \right\}.$$

Результат применения критерия:

n = 5			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 10			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 100			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 200			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 400			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 600			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H1	H1	H1
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 800			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H1	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 1000			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H1	H1	H1
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

11.4.2 Распределение Лапласа

Критерий согласия хи-квадрат в случае сложной гипотезы в условиях, когда неизвестен параметр распределения, также применим.

Аналогично случаю с простой гипотезы, разобьем область значений случайной величины на $N \in \{3, 5, 10\}$ равновероятных отрезков.

Учтем параметр, найденный ранее:

$$\hat{\theta} = \frac{n}{\sum_{i=1}^n |X_i - 22|}.$$

Квантили распределения кси-квадрат:

$$\chi^2_{1-0.1,3-2} = 2.71, \quad \chi^2_{1-0.05,3-2} = 3.84, \quad \chi^2_{1-0.01,3-2} = 6.63,$$

$$\chi^2_{1-0.1,5-2} = 6.25, \quad \chi^2_{1-0.05,5-2} = 7.81, \quad \chi^2_{1-0.01,5-2} = 11.34,$$

$$\chi^2_{1-0.1,10-2} = 13.36, \quad \chi^2_{1-0.05,10-2} = 15.51, \quad \chi^2_{1-0.01,10-2} = 20.01$$

Результат применения критерия:

n = 5 :

N = 3			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

N = 5			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

N = 10			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H1	H1	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

n = 10 :

N = 3			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

N = 5			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H1	H0	H0
5	H1	H1	H0

N = 10			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H1	H1	H1
2	H1	H1	H0
3	H1	H1	H0
4	H1	H1	H1
5	H1	H1	H1

n = 100 :

N = 3			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

N = 5			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H1	H1	H1
2	H1	H1	H1
3	H1	H1	H1
4	H1	H1	H1
5	H1	H1	H1

N = 10			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H1	H1	H1
2	H1	H1	H1
3	H1	H1	H1
4	H1	H1	H1
5	H1	H1	H1

n = 200 :

N = 3			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

N = 5			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H1	H1	H1
2	H1	H1	H1
3	H1	H1	H1
4	H1	H1	H1
5	H1	H1	H1

N = 10			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H1	H1	H1
2	H1	H1	H1
3	H1	H1	H1
4	H1	H1	H1
5	H1	H1	H1

n = 400 :

N = 3			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

N = 5			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H1	H1	H1
2	H1	H1	H1
3	H1	H1	H1
4	H1	H1	H1
5	H1	H1	H1

N = 10			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H1	H1	H1
2	H1	H1	H1
3	H1	H1	H1
4	H1	H1	H1
5	H1	H1	H1

n = 600 :

N = 3			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

N = 5			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H1	H1	H1
2	H1	H1	H1
3	H1	H1	H1
4	H1	H1	H1
5	H1	H1	H1

N = 10			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H1	H1	H1
2	H1	H1	H1
3	H1	H1	H1
4	H1	H1	H1
5	H1	H1	H1

n = 800 :

N = 3			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H0	H0	H0
2	H0	H0	H0
3	H0	H0	H0
4	H0	H0	H0
5	H0	H0	H0

N = 5			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H1	H1	H1
2	H1	H1	H1
3	H1	H1	H1
4	H1	H1	H1
5	H1	H1	H1

N = 10			
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1	H1	H1	H1
2	H1	H1	H1
3	H1	H1	H1
4	H1	H1	H1
5	H1	H1	H1

12 Блок №2. Проверка гипотезы об однородности выборок

Пусть $X = (X_1, \dots, X_n)$ из распределения $\mathcal{L}(\xi)$ с неизвестной функцией распределения $F_1(x)$ и $Y = (Y_1, \dots, Y_n)$ из распределения $\mathcal{L}(\eta)$ также с неизвестной функцией распределения $F_2(x)$. Гипотеза однородности формулируется следующим образом $H_0 : F_1(x) = F_2(x)$ и заключается в проверке гипотезы о том, что рассматриваются две выборки из одного и того же распределения.

12.1 Распределение Лапласа

Критерий однородности Смирнова

Пусть X, Y — две выборки объема n и m соответственно, \hat{F}_{1n} — эмпирическая функция распределения, построенная по выборке X , \hat{F}_{2m} — эмпирическая функция распределения, построенная по выборке Y . Рассмотрим статистику:

$$D_{n,m} = \sup_{x \in \mathbb{R}} |\hat{F}_{1n}(x) - \hat{F}_{2m}(x)|$$

В случае если F_1 и F_2 — непрерывные функции распределения, то по теореме Смирнова статистика

$$\sqrt{\frac{n \cdot m}{n + m}} D_{n,m}$$

имеет распределение Колмогорова. Тогда критерий проверки гипотезы однородности можно сформулировать следующим образом: если $D_{n,m} > t_\alpha(n, m)$, то гипотезу H_0 отвергаем, где

$$t_\alpha(n, m) = \sqrt{\frac{1}{n} + \frac{1}{m}} t_\alpha, \quad K(t_\alpha) = 1 - \alpha.$$

При этом:

$$\mathbb{P} \left(D_{n,m} > \sqrt{\frac{1}{n} + \frac{1}{m}} t_\alpha \middle| H_0 \right) = \mathbb{P} \left(\sqrt{\frac{m \cdot n}{n + m}} D_{n,m} > t_\alpha \middle| H_0 \right) \underset{n, m \rightarrow \infty}{\simeq} 1 - K(\lambda_\alpha) = \alpha.$$

Возьмем $\alpha = 0.05$, тогда $\lambda_\alpha \approx 1.36$. Применим критерий.

Результаты для $n_1 = 5, n_2 = 10$

Пара выборок	D_{nm}	t_α	Принимаем H_0
Выборка 1 vs Выборка 1	0.4	0.744903	True
Выборка 1 vs Выборка 2	0.5	0.744903	True
Выборка 1 vs Выборка 3	0.3	0.744903	True
Выборка 1 vs Выборка 4	0.5	0.744903	True
Выборка 1 vs Выборка 5	0.4	0.744903	True
Выборка 2 vs Выборка 1	0.3	0.744903	True
Выборка 2 vs Выборка 2	0.3	0.744903	True
Выборка 2 vs Выборка 3	0.3	0.744903	True
Выборка 2 vs Выборка 4	0.5	0.744903	True
Выборка 2 vs Выборка 5	0.3	0.744903	True
Выборка 3 vs Выборка 1	0.6	0.744903	True
Выборка 3 vs Выборка 2	0.7	0.744903	True
Выборка 3 vs Выборка 3	0.4	0.744903	True
Выборка 3 vs Выборка 4	0.3	0.744903	True
Выборка 3 vs Выборка 5	0.4	0.744903	True
Выборка 4 vs Выборка 1	0.4	0.744903	True
Выборка 4 vs Выборка 2	0.3	0.744903	True
Выборка 4 vs Выборка 3	0.5	0.744903	True
Выборка 4 vs Выборка 4	0.5	0.744903	True
Выборка 4 vs Выборка 5	0.3	0.744903	True
Выборка 5 vs Выборка 1	0.3	0.744903	True
Выборка 5 vs Выборка 2	0.4	0.744903	True
Выборка 5 vs Выборка 3	0.2	0.744903	True
Выборка 5 vs Выборка 4	0.4	0.744903	True
Выборка 5 vs Выборка 5	0.4	0.744903	True

Часть V

Различение статистических гипотез

13 Блок №1. Вычисление функции отношения правдоподобия

Рассмотрим распределение Лапласа и следующие гипотезы:

$$\mathcal{F}_0 = L(\theta_1, \mu)$$

$$\mathcal{F}_1 = L(\theta_2, \mu),$$

где $\mu = 22$.

Причем для определенности будем считать $\theta_2 > \theta_1$. Отметим, что выбор из двух простых гипотез можно представить в виде параметрической гипотезы:

Пусть $\Theta = \{0, 1\}$; $F_\theta(x) = (1 - \theta)F_0(x) + \theta F_1(x)$.

В случае, когда H_0 и H_1 — простые гипотезы, вводятся понятия ошибок критерия. Критерий описывается вероятностями:

1. $\mathbb{P}(X \in \mathcal{X}_1 | H_0) = \alpha$ — ошибка первого рода;

2. $\mathbb{P}(X \in \mathcal{X}_0 | H_1) = \beta$ — ошибка второго рода.

Функцией мощности критерия W назовем функционал на множестве допустимых распределений \mathcal{F} и выборке X :

$$W(F_X) = W(F_X; \mathcal{X}_{1,\alpha}) = \mathbb{P}(X \in \mathcal{X}_{1,\alpha} | F_X),$$

где $\mathbb{P}(x \in \mathcal{X}_{1,\alpha} | F_X)$ — вероятность попасть в $\mathcal{X}_{1,\alpha}$, если F_X — истинное распределение.

В случае параметрических гипотез функция мощности может быть представлена в виде:

$$W(\theta) = W(\theta, \mathcal{X}_{1,\alpha}) = P_\theta(X \in \mathcal{X}_{1,\alpha}).$$

Через функцию мощности критерия легко можно выразить вероятности ошибок первого и второго рода:

$$\alpha = W(F_X), \quad \beta = 1 - W(F_X).$$

Определение. Функция, имеющая вид:

$$l(\bar{x}) = \frac{L(\bar{x}, \theta_1)}{L(\bar{x}, \theta_0)} = \frac{\prod_{i=1}^n f_1(x_i)}{\prod_{i=1}^n f_0(x_i)}$$

называется функцией отношения правдоподобия.

Описание критерия отношения правдоподобия.

Определение. Критическим множеством критерия Неймана-Пирсона называется множество $\mathcal{X}_{1,\alpha}^*$, имеющее вид:

$$\mathcal{X}_{1,\alpha}^* = \{\bar{x} \in \mathcal{X} : l(\bar{x}) \geq c_\alpha\},$$

где c_α такое, что ошибка 1 рода равна α .

Предположим далее, что $f_i(x) > 0$, так как из $f_i(x) = 0$ следует $L(x_i, \theta_0) = 0$, что говорит о верности другой гипотезы. Здесь и далее для сокращения записи вместо \mathbb{P}_{θ_i} будем писать \mathbb{P}_i . Определим вспомогательную функцию $\varphi(c)$:

$$\varphi(c) = \mathbb{P}_0(l(\bar{x}) \geq c) = \int_{\bar{x}:l(\bar{x}) \geq c} L(\bar{x}, \theta_0) d\bar{x}.$$

Заметим, что при $c = 0$ функция принимает значение равное $\varphi(0) = 1$. Покажем, что чем больше c , тем меньше значение $\varphi(c)$:

$$1 \geq \mathbb{P}_1(l(\bar{x}) \geq c) = \int_{\bar{x}:l(\bar{x}) \geq c} L(\bar{x}, \theta_1) d\bar{x} \geq c \cdot \int_{\bar{x}:l(\bar{x}) \geq c} L(\bar{x}, \theta_0) d\bar{x} = c \cdot \varphi(c).$$

Следовательно:

$$\varphi(c) \leq \frac{1}{c}.$$

Отсюда следует, что при $c \rightarrow \infty$ значение $\varphi(c) \rightarrow 0$. Предположим, что $\varphi(c)$ — непрерывная функция. Тогда для любого $\alpha \in (0; 1)$ можно найти такое c_α , что $\varphi(c_\alpha) = \alpha$. Соответственно для множества $\mathcal{X}_{1,\alpha}^* = \{\bar{x} : l(\bar{x}) \geq c_\alpha\}$ верно равенство:

$$W(\theta_0, \mathcal{X}_{1,\alpha}^*) = \mathbb{P}_0(\mathcal{X}_{1,\alpha}^*) = \varphi(c_\alpha) = \alpha.$$

Теорема (Лемма Неймана-Пирсона). Пусть для фиксированного α существует такое c_α , что $\varphi(c_\alpha) = \alpha$. Тогда критическая область $\mathcal{X}_{1,\alpha}^* = \{\bar{x} : l(\bar{x}) \geq c_\alpha\}$ задает наиболее мощный критерий для гипотезы $H_0 : F_\xi = F_0$ относительно альтернативы $H_1 : F_\xi = F_1$ среди всех критериев с уровнем значимости α .

13.1 Распределение Лапласа

Вычисление функции отношения правдоподобия

Функция правдоподобия имеет следующий вид:

$$L(\bar{x}, \theta) = \prod_{i=1}^n f(x_i) = \prod_{i=1}^n \frac{\theta}{2} * \exp\{-\theta|x_i - \mu|\} = \frac{\theta^n}{2^n} * \prod_{i=1}^n \exp\{-\theta|x_i - \mu|\}.$$

Тогда функция отношения правдоподобия:

$$\begin{aligned} l(\bar{x}) &= \frac{L(\bar{x}, \theta_1)}{L(\bar{x}, \theta_0)} = \frac{\frac{\theta_1^n}{2^n} \prod_{i=1}^n \exp\{-\theta_1|x_i - \mu|\}}{\frac{\theta_0^n}{2^n} \prod_{i=1}^n \exp\{-\theta_0|x_i - \mu|\}} = \frac{\theta_1^n}{\theta_0^n} \cdot \frac{\prod_{i=1}^n \exp\{-\theta_1|x_i - \mu|\}}{\prod_{i=1}^n \exp\{-\theta_0|x_i - \mu|\}} = \\ &= \left(\frac{\theta_1}{\theta_0}\right)^n \prod_{i=1}^n \frac{\exp\{-\theta_1|x_i - \mu|\}}{\exp\{-\theta_0|x_i - \mu|\}} = \left(\frac{\theta_1}{\theta_0}\right)^n \prod_{i=1}^n \exp\{-(\theta_1 - \theta_0)|x_i - \mu|\}. \end{aligned}$$

Получили итоговое выражение:

$$l(\bar{x}) = \left(\frac{\theta_1}{\theta_0}\right)^n \exp\left\{-(\theta_1 - \theta_0) \sum_{i=1}^n |x_i - \mu|\right\}.$$

Вычисление критической области

Основываясь на критерий Неймана-Пирсона (гипотеза H_0 отвергается, если функция отношения правдоподобия превышает некоторый порог c), критическую область определим следующим образом:

$$l(X) \geq c \iff \ln l(X) \geq \ln c.$$

Значение $\ln l(X)$:

$$\ln l(X) = n \ln \left(\frac{\theta_1}{\theta_0}\right) - (\theta_1 - \theta_0) \sum_{i=1}^n |X_i - \mu|.$$

Таким образом, условие становится:

$$n \ln \left(\frac{\theta_1}{\theta_0} \right) - (\theta_1 - \theta_0) \sum_{i=1}^n |X_i - \mu| \geq \ln c.$$

Рассмотрим упрощение для суммы $\sum_{i=1}^n |X_i - \mu|$. Мы знаем, что статистика:

$$T = \frac{\sum_{i=1}^n |X_i - \mu|}{n},$$

является оценкой абсолютного среднего отклонения для распределения Лапласа.

Условие принимает вид:

$$\sum_{i=1}^n |X_i - \mu| \leq \frac{n}{\theta_1 - \theta_0} \left(\ln \left(\frac{\theta_1}{\theta_0} \right) - \frac{\ln c}{n} \right).$$

С учетом перехода к среднему отклонению T :

$$T \leq \frac{1}{\theta_1 - \theta_0} \left(\ln \left(\frac{\theta_1}{\theta_0} \right) - \frac{\ln c}{n} \right) = k.$$

Где k - критическое значение, зависящее от порога c , объема выборки n , и значений θ_0, θ_1 .

Рассмотрим нормализацию среднего отклонения T . По центральной предельной теореме, для больших n , T будет асимптотически нормально распределена:

$$T \sim N \left(E(|X - \mu|), \frac{1}{n} D(|X - \mu|) \right) \sim N \left(\frac{1}{\theta}, \frac{1}{n\theta^2} \right).$$

Нормируем T , получаем стандартное нормальное распределение:

$$Z = \sqrt{n}\theta_1(T - \frac{1}{\theta_1}), \quad Z \sim N(0, 1).$$

$$\phi(c) = P_0(l(\bar{X}) \geq c) = P_0 \left(\sqrt{n}\theta_1 \left(T - \frac{1}{\theta_1} \right) \geq t(c) \right),$$

Так как $t(c)$ является непрерывной функцией для $c > 0$, функция $\phi(c)$ также непрерывна. В соответствии с теорией, для любого уровня значимости $\alpha \in (0, 1)$,

существует критическое значение c_α , такое что:

$$\phi(c_\alpha) = F(-t(c_\alpha)) = \alpha.$$

Таким образом, критическая область может быть записана как:

$$\mathcal{X}_{1,\alpha} = \left\{ \bar{x} : \sqrt{n}\theta_1 \left(\frac{\sum_{i=1}^n |X_i - \mu|}{n} - \frac{1}{\theta_1} \right) \geq t_\alpha \mid \bar{x} \in \mathcal{X} \right\},$$

где

- t_α — (-1) α -квантиль стандартного нормального распределения $N(0, 1)$.

Найдем ошибку второго рода:

$$\beta = P_1 \left(\sqrt{n}\theta_1 \left(T - \frac{1}{\theta_1} \right) < t_\alpha \right).$$

Воспользуемся центрированием и нормированием T . Из ЦПТ для распределения T под гипотезой H_1 :

$$T \sim N \left(\frac{1}{\theta_1}, \frac{1}{n\theta_1^2} \right).$$

Нормируем T так, чтобы распределение стало стандартным нормальным:

$$Z = \sqrt{n}\theta_1 \left(T - \frac{1}{\theta_1} \right) \sim N(0, 1); \quad \beta = P_1 \left(\sqrt{n}\theta_1 \left(T - \frac{1}{\theta_1} \right) < t_\alpha \right).$$

Перепишем через стандартное нормальное распределение

$$\beta = P_1 \left(\xi < t_\alpha - \sqrt{n} \frac{\theta_1 - \theta_0}{\theta_1} \right),$$

где $\xi \sim N(0, 1)$.

Ошибка второго рода β принимает вид:

$$\beta = F \left(t_\alpha - \sqrt{n} \frac{\theta_1 - \theta_0}{\theta_1} \right),$$

где:

- $F(\cdot)$ — функция распределения стандартного нормального распределения $N(0, 1)$,

- t_α — α -квантиль $N(0, 1)$,
- n — объем выборки.

Минимальное количество материала

В данном разделе займемся подсчетом минимального необходимого количества материала при фиксации наименьших допустимых значений ошибок первого и второго рода.

Пусть $n' = n'(\alpha', \beta')$ - искомая величина, т.е. минимальное необходимое количество материала, при котором ошибочные гипотезы принимаются с вероятностями, не превышающими α' и β' .

Обозначим ζ_γ - γ -квантиль распределения $N(0, 1)$. Тогда согласно предыдущим вычислениям:

$$t_{\alpha'} = -\zeta_{\alpha'} \text{ и } t_{\alpha'} - \frac{\sqrt{n}(\theta_1 - \theta_0)}{\theta_1} = \zeta_{\beta'}.$$

Следовательно:

$$n' = \frac{\theta_1^2}{(\theta_1 - \theta_0)^2} (\zeta_{\alpha'} + \zeta_{\beta'})^2.$$

Очевидно, что $n' \in \mathbb{N} \Rightarrow$ полученный результат нужно округлить. Заметим, что функция распределения $N(0, 1)$ - неубывающая \Rightarrow чем больше n , тем меньше $\beta \Rightarrow$ чтобы получаемая при количестве материала n' ошибка 2-го рода не превосходила β' , необходимо округлить полученное значение вверх, т.е.:

$$n' = \left\lceil \frac{\theta_1^2}{(\theta_1 - \theta_0)^2} (\zeta_{\alpha'} + \zeta_{\beta'})^2 \right\rceil.$$

13.2 Дискретное равномерное I

Для дискретного равномерного распределения на $\{1, \dots, \theta\}$, функция плотности вероятности задается как:

$$f(x; \theta) = \begin{cases} \frac{1}{\theta}, & x \in \{1, 2, \dots, \theta\}, \\ 0, & \text{иначе.} \end{cases}$$

Функция правдоподобия

Пусть у нас есть выборка $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$. Тогда функция правдоподобия для гипотезы $\theta = \theta_0$ записывается как:

$$L(\mathbf{x}; \theta) = \prod_{i=1}^n f(x_i; \theta).$$

Так как $f(x; \theta) = \frac{1}{\theta}$ для всех $x_i \leq \theta$, функция правдоподобия равна:

$$L(\mathbf{x}; \theta) = \begin{cases} \theta^{-n}, & \text{если } \max(x_i) \leq \theta, \\ 0, & \text{иначе.} \end{cases}$$

Функция отношения правдоподобия

Функция отношения правдоподобия для двух гипотез θ_1 и θ_0 задается как:

$$l(\mathbf{x}) = \frac{L(\mathbf{x}; \theta_1)}{L(\mathbf{x}; \theta_0)}.$$

Подставим $L(\mathbf{x}; \theta)$ в это выражение:

$$l(\mathbf{x}) = \begin{cases} \left(\frac{\theta_0}{\theta_1}\right)^n, & \text{если } \max(x_i) \leq \min(\theta_0, \theta_1), \\ 0, & \text{если } \max(x_i) > \max(\theta_0, \theta_1). \end{cases}$$

Критическая область

Критическая область $\mathcal{X}_{1,\alpha}$ определяется как множество значений выборки, для которых отношение правдоподобия меньше некоторого порога c :

$$\mathcal{X}_{1,\alpha} = \{\mathbf{x} : l(\mathbf{x}) \leq c\}.$$

Подставим выражение для $\ell(\mathbf{x})$:

$$\ell(\mathbf{x}) \leq c \implies \left(\frac{\theta_0}{\theta_1}\right)^n \leq c.$$

Отсюда:

$$\ln \ell(\mathbf{x}) = n \ln \left(\frac{\theta_0}{\theta_1}\right) \leq \ln c.$$

Условия для критической области

1. Если $\max(\mathbf{x}) > \max(\theta_0, \theta_1)$, то $\ell(\mathbf{x}) = 0$, и выборка автоматически принадлежит критической области, если $c > 0$.
2. Если $\max(\mathbf{x}) \leq \min(\theta_0, \theta_1)$, критерий сравнивает $\left(\frac{\theta_0}{\theta_1}\right)^n$ с c .

Тогда для выборки \mathbf{x} критическая область $\mathcal{X}_{1,\alpha}$ определяется как:

$$\mathcal{X}_{1,\alpha} = \{x : \max(x) > k_\alpha\}$$

Критическая область зависит от максимального значения выборки $\max(\mathbf{x})$.

Если максимальное значение превышает θ_1 , гипотеза H_0 отвергается.

$k_\alpha = \lfloor \theta_0(1-\alpha) \rfloor$ - определяет максимальное значение выборки, при котором гипотеза H_0 отвергается.

Найдем ошибку второго рода.

1. **Определение распределения $\max(\mathbf{x})$ под H_1 :**

Для дискретного равномерного распределения с параметром θ_1 , максимальное значение выборки $\max(\mathbf{x})$ из n независимых наблюдений имеет распределение:

$$P(\max(\mathbf{x}) \leq k) = \left(\frac{k}{\theta_1}\right)^n,$$

где $k \in \{1, 2, \dots, \theta_1\}$.

2. **Вычисление β :**

Ошибка второго рода равна вероятности того, что $\max(\mathbf{x}) \leq k_\alpha$ под гипотезой H_1 . Это записывается как:

$$\beta = P(\max(\mathbf{x}) \leq k_\alpha | H_1) = \left(\frac{k_\alpha}{\theta_1}\right)^n,$$

где $k_\alpha = \lfloor \theta_0(1 - \alpha) \rfloor$.

3. Итоговая формула:

Подставляя k_α , получаем:

$$\beta = \left(\frac{\lfloor \theta_0(1 - \alpha) \rfloor}{\theta_1} \right)^n.$$

Минимальное количество материала:

$$n' = \left\lceil \frac{\ln(\beta')}{\ln \left(\frac{\lfloor \theta_0(1 - \alpha') \rfloor}{\theta_1} \right)} \right\rceil$$

Часть VI

Ссылки

Ссылка на файлы