Национальный исследовательский университет
Высшая школа экономики
Московский институт электроники и математики

Департамент прикладной математики
кафедра компьютерной безопасности

Домашнее задание №4 по математической статистике
Проверка статистических гипотез

Дискретное распределение: *дискретное равномерное I*
Неизвестный параметр: $\theta = 121$
Непрерывное распределение: *распределение Парето*
Неизвестный параметр: $\theta = 12$

Выполнила
Мазитова Е.А.

Проверил
Богданов Д.С.

Москва 2025

Содержание

1. Проверка гипотезы о виде распределения

Дискретное распределение (дискретное равномерное I)

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

1. **Критерий согласия хи-квадрат**

   Для каждой выборки вычисляется статистика хи-квадрат с предварительной группировкой значений:

   (a) **Определение числа интервалов**: По правилу Старджесса

   $$k = 1 + \lfloor \log_2 n \rfloor$$

   где n — объем выборки.

   (b) **Разбиение области значений**: Исходная область $\{1, 2, \ldots, 121\}$ разбивается на k интервалов. Длина j-го интервала:

   $$L_j = \left\lfloor \frac{121}{k} \right\rfloor + \delta_j, \quad \delta_j = \begin{cases} 1, & j \leq (121 \bmod k) \\ 0, & \text{иначе} \end{cases}$$

   (c) **Наблюдаемые частоты**: Для каждого интервала $I_j = [a_j, b_j]$ вычисляется

   $$O_j = \#\{i : X_i \in I_j\}.$$

   (d) **Ожидаемые частоты**: При справедливости H_0 для выборки объема n:

   $$E_j = n \cdot \frac{L_j}{121}, \quad j = 1, \ldots, k.$$

   (e) **Статистика хи-квадрат**:

   $$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}.$$

   Правило проверки гипотезы

   Статистика χ^2 при справедливости H_0 имеет распределение хи-квадрат с числом степеней свободы:

   $$\mathrm{df} = k - 1.$$

Для уровня значимости $\alpha = 0.05$ критическое значение вычисляется как:

$$\chi^2_{k-1,0.95}.$$

Правило принятия решения:

$$\begin{cases} \chi^2 \leq \chi^2_{k-1,0.95} & \Rightarrow \text{принимаем } H_0, \\ \chi^2 > \chi^2_{k-1,0.95} & \Rightarrow \text{отвергаем } H_0. \end{cases}$$

Для каждой выборки, сгенерированной в пункте 2.1, приведем значения полученных $\chi^2_{\text{набл}}$, $\chi^2_{\text{крит}}$. Значения ниже были вычислены на основе сгенерированных в ДЗ2 выборок с помощью написания программы на языке программирования Python (см. приложение):

Таблица 1: Значения статистики хи-квадрат с группировкой по Старджессу ($\theta = 121$, $\alpha = 0.05$)

n	Серия	k	$\chi^2_{\text{набл}}$	$\chi^2_{\text{крит}}$	Вывод
5*10	1	4	3.51	7.81	Принимаем
	2	4	2.05	7.81	Принимаем
	3	4	1.98	7.81	Принимаем
	4	4	1.98	7.81	Принимаем
	5	4	1.98	7.81	Принимаем
5*100	1	7	10.92	12.59	Принимаем
	2	7	4.79	12.59	Принимаем
	3	7	7.31	12.59	Принимаем
	4	7	7.28	12.59	Принимаем
	5	7	9.34	12.59	Принимаем
5*200	1	8	5.50	14.07	Принимаем
	2	8	6.01	14.07	Принимаем
	3	8	4.88	14.07	Принимаем
	4	8	7.90	14.07	Принимаем
	5	8	4.09	14.07	Принимаем

Таблица 2: Значения статистики хи-квадрат с группировкой по Старджессу (продолжение)

n	Серия	k	$\chi^2_{\text{набл}}$	$\chi^2_{\text{крит}}$	Вывод
5*400	1	9	20.97	15.51	Отвергаем
	2	9	5.16	15.51	Принимаем
	3	9	5.32	15.51	Принимаем
	4	9	6.61	15.51	Принимаем
	5	9	6.35	15.51	Принимаем
5*600	1	10	8.55	16.92	Принимаем
	2	10	6.58	16.92	Принимаем
	3	10	5.52	16.92	Принимаем
	4	10	4.25	16.92	Принимаем
	5	10	9.06	16.92	Принимаем
5*800	1	10	6.71	16.92	Принимаем
	2	10	8.57	16.92	Принимаем
	3	10	8.06	16.92	Принимаем
	4	10	4.43	16.92	Принимаем
	5	10	10.93	16.92	Принимаем
5*1000	1	10	8.80	16.92	Принимаем
	2	10	9.35	16.92	Принимаем
	3	10	8.52	16.92	Принимаем
	4	10	6.43	16.92	Принимаем
	5	10	8.86	16.92	Принимаем

Таблица 3: Сводные результаты критерия хи-квадрат с группировкой

n	Среднее χ^2	Процент отклонений
10	2.30	0.0%
100	7.93	0.0%
200	5.68	0.0%
400	8.88	20.0%
600	6.79	0.0%
800	7.74	0.0%
1000	8.39	0.0%

Видим, что средние значения χ^2 близки к математическим ожиданиям соответствующих распределений: $\mathbb{E}[\chi^2_{k-1}] = k - 1$. Например, для $n = 1000$ ($k = 10$): $\mathbb{E}[\chi^2_9] = 9$, получено 8.39.

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

Для проверки сложной гипотезы о принадлежности выборки дискретному равномерному распределению с неизвестным параметром θ применяется критерий хи-квадрат:

(a) **Оценка параметра** θ: Для каждой проверяемой выборки параметр θ оценивается как максимальное значение выборки:

$$\hat{\theta} = \max\{X_1, X_2, \ldots, X_n\}$$

(b) **Определение числа интервалов**: По правилу Старджесса

$$k = 1 + \lfloor \log_2 n \rfloor$$

(c) **Разбиение на интервалы**: Область $\{1, 2, \ldots, \hat{\theta}\}$ разбивается на k приблизительно равных интервалов. Длина j-го интервала:

$$L_j = \left\lfloor \frac{\hat{\theta}}{k} \right\rfloor + \delta_j, \quad \delta_j = \begin{cases} 1, & j \leq (\hat{\theta} \bmod k) \\ 0, & \text{иначе} \end{cases}$$

(d) **Наблюдаемые и ожидаемые частоты**:

$$O_j = \#\{X_i \in I_j\}, \quad E_j = n \cdot \frac{L_j}{\hat{\theta}}$$

(e) **Статистика хи-квадрат**:

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}$$

Правило проверки гипотезы

При оценке одного параметра число степеней свободы уменьшается на единицу:

$$\mathrm{df} = k - 2.$$

Для уровня значимости $\alpha = 0.05$ критическое значение:

$$\chi^2_{\text{крит}} = \chi^2_{k-2,0.95}.$$

Правило принятия решения:

$$\begin{cases} \chi^2 \leq \chi^2_{\text{крит}} & \Rightarrow \text{принимаем } H_0, \\ \chi^2 > \chi^2_{\text{крит}} & \Rightarrow \text{отвергаем } H_0. \end{cases}$$

Для каждой выборки, сгенерированной в пункте 2.1, приведем значения полученных $\chi^2_{\text{набл}}$, $\chi^2_{\text{крит}}$. Значения ниже были вычислены на основе сгенерированных в ДЗ2 выборок с помощью написания программы на языке программирования Python (см. приложение):

Таблица 4: Значения статистики хи-квадрат для сложной гипотезы (дискретное равномерное, $\alpha = 0.05$)

n	Серия	k	$\hat{\theta}$	$\chi^2_{\text{набл}}$	$\chi^2_{\text{крит}}$	df	Вывод
5*10	1	4	120	6.80	5.99	2	Отвергаем
	2	4	116	2.00	5.99	2	Принимаем
	3	4	103	0.36	5.99	2	Принимаем
	4	4	106	4.53	5.99	2	Принимаем
	5	4	120	2.00	5.99	2	Принимаем
5*100	1	7	121	10.92	11.07	5	Принимаем
	2	7	121	4.79	11.07	5	Принимаем
	3	7	120	7.37	11.07	5	Принимаем
	4	7	121	7.28	11.07	5	Принимаем
	5	7	121	9.34	11.07	5	Принимаем
5*200	1	8	121	5.50	12.59	6	Принимаем
	2	8	121	6.01	12.59	6	Принимаем
	3	8	121	4.88	12.59	6	Принимаем
	4	8	121	7.90	12.59	6	Принимаем
	5	8	121	4.09	12.59	6	Принимаем

Таблица 5: Значения статистики хи-квадрат для сложной гипотезы (продолжение)

n	Серия	k	$\hat{\theta}$	$\chi^2_{\text{набл}}$	$\chi^2_{\text{крит}}$	df	Вывод
5*400	1	9	121	20.97	14.07	7	Отвергаем
	2	9	121	5.16	14.07	7	Принимаем
	3	9	121	5.32	14.07	7	Принимаем
	4	9	121	6.61	14.07	7	Принимаем
	5	9	121	6.35	14.07	7	Принимаем
5*600	1	10	121	8.55	15.51	8	Принимаем
	2	10	121	6.58	15.51	8	Принимаем
	3	10	121	5.52	15.51	8	Принимаем
	4	10	121	4.25	15.51	8	Принимаем
	5	10	121	9.06	15.51	8	Принимаем
5*800	1	10	121	6.71	15.51	8	Принимаем
	2	10	121	8.57	15.51	8	Принимаем
	3	10	121	8.06	15.51	8	Принимаем
	4	10	121	4.43	15.51	8	Принимаем
	5	10	121	10.93	15.51	8	Принимаем
5*1000	1	10	121	8.80	15.51	8	Принимаем
	2	10	121	9.35	15.51	8	Принимаем
	3	10	121	8.52	15.51	8	Принимаем
	4	10	121	6.43	15.51	8	Принимаем
	5	10	121	8.86	15.51	8	Принимаем

Таблица 6: Средние оценки параметра θ по объёму выборки

n	Среднее $\hat{\theta}$
10	113.0
100	120.8
200	121.0
400	121.0
600	121.0
800	121.0
1000	121.0

Таблица 7: Сводные результаты критерия хи-квадрат для сложной гипотезы

n	Среднее χ^2	Процент отклонений
10	3.14	20.0%
100	7.94	0.0%
200	5.68	0.0%
400	8.88	20.0%
600	6.79	0.0%
800	7.74	0.0%
1000	8.39	0.0%

Непрерывное распределение (распределение Парето)

$$f(x) = \theta \cdot x^{-(\theta+1)}, \quad x \in [1, +\infty), \quad \theta > 0$$

1. **Критерий согласия Колмогорова (Смирнова)**

   Посчитаем S — статистику Колмогорова с поправкой Большева:

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

   Правило проверки:

   - ▷ Критическое значение для уровня значимости $\alpha = 0.05$: $K_{0.95} \approx 1.358$
   - ▷ Если $S > 1.358$ — отвергаем гипотезу H_0 (выборка не из Парето с $\theta = 12$)
   - ▷ Если $S \leq 1.358$ — принимаем гипотезу H_0

   Для каждой выборки, сгенерированной в пункте 2.1, приведем значения полученных D_n, S и сравним с критическим значением. Значения ниже были вычислены на основе сгенерированных в ДЗ2 выборок с помощью написания программы на языке программирования Python (см. приложение):

Таблица 8: Значения статистик критерия Колмогорова для распределения Парето ($\theta = 12$)

n	Серия	D_n	S	Вывод
5*5	1	0.251260	0.636370	Принимаем
	2	0.505256	1.204323	Принимаем
	3	0.388123	0.942406	Принимаем
	4	0.240054	0.611313	Принимаем
	5	0.325217	0.801743	Принимаем
5*10	1	0.188297	0.648151	Принимаем
	2	0.396006	1.304985	Принимаем
	3	0.244872	0.827057	Принимаем
	4	0.240054	0.811823	Принимаем
	5	0.425217	**1.397360**	**Отвергаем**
5*100	1	0.064916	0.665828	Принимаем
	2	0.110848	1.125152	Принимаем
	3	0.108079	1.097458	Принимаем
	4	0.077751	0.794172	Принимаем
	5	0.066431	0.680978	Принимаем
5*200	1	0.060092	0.861608	Принимаем
	2	0.085823	1.225512	Принимаем
	3	0.051485	0.739897	Принимаем
	4	0.046189	0.665000	Принимаем
	5	0.082550	1.179213	Принимаем
5*400	1	0.043350	0.875332	Принимаем
	2	0.048757	0.983481	Принимаем
	3	0.033323	0.674797	Принимаем
	4	0.044727	0.902880	Принимаем
	5	0.056017	1.128670	Принимаем

Таблица 9: Значения статистик критерия Колмогорова для распределения Парето (продолжение)

n	Серия	D_n	S	Вывод
5*600	1	0.037564	0.926935	Принимаем
	2	0.028017	0.693072	Принимаем
	3	0.027403	0.678034	Принимаем
	4	0.041150	1.014762	Принимаем
	5	0.034702	0.856833	Принимаем
5*800	1	0.037100	1.055237	Принимаем
	2	0.027073	0.771645	Принимаем
	3	0.026272	0.748966	Принимаем
	4	0.031813	0.905696	Принимаем
	5	0.031755	0.904073	Принимаем
5*1000	1	0.022850	0.727849	Принимаем
	2	0.024135	0.768482	Принимаем
	3	0.028935	0.920272	Принимаем
	4	0.025446	0.809948	Принимаем
	5	0.028631	0.910656	Принимаем

Таблица 10: Сводные результаты критерия Колмогорова

n	Среднее D_n	Среднее S	Процент отклонений
5	0.341982	0.839231	0.0%
10	0.298889	0.997875	20.0%
100	0.085605	0.872717	0.0%
200	0.065228	0.934246	0.0%
400	0.045235	0.913032	0.0%
600	0.033767	0.833927	0.0%
800	0.030803	0.877123	0.0%
1000	0.025999	0.827441	0.0%

2. **Критерий согласия хи-квадрат**

   Для каждой выборки вычисляется статистика хи-квадрат с предварительной группировкой значений:

   (a) **Определение числа интервалов**: По правилу Старджесса

   $$k = 1 + \lfloor \log_2 n \rfloor$$

   где n — объем выборки.

(b) **Разбиение области значений**: Область $[1, +\infty)$ разбивается на k интервалов по квантилям распределения Парето. Границы интервалов определяются через обратную функцию распределения:

$$x_j = (1 - p_j)^{-1/\theta}, \quad p_j = \frac{j}{k}, \quad j = 1, 2, \ldots, k-1$$

где $\theta = 12$ — параметр распределения Парето.

(c) **Наблюдаемые частоты**: Для каждого интервала $I_j = [a_j, b_j]$ вычисляется

$$O_j = \#\{i : X_i \in I_j\}.$$

(d) **Ожидаемые частоты**: При справедливости H_0 для выборки объема n:

$$E_j = n \cdot [F(b_j) - F(a_j)], \quad j = 1, \ldots, k$$

где $F(x) = 1 - x^{-\theta}$ — функция распределения Парето.

(e) **Статистика хи-квадрат**:

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}.$$

Правило проверки гипотезы

Статистика χ^2 при справедливости H_0 имеет распределение хи-квадрат с числом степеней свободы:

$$\mathrm{df} = k - 1.$$

Для уровня значимости $\alpha = 0.05$ критическое значение вычисляется как:

$$\chi^2_{k-1,0.95}.$$

Правило принятия решения:

$$\begin{cases} \chi^2 \leq \chi^2_{k-1,0.95} & \Rightarrow \text{принимаем } H_0, \\ \chi^2 > \chi^2_{k-1,0.95} & \Rightarrow \text{отвергаем } H_0. \end{cases}$$

Для каждой выборки, сгенерированной в пункте 2.1, приведем значения полученных $\chi^2_{\text{набл}}$, $\chi^2_{\text{крит}}$. Значения ниже были вычислены на основе сгенерированных в ДЗ2 выборок с помощью написания программы на языке программирования Python (см. приложение):

Таблица 11: Значения статистики хи-квадрат для распределения Парето ($\theta = 12$, $\alpha = 0.05$)

n	Серия	k	$\chi^2_{\text{набл}}$	$\chi^2_{\text{крит}}$	Вывод
5*5	1	3	0.30	5.99	Принимаем
	2	3	3.14	5.99	Принимаем
	3	3	5.23	5.99	Принимаем
	4	3	0.45	5.99	Принимаем
	5	3	0.59	5.99	Принимаем
5*10	1	4	3.51	7.81	Принимаем
	2	4	4.11	7.81	Принимаем
	3	4	4.00	7.81	Принимаем
	4	4	4.75	7.81	Принимаем
	5	4	5.65	7.81	Принимаем
5*100	1	7	3.66	12.59	Принимаем
	2	7	9.22	12.59	Принимаем
	3	7	8.60	12.59	Принимаем
	4	7	4.86	12.59	Принимаем
	5	7	1.19	12.59	Принимаем
5*200	1	8	6.05	14.07	Принимаем
	2	8	7.55	14.07	Принимаем
	3	8	16.94	14.07	Отвергаем
	4	8	8.96	14.07	Принимаем
	5	8	13.02	14.07	Принимаем

Таблица 12: Значения статистики хи-квадрат для распределения Парето (продолжение)

n	Серия	k	$\chi^2_{\text{набл}}$	$\chi^2_{\text{крит}}$	Вывод
5*400	1	9	5.63	15.51	Принимаем
	2	9	6.98	15.51	Принимаем
	3	9	12.37	15.51	Принимаем
	4	9	6.13	15.51	Принимаем
	5	9	7.56	15.51	Принимаем
5*600	1	10	5.95	16.92	Принимаем
	2	10	12.17	16.92	Принимаем
	3	10	4.29	16.92	Принимаем
	4	10	5.58	16.92	Принимаем
	5	10	5.16	16.92	Принимаем
5*800	1	10	5.02	16.92	Принимаем
	2	10	13.59	16.92	Принимаем
	3	10	7.79	16.92	Принимаем
	4	10	3.97	16.92	Принимаем
	5	10	6.56	16.92	Принимаем
5*1000	1	10	3.95	16.92	Принимаем
	2	10	12.64	16.92	Принимаем
	3	10	7.36	16.92	Принимаем
	4	10	4.52	16.92	Принимаем
	5	10	6.44	16.92	Принимаем

Таблица 13: Сводные результаты критерия хи-квадрат для распределения Парето

n	Среднее χ^2	Процент отклонений
5	1.94	0.0%
10	4.40	0.0%
100	5.51	0.0%
200	10.50	20.0%
400	7.73	0.0%
600	6.63	0.0%
800	7.39	0.0%
1000	6.98	0.0%

Видим, что математическое ожидание статистики χ^2 при справедливой H_0 равно $\mathbb{E}[\chi^2_{k-1}] = k - 1$. Наблюдаемые средние близки к теоретическим:

- $n = 1000$, $k = 10$: $\mathbb{E} = 9$, получено 6.98

- $n = 100$, $k = 7$: $\mathbb{E} = 6$, получено 5.51
- $n = 10$, $k = 4$: $\mathbb{E} = 3$, получено 4.40

3. **Критерий согласия Колмогорова (Смирнова) для сложной гипотезы (в условиях когда неизвестен параметр распределения)**

   Для проверки сложной гипотезы о принадлежности выборки распределению Парето с неизвестным параметром θ применяется следующий подход:

   (a) **Оценка параметра**: По одной выборке достаточного объема ($n = 10000$), сгенерированной из распределения Парето с $\theta = 12$, методом максимального правдоподобия оценивается параметр:

   $$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} \ln X_i} = 11.8956$$

   (b) **Эмпирическая функция распределения**: Для проверяемой выборки объема n:

   $$F_n(x) = \frac{\#\{X_i \le x\}}{n}$$

   (c) **Теоретическая функция распределения**: С использованием оценки параметра $\hat{\theta}$:

   $$F(x; \hat{\theta}) = 1 - x^{-\hat{\theta}}, \quad x \ge 1$$

   (d) **Статистика Колмогорова**:

   $$D_n = \sup_{x \ge 1} |F_n(x) - F(x; \hat{\theta})|$$

   (e) **Статистика с поправкой Большева**:

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

   Правило проверки гипотезы

   Поскольку параметр θ оценивается по независимой большой выборке, распределение статистики S приближается к распределению Колмогорова. Для уровня значимости $\alpha = 0.05$ критическое значение:

   $$K_{0.95} \approx 1.358.$$

   Правило принятия решения:

   $$\begin{cases} S \le 1.358 & \Rightarrow \text{принимаем } H_0, \\ S > 1.358 & \Rightarrow \text{отвергаем } H_0. \end{cases}$$

Для каждой выборки, сгенерированной в пункте 2.1, приведем значения полученных D_n, S, критического значения. Значения ниже были вычислены на основе сгенерированных в ДЗ2 выборок с помощью написания программы на языке программирования Python (см. приложение):

Таблица 14: Значения статистик критерия Колмогорова для сложной гипотезы (Парето, $\hat{\theta} = 11.8956$, $\alpha = 0.05$)

n	Серия	D_n	S	Крит.	Вывод
5*5	1	0.254395	0.643379	1.358	Принимаем
	2	0.302218	0.750315	1.358	Принимаем
	3	0.185244	0.488753	1.358	Принимаем
	4	0.242544	0.616880	1.358	Принимаем
	5	0.325843	0.803143	1.358	Принимаем
5*10	1	0.154395	0.540943	1.358	Принимаем
	2	0.292993	0.979228	1.358	Принимаем
	3	0.141822	0.501184	1.358	Принимаем
	4	0.242544	0.819695	1.358	Принимаем
	5	0.425843	1.399339	1.358	Отвергаем
5*100	1	0.068066	0.697324	1.358	Принимаем
	2	0.114016	1.156828	1.358	Принимаем
	3	0.110954	1.126206	1.358	Принимаем
	4	0.080299	0.819652	1.358	Принимаем
	5	0.069644	0.713111	1.358	Принимаем
5*200	1	0.052507	0.754350	1.358	Принимаем
	2	0.089022	1.270753	1.358	Принимаем
	3	0.045495	0.655179	1.358	Принимаем
	4	0.048697	0.700469	1.358	Принимаем
	5	0.085764	1.224670	1.358	Принимаем

Таблица 15: Значения статистик критерия Колмогорова для сложной гипотезы (продолжение)

n	Серия	D_n	S	Крит.	Вывод
5*400	1	0.037837	0.765066	1.358	Принимаем
	2	0.051931	1.046958	1.358	Принимаем
	3	0.027894	0.566215	1.358	Принимаем
	4	0.039378	0.795897	1.358	Принимаем
	5	0.059230	1.192934	1.358	Принимаем
5*600	1	0.032890	0.812450	1.358	Принимаем
	2	0.031224	0.771625	1.358	Принимаем
	3	0.022650	0.561602	1.358	Принимаем
	4	0.036528	0.901556	1.358	Принимаем
	5	0.037915	0.935531	1.358	Принимаем
5*800	1	0.032837	0.934653	1.358	Принимаем
	2	0.030272	0.862127	1.358	Принимаем
	3	0.027875	0.794308	1.358	Принимаем
	4	0.027591	0.786270	1.358	Принимаем
	5	0.034966	0.994884	1.358	Принимаем
5*1000	1	0.018837	0.600937	1.358	Принимаем
	2	0.027320	0.869202	1.358	Принимаем
	3	0.024812	0.789896	1.358	Принимаем
	4	0.021472	0.684286	1.358	Принимаем
	5	0.031831	1.011867	1.358	Принимаем

Таблица 16: Сводные результаты критерия Колмогорова для сложной гипотезы

n	Среднее D_n	Среднее S	Процент отклонений
5	0.262049	0.660494	0.0%
10	0.251519	0.848078	20.0%
100	0.088596	0.902624	0.0%
200	0.064297	0.921084	0.0%
400	0.043254	0.873414	0.0%
600	0.032241	0.796553	0.0%
800	0.030708	0.874448	0.0%
1000	0.024854	0.791238	0.0%

Видим, что средние значения S для больших выборок ($n \geq 400$) находятся в диапазоне 0.79–0.92, что значительно ниже критического значения 1.358, что свидетельствует о хорошем соответствии выборок распределению Парето с параметром $\hat{\theta} = 11.8956$.

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

   Для проверки сложной гипотезы о принадлежности выборки распределению Парето с неизвестным параметром θ применяется критерий хи-квадрат с группировкой:

   (a) **Оценка параметра** θ: Для каждой проверяемой выборки методом максимального правдоподобия оценивается параметр:

   $$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} \ln X_i}$$

   (b) **Определение числа интервалов**: По правилу Старджесса

   $$k = 1 + \lfloor \log_2 n \rfloor$$

   (c) **Разбиение на интервалы**: Границы интервалов определяются через квантили распределения Парето с оцененным параметром:

   $$x_j = (1 - p_j)^{-1/\hat{\theta}}, \quad p_j = \frac{j}{k}, \quad j = 1, 2, \ldots, k-1$$

   (d) **Наблюдаемые и ожидаемые частоты**:

   $$O_j = \#\{X_i \in I_j\}, \quad E_j = n \cdot [F(b_j; \hat{\theta}) - F(a_j; \hat{\theta})]$$

   где $F(x; \hat{\theta}) = 1 - x^{-\hat{\theta}}$.

   (e) **Статистика хи-квадрат**:

   $$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}$$

   Правило проверки гипотезы

   При оценке одного параметра число степеней свободы уменьшается на единицу:

   $$\mathrm{df} = k - 2.$$

   Для уровня значимости $\alpha = 0.05$ критическое значение:

   $$\chi^2_{\text{крит}} = \chi^2_{k-2,0.95}.$$

   Правило принятия решения:

   $$\begin{cases} \chi^2 \leq \chi^2_{\text{крит}} & \Rightarrow \text{принимаем } H_0, \\ \chi^2 > \chi^2_{\text{крит}} & \Rightarrow \text{отвергаем } H_0. \end{cases}$$

Для каждой выборки, сгенерированной в пункте 2.1, приведем значения полученных $\chi^2_{\text{набл}}$, $\chi^2_{\text{крит}}$. Значения ниже были вычислены на основе сгенерированных в ДЗ2 выборок с помощью написания программы на языке программирования Python (см. приложение):

Таблица 17: Значения статистики хи-квадрат для сложной гипотезы (Парето, $\alpha = 0.05$)

n	Серия	k	$\hat{\theta}$	$\chi^2_{\text{набл}}$	$\chi^2_{\text{крит}}$	df	Вывод
5*5	1	3	11.4390	0.29	3.84	1	Принимаем
	2	3	6.8682	1.24	3.84	1	Принимаем
	3	3	9.4250	8.90	3.84	1	Отвергаем
	4	3	9.7534	2.93	3.84	1	Принимаем
	5	3	10.7538	0.71	3.84	1	Принимаем
5*10	1	4	12.2197	3.52	5.99	2	Принимаем
	2	4	9.1263	1.91	5.99	2	Принимаем
	3	4	9.5008	2.39	5.99	2	Принимаем
	4	4	10.8064	2.18	5.99	2	Принимаем
	5	4	13.5299	5.46	5.99	2	Принимаем
5*100	1	7	12.2589	4.07	11.07	5	Принимаем
	2	7	13.5828	6.79	11.07	5	Принимаем
	3	7	12.5704	10.99	11.07	5	Принимаем
	4	7	11.3617	4.27	11.07	5	Принимаем
	5	7	12.6844	2.15	11.07	5	Принимаем

Таблица 18: Значения статистики хи-квадрат для сложной гипотезы (продолжение)

n	Серия	k	$\hat{\theta}$	$\chi^2_{\text{набл}}$	$\chi^2_{\text{крит}}$	df	Вывод
5*200	1	8	11.6314	4.12	12.59	6	Принимаем
	2	8	13.4781	4.66	12.59	6	Принимаем
	3	8	11.6265	12.63	12.59	6	Отвергаем
	4	8	11.6682	7.11	12.59	6	Принимаем
	5	8	13.5010	7.65	12.59	6	Принимаем
5*400	1	9	11.2223	2.82	14.07	7	Принимаем
	2	9	12.4730	5.36	14.07	7	Принимаем
	3	9	12.0676	11.66	14.07	7	Принимаем
	4	9	11.4388	1.58	14.07	7	Принимаем
	5	9	12.8376	9.98	14.07	7	Принимаем
5*600	1	10	11.5030	4.57	15.51	8	Принимаем
	2	10	12.0032	12.17	15.51	8	Принимаем
	3	10	12.0277	4.29	15.51	8	Принимаем
	4	10	11.4259	6.66	15.51	8	Принимаем
	5	10	12.6708	5.48	15.51	8	Принимаем

Таблица 19: Значения статистики хи-квадрат для сложной гипотезы (продолжение)

n	Серия	k	$\hat{\theta}$	$\chi^2_{\text{набл}}$	$\chi^2_{\text{крит}}$	df	Вывод
5*800	1	10	11.5633	3.06	15.51	8	Принимаем
	2	10	12.0477	13.45	15.51	8	Принимаем
	3	10	12.1070	9.26	15.51	8	Принимаем
	4	10	11.6783	3.98	15.51	8	Принимаем
	5	10	12.5079	4.28	15.51	8	Принимаем
5*1000	1	10	11.7606	4.43	15.51	8	Принимаем
	2	10	12.1558	12.45	15.51	8	Принимаем
	3	10	11.8347	6.81	15.51	8	Принимаем
	4	10	11.7598	3.19	15.51	8	Принимаем
	5	10	12.5243	5.77	15.51	8	Принимаем

Таблица 20: Средние оценки параметра θ по объёму выборки

n	Среднее $\hat{\theta}$
5	9.6479
10	11.0366
100	12.4916
200	12.3811
400	12.0078
600	11.9261
800	11.9809
1000	12.0071

Таблица 21: Сводные результаты критерия хи-квадрат для сложной гипотезы

n	Среднее χ^2	Процент отклонений
5	2.81	20.0%
10	3.09	0.0%
100	5.65	0.0%
200	7.23	20.0%
400	6.28	0.0%
600	6.63	0.0%
800	6.81	0.0%
1000	6.53	0.0%

2. Проверка гипотезы об однородности выборок

Метод проверки

Для проверки гипотезы об однородности двух выборок используется двухвыборочный критерий Смирнова. Статистика критерия вычисляется по формуле:

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

где $\mathcal{F}_n(x)$ и $\mathcal{F}_m(x)$ — эмпирические функции распределения выборок объёмов n и m соответственно. Эти значения были ранее посчитаны в ДЗ2.

Критические значения $D_{m,n,\text{крит}}$ для уровня значимости $\alpha = 0.05$ взяты из таблиц Масси.

Правило принятия решения:

$$\begin{cases} D_{m,n} \leq D_{m,n,\text{крит}} & \Rightarrow \text{выборки однородны (принимаем } H_0) \\ D_{m,n} > D_{m,n,\text{крит}} & \Rightarrow \text{выборки неоднородны (отвергаем } H_0) \end{cases}$$

Дискретное равномерное распределение

Таблица 22: Проверка однородности: дискретное равномерное распределение

Выборки		$D_{m,n}$	$D_{\text{крит}}$	Вывод
m	n			
5	10	0.2921	1.73	Принимаем
5	100	0.3317	1.63	Принимаем
5	200	0.3644	1.62	Принимаем
5	400	0.3511	1.61	Принимаем
5	600	0.3615	1.61	Принимаем
5	800	0.3600	1.61	Принимаем
5	1000	0.3604	1.61	Принимаем
10	100	0.2352	1.52	Принимаем
10	200	0.2191	1.51	Принимаем
10	400	0.2155	1.50	Принимаем
10	600	0.2279	1.50	Принимаем
10	800	0.2357	1.50	Принимаем
10	1000	0.2228	1.50	Принимаем
100	200	0.1878	1.36	Принимаем
100	400	0.2147	1.35	Принимаем
100	600	0.2191	1.35	Принимаем
100	800	0.2522	1.35	Принимаем
100	1000	0.2498	1.35	Принимаем
200	400	0.1328	1.36	Принимаем
200	600	0.1225	1.35	Принимаем
200	800	0.1550	1.35	Принимаем
200	1000	0.1627	1.35	Принимаем
400	600	0.0852	1.36	Принимаем
400	800	0.1347	1.36	Принимаем
400	1000	0.1572	1.36	Принимаем
600	800	0.1034	1.36	Принимаем
600	1000	0.1510	1.36	Принимаем
800	1000	0.1117	1.36	Принимаем

Таблица 23: Проверка однородности: распределение Парето

Выборки		$D_{m,n}$	$D_{\text{крит}}$	Вывод
m	n			
5	10	0.1826	1.73	Принимаем
5	100	0.4670	1.63	Принимаем
5	200	0.4506	1.62	Принимаем
5	400	0.4144	1.61	Принимаем
5	600	0.4031	1.61	Принимаем
5	800	0.4118	1.61	Принимаем
5	1000	0.4158	1.61	Принимаем
10	100	0.3859	1.52	Принимаем
10	200	0.3611	1.51	Принимаем
10	400	0.3123	1.50	Принимаем
10	600	0.2885	1.50	Принимаем
10	800	0.3009	1.50	Принимаем
10	1000	0.3115	1.50	Принимаем
100	200	0.2776	1.36	Принимаем
100	400	0.3309	1.35	Принимаем
100	600	0.4290	1.35	Принимаем
100	800	0.4196	1.35	Принимаем
100	1000	0.4195	1.35	Принимаем
200	400	0.2656	1.36	Принимаем
200	600	0.3552	1.35	Принимаем
200	800	0.3384	1.35	Принимаем
200	1000	0.3382	1.35	Принимаем
400	600	0.1988	1.36	Принимаем
400	800	0.1837	1.36	Принимаем
400	1000	0.1809	1.36	Принимаем
600	800	0.1435	1.36	Принимаем
600	1000	0.1962	1.36	Принимаем
800	1000	0.0727	1.36	Принимаем

Для **дискретного равномерного распределения**: Все 28 пар выборок однородны.

Для **распределения Парето**: Все 28 пар выборок однородны.

Все выборки действительно происходят из одного распределения, что подтверждает корректность процедуры генерации.

3. Приложения

1. https://github.com/faisvire/mathstat-hw