Национальный исследовательский университет
Высшая школа экономики
Московский институт электроники и математики

Департамент прикладной математики
кафедра компьютерной безопасности

Домашнее задание №2 по математической статистике
Основные понятия математической статистики

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.

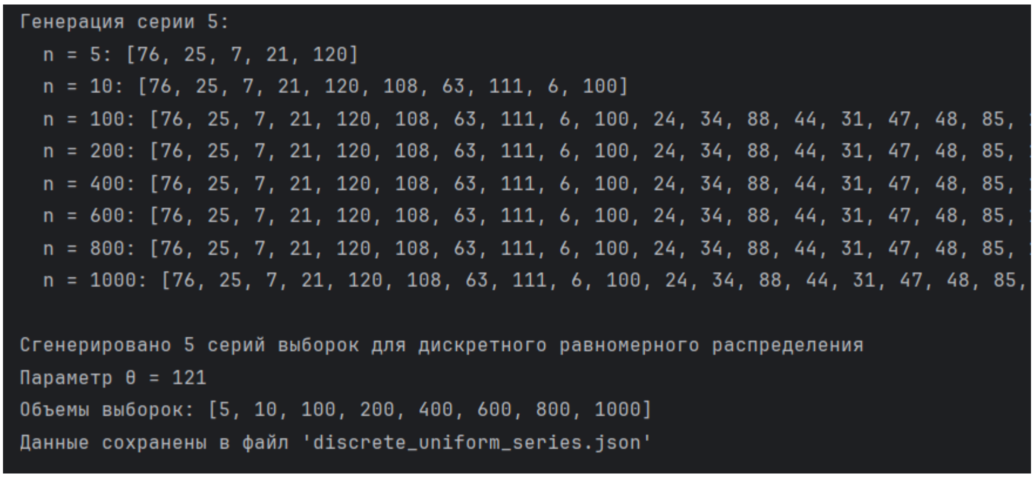
Москва 2025

Содержание

1. Генерация выборок выбранных случайных величин

Для генерации выборок случайных величин обоих распределений воспользуемся алгоритмами моделирования, описанными в задании 3 домашнего задания №1 и модифицируем написанный на языке программирования Python код так, чтобы для каждой из выбранных случайных величин строились по 5 выборок следующих объемов $n = \{5, 10, 100, 200, 400, 600, 800, 1000\}$.

Заданные параметры: $\theta = 121$ для дискретного распределения и $\theta = 12$ для непрерывного.



```
Генерация серии 5:
  n = 5: [76, 25, 7, 21, 120]
  n = 10: [76, 25, 7, 21, 120, 108, 63, 111, 6, 100]
  n = 100: [76, 25, 7, 21, 120, 108, 63, 111, 6, 100, 24, 34, 88, 44, 31, 47, 48, 85,
  n = 200: [76, 25, 7, 21, 120, 108, 63, 111, 6, 100, 24, 34, 88, 44, 31, 47, 48, 85,
  n = 400: [76, 25, 7, 21, 120, 108, 63, 111, 6, 100, 24, 34, 88, 44, 31, 47, 48, 85,
  n = 600: [76, 25, 7, 21, 120, 108, 63, 111, 6, 100, 24, 34, 88, 44, 31, 47, 48, 85,
  n = 800: [76, 25, 7, 21, 120, 108, 63, 111, 6, 100, 24, 34, 88, 44, 31, 47, 48, 85,
  n = 1000: [76, 25, 7, 21, 120, 108, 63, 111, 6, 100, 24, 34, 88, 44, 31, 47, 48, 85,

Сгенерировано 5 серий выборок для дискретного равномерного распределения
Параметр θ = 121
Объемы выборок: [5, 10, 100, 200, 400, 600, 800, 1000]
Данные сохранены в файл 'discrete_uniform_series.json'
```

Рис. 1: Пример работы программы и сгенерированные выборки для дискретного равномерного I распределения

Результаты генерации (5 выборок указанных объемов) хранятся в файлах 'discrete_uniform_series.json' и 'pareto_series.json' для дискретной и непрерывной случайных величин соответственно.

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

Данные файлы и код можно посмотреть в Приложении 1.

2. Построение эмпирической функции распределения

▷ Дискретное равномерное I распределение

Эмпирическую функцию распределения можно посчитать по формуле:

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

Для каждого объема выборки и каждого целого t от 1 до 121 мы вычислили значение эмпирической функции распределения, которое посчитали как среднее арифметическое по 5 сериям выборок с помощью кода на языке программирования Python.

На графиках ниже представлены сравнения эмпирической и теоретической функций распределения для случайной величины с дискретным равномерным I распределением.

- Распределение Парето

Для распределения Парето с параметром $\theta = 12$ также было сгенерировано 5 серий выборок объемов: $n = \{5, 10, 100, 200, 400, 600, 800, 1000\}$

Для каждого объема выборки собирались все уникальные значения из всех 5 серий, а затем для каждой точки t из этих уникальных значений вычислялась эмпирическая функция распределения по формуле:

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

Получившиеся значения усреднялись по 5 сериям методом среднего арифметического. Все осуществлялось с помощью кода на языке программирования Python.

На графиках ниже представлены сравнения эмпирической и теоретической функций распределения для случайной величины с непрерывным распределением Парето.

- Двухвыборочные статистики

Были вычислены двухвыборочные статистики для всех пар объемов выборок дискретного равномерного распределения по формуле:

Для каждой пары построенных эмпирических $\mathcal{F}_n(x), \mathcal{F}_m(x), n, m \in= \{5, 10, 100, 200, 400, 600, 800, 1000\}$ необходимо вычислить двухвыборочную статистику. Это реализовано с помощью кода на языке программирования Python.

Вычислим двухвыборочную статистику по следующей формуле:

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

Написанный код можно посмотреть в Приложении 1.

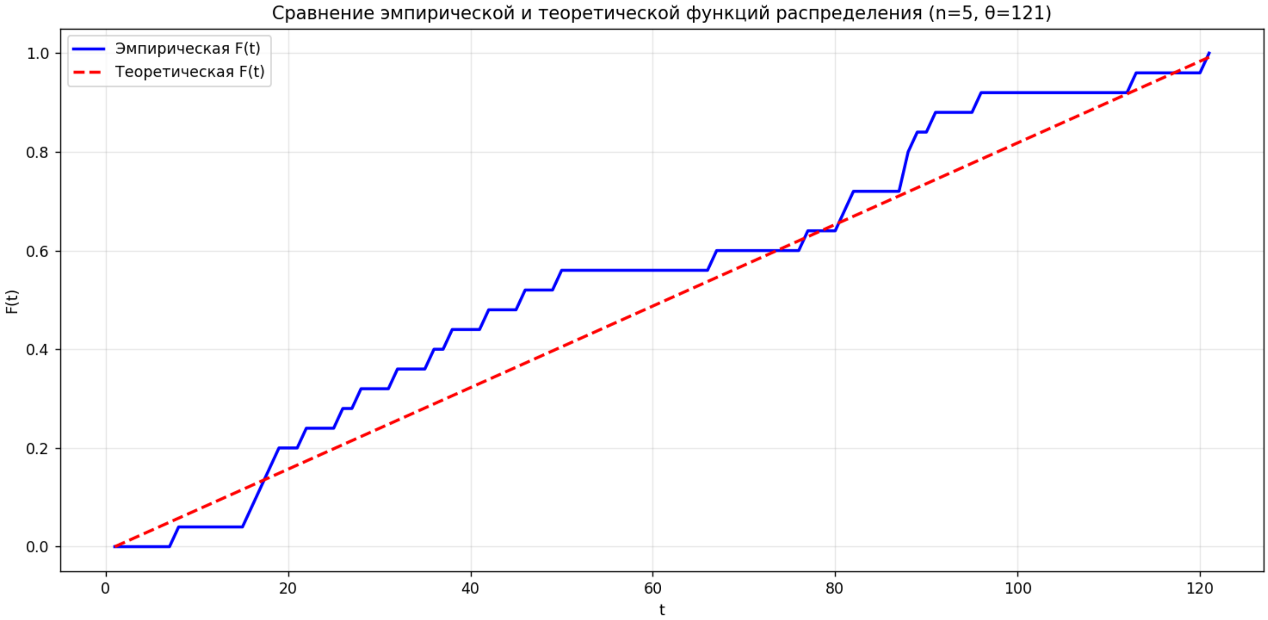


Рис. 2: График для выборки дискретного распределения объёма 5

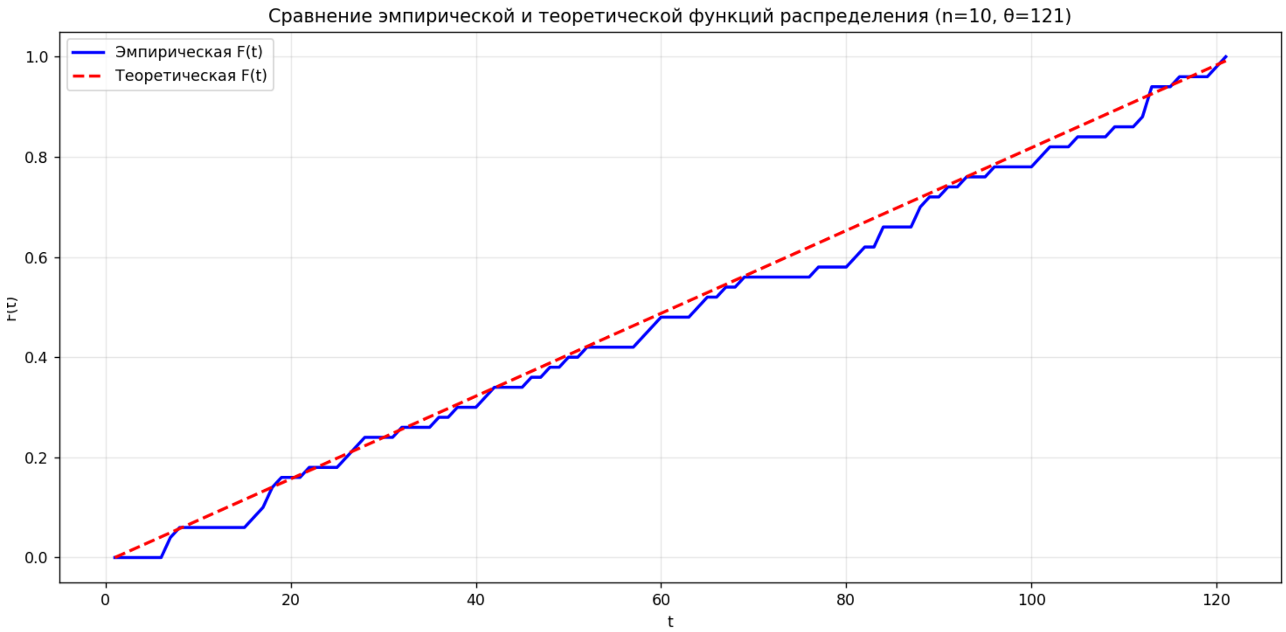


Рис. 3: График для выборки дискретного распределения объёма 10

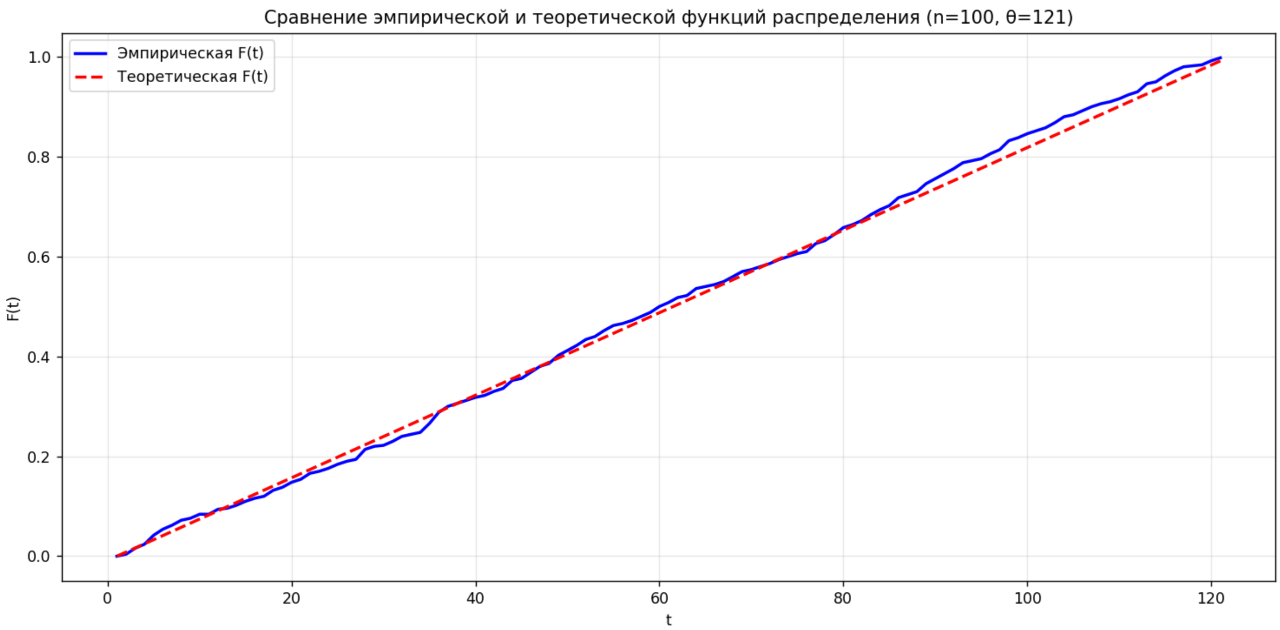


Рис. 4: График для выборки дискретного распределения объёма 100

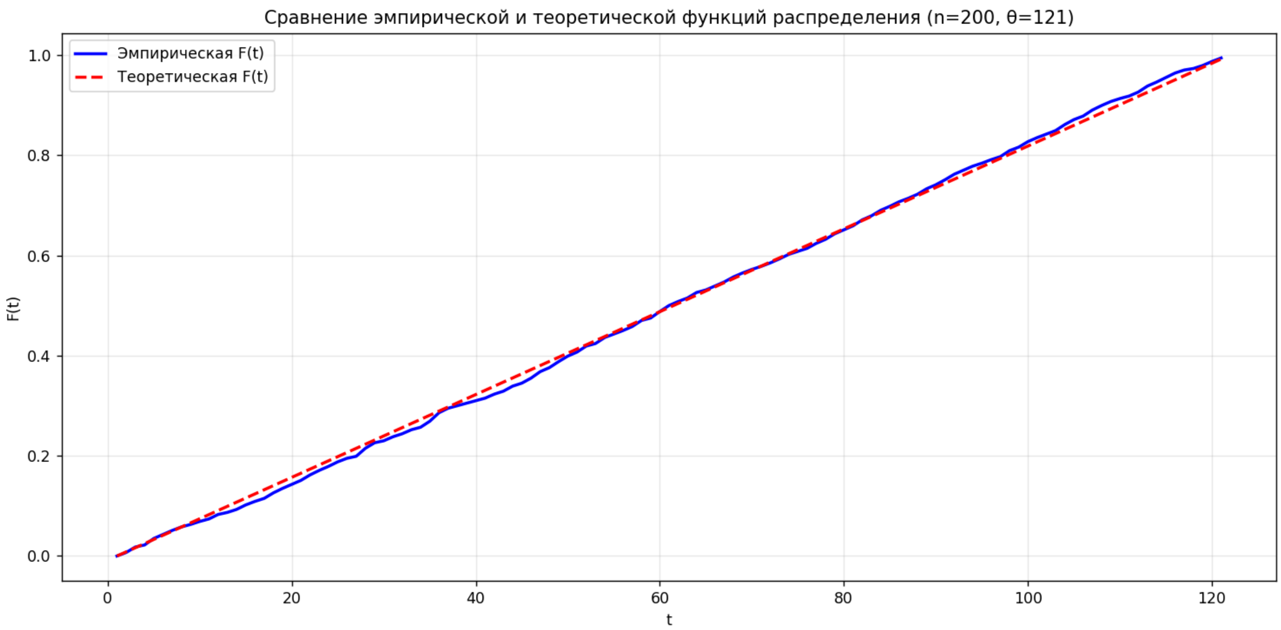


Рис. 5: График для выборки дискретного распределения объёма 200

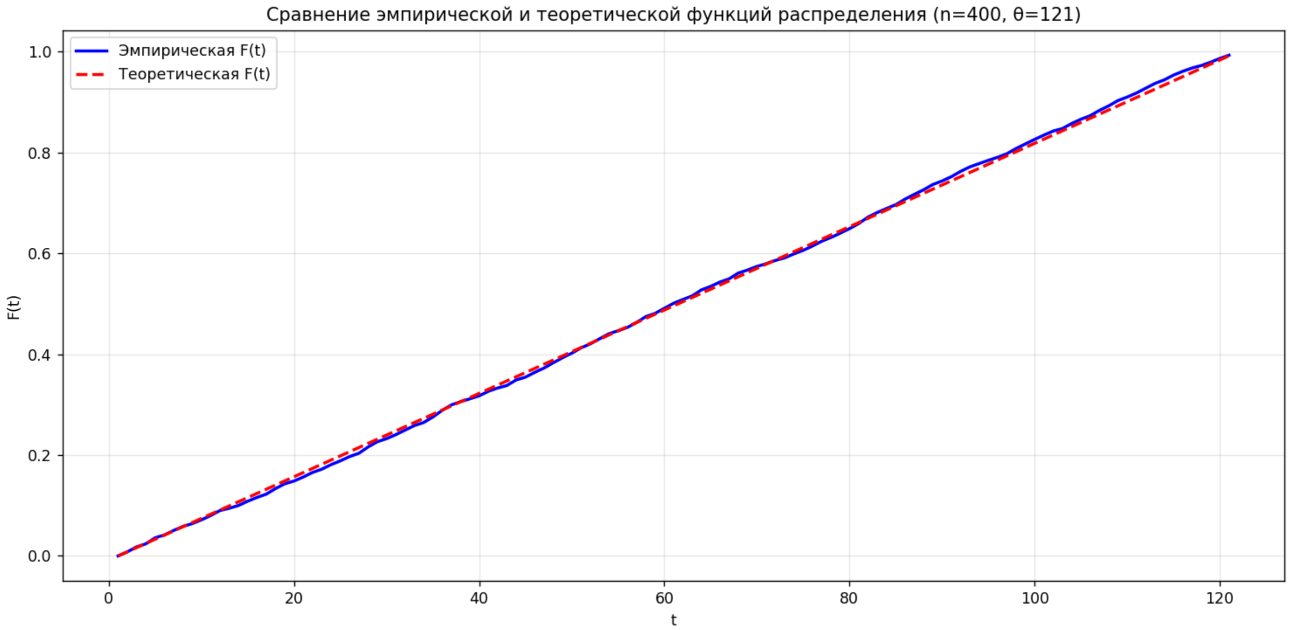


Рис. 6: График для выборки дискретного распределения объёма 400

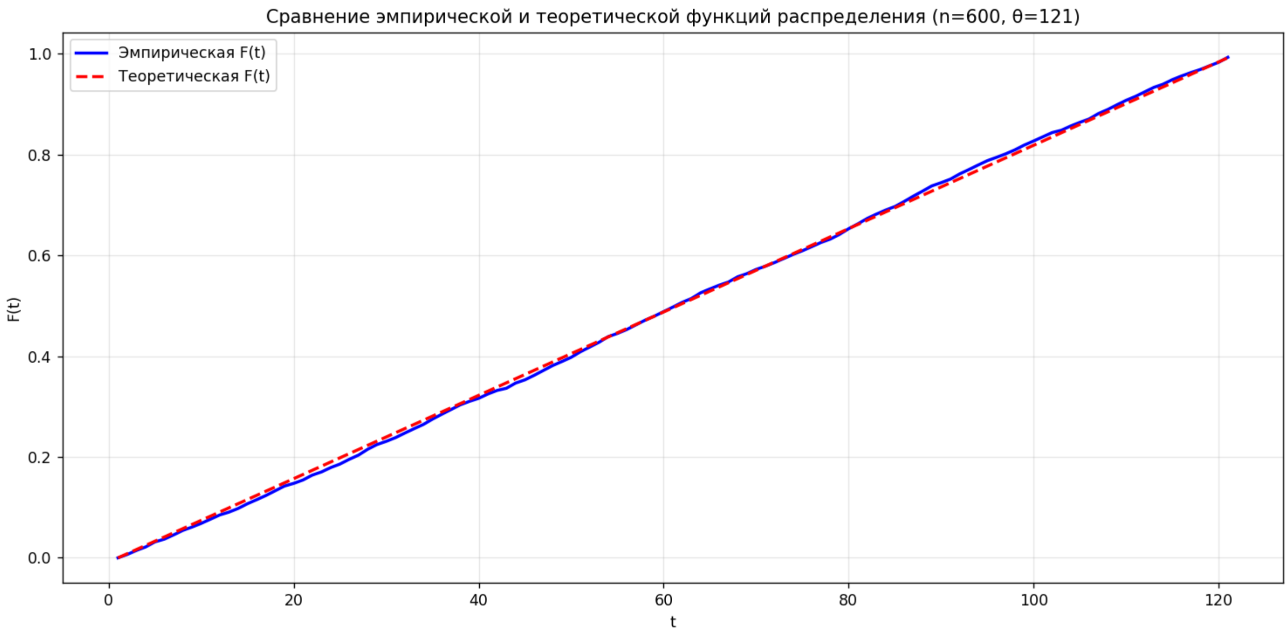


Рис. 7: График для выборки дискретного распределения объёма 600

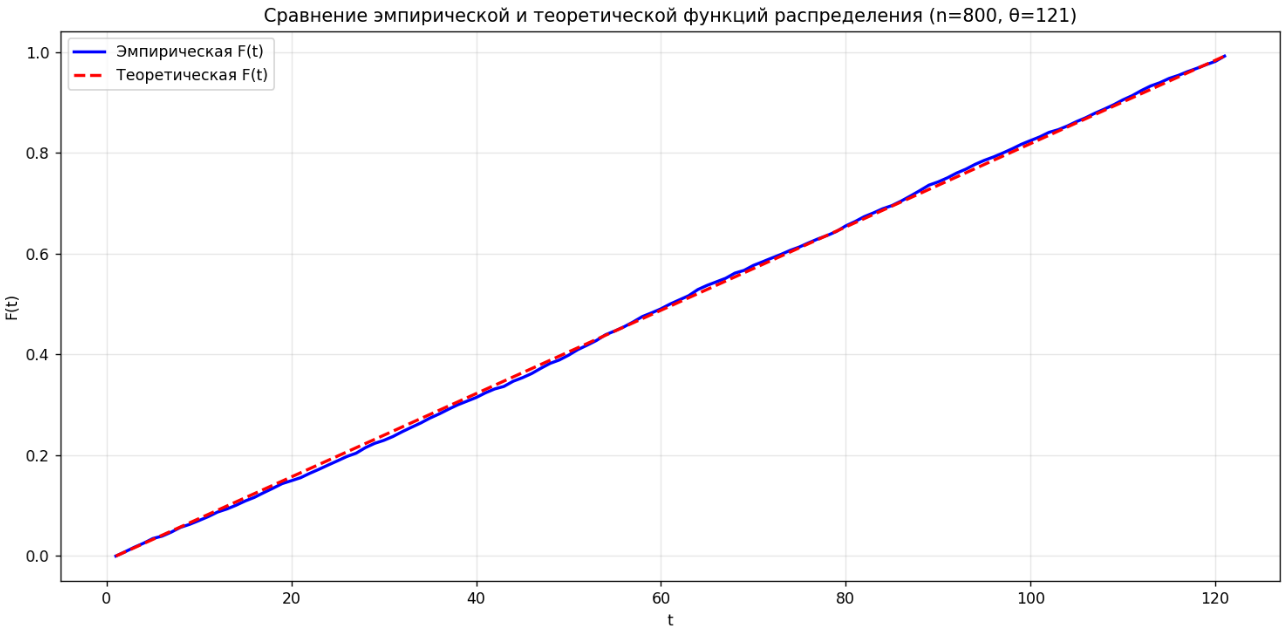


Рис. 8: График для выборки дискретного распределения объёма 800

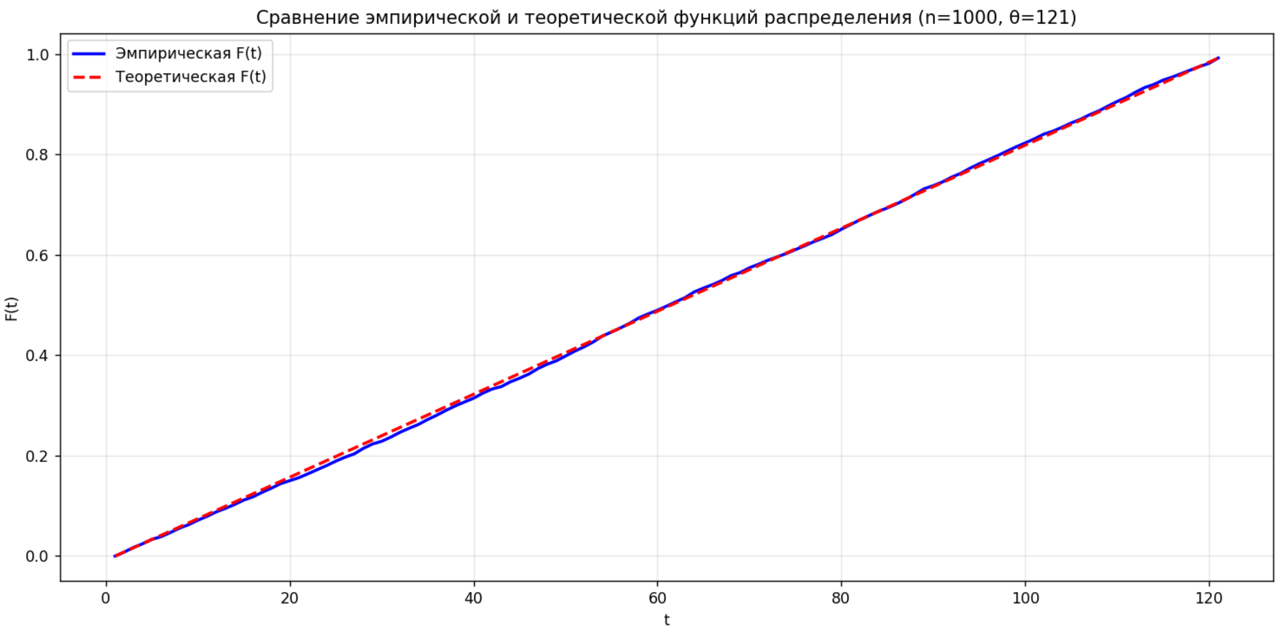


Рис. 9: График для выборки дискретного распределения объёма 1000

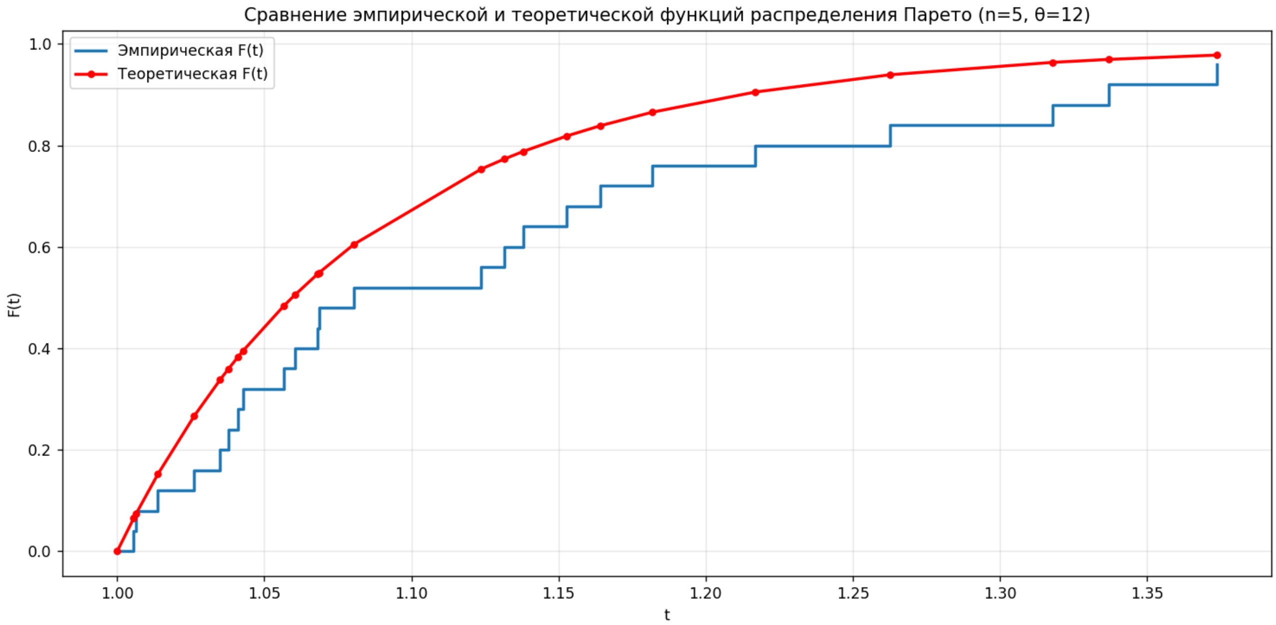


Рис. 10: График для выборки непрерывного распределения объёма 5

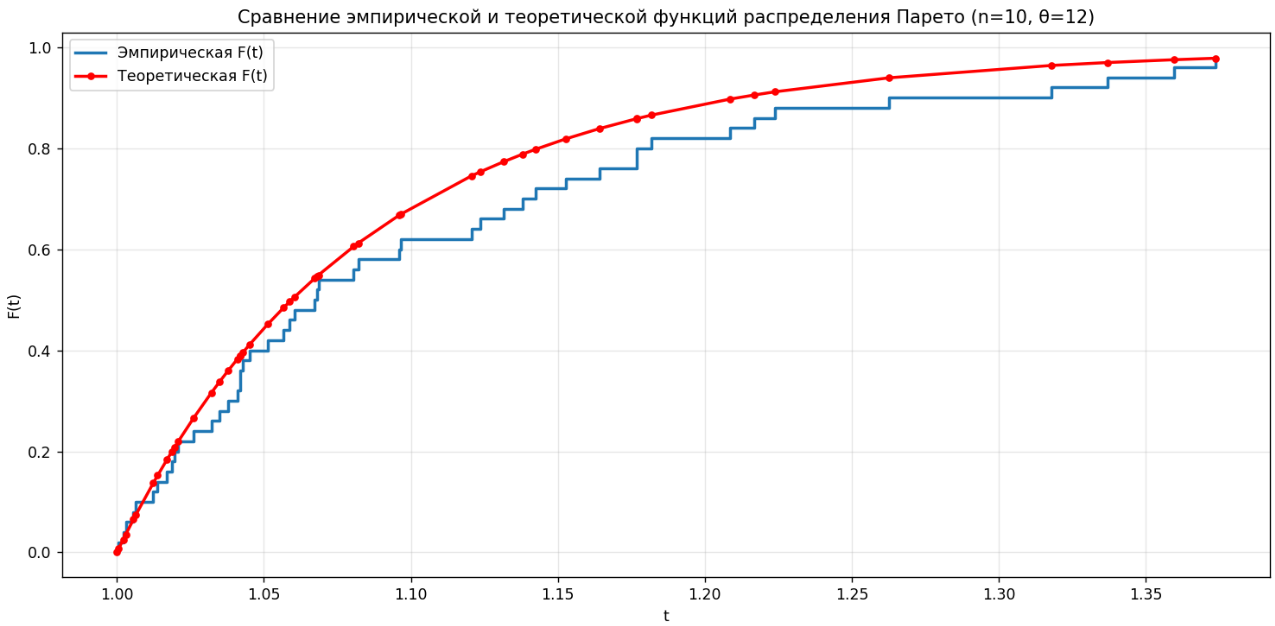


Рис. 11: График для выборки непрерывного распределения объёма 10

Рис. 12: График для выборки непрерывного распределения объёма 100

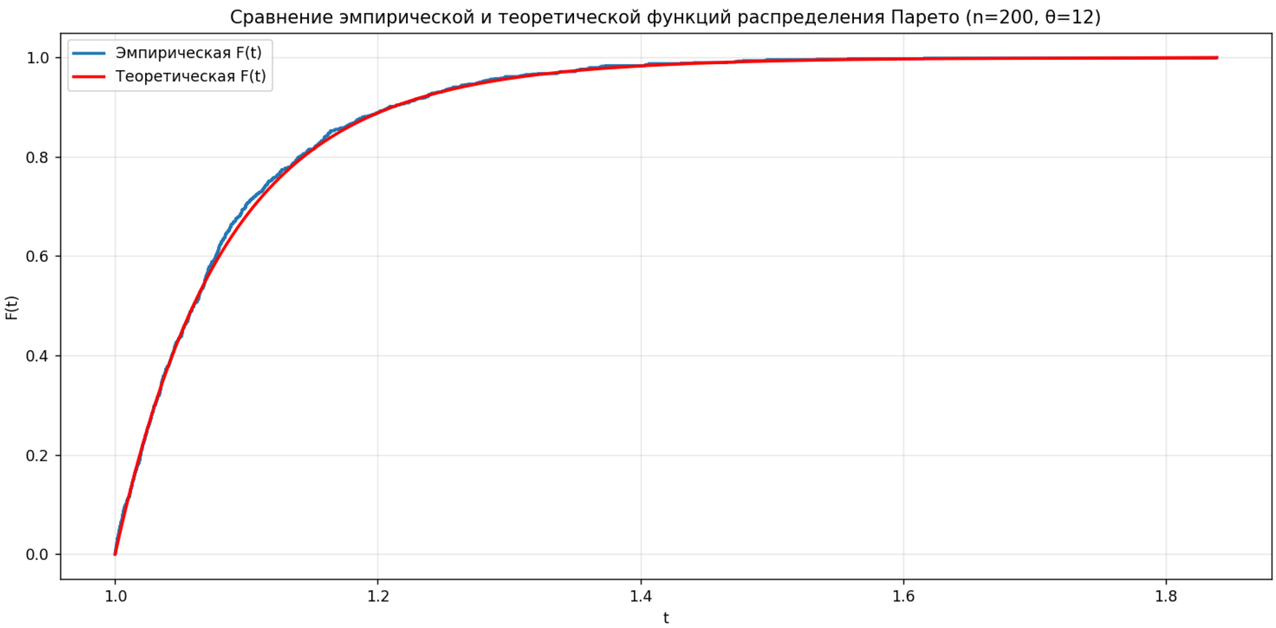


Рис. 13: График для выборки непрерывного распределения объёма 200

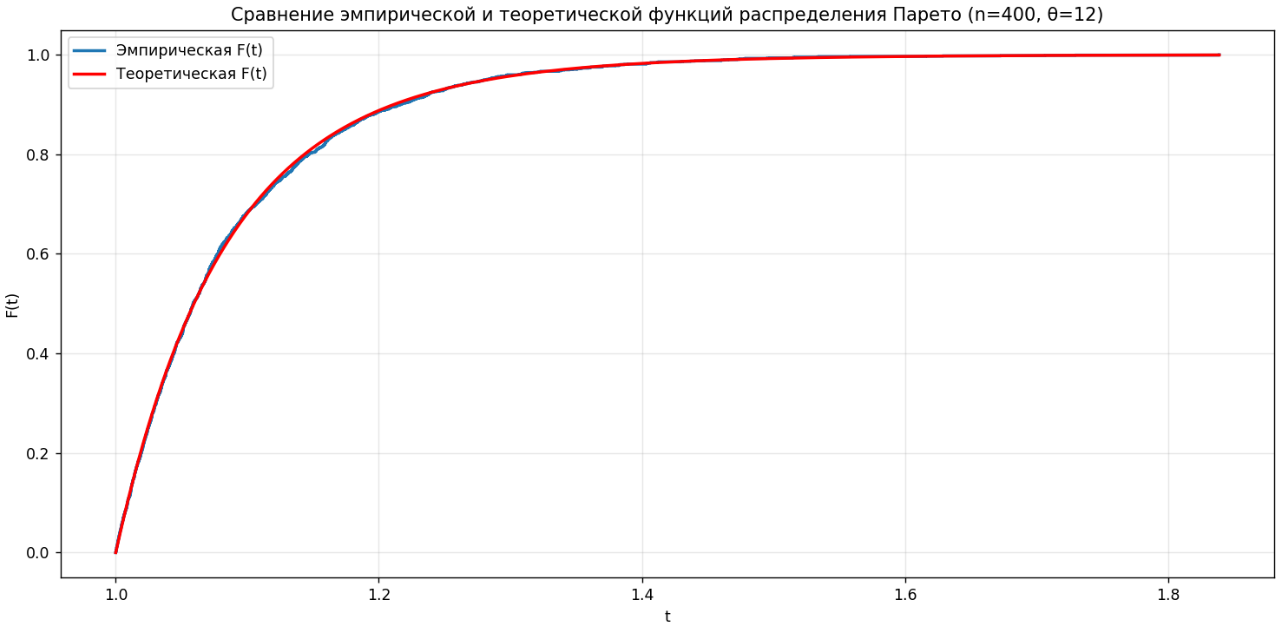


Рис. 14: График для выборки непрерывного распределения объёма 400

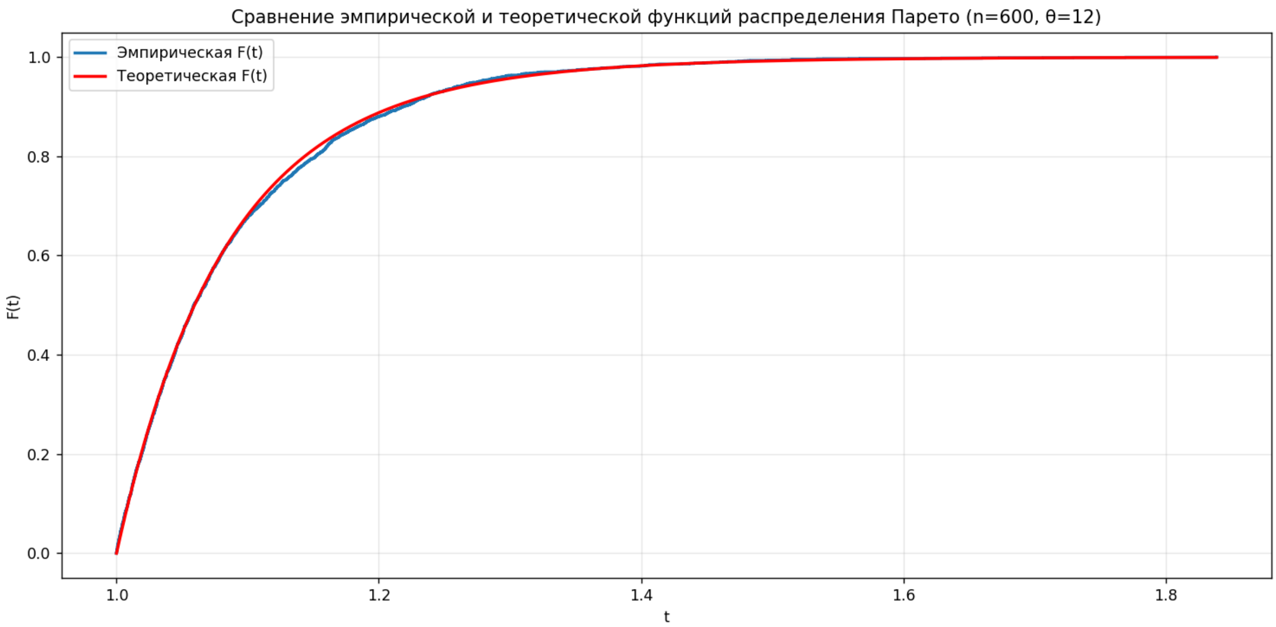


Рис. 15: График для выборки непрерывного распределения объёма 600

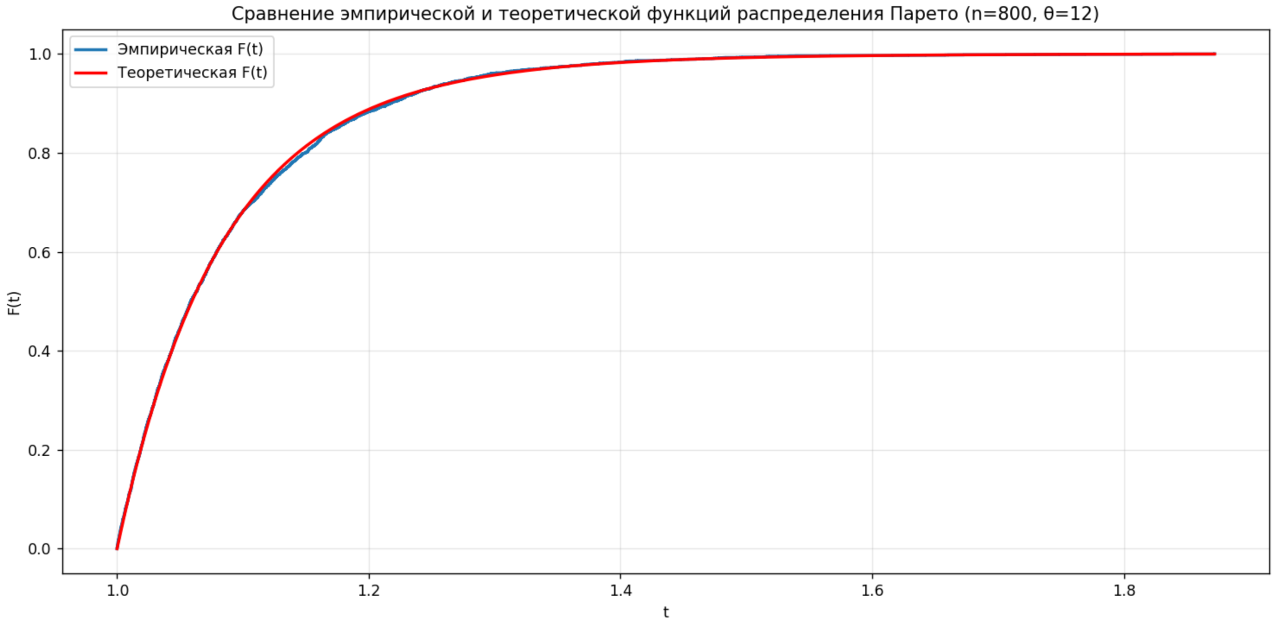


Рис. 16: График для выборки непрерывного распределения объёма 800

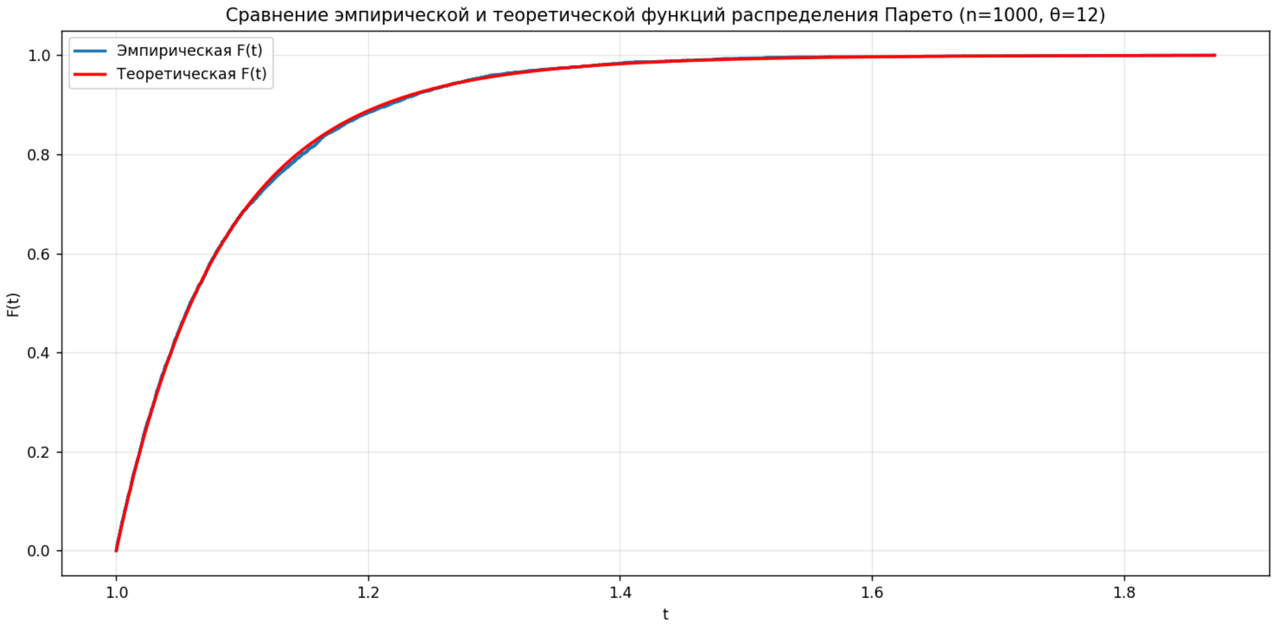


Рис. 17: График для выборки непрерывного распределения объёма 1000

Таблица 1: Значения двухвыборочных статистик $D_{m,n}$ для дискретного распределения (дискретное равномерное I)

n/m	5	10	100	200	400	600	800	1000
5	-	0.2921	0.3317	0.3644	0.3511	0.3615	0.3600	0.3604
10	-	-	0.2352	0.2191	0.2155	0.2279	0.2357	0.2228
100	-	-	-	0.1878	0.2147	0.2191	0.2522	0.2498
200	-	-	-	-	0.1328	0.1225	0.1550	0.1627
400	-	-	-	-	-	0.0852	0.1347	0.1572
600	-	-	-	-	-	-	0.1034	0.1510
800	-	-	-	-	-	-	-	0.1117
1000	-	-	-	-	-	-	-	-

Таблица 2: Значения двухвыборочных статистик $D_{m,n}$ для непрерывного распределения (распределение Парето)

n/m	5	10	100	200	400	600	800	1000
5	-	0.1826	0.4670	0.4506	0.4144	0.4031	0.4118	0.4158
10	-	-	0.3859	0.3611	0.3123	0.2885	0.3009	0.3115
100	-	-	-	0.2776	0.3309	0.4290	0.4196	0.4195
200	-	-	-	-	0.2656	0.3552	0.3384	0.3382
400	-	-	-	-	-	0.1988	0.1837	0.1809
600	-	-	-	-	-	-	0.1435	0.1962
800	-	-	-	-	-	-	-	0.0727
1000	-	-	-	-	-	-	-	-

3. Построение гистограммы и полигона частот

На графиках ниже представлено сравнение усреднённых полигонов частот с теоретической плотностью распределения для дискретного равномерного I распределения и распределения Парето. Для каждой из 5 серий выборок и каждого объёма выборки $n = \{5, 10, 100, 200, 400, 600, 800, 1000\}$ вычислялись относительные частоты появления каждого значения t, и полученные частоты усреднялись по 5 сериям выборок методом среднего арифметического. Расчёт осуществлялся с помощью написания кода на языке программирования Python.

Для каждого объема выборки n и каждого значения t усредненная частота для дискретного распределения вычислялась по формуле:

$$p_n(t) = \frac{1}{5 \cdot n} \sum_{k=1}^{5} \sum_{i=1}^{n} \mathbb{I}\{x_i^{(k)} = t\}$$

Для каждого объема выборки n и каждого значения t усредненная частота для непрерывного распределения вычислялась по формуле:

$$p_n(w) = \frac{1}{5 \cdot n} \sum_{k=1}^{5} \sum_{i=1}^{n} \mathbb{I}\{x_i^{(k)} \in w\},$$

где w - интервал. На графиках непрерывного распределения ось абсцисс разбита на 15 равных интервалов.

Полученные графики для рассматриваемых распределений демонстрируют следующие теоремы из курса теории вероятностей и математической статистики:

1. **Усиленный закон больших чисел**

   Среднее арифметическое последовательности независимых случайных величин сходится почти наверное к их теоретическим значениям по мере увеличения количества испытаний. Это иллюстрируют приведенные ниже графики в задании 3.

2. **Теорема Гливенко-Кантелли**

   Теорема утверждает, что выборочная функция распределения (наблюдаемая на основе выборки) равномерно сходится к истинной (теоретической) функции распределения при увеличении размера выборки. Это иллюстрируют графики сравнения эмпирической и теоретической функций распределений, сделанные в задании 2.

Написанный код можно посмотреть в Приложении 1.

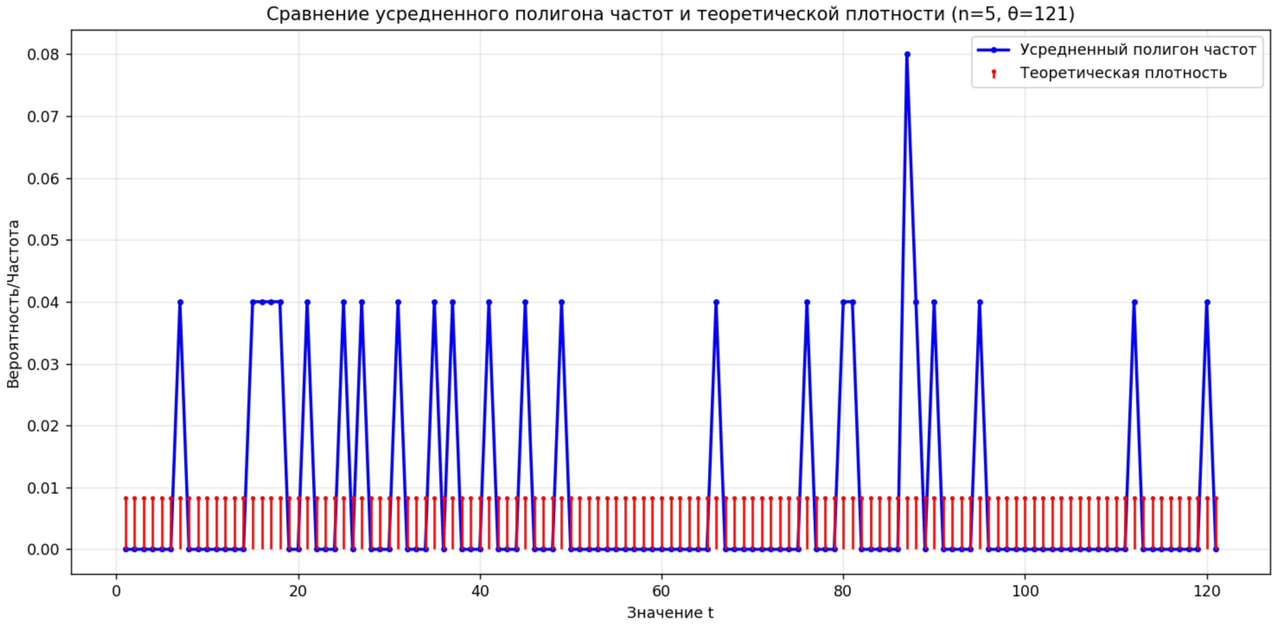


Рис. 18: График для выборки дискретного распределения объёма 5

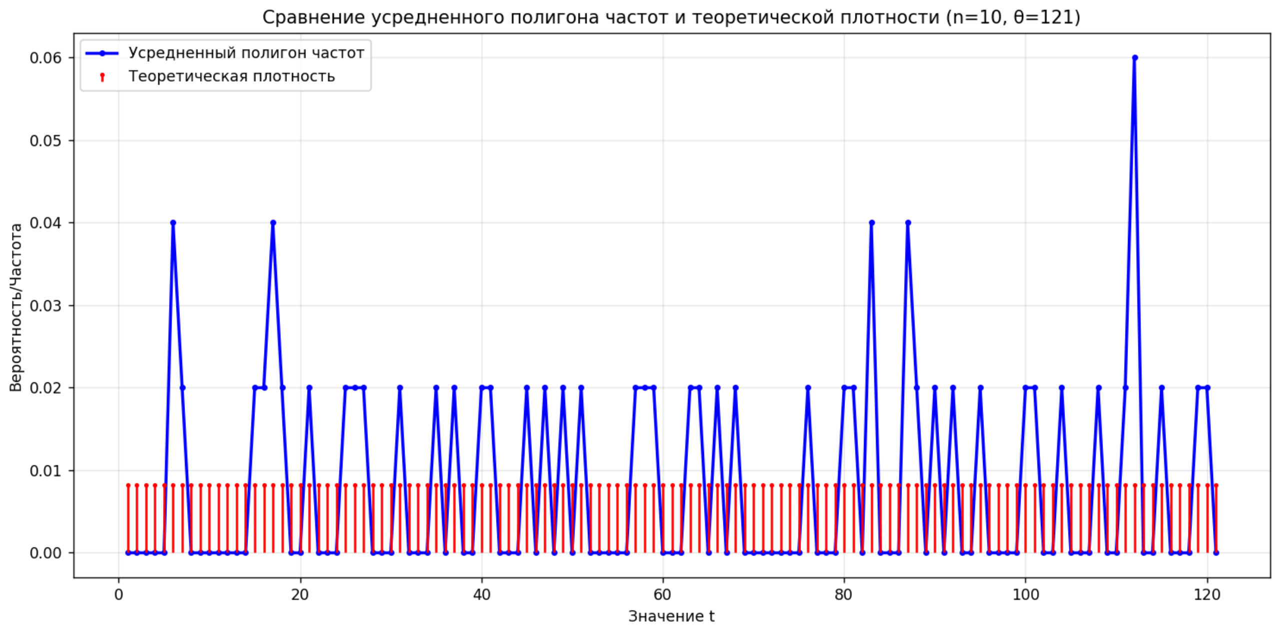


Рис. 19: График для выборки дискретного распределения объёма 10

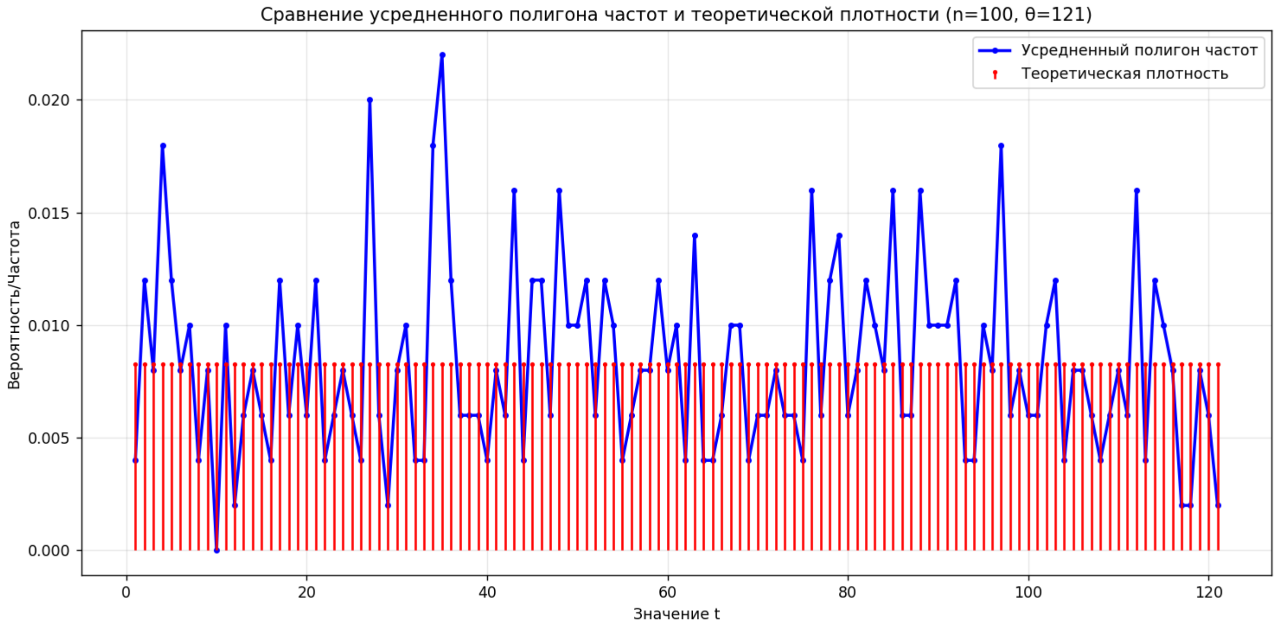


Рис. 20: График для выборки дискретного распределения объёма 100

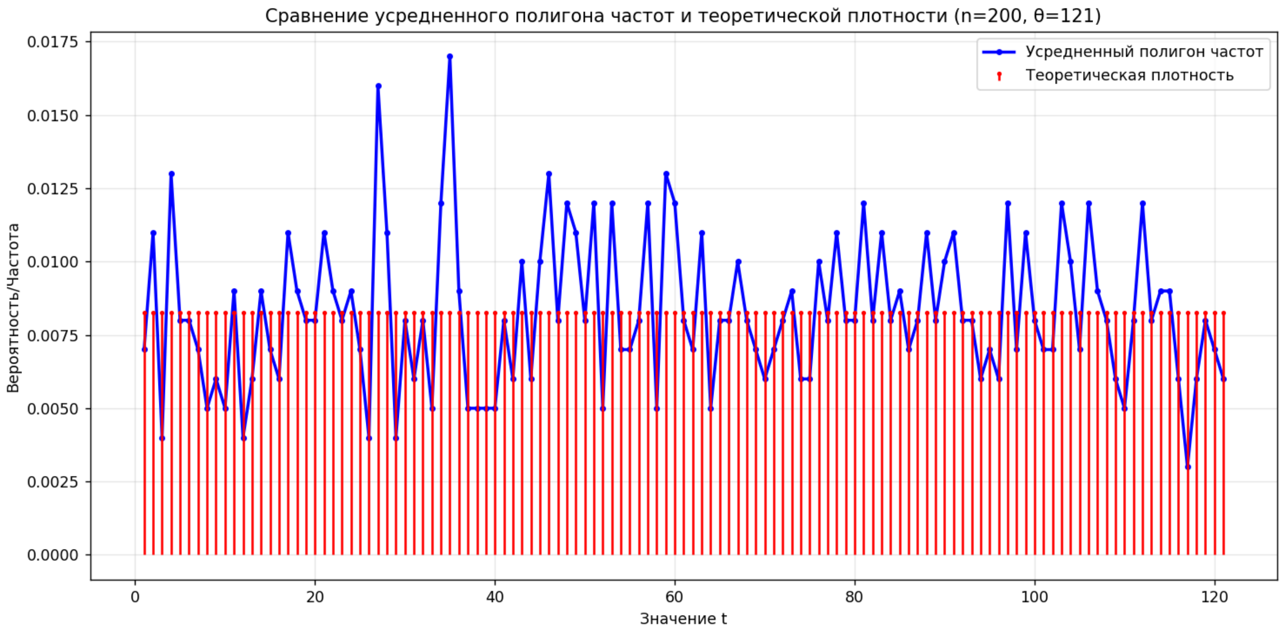


Рис. 21: График для выборки дискретного распределения объёма 200

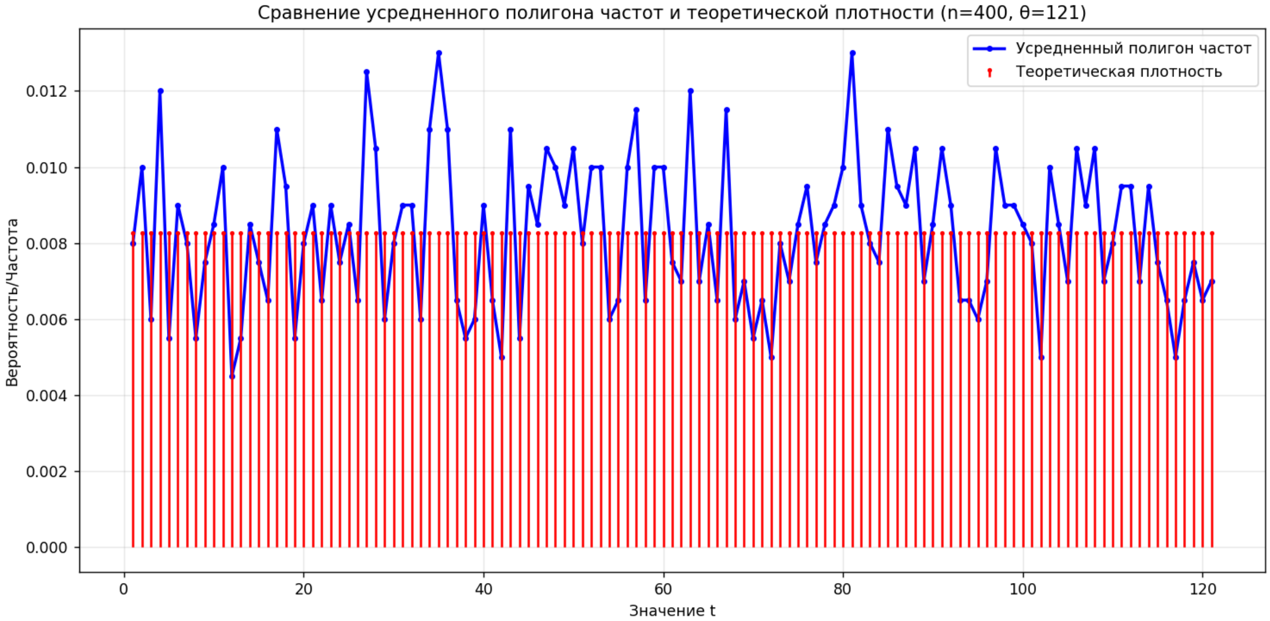


Рис. 22: График для выборки дискретного распределения объёма 400

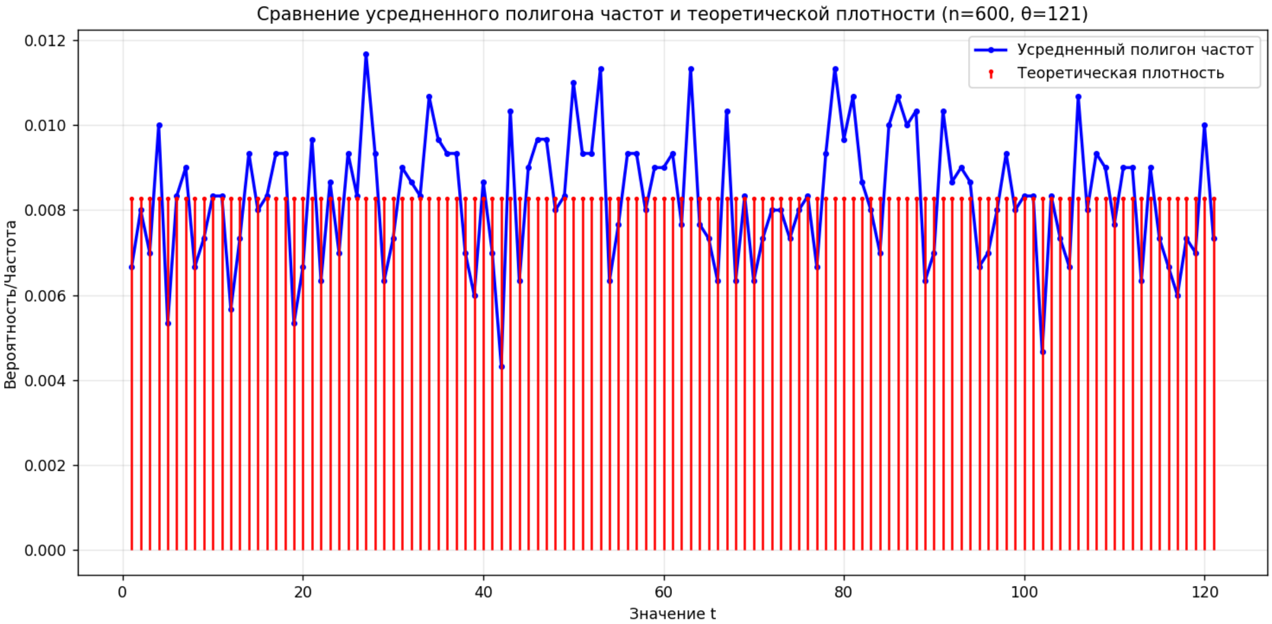


Рис. 23: График для выборки дискретного распределения объёма 600

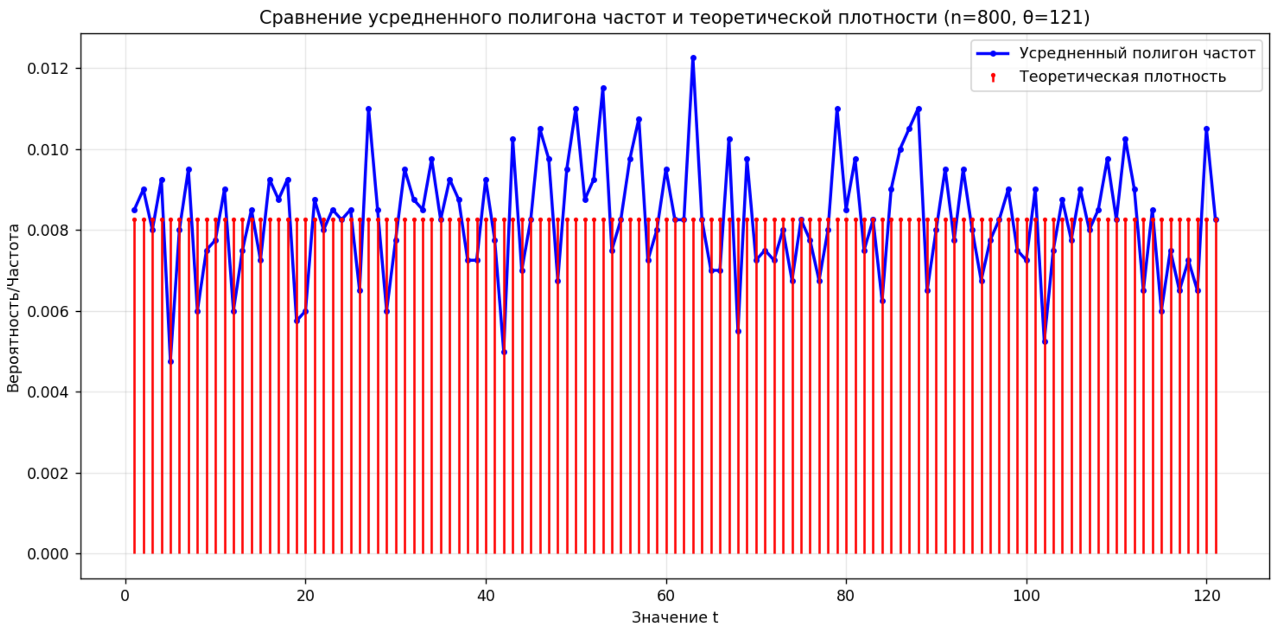


Рис. 24: График для выборки дискретного распределения объёма 800

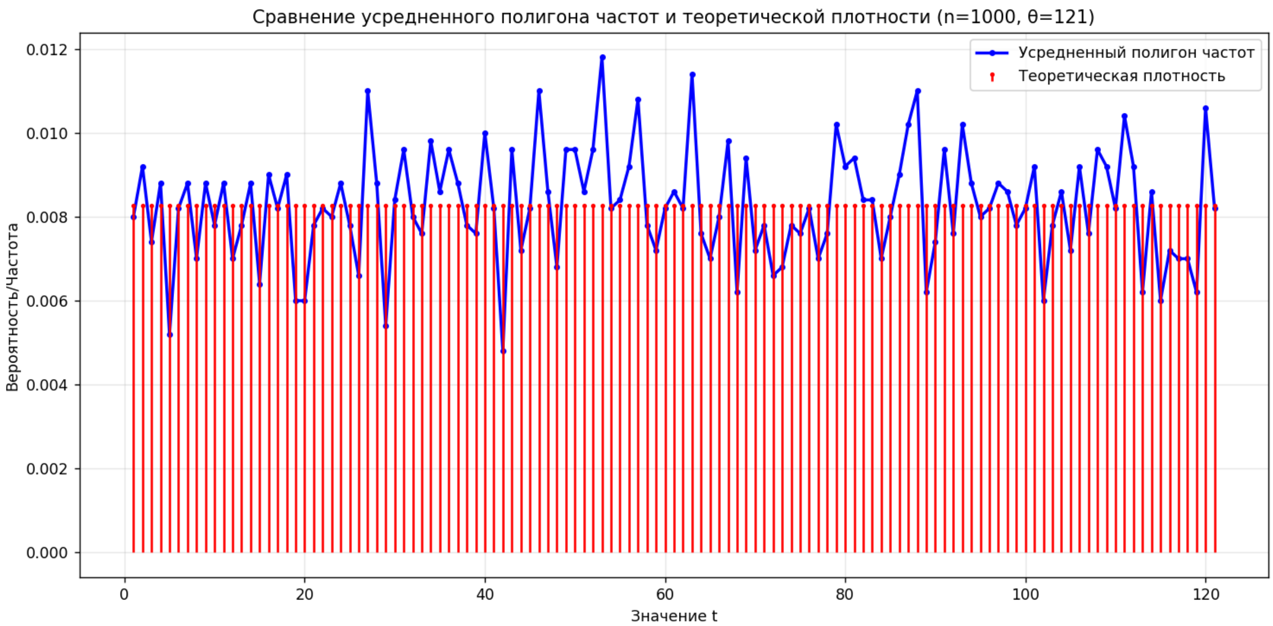


Рис. 25: График для выборки дискретного распределения объёма 1000

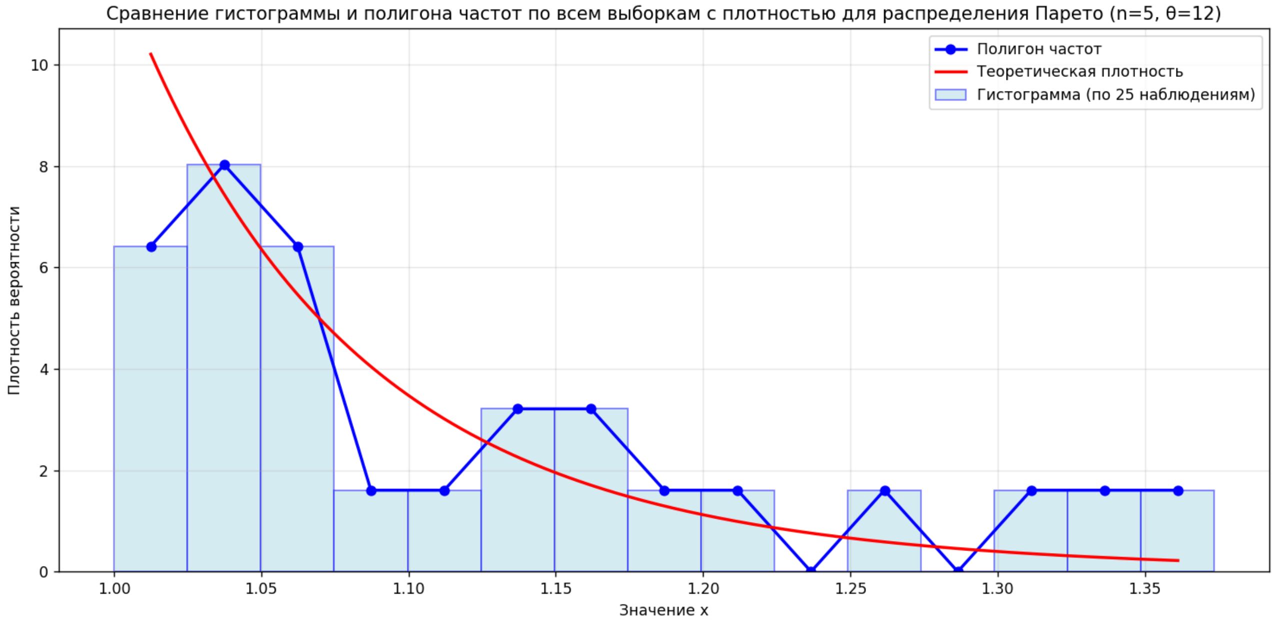


Рис. 26: График для выборки непрерывного распределения объёма 5

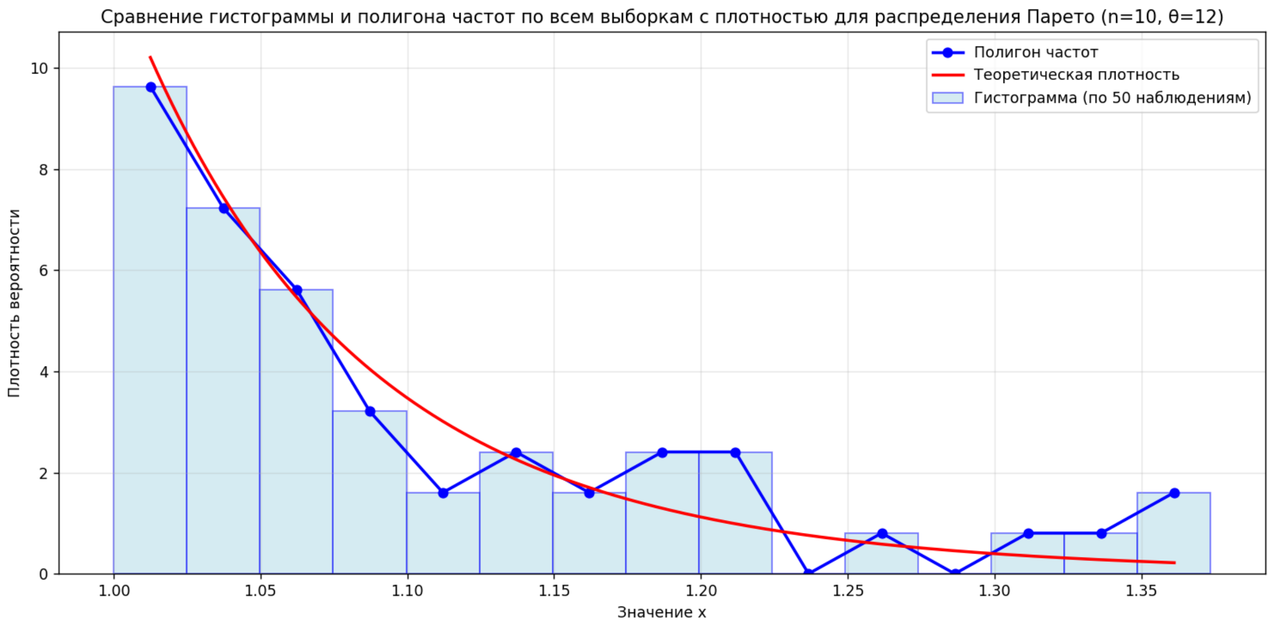


Рис. 27: График для выборки непрерывного распределения объёма 10

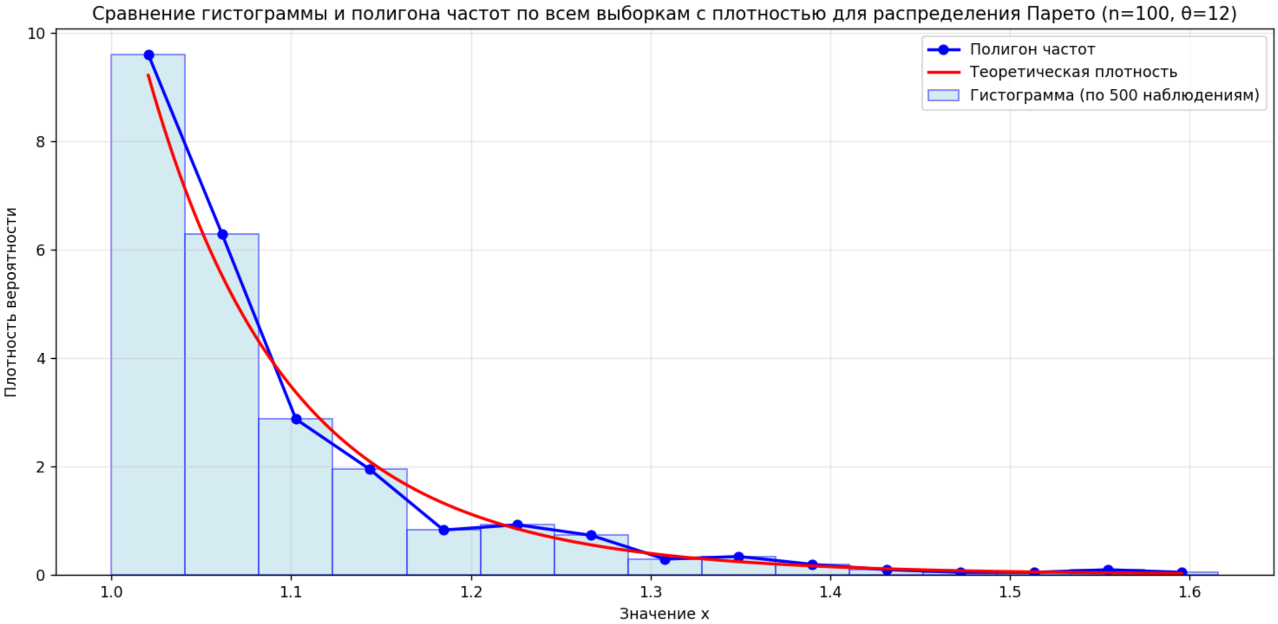


Рис. 28: График для выборки непрерывного распределения объёма 100

Рис. 29: График для выборки непрерывного распределения объёма 200

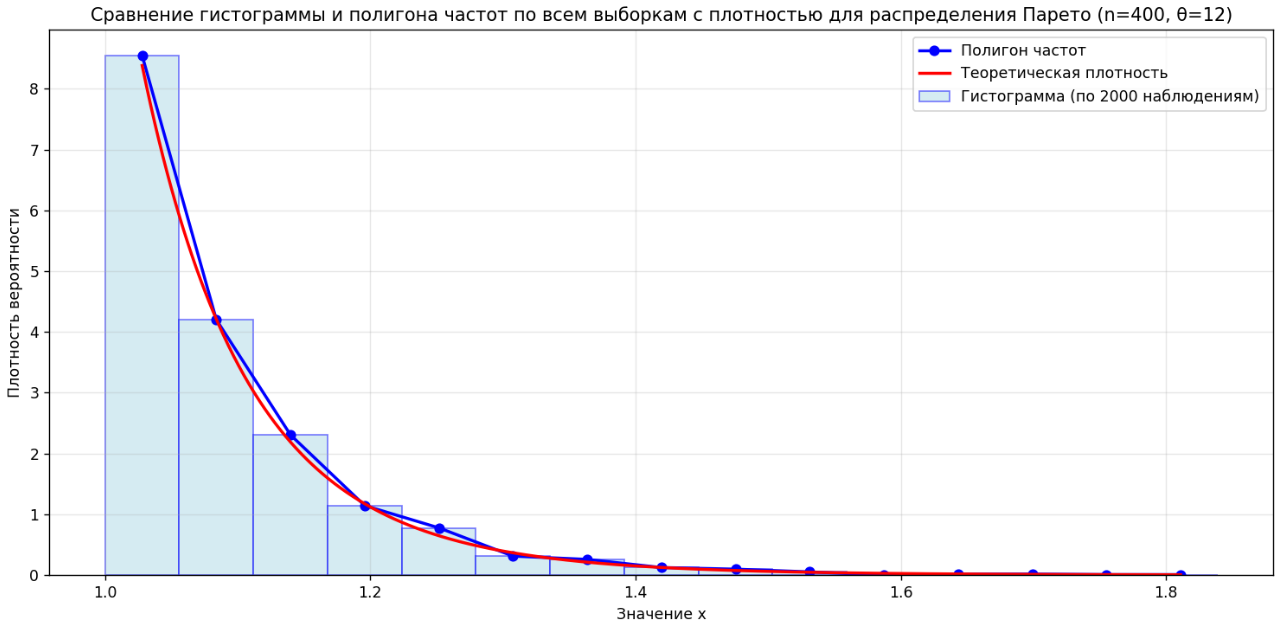


Рис. 30: График для выборки непрерывного распределения объёма 400

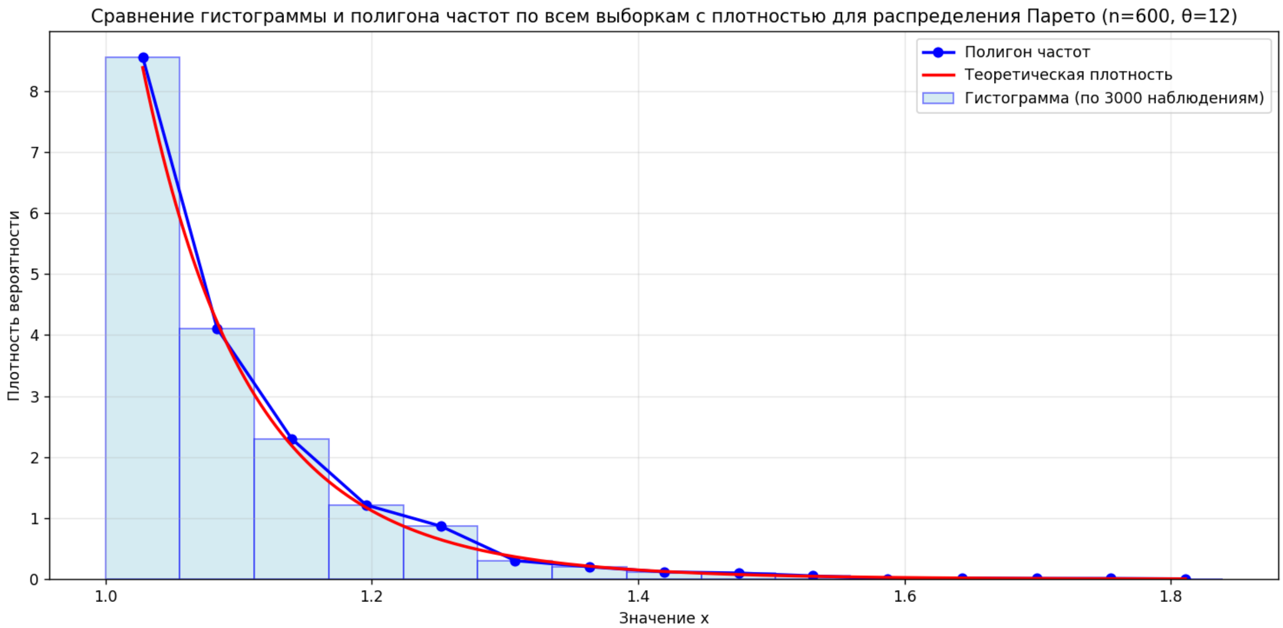


Рис. 31: График для выборки непрерывного распределения объёма 600

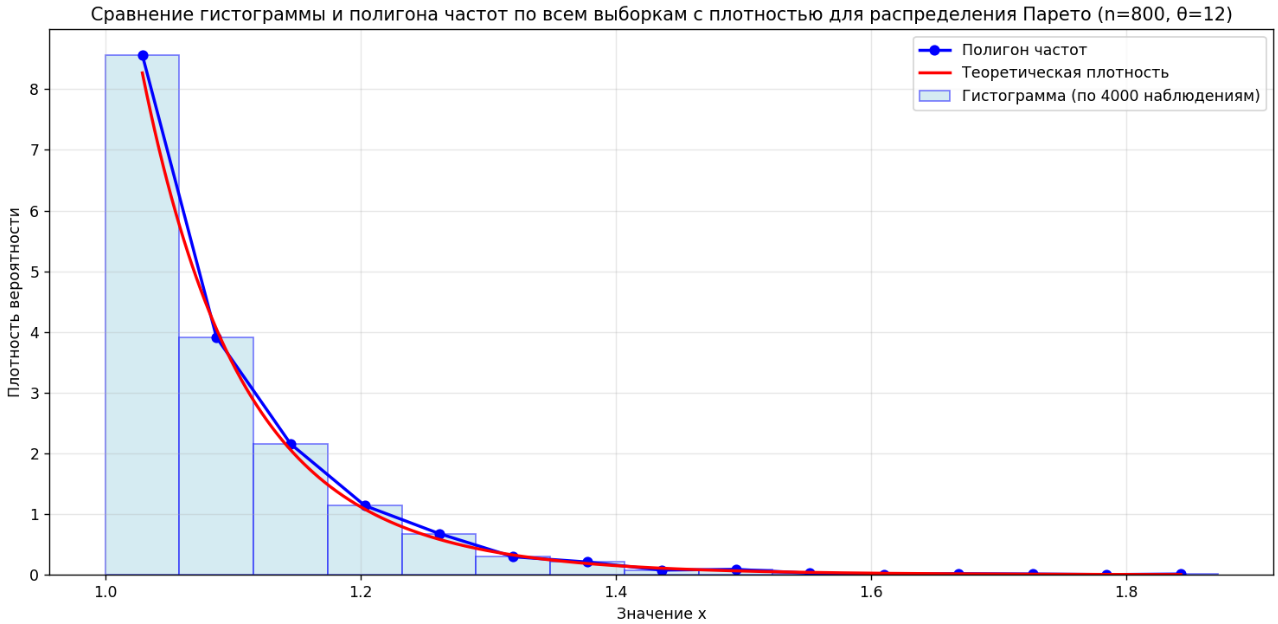


Рис. 32: График для выборки непрерывного распределения объёма 800

Рис. 33: График для выборки непрерывного распределения объёма 1000

4. Вычисление выборочных моментов

Вычислим значения выборочного среднего $\overline{X}$ и выборочной дисперсии $\overline{S}$ для каждой серии выборок, а также покажем среднее арифметическое по следующим формулам:

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

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

1. **Дискретное распределение (дискретное равномерное I)**

   (a) Объем выборки $n = 5$:

   - ▷ Серия 1: $\overline{X} = 65.600000$, $\overline{S}^2 = 958.240000$
   - ▷ Серия 2: $\overline{X} = 57.400000$, $\overline{S}^2 = 1297.840000$
   - ▷ Серия 3: $\overline{X} = 50.000000$, $\overline{S}^2 = 653.200000$
   - ▷ Серия 4: $\overline{X} = 50.400000$, $\overline{S}^2 = 723.040000$
   - ▷ Серия 5: $\overline{X} = 49.800000$, $\overline{S}^2 = 1778.160000$
   - ▷ Среднее арифметическое: $\overline{X} = 54.640000$, $\overline{S}^2 = 1082.096000$

   (b) Объем выборки $n = 10$:

   - ▷ Серия 1: $\overline{X} = 67.000000$, $\overline{S}^2 = 924.800000$
   - ▷ Серия 2: $\overline{X} = 63.100000$, $\overline{S}^2 = 1310.090000$
   - ▷ Серия 3: $\overline{X} = 53.500000$, $\overline{S}^2 = 905.450000$
   - ▷ Серия 4: $\overline{X} = 69.500000$, $\overline{S}^2 = 1027.650000$
   - ▷ Серия 5: $\overline{X} = 63.700000$, $\overline{S}^2 = 1870.410000$
   - ▷ Среднее арифметическое: $\overline{X} = 63.360000$, $\overline{S}^2 = 1207.680000$

   (c) Объем выборки $n = 100$:

   - ▷ Серия 1: $\overline{X} = 62.940000$, $\overline{S}^2 = 1267.576400$
   - ▷ Серия 2: $\overline{X} = 66.390000$, $\overline{S}^2 = 1142.657900$
   - ▷ Серия 3: $\overline{X} = 56.330000$, $\overline{S}^2 = 986.661100$
   - ▷ Серия 4: $\overline{X} = 55.920000$, $\overline{S}^2 = 1102.873600$
   - ▷ Серия 5: $\overline{X} = 60.100000$, $\overline{S}^2 = 1150.630000$
   - ▷ Среднее арифметическое: $\overline{X} = 60.336000$, $\overline{S}^2 = 1130.079800$

   (d) Объем выборки $n = 200$:

- ▷ Серия 1: $\overline{X} = 63.095000$, $\overline{S}^2 = 1296.085975$
- ▷ Серия 2: $\overline{X} = 65.730000$, $\overline{S}^2 = 1137.147100$
- ▷ Серия 3: $\overline{X} = 59.915000$, $\overline{S}^2 = 1063.717775$
- ▷ Серия 4: $\overline{X} = 58.850000$, $\overline{S}^2 = 1158.137500$
- ▷ Серия 5: $\overline{X} = 58.430000$, $\overline{S}^2 = 1122.555100$
- ▷ Среднее арифметическое: $\overline{X} = 61.204000$, $\overline{S}^2 = 1155.528690$

(e) Объем выборки $n = 400$:

- ▷ Серия 1: $\overline{X} = 62.352500$, $\overline{S}^2 = 1228.473244$
- ▷ Серия 2: $\overline{X} = 62.042500$, $\overline{S}^2 = 1176.055694$
- ▷ Серия 3: $\overline{X} = 59.455000$, $\overline{S}^2 = 1133.167975$
- ▷ Серия 4: $\overline{X} = 60.177500$, $\overline{S}^2 = 1229.290994$
- ▷ Серия 5: $\overline{X} = 60.700000$, $\overline{S}^2 = 1135.255000$
- ▷ Среднее арифметическое: $\overline{X} = 60.945500$, $\overline{S}^2 = 1180.448581$

(f) Объем выборки $n = 600$:

- ▷ Серия 1: $\overline{X} = 62.131667$, $\overline{S}^2 = 1211.747664$
- ▷ Серия 2: $\overline{X} = 62.261667$, $\overline{S}^2 = 1168.146531$
- ▷ Серия 3: $\overline{X} = 60.068333$, $\overline{S}^2 = 1161.666997$
- ▷ Серия 4: $\overline{X} = 61.201667$, $\overline{S}^2 = 1193.494331$
- ▷ Серия 5: $\overline{X} = 59.938333$, $\overline{S}^2 = 1150.767864$
- ▷ Среднее арифметическое: $\overline{X} = 61.120333$, $\overline{S}^2 = 1177.164677$

(g) Объем выборки $n = 800$:

- ▷ Серия 1: $\overline{X} = 60.845000$, $\overline{S}^2 = 1209.655975$
- ▷ Серия 2: $\overline{X} = 62.912500$, $\overline{S}^2 = 1128.404844$
- ▷ Серия 3: $\overline{X} = 59.812500$, $\overline{S}^2 = 1167.297344$
- ▷ Серия 4: $\overline{X} = 61.686250$, $\overline{S}^2 = 1233.515311$
- ▷ Серия 5: $\overline{X} = 60.062500$, $\overline{S}^2 = 1186.521094$
- ▷ Среднее арифметическое: $\overline{X} = 61.063750$, $\overline{S}^2 = 1185.078913$

(h) Объем выборки $n = 1000$:

- ▷ Серия 1: $\overline{X} = 60.820000$, $\overline{S}^2 = 1194.321600$
- ▷ Серия 2: $\overline{X} = 63.521000$, $\overline{S}^2 = 1135.825559$
- ▷ Серия 3: $\overline{X} = 60.201000$, $\overline{S}^2 = 1173.532599$
- ▷ Серия 4: $\overline{X} = 61.075000$, $\overline{S}^2 = 1237.039375$
- ▷ Серия 5: $\overline{X} = 60.370000$, $\overline{S}^2 = 1206.991100$

- ▷ Среднее арифметическое: $\overline{X} = 61.197400$, $\overline{S}^2 = 1189.542047$

Как мы видим, данные значения сходятся к значениям теоретических математического ожидания и дисперсии для равномерного распределения с заданным параметром $\theta = 121$:

$$E[X] = \frac{\theta + 1}{2} = \frac{122}{2} = 61$$
$$D[X] = \frac{\theta^2 - 1}{12} = \frac{121^2 - 1}{12} = 1220$$

2. **Непрерывное распределение (распределение Парето)**

   (a) Объем выборки $n = 5$:

   - ▷ Серия 1: $\overline{X} = 1.096638$, $\overline{S}^2 = 0.012632$
   - ▷ Серия 2: $\overline{X} = 1.161239$, $\overline{S}^2 = 0.010834$
   - ▷ Серия 3: $\overline{X} = 1.113574$, $\overline{S}^2 = 0.003615$
   - ▷ Серия 4: $\overline{X} = 1.115615$, $\overline{S}^2 = 0.018546$
   - ▷ Серия 5: $\overline{X} = 1.101622$, $\overline{S}^2 = 0.009451$
   - ▷ Среднее арифметическое: $\overline{X} = 1.117738$, $\overline{S}^2 = 0.011016$

   (b) Объем выборки $n = 10$:

   - ▷ Серия 1: $\overline{X} = 1.088552$, $\overline{S}^2 = 0.007757$
   - ▷ Серия 2: $\overline{X} = 1.119158$, $\overline{S}^2 = 0.008009$
   - ▷ Серия 3: $\overline{X} = 1.115099$, $\overline{S}^2 = 0.009792$
   - ▷ Серия 4: $\overline{X} = 1.102710$, $\overline{S}^2 = 0.013635$
   - ▷ Серия 5: $\overline{X} = 1.080461$, $\overline{S}^2 = 0.008443$
   - ▷ Среднее арифметическое: $\overline{X} = 1.101196$, $\overline{S}^2 = 0.009527$

   (c) Объем выборки $n = 100$:

   - ▷ Серия 1: $\overline{X} = 1.089071$, $\overline{S}^2 = 0.010034$
   - ▷ Серия 2: $\overline{X} = 1.079443$, $\overline{S}^2 = 0.007204$
   - ▷ Серия 3: $\overline{X} = 1.086723$, $\overline{S}^2 = 0.009738$
   - ▷ Серия 4: $\overline{X} = 1.097229$, $\overline{S}^2 = 0.012672$
   - ▷ Серия 5: $\overline{X} = 1.084949$, $\overline{S}^2 = 0.006752$
   - ▷ Среднее арифметическое: $\overline{X} = 1.087483$, $\overline{S}^2 = 0.009280$

   (d) Объем выборки $n = 200$:

   - ▷ Серия 1: $\overline{X} = 1.093842$, $\overline{S}^2 = 0.009941$

- ▷ Серия 2: $\overline{X} = 1.079821$, $\overline{S}^2 = 0.006580$
- ▷ Серия 3: $\overline{X} = 1.093554$, $\overline{S}^2 = 0.008991$
- ▷ Серия 4: $\overline{X} = 1.094386$, $\overline{S}^2 = 0.012409$
- ▷ Серия 5: $\overline{X} = 1.079574$, $\overline{S}^2 = 0.006223$
- ▷ Среднее арифметическое: $\overline{X} = 1.088236$, $\overline{S}^2 = 0.008829$

(e) Объем выборки $n = 400$:

- ▷ Серия 1: $\overline{X} = 1.097680$, $\overline{S}^2 = 0.010980$
- ▷ Серия 2: $\overline{X} = 1.087006$, $\overline{S}^2 = 0.008392$
- ▷ Серия 3: $\overline{X} = 1.089870$, $\overline{S}^2 = 0.008313$
- ▷ Серия 4: $\overline{X} = 1.095797$, $\overline{S}^2 = 0.011176$
- ▷ Серия 5: $\overline{X} = 1.084453$, $\overline{S}^2 = 0.008297$
- ▷ Среднее арифметическое: $\overline{X} = 1.090961$, $\overline{S}^2 = 0.009432$

(f) Объем выборки $n = 600$:

- ▷ Серия 1: $\overline{X} = 1.095266$, $\overline{S}^2 = 0.010923$
- ▷ Серия 2: $\overline{X} = 1.090619$, $\overline{S}^2 = 0.008996$
- ▷ Серия 3: $\overline{X} = 1.090162$, $\overline{S}^2 = 0.008257$
- ▷ Серия 4: $\overline{X} = 1.095703$, $\overline{S}^2 = 0.010475$
- ▷ Серия 5: $\overline{X} = 1.085442$, $\overline{S}^2 = 0.007946$
- ▷ Среднее арифметическое: $\overline{X} = 1.091438$, $\overline{S}^2 = 0.009320$

(g) Объем выборки $n = 800$:

- ▷ Серия 1: $\overline{X} = 1.094757$, $\overline{S}^2 = 0.011002$
- ▷ Серия 2: $\overline{X} = 1.090426$, $\overline{S}^2 = 0.009490$
- ▷ Серия 3: $\overline{X} = 1.089691$, $\overline{S}^2 = 0.008615$
- ▷ Серия 4: $\overline{X} = 1.093521$, $\overline{S}^2 = 0.010102$
- ▷ Серия 5: $\overline{X} = 1.086769$, $\overline{S}^2 = 0.008515$
- ▷ Среднее арифметическое: $\overline{X} = 1.091033$, $\overline{S}^2 = 0.009545$

(h) Объем выборки $n = 1000$:

- ▷ Серия 1: $\overline{X} = 1.093026$, $\overline{S}^2 = 0.010568$
- ▷ Серия 2: $\overline{X} = 1.089544$, $\overline{S}^2 = 0.009233$
- ▷ Серия 3: $\overline{X} = 1.092028$, $\overline{S}^2 = 0.009322$
- ▷ Серия 4: $\overline{X} = 1.092811$, $\overline{S}^2 = 0.009920$
- ▷ Серия 5: $\overline{X} = 1.086495$, $\overline{S}^2 = 0.008089$
- ▷ Среднее арифметическое: $\overline{X} = 1.090781$, $\overline{S}^2 = 0.009427$

Как мы видим, данные значения сходятся к значениям теоретических математического ожидания и дисперсии для распределения Парето с заданным параметром $\theta = 12$:

$$E[X] = \frac{\theta}{\theta - 1} = \frac{12}{12 - 1} \approx 1.09091$$
$$D[X] = \frac{\theta}{(\theta - 2)(\theta - 1)^2} = \frac{12}{(12 - 2)(12 - 1)^2} \approx 0.009917$$

Выборочные моменты обладают следующими свойствами:

▷ Выборочное среднее $\overline{X}$:

- Несмещённость выполняется: выборочное среднее является несмещённой оценкой теоретического математического ожидания случайной величины
- Состоятельность выполняется: при $n \to \infty$ выборочное среднее сходится к теоретическому математическому ожиданию
- Эффективность выполняется: выборочное среднее имеет наименьшую дисперсию среди всех несмещённых линейных оценок

▷ Выборочная дисперсия $\overline{S}$:

- Несмещённость не выполняется
- Состоятельность выполняется: при $n \to \infty$ выборочная дисперсия сходится к теоретической дисперсии
- Эффективность не выполняется: оценка смещённая, поэтому не может быть эффективной

5. Приложения

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