

Санкт-Петербургский государственный университет

Кафедра системного программирования

Группа 23.Б10-мм

Графовые критерии проверки гипотез согласия

Похабов Иван Ильич

Отчёт по учебной практике
в форме «Теоретическое исследование»

Научный руководитель:
доцент кафедры системного программирования, к. ф.-м. н., Гориховский В. И.

Санкт-Петербург
2025

Оглавление

Введение	3
1. Постановка задачи	5
2. Обзор	6
2.1. Постановка задачи	6
2.2. Обзор существующих критериев согласия	6
2.3. Графовые методы проверки гипотезы согласия	7
2.4. Выводы	7
3. Реализация исследовательского окружения	8
3.1. Построение графа	9
3.2. Визуализация графов	9
3.3. Анализ графовых характеристик	10
4. Исследование	11
5. Эксперимент	13
5.1. Дизайн эксперимента	13
5.2. Условия эксперимента	13
5.3. Исследовательские вопросы	13
5.4. Результаты	14
Заключение	18
Список литературы	19

Введение

Математическая статистика применяется в большом спектре задач, например, при проверке биологических, медицинских и прочих гипотез:

1. *Evaluating the statistical power of goodness-of-fit tests for health and medicine survey data* [4] — статья анализирует эффективность различных критериев согласия в контексте анализа данных, полученных в медицинских и здравоохранительных опросах.
2. *Impact of Criticism of Null-Hypothesis Significance Testing on Statistical Reporting Practices in Conservation Biology* [2] — статья рассматривает влияние критики традиционного подхода к проверке нулевой гипотезы на методы статистической отчетности в области охраны природы.
3. *The Scientology of Hypothesis Testing in Empirical Research: Emphasizing Economic Significance* [3] — статья исследует подходы к проверке гипотез, используемые в различных научных и социальных науках, особенно в экономике, и акцентирует внимание на значимости статистических результатов с точки зрения их реального воздействия.

Одной из таких задач является проверка гипотезы согласия. Критерием согласия называют функцию, которая по выборке выдаёт вероятность опровергнуть гипотезу согласия. Существует множество критериев, например, критерий Колмогорова-Смирнова или критерий хи-квадрат. Но ни один из них не является идеальным, поэтому постоянно появляются новые критерии.

Графовый подход в статистике уже применяется, например, для решения задачи обнаружения разладки [1]. Однако в научной литературе не встречается его применения для формирования новых критериев согласия. В данном случае мы будем рассматривать графы, в которых вершины соотносятся с наблюдениями, а ребра между ними показывают некоторое отношение между наблюдениями. Подход является перспек-

тивным, так как графы позволяют найти визуальные отличия в структуре различных распределений, а так же легко собирать различные статистические данные.

В рамках данной работы ставится задача изучения взаимосвязи графовых характеристик и различных функций распределения для оценки применимости данного подхода к созданию новых критериев согласия.

1. Постановка задачи

Целью семестровой работы является экспериментальный анализ применимости простых численных характеристик дистанционных графов как статистик в критериях согласия

Задачи:

- Выбор графовых статистик
- Создание экспериментального окружения
 - генерации и визуализации дистанционных графов по выборке
 - вычисления численных характеристик графов
 - проведения Монте-Карло моделирования для создания собственных критериев
- Проведение корреляционного анализа между графовыми характеристиками и семействами распределений

2. Обзор

2.1. Постановка задачи

Пусть дана выборка $X_1, X_2, \dots, X_n \sim P$. Гипотеза $H_0 : P \sim P_0$, где P_0 — это известное распределение, называется простой гипотезой согласия. Критерием согласия называют функцию, которая по выборке выдаёт вероятность опровергнуть гипотезу согласия.

2.2. Обзор существующих критериев согласия

Существует 2 вида критериев согласия: параметрические и непараметрические. Параметрические критерии предполагают, что данные следуют определённому распределению и требуют, чтобы параметры этого распределения были известны или оценены. Эти тесты наиболее эффективны, когда предположения о распределении данных выполняются. Примерами таких распределений являются:

- **Критерий Хи-квадрат.**
- **Критерий Колмогорова-Смирнова.**
- **Критерий Шапиро-Уилка.**

Этот вид имеет ряд минусов таких как:

1. *Предположение о распределении данных.*
2. *Необходимость в больших выборках.*
3. *Необходимость в точном определении параметров.*
4. *Трудности с интерпретацией при несовпадении с реальной ситуацией.*

Непараметрические критерии не требуют строгих предположений о распределении данных и их параметрах. Эти тесты намного более универсальны и подходят в тех случаях, когда распределение данных неизвестно. Примерами таких распределений являются:

- Критерий Манна-Уитни.
- Критерий Колмогорова-Смирнова.
- Критерий Краскала-Уоллиса.

Этот вид распределений так же имеет свои недостатки:

1. *Меньшая мощность теста по сравнению с параметрическими.*
2. *Ограниченная возможность для детальной интерпретации.*
3. *Трудности с интерпретацией при больших выборках.*

Так же в мире постоянно возникают новые критерии, потому что ни один не является идеальным, а также каждый из них имеет свою специфику.

2.3. Графовые методы проверки гипотезы согласия

Графовые методы уже применяются в более сложных статистических задачах, например, для обнаружения разладки [1]. Но для проверки гипотезы согласия ещё не было никаких опубликованных исследований.

Подход, основанный на графах, будет непараметрическим и может быть применён к любому набору данных. Он может стать более мощным, чем непараметрические методы, при этом не требуя точных определений параметров как параметрические.

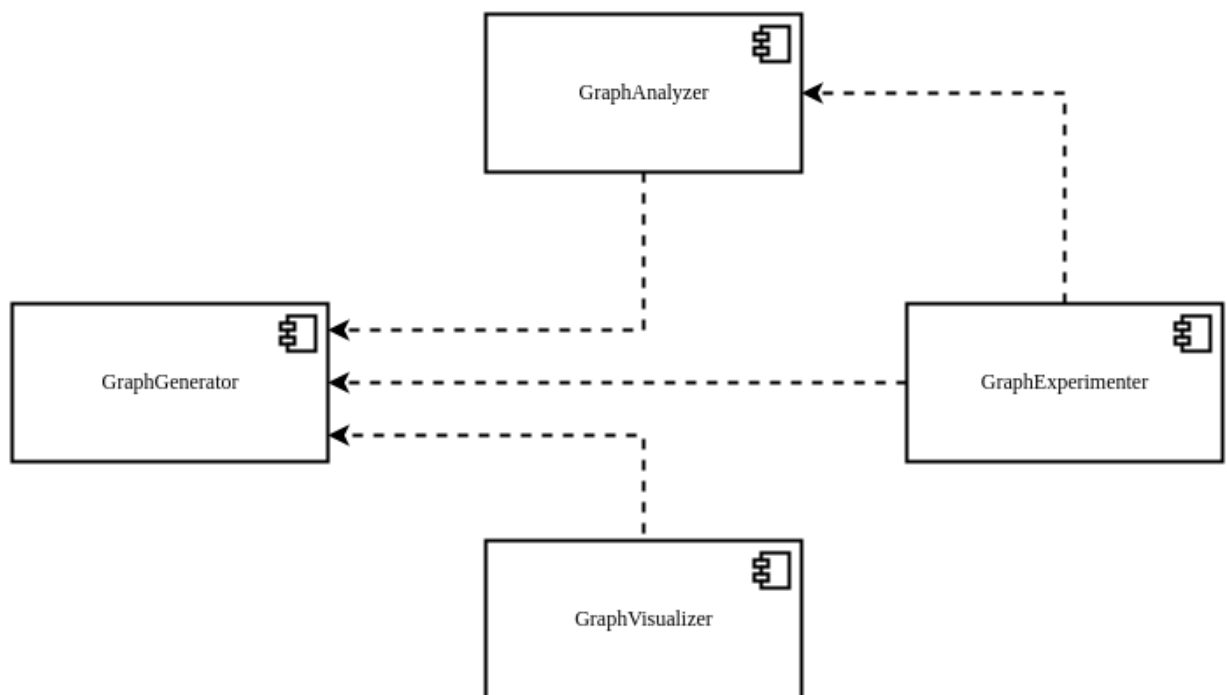
2.4. Выводы

Существует множество различных методов проверки гипотез согласия, которые применяются в зависимости от задачи и вида данных. Графовые критерии могут прийти на замену ранее известным методам в каких-то задачах.

3. Реализация исследовательского окружения

В данном разделе представлены основные моменты разработки инструмента, для построения и отрисовки графов и анализа графовых данных.

Для разработки был выбран язык *Python*, потому что работа выполняется в рамках проекта **PySATL**, в котором он является основным языком.



Архитектура окружения

В ходе работы было создано 4 модуля для

- генерации дистанционных графов по выборке
- визуализации графов
- вычисления графовых статистик
- постановки экспериментов

3.1. Построение графа

В работе используются дистанционные неориентированные графы, которые строятся по массиву вещественных чисел (выборке) и заданному расстоянию. Элемент массива будет являться вершиной. Ребра в графе будут строиться по правилу: Между вершинами x_1 и x_2 существует ребро $\Leftrightarrow |x_1 - x_2| < \text{dist}$.

Для этой задачи был реализован модуль *generation*, который по заданному распределению генерирует выборку, а после строит граф в виде списка рёбер.

3.2. Визуализация графов

Важная часть исследования — работа с видимым представлением графов, так как она позволяет определить многие особенности с помощью зрительной оценки.

Для визуализации графов был создан модуль *draw*, использующий *CLI*. Аргументом принимается функция распределения, а необязательными опциями являются количество графов, сдвиг распределения, количество элементов генерируемой выборки, и строка, которая хранит в себе путь и название картинок, которые будут созданы, относительно модуля.

Для отрисовки графов была использована библиотека *pyvis.network*, генерирующая *html-файлы*, которые при открытии создают визуализацию графов с интерактивным интерфейсом, который позволяет изменять положение ребра и вершин и многое другое.

Выбор данной библиотеки обусловлен простотой использования и функциональностью, которая позволяет физически изменять представление графа, так как это важно для детального рассмотрения и выявления различных численных характеристик, например для поиска вершин, имеющих ровно 1 ребро.

3.3. Анализ графовых характеристик

Основная часть работы, проделанная в этом семестре заключается в изучении влияния выбора распределения на различные графовые характеристики. Для этого реализован модуль *analyze*, который работает с помощью *CLI* и позволяет вычислять средние значения реализованных в нём функций, подсчитывающих данных.

Вычисления характеристик проводится с помощью указанной ранее библиотеки *networkx*. Преимуществами библиотеки являются:

- Качественная документация с большим количеством примерами использования
- Совместимость с библиотеками *matplotlib* и *pyvis.network*, использующимися в работе
- Большое количество реализованных функций вычисления графовых численных характеристик

4. Исследование

В начале работы были выбраны следующие распределения, данные которых будут сравниваться:

- Нормальное
- Равномерное непрерывное
- Экспоненциальное

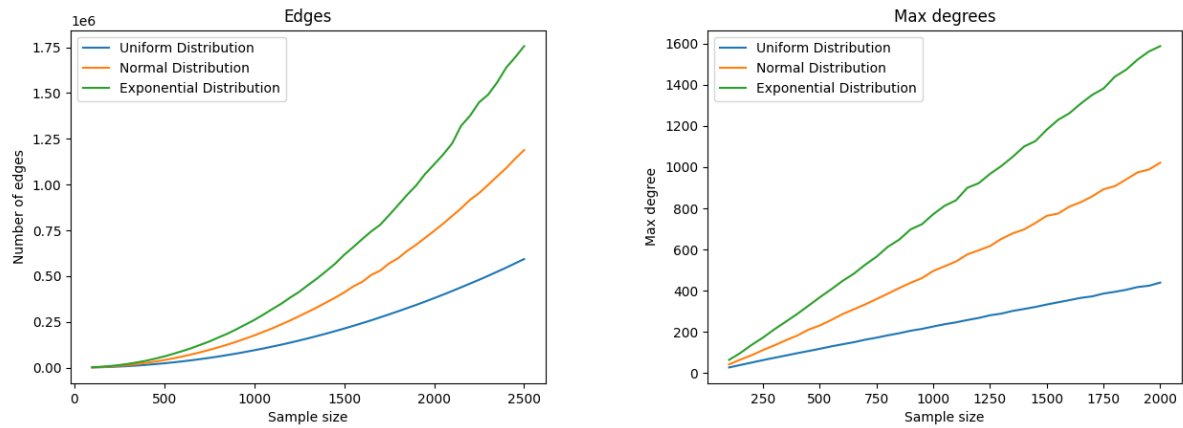
Для изучения влияния различных распределений на дистанционные графы были выбраны следующие характеристики:

- Количество рёбер
- Максимальная степень вершины
- Количество компонент связности

Эти показатели были выбраны из-за визуальных при первичном зрительном анализе графов с данными распределениями. В данной работе для построения был выбран параметр дистанции, вычисляемый для каждой выборке по следующей формуле: $dist = \frac{max(S)-min(S)}{const}$, которая позволяет нормировать значения для графов сгенерированных с помощью функций распределения, принадлежащими одному семейству, но с разными параметрами (в данной работе все исследования и эксперименты проводились для $const = 10$).

После проведения ряда измерений было принято решение отложить изучение количества компонент связности, так как не был подобран параметр дистанции, который бы показывал явную корреляцию между данными и выбранным распределением.

Были вычислены средние значения выбранных характеристик для выборок размером 100...2500 с шагом 50 используя выбранные распределения со случайными параметрами. Графики количества рёбер и максимальной степени вершины:



На графиках видно, что данные функции явно зависят от семейства распределения и могут быть полезны для графовых критериев согласия:

- Полученный график максимальной степени вершины позволяет сказать, что эта характеристика может быть представлена как линейная функция относительно размера выборки
- Полученный график количества ребер позволяет сказать, что эта характеристика может быть представлена как квадратичная функция относительно размера выборки

На данном этапе эти функции для каждого семейства были аппроксимированы *методом наименьших квадратов*, встроенным в *Python* в библиотеку *numpy*, как линейная и квадратичная функции без доказательств.

5. Эксперимент

5.1. Дизайн эксперимента

Оценим отклонение реальных значений характеристик от полученных функций средних значений, а так же проверим как на это влияет параметр масштаба.

5.2. Условия эксперимента

- Версии библиотек и инструментов

- *Python* — 3.12

- *networkx* — 3.3

- *numpy* — 2.1.1

- *scipy* — 1.14.1

- Параметры распределений

- 1. $loc = 0, scale = 1$

- 2. $loc = 0, scale = 100$

- 3. $loc = 0, scale = 10^9$

- Параметр дистанции $dist = \frac{max(S) - min(S)}{10}$, где S -выборка

- Размеры выборок — 100, 200, 500, 1000

- Количество выборок — 10000

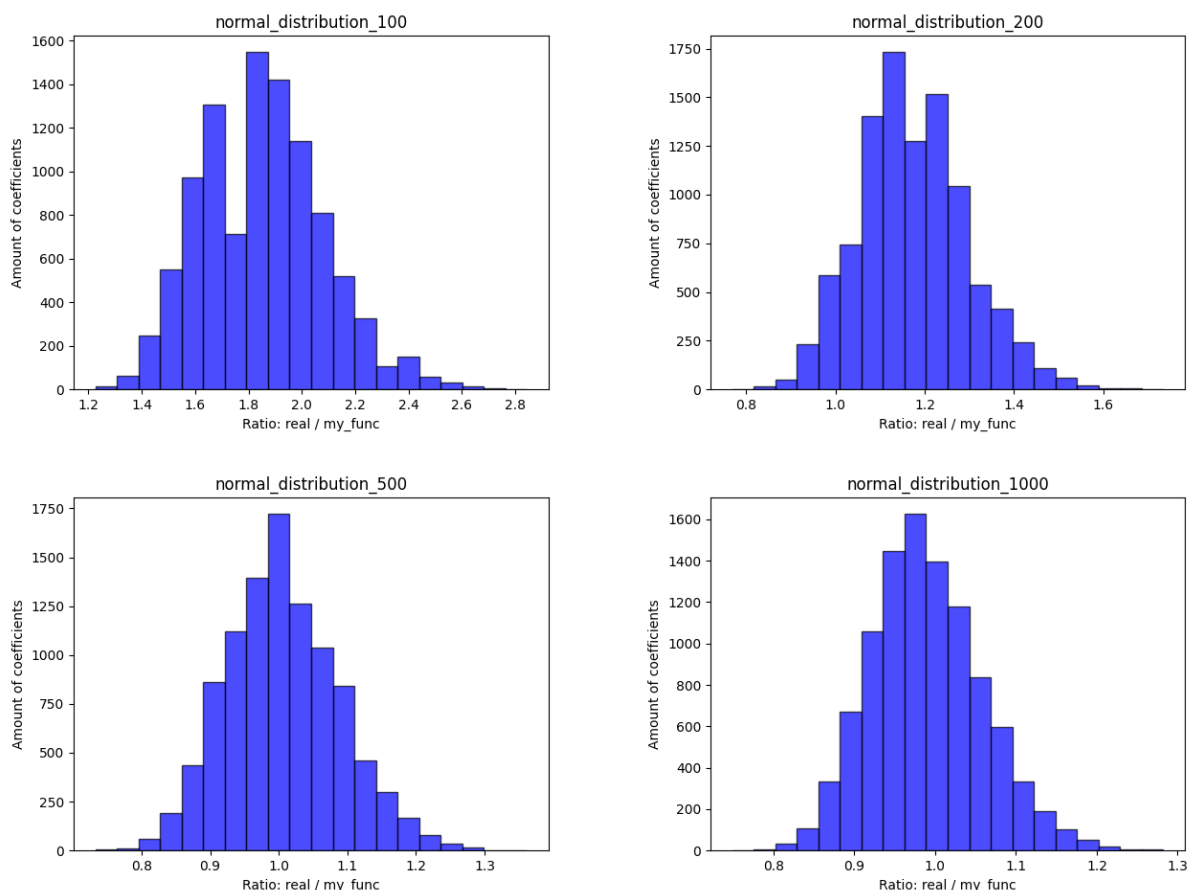
5.3. Исследовательские вопросы

RQ1 : каково отклонение реальных значений характеристик от среднего значения в зависимости от размера выборки?

RQ2 : как нормировка графа, с помощью параметра дистанции, влияет на получаемые значения?

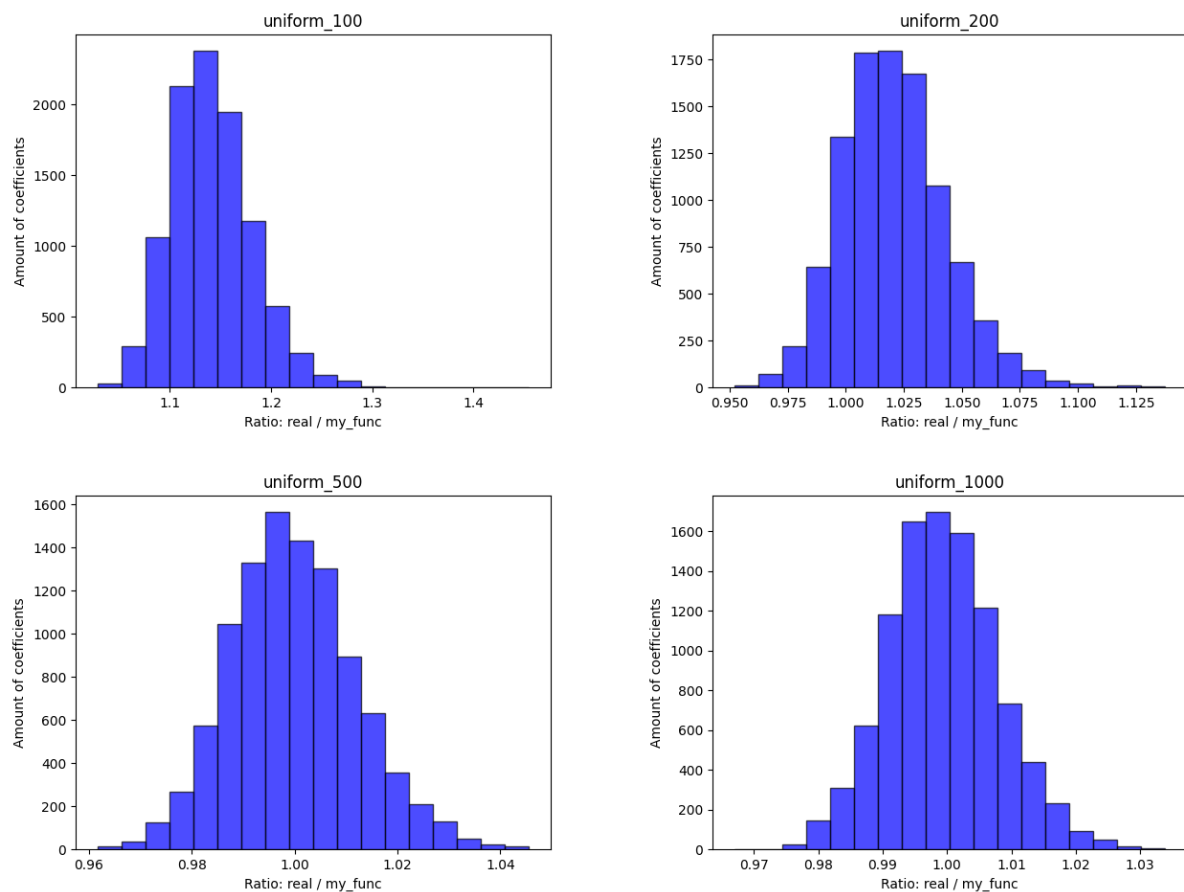
5.4. Результаты

По результате эксперимента было выявлено, что максимальная степень вершины и количество рёбер в графе имеют какое-то предельное распределение, а так же для каждого семейства имеют свои пороги. При увеличении размеров выборки разброс уменьшается, о чём свидетельствуют следующие графики:



Гистограммы распределения коэффициентов отношения
максимальных степеней вершин стандартного нормального
распределения

Лучшую точность показала функция количества рёбер для равномерного распределения. Разброс значений относительно неё в разы меньше, чем у других семейств или функции максимальной степени вершины.



Гистограммы распределения коэффициентов отношения числа ребер стандартного равномерного распределения

5.4.1. RQ1

	Нормальное	Равномерное	Экспоненциальное
100	1.2-2.8	0.7-1.2	0.8-2.2
200	0.8-1.6	0.8-1.2	0.7-1.5
500	0.8-1.3	0.9-1.2	0.7-1.3
1000	0.8-1.2	0.9-1.2	0.8-1.2

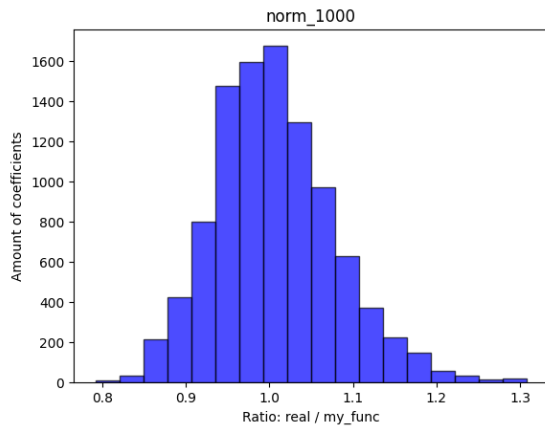
Таблица 1: Примерный разброс значений максимальной степени вершины со стандартными параметрами

	Нормальное	Равномерное	Экспоненциальное
100	0.2-0.45	1.00-1.30	0.15-0.45
200	0.60-1.20	0.95-1.12	0.51-1.35
500	0.80-1.35	0.96-1.05	0.78-1.60
1000	0.80-1.3	0.97-1.03	0.76-1.50

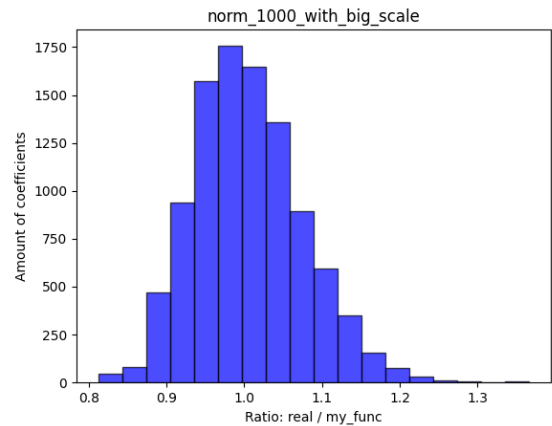
Таблица 2: Примерный разброс значений числа ребер со стандартными параметрами

5.4.2. RQ2

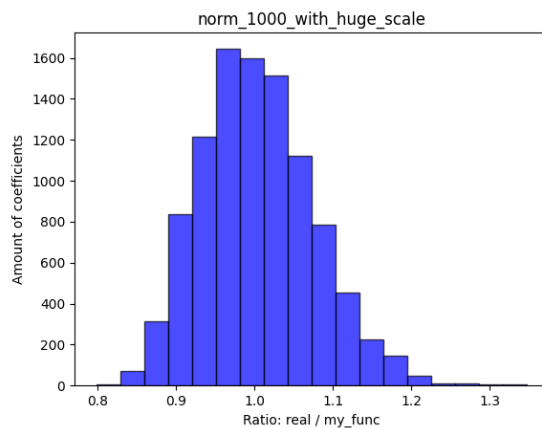
Аналитически выявлено, что нормировка с помощью формулы дистанции для графа идеально нормирует только выборки равномерного распределения, для других распределений проводился численный эксперимент. Параметр масштаба действительно влияет на отклонение реальных значений от функции средних значений для нормальных и экспоненциальных распределений, учитывая нормировку, но при этом практически не изменяет пороги, что можно увидеть на следующих графиках:



scale=1

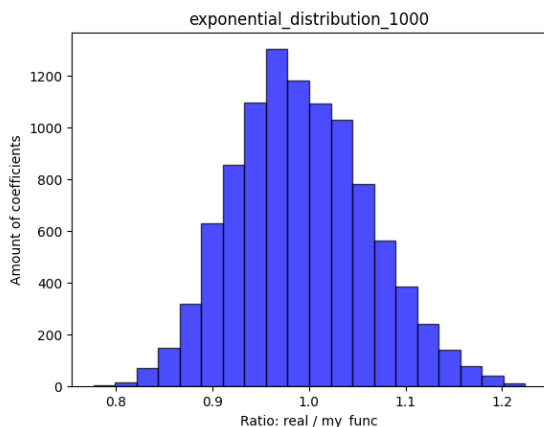


scale=100

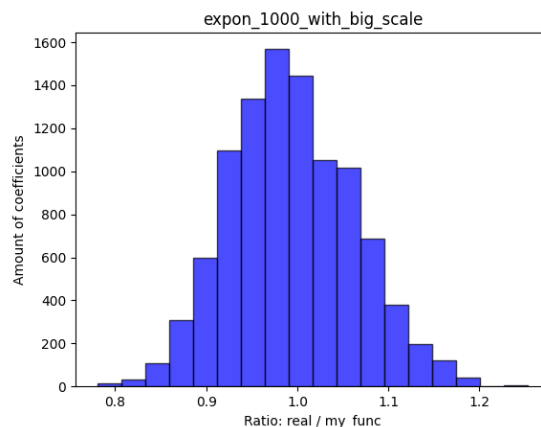


scale= 10^9

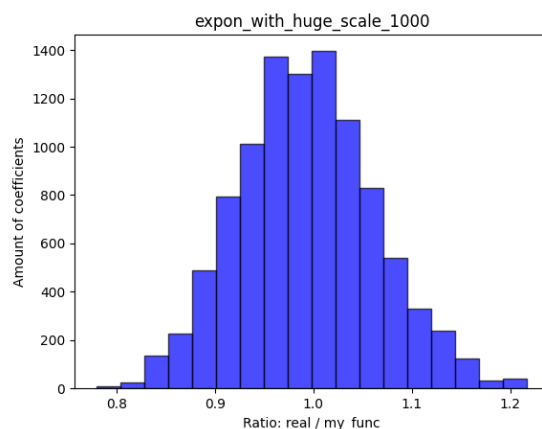
Гистограммы распределения коэффициентов отношения числа ребер нормального распределения для выборок размером 1000



scale=1



scale=100



scale= 10^9

Гистограммы распределения коэффициентов отношения максимальной степени вершины экспоненциального распределения для выборок размером 1000

Заключение

В ходе данной работы были выполнены следующие задачи:

- Произведен выбор графовых характеристик для исследования
- Реализован прототип экспериментального окружения¹
 - генерации и визуализации дистанционных графов по выборке
 - вычисления численных характеристик графов
 - проведения Монте-Карло моделирования для создания собственных критериев
- Проведено экспериментальное исследование корреляции между численными характеристиками дистанционных графов и семействами распределения
- Проведен эксперимент оценивающий отклонение реальных значений характеристик от полученных функций средних значений и влияние параметра масштаба

¹<https://github.com/Ivan-Pokhabov/GraphResearch-ForCheckingCriterionOfGoodnessOfFit>

Список литературы

- [1] Chen Hao, Zhang Nancy. Graph-based change-point detection. — URL: <https://www.projecteuclid.org/journals/annals-of-statistics/volume-43/issue-1/Graph-based-change-point-detection/10.1214/14-AOS1269.full> (дата обращения: 12 ноября 2024 г.).
- [2] Fidler Fiona, Burgman Mark A., Cumming Geoff, Buttrose Robert. Impact of Criticism of Null-Hypothesis Significance Testing on Statistical Reporting Practices in Conservation Biology. — URL: https://www.researchgate.net/publication/6791353_Impact_of_Criticism_of_Null-Hypothesis_Significance_Testing_on_Statistical_Reporting_Practices_in_Conservation_Biology (дата обращения: 12 ноября 2024 г.).
- [3] Mayondi Moses, Mulenga Richard. The Scientology of Hypothesis Testing in Empirical Research: Emphasizing Economic Significance. — URL: https://www.researchgate.net/publication/377757271_The_Scientology_of_Hypothesis_Testing_in_Empirical_Research_Emphasizing_Economic_Significance (дата обращения: 12 ноября 2024 г.).
- [4] Steele Michael Craig, Neil A Smart Cameron Paul Hurst, Chaseling Janet. Evaluating the statistical power of goodness-of-fit tests for health and medicine survey data. — URL: https://www.researchgate.net/publication/44929281_Evaluating_the_statistical_power_of_goodness-of-fit_tests_for_health_and_medicine_survey_data (дата обращения: 12 ноября 2024 г.).