Исследование свойств случайных графов на выборках из распределений и проверка гипотез

Цель

Построить случайные графы на основе выборок из различных распределений и исследовать поведение числовых характеристик графа в зависимости от параметров построения. Кроме того, реализовать проверку гипотезы H_0 против H_1 на основе статистик графа.

Используемые распределения

- **Экспоненциальное распределение:** $\mathrm{Exp}(\lambda)$, где $\lambda = \frac{1}{\sqrt{e^2 - e}}$;
- **Логнормальное распределение:** $\mathrm{LogNormal}(0, \sigma = 1)$;
- **Нормальное распределение:** $\mathrm{Normal}(0, \sigma = 1)$:
- **Несимметричное нормальное распределение:** $\mathrm{SkewNormal}(\sigma = 1)$

Функции и методы

1. Построение графов

- **build_knn_graph(data, k)** — построение графа ближайших соседей (kNN): каждая вершина соединяется с k ближайшими по расстоянию.
- **build_distance_graph(data, d)** — построение графа расстояния: вершины соединяются ребром, если $|x_i - x_j| \leq d$.

2. Характеристики графа

- $\delta(G)$ — минимальная степень вершины в графе;
- $\chi(G)$ — приближенное хроматическое число (оценка с помощью жадной раскраски);
- $\Delta(G)$ — максимальная степень вершины в графе;
- $\alpha(G)$ — размер максимального независимого множества.

Эксперименты

Грид-серч по параметрам графа

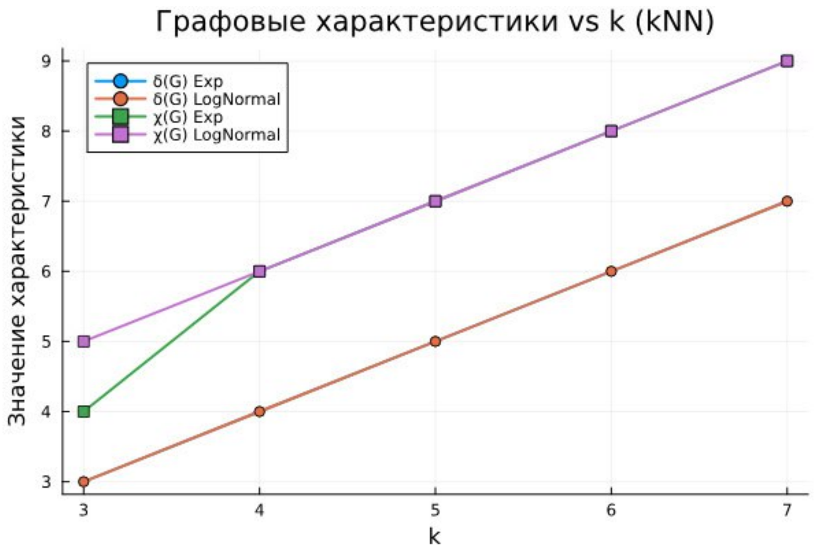
Были проведены переборы по параметрам:

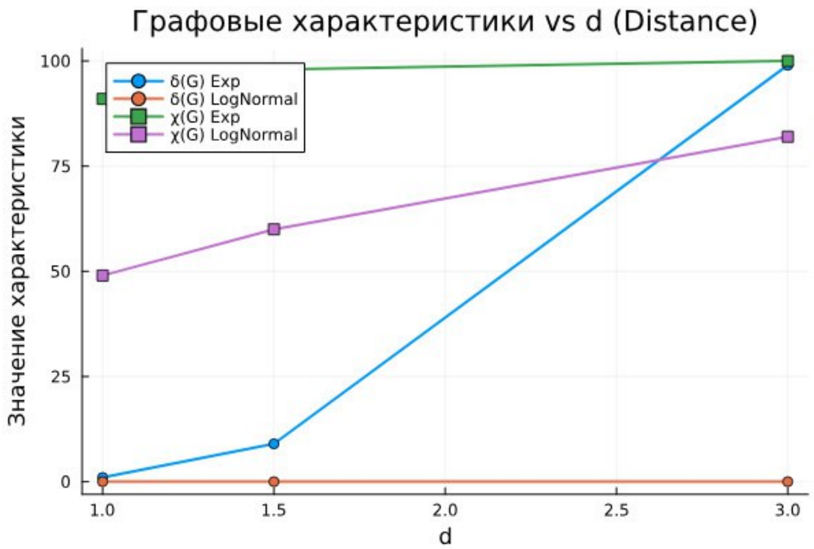
- $k \in \{3, 4, 5, 6, 7\}$ для kNN-графов;
- $d \in \{1.0, 1.5, 3.0\}$ для dist-графов.

Для каждого значения параметра строились графы на выборках из $\text{Exp}(\lambda_0)$ и $\text{LogNormal}(0, 1)$, после чего вычислялись $\delta(G)$ и $\chi(G)$.

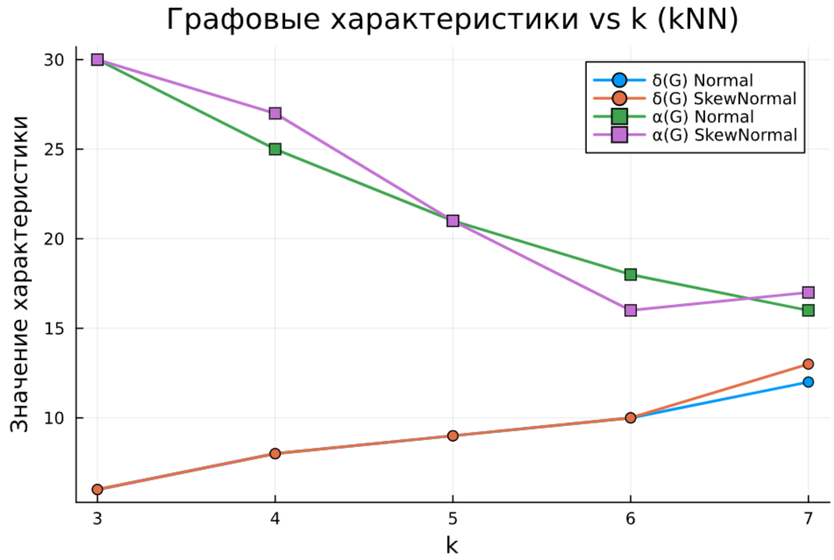
Результаты

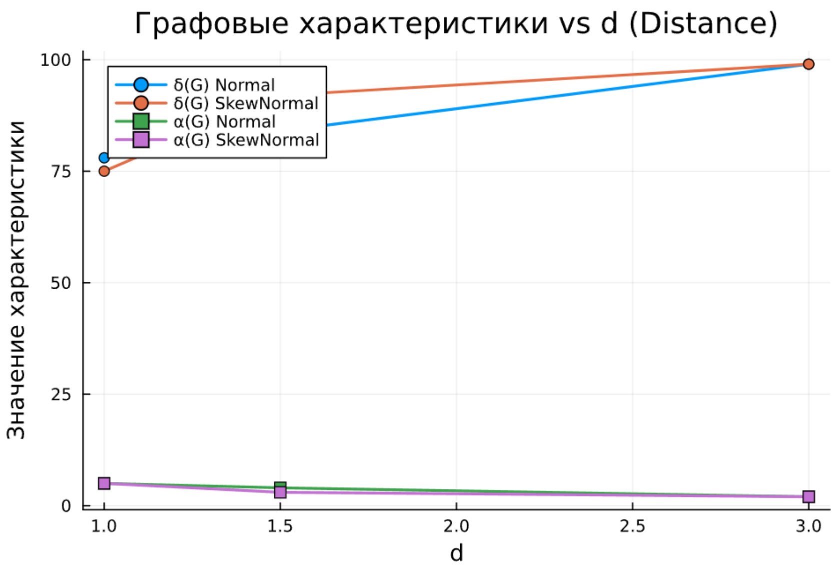
Графовые характеристики существенно различаются для разных распределений, особенно при росте k или d. Логнормальное распределение, как правило, даёт более плотные графы с большими $\chi(G)$ и $\delta(G)$.





Асимметричное распределение формирует графы с большей плотностью связей и меньшими независимыми множествами, чем нормальное распределение.





Проверка гипотез

Описание

Рассматривается задача проверки гипотезы:

$$H_0 : \xi \sim f(x, \theta) \quad \text{vs} \quad H_1 : \xi \sim h(x, \nu),$$

где f и h — плотности экспоненциального/нормального и логнормального/ассиметричного распределений соответственно.

Методика

1. Генерируется $N = 1000$ выборок из H_0 (Exp), строятся графы и считается $T(G)$;

2. Вычисляется критическое значение $T^* = \text{quantile}(T_{H_0}, \alpha)$ при уровне значимости $\alpha = 0.05$;

3. Считается доля выборок из H_1 (LogNormal), у которых $T(G) < T^*$ — это оценка мощности критерия.

Результаты проверки

При $k = 5$ и $n = 100$ для kNN-графа обе гипотезы не отвергаются.

30 мая 2025 г.

Анализ результатов классификации (Normal и SkewNormal)

1. Средние значения характеристик Δ и α

- **Δ:** для обоих распределений наблюдается рост значений Δ с увеличением размера выборки. Однако значения Δ для SkewNormal практически совпадают со значениями для Normal и заметно ниже соответствующего порогового значения Δ^*, рассчитанного по правому квантилю распределения при H_0.
  - Это говорит о том, что характеристика Δ не чувствительна к различиям между $\mathrm{Normal}(0,1)$ и $\mathrm{SkewNormal}(0,1,\xi=5)$.
- **α:** значения α растут линейно с n, как и ожидалось. Между Normal и SkewNormal различия минимальны, но α для SkewNormal в среднем чуть ниже. Пороговое значение α^* устанавливается как левый квантиль уровня $\alpha = 0.05$.
  - Различие между α-значениями наблюдается, но не является выраженным.

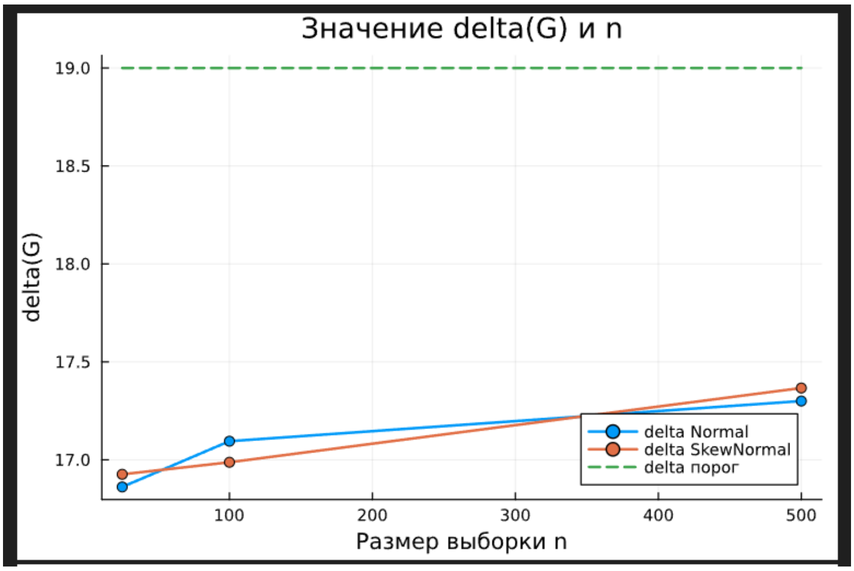
2. Мощность критериев

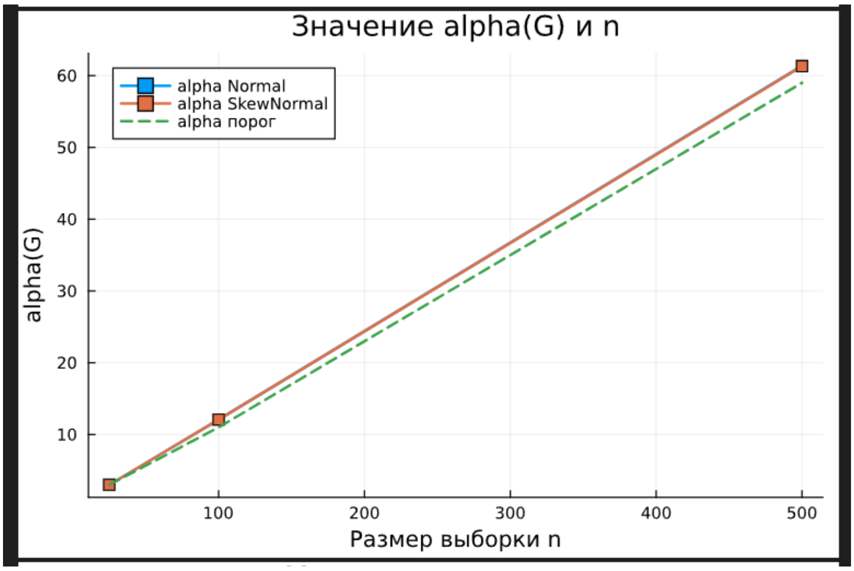
- **Мощность критерия на основе Δ:** во всех случаях остаётся близкой к нулю (не превышает 0.0115), что свидетельствует о его неэффективности в задаче различения Normal и SkewNormal.
- **Мощность критерия на основе α:** возрастает с увеличением n, достигая 0.048 при $n = 500$, что демонстрирует некоторое различие, но всё ещё недостаточное для уверенного отклонения H_0.

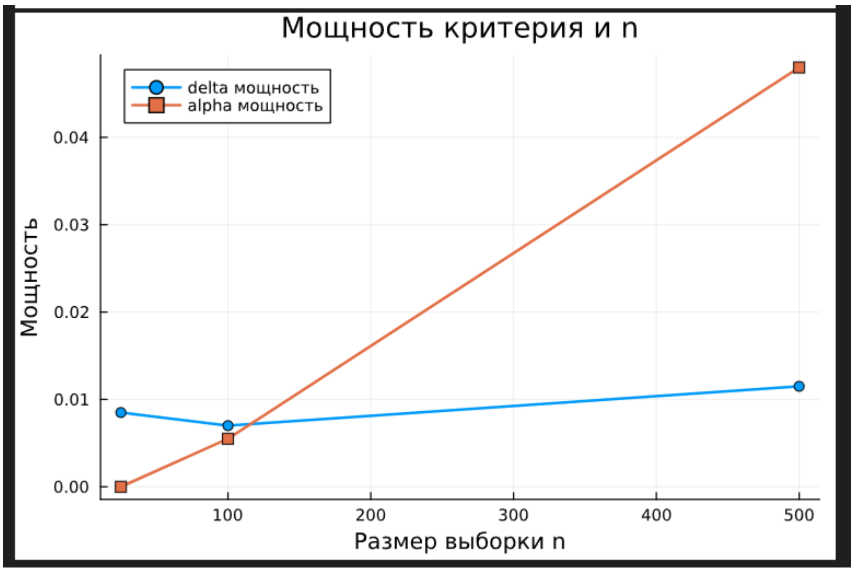
3. Общий вывод

- Ни $\Delta(G)$, ни $\alpha(G)$ не обладают достаточной чувствительностью и мощностью для различения распределений $\mathrm{Normal}(0,1)$ и $\mathrm{SkewNormal}(0,1,\xi=5)$ при использовании k-ближайших соседей ($k = 10$) и уровне значимости $\alpha = 0.05$.
- Однако α показывает потенциальную полезность как характеристика: при большем размере выборки и оптимальном выборе k она может стать основой более мощного критерия.

Ниже представлены результаты исследований:


Значение delta(G) и n
delta(G)
19.0
18.5
18.0
17.5
17.0
100
200
300
400
500
Размер выборки n
delta Normal
delta SkewNormal
delta порог


Значение alpha(G) и n
alpha(G)
60
50
40
30
20
10
100
200
300
400
500
Размер выборки n
alpha Normal
alpha SkewNormal
alpha порог


Мощность критерия и n
Мощность
0.04
0.03
0.02
0.01
0.00
100
200
300
400
500
Размер выборки n
delta мощность
alpha мощность

Анализ результатов классификации (Exponential и LogNormal)

1. Средние значения характеристик δ и χ

- **δ:** средние значения δ практически не различаются между выборками из распределений Exp и LogNormal, а также близки к пороговым значениям, вычисленным по критерию уровня значимости $\alpha = 0.05$.
  - Это свидетельствует о слабой чувствительности признака δ к различию между распределениями.
- **χ:** значения метрики χ для распределения LogNormal стабильно выше, чем для Exp. Однако значения χ для LogNormal по-прежнему ниже соответствующих порогов, и различие между ними недостаточно велико.
  - Это означает, что χ является более чувствительным признаком, чем δ, но не обладает достаточной статистической силой.

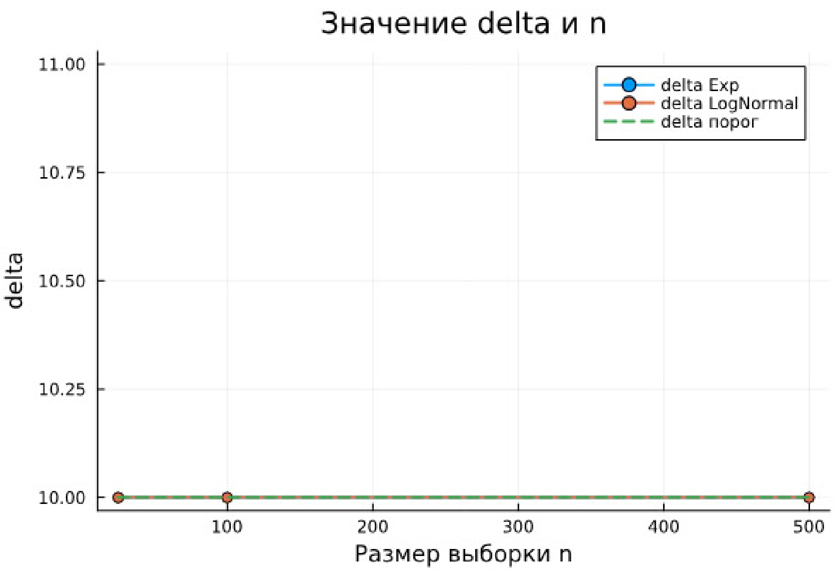
2. Мощность критериев

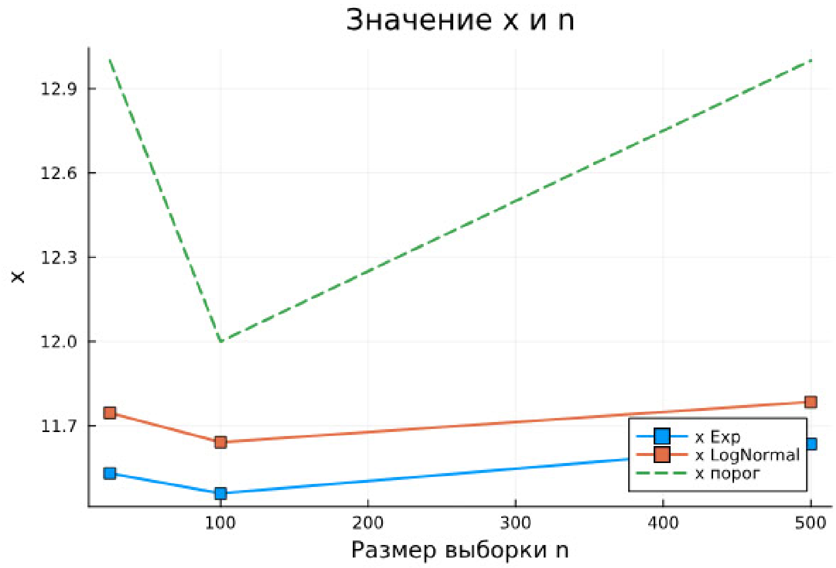
- **Мощность критерия на основе** δ во всех экспериментах оказалась равна 0, что указывает на полное отсутствие способности различать распределения на основе этой метрики.
- **Мощность критерия на основе** χ была положительной лишь при $n = 100$, где достигала ≈ 0.078, но в остальных случаях также близка к нулю.

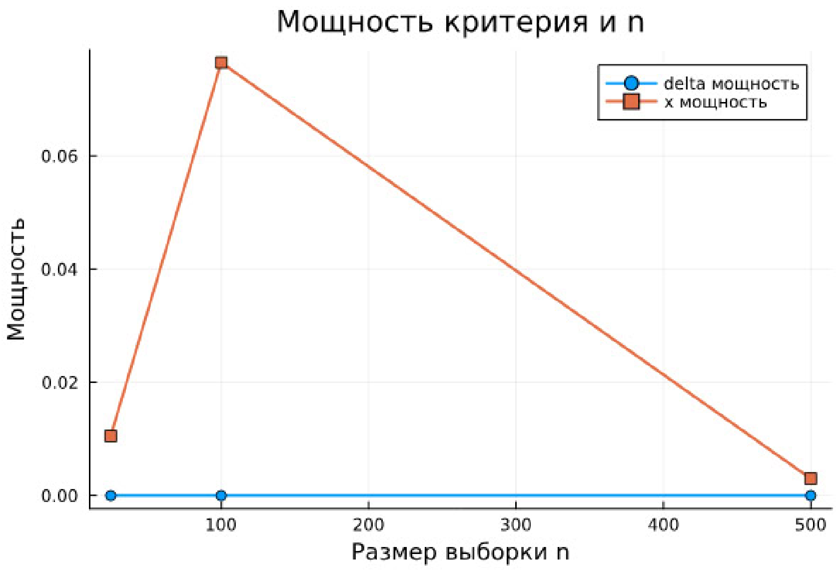
3. Общий вывод

- Ни одна из двух графовых характеристик (δ, χ) не показала достаточную эффективность в задаче различения распределений $\text{Exp}(\lambda_0)$ и $\text{LogNormal}(0, 1)$ при фиксированном числе соседей $k = 10$ и уровне значимости $\alpha = 0.05$.

Ниже представлены результаты исследований:




Значение x и n
12.9
12.6
12.3
12.0
11.7
x
100
200
300
400
500
Размер выборки n
x Exp
x LogNormal
x порог


Мощность критерия и n
0.06
0.04
0.02
0.00
Мощность
100
200
300
400
500
Размер выборки n
delta мощность
x мощность