Отчет по части I

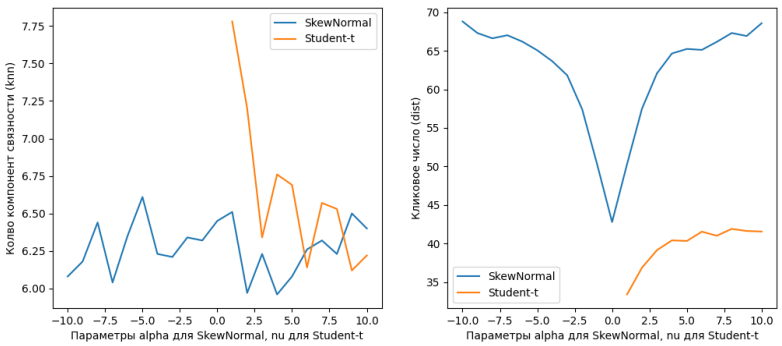
Отчет Аскара

Часть 1. Зависимость от параметров распределений

Значения на графиках — это среднее по $M = 100$ независимым реализациям для каждого набора параметров. Число вершин графа $n = 100$, параметр $k = 5$ для kNN-графа и порог $d = 1$ для DIST-графа.

1. **kNN-граф:** Среднее число компонент связности практически не зависит от параметра α SkewNormal (почти горизонтальная кривая около 6–6.5). Для Student-t с ростом ν число компонент убывает, то есть при «тяжёлых хвостах» (ν — меньше) граф рассоединён сильнее.

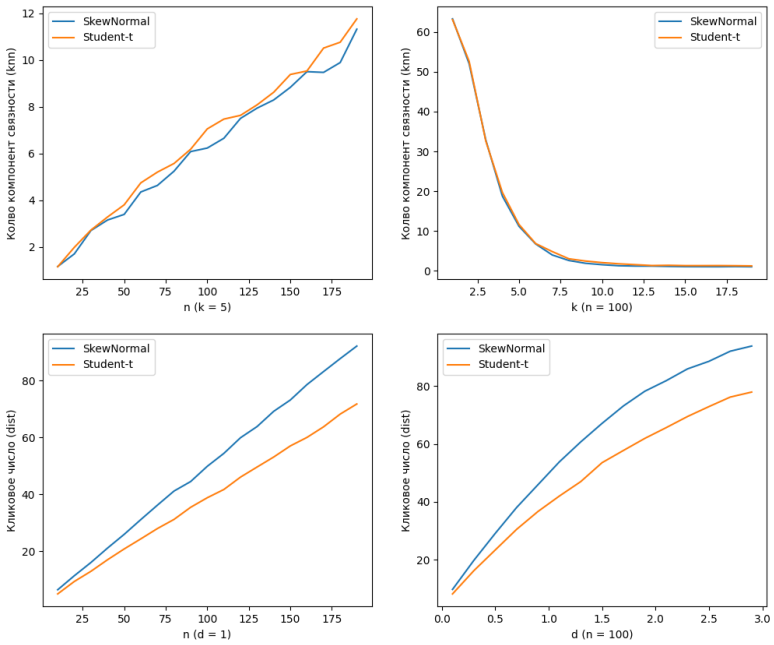
2. **DIST-граф:** Среднее кликовое число минимально при $\alpha = 0$ и симметрично растёт при удалении от нуля (от ~ 40 до ~ 70). Для Student-t кликовое число увеличивается с ν (от ~ 30 при $\nu \approx 1$ до ~ 40–45 при $\nu \approx 10$).



Часть 2. Зависимость от n, k и d

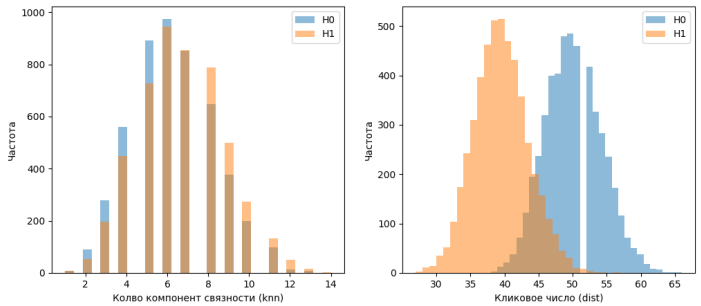
Значения на графиках — это среднее по $M = 100$ независимым реализациям для каждого набора параметров.

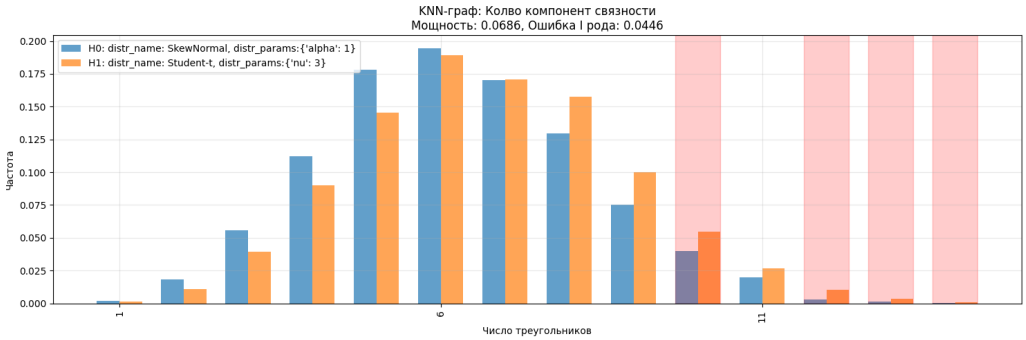
- **kNN-граф:**
  - При увеличении числа вершин n (при $\alpha = \alpha_0$, $\nu = \nu_0$, $k = 5$) среднее число компонент связности возрастает.
  - При увеличении числа соседей k (при $\alpha = \alpha_0$, $\nu = \nu_0$, $n = 100$) число компонент резко убывает.
- **DIST-граф:**
  - При увеличении числа вершин n (при $\alpha = \alpha_0$, $\nu = \nu_0$, $d = 1$) среднее кликовое число растёт, причём скорость роста выше для SkewNormal-графов.
  - При увеличении d (при $\alpha = \alpha_0$, $\nu = \nu_0$, $n = 100$) кликовое число также увеличивается, и для SkewNormal-графов этот рост быстрее. Рост вызван тем, что точки чаще попадают в радиус d.

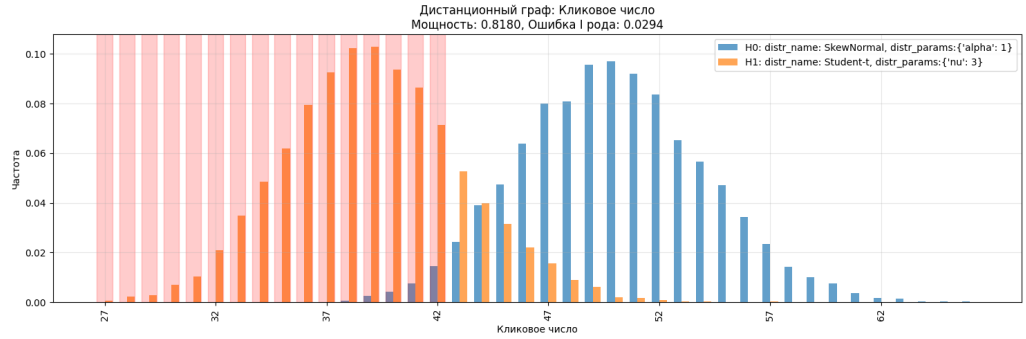


Часть 3. Разделяющая способность статистик

Построено по $M_{\text{large}} = 5000$ реализаций каждого распределения.







- **kNN-граф:** Распределения числа компонент при H_0 и H_1 сильно перекрываются — низкая разделяющая способность, мощность маленькая.
- **DIST-граф:** Распределения кликового числа сдвинуты друг от друга: для SkewNormal значения пик около 50, для Student-t — около 39. Красная зона — область принятия H_1: мощность выше.

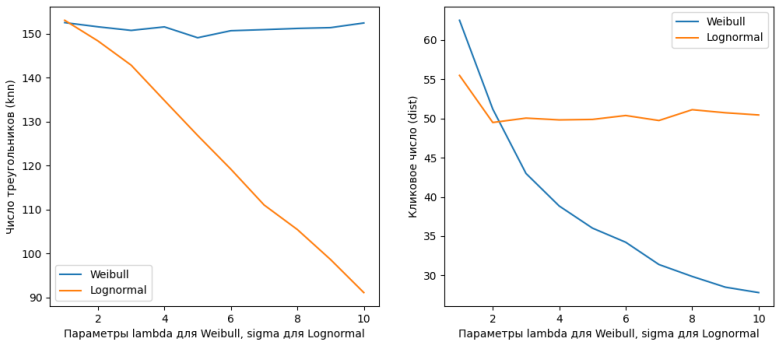
Отчет Ярослава

Часть 1. Влияние параметров распределений

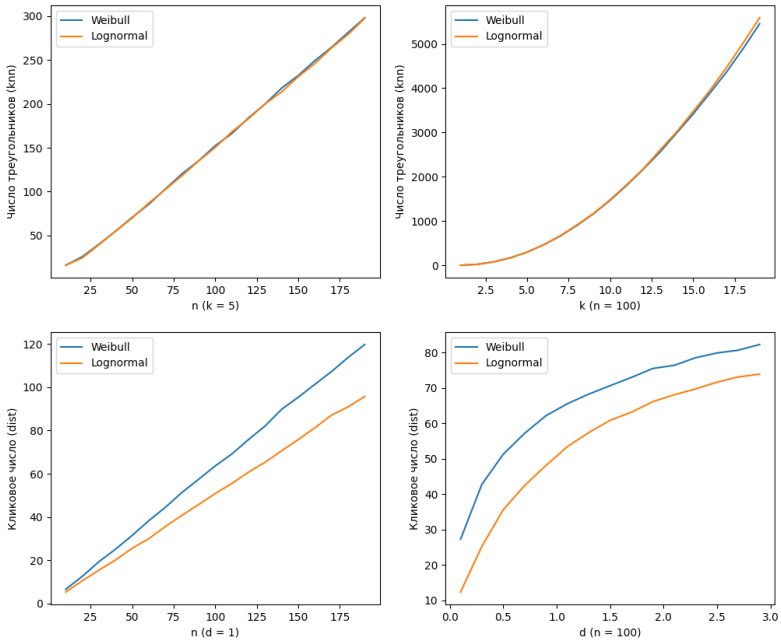
Среднее по $M = 100$ реализациям, $n = 100$, $k = 5$ (kNN) / $d = 1$ (DIST).

1. **kNN-граф:** Число треугольников почти не меняется при изменении λ Weibull (около 150–151); при увеличении дисперсии σ у Lognormal падает с 151 до 91.

2. **DIST-граф:** Кликовое число уменьшилось с 62 до 27 при росте λ (Weibull) и с 55 до 50 при росте σ (Lognormal).



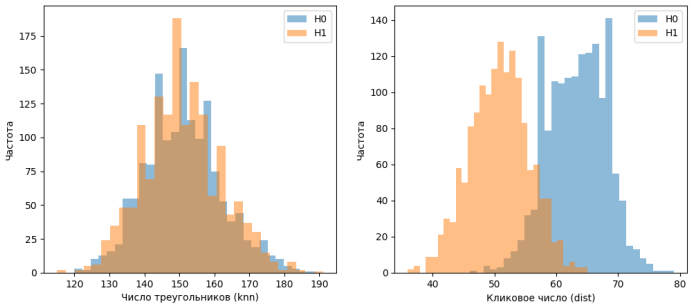
Часть 2. Зависимость от n, k и d

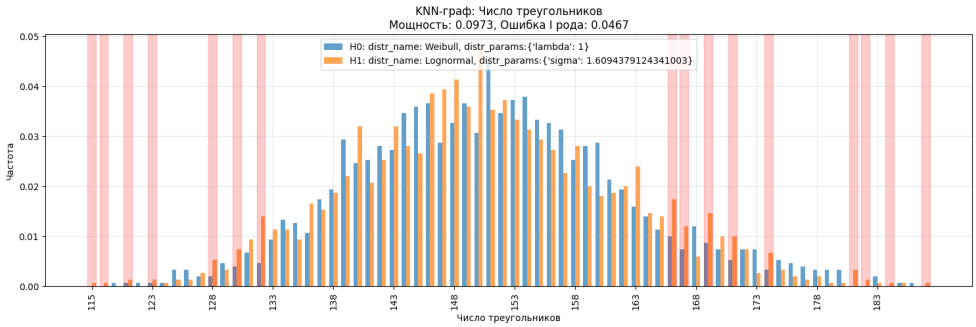


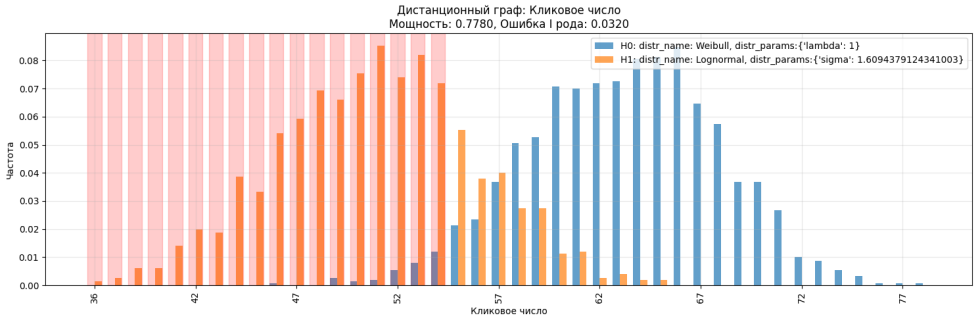
Выводы

1. Для метрики «треугольники» оба распределения ведут себя почти одинаково — выбор распределения практически не влияет на итог.
2. Для «кликового числа» Weibull формирует более плотные графы: прирост относительно Lognormal усиливается с ростом n и d.
3. Чувствительность метрик:
   - «Треугольники» — сильнее реагируют на увеличение k (приблизительно $\propto k^3$), чем на n (приблизительно $\propto n$).
   - «Клики» — линейны по n, но по d быстро достигают плато.

Часть 3. Проверка статистических гипотез







- Мощность теста по треугольникам (kNN) составляет 0.1 при ошибке I рода 0.05.
- Мощность теста по кликовому числу (DIST) — 0.78 при ошибке 0.03.

Отчет по части II

Шаяхметов Аскар

Гипотезы:

- H_0: данные из распределения skewnorm с параметром $\alpha = 1$
- H_1: данные из распределения student_t с параметром $\nu = 3$

Параметры исследования:

- Тип графа: dist-граф с параметром $d = 0.5$
- Размеры выборок: $n = 25, 100, 500$
- Количество выборок на класс: 500

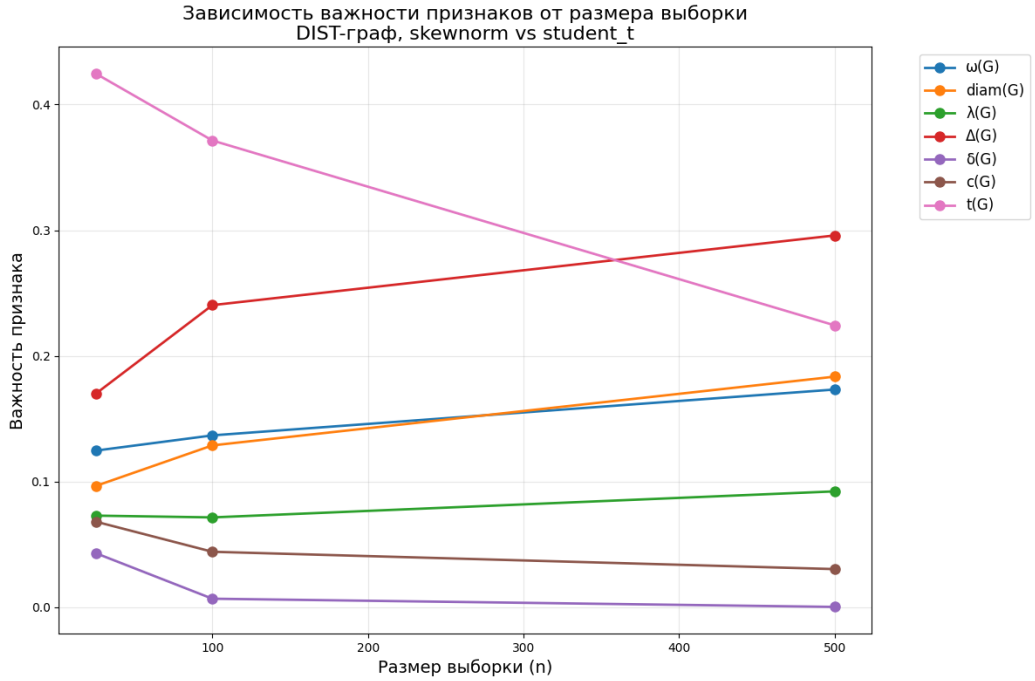
Исследуемые характеристики графов:

- $\Delta(G)$ — максимальная степень вершины
- $\delta(G)$ — минимальная степень вершины
- $c(G)$ — количество компонент связности
- $t(G)$ — количество треугольников
- $\mathrm{diam}(G)$ — диаметр графа
- $\lambda(G)$ — рёберная связность
- $\omega(G)$ — кликовое число

1 Результаты

1.1 Анализ важности характеристик

Анализ важности характеристик с использованием Random Forest показал следующие результаты:

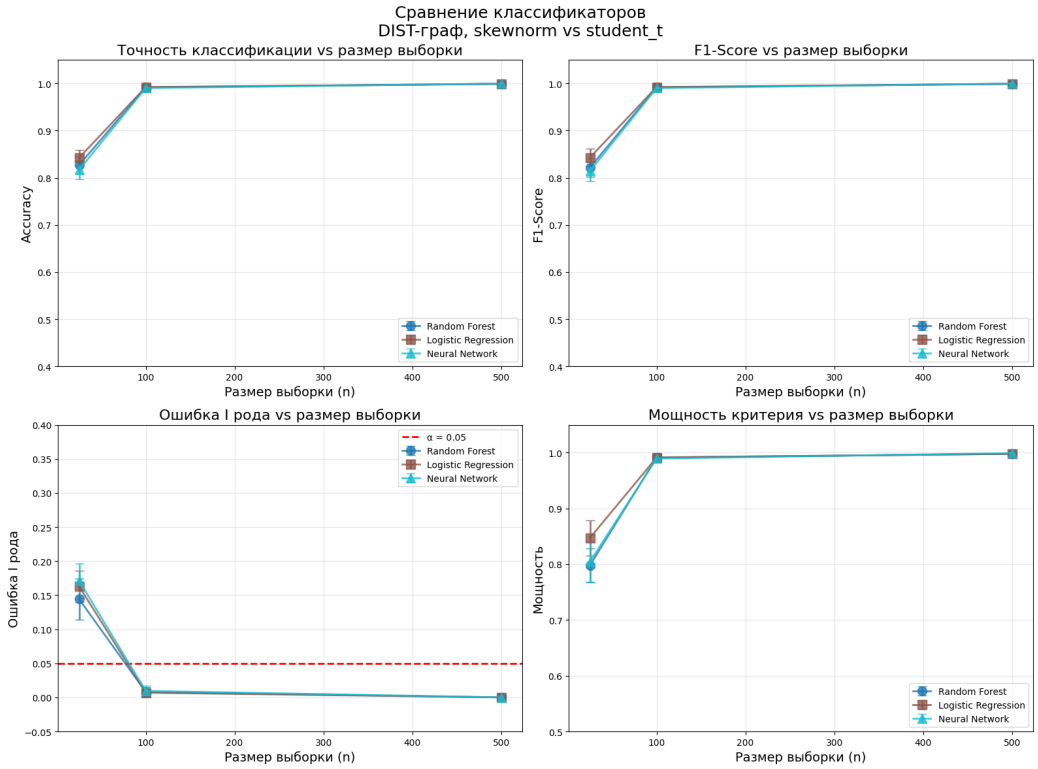


Основные наблюдения:

- Для малых выборок ($n = 25$) наиболее важной характеристикой является количество треугольников $t(G)$ (42.5% важности)
- С ростом размера выборки важность максимальной степени $\Delta(G)$ увеличивается: от 17% при $n = 25$ до 29.6% при $n = 500$
- Минимальная степень $\delta(G)$ практически теряет значение с ростом n

1.2 Сравнение классификаторов

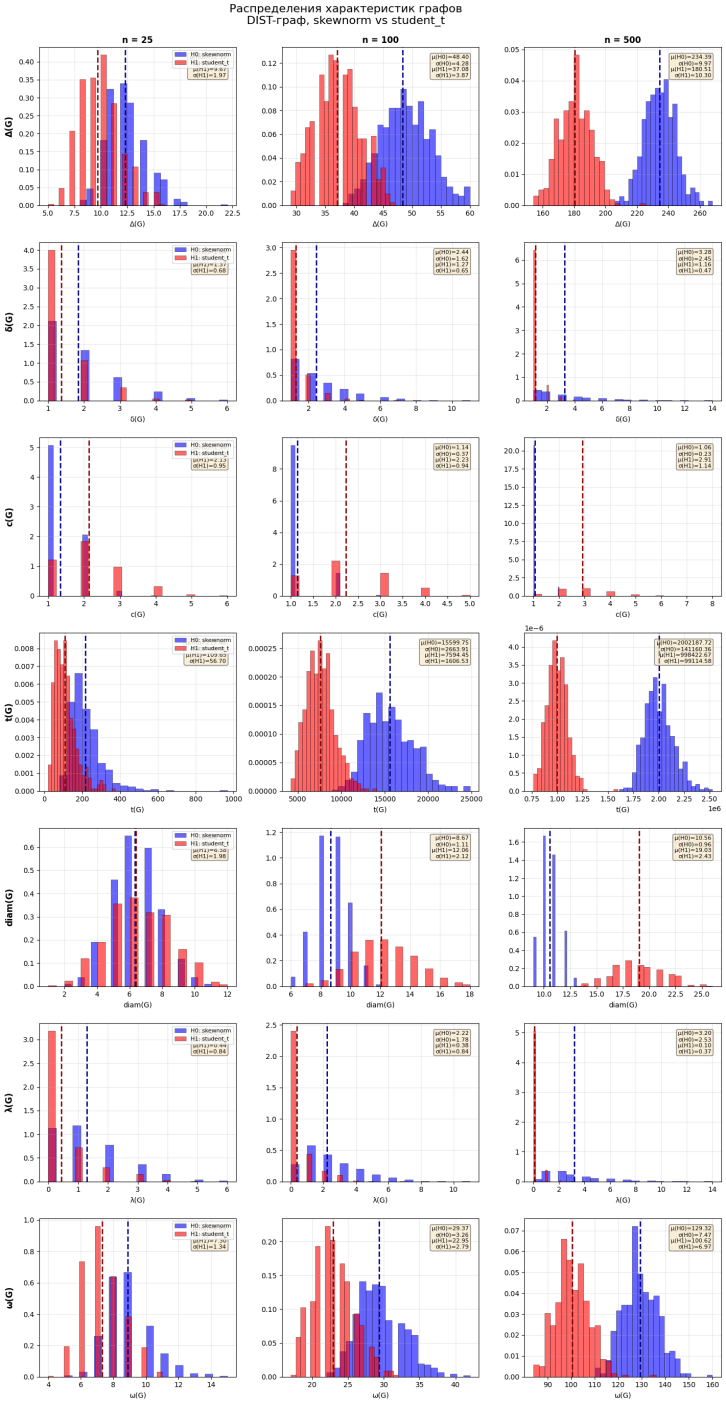
Для оценки качества классификации использовались следующие алгоритмы: Random Forest, Logistic Regression и Neural Network. Результаты представлены на графике:



Основные выводы по классификации:

- Для малых выборок ($n = 25$) все классификаторы показывают умеренное (≈ 0.83) качество с высокой ошибкой первого рода ($\alpha > 0.14$)
- При $n = 100$ качество классификации резко улучшается, ошибка первого рода снижается до уровня ($\alpha \approx 0.01$)
- Для больших выборок ($n = 500$) все классификаторы показывают практически идеальное качество

1.3 Анализ распределений характеристик



Гистограммы распределений характеристик графов показывают четкое разделение между гипотезами H_0 и H_1 для некторых характеристик.

- Максимальной степени $\Delta(G)$ — разделение улучшается при увеличении n
- Количества треугольников $t(G)$ — четкое разделение для $n = 500$
- Диаметра графа $\mathrm{diam}(G)$ — приемлемое разделение
- Кликового числа $\omega(G)$ — для $n = 500$ хорошее разделение

С увеличением размера выборки разделение между распределениями становится более выраженным, что объясняет улучшение качества классификации.

2 Выводы

Анализ результатов показал следующее:

- **Для** $n = 25$: ни один классификатор не удовлетворяет условию $\alpha \leq 0.05$
- **Для** $n = 100$: лучший классификатор — Random Forest с ошибкой первого рода $\alpha = 0.008$ и мощностью 0.991
- **Для** $n = 500$: лучший классификатор — Neural Network (два скрытых слоя размерами 50 и 30) с ошибкой первого рода $\alpha = 0.000$ и мощностью 0.999

Отчет по части II

Богданов Ярослав

Гипотезы:

- H_0: данные из распределения weibull с параметром $\lambda = 1$
- H_1: данные из распределения lognormal с параметром $\sigma = log(5)$

Параметры исследования:

- Тип графа: dist-граф с параметром $d = 0.5$
- Размеры выборок: $n = 25, 100, 500$
- Количество выборок на класс: 100

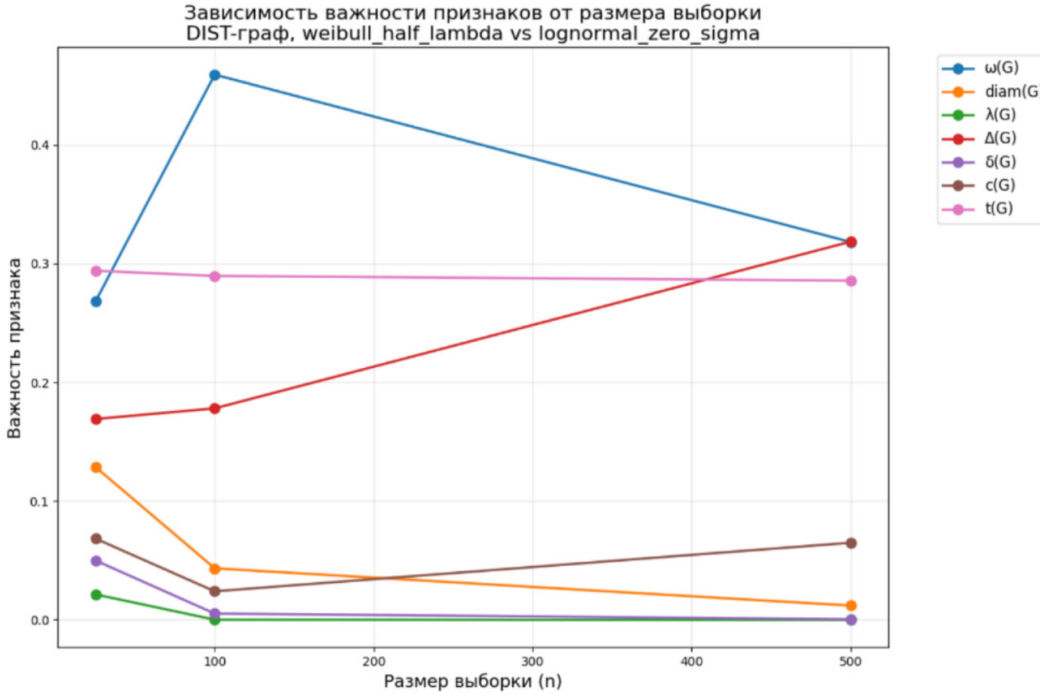
Исследуемые характеристики графов:

- $\Delta(G)$ — максимальная степень вершины
- $\delta(G)$ — минимальная степень вершины
- $c(G)$ — количество компонент связности
- $t(G)$ — количество треугольников
- $\mathrm{diam}(G)$ — диаметр графа
- $\lambda(G)$ — рёберная связность
- $\omega(G)$ — кликовое число

1 Результаты

1.1 Анализ важности характеристик

Анализ важности характеристик с использованием Random Forest показал следующие результаты:



Основные наблюдения:

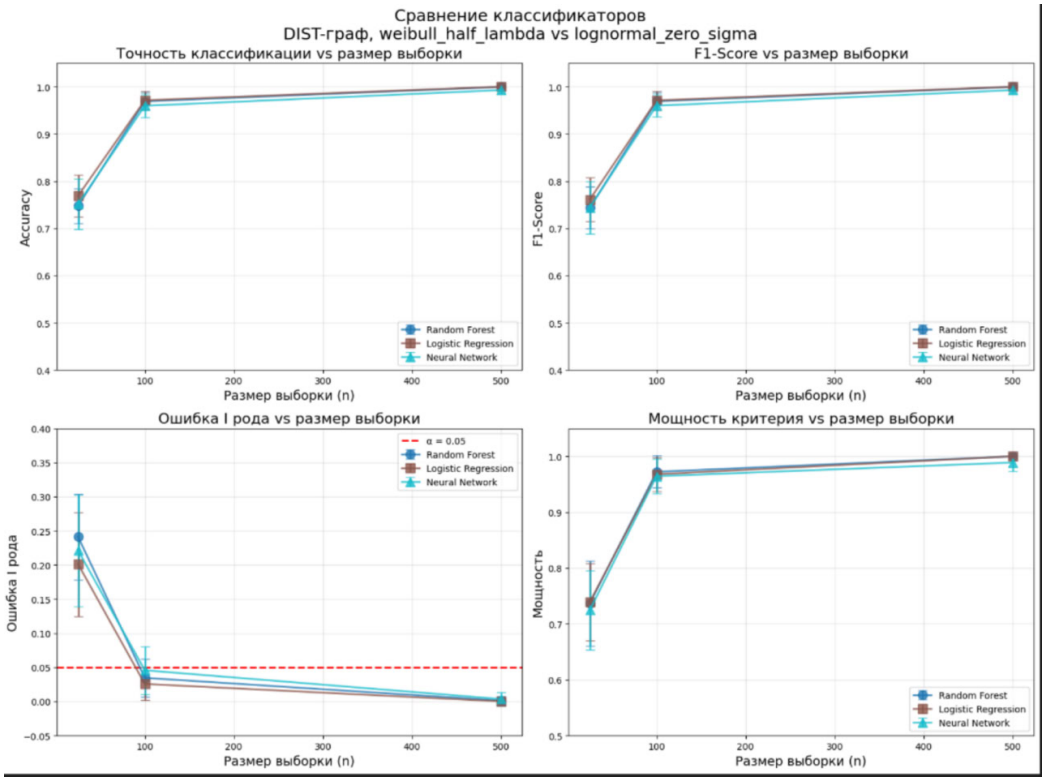
- При $n = 25$ наибольший вклад дают число треугольников $t(G) \approx 29\%$ и кликовое число $\omega(G) \approx 27\%$, за ними следуют максимальная степень $\Delta(G) \approx 17\%$ и диаметр $\mathrm{diam}(G) \approx 13\%$.
- При $n = 100$ доминирует $\omega(G) \approx 46\%$, тогда как важность диаметра $\mathrm{diam}(G)$ падает до $\approx 4\%$, а $\delta(G)$ и $\lambda(G)$ практически сходят на нет ($< 1\%$).
- При увеличении до $n = 500$ максимальная степень $\Delta(G)$ растёт до $\approx 32\%$ и выравнивается с $\omega(G) \approx 32\%$, число треугольников $t(G)$ остаётся стабильным ($\approx 29\%$).
- Диаметр $\mathrm{diam}(G)$ продолжает снижаться (до $\approx 1.3\%$), а минимальная степень $\delta(G)$ и реберная связность $\lambda(G)$ практически теряют

значение.

- Число компонент связности $c(G)$ демонстрирует U-образную динамику: $\approx 7\% \to 2\% \to 6.5\%$ при росте n.

1.2 Сравнение классификаторов

Для оценки качества классификации использовались следующие алгоритмы: Random Forest, Logistic Regression и Neural Network. Результаты представлены на графике:

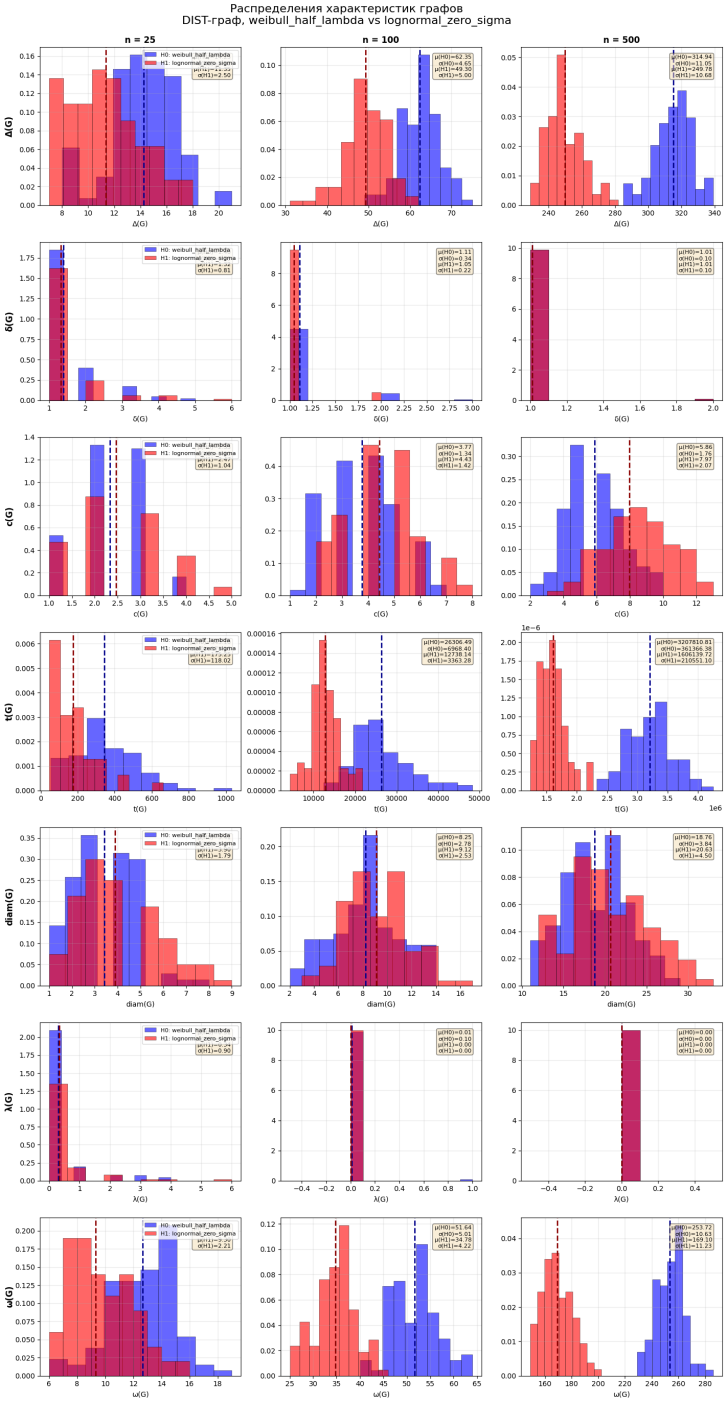


Основные выводы по классификаторам:

- Для малых выборок ($n = 25$) все модели демонстрируют среднюю точность (≈ 0.75–0.77) и аналогичный F1-Score, при этом ошибка I рода значительно превышает уровень значимости ($\alpha = 0.05$), достигая 20%–24%, а мощность критерия находится на уровне 0.72–0.74.
- При увеличении выборки до $n = 100$ точность и F1-Score резко возрастают до 0.95–0.97, ошибка I рода падает ниже 5% (до $\approx$ 2%–3%), а мощность критерия достигает 0.94–0.97.

- Для больших выборок ($n = 500$) все три алгоритма достигают практически идеальных показателей: точность и F1-Score близки к 1.00, ошибка I рода стремится к нулю, мощность критерия приближается к единице.
- Различия между алгоритмами минимальны: Logistic Regression чуть опережает Random Forest на средних выборках, Neural Network демонстрирует чуть больший разброс оценок.

1.3 Анализ распределений характеристик



Гистограммы распределений характеристик графов показывают, как изменяется разделимость между гипотезами H_0 и H_1 с ростом размера выборки:

- $\Delta(G)$ (максимальная степень) — при $n = 25$ видна лишь слабая тенденция к сдвигу, при $n = 100$ распределения уже хорошо разделяются, а при $n = 500$ их разделение становится почти полным.
- $t(G)$ (число треугольников) — умеренное разделение для $n = 25$ и $n = 100$, для $n = 500$ гистограммы практически не перекрываются.
- $\mathrm{diam}(G)$ (диаметр) — заметное, но неполное разделение; с ростом n средние значения расходятся, но хвосты всё ещё пересекаются.
- $\omega(G)$ (кликовое число) — при $n = 100$ уже явное разделение, при $n = 500$ гистограммы хорошо разделены.
- $c(G)$ (число компонент связности) — небольшое смещение средних при $n \geq 100$, сильнее выраженное при $n = 500$, но перекрытие сохраняется.
- $\delta(G)$ (минимальная степень) и $\lambda(G)$ (рёберная связность) — при любых n распределения почти совпадают, разделения не наблюдается.

С увеличением размера выборки разделение между распределениями становится более выраженным, что объясняет улучшение качества классификации на больших n.

2 Выводы

Анализ итоговых показателей классификации даёт следующие выводы:

- **Для** $n = 25$: ни один классификатор не удовлетворяет условию $\alpha \leq 0.05$.
- **Для** $n = 100$: лучший классификатор — Random Forest с ошибкой I рода $\alpha = 0.0344$ и мощностью 0.9722.
- **Для** $n = 500$: лучший классификатор — Random Forest с ошибкой I рода $\alpha = 0.0011$ и мощностью 1.0000.