Научная статья
УДК 004.056.53, 519.213, 519.224.22
DOI: 10.17072/1993-0550-2026-2-112-129
https://elibrary.ru/iavrsm

Статистический анализ данных поведения сетевого трафика при DDoS-атаках

Алексей Леонидович Шкерин[1], Дмитрий Иванович Понаётов[2], Фёдор Владимирович Пискун[3]

[1, 2, 3] Российский технологический университет − МИРЭА, Москва, Россия
[1]shkerin@mirea.ru
[2]dmiponayotov@gmail.com
[3]piskun.fv@gmail.com

Аннотация. Из-за прошедшего этапа информатизации общества в последнее время участились случаи атак злоумышленников, которые воруют данные и перегружают сеть. Цель работы состоит в том, чтобы исследовать статистические особенности *DDoS*-трафика, оценить устойчивость точечных оценок и проверить соответствие данных известным распределениям. Построение гистограммы и эмпирической функции распределения; разбиение выборки на пять случайных подвыборок ($\sim 20\%$ от объема); расчет точечных оценок (медиана, среднее, дисперсия, СКО); построение 95%-ных доверительных интервалов для математического ожидания и дисперсии; вычисление усеченного среднего; проверка гипотезы о симметрии с использованием коэффициента асимметрии; применение критерия Колмогорова–Смирнова и графического метода анаморфоз с оценкой соответствия по коэффициенту детерминации R^2для трех распределений: логнормального, Максвелла–Больцмана и Лапласа. В результате анализа была выявлена ярко выраженная правосторонняя асимметрия и тяжелый "хвост". Выборочное среднее оказалось заметно смещено вправо по сравнению с медианой и усеченным средним, что типично для подобной структуры распределения. Все три рассмотренные параметрические модели (логнормальное, Максвелла–Больцмана и Лапласа) были отвергнуты на уровне значимости $\alpha = 0{,}05$, несмотря на то, что распределение Лапласа формально продемонстрировало наивысшее качество соответствия. В ходе исследования подтверждено, что *DDoS*-трафик обладает сложной структурой, которая не описывается классическими параметрическими моделями. Полученные результаты обосновывают необходимость применения сложных и непараметрических подходов при предобработке данных и разработке систем обнаружения вредоносного трафика.

Ключевые слова: *статистический анализ; DDoS; сетевой трафик; критерий Колмогорова–Смирнова; распределение Максвелла–Больцмана; распределение Лапласа; анаморфоза.*

Для цитирования: *Шкерин А. Л., Понаётов Д. И., Пискун Ф.В.* Статистический анализ данных поведения сетевого трафика при DDoS-атаках // Вестник Пермского университета. Математика. Механика. Информатика. 2026. № 2(73). С. 112−129. DOI: 10.17072/1993-0550-2026-2-112-129. https://elibrary.ru/iavrsm.

Благодарности: работа выполнена при поддержке научного сайта с данными IEEE Dataport, The Bot-IoT dataset; авторы выражают благодарность Нуру Мустафе за предоставление данных о DDoS-атаках.

Статья поступила в редакцию 09.01.2026; одобрена после рецензирования 20.05.2026; принята к публикации 20.10.2026.

Research article

Statistical Data Analysis of Network Traffic Behavior Under DDoS-Attack

Alexey L. Shkerin[1], Dmitriy I. Ponayotov[2], Fedor V. Piskun[3]

[1, 2, 3] Russian Technological University − MIREA, Moscow, Russia

[1]shkerin@mirea.ru

[2]dmiponayotov@gmail.com

[3]piskun.fv@gmail.com

Abstract. Due to the recent advancement of informatization in society, there has been a recent increase in the number of attacks by malicious users who steal data and overload the network. The aim of this work is to investigate the statistical characteristics of DDoS traffic, evaluate the robustness of point estimates, and check the data's conformity with known distributions. Constructing a histogram and empirical distribution function; splitting the sample into five random subsamples ($\sim 20\%$ of the volume); calculating point estimates (median, mean, variance, standard deviation); constructing 95% confidence intervals for the expectation and variance; calculating the trimmed mean; testing the symmetry hypothesis using the asymmetry coefficient; applying the Kolmogorov–Smirnov test and the graphical goodness of fit method with conformity assessment using the R^2 determination coefficient for three distributions: lognormal, Maxwell–Boltzmann, and Laplace. The analysis revealed pronounced right-sided asymmetry and a heavy tail. The sample mean was significantly shifted to the right compared to the median and trimmed mean, which is typical for this type of distribution structure. Although the true mean fell within the 95% confidence intervals of all subsamples, the true variance did not fall within any of them, demonstrating the instability of this estimate. All three considered parametric models (lognormal, Maxwell-Boltzmann, and Laplace) were rejected at the significance level of $\alpha = 0.05$, despite the fact that the Laplace distribution formally demonstrated the highest quality of fit. The study confirmed that DDoS traffic has a complex structure that is not described by classical parametric models. The obtained results substantiate the need to apply complex and nonparametric approaches in data preprocessing and the development of malicious traffic detection systems.

Keywords: *statistical analysis; DDoS; network traffic; Kolmogorov-Smirnov test; Maxwell–Boltzmann distribution; Laplace distribution; goodness of fit.*

For citation: Shkerin, A. L., Ponayotov, D. I. and Piskun, F. V. (2026), "Statistical Data Analysis of Network Traffic Behavior Under DDoS-Attack", *Bulletin of Perm University. Mathematics. Mechanics. Computer Science*, no 2(73), pp. 112−129, DOI: 10.17072/1993-0550-2026-2-112-129, https://elibrary.ru/iavrsm.

Acknowledgments: the work was supported by the dataset files from IEEE Dataport, The Bot-IoT dataset; the authors are grateful to Nour Moustafa for providing the data of DDoS-attacks.

The article was submitted 09.01.2026; approved after reviewing 20.05.2026; accepted for publication 20.10.2026.

Введение

Современные сетевые сервисы все чаще сталкиваются с различными мощными видами *DDoS*-атак, которые ставят под угрозу стабильность их работы. Для эффективного противодействия этой проблеме важно понимать природу вредоносного трафика. Одной из выделяющихся характеристик таких атак является длительность $HTTP$-соединений – она может стать основой для адаптивных систем обнаружения аномалий [1–2].

В работе выявляются и исследуются характерные особенности *DDoS*-трафика, включая проверку его соответствия трем параметрическим моделям (логнормальная, Максвелла–Больцмана, Лапласа), при помощи визуально сложных и непараметрических методах.

Постановка задачи

Пусть дана выборка $X = \{x_1, x_2, \cdots, x_n\}$, где $x_i > 0$ – длительность i-го *HTTP*-соединения *DDoS*-атаки в секундах. Поставим следующие задачи:

– визуально оценить структуру выборки;

– оценить устойчивость статистических характеристик;

– оценить соответствие распределения выборки классическим.

Методы исследования

Для решения поставленных задач используем комбинированный подход:

1. Выполним визуальный анализ при помощи гистограммы и эмпирической функции распределения.
2. Оценим устойчивость статистик – разобьем выборку на 5 случайных, не пересекающихся, подвыборок и вычислим основные статистики по ним: медиану, выборочные среднее и дисперсию, среднеквадратичное отклонение (СКО). После построим доверительные интервалы для среднего и дисперсии с надежностью 95%. Также вычислим усеченное среднее и коэффициент асимметрии выборки для численного подтверждения результатов визуального анализа.
3. Проверим соответствие распределения выборки трем классическим моделям (логнормальной, Максвелла–Больцмана и Лапласа) при помощи критерия Колмогорова–Смирнова и графического метода анаморфоз с оценкой по коэффициенту детерминации R^2.

1. Выявление классических параметров и точечных оценок для анализируемых данных

Исследуемые наборы данных относится к анализу сетевого трафика при *DDoS*-атаках (*Distributed Denial of Service*) [3–4]. Выбор датасетов *IEEE Dataport* (*The Bot-IoT dataset*) и *CAIDA UCSD "DDoS Attack 2007"* обусловлен их репрезентативностью и широким признанием в научном сообществе. Набор данных *Bot-IoT* содержит информацию о трафике *IoT*-устройств, включающую как легитимную активность, так вредоносную. Датасет *CAIDA* представляет собой реальные данные перехвата трафика во время крупномасштабной *DDoS*-атаки. Использование двух независимых источников данных позволит сравнить их специфику и сделать универсальные выводы для аналогичных исследований.

Первые записи датасета от *IEEE-dataport*, с интересующими нас признаками, представлены в табл. 1.

Таблица 1. *Обзор строк исследуемого датасета IEEE-dataport*

Протокол	Категория	Длительность
TCP	DDoS	12,96526
UDP	DDoS	14,4117
TCP	Normal	1097,528
UDP	DDoS	13,81165
UDP	DDoS	12,1611
TCP	DDoS	14,71928
TCP	DDoS	0,258019
TCP	DDoS	0,298577
TCP	DDoS	34,36997
TCP	DDoS	0,282785
TCP	Normal	1478,756
TCP	DDoS	113,5577

В табл. 2 представлены записи датасета от *CAIDA*, с интересующими нас признаками.

Таблица 2. *Обзор строк исследуемого датасета CAIDA*

Протокол	Информация	Длительность
TCP	50450 > 35451 [SYN] Seq=0 Win=64512 Len=0 MS...	138.717388
TCP	50452 > 35451 [SYN] Seq=0 Win=64512 Len=0 MS...	138.716784
TCP	50454 > 35451 [SYN] Seq=0 Win=64512 Len=0 MS...	210.258711
TCP	50456 > 35451 [SYN] Seq=0 Win=64512 Len=0 MS...	213.173783
TCP	50458 > 35451 [SYN] Seq=0 Win=64512 Len=0 MS...	213.173822
TCP	50460 > 35451 [SYN] Seq=0 Win=64512 Len=0 MS...	213.174086

Колонка "Информация" содержит следующие данные о сетевом пакете:

- исходный и целевой порты;
- TCP-флаг (к примеру, [SYN]);
- номер последовательности пакета (Seq);
- размер окна приема – объем байт данных, который может принять получатель (Win);
- длина полезной нагрузки пакета в байтах (Len);
- прочие TCP-опции.

Для анализа выберем признак длительности, который демонстрирует время жизни *HTTP*-соединения в секундах. Данный признак выбран поскольку *DDoS*-атаки на прикладном уровне часто характеризуются аномальными паттернами длительности соединений − либо массовыми короткими запросами, либо удержанием соединений в течение продолжительного времени.

Протокол *HTTP* выбран, поскольку, согласно статистике *Cloudflare*, на долю *HTTP*-запросов приходится около 75% всех зафиксированных *DDoS*-атак прикладного уровня [5].

Доля *HTTP*-трафика в структуре *DDoS*-атак представлена на рис. 1.

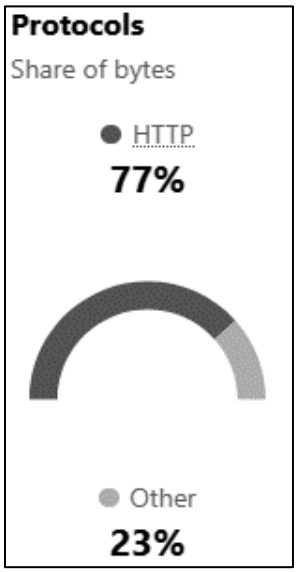

Рис. 1. *Доля HTTP-трафика*

Несмотря на то, что в таблице указаны протоколы *TCP* и *UDP*, исследуемый *HTTP*-трафик передается исключительно поверх *TCP* – это обусловлено самой архитектурой протокола *HTTP*, который использует *TCP* в качестве транспортного уровня. UDP же используется в основном для амплификационных атак, поэтому рассматриваться не будет. Перейдем непосредственно к анализу признака длительности соединения по первому датасету.

Тип признака – числовой, непрерывный, содержит только положительные значения. Объем выборки составляет 367307 наблюдений. Диапазон – от $2 \cdot 10^{-5}$ до 113.56 секунд (~2 минуты).

Построим гистограмму вариационного ряда по формуле Стёрджесса [6]:

$$k = 1 + \log_2 n = 1 + 3.222 \cdot \log_{10} 367307 \approx 19\,. \tag{1}$$

Разделим данные на 19 интервалов с шагом h и рассчитаем их границы по формулам:

$$h = \frac{x_{max} - x_{min}}{k}, \tag{2}$$

$$Z_i = x_{min} + i \cdot h, i \in [0, k], \tag{3}$$

$$x_i = \frac{Z_i + Z_{i+1}}{2}, i \in [1, k], \tag{4}$$

где n – объем выборки;
Z_i – границы интервалов;
$x_{max} = 113.56$ – максимальное значение в выборке;
$x_{min} = 2 \cdot 10^{-5}$ – минимальное значение в выборке.

После разделения на интервалы построим гистограмму частот. Гистограмма представлена на рис. 2.

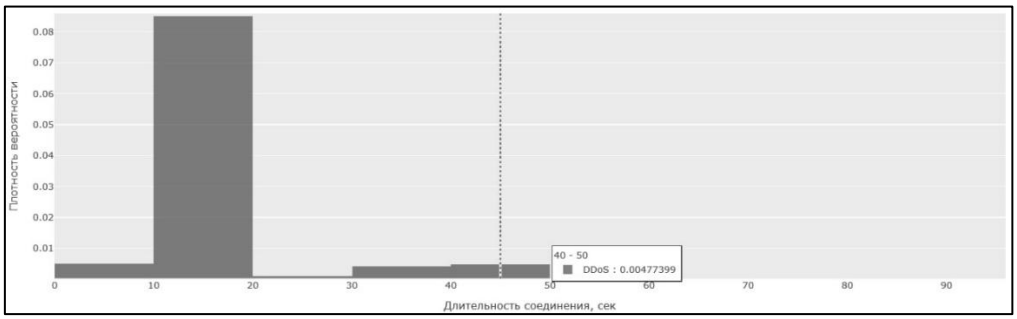

Рис. 2. *Гистограмма длительности соединений DDoS трафика*

Таким образом, гистограмма показывает сильную правостороннюю асимметрию.

Для демонстрации различия с легитимным трафиком, построим его гистограмму длительности соединений. Гистограмма продемонстрирована на рис. 3.

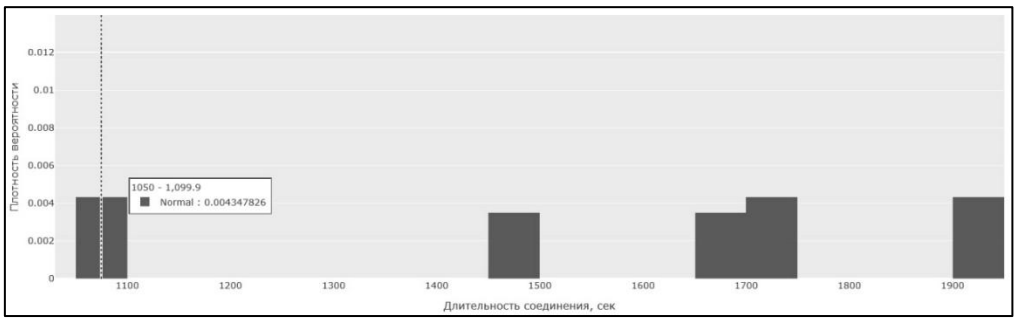

Рис. 3. *Гистограмма длительности соединений легитимного трафика*

Как видно из гистограммы, легитимному трафику исследуемого датасета характерны более длительные соединения до ~ 32 минут.

Перейдем к вычислению эмпирической функции (ЭФР) распределения по исследуемому вариационному ряду. Вычислим функцию по формуле:

$$F_n(x) = \frac{1}{n}\sum_{x_i<x}^{n} n_i = \frac{n_x}{n}, \tag{5}$$

где $n_x = \sum_{x_i<x}^{n} n_i$ – число вариант, значения которых меньше x [7].

График ЭФР на отрезке от 0 до 50 секунд продемонстрирован на рис. 4.

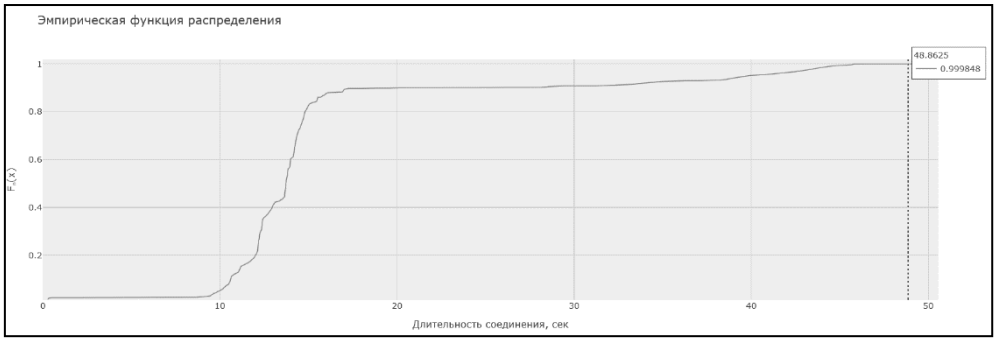

Рис. 4. *ЭФР для признака длительности до 50 сек*

Такой вид ЭФР подтверждает наличие сильной правосторонней асимметрии и тяжелого "хвоста", что согласуется с результатами визуального анализа гистограммы. Таким образом, выявлены асимметрия и тяжелый "хвост" в распределении.

Перейдем к анализу второго датасета от *CAIDA*. Тип исследуемого признака абсолютно такой же. Объем данных составил 1177442 наблюдений. Диапазон – от $1{,}9 \cdot 10^{-6}$ до 299.839 секунд (~5 минут). Как видно, данный датасет существенно отличается от предыдущего по масштабу и структуре. Построим гистограмму для визуальной оценки распределения данных. Гистограмма частот представлена на рис. 5.

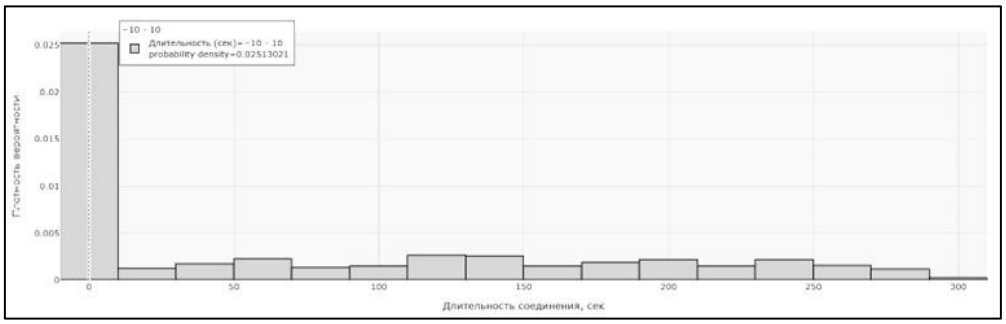

Рис. 5. *Гистограмма длительности соединений*

Также построим график ЭФР. Он продемонстрирован на рис. 6.

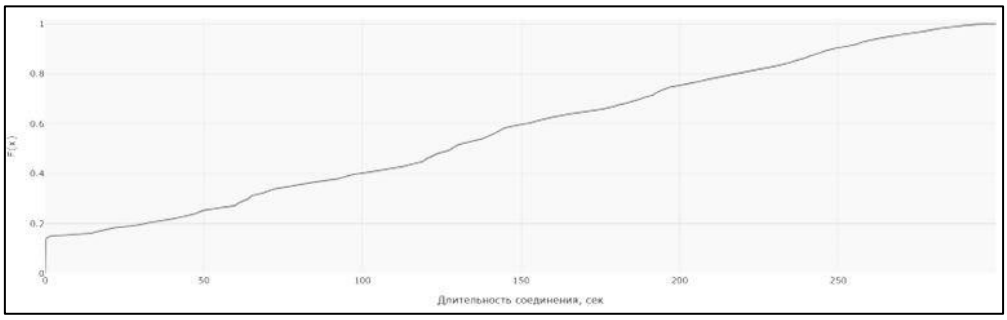

Рис. 6. *ЭФР для признака длительности*

Несмотря на структурные различия датасетов, оба демонстрируют ярко выраженную правостороннюю асимметрию и тяжелый "хвост", т.е. основная масса значений сосредоточена в диапазоне от 0 до 10 секунд, при этом редкие, но более длительные соединения формируют длинный правый "хвост" распределения. Перейдем к оценке устойчивости статистических характеристик исследуемого признака.

Разобьем каждую исследуемую совокупность по признаку на пять непересекающихся случайных выборок. Каждая из них составит ~20% от общего объема наблюдений. Для вычисления медианы используем следующую формулу:

$$Med = \begin{cases} x_{\frac{n+1}{2}}, \text{если } n - \text{нечетное} \\ \dfrac{\left[x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right]}{2}, \text{если } n - \text{четное} \end{cases}, \tag{6}$$

для математического ожидания:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \tag{7}$$

где x_i – значение i-го элемента выборки;
n – объем выборки.

Для дисперсии будет использоваться следующая формула:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\,, \tag{8}$$

Для СКО:

$$s = \sqrt{\sigma^2}. \tag{9}$$

Поскольку нам известна дисперсия (σ), то для расчета доверительного интервала математического ожидания [6] будет использоваться формула:

$$\mu = \overline{X} \pm t_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \tag{10}$$

где $t_{1-\alpha/2}$ – квантиль стандартного нормального распределения соответствующего уровня;
α – уровень значимости ($\alpha = 1 - 0.95 = 0.05$) [8].

Для расчета доверительного интервала дисперсии при тех же условиях будет использоваться следующая формула:

$$(\sigma^2)^{\pm} = \frac{n \cdot \sigma^2}{t_{\alpha}^{\mp}}, \tag{11}$$

где t_{α}^{+} – квантиль распределения хи-квадрат с n степенями свободы уровня $\frac{1-\alpha}{2}$;
t_{α}^{-} – квантиль уровня $\alpha/2$ того же распределения;
σ^2 – выборочная дисперсия, вычисляемая.

Начнем с первого датасета. Истинные среднее и дисперсия генеральной совокупности составили 15.528 и 149.845 соответственно. Результаты вычислений по выборкам представлены в табл. 3.

Таблица 3. *Точечные оценки и доверительные интервалы по выборкам датасета IEEE*

Выборка	I	II	III	IV	V
$\boldsymbol{Med}$	13.712	13.713	13.713	13.713	13.712
$\boldsymbol{\overline{x}}$	15.521	15.540	15.533	15.540	15.507
$\boldsymbol{\sigma^2}$	155.350	147.357	156.161	166.252	124.106
$\boldsymbol{s}$	12.464	12.139	12.496	12.894	11.140
Верхняя граница ДИ для среднего	15.562	15.579	15.574	15.581	15.543
Верхняя граница ДИ для дисперсии	156.061	148.032	156.877	167.014	124.675
Нижняя граница ДИ для среднего	15.481	15.501	15.493	15.498	15.471

Выборка	I	II	III	IV	V
Нижняя граница ДИ для дисперсии	154.643	146.686	155.451	165.496	123.542
Попадание истинного среднего в ДИ	Да				
Попадание истинной дисперсии в ДИ	Нет				

Проанализируем полученные результаты. Медиана во всех выборках лежит в узком диапазоне $13.712 - 13.713$ секунд, а выборочные средние − в пределах $15.507 - 15.540$ секунд. Истинное среднее по всей совокупности (15.528 сек) попало в 95%-ные доверительные интервалы во всех пяти выборках, что подтверждает их репрезентативность.

Оценим результаты по дисперсии. Истинное значение (149.845 сек^2) не вошло ни в один доверительный интервал. Оценки дисперсии сильно варьируются − от 124.106 до 166.252, что связано с высокой чувствительностью этой величины к выбросам. Таким образом, среднее оценивается надежно, а дисперсия − нет.

Теперь перейдем к вычислению усеченного среднего − оценке, устойчивой к выбросам. Для этого из выборки отбросим по 10% самых коротких и самых длинных значений (слева и справа распределения), а затем посчитаем среднее по оставшейся части. Усеченное среднее будем вычислять по следующей формуле:

$$\overline{X_\alpha} = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} x_i\,, \tag{12}$$

где $k = \alpha \cdot n, \alpha$ − доля отсечения [9].

Усеченное среднее (с отсечением по 10% с каждого конца) составило 13.485 секунд, что заметно меньше обычного среднего ($\sim$15.5 секунд) и меньше медианы ($\sim$13.7 секунд).

Перейдем к оценке второго набора данных. Истинные среднее и дисперсия составили 126.912 и 7832.156 соответственно. Результаты вычислений по выборкам представлены в табл. 4.

Таблица 4. *Точечные оценки и доверительные интервалы по выборкам датасета CAIDA*

Выборка	I	II	III	IV	V
$\boldsymbol{Med}$	128.0	128.180	128.188	127.938	128.213
$\boldsymbol{\overline{x}}$	126.794	127.167	127.001	126.712	126.886
$\boldsymbol{\sigma^2}$	7835.757	7840.121	7803.541	7843.536	7837.828
$\boldsymbol{s}$	88.520	88.544	88.338	88.564	88.532
Верхняя граница ДИ для среднего	126.436	126.809	126.644	126.355	126.529
Верхняя граница ДИ для дисперсии	7791.192	7795.531	7759.159	7798.927	7793.252
Нижняя граница ДИ для среднего	127.152	127.524	127.358	127.070	127.244

Выборка	I	II	III	IV	V
Нижняя граница ДИ для дисперсии	7880.708	7885.096	7848.306	7888.531	7882.790
Попадание истинного среднего в ДИ	Да				
Попадание истинной дисперсии в ДИ	Да				

Медиана во всех выборках также лежит в узком диапазоне $127.938 - 128.213$ секунд, а выборочные средние − в пределах $126.712 - 127.167$ секунд. Истинное среднее (126.912 сек) и дисперсия (7832.156 сек) по всей совокупности попали в 95%-ные доверительные интервалы всех пяти подвыборок, что свидетельствует о высокой репрезентативности и устойчивости оценок.

Стоит заметить, что, в отличие от первого датасета, здесь дисперсия оценивается надежно, что указывает на отсутствие выраженных экстремальных выбросов.

Усеченное среднее составило 125.049 секунд. Оно устойчиво ниже как обычного среднего (~ 126.912 секунд), так и медианы (~ 128.071 секунд), что согласуется с правосторонней асимметрией распределения и повторяет тенденцию, наблюдаемую в первом датасете.

Перейдем к проверке на соответствие исследуемых данных известным теоретическим распределениям.

2. Анализ представленных распределений и исследование закономерностей данных

Для начала проведем анализ наличия асимметрии. Воспользуемся коэффициентом асимметрии. Сформулируем следующие гипотезы:

- H_0: распределение симметрично;
- H_1: распределение асимметрично;
- p-value < 0.0001.

Для проверки гипотезы воспользуемся коэффициентом асимметрии [10]:

$$K_a = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^3. \quad (13)$$

Значение коэффициента для первого датасета составило 1007.141. p-value составил 0. Высокое значение статистики и практически нулевой p-value четко указывают на сильную правостороннюю асимметрию. Нулевая гипотеза отвергнута.

Значение коэффициента для второго набора данных составило 0.061. p-value также составил 0. Нулевая гипотеза отвергнута.

Проверим, подчиняется ли выборка следующим законам распределений: логнормальное, Максвелла–Больцмана и Лапласа при помощи непараметрического и графического методов.

Для проверки на принадлежность к логнормальному распределению будем использовать формулу плотности:

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right)}, \quad (14)$$

где $x > 0$ [7].

Распределение Максвелла–Больцмана описывает распределение скоростей молекул в идеальном газе. Имеет фиксированную форму с пиком и быстрым спадом. Формула теоретической плотности:

$$f(x;a) = \sqrt{\frac{2}{\pi}} \cdot \frac{x^2}{a^3} \cdot e^{\left(-\frac{x^2}{2a^2}\right)}, \quad (15)$$

где $a = \frac{x}{\sqrt{2}}$ и $x \geq 0$ [11].

Распределение Лапласа – это двустороннее экспоненциальное распределение с пиком в нуле. Оно симметрично, но имеет более тяжелые "хвосты", чем нормальное [12, с. 189]. Формула теоретической плотности:

$$f(x;\mu,b) = \frac{1}{2b} \cdot e^{\left(-\frac{|x-\mu|}{b}\right)}, \quad (16)$$

где $b = \frac{\sigma}{\sqrt{2}}$ и $x \in (-\infty;\infty)$.

В качестве непараметрического метода воспользуемся критерием Колмогорова–Смирнова. Данная статистика сравнивает эмпирическую и теоретическую функции распределения, измеряя максимальное отклонение между ними и не требует группировки данных. Для вычисления воспользуемся формулой:

$$D_n = \sqrt{n} \cdot \sup_x |F_n(x) - F(x)|, \quad (17)$$

где D_n – статистика критерия Колмогорова;
$F_n(x)$ – выборочная функция распределения;
$F(x)$ – функция распределения [13].

Статистики по гипотезам о соответствии распределениям по первому набору данных представлены в табл. 5.

Таблица 5. *Статистики по гипотезам о соответствии распределениям по датасету IEEE*

Распределение	Статистика $\boldsymbol{D_n}$	p-value	Решение по гипотезе
Логнормальное	0.292	0	Отвергается
Максвелла–Больцмана	0.327	0	Отвергается
Лапласа	0.192	0	Отвергается

Как видно, ни одно из трёх предполагаемых распределений не прошло тесты. Перейдём к проверке второго датасета. Статистики представлены в табл. 6.

Таблица 6. *Статистики по гипотезам о соответствии распределениям по датасету CAIDA*

Распределение	Статистика $\boldsymbol{D_n}$	p-value	Решение по гипотезе
Логнормальное	0.308	0	Отвергается
Максвелла–Больцмана	0.195	0	Отвергается
Лапласа	0.100	0	Отвергается

Результаты теста аналогичны предыдущим – ни одно распределение не прошло тесты. Это в очередной раз указывает на то, что исследуемый *DDoS*-трафик обладает более сложной структурой, чем предполагают классические модели распределений.

Перейдем к идентификации распределений графическим способом при помощи анаморфоз. Анаморфоза – графический метод, при котором эмпирическая функция распределения отображается в координатах, соответствующих теоретическому закону. При соответствии точки лежат на прямой.

Оценка соответствия будет проводиться при помощи коэффициента детерминации R^2, поскольку данная статистическая мера позволяет объективно сравнить, насколько хорошо разные распределения совпадают с эмпирическими данными. Вычисляют R^2 следующей формулой:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}, \tag{18}$$

где SSE – сумма квадратов остатков;
SST – общая сумма квадратов;
Y_i – наблюдаемые значения;
$\hat{Y}_i$ – предсказанные значения;
$\overline{Y}$ – среднее значение Y;
n – число точек.

Чем ближе $R^2 \to 1$, тем лучше распределение подходит [14].

Проанализируем первый набор данных. Статистики R^2 по распределениям представлены в табл. 7.

Таблица 7. *Статистики R^2 по распределениям по датасету IEEE*

Распределение	R^2
Логнормальное	0.760
Максвелла-Больцмана	0.640
Лапласа	0.899

График анаморфозы для логнормального распределения представлен на рис. 7.

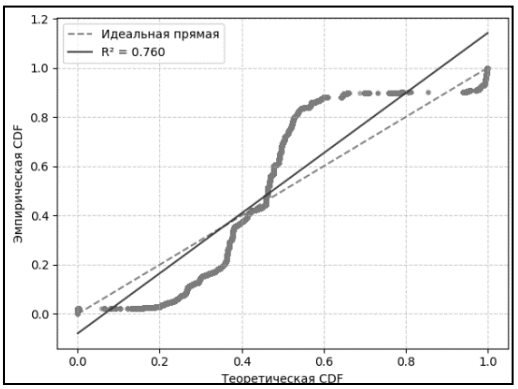

Рис. 7. *Анаморфоза логнормального распределения*

Точки на графике принимают *S*-образную форму, что говорит об отклонении в "хвостах" распределения причем как справа, так и слева от него. Несмотря на высокое значение коэффициента детерминации ($R^2 = 0.76$), ассимметрия данных не учитывается корректно, это будет оговорено в следующих этапах анализа.

График анаморфозы для распределения Максвелла–Больцмана представлен на рис. 8.

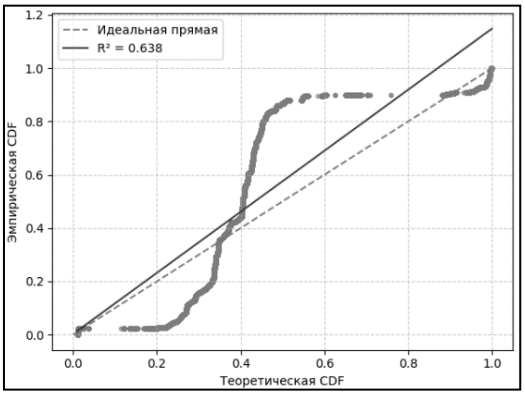

Рис. 8. *Анаморфоза распределения Максвелла–Больцмана*

График демонстрирует схожую с предыдущим *S*-образную форму. Отличительная особенность наблюдается в более выраженном отклонении в правом "хвосте" и смещении центра влево. Данное отклонение ухудшает R^2 – здесь его значение уже составляет 0.638.

График анаморфозы для распределения Лапласа представлен на рис. 9.

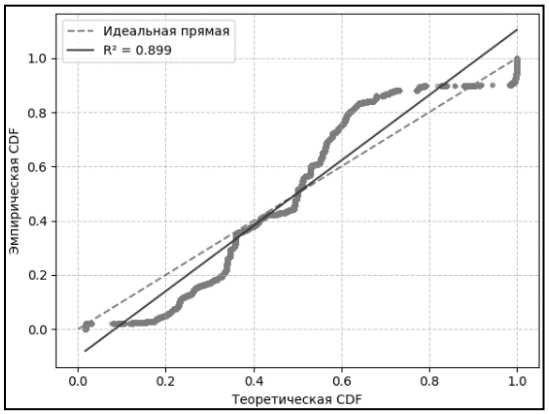

Рис. 9. *Анаморфоза распределения Лапласа*

Здесь точки также принимают *S*-образную форму, но менее выраженную, по сравнению с предыдущими анаморфозами. Также данная анаморфоза имеет наивысший среди всех $R^2 = 0.899$.

Перейдем к анализу второго датасета. Статистики R^2 по распределениям представлены в табл. 8.

Таблица 8. *Статистики R^2 по распределениям по датасету CAIDA*

Распределение	R^2
Логнормальное	0.665
Максвелла-Больцмана	0.956
Лапласа	0.974

График анаморфозы для логнормального распределения представлен на рис. 10.

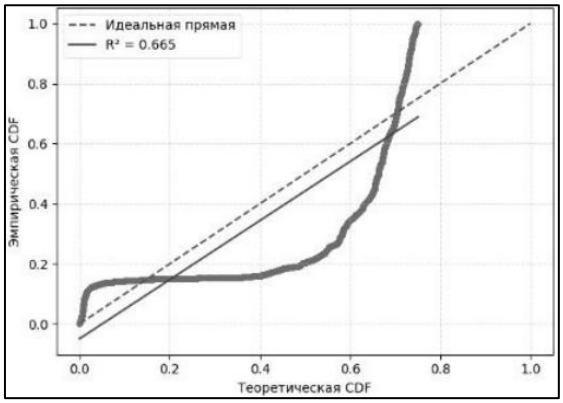

Рис. 10. *Анаморфоза логнормального распределения*

Точки образуют выраженную *S*-образную кривую, что указывает на систематическое отклонение в "хвостах" и отражается на значении R^2 (0.665). По сравнению с первым датасетом отклонение менее резкое, однако структура несоответствия сохраняется.

График анаморфозы для распределения Максвелла–Больцмана представлен на рис. 11.

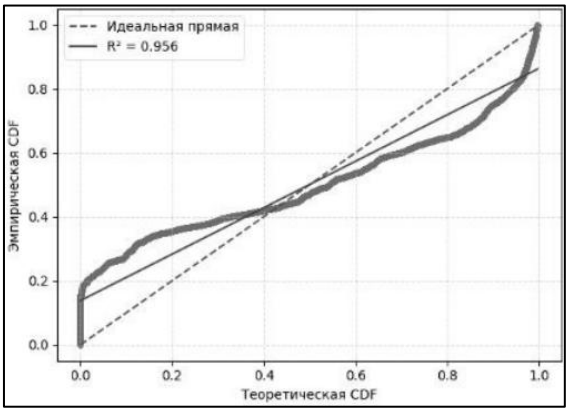

Рис. 11. *Анаморфоза распределения Максвелла–Больцмана*

На данном графике отклонения более линейны, в отличие от первого датасета, что отражается на более высоком значении R^2 (0.956).

График анаморфозы для распределения Лапласа представлен на рис. 12.

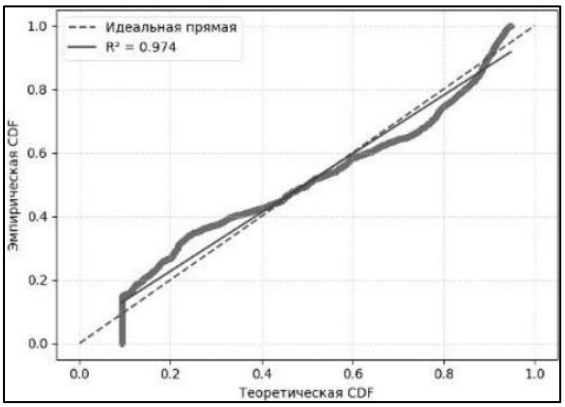

Рис. 12. *Анаморфоза распределения Лапласа*

Здесь можно наблюдать наиболее близкое приближение к прямой среди трех моделей. Значение коэффициента детерминации здесь также наиболее приближенно к 1 ($R^2 = 0.974$). Однако наблюдается слабая *S*-образная девиация, как и в первом наборе данных.

Тем не менее, распределение Лапласа демонстрирует наилучшее соответствие среди всех моделей как в первом, так и во втором наборах данных.

Несмотря на различия в масштабе и форме, оба датасета демонстрируют схожую структурную несогласованность с классическими параметрическими моделями, что подтверждает универсальность их статистической сложности.

Заключение

Таким образом, из трех рассмотренных моделей распределение Лапласа показало наилучшее соответствие как для первого, так и для второго датасета – у него самый высокий коэффициент детерминации. Однако даже в этом случае модель не может являться подходящей для описания *DDoS*-трафика, поскольку для нее характерна симметричность относительно нуля, а оба набора данных явно смещены вправо.

Логнормальное распределение и распределение Максвелла–Больцмана описывают данные немного хуже, что согласуется с общим выводом о структуре трафика – она сложная, с тяжелым "хвостом", и простые параметрические модели ее не передают.

Результаты работы показали, что классические модели плохо описывают поведение *DDoS*-трафика, что указывает на необходимость применять робастные и непараметрические методы при построении систем обнаружения вторжений (*IDS/IPS*).

Результаты данного анализа подтверждают выводы С.В. Кочергина о необходимости учета структуры данных для эффективного обнаружения аномалий в киберфизических системах. Методы, подобные *Isolation Forest*, оказываются предпочтительными именно в условиях резких, но редких отклонений – каковыми и являются длительные соединения в *DDoS*-трафике [15].

Проведенный статистический анализ закладывает методологическую основу для более точной и осмысленной предобработки длительности сетевого трафика, что позволит корректно применять методы машинного обучения в задачах кибербезопасности.

Список источников

1. *Кодацкий Н. М., Муратов И. А.* Решения в кибербезопасности // StudNet. 2022. № 1. URL: https://cyberleninka.ru/article/n/resheniya-v-kiberbezopasnosti (дата обращения: 10.22.2025).

2. *Крыгин Н. Д.* DDoS-атаки и защиты от них // Столыпинский вестник. 2022. № 4. URL: https://cyberleninka.ru/article/n/ddos-ataki-i-zaschity-ot-nih (дата обращения: 10.22.2025).
3. *Nour Moustafa* // The Bot-IoT dataset from IEEE-DataPort. URL: https://dx.doi.org/10.21227/r7v2-x988 (дата обращения: 10.22.2025).
4. The CAIDA UCSD "DDoS Attack 2007" Dataset. URL: https://www.caida.org/catalog/datasets/ddos-20070804_dataset/ (дата обращения: 27.01.2026).
5. Cloudflare Radar. URL: https://radar.cloudflare.com/?dateRange=7d (дата обращения: 27.01.2026).
6. *Коновалова М. Е., Жиронкин С. А., Сагдеев Р. Р.* Особенности формирования инвестиционного портфеля розничных инвесторов // Вестник РГЭУ (РИНХ). 2024. № 2.
7. *Дьячкова О. Н., Михайлов А. Е.* Опыт применения эмпирических функций распределения для описания градостроительной деятельности (на примере Санкт-Петербурга) // Экология урбанизированных территорий. 2023. № 2.
8. *Mokhtar S. f., Yusof Z. Md, Sapiri H.* Confidence Interval Estimating the Mean of Normal Distribution and Skewed Distribution // Malaysian Journal of Fundamental and Applied Sciences. 2024. Vol. 20(5). P. 1124−1135. URL: https://www.researchgate.net/publication/384959202_Confidence_Interval_Estimating_the_Mean_of_Normal_Distribution_and_Skewed_Distribution (дата обращения: 10.22.2025).
9. *Галалу В. Г., Киракосян С. А., Аль-Карави Х. Ш. М., Турулин И. И*. Сравнительная оценка методов усреднений для фильтрации измерительных сигналов // Известия ЮФУ. Технические науки. 2023. № 2 (232). URL: https://cyberleninka.ru/article/n/sravnitelnaya-otsenka-metodov-usredneniy-dlya-filtratsii-izmeritelnyh-signalov (дата обращения: 10.22.2025).
10. *Буре В. М.* Теория вероятностей и вероятностные модели: учебник. В. М. Буре, Е. М. Парилина, А. А. Седаков. Санкт-Петербург: Лань, 2020. 296 с. ISBN 978-5-8114-3168-7. URL: https://e.lanbook.com/book/108328 (дата обращения: 10.22.2025).
11. *Тарханова А. М., Шулдикова Н. С*. Компьютерное моделирование процесса установления распределения молекул по энергиям // МНКО. 2023. № 2 (99). URL: https://cyberleninka.ru/article/n/kompyuternoe-modelirovanie-protsessa-ustanovleniya-raspredeleniya-molekul-po-energiyam (дата обращения: 10.22.2025).
12. *Chattamvelli R., ShanmugamR*. Laplace Distribution. In: Continuous Distributions in Engineering and the Applied Sciences. Part II. Synthesis Lectures on Mathematics & Statistics. Springer, Cham. URL: https://doi.org/10.1007/978-3-031-02435-1_4 P. 189–199 (дата обращения: 10.22.2025).
13. *Воробьёв Р. И., Брагин А. С.* Исследование статистических критериев проверки гипотез для целей определения местоположения // Экономика и качество систем связи. 2022. № 3 (25). URL: https://cyberleninka.ru/article/n/issledovanie-statisticheskih-kriteriev-proverki-gipotez-dlya-tseley-opredeleniya-mestopolozheniya (дата обращения: 10.22.2025).
14. *Шукуров И. А.* Парная регрессия и корреляция // Universum: технические науки. 2021. № 5–1 (86). URL: https://cyberleninka.ru/article/n/parnaya-regressiya-i-korrelyatsiya (дата обращения: 10.22.2025).
15. *Кочергин С. В., Артемова С. В., Бакаев А. А., Митяков Е. С., Вегера Ж. Г., Максимова Е. А*. Кибербезопасность смарт-сетей: сравнение подходов машинного обучения для обнаружения аномалий // Russian Technological Journal. 2024.

Т. 12(6). С. 7–19. URL: https://doi.org/10.32362/2500-316X-2024-12-6-7-19 (дата обращения: 10.22.2025).

Reference

1. Kodyatsky, N. M. and Muratov, I. A. (2022), "Solutions in Cybersecurity" [Resheniya v kiberbezopasnosti], *StudNet*, no 1, URL: https://cyberleninka.ru/article/n/resheniya-v-kiberbezopasnosti (accessed: 10.22.2025).
2. Krygin, N. D. (2022), "DDoS Attacks and Defenses Against Them" [DDoS-ataki i zashchity ot nikh], Stolypinsky Vestnik [Stolypinskii vestnik], no 4, URL: https://cyberleninka.ru/article/n/ddos-ataki-i-zaschity-ot-nih (accessed: 10.22.2025).
3. Moustafa, N. The Bot-IoT dataset from IEEE-DataPort, URL: https://dx.doi.org/10.21227/r7v2-x988 (accessed: 10.22.2025).
4. The CAIDA UCSD "DDoS Attack 2007" Dataset, URL: https://www.caida.org/catalog/datasets/ddos-20070804_dataset/ (accessed: 27.01.2026).
5. Cloudflare Radar, URL: https://radar.cloudflare.com/?dateRange=7d (accessed: 27.01.2026).
6. Konovalova, M. E., Zhironkin, S. A. and Sagdeev, R. R. (2024), "Peculiarities of Forming an Investment Portfolio for Retail Investors" [Osobennosti formirovaniya investitsionnogo portfelya roznichnyh investorov], *Vestnik of Rostov State University of Economics* [Vestnik RGEU (RINKh)], no 2, URL: https://cyberleninka.ru/article/n/osobennosti-formirovaniya-investitsionnogo-portfelya-roznichnyh-investorov (accessed: 10.22.2025).
7. Diachkova, O. N. and Mikhailov, A. E. (2023), "Experience in Applying Empirical Distribution Functions to Describe Urban Development Activities (on the Example of St. Petersburg)" [Opyt primeneniya empiricheskih funktsiy raspredeleniya dlya opisaniya gradostroitelnoy deyatelnosti na primere Sankt-Peterburga], *Ecology of Urbanized Territories* [Ekologiya urbanizirovannykh territoriy], no 2, URL: https://cyberleninka.ru/article/n/opyt-primeneniya-empiricheskih-funktsiy-raspredeleniya-dlya-opisaniya-gradostroitelnoy-deyatelnosti-na-primere-sankt-peterburga (accessed: 10.22.2025).
8. Mokhtar, S. F., Yusof, Z. M. and Sapiri, H. (2024), "Confidence Interval Estimating the Mean of Normal Distribution and Skewed Distribution", *Malaysian Journal of Fundamental and Applied Sciences*, no 20(5), pp. 1124−1135. URL: https://www.researchgate.net/publication/384959202_Confidence_Interval_Estimating_the_Mean_of_Normal_Distribution_and_Skewed_Distribution (accessed: 10.22.2025).
9. Galalu, V. G., Kirakosyan, S. A., Al-Karavi, Kh. Sh. M. and Turulin, I. I. (2023), "Comparative Evaluation of Averaging Methods for Filtering Measurement Signals" [Sravnitelnaya otsenka metodov usredneniy dlya filtratsii izmeritelnyh signalov], *Izvestiya SFedU. Engineering Sciences* [Izvestiya YuFU. Tekhnicheskie nauki], no 2(232), URL: https://cyberleninka.ru/article/n/sravnitelnaya-otsenka-metodov-usredneniy-dlya-filtratsii-izmeritelnyh-signalov (accessed: 10.22.2025).
10. Bure, V. M., Parilina, E. M. and Sedakov, A. A. (2020), *Probability Theory and Probabilistic Models* [Teoriya veroyatnostei i veroyatnostnye modeli], Available at: https://e.lanbook.com/book/108328 (accessed: 10.22.2025).
11. Tarkhanova, A. M. and Shuldykova, N. S. (2023), "Computer Simulation of the Process of Establishing the Distribution of Molecules by Energy" [Komp'yuternoe modelirovanie protsessa ustanovleniya raspredeleniya molekul po energiyam], *International Scientific Research Journal* [Mezhdunarodnyi nauchno-issledovatel'skii zhurnal], no 2(99), URL: https://cyberleninka.ru/article/n/kompyuternoe-modelirovanie-protsessa-ustanovleniya-raspredeleniya-molekul-po-energiyam (accessed: 10.22.2025).

12. Chattamvelli, R. and Shanmugam, R. (2021), *Laplace Distribution. In: Continuous Distributions in Engineering and the Applied Sciences – Part II. Synthesis Lectures on Mathematics & Statistics*, Springer, Cham, URL: https://doi.org/10.1007/978-3-031-02435-1_4 (accessed: 10.22.2025).
13. Vorobyov, R. I. and Bragin, A. S. (2022), "Research of Statistical Criteria for Hypothesis Testing for the Purposes of Location Determination" [Issledovanie statisticheskikh kriteriev proverki gipotez dlya tselei opredeleniya mestopolozheniya], *Economics and Quality of Communication Systems* [Ekonomika i kachestvo sistem svyazi], no 3(25), URL: https://cyberleninka.ru/article/n/issledovanie-statisticheskih-kriteriev-proverki-gipotez-dlya-tseley-opredeleniya-mestopolozheniya (accessed: 10.22.2025).
14. Shukurov, I. A. (2021), "Paired Regression and Correlation" [Parnaya regressiya i korrelyatsiya], *Universum: Technical Sciences* [Universum: tekhnicheskie nauki], no 5-1(86), URL: https://cyberleninka.ru/article/n/parnaya-regressiya-i-korrelyatsiya (accessed: 10.22.2025).
15. Kochergin, S. V., Artemova, S. V., Bakaev, A. A., Mityakov, E. S., Vegera, Zh. G. and Maksimova, E. A. (2024), "Cybersecurity of Smart Grids: A Comparison of Machine Learning Approaches for Anomaly Detection", *Russian Technological Journal* [Russ. Technol. J.], no 12(6), pp. 7–19, URL: https://doi.org/10.32362/2500-316X-2024-12-6-7-19 (accessed: 10.22.2025).

Информация об авторах:
А. Л. Шкерин – ассистент кафедры прикладной математики института информационных технологий РТУ – МИРЭА (119454, Россия, г. Москва, проспект Вернадского д. 78);
Д. И. Понаётов – студент 3-го курса направления 09.03.03 "Прикладная информатика (Управление данными)" института информационных технологий РТУ – МИРЭА (119454, Россия, г. Москва, проспект Вернадского д. 78);
Ф. В. Пискун – студент 3-го курса направления 09.03.03 "Прикладная информатика (Управление данными)" института информационных технологий РТУ – МИРЭА (119454, Россия, г. Москва, проспект Вернадского д. 78).

Information about the authors:
A. L. Shkerin – assistant at the Department of Applied Mathematics at the Institute of Information Technology RTU – MIREA (78 Vernadsky Avenue, Moscow, Russia, 119454);
D. I. Ponayotov – third-year student of the applied computer science program at the Institute of Information Technology RTU – MIREA (78 Vernadsky Avenue, Moscow, Russia, 119454);
F. V. Piskun – third-year student of the applied computer science program at the Institute of Information Technology RTU – MIREA (78 Vernadsky Avenue, Moscow, Russia, 119454).