DOI 10.37882/2223-2966.2024.10-2.07

МОНИТОРИНГ ДЛИТЕЛЬНОСТИ ТЕЛЕФОННЫХ ПЕРЕГОВОРОВ С ИСПОЛЬЗОВАНИЕМ ГАММА-РАСПРЕДЕЛЕНИЯ

MONITORING THE DURATION OF TELEPHONE CONVERSATIONS USING ПАММА DISTRIBUTION

K. Zubarev
G. Makarichev

Summary. This work is devoted to the construction of a model based on the gamma distribution, which allows identifying abnormal patterns in the data on the duration of telephone calls, which may indicate the presence of spam activity. The study used telephone call data that had undergone preliminary filtering and processing. A histogram of relative frequencies was constructed for the data, and point estimates of the gamma distribution parameters were obtained based on the method of moments. An interval estimate of the gamma distribution parameters was also made. At the final stage of the study, a method for identifying spam activity was proposed, based on checking the compliance of the sample of telephone calls with the gamma distribution with parameters within the constructed confidence intervals.

Keywords: gamma distribution, method of moments, confidence intervals, Pearson criterion, phone calls, spam.

Зубарев Кирилл Михайлович
Старший преподаватель, Московский государственный технический университет имени Н.Э. Баумана
zubarev.bmstu@mail.ru
Макаричев Георгий Олегович
Московский государственный технический университет имени Н.Э. Баумана (Москва)
makarichevgo@student.bmstu.ru

Аннотация. Настоящая работа посвящена построению основанной на гамма-распределении модели, позволяющей выявлять в данных о продолжительности телефонных звонков аномальные паттерны, которые могут свидетельствовать о наличии спам-активности.

В рамках исследования использовались данные телефонных звонков, которые прошли предварительную фильтрацию и обработку. Для данных была построена гистограмма относительных частот, также были получены точечные оценки параметров гамма-распределения, основанные на методе моментов. Также была произведена интервальная оценка параметров гамма-распределения. На завершающем этапе исследования предложена методика выявления спам-активности, основанная на проверке соответствия выборки телефонных звонков гамма-распределению с параметрами в пределах построенных доверительных интервалов.

Ключевые слова: гамма-распределение, метод моментов, доверительные интервалы, критерий Пирсона, телефонные звонки, спам.

Введение

В современных условиях роста объёма телефонных коммуникаций неизбежным становится увеличение числа мошеннических или спам-звонков [1,2]. Такая нежелательная активность может нарушать работу коммуникационных систем, увеличивать затраты как пользователей, так и операторов связи [1,2,3]. В связи с этим эффективный анализ данных становится крайне важным, а методы для выявления аномалий в данных телефонных переговоров становятся всё более актуальными и востребованными [2,3].

Гипотеза исследования состоит в том, что длительности звонков подчиняются гамма-распределению [4] с параметрами, находящимися в пределах определённых доверительных интервалов. Отклонение от ожидаемых параметров гамма-распределения может указывать на наличие в выборке спам-звонков, что позволит эффективно идентифицировать потенциальных спамеров и составлять списки кандидатов для проверки на спам-активность.

Таким образом, данная работа предлагает новый подход к выявлению спам-активности, что может способствовать улучшению методов мониторинга и анализа телефонных переговоров.

Первичная обработка данных

В качестве исходных данных будем использовать датасет телефонных звонков, объемом в 51411 запись. Произведя фильтрацию данных и объединив звонки на один номер, для которых разница между временем окончания одного и началом другого не превосходит 20 секунд, получим результирующий датасет объемом в $n = 6093$ запись. Извлекая данные о длительности звонков (в секундах), формируем конечную выборку.

Объединение звонков на один номер в данном случае обосновано тем, что короткие промежутки времени между ними естественным образом позволяют считать их частью одной телефонной сессии. Такой подход помогает уменьшить избыточность данных и предоставляет более точную картину поведения пользователей.

Оценка параметров закона распределения

Принято считать, что длительности звонков подчиняются экспоненциальному распределению. Однако

известно, что данное распределение является частным случаем гамма-распределения, плотность вероятности для гамма-распределения:

$$p_{\xi}(x,\alpha,\beta) = \begin{cases} \dfrac{x^{\alpha-1}exp\left(-\dfrac{x}{\beta}\right)}{\beta^{\alpha}\hat{}\,(\alpha)}, x \geq 0 \\ 0, x < 0 \end{cases} \quad (1)$$

где p_{ξ} — плотность вероятности; α — параметр формы гамма-распределения; β — параметр масштаба гамма распределения; $\Gamma(\alpha)$ — гамма-функция

Используя гамма-распределение для моделирования длительности телефонных переговоров, можно более гибко и точно описывать наблюдаемые данные. Помимо этого, оно имеет достаточно простые аналитические формулы для вычисления теоретических моментов. Найдем оценку параметра λ по методу моментов для выборки

$$\begin{cases} \hat{\alpha} = \dfrac{\overline{X}^2}{S^2} = 0.364128 \\ \hat{\beta} = \dfrac{S^2}{\overline{X}} = 1498.998938 \end{cases}$$

Дополним гистограмму относительных частот кривыми плотности вероятности показательного и гамма-распределений, считая параметры распределений равными своим точечным оценкам по методу моментов:

На рисунке 1 можно заметить, что кривая плотности вероятности показательного распределения является более пологой и описывает поведение выборки хуже, чем кривая плотности вероятности гамма-распределения. Воспользуемся критерием согласия Пирсона для численной оценки разницы между наблюдаемыми данными и предсказанными по гамма и показательному законам [5]. Зададимся уровнем значимости $a_s = 0.01$ и проверим гипотезы о виде распределения для показательного и гамма законов. Критическое значение статистики в случае показательного закона равно 24.72, вычисленная статистика $\chi^2 = 5678,56$, что говорит о том, что гипотеза отклоняется, для гамма-распределения получили следующее значения $\chi^2_{cr} = 23.21$, $\chi^2 = 49,4$, что также является основанием отклонить гипотезу, но значение выборочной статистики существенно меньше по сравнению с показательным законом. Это служит еще одним подтверждением того, что гамма-распределение лучше описывает наблюдаемые данные.

Тем не менее, на заданном уровне значимости обе гипотезы не выполняются. На других стандартных уровнях значимости выше заданного (0.05, 0.1) гипотезы также не будут выполняться.

Оптимизация ширины интервалов

Ввиду выраженной неравномерности плотности данных, полезным будет учесть их специфику оптимизируя

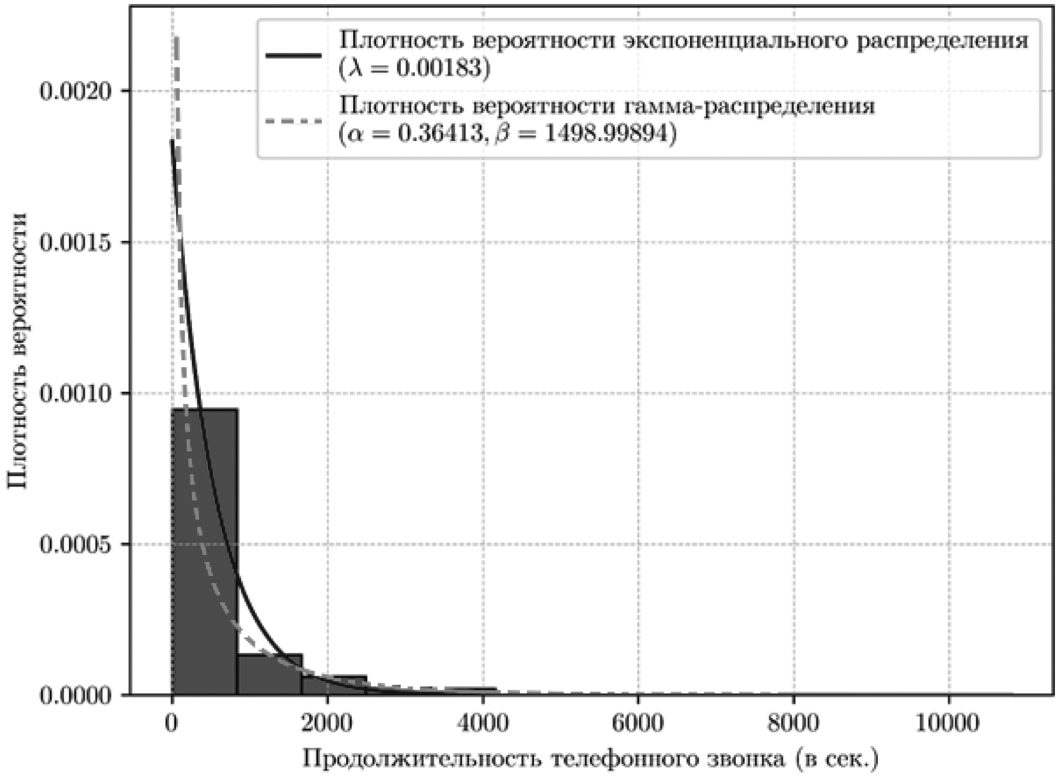

Рис. 1. Гистограмма относительных частот и кривые плотности показательного и гамма распределений

число и ширину интервалов [6,7]. Для решения данной задачи составим скалярную функцию следующего вида:

$$f_{\chi^2}\left(b_1, b_2, \ldots, b_k, b_{k+1}\right) = n \cdot \sum_{i=1}^{k} \frac{\left(p_i - p_i^*\right)^2}{p_i^*} \qquad (2)$$

Где n — объем выборки; k — количество интервалов гистограммы;
$b_1 = X_{min}, b_{k+1} = X_{max}$ — фиксированные граничные точки; $b_2, b_3, \ldots, b_k \left(b_1 < b_2 < b_3 < \ldots < b_k < b_{k+1}\right)$ — точки, определяющие границы интервалов гистограммы;
p_i–наблюдаемая относительная частота в интервале $\left[b_i, b_{i+1}\right)$; p_i^* — ожидаемая относительная частота в интервале $[b_i, b_{i+1})$

Стоит уточнить, что в случае $i = k$ правая граница интервала включается в него.

По виду функции становится ясно, что она возвращает значение статистики критерия Пирсона, характерное для набора интервалов с границами в точках $b_1, b_2, \ldots, b_k, b_{k+1}$. Наложив дополнительные естественные ограничения (обусловленные спецификой данных и здравым смыслом) и используя методы оптимизации (например, метод дифференциальной эволюции), минимизируем данную функцию. Найденная точка минимума функции будет определять оптимальные границы интервалов для наблюдаемых данных, учитывая их специфику.

Проверим значение статистики Пирсона, используя уровень значимости $a_s = 0.1$ и получим следующие значения $\chi^2_{cr} = 21.1$, $\chi^2 = 20{,}55$, и тогда, гипотеза о том, что данные подчиняются гамма-распределению, выполняется в соответствии с критерием согласия Пирсона. Так как гипотеза выполняется на уровне $a_s = 0.1$, она будет выполнятся и на других стандартных уровнях значимости меньше заданного.

Визуализируем полученный результат, построив гистограмму относительных частот и наложив на нее кривую плотности гамма-распределения (рис. 2).

Интервальные оценки

Построим доверительные интервалы для параметров гамма-распределения по имеющимся данным. Тогда если при анализе телефонной активности абонента возникнут существенные отклонения от гамма-распределения или оценки параметров не попадут в доверительные интервалы, которые будут вычислены ниже, то это может служить сигналом для дополнительной проверки абонента. Такие методы анализа находят применение и в других областях [8], в том числе при выявлении фальсификации на выборах [9]

Для построения интервальных оценок параметров гамма-распределения нами были использованы два различных способа: с применением центральной предель-

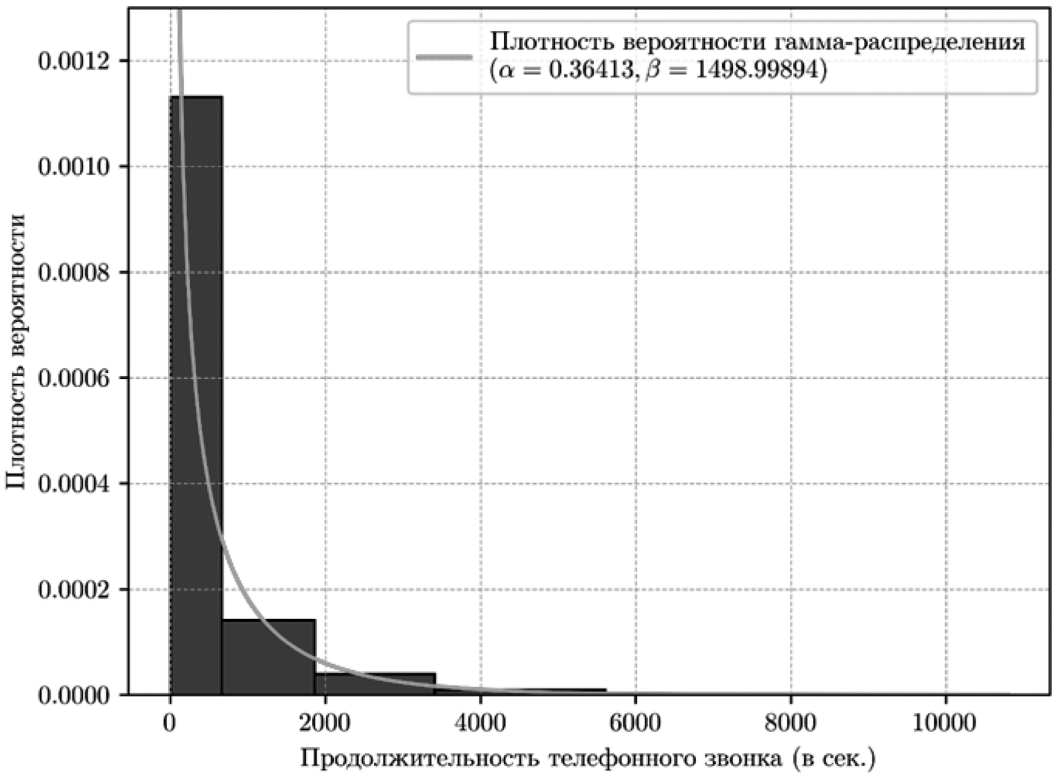

Рис. 2. Гистограмма относительных частот при оптимальных интервалах разбиения

ной теоремы (ЦПТ), а также на основе свойства устойчивости по суммированию гамма-распределения.

Посредством ЦПТ возможно построение только приближенных интервальных оценок и только в случае, если один из параметров распределения уже известен.

$$X_i \sim \Gamma(\alpha,\beta) \Rightarrow$$
$$\Rightarrow \sum_{i=1}^{n} X_i \overset{\text{цпт}}{\sim} N(n \cdot MX_i, n \cdot DX_i) = N(n\alpha\beta, n\alpha\beta^2) \Rightarrow$$
$$\Rightarrow \frac{\sum_{i=1}^{n} X_i - n\alpha\beta}{\beta\sqrt{n\alpha}} = \frac{n\bar{x} - n\alpha\beta}{\beta\sqrt{n\alpha}} = \sqrt{n}\frac{\bar{x} - \alpha\beta}{\beta\sqrt{\alpha}} \overset{\text{цпт}}{\sim} N(0,1)$$

Тогда, задавшись некоторым уровнем значимости и выбрав значениями двух некоторых точек соответствующие квантили стандартного нормального распределения:

$$P\left(-z_{1-\frac{a_s}{2}} < \sqrt{n}\frac{\bar{x} - \alpha\beta}{\beta\sqrt{\alpha}} < z_{1-\frac{a_s}{2}}\right) = 1 - a_s \qquad (3)$$

Полученное двойное неравенство из формулы (3) можно использовать для построения приближенных доверительных интервалов для параметров гамма-распределения на некотором заданном уровне значимости a_s:

$$-z_{1-\frac{a_s}{2}} < \sqrt{n}\frac{\bar{x} - \alpha\beta}{\beta\sqrt{\alpha}} < z_{1-\frac{a_s}{2}}, \qquad (4)$$

Где z_l — квантиль уровня l стандартного нормального распределения;

Считая, что параметр распределения α известен и равен своей оценке по методу моментов, будем интервально оценивать второй параметр распределения β. Аналогичную интервальную оценку произведем для параметра α, результаты отображены в таблице 1.

Таблица 1.

Доверительные интервалы для параметров гамма-распределения по ЦПТ, основанные на точечных оценках по методу моментов

Уровень значимости a_s	Значение квантиля $z_{1-\frac{a_s}{2}}$	Доверительный интервал для параметра α	Доверительный интервал для параметра β
0.1	1.64	(0.35, 0.38)	(1448.42, 1553.24)
0.05	1.96	(0.35, 0.38)	(1439.12, 1564.08)
0.01	2.58	(0.34, 0.38)	(1421.28, 1585.71)

Как говорилось ранее, полученные таким образом интервальные оценки являются приближенными и неточны по отношению к истинной генеральной совокупности.

Для параметра масштаба β однако можно построить более точный доверительный интервал, используя тот факт, что гамма-распределение устойчиво по суммированию. Но как и в случае с построением интервалов по ЦПТ, мы должны знать значение другого параметра α.

$$X_i \sim \Gamma(\alpha,\beta) \Rightarrow \sum_{i=1}^{n} X_i = n\bar{x} \sim \hat{}(n\alpha,\beta) \Rightarrow$$
$$\Rightarrow \frac{n\bar{x}}{\beta} \sim \Gamma(n\alpha,1) \qquad (5)$$

Тогда, задавшись некоторым уровнем значимости и выбрав значениями двух некоторых точек соответствующие квантили гамма-распределения с параметрами $n\alpha$ и 1:

$$P\left(q_{\frac{a_s}{2}} < \frac{n\bar{x}}{\beta} < q_{1-\frac{a_s}{2}}\right) = 1 - a_s \qquad (6)$$

Таким образом, следующее двойное неравенство из формулы (6) можно использовать для построения приближенных доверительных интервалов для параметра β гамма-распределения на некотором заданном уровне значимости a_s:

$$q_{\frac{a_s}{2}} < \frac{n\bar{x}}{\beta} < q_{1-\frac{a_s}{2}}, \qquad (7)$$

Где q_l — квантиль уровня l гамма-распределения с параметрами $n\alpha$ и 1;

Произведем оценивание параметра β, Значение параметра α полагаем равным оценке по методу моментов, результаты отображены в таблице 2.

Таблица 2.

Доверительные интервалы, построенные с использованием свойств гамма-распределения

Уровень значимости a_s	Значение квантиля $q_{\frac{a_s}{2}}$	Значение квантиля $q_{1-\frac{a_s}{2}}$	Доверительный интервал для параметра β
0.1	2141.73	2296.67	(1448.06, 1552.82)
0.05	2127.27	2311.90	(1438.53, 1563.38)
0.01	2099.18	2341.84	(1420.14, 1584.30)

Как видно из таблицы 2, интервалы для параметра β, построенные вторым способом, имеют меньшую ширину и обеспечивают более высокую точность, так как не используют приближений, связанных с центральной предельной теоремы.

Полученные доверительные интервалы для параметров распределения можно использовать для выявле-

ния в данных аномалий, которые могут свидетельствовать о наличии мошеннической или спам-активности. Если гистограмма длительностей звонков абонента не соответствует гамма-распределению с параметрами, попадающими в эти интервалы, оператору связи следует рассмотреть возможность проверки такого абонента.

Выводы

В результате проведенного статистического исследования была разработана модель на основе гамма-распределения и его доверительных интервалов, позволяющая выявлять в данных аномалии, потенциально свидетельствующие о наличии спам-активности.

Анализ доверительных интервалов показал, что для интервальной оценки параметра β предпочтительней использовать способ на основе свойства устойчивости по суммированию, присущему гамма-распределению. Это может увеличить точность предложенной модели.

ЛИТЕРАТУРА

1. Смирнов, В.М., Захарова А.И. Методы защиты от спам-звонков // Тенденции развития науки и образования. — 2023. — № 96-8. — С. 94–97.
2. Ковалев, С.С., Шишаев М.Г. Современные методы защиты от нежелательных почтовых рассылок // Труды Кольского научного центра РАН. — 2011. — № 4(7). — С. 100–111.
3. Сулейменова, Р.Д., Антонов И.В., Патутин В.В. Способы и методы защиты от СПАМ звонков в современном обществе // Развитие науки и практики в глобально меняющемся мире в условиях рисков: Сборник материалов XIX Международной научно-практической конференции, Москва, 30 мая 2023 года. — Москва: Алеф, 2023. — С. 153–160.
4. Облакова Т.В., Зубарев К.М., Сальникова А.А., Шинаков Д.С. Математические и инженерные примеры законов распределний случайных величин в ЦОС Nomotex // Дневник науки. — 2022. — № 12(72).
5. Гателюк, О.В., Манюкова Н.В. Проверка статистических гипотез — Санкт-Петербург: Издательство «Лань», 2022. — 112 с.
6. Лемешко, Б.Ю. Асимптотически оптимальное группирование наблюдений в критериях согласия // Заводская лаборатория. Диагностика материалов. — 1998. — Т. 64, № 1. — С. 56–64.
7. Никулин М.С. Критерий хи-квадрат для непрерывных распределений с параметрами сдвига и масштаба/Теория вероятностей и ее применение. 1973. Т. XVIII. № 3. С.583–591.
8. Облакова, Т.В., Зубарев К.М., Яковлев Д.Ю. Анализ распределения высоты морских волн. Сравнение оценок и применение критерия согласия Пирсона // Дневник науки. — 2023. — № 12(84).
9. Подлазов, А.В. Выборы депутатов Государственной Думы VII созыва: Выявление фальсификаций результатов и их реконструкция // Социологические исследования. — 2018. — № 1(405). — С. 59–72.

Журнал «Современная наука: актуальные проблемы теории и практики»