

Тунников Т.А.,
студент группы АИС-31, Стерлитамакский филиал УУНиТ,
Россия, г. Стерлитамак
Гнатенко Ю.А.,
кандидат физико-математических наук, доцент, Стерлитамакский
филиал УУНиТ,
Россия, г. Стерлитамак

СТАТИСТИЧЕСКИЙ АНАЛИЗ ПОЛЬЗОВАТЕЛЬСКОЙ АКТИВНОСТИ В ИНФОРМАЦИОННЫХ СИСТЕМАХ

Аннотация: в статье рассматривается задача количественной оценки пользовательской активности в информационных системах по данным журналов регистрации. Актуальность обусловлена необходимостью отличать обычные колебания нагрузки от статистически значимых отклонений. Сформулирована задача выбора показателей для характеристики интенсивности и устойчивости обращений к системе. На модельной выборке показано применение относительной частоты, среднего значения, дисперсии и среднеквадратического отклонения. Рассмотрена проверка статистической значимости изменений и выявление аномальных наблюдений. Показана практическая применимость вероятностно-статистических методов при первичном анализе журналов информационных систем.

Ключевые слова: теория вероятностей, математическая статистика, информационные системы, журналы событий, пользовательская активность, статистический анализ, дисперсия, аномалии, проверка гипотез, прогнозирование.

Annotation: the article considers the problem of quantitatively assessing user

activity in information systems using event logs. The relevance is related to the need to distinguish ordinary workload fluctuations from statistically significant deviations. The study formulates a set of indicators for describing the intensity and stability of system requests. A model sample is used to demonstrate relative frequency, mean value, variance, and standard deviation. Statistical significance testing and anomaly detection are also considered. The results demonstrate the practical applicability of probabilistic and statistical methods for the initial analysis of information system logs.

Keywords: *probability theory, mathematical statistics, information systems, event logs, user activity, statistical analysis, variance, anomalies, hypothesis testing, forecasting.*

Современные информационные системы формируют значительный поток данных о действиях пользователей, работе программных компонентов и состоянии инфраструктуры. Журналы регистрации позволяют фиксировать время обращения, тип операции, продолжительность сеанса и время отклика. Однако наличие данных само по себе не показывает, какие изменения являются закономерными, а какие требуют проверки. Поэтому практическая задача состоит не только в расчете статистических характеристик, но и в обоснованной интерпретации отклонений пользовательской активности.

Научная проблема исследования заключается в выборе простого и интерпретируемого набора вероятностно-статистических показателей, позволяющего по ограниченной выборке журналов выявлять существенные изменения пользовательской активности без сложных моделей машинного обучения. Объектом исследования являются процессы взаимодействия пользователей с информационной системой, предметом — статистические характеристики зарегистрированных событий. Цель работы — обосновать применение базовых статистических методов для обнаружения и интерпретации отклонений активности [1].

Для количественной оценки частоты наступления исследуемого события используется вероятность. В рамках анализа пользовательской активности событием А может быть, например, обращение к информационной системе в заданный временной интервал. Для дискретной ситуации вероятность определяется отношением числа благоприятных исходов к общему числу равновозможных исходов:

$$P(A) = m / n, (1)$$

где m — число наблюдений, соответствующих событию А, n — общее число наблюдений. В реальных журналах регистрации данная оценка интерпретируется как относительная частота события. При увеличении объема наблюдений она становится более устойчивой и может использоваться для сравнения различных временных периодов.

Математическая статистика позволяет перейти от отдельных наблюдений к количественной характеристике выборки. Исходными данными могут быть журналы событий, записи базы данных и результаты мониторинга серверов. Для оценки пользовательской активности целесообразно анализировать число обращений, среднюю интенсивность запросов, время отклика и разброс значений. Совместное рассмотрение этих показателей позволяет отделить изменение типичного уровня активности от увеличения неустойчивости процесса [2].

В качестве исследовательской задачи рассмотрим определение периода повышенной активности. Само по себе увеличение числа обращений не означает возникновения проблемы: в рабочие часы оно может быть нормальным. Поэтому наблюдаемое значение необходимо сопоставлять с характерным уровнем и вариативностью показателя. Существенное превышение обычного разброса служит основанием для дополнительной проверки причин изменения нагрузки.

Для иллюстрации подхода рассмотрим модельную выборку из 5000 зарегистрированных входов в информационную систему за исследуемый период. Из них 3200 входов были выполнены в рабочее время. В данном случае событием

А является вход пользователя в рабочее время, а относительная частота этого события используется как оценка его вероятности. Пример является модельным и предназначен для демонстрации методики статистического анализа.

Используя формулу (1), получим:

$$P(A) = 3200 / 5000 = 0,64. (2)$$

Оценка вероятности входа в рабочее время составляет 0,64, или 64 %. Однако интерпретировать ее как точную характеристику всей совокупности без учета статистической неопределенности некорректно. Поэтому далее оцениваются устойчивость результата и разброс наблюдений.

При $n = 5000$ стандартная ошибка оценки доли для $P = 0,64$ составляет приблизительно 0,0068. Приближенный 95%-й доверительный интервал находится в пределах 0,627–0,653. Следовательно, даже при большом числе наблюдений оценка имеет статистическую неопределенность. Доверительный интервал позволяет корректнее сравнивать периоды и не считать небольшую разницу относительных частот доказательством изменения активности.

Для характеристики уровня пользовательской активности необходимо учитывать и количественные значения случайной величины. Среднее значение отражает типичный уровень показателя, а дисперсия и среднееквадратическое отклонение характеризуют степень его изменчивости. Это особенно важно для параметров, связанных с производительностью информационной системы: два периода могут иметь одинаковое среднее время отклика, но существенно различаться по стабильности работы.

Для вычисления среднего значения используется формула: $\bar{x} = (x_1 + x_2 + \dots + x_n) / n$, (3)

где $x_1, x_2, \dots, x_n$ — значения выборки, n — количество наблюдений. Среднее значение характеризует типичный уровень нагрузки, дисперсия — степень отклонения наблюдений от него, а среднееквадратическое отклонение имеет ту же размерность, что и исходный показатель, поэтому удобно для практической интерпретации [3].

При анализе времени отклика веб-приложения одного среднего значения недостаточно. Если среднее остается приемлемым, но среднеквадратическое отклонение возрастает, это указывает на увеличение нестабильности обслуживания запросов. Поэтому при мониторинге целесообразно одновременно отслеживать центральную тенденцию и разброс показателя.

Следующий этап связан с проверкой статистических гипотез. После обновления программного обеспечения, изменения конфигурации сервера или пользовательского сценария могут измениться число ошибок и время отклика. Возникает вопрос, вызвана ли разница проведенным изменением или она объясняется случайными колебаниями. Нулевая и альтернативная гипотезы позволяют перевести этот вопрос в процедуру статистической проверки.

Например, нулевая гипотеза H_0 может предполагать отсутствие различий между средним временем отклика до и после обновления, а альтернативная H_1 — наличие различий. Конкретный критерий выбирается с учетом объема выборок, характера распределения и независимости наблюдений. Само изменение среднего показателя еще не доказывает наличие эффекта от модернизации [4].

Практическое значение имеет выявление аномальных наблюдений. Значение показателя можно сравнивать со средним уровнем и среднеквадратическим отклонением. Резкое увеличение числа запросов, необычно высокая активность учетной записи или значительное увеличение времени отклика могут быть отмечены как отклонения для дальнейшей проверки. При этом статистический метод не устанавливает причину аномалии, а лишь помогает своевременно выделить ее.

Таким образом, поставленная задача решается последовательным применением статистических методов: формируется выборка из журналов, рассчитываются относительные частоты и описательные показатели, оценивается разброс, а при необходимости проверяется статистическая значимость различий. Такой подход позволяет перейти от накопления журналов

к обоснованной интерпретации данных. Перспективным продолжением является сравнение классической статистики с методами обнаружения аномалий и прогнозирования.

Список литературы:

1. Гмурман В. Е. Теория вероятностей и математическая статистика : учебное пособие. – М. : Юрайт, 2022. – 479 с.
2. Вентцель Е. С. Теория вероятностей : учебник. – М. : КНОРУС, 2021. – 658 с.
3. Кремер Н. Ш. Теория вероятностей и математическая статистика : учебник для вузов. – М. : ЮНИТИ-ДАНА, 2022. – 573 с.
4. Орлов А. И. Прикладная статистика : учебник. – М. : Экзамен, 2021. – 687 с.
5. Фатхутдинова Р.Р., Гнатенко Ю.А. ОБЗОР СТАТИСТИЧЕСКОГО АНАЛИЗА ПОКАЗАТЕЛЕЙ // Тенденции развития науки и образования. 2024 №110-18. С. 148-151
6. Янбердин И.Р., Гнатенко Ю.А. Применение теории вероятности в статистике и машинном обучении // "Современные тенденции развития науки и мирового сообщества в эпоху цифровизации" (шифр –МКСТР). Сборник материалов XXXVI Международной научно-практической конференции. Москва, 2025. С. 295-297.
7. Исмагилов А.Р., Гнатенко Ю.А. Роль теории вероятности и математической статистики в повседневной жизни // "Современные тенденции развития науки и мирового сообщества в эпоху цифровизации" (шифр – МКСТР). Сборник материалов XXXVI Международной научно-практической конференции. Москва, 2025. С. 298-301.