УДК 004.942:519.25

**С.Б. Приходько, Л.Н. Макарова**

# ВЫБОР НОРМАЛИЗИРУЮЩЕГО ПРЕОБРАЗОВАНИЯ ДЛЯ ОЦЕНКИ ВРЕМЕНИ НАРАБОТКИ МЕЖДУ ОТКАЗАМИ УСТРОЙСТВ ТЕРМИНАЛЬНОЙ СЕТИ В СЛУЧАЕ МАЛОЙ ВЫБОРКИ

**Постановка проблемы.** В практике статистического анализа эмпирических данных часто приходится работать с малыми выборками, когда количество значений в выборке $n \leq 30$. Предметная область применения малых выборок достаточно широка: это и вопросы управления предприятием в случае мелкосерийного производства [1], и испытания электронной аппаратуры в разнообразных режимах [2]. Аналогичные примеры можно привести и в других отраслях знаний. Особенности использования малых выборок возникают не столько при организации испытаний, сколько при анализе полученных результатов.

Например, при выборе аналитической модели закона распределения времени наработки между отказами устройств терминальной сети [3] из 284 обработанных выборок 38% составили малые выборки.

При обработке малых выборок задача нахождения функции распределения случайной величины (СВ) и ее параметров принимает проблематичный характер, т.к. невозможно найти закон распределения традиционным способом, поскольку формально невозможно построить гистограмму распределения. Применение нормализирующих преобразований в данном случае частично решает проблему, т.к. дает возможность если не найти закон распределения СВ, то сделать выводы о параметрах распределения этой СВ, например, определить доверительные интервалы для точечной оценки математического ожидания и среднеквадратического отклонения СВ.

**Анализ последних исследований и публикаций.** Понятие «малая выборка» не имеет универсального определения. Если рассматривать один из основных вопросов математической статистики — какова должна быть минимально необходимая информация для получения требуемой достоверности при отсутствии каких-либо ограничений по точности конечного результата статистического анализа — то ответ на поставленный вопрос дал Р. Фишер в своей работе [4]. Минимальный объем выборки не может быть меньше 4 значений, иначе неизбежно возникает систематическая ошибка — смещение — наличие которой является первым признаком отсутствия достаточности статистики [5].

Теория малых выборок была разработана английским статистиком В. Госсетом (Стьюдентом), который построил специальное распределение, позволившее и при малых выборках соотносить СВ и доверительную вероятность. При $n > 100$ таблицы распределения Стьюдента дают те же результаты, что и таблицы интеграла вероятностей Лапласа, при $30 \leq n \leq 100$ различия незначительны. Поэтому практически к малым выборкам относят выборки объемом менее 30 значений [6]. Исследования, касающиеся малых выборок, связаны с именами А.Н. Колмогорова, Дж. Неймана и А. Вальда.

На сегодняшний день для оценки времени наработки между отказами устройств терминальной сети, как это принято в теории надежности, в качестве аналитической модели применяется экспоненциальный закон распределения СВ, однако, как это было показано в [3], эта модель не является адекватной эмпирическому распределению. Поэтому необходимо выполнить нормализацию эмпирических данных. К тому же, существует проблема — оценка доверительных интервалов математического ожидания и среднеквадратического отклонения, решение которой также потребует нормализации исходных данных.

Для нормализации эмпирических данных на сегодняшний день наиболее часто используются преобразования Джонсона, Бокса-Кокса и различные его модификации, например, Менли, Иео-Джонсона [7]. Преобразование Бокса-Кокса не всегда позволяет нормализировать эмпирические данные, особенно с бимодальным и U-образным законами распределения СВ в силу того, что оно является однопараметрическим. Кроме того, определенным минусом преобразования Бокса-Кокса является отсутствие единой методики определения параметров распределения. В таком случае лучше использовать четырехпараметрическое преобразование Джонсона, для которого существует единая методика определения параметров распределения. Для малой выборки был предложен непараметрический подход нахождения неизвестных параметров распределения.

Выбор конкретного нормализирующего преобразования необходимо выполнять в зависимости от обрабатываемых эмпирических данных. Для времени наработки между отказами устройств терминальной сети этот вопрос не исследовался.

**Целью данной статьи** является выбор нормализирующего преобразования для оценки времени наработки между отказами устройств терминальной сети в случае малой выборки ($n \leq 30$).

**Изложение основного материала.** Преобразование Джонсона в общем случае имеет вид [8]:

$$z = \gamma + \eta q(x, \phi, \lambda); -\infty < \gamma < \infty; \eta > 0; -\infty < \phi < \infty; \lambda > 0, \quad (1)$$

где $q$ — нелинейная функция; γ, η, φ, λ — параметры распределения, причем γ и η — параметры формы, φ — параметр смещения, λ — параметр масштаба; $x$ — СВ, которая нормализируется; $z$ — нормированная нормально распределенная СВ.

Семейству функций $q_1$ соответствует логарифмически нормальное распределение $S_L$ Джонсона, семейству функций $q_2$ соответствует семейство распределений $S_B$ Джонсона, семейству функций $q_3$ соответствует семейство распределений $S_U$ Джонсона:

$$\begin{aligned} q_1(x,\phi,\lambda) &= \ln\left(\frac{x-\phi}{\lambda}\right), x > \phi; \\ q_2(x,\phi,\lambda) &= \ln\left(\frac{x-\phi}{\lambda+\phi-x}\right), \phi < x < \phi + \lambda; \\ q_3(x,\phi,\lambda) &= Arsh\left(\frac{x-\phi}{\lambda}\right), -\infty \le x \le +\infty. \end{aligned} \quad (2)$$

Конкретное семейство распределений Джонсона выбирается исходя из значений квадрата асимметрии $A^2$ и эксцесса ε исходной выборки [9]. Для анализируемых эмпирических данных о времени наработки между отказами устройств терминальной сети, как было показано в [3], таким семейством является семейство распределений $S_B$ Джонсона.

В случае малой выборки значения неизвестных параметров распределения можно найти с помощью непараметрического метода решения задачи математического программирования, описанного в [7]:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\left\{A_z^2 + (\varepsilon_z - 3)^2 + m_z^2 + (D_z - 1)^2\right\}, \quad (3)$$

где θ — вектор неизвестных параметров, θ={γ,η,φ,λ}, $A_z = \frac{1}{n\sigma_z^3}\sum_{i=1}^{n}(z_i - m_z)^3$ — асимметрия нормализированной выборки, $\varepsilon_z = \frac{1}{n\sigma_z^4}\sum_{i=1}^{n}(z_i - m_z)^4$ — эксцесс нормализированной выборки, $m_z = \frac{1}{n}\sum_{i=1}^{n} z_i$ — математическое ожидание нормализированной выборки, $D_z = \sigma_z^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - m_z)^2$ — дисперсия нормализированной выборки, $n$ — количество значений в выборке.

Преобразование Бокса-Кокса имеет вид:

$$z = x(\lambda) = \begin{cases} \frac{x^\lambda - 1}{\lambda}, \text{при } \lambda \neq 0 \\ \ln(x), \text{при } \lambda = 0 \end{cases}, \quad (4)$$

где $z$ — нормированная нормально распределенная СВ, $x$ — СВ, которая нормализируется, λ — параметр распределения.

Значение неизвестного параметра λ можно определить двумя путями [7]:

– по максимуму логарифма функции правдоподобия:

$$f(x,\lambda) = -\frac{n}{2}\ln\sum_{i=1}^{n}\frac{(z_i(\lambda) - \bar{z}(\lambda))^2}{n} + (\lambda - 1)\sum_{i=1}^{n}\ln(x_i), \quad (5)$$

где $\bar{z}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} z_i(\lambda)$, $n$ — количество значений в выборке;

– в результате решения следующей задачи математического программирования:

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmin}}\left\{A_z^2 + (\varepsilon_z - 3)^2\right\}. \quad (6)$$

В качестве исходных эмпирических данных были рассмотрены 150 выборок из 2509 случаев отказа в обслуживании 231 устройства терминальной сети за период 36 месяцев, в которых количество значений в выборке $10 \le n \le 30$. Данные группировались по тем же признакам, что и для анализа, проведенного в [3]. Статистические параметры типичных выборок приведены в табл. 1.

Таблица 1

Статистические параметры выборок эмпирических данных

| Параметр | Выборка 1 | Выборка 2 |
|---|---|---|
| Количество значений | 27 | 20 |
| Выборочное среднее | 35059 | 282011 |
| Дисперсия | $2{,}08\cdot10^{9}$ | $5{,}67\cdot10^{10}$ |
| Среднеквадратическое отклонение | 45624 | 238075 |
| Асимметрия | 1,67 | 0,21 |
| Эксцесс | 5,40 | 1,52 |

Как видно из приведенных данных, выборки не являются нормальными, и их можно нормализировать с помощью описанных выше преобразований.

Нормализация с помощью преобразования Джонсона. Исходя из значений квадрата асимметрии $A^2$ и эксцесса ε исходных выборок, было выбрано семейство распределений $S_B$ Джонсона. В результате решения задачи (3) были найдены следующие параметры:

– для выборки 1: γ = 1,1121; η = 0,3997; φ = 436,1284; λ = 193195;

– для выборки 2: γ = 0,2844; η =0,3531; φ = 0,0001; λ = 691797.

Нормализация СВ x выполнена по преобразованиям (1), (2).

Нормализация с помощью преобразования Бокса-Кокса (вариант 1). В результате решения задачи (5) было найдено значение параметра λ:

– для выборки 1: λ = 0,5084;

– для выборки 2: λ = 0,1373.

Нормализация СВ *x* выполнена по преобразованию (4).

Нормализация с помощью преобразования Бокса-Кокса (вариант 2). В результате решения задачи (6) было найдено значение параметра λ:

– для выборки 1: λ = 0,5368;

– для выборки 2: λ = 0,1352.

Нормализация СВ *x* выполнена по преобразованию (4).

Сводные результаты нормализации приведенных выборок эмпирических данных представлены в табл. 2.

Проверка соответствия преобразованных выборок нормальному распределению выполнена с помощью критерия согласия Колмогорова-Смирнова [10]. Применяют несколько различных вариантов расчета соответствующих статистик, однако, если параметры закона распределения СВ заранее неизвестны, а оцениваются по данным выборки, то указанный критерий согласия может быть использован для проверки соответствия только некоторым законам распределения СВ. В частности, для проверки соответствия преобразованных данных нормальному распределению, использовались следующие формулы для вычисления статистики Колмогорова [11]:

$$D_n^+ = \max_{1\le i\le n}\left[\frac{i}{n} - \Phi(\hat{z}_i)\right], D_n^- = \max_{1\le i\le n}\left[\Phi(\hat{z}_i) - \frac{i-1}{n}\right], D_n = \max[D_n^+, D_n^-],$$
$$D_n\left(\sqrt{n} - 0{,}01 + \frac{0{,}85}{\sqrt{n}}\right) \le \lambda_\alpha, \quad (7)$$

где $\Phi(\hat{z}_i)$ — значение функции Лапласа для $\hat{z}_i = \frac{x_i - \bar{x}}{\sigma_x}$.

Статистика Смирнова была вычислена по следующим формулам, приведенным в [12]:

$$nw_n^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left(\Phi(\hat{z}_i) - \frac{2i-1}{2n}\right)^2,$$
$$nw_n^2\left(1 + \frac{1}{2n}\right) \le W_\alpha^2. \quad (8)$$

Критические значения статистики Колмогорова $\lambda_{\alpha кр}$ и статистики Смирнова $W^2_{\alpha кр}$ для уровня значимости α = 0,05 можно взять из [11, 12, 13].

Значения соответствующих статистик нормализированных выборок и их критические значения приведены в табл. 2.

Таблица 2

Результаты нормализации выборок эмпирических данных

| Параметр | Преобразование Джонсона | | Преобразование Бокса-Кокса | | | |
|---|---|---|---|---|---|---|
| | Выборка 1 | Выборка 2 | Выборка 1 | Выборка 2 | Выборка 1 | Выборка 2 |
| Формулы преобразования | (1), (2), (3) | | (4), (5) | | (4), (6) | |
| Значение целевой функции преобразования | $6{,}36\cdot10^{-6}$ | $2{,}65\cdot10^{-3}$ | $-2{,}66\cdot10^{2}$ | $-2{,}42\cdot10^{2}$ | $8{,}01\cdot10^{-2}$ | $7{,}14\cdot10^{-2}$ |
| Математическое ожидание | $-2{,}03\cdot10^{-3}$ | $-5{,}00\cdot10^{-8}$ | 316,6412 | 29,9601 | 406,1322 | 29,4796 |
| Дисперсия | 0,9987 | 1,0000 | 62213 | 64,6703 | 111406 | 61,7401 |
| Среднеквадратическое отклонение | 0,9994 | 1,0000 | 249,4253 | 8,0418 | 333,7752 | 7,8575 |
| Асимметрия | $5{,}59\cdot10^{-4}$ | $-2{,}73\cdot10^{-2}$ | 0,7727 | -0,7937 | 0,8231 | -0,8002 |
| Эксцесс | 3,0005 | 3,0436 | 2,5339 | 2,7096 | 2,6490 | 2,7290 |
| Статистика Колмогорова | 0,6747 | 0,8047 | 0,9325 | 1,0992 | 1,0435 | 1,0725 |
| Критическое значение статистики Колмогорова | 0,8950 | 0,8950 | 0,8950 | 0,8950 | 0,8950 | 0,8950 |
| Статистика Смирнова | 0,0498 | 0,0917 | 0,1324 | 0,1890 | 0,1984 | 0,1902 |
| Критическое значение статистики Смирнова | 0,1260 | 0,1260 | 0,1260 | 0,1260 | 0,1260 | 0,1260 |

Исходя из полученных значений статистик, с доверительной вероятностью 0,95 гипотеза о соответствии преобразованных выборок нормальному распределению принимается в случае преобразования Джонсона и отвергается в случае преобразования Бокса-Кокса. По данным, приведенным в табл. 2, можно сделать вывод, что преобразование Джонсона, выполненное с помощью формул (1) — (3), дает результат, наиболее близкий к нормальному распределению СВ.

**Выводы.** Получила дальнейшее развитие модель распределения времени наработки между отказами устройств терминальной сети на основе нормализирующего преобразования Джонсона.

Для выполнения расчетов было разработано соответствующее программное обеспечение на языке программирования Java. На основании проведенных расчетов показано, что нормализирующее преобразование Джонсона дает лучший результат, чем нормализирующее преобразование Бокса-Кокса.

В дальнейшем планируется использование полученных результатов для оценки доверительных интервалов математического ожидания и среднеквадратического отклонения времени наработки между отказами устройств терминальной сети.

ЛИТЕРАТУРА

1. Коваленко, И. И. Оценивание статистических характеристик вероятностных распределений малых выборок данных [Текст] / И. И. Коваленко, Т. С. Гавриш // Наукові праці: науково-методичний журнал. — Вип. 93. Т. 106. Комп`ютерні технології. — Миколаїв: Вид-во ЧДУ ім. Петра Могили, 2009. — С.32-38.
2. Гусев, А. Малые выборки при оценке работоспособности и надежности электронных компонентов [Текст] / А. Гусев, Э. Лидский, О. Мироненко // CHIP NEWS, 2002. — №1(64). — С.52-55.
3. Приходько, С. Б. Выбор аналитической модели закона распределения времени наработки между отказами устройств терминальной сети [Текст] / С. Б. Приходько, Л. Н. Макарова // Наукові праці: науково-методичний журнал. — Вип. 179. Т. 191. Комп`ютерні технології. — Миколаїв: Вид-во ЧДУ ім. Петра Могили, 2012 — С.42-45.
4. Фишер, Р. Статистические методы для исследователей [Текст] / Р. Фишер — М: Госстатиздат, 1958. — 267 с.
5. Леман, Э. Проверка статистических гипотез [Текст] / Э. Леман — М: Наука. Главная редакция физико-математической литературы, 1979. — 408 с.
6. Елисеева, И. И. Общая теория статистики: Учебник. Под. ред. чл.-корр. РАН И.И. Елисеевой [Текст] /И. И. Елисеева, М. М. Юзбашев — М.: Финансы и статистика, 2003. — 480 с.
7. Приходько, С. Б Інтервальне оцінювання статистичних моментів негаусівських випадкових величин на основі нормалізуючих перетворень [Текст] / С. Б. Приходько // Математичне моделювання: науковий журнал. — Дніпродзержинськ: ДДТУ, 2011. — №1 (24). — С.9-13.
8. Кендалл, М. Теория распределений [Текст] / М. Кендалл, А. Стьюарт — М.: Наука, 1966. — 588 с.

9. Коваленко, І. І. Сучасні методи статистичного аналізу даних: Навчальний посібник [Текст] / І. І. Коваленко, С. Б. Приходько, Л. О. Латанська. — Миколаїв: НУК, 2011. — 192 с.
10. Вентцель, Е. С. Теория вероятностей: Учеб. для вузов [Текст] / Е. С. Вентцель — М.: Высш. шк., 1999. — 576 с.
11. Степнов, М. Н. Статистические методы обработки результатов механических испытаний: Справочник [Текст] / М. Н. Степнов — М.: Машиностроение, 1985. — 232 с.
12. 12. Тюрин, Ю. Н. Непараметрические методы статистики [Текст] / Ю. Н. Тюрин — М.: Знание, 1978. — 64 с.
13. Большев, Л. Н. Таблицы математической статистики [Текст] / Л. Н. Большев, Н. В. Смирнов — М.: Наука. Главная редакция физико-математической литературы, 1983. — 416 с.

ПРИХОДЬКО Сергей Борисович – к.т.н., доцент, заведующий кафедрой программного обеспечения автоматизированных систем, Национальный университет кораблестроения имени адмирала Макарова.

МАКАРОВА Лидия Николаевна – соискатель, Национальный университет кораблестроения имени адмирала Макарова.

Научные интересы: математическое моделирование случайных процессов в информационных технологиях.