УДК 621.396.6
DOI: 10.17586/0021-3454-2023-66-11-899-906

ПРОВЕРКА СОГЛАСИЯ ТЕОРЕТИЧЕСКОГО И ЭМПИРИЧЕСКОГО РАСПРЕДЕЛЕНИЙ ПАРЕТО ДЛЯ КОМАНД И МИКРОКОМАНД ЭВМ С ПОМОЩЬЮ КРИТЕРИЯ КОЛМОГОРОВА

А. В. Аверьянов*, В. Т. Нгуен

Военно-космическая академия им. А. Ф. Можайского, Санкт-Петербург, Россия
**vka_24kaf@mail.ru*

Аннотация. Выполнена проверка гипотезы о согласии теоретического и эмпирического распределений Парето применительно к кумулятивным кривым — диаграммам для команд и микрокоманд учебной ЭВМ. В качестве статистического критерия согласия использован критерий Колмогорова. Получены значения параметров для функций распределения Парето, описывающих вероятностные свойства случайных величин, которыми являются порядковые номера команд и реализующих их микрокоманд. Построены графики теоретических и практических функций распределения, позволяющие исключить из системы команд ЭВМ редко используемые команды, что способствует упрощению архитектуры процессоров ЭВМ. Отдельные теоретические положения и полученные практические результаты являются дальнейшим развитием статистического метода повышения качества — анализа Парето применительно к количественному оцениванию метрик машинных команд и микрокоманд ЭВМ.

Ключевые слова: *критерий Колмогорова, функция распределения Парето, кумулятивная кривая Парето, уровень значимости, квантиль распределения, доверительный интервал, команды и микрокоманды ЭВМ, устройство управления ЭВМ*

Ссылка для цитирования: *Аверьянов А. В., Нгуен В. Т.* Проверка согласия теоретического и эмпирического распределений Парето для команд и микрокоманд ЭВМ с помощью критерия Колмогорова // Изв. вузов. Приборостроение. 2023. Т. 66, № 11. С. 899—906. DOI: 10.17586/0021-3454-2023-66-11-899-906.

CHECKING THE AGREEMENT OF THEORETICAL AND EMPIRICAL PARETO DISTRIBUTIONS FOR COMPUTER COMMANDS AND MICROCOMMANDS USING THE KOLMOGOROV CRITERION

A. V. Averyanov*, V. T. Nguyen

A. F. Mozhaisky Military Space Academy, St. Petersburg, Russia
**vka_24kaf@mail.ru*

Abstract. The hypothesis about the agreement of theoretical and empirical Pareto distributions is tested in relation to cumulative curves - diagrams for instructions and microinstructions of an educational computer. The Kolmogorov criterion is used as a statistical criterion for agreement. The values of the parameters are obtained for the Pareto distribution functions that describe the probabilistic properties of random variables, which are the ordinal numbers of commands and microcommands implementing them. Constructed graphs of theoretical and practical distribution functions make it possible to exclude rarely used commands from the computer command system, which helps to simplify the architecture of computer processors. Certain theoretical principles and the obtained practical results are a further development of the statistical method of improving quality, i.e. Pareto analysis, in relation to quantitative assessment of metrics of machine commands and microcommands of a computer.

Keywords: Kolmogorov criterion, Pareto distribution function, cumulative Pareto curve, significance level, distribution quantile, confidence interval, computer commands and microcommands, computer control device

For citation: Averyanov A. V., Nguyen V. T. Checking the agreement of theoretical and empirical Pareto distributions for computer commands and microcommands using the Kolmogorov criterion. *Journal of Instrument Engineering.* 2023. Vol. 66, N 11. P. 899—906 (in Russian). DOI: 10.17586/0021-3454-2023-66-11-899-906.

При решении ряда статистических задач вводится предположение, что некоторые случайные величины имеют заданное распределение (нормальное, экспоненциальное, равномерное и др.) с известными или неизвестными параметрами. При этом необходимо ответить на вопрос: насколько сделанные предположения о распределении случайных величин соответствуют опытным данным? Для решения этой задачи разработаны разные способы, иначе говоря, статистические критерии. Критериями согласия называют статистические критерии, предназначенные для обнаружения расхождений между гипотетической статистической моделью и реальными данными, которые эта модель призвана описать [1]. Важнейшим из этих критериев является критерий Колмогорова.

При использовании критерия Колмогорова в качестве меры расхождения теоретического и эмпирического распределений принимают число D_N [2], которое вычисляется по формуле

$$D_N = \sqrt{N} \cdot D_{\max}, \tag{1}$$

где N — количество испытаний, $D_{\max}$ — максимальная абсолютная разность теоретической и эмпирической функций распределения случайной величины x, т.е.

$$D_{\max} = \max_j \left| F\left(x_j\right) - \hat{F}\left(x_j\right) \right|. \tag{2}$$

В связи с тем, что значения эмпирической функции $\hat{F}\left(x_j\right)$ являются случайными, мера расхождения D_N есть случайная величина.

Для проверки гипотезы о правильности выбора теоретической функции распределения $F\left(x_j\right)$ задаются уровнем значимости α и находят квантиль распределения Колмогорова $z_{1-\alpha}$. Уровень значимости α определяет максимальное значение меры расхождения (1), которое можно считать случайным. Если в результате сравнения окажется, что $D_N < z_{1-\alpha}$, то делается заключение, что нет основания отвергнуть принятую гипотезу о виде функции распределения. В противном случае принятая на начальном этапе гипотеза отвергается и вся последовательность обработки информации повторяется, начиная с уточнения гипотезы о виде функции распределения. Квантили распределения Колмогорова приведены в табл. 1.

Таблица 1

α	$z_{1-\alpha}$	α	$z_{1-\alpha}$
0,01	1,63	0,20	1,07
0,025	1,48	0,25	1,02
0,05	1,36	0,30	0,97
0,10	1,22	0,35	0,93
0,15	1,14	0,40	0,89

Процедура проверки гипотезы по критерию Колмогорова состоит из следующих этапов.

1. Построение эмпирической функции распределения.

2. Выдвижение гипотезы о функции распределения и оценивание параметров этого распределения.

3. Вычисление значения теоретической функции распределения в точках x_j, которые соответствуют скачкам эмпирической функции распределения.

4. Вычисление в каждой из точек абсолютной разности

$$D_j = \left| F\left(x_j\right) - \hat{F}\left(x_j\right) \right|. \tag{3}$$

5. Выбор максимальной разности $D_{\max}$ и определение величины D_N.

6. Сравнение меры расхождения D_N с квантилем $z_{1-\alpha}$ распределения Колмогорова.

Применим рассмотренную процедуру для проверки гипотезы о правильности выбора функций распределений, соответствующих кумулятивным кривым-диаграммам Парето для команд и микрокоманд учебной ЭВМ. Критерий Колмогорова прост, но его применение на практике осложняется следующими обстоятельствами. Для сопоставления функций опытного и теоретического распределений необходимо знать параметры распределения, которые обычно определяются по результатам испытаний [2].

Кумулятивная кривая математически соответствует функции распределения Парето, которая задается равенством

$$F_X\left(x\right) = P\left(X < x\right) = 1 - \left(\frac{x_m}{x}\right)^k \text{ для } x \ge x_m;\ x_m > 0;\ k > 0, \tag{4}$$

где X — случайная величина; x — значение случайной величины X; k и x_m — параметры распределения Парето.

Для проверки согласия теоретического и эмпирического распределений с помощью критерия Колмогорова были выполнены 50 программ в мнемокодах учебной ЭВМ [3, 4]. Все использованные команды, реализующие выполненные программы, указаны в таблице [5, табл. 1], на основе которой строится диаграмма Парето. Первая графа таблицы соответствует 37 командам, используемым в 50 программах. Система команд учебной ЭВМ состоит из 42 команд. Пять команд при программировании не использовались. В последней графе таблицы показана накопленная доля от суммарного числа повторений, выраженная в процентах, для каждой команды. Значения накопленных долей соответствуют значениям $\hat{F}\left(x\right)$ — эмпирической функции распределения. В данном случае x — порядковый номер команды в указанной таблице [5, табл. 1], являющийся значением случайной величины X в формуле (4).

Команды, используемые при программировании, реализуются на микропрограммном уровне 48 микрокомандами, представленными в другой таблице [5, табл. 2], которая также была использована для построения кумулятивной кривой Парето.

Выдвигаем гипотезу о том, что поведение случайной величины X, соответствующей порядковым номерам команд и микрокоманд, описывается двухпараметрическим распределением Парето. Для сравнения эмпирической функции (кумулятивной кривой) и теоретической функции распределения необходимо определить значения параметров k и x_m.

Параметр x_m должен быть задан таким, чтобы любое значение случайной величины X было не меньше x_m, т.е. $x \ge x_m$. Как следует из собранного статистического материала, возможное минимальное значение случайной величины X соответствует первой команде или микрокоманде в рассмотренных таблицах [5, табл. 1, 2]. Поэтому задаем значение $x_m = 1$.

Параметр k является параметром масштаба, который имеет разные значения, зависящие от конкретного испытания. Задача нахождения параметра теоретической функции распределения при наличии совокупности опытных данных — точек — называется задачей „подгонка кривой“ [6, 7]. Решение подобной задачи заключается в вычислении наименьшей суммы квадратов разностей между значениями теоретической и эмпирической функций распределений.

Для представленных в указанных таблицах [5] статистических данных на основе решения задачи „подгонка кривой“ получены следующие значения параметров масштаба: для мнемокоманд $k = 0,66$ и реализующих их микрокоманд $k = 0,70$.

На рис. 1, *а*, *б* представлены графики эмпирической и теоретической функций распределения для команд с $k = 0,66$ (*а*) и микрокоманд с $k = 0,70$ (*б*) и заданным параметром x_m.

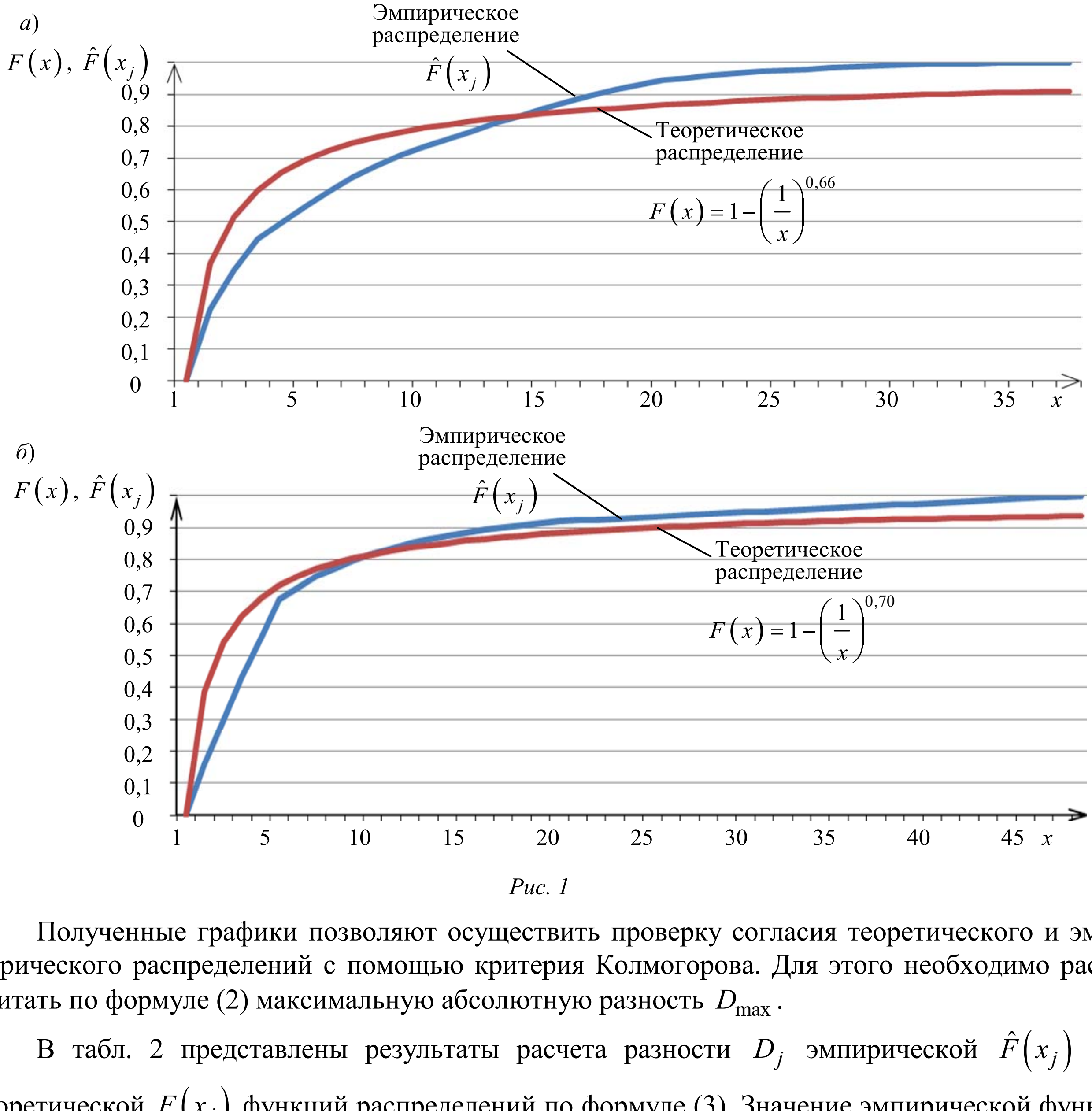

Рис. 1

Полученные графики позволяют осуществить проверку согласия теоретического и эмпирического распределений с помощью критерия Колмогорова. Для этого необходимо рассчитать по формуле (2) максимальную абсолютную разность $D_{\max}$.

В табл. 2 представлены результаты расчета разности D_j эмпирической $\hat{F}(x_j)$ и теоретической $F(x_j)$ функций распределений по формуле (3). Значение эмпирической функции — накопленные доли от суммарного числа повторений команд в 50 программах.

По табл. 2 легко определить максимальную разность $D_{\max} = 0,17$. Значение D_N рассчитывается по формуле (1) для количества испытаний $N = 50$, поскольку были рассмотрены 50 программ. Под испытанием понимается процесс реализации одной программы:

$$D_N = \sqrt{N} \cdot D_{\max} = \sqrt{50} \cdot 0,17 = 1,202.$$

Таблица 2

Случайная величина x	Значение функции $\hat{F}(x_j)$	Значение функции $F(x_j)$	Разность D_j	Случайная величина x	Значение функции $\hat{F}(x_j)$	Значение функции $F(x_j)$	Разность D_j
1	0	0	0	20	0,93	0,862	0,068
2	0,225	0,367	0,142	21	0,944	0,866	0,078
3	0,346	0,516	0,17	22	0,952	0,87	0,082
4	0,445	0,6	0,155	23	0,959	0,874	0,085
5	0,497	0,655	0,158	24	0,966	0,878	0,088
6	0,547	0,694	0,147	25	0,971	0,881	0,09
7	0,596	0,724	0,128	26	0,975	0,884	0,091
8	0,639	0,747	0,108	27	0,979	0,887	0,092
9	0,677	0,766	0,089	28	0,983	0,889	0,094
10	0,709	0,782	0,073	29	0,987	0,892	0,095
11	0,736	0,795	0,059	30	0,99	0,894	0,096
12	0,76	0,806	0,046	31	0,994	0,897	0,097
13	0,785	0,816	0,031	32	0,995	0,899	0,096
14	0,81	0,825	0,015	33	0,996	0,901	0,095
15	0,833	0,833	0	34	0,997	0,903	0,094
16	0,856	0,84	0,016	35	0,998	0,905	0,093
17	0,877	0,846	0,031	36	0,998	0,906	0,092
18	0,896	0,852	0,044	37	0,999	0,908	0,091
19	0,915	0,857	0,058	38	1	0,91	0,09

При решении многих практических задач для уровня значимости α обычно используют стандартные значения: $\alpha = 0{,}1$, $\alpha = 0{,}05$, $\alpha = 0{,}01$ [1, 8]. Из табл. 1 следует:

$$\left.\begin{aligned} D_N &< z_{1-\alpha} = 1{,}63 \quad \text{при} \quad \alpha = 0{,}01; \\ D_N &< z_{1-\alpha} = 1{,}36 \quad \text{при} \quad \alpha = 0{,}05; \\ D_N &< z_{1-\alpha} = 1{,}22 \quad \text{при} \quad \alpha = 0{,}1. \end{aligned}\right\} \tag{5}$$

Полученные результаты позволяют сделать заключение, что при всех трех рекомендованных уровнях значимости нет основания отвергнуть принятую гипотезу о виде функции распределения — распределении Парето.

При необходимости можно построить доверительную область для теоретического распределения — с этой целью при выбранном уровне значимости ($\alpha = 0{,}01$) вычисляют величину

$$D_N^* = \frac{z_{1-\alpha}}{\sqrt{N}} = \frac{1{,}63}{\sqrt{50}} = 0{,}231$$

и наносят на график доверительной области:

$$F_{\text{в}}(x) = F(x) + D_N^*, \; F_{\text{н}}(x) = F(x) - D_N^*,$$

где $F_{\text{в}}(x)$, $F_{\text{н}}(x)$ — верхняя и нижняя границы доверительной области.

Нанесенная на графике (рис. 2) эмпирическая функция распределения $\hat{F}(x_j)$ команд учебной ЭВМ не выходит за доверительные границы. Это позволяет сделать вывод, что проверка гипотезы о согласии теоретического и эмпирического распределений с помощью критерия Колмогорова прошла успешно.

Такой же подход был применен для микрокоманд с заданным числом испытаний $N = 42$ (были рассмотрены 42 команды). Результаты расчетов значений эмпирической

$\hat{F}(x_j)$ и теоретической $F(x_j)$ функций распределения и разности D_j представлены в табл. 3.

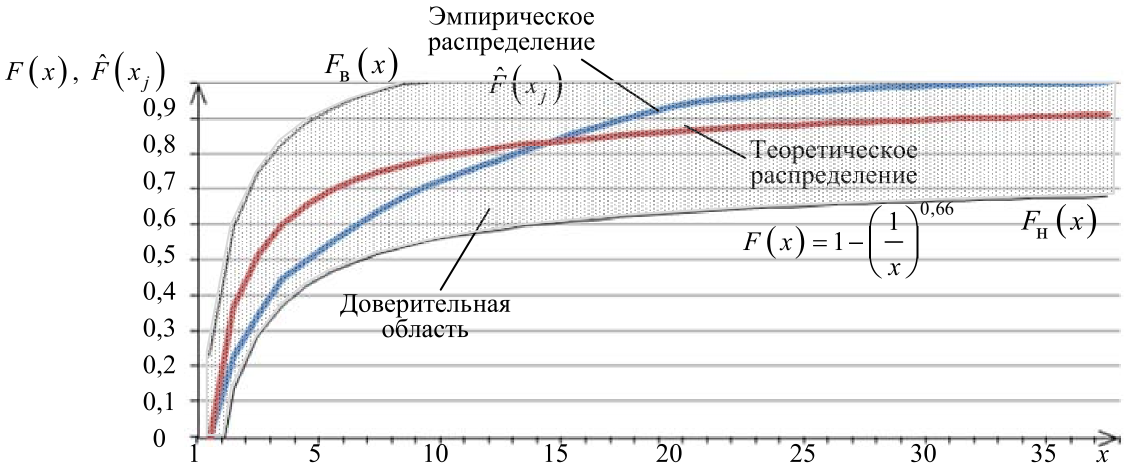

Рис. 2

Таблица 3

Случайная величина x	Значение функции $\hat{F}(x_j)$	Значение функции $F(x_j)$	Разность D_j	Случайная величина x	Значение функции $\hat{F}(x_j)$	Значение функции $F(x_j)$	Разность D_j
1	0	0	0	26	0,933	0,9	0,033
2	0,16	0,387	0,227	27	0,936	0,903	0,033
3	0,297	0,54	0,243	28	0,939	0,905	0,034
4	0,433	0,625	0,192	29	0,942	0,907	0,035
5	0,555	0,679	0,124	30	0,945	0,91	0,035
6	0,677	0,718	0,041	31	0,948	0,912	0,036
7	0,712	0,747	0,035	32	0,951	0,914	0,037
8	0,747	0,77	0,023	33	0,953	0,915	0,038
9	0,773	0,788	0,015	34	0,956	0,917	0,039
10	0,797	0,804	0,007	35	0,959	0,919	0,04
11	0,817	0,816	0,001	36	0,962	0,921	0,041
12	0,834	0,827	0,007	37	0,965	0,922	0,043
13	0,852	0,837	0,015	38	0,968	0,924	0,044
14	0,863	0,845	0,018	39	0,971	0,925	0,046
15	0,875	0,852	0,023	40	0,974	0,926	0,048
16	0,884	0,859	0,025	41	0,977	0,928	0,049
17	0,892	0,865	0,027	42	0,98	0,929	0,051
18	0,901	0,87	0,031	43	0,983	0,93	0,053
19	0,907	0,875	0,032	44	0,985	0,931	0,054
20	0,913	0,88	0,033	45	0,988	0,932	0,056
21	0,919	0,884	0,035	46	0,991	0,933	0,058
22	0,922	0,887	0,035	47	0,994	0,934	0,06
23	0,924	0,891	0,033	48	0,997	0,935	0,062
24	0,927	0,894	0,033	49	1	0,936	0,064
25	0,93	0,897	0,033	—	—	—	—

На основе табл. 3 по формулам (1), (2) получаем

$$D_{\max} = 0,243\,, \quad D_N = \sqrt{N} \cdot D_{\max} = \sqrt{42} \cdot 0,243 = 1,575.$$

Сравнение полученного значения D_N с числом $z_{1-\alpha}$ при $\alpha = 0,1$, $\alpha = 0,05$, $\alpha = 0,01$ — см. выражения (5) — позволяет сделать вывод, что при $\alpha = 0,1$ и $\alpha = 0,05$ эмпирическое распределение микрокоманд не согласуется с теоретическим распределением Парето, а при $\alpha = 0,01$ гипотеза о согласии принимается.

Для $\alpha = 0{,}01$ доверительные границы для теоретического распределения представлены на рис. 3. Нанесенная на графике эмпирическая функция $\hat{F}\left(x_j\right)$ распределения микрокоманд учебной ЭВМ не выходит за доверительные границы. Таким образом, гипотеза о согласии теоретического и эмпирического распределений Парето принимается.

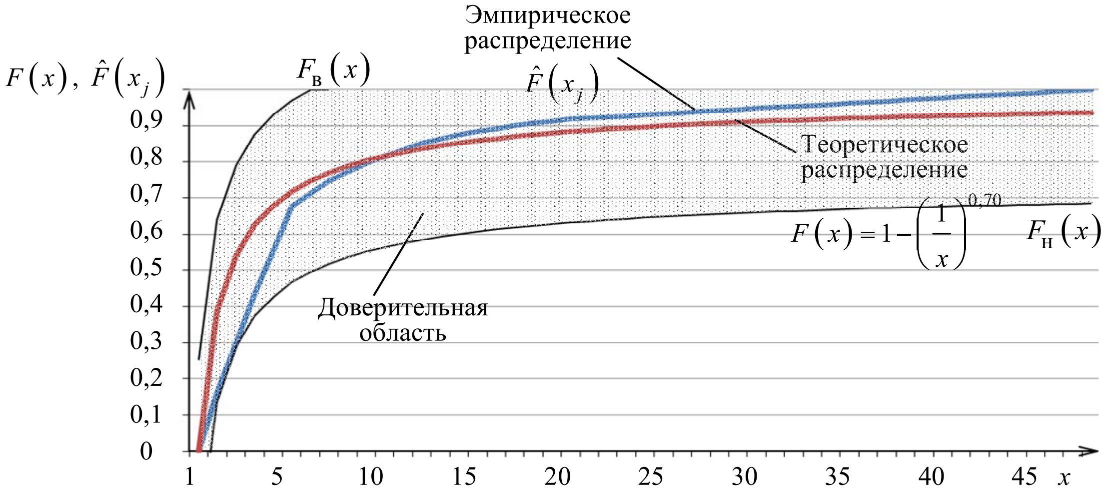

Рис. 3

Итак, вышеизложенное позволяет сделать следующие обобщающие выводы.

1. Гипотетическая статистическая модель в виде функции распределения Парето соответствует реальным практическим данным, полученным для количественного оценивания частоты использования машинных команд ЭВМ и реализующих их микрокоманд.

2. Кумулятивные кривые математически характеризуются функциями распределения Парето с параметрами $x_m = 1$, $k = 0{,}66$ для команд и $x_m = 1$, $k = 0{,}70$ для микрокоманд. При этом случайными величинами являются номера команд и микрокоманд, представляющие собой члены ряда натуральных чисел.

3. Анализ графиков теоретических и практических функций распределения позволяет осуществить выбор команд и микрокоманд, редко используемых при работе процессора. Исключение из системы команд редко используемых машинных инструкций позволяет упростить состав и структуру устройств управления как аппаратного, так и микропрограммного типа процессоров ЭВМ.

Рассмотренные в статье отдельные теоретические положения и полученные практические результаты являются развитием представленного в работах [5, 8] статистического метода повышения качества — анализа Парето применительно к количественному оцениванию метрик машинных команд и микрокоманд ЭВМ.

СПИСОК ЛИТЕРАТУРЫ

1. *Шуленин В. П.* Математическая статистика. Ч. 1. Параметрическая статистика: Учебник. Томск: Изд-во НТЛ, 2012. 540 с.

2. *Голяков А. Д., Миронов В. И., Смирнов В. В.* Испытания систем ракетно-космической техники. СПб: ВИКИ им. А. Ф. Можайского, 1992. 398 с.

3. *Басыров А. Г.* Организация ЭВМ и систем: Практикум. СПб: ВКА им. А. Ф. Можайского, 2012. 83 с.

4. *Жмакин А. П.* Архитектура ЭВМ. СПб: БХВ-Петербург, 2010. 352 с.

5. *Аверьянов А. В., Кошель И. Н., Кузнецов В. В., Нгуен В. Т.* Статистическое оценивание метрик машинных команд ЭВМ и реализующих их микрокоманд на основе анализа Парето // Изв. вузов. Приборостроение. 2023. Т. 66, № 4. С. 259—265.

6. *Brunton S. L., Kutz J. N.* Data-driven science and engineering : machine learning, dynamical systems, and control. Cambridge Univ. Press, 2019. 472 p.
7. *Жадан В. Г.* Методы оптимизации. Ч. II. Численные алгоритмы: Учеб. пособие. М.: МФТИ, 2015. 320 с.
8. *Аверьянов А. В., Калюжный А. В.* Применение анализа Парето для количественного оценивания частоты использования машинных команд ЭВМ // Изв. вузов. Приборостроение. 2019. Т. 62, № 2. С. 101—105.

Сведения об авторах

Алексей Васильевич Аверьянов — канд. техн. наук, доцент; ВКА им. А. Ф. Можайского, кафедра информационно-вычислительных систем и сетей; преподаватель; E-mail: vka_24kaf1@mail.ru

Ван Тиен Нгуен — ВКА им. А. Ф. Можайского, кафедра информационно-вычислительных систем и сетей; курсант; E-mail: vka_24kaf1@mail.ru

Поступила в редакцию 02.06.2023; одобрена после рецензирования 09.06.2023; принята к публикации 27.09.2023.

REFERENCES

1. Shulenin V.P. *Matematicheskaya statistika. Chast' 1. Parametricheskaya statistika* (Math Statistics. Part 1. Parametric Statistics), Tomsk, 2012, 540 p. (in Russ.)
2. Golyakov A.D., Mironov V.I., Smirnov V.V. *Ispytaniya sistem raketno-kosmicheskoy tekhniki* (Testing of Rocket and Space Technology Systems), St. Petersburg, 1992, 398 p. (in Russ.)
3. Basyrov A.G. *Organizatsiya EVM i sistem: praktikum* (Organization of Computers and Systems: Workshop), St. Petersburg, 2012, 83 p. (in Russ.)
4. Zhmakin A.P. *Arkhitektura EVM* (Computer Architecture), St. Petersburg, 2010, 352 p. (in Russ.)
5. Averyanov A.V., Koshel I.N., Kuznetsov V.V., Nguyen V.T. *Journal of Instrument Engineering*, 2023, no. 4(66), pp. 259–265. (in Russ.)
6. Brunton S.L., Kutz J.N. *Data-driven science and engineering: machine learning, dynamical systems, and control*, Cambridge, Cambridge University Press, 2019, 472 p.
7. Zhadan V.G. *Metody optimizatsii. Chast' 2. Chislennyye algoritmy* (Optimization Methods. Part 2. Numerical Algorithms), Moscow, 2015, 320 p. (in Russ.)
8. Averyanov A.V., Kalyuzhny A.V. *Journal of Instrument Engineering*, 2019, no. 2(62), pp. 101–105. (in Russ.)

Data on authors

Alexey V. Averyanov — PhD, Associate Professor; A. F. Mozhaisky Military Space Academy, Department of Information Systems and Networks; Lecturer; E-mail: vka_24kaf1@mail.ru

Van Tien Nguyen — A. F. Mozhaisky Military Space Academy, Department of Information Systems and Networks; Student; E-mail: vka_24kaf1@mail.ru

Received 02.06.2023; approved after reviewing 09.06.2023; accepted for publication 27.09.2023.