ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ

УДК 519.23

В. В. Мартынов, П. В. Мартынов

МЕТОД ОБРАБОТКИ И АНАЛИЗА ВЫБОРОЧНЫХ ДАННЫХ

Аннотация. Представлены результаты разработки метода установления закона распределения выборочных данных.

Ключевые слова: выборка, закон распределения, метод, энтропия, количество информации, коэффициент избыточности.

Abstract. The article introduces the results of the development of a method of establishing the law of distribution of sample data.

Key words: sample, distribution law, method, entropy, amount of information, redundancy factor.

Введение

Одной из задач, часто возникающих в практической деятельности, является обработка экспериментальных данных, представляющих собой часть членов некоторой достаточно большой совокупности (называемой генеральной), отобранных из нее для получения сведений обо всей совокупности. Обработкой этих данных занимается раздел математической статистики, называемый теорией выборок.

Основная цель выравнивания выборочных данных состоит в установлении закона их распределения. Это позволяет решать разнообразные практические задачи, в том числе прогнозировать вероятность появления различных событий. Существующие методы [1–5] не всегда позволяют установить вид закона распределения однозначно. Это связано в том числе и с тем, что распределения имеют близкие статистические свойства, в частности, широко используемые в различных приложениях гамма-распределение, логарифмически нормальное и Вейбулла. Кроме этого, при определенных значениях параметра, характеризующего форму, гамма-распределение и распределение Вейбулла приближаются к нормальному распределению. В связи с этим актуальной является разработка метода, в основе которого лежат критерии, учитывающие специфические свойства распределений.

Теоретическая основа

Пусть по выборке x, содержащей информацию о законе распределения ее данных $x_1, \ldots, x_n$, требуется отдать предпочтение одной из конкурирующих гипотез G_j, $j = 1, \ldots, k$, если известны распределения данных для каждой из них, т.е. $p(x/G_j)$. Если согласиться с утверждением о том, что никакая обра-

ботка не может увеличить количество содержащейся в выборке информации, то тогда наилучшим образом отдать предпочтение можно, вычислив количество информации для каждой из гипотез, т.е. поставив в соответствие каждому распределению некоторое число. Последующее сравнение чисел между собой либо с каким-то эталоном позволит выбрать распределение, воспроизводящее максимум исходной информации или (что то же самое) обеспечивающее ее минимальные потери.

Из теории информации известно, что основной информационной числовой характеристикой случайной величины является энтропия (H), которую можно трактовать и как меру рассеяния случайной величины; в этом смысле она подобна дисперсии. Но если дисперсия является адекватной мерой рассеяния лишь для специальных распределений вероятностей случайных величин, то энтропия не зависит от типа распределения, поэтому может использоваться и в качестве его количественной характеристики.

Меру уменьшения энтропии (в нашем случае выборки x) определяет количество информации (I). Из этого следует, что количество информации и энтропия характеризуют одну и ту же ситуацию, но с качественно противоположенных сторон: величина I – это количество информации, которое требуется для снятия неопределенности H. Если неопределенность снимается полностью, то количество полученной информации равно изначально существовавшей неопределенности. При частичном снятии неопределенности ее исходное значение представляет собой сумму полученного количества информации I_t и оставшейся неснятой неопределенности H_t, т.е. потерь информации

$$H = I_t + H_t. \tag{1}$$

Таким образом, задача выбора закона распределения сводится к вычислению величины $I_{t,\,j}$ для каждой из гипотез, поэтому требует определенного представления данных выборки x. В теории информации для этого обычно выполняется процедура их квантования по уровню и по времени с последующим вычислением вероятностей $p(x_i)$ появления любого из m уровней при снятии любого из n отсчетов при условии, что одни и те же уровни могут появляться неоднократно, т.е. a_i раз. Тогда вероятности появления квантованных уровней определятся как

$$p\left(x_i\right) \approx \frac{a_i}{n} \;\; i = 1, ..., m\,, \tag{2}$$

а выражение для I_t при условиях, что $\sum_{i=1}^{m} p(x_i) = 1$ и отсчеты являются независимыми, будет иметь следующий вид [6]:

$$I_t\left(x\right) = -n\sum_{i=1}^{m} p(x_i) \times \log p(x_i)\,. \tag{3}$$

Поскольку все $p(x_i) \neq 0$, функция I_t всегда определена. В тех же случаях, когда необходимо включить в рассмотрение значения $p(x_i) = 0$, принимается, что соответствующее произведение $p(x_i) \times \log p(x_i) = 0$.

Основание логарифма в (2) выбирается произвольно. В теории информации, например, для этого используется число 2.

Рассматривая изложенное применительно к решаемой задаче, заметим, что процедура квантования практически полностью совпадет с применяемой в статистике процедурой построения эмпирического распределения данных на основе подсчета частот их попадания в интервалы, число которых зависит от объема выборки, а цена одного интервала определяется отношением диапазона варьирования данных к числу интервалов. Значения частот попадания данных в каждый интервал и есть эмпирическая кривая распределения выборки x.

Сопоставление процедур показывает, что частоты попадания данных выборки в интервалы есть не что иное, как величина a_i, число интервалов – число уровней их квантования m, цена интервала – шаг квантования по уровню Δx, а количество отсчетов – объем выборки n. Это делает корректным представление эмпирического распределения данных в терминах теории информации (рис. 1) и позволяет использовать выражение (3) для определения количества содержащейся в них информации.

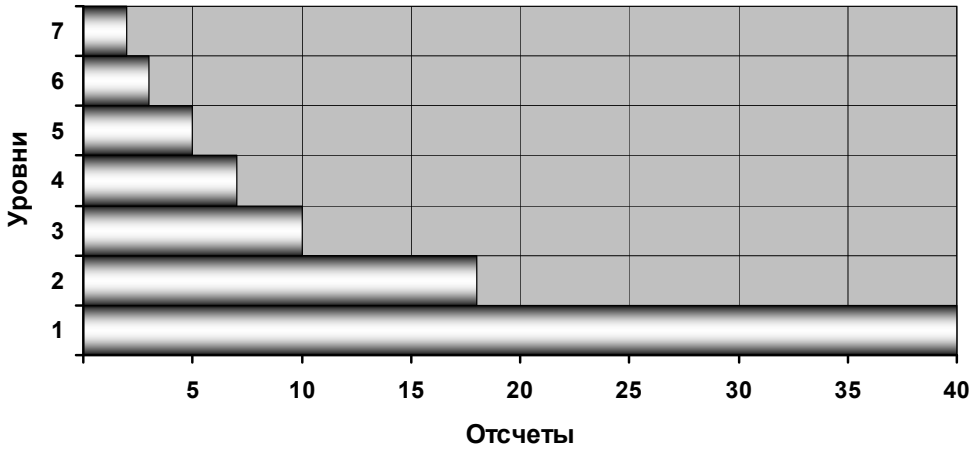


Рис. 1. Информационное представление эмпирического распределения выборочных данных

Для определения значений I_t данных, относящихся к каждой из гипотез G_j, необходимо определить значения вероятностей $p_j(x_i)$. Это можно сделать, используя значения плотностей вероятностей $f_j(x_i)$ каждого из распределений, вычисленные для среднего значения $x_{c,\,i}$ каждого из m уровней, т.е.

$$p_j(x_{c,\,i}) = f_i(x_{c,\,i}) \times \Delta x. \tag{4}$$

Если затем сопоставить значения $I_{t,\,j}$ со значением $I_t(x)$, являющимся максимальным в силу того, максимальным количеством информации обладает сама выборка x, то на основании фундаментальной леммы Неймана – Пирсона [7] распределение, для которого отклонение $I_{t,\,j}$ от $I_t(x)$, называемое коэффициентом избыточности:

$$k_{t,\,j} = \frac{I_t(x) - I_{t,\,j}}{I_t(x)}, \tag{5}$$

окажется минимальным и будет искомым как наиболее правдоподобное, имеющее наибольшую «плотность информации», т.е. количество информации, приходящееся на один отсчет, равно

$$i_{t,j} = \frac{I_t}{n} = -\sum_{i=1}^{m} p(x_i) \times \log p(x_i), \qquad (6)$$

и потому снимающее неопределенность о содержащихся в выборке данных в наибольшей степени. В связи с этим назовем разработанный метод методом информационного критерия.

Практическая реализация

Практическую реализацию метода вначале рассмотрим на данных, имеющих наиболее часто используемые в различных технических приложениях распределения: логарифмически нормальное, гамма, Вейбулла, экспоненциальное. Данные получены методом генерации на ЭВМ и с целью внесения искажений, имитирующих помехи, округлены до целочисленных значений.

Типичные примеры полученных результатов представлены на рис. 2–5 и убедительно свидетельствуют о том, что применение метода, даже при наличии помех, позволяет найти именно то распределение, которое было использовано для получения исходных данных.

Вместе с тем анализ результатов позволил установить, что эффективность метода в значительной степени зависит от того, насколько вычисленные по выборочным данным оценки параметров распределений соответствуют их истинным значениям, т.е. параметрам распределения генеральной совокупности, или, говоря иначе, в какой степени они обладают свойствами состоятельности, несмещенности и эффективности. Широко применяемый на практике метод моментов, основанный на представлении параметров распределений через моменты низких порядков (обычно первых трех), замене их оценками моментов, найденных с использованием выборочных данных, и решении полученных уравнений, дает приемлемые результаты только при значительном объеме выборочных данных. В противном случае погрешность оценок параметров будет большой, а иногда оценки могут оказаться вне допустимой области. Кроме этого, оценки, найденные по методу моментов, не всегда извлекают из данных всю имеющуюся в них информацию, поэтому более целесообразным является использование метода максимального правдоподобия, поскольку известно, что если несмещенные эффективные оценки параметров существуют, то уравнения правдоподобия имеют единственное решение и позволяют найти оценки параметров, которые имеют бóльшую вероятность оказаться ближе к их истинным значениям и делают данные, содержащихся в выборке, более близкими к реальным.

Заметим, что для некоторых распределений, например нормального и экспоненциального, оценки параметров, найденные с помощью метода моментов, совпадают с соответствующими оценками максимального правдоподобия. В связи с этим оценки, полученные методом моментов, могут быть использованы как первое приближение для метода максимального правдоподобия.

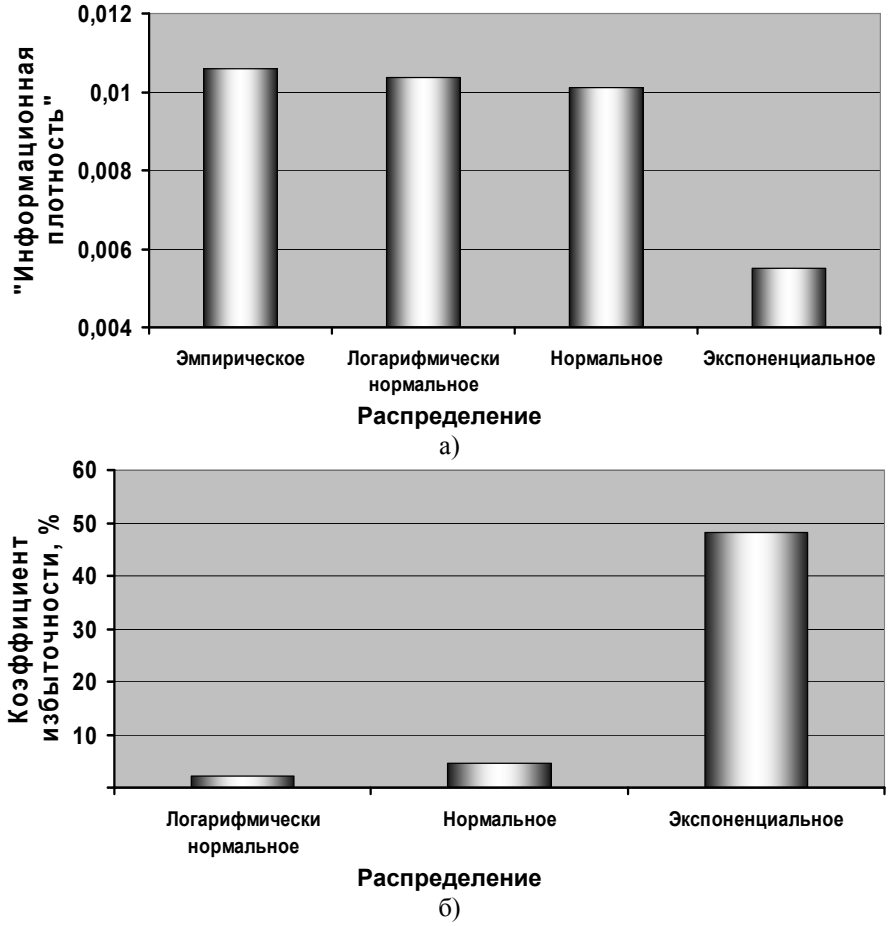
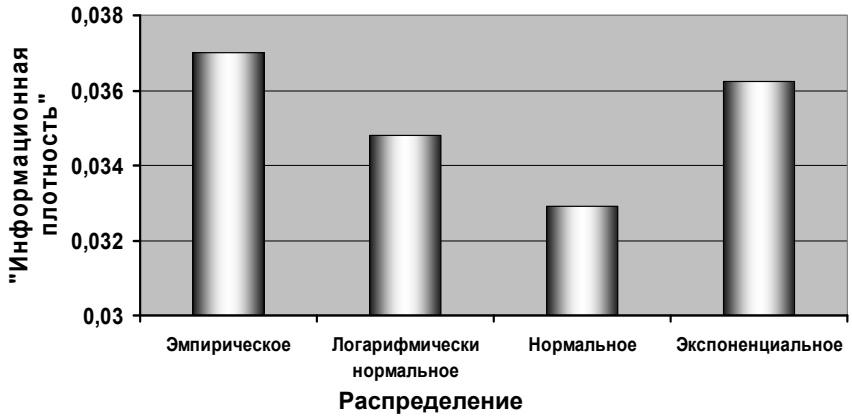


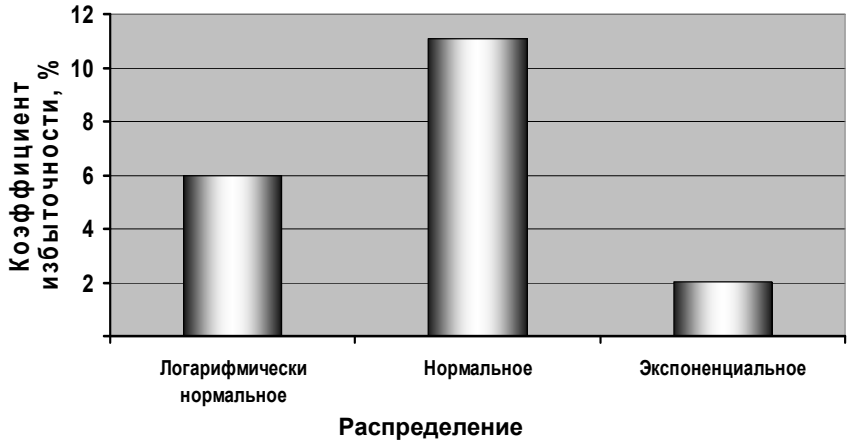
Рис. 2. Пример результатов применения метода к обработке данных, имеющих логарифмически нормальное распределение

Рассмотрим применение метода к обработке данных, полученных в ОАО «Тантал» г. Саратова по результатам проведения статистических исследований надежности автоматических токарных станков ТПАРМ-100 с целью оценки значимости изменения их состояния при эксплуатации в условиях автоматизированного многономенклатурного производства. Считается, что, поскольку такое оборудование относится к классу сложных управляемых объектов, объединяющих большое число различных по физической природе элементов, каждый из которых в отдельности не оказывает большого влияния на вероятность возникновения отказов объекта в целом, то статистической моделью времени его безотказной работы должно быть экспоненциальное распределение [1, 4, 8]. Если гипотеза окажется достоверной, то это означает, что отказы являются следствием не значимых изменений состояния объекта, а лишь неблагоприятного сочетания внешних и внутренних факторов, т.е. носят чисто случайный (внезапный) характер и не влияют на его эксплуатационные свойства.

Результаты обработки представлены на рис. 6 и позволяют констатировать следующее.

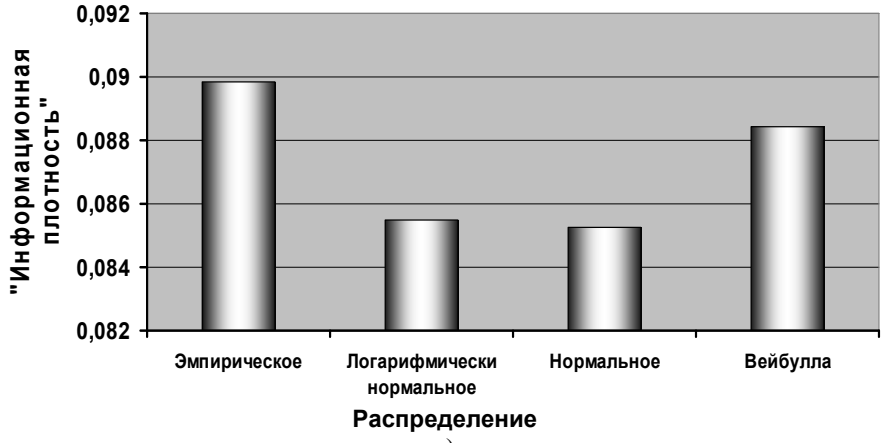


а)



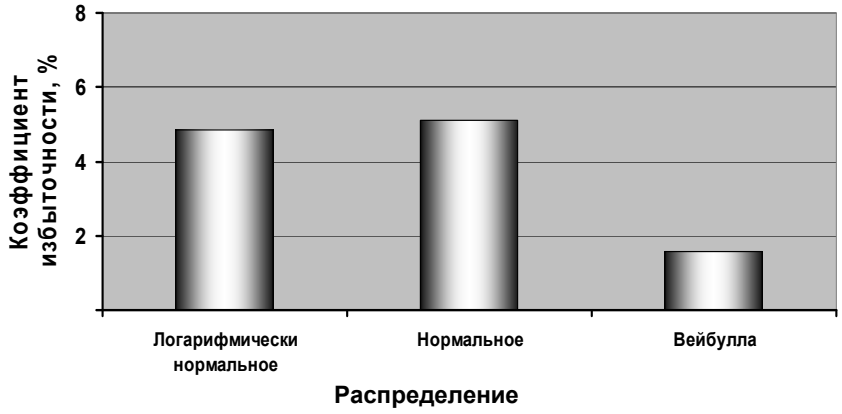
б)

Рис. 3. Пример результатов применения метода к обработке данных, имеющих экспоненциальное распределение



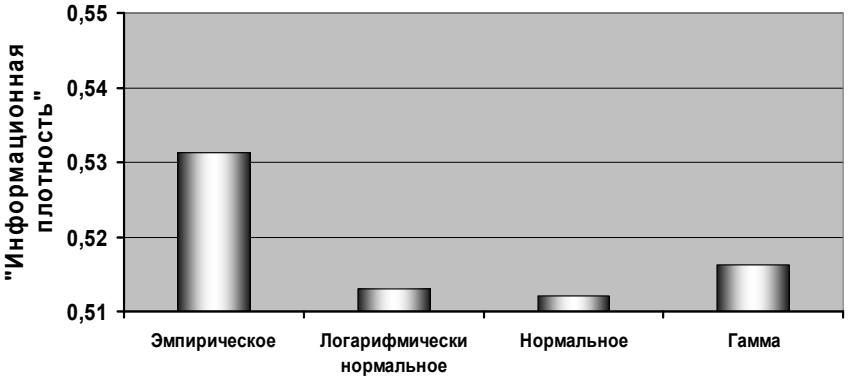
а)

Рис. 4. Пример результатов применения метода к обработке данных, имеющих распределение Вейбулла

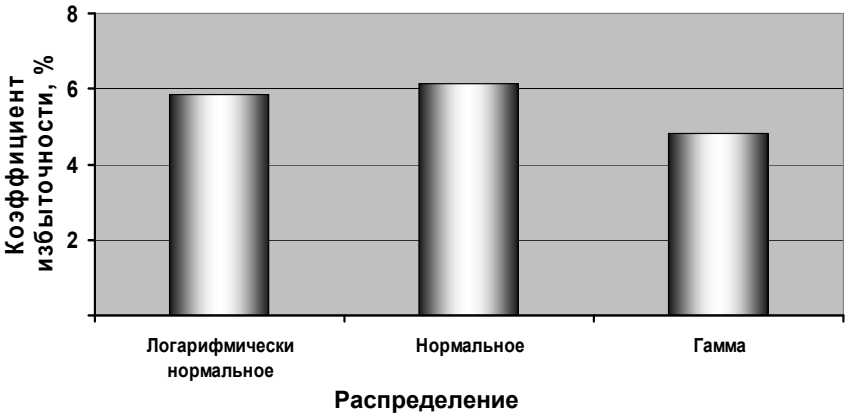


б)

Рис. 4. Окончание



а)



б)

Рис. 5. Пример результатов применения метода к обработке данных, имеющих гамма-распределение

1. Статистической моделью данных во всех случаях является распределение Вейбулла с параметром, характеризующим форму, меньше 1. Это означает, что интенсивность отказов является монотонно убывающей и объясня-

ется присутствием в данных отказов, возникновение которых связано, во-первых, с неоптимальностью алгоритмов анализа системой управления информации от датчиков обратной связи при управлении рабочими органами станков; во-вторых, с несовершенством конструкции подсистемы подачи технологической жидкости в зону резания. Наработки станков на эти отказы имели небольшие значения, а их количество в общей совокупности в большинстве случаев было преобладающим, что и привело к распределению, в котором наибольшая частота попаданий пришлась на первый интервал (рис. 7).

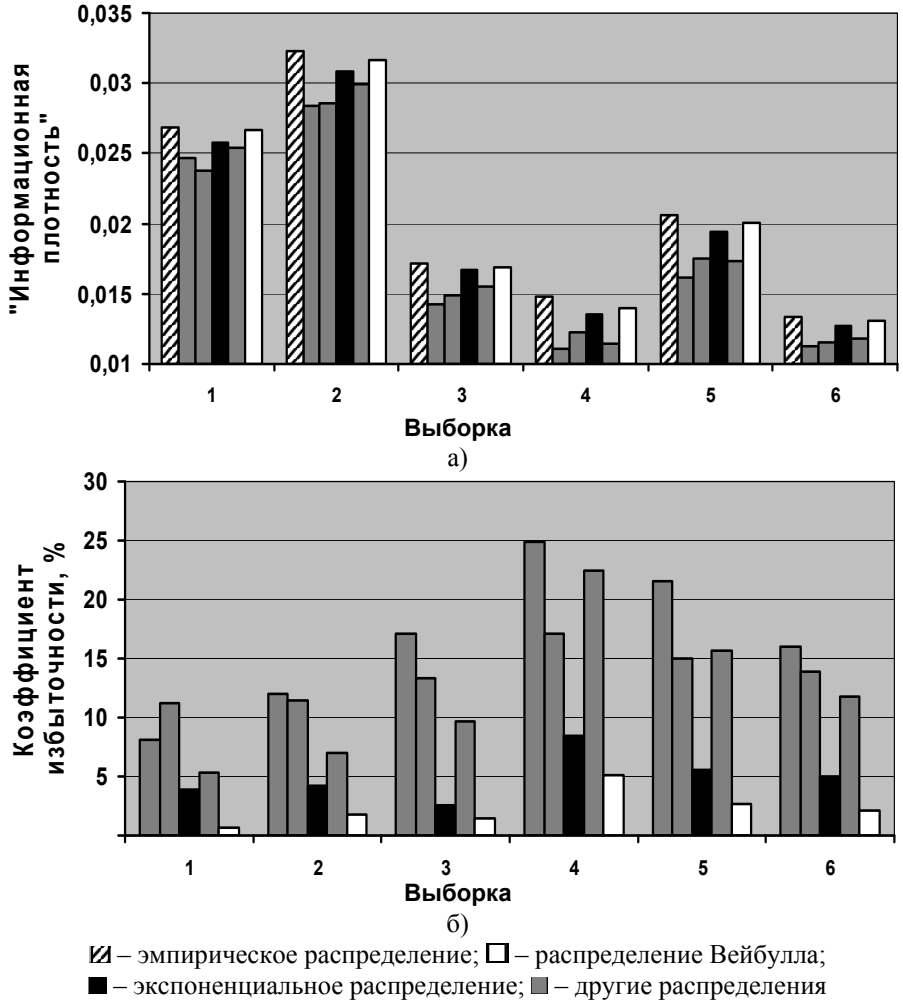


Рис.6. Результаты применения метода к данным об отказах станков ТПРАМ-100

2. Экспоненциальное распределение обеспечивает представление данных с погрешностью, в среднем всего на 2,7 % большей, чем распределение Вейбулла. С учетом изложенного в п. 1, а также того, что распределения, которые в теории надежности используются в качестве статистических моделей времени безотказной работы в условиях возникновения постепенных отказов, вызванных процессами износа и старения (логарифмически нормальное и нормальное), имели наименьшую «информационную ценность», использова-

ние экспоненциального распределения в качестве статистической модели времени безотказной работы сложных управляемых объектов является вполне корректным. Причем это касается не только решения простых задач, связанных с вычислением показателей надежности и эффективности их использования в составе станочных комплексов, но и задач управления самими комплексами на организационно-технологическом уровне (оперативно-производственное планирование) и уровне оперативно-диспетчерского регулирования хода производственного процесса (выполнение плановых заданий).

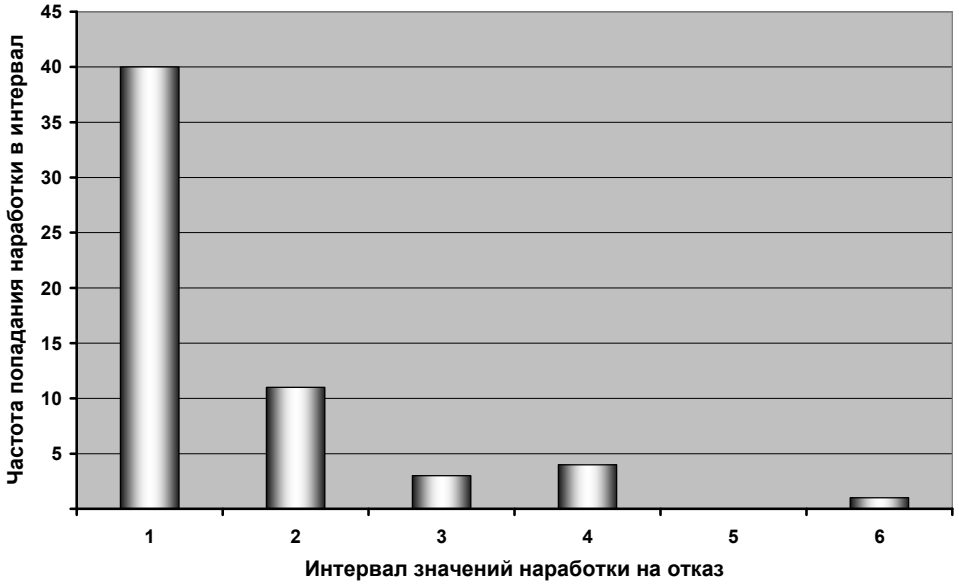


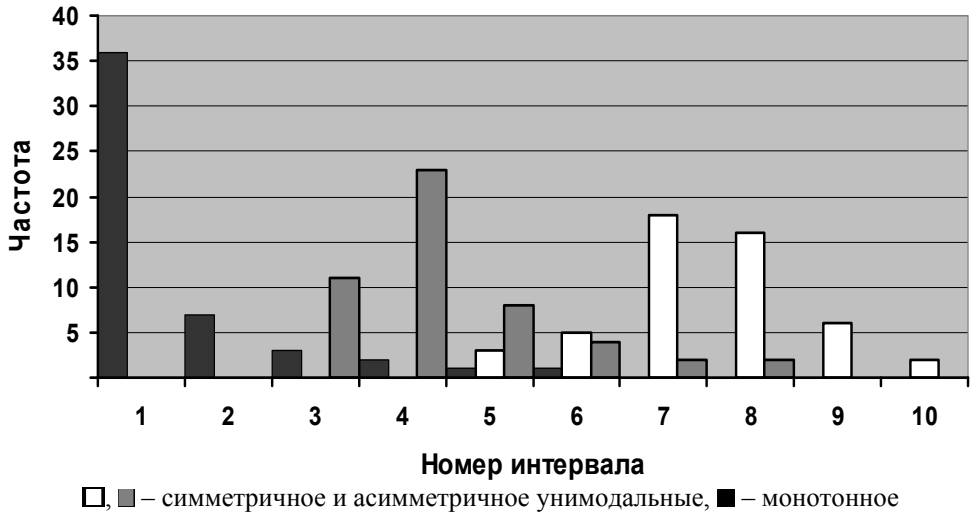
Рис. 7. Типичная гистограмма распределения отказов станков ТПРАМ-100

Заключение

Материалы выполненных исследований позволяют сделать вывод о том, какие специфические свойства распределений позволяют учесть метод информационного критерия. Эти свойства связаны с адекватным количественным отображением степени отклонения (т.е. деформации) кривой эмпирического распределения выборочных данных от идеального распределения абсолютно случайных величин, которым в соответствии с центральной предельной теоремой математической статистики является нормальное распределение, т.е. от симметричного унимодального распределения к распределению монотонно убывающего вида (рис. 8) [4, 9]. Это означает, что метод позволяет дать объективную количественную оценку способности распределений отображать содержащуюся в выборке информацию о физическом механизме, лежащем в основе процесса деформации. Прежде всего это касается распределений, имеющих близкие статистические свойства: логарифмически нормального, гамма и Вейбулла (рис. 9).

Логарифмически нормальное распределение применяется для отображения процессов, в которых наблюдаемое значение случайной величины со-

ставляет случайную долю ее предыдущего значения и является результатом умножения большого числа небольших ошибок.



□, ■ – симметричное и асимметричное унимодальные, ■ – монотонное

Рис. 8. Специфика распределений, отображаемая методом информационного критерия

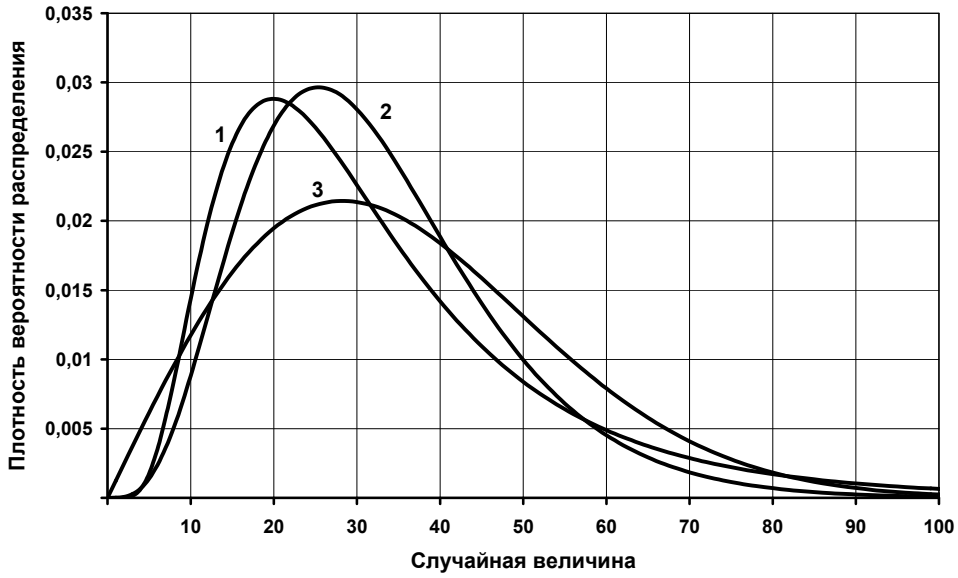


Рис. 9. Плотности логарифмически нормального распределения (1), гамма-распределения (2) и распределения Вейбулла (3)

Гамма-распределение хорошо отображает процессы, в которых появление случайной величины есть результат n независимых событий, происходящих с постоянной интенсивностью.

Распределение Вейбулла является предельной моделью отображения распределения минимальных значений n случайных величин (при $n \to \infty$), имеющих различные исходные распределения, ограниченные слева.

Изложенное означает, что степень деформации у каждого из распределений будет различной и соответственно при выравнивании выборочных данных каждым из них будет различной степень искажения содержащейся в данных информации. Из этого следует, что имеется возможность дальнейшего развития предложенного метода в направлении поиска показателя, отображающего степень искажения содержащейся в выборке информации количественно. Найденный показатель может быть положен в основу конструирования критерия, позволяющего проводить проверку принадлежности выборочных данных к распределению конкретного вида.

Список литературы

1. **Баублис, А. Б.** Статистические модели в АСУ машиностроительного предприятия / А. Б. Баублис. – М. : Машиностроение, 1984. – 245 с.
2. **Солонин, И. С.** Математическая статистика в технологии машиностроения / И. С. Солонин. – М. : Машиностроение, 1972. – 208 с.
3. **Кокс, Д. Г.** Анализ данных типа времени жизни / Д. Г. Кокс, Д. Оукс ; пер. с англ. – М. : Финансы и статистика, 1988. – 191 с.
4. **Хан, Г.** Статистические модели в инженерных задачах / Г. Хан, С. Шапиро; пер. с англ. – М. : Мир, 1968. – 396 с.
5. **Герцбах, И. Б.** Модели отказов / И. Б. Герцбах, Х. Б. Кордонский ; под ред. Б. В. Гнеденко. – М. : Советское радио, 1966. – 166 с.
6. Вероятностные методы в вычислительной технике / А. В. Крайников, Б. А. Курдиков, А. Н. Лебедев и др. ; под ред. А. Н. Лебедева и Е. А. Чернявского. – М. : Высш. шк., 1986. – 312 с.
7. **Ван Трис, Г.** Теория обнаружения, оценок и модуляции / Г. Ван Трис. – М. : Советское радио, 1972. – Т. 1. – 744 с.
8. **Хазов, Б.Ф.** Справочник по расчету надежности машин на стадии проектирования / Б. Ф. Хазов, Б. А. Дидусев. – М. : Машиностроение, 1986. – 224 с.
9. **Кармадонов, А. Ф.** Деформация кривых плотности распределения наработки изделий в зависимости от режимов загружения и качества изготовления / А. Ф. Кармадонов, В. М. Ляховецкий // Надежность и контроль качества. – 1976. – № 1. – С. 69–75.

Мартынов Владимир Васильевич
доктор технических наук, профессор, кафедра конструирования и компьютерного моделирования технологического оборудования в машино- и приборостроении, Саратовский государственный технический университет имени Ю. А. Гагарина

E-mail: v_martynov@mail.ru

Martynov Vladimir Vasilyevich
Doctor of engineering sciences, professor, sub-department of processing equipment design and computer simulation in mechanical and instrument engineering, Saratov State Technical University named after Y. A. Gagarin

Мартынов Павел Владимирович
аспирант, Саратовский государственный технический университет имени Ю. А. Гагарина

E-mail: mpv1990@mail.ru

Martynov Pavel Vladimirovich
Postgraduate student, Saratov State Technical University named after Y. A. Gagarin

УДК 519.23
Мартынов, В. В.
Метод обработки и анализа выборочных данных / В. В. Мартынов, П. В. Мартынов // Известия высших учебных заведений. Поволжский регион. Технические науки. – 2012. – № 3 (23). – С. 3–14.