

Параметрический метод обработки результатов наблюдений с учетом пропущенных данных

Дмитрий А. Николаев, Обнинский институт атомной энергетики НИЯУ МИФИ, Обнинск, Россия,
e-mail: dafanday@gmail.com



Дмитрий А.
Николаев

Резюме. Вопросы обеспечения надёжного и безопасного функционирования объектов АЭС являются актуальными. Это связано с тем, что доля оборудования, находящегося на грани истощения назначенного ресурса, в атомной энергетике очень велика, поэтому необходимо проводить анализ надёжности элементов и систем АЭС. В процессе анализа характеристик надёжности приходится сталкиваться с различными проблемами, например, оценкой остаточного ресурса оборудования, обоснованием решения о продлении его срока службы. Также необходимо решать задачи обеспечения элементов и систем запасными частями и изделиями, выбора стратегий обслуживания оборудования и т.п. В связи с этим возрастает ценность работ по анализу надёжности объектов ядерной энергетики и, как следствие, необходимость разработки методов анализа статистической информации о функционировании элементов, подсистем и систем АЭС с целью определения параметров их работоспособности. На атомных станциях организуются работы по сбору информации о функционировании различных объектов: об отказах и дефектах составных частей системы, порядке их обслуживания, режимах работы, условиях хранения и т.п. Предоставляемая информация с АЭС имеет ряд особенностей. Это связано с несколькими основными моментами: наличием цензурирования в данных об отказах, отсутствием наработок на заданном интервале наблюдения и ограниченным объёмом этих данных. Все эти факторы приводят к наличию неопределённости в получаемых оценках и, как следствие, к невысокой точности расчетов характеристик надёжности. В процессе проведения анализа надёжности эксплуатируемых объектов сталкиваются с ситуациями, когда определенная часть объектов или систем не отказала за период наблюдения. В таких ситуациях возникает необходимость проведения статистического анализа надёжности на основе так называемых цензурированных выборок справа, основной особенностью которых является тот факт, что контролируемое изделие не вышло из строя за время наблюдения.

Также имеют место случаи, когда неизвестны наработки конкретных объектов. Например, в начале эксплуатации объекта информация о его работе не собиралась, в дальнейшем было принято решение осуществлять сбор данных. В этом случае необходимо разработать такой подход, чтобы учесть недостающую информацию, которая не фиксировалась на первоначальном этапе. Ограниченность объёма данных объясняется тем, что объекты ядерных энергетических установок относятся к категории высоконадёжного оборудования. Отказы их — события редкие. Поэтому для повышения достоверности оценок показателей надёжности необходимо использовать всю имеющуюся информацию. Таким образом, учёт всей имеющейся информации приводит к более точным результатам, на основе которых можно рассчитать срок службы объекта АЭС. Целью работы является демонстрация применения метода повторной выборки и исследование его эффективности. Основное внимание в работе уделяется пропущенным данным, подлежащим восстановлению. Приводятся результаты оценки параметра экспоненциального закона распределения с учетом цензурированных справа и пропущенных данных. Производится сравнение предложенного метода повторной выборки с бутстреп методом и методом замены средним арифметическим. В качестве подхода к оценке параметра экспоненциального закона распределения предложено использовать метод максимального правдоподобия. Производится расчет статистических характеристик. Все расчеты и полученные результаты выполнены на основе тестовых примеров.

Ключевые слова: метод однократного замещения, метод повторной выборки, бутстреп метод, метод максимального правдоподобия, цензурированные данные, пропущенные данные, восстановление данных.

Формат цитирования: Николаев Д.А. Параметрический метод обработки результатов наблюдений с учетом пропущенных данных // Надежность. 2017. Т. 17, № 1. С. 53-58. DOI: 10.21683/1729-2646-2017-17-1-53-58

Введение

Вопросы обеспечения надёжного и безопасного функционирования объектов АЭС являются актуальными. Это связано с тем, что доля оборудования, находящегося на грани исчерпания назначенного ресурса, в атомной энергетике очень велика, поэтому необходимо проводить анализ надёжности элементов и систем АЭС. Для этого рассматривается современное состояние проблемы статистического анализа информации, содержащей наработки до отказа, наработки до цензурирования и пропущенные данные. Учёт всей имеющейся информации приводит к более точным результатам.

Таким образом, задача состоит в разработке параметрического подхода к восстановлению плотности распределения наработок на основании наработок до отказа, наработок до цензурирования и пропущенных данных, что позволит определить характер поведения объекта. Следовательно, необходимо оценить параметр восстановленной плотности распределения и определить статистические показатели надёжности.

Описание данных

Стоит отметить, что предметом исследования являются восстанавливаемые объекты, работоспособность которых в случае возникновения отказа подлежит восстановлению. Перед тем как непосредственно перейти к решению поставленной задачи, определим виды данных, поступающих на обработку. Первый и основной тип данных – это наработки до отказа.

В практических ситуациях, в особенности при проведении обследования функционирующих объектов, информация, поступающая на обработку, крайне ограничена. В таких ситуациях возникает необходимость проведения статистического анализа надёжности на основе специфических выборок, основной особенностью которых является отсутствие сведений о наработках контролируемого объекта. К таким данным относятся цензурированные данные. Цензурирование – это событие, приводящее к прекращению наблюдений за изделием до наступления системного события, либо к свершению события в неизвестный момент времени в пределах некоторого интервала. Нас интересует случай, когда этот интервал может быть неограничен справа, т.е. выборка цензурирована справа [1, 2, 5]. Следующий тип данных – пропущенные данные.

Пропущенные данные очень часто встречаются при анализе и могут оказать существенное влияние на выводы, которые можно получить на основании этих данных.

Возможные ситуации реализации наблюдаемой информации приведены на рисунке 1.

Этот график показывает, что происходит с объектами за время их функционирования, данные изображаются как непрерывный поток данных для конкретного объекта, где знаком $\times$ обозначается отказ, а знаком $\triangleright$ – цен-

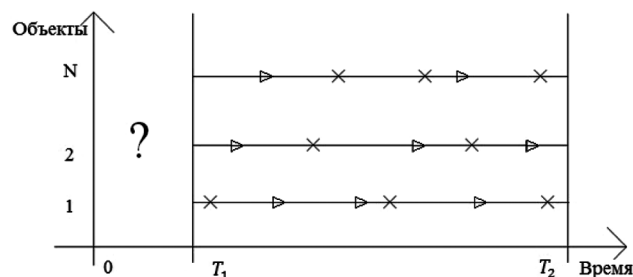


Рисунок 1 – График событий для потока данных

зурирование справа. На графике представлен также интервал $[0, T_1]$, на котором сбор данных не осуществлялся, т.е. на этом интервале наблюдения не проводились. На интервале $[T_1, T_2]$ велись наблюдения, и фиксировалась информация по каждому из объектов. Следовательно, задача состоит в необходимости восстановления данных на первом интервале для того, чтобы иметь возможность произвести оценку параметра закона распределения наработок на интервале $[0, T_2]$.

Разработка методики учета пропущенных данных с использованием метода повторной выборки

При решении вопросов статистической оценки показателей надёжности элементов и систем особую актуальность имеет задача сбора и представления исходной информации о поведении объектов анализа. От точности и достоверности исходной информации зависит точность оценок параметров плотности распределения и результаты расчета характеристик надёжности.

Как было отмечено ранее, при проведении статистического анализа приходится сталкиваться с ситуацией, когда на определенных временных интервалах информация о поведении объекта отсутствует. Возникает ситуация данных с пропусками (рисунок 1), что значительно осложняет математическую обработку, так как это приводит к смещению основных статистических характеристик, таких, как математическое ожидание или дисперсия.

Постановка задачи. Пусть под наблюдением находится N объектов (рисунок 1). Для каждого объекта имеется набор данных за определенный период времени. За время функционирования объекта на интервале от 0 до T_1 информация о его поведении не фиксировалась. С момента T_1 до T_2 осуществлялся сбор информации. На основании результатов наблюдений на интервале $[T_1, T_2]$ фиксируются для каждого объекта наработки до отказа, наработки до цензурирования. Задача состоит в том, чтобы восстановить данные на интервале $[0, T_1]$, которые могут быть как отказами, так и цензурированными данными, и оценить параметр закона распределения наработок на интервале $[0, T_2]$.

Задачу восстановления пропущенных данных предлагается решать с помощью методики, основанной на методе повторной выборки.

Методика заключается в следующем:

1. По результатам наблюдений на интервале $[T_1, T_2]$ производится оценка параметра закона распределения отдельно для наработок до отказа и отдельно для наработок до цензурирования с помощью метода максимального правдоподобия.

2. Оценивается количество наработок n на интервале $[0, T_1]$ для каждого вида наработок: $n = mT_1/(T_2 - T_1)$, где m – количество наработок на интервале $[T_1, T_2]$, в случае однородного потока данных.

3. Моделируется n наработок по заданному закону распределения $F(t)$ (рисунок 2) для каждого вида наработок. Т.е. на вероятностной оси моделируем равномерно распределенное случайное число γ_i . Выполняем взаимно однозначное отображение на ось времени t .

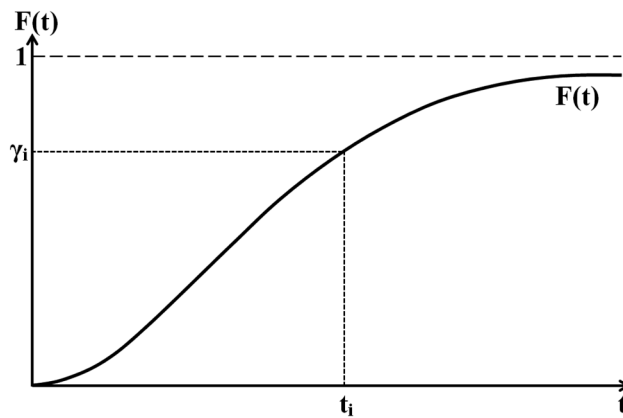


Рисунок 2 – Метод восстановления пропущенных данных по функции распределения

4. Объединяются выборки, полученные на интервале $[0, T_1]$ и $[T_1, T_2]$.

5. Производится оценка параметра закона распределения θ_i наработок на интервале $[0, T_2]$, а также расчет среднеквадратичного отклонения σ_i с помощью метода максимального правдоподобия.

6. Далее рассчитывается оценка параметра закона распределения отдельно для наработок до отказа и отдельно для наработок до цензурирования на интервале $[0, T_2]$.

7. Полученные оценки на шаге 6 используются для повторного моделирования наработок на интервале $[0, T_1]$. Шаги с 3 по 7 повторяются k раз. Количество итераций k определяется исследователем.

8. По завершению шага 7 рассчитываются средняя оценка параметра закона распределения:

$$\hat{\theta} = \frac{1}{k} \sum_{i=1}^k \theta_i \quad (1)$$

и среднее среднеквадратичное отклонение:

$$\hat{\sigma} = \frac{1}{k} \sum_{i=1}^k \sigma_i \quad (2)$$

Найденные значения $\hat{\theta}$ и $\hat{\sigma}$ являются результатом применения метода повторной выборки.

Тестовый пример

Шаг 1: В качестве исходной информации используется информация, полученная в результате моделирования случайной величины в соответствии с экспоненциальным законом распределения. Общее число данных на интервале $[0, T_2]$ составляет 1000 наработок из них

Таблица 1. Результаты работы метода повторной выборки

Процент пропусков	Алгоритм заполнения пропусков	Оценка параметра закона распределения для восстановленной выборки, * 10^{-3}	Среднеквадратичное отклонение, * 10^{-3}
10%	Без восстановления	2,908	0,110
	С помощью среднего арифметического	2,959	0,107
	Бутстреп заполнения	2,909	0,104
	Метод повторной выборки	2,911	0,105
20%	Без восстановления	2,874	0,116
	С помощью среднего арифметического	2,917	0,105
	Бутстреп заполнения	2,877	0,103
	Метод повторной выборки	2,877	0,103
30%	Без восстановления	2,877	0,125
	С помощью среднего арифметического	2,923	0,106
	Бутстреп заполнения	2,881	0,104
	Метод повторной выборки	2,878	0,103
40%	Без восстановления	2,894	0,136
	С помощью среднего арифметического	2,946	0,106
	Бутстреп заполнения	2,895	0,104
	Метод повторной выборки	2,897	0,104
50%	Без восстановления	2,963	0,151
	С помощью среднего арифметического	3,015	0,109
	Бутстреп заполнения	2,966	0,106
	Метод повторной выборки	2,968	0,106

581 наработка до отказа и 419 цензурированных данных, количество определяется случайным образом. Для моделирования данных использовали экспоненциальное распределение со следующими параметрами $\lambda_i = 0,003$ и $\lambda_c = 0,002$ для наработок до отказа и наработок до цензурирования соответственно. Из полученного набора искусственно удаляются данные (10%, 20%, ..., 50%) для каждого эксперимента. На рисунке 1 схематично показан полученный набор данных.

Шаг 2: Используя разработанную методику, получаем среднюю оценку параметра экспоненциального закона распределения λ и среднее среднеквадратичное отклонение $\hat{\sigma}$. Характеристики надёжности рассчитывали с использованием метода максимального правдоподобия. Результаты записаны в таблице 1.

В таблице 1 отражены не только результаты работы метода повторной выборки, но и предоставлены результаты, полученные с помощью таких методов, как метод заполнения средним арифметическим (однократное замещение) [3], [4] и бутстреп метод [9].

При использовании метода заполнения средним арифметическим вместо пропущенных данных подставляется значение среднего арифметического выборки, рассчитанного на интервале $[T_1, T_2]$, вместо каждой пропущенной величины на интервале $[0, T_1]$. Расчёт характеристик надёжности производится с помощью метода максимального правдоподобия.

Бутстреп метод – это практический компьютерный метод исследования распределения статистик вероятностных распределений, основанный на многократной

генерации выборок методом Монте-Карло на базе имеющейся выборки. Т.е. берутся данные на интервале $[T_1, T_2]$ откуда на каждом шаге из n последовательных итераций, равномерно распределённых на интервале $[1, n]$, случайным образом достаётся произвольный элемент, который затем снова возвращается в исходную выборку (т.е. может быть извлечен повторно). Здесь n – количество наработок на интервале $[0, T_1]$. Полученные элементы формируют набор данных на интервале $[0, T_1]$. Затем выборка на интервале $[0, T_2]$ оценивается методом максимального правдоподобия. Количество итераций для генерации новой выборки устанавливается исследователем. В нашем случае при проведении исследования было проведено 1000 итераций. Средняя оценка параметра экспоненциального закона распределения и среднее среднеквадратичное отклонение находят по формулам (1) и (2).

Результаты обоих методов записаны в таблице 1.

Чтобы легче было сравнивать результаты, отобразим наши данные на следующих рисунках (3, 4, 5) и сделаем выводы по каждому из них.

На рисунке 3 изображена оценка параметра закона распределения для эталонной выборки и оценка для потока данных с пропусками (10% ÷ 50%). Как можно видеть из рисунка 3 и таблицы 1, чем больше процент пропущенных данных, тем больше среднеквадратичное отклонение. Следовательно, появляется необходимость восстановления пропущенных данных, чтобы уменьшить отклонение.

Проанализируем полученную информацию на рисунке 4. Сравним для начала оценки параметра закона

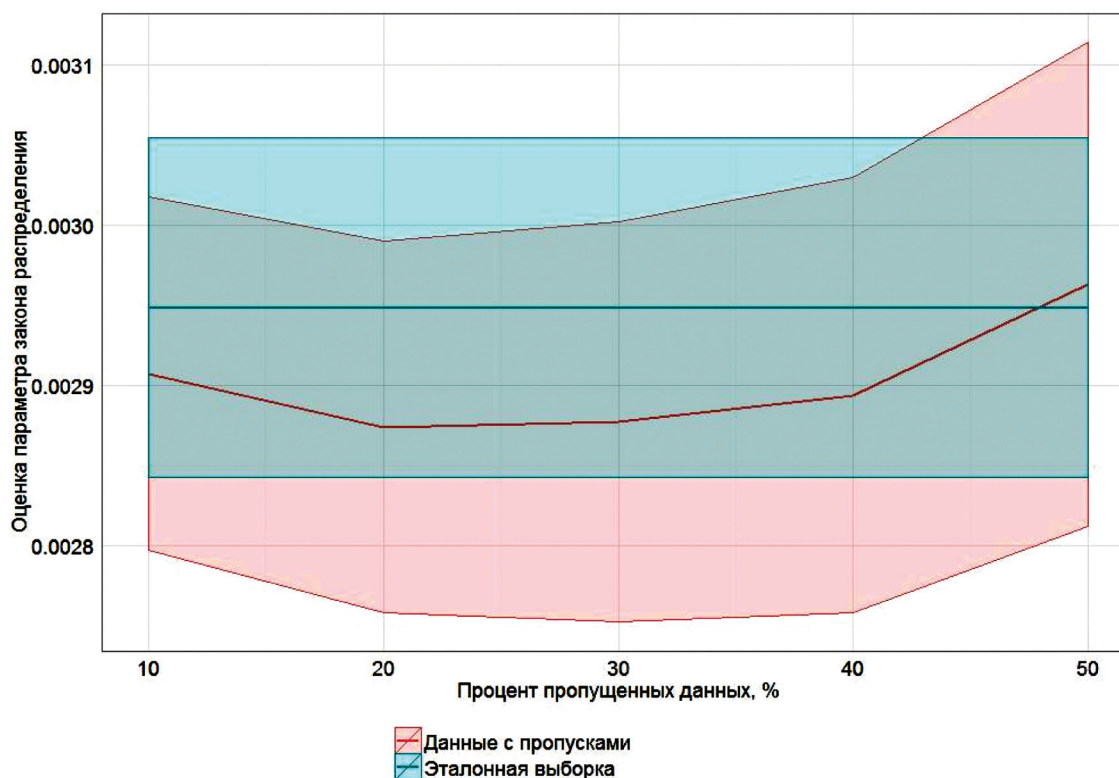


Рисунок 3 – Сравнение оценок эталонных данных и данных с пропусками

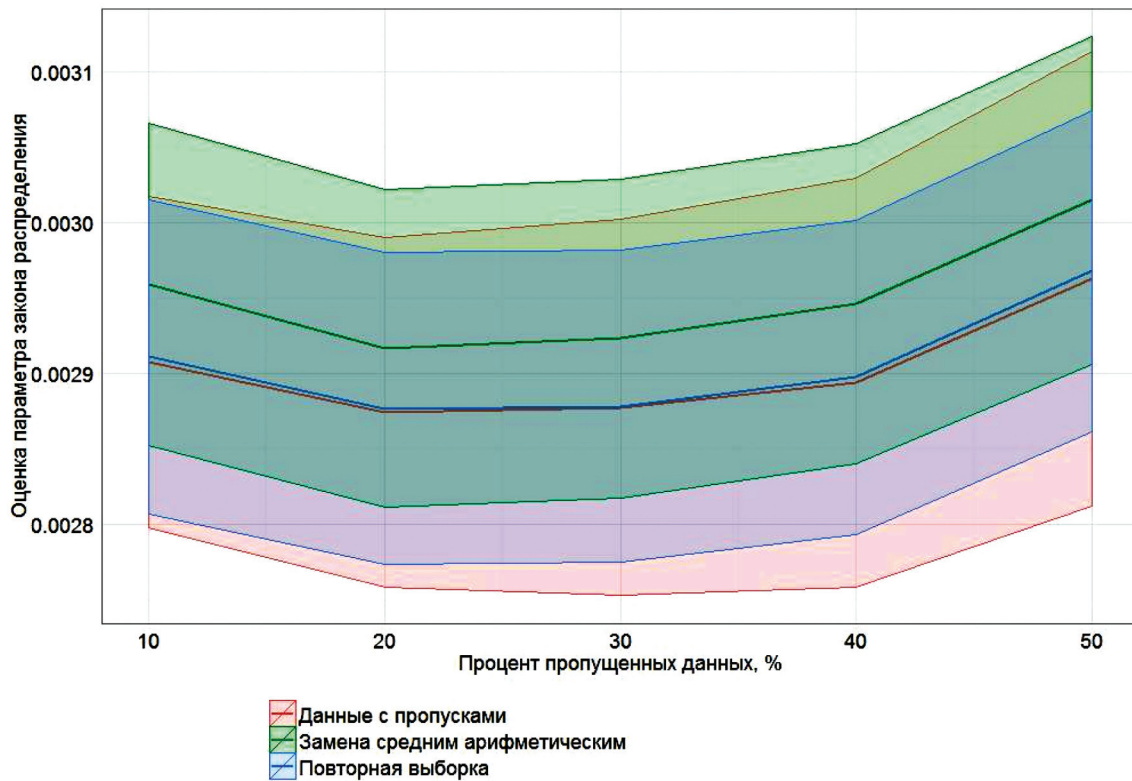


Рисунок 4 – Сравнение оценок, полученных методом повторной выборки и заменой средним арифметическим

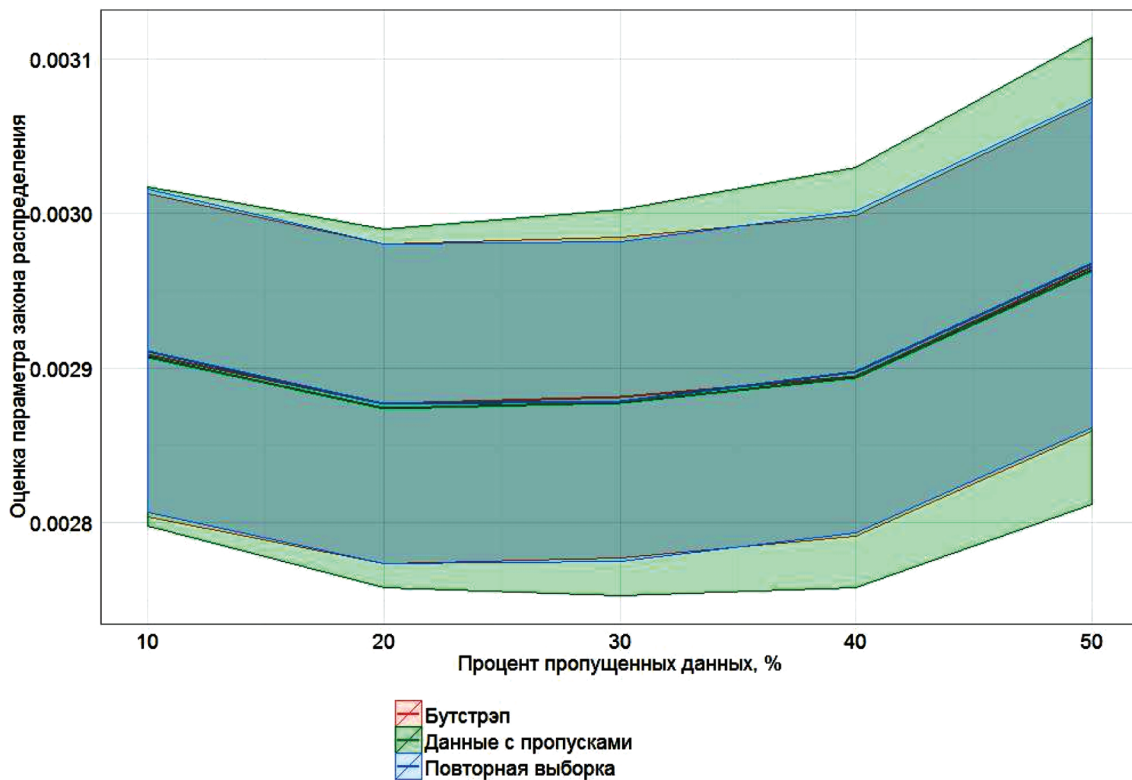


Рисунок 5 – Сравнение оценок, полученных методом повторной выборки и бутстрэп методом

распределения. Как можно видеть, оценка, полученная методом повторной выборки, и оценка данных с пропусками совпадают, это говорит нам о том, что восстановление пропусков не влияет на саму оценку. Оценка,

полученная методом замены средним арифметическим, получилась смещенной, это естественное следствие заполнения пропусков средним, т.е. значением в центре распределения. Следовательно, среднеквадратичное

отклонение получилось смещенным. Рассмотрим отклонение для метода повторной выборки. Используя этот метод для восстановления пропусков, мы смогли уменьшить среднеквадратичное отклонение.

Результаты, полученные методом повторной выборки, совпадают с результатами, полученными бутстреп методом (рисунок 5). Отсюда можно сделать два вывода. Во-первых, предложенный метод повторной выборки не уступает в точности бутстреп методу. Во-вторых, оба метода относятся к классу методов моделирования выборки, с одним лишь отличием, что предложенный метод является параметрическим методом, а бутстреп метод является непараметрическим методом.

Заключение

В ходе выполнения работы были получены следующие основные результаты и выводы.

1. Произведено описание различных видов данных. Разработан параметрический метод восстановления пропущенных данных с учетом цензурированной информации, который был исследован на тестовом примере. Проведено сравнение с другими методами (замена средним и бутстреп метод), которое показало эффективность предложенного метода.

2. Реализованы процедуры восстановления информации и оценки параметра экспоненциального закона распределения: метод повторной выборки, метод замещения средним арифметическим, метод максимального правдоподобия. Проведенные расчеты позволяют сделать вывод о том, что предложенный метод повторной выборки не уступает в точности бутстреп методу. Также особенностью является то, что метод повторной выборки является параметрическим методом, в то время как бутстреп – непараметрическим методом.

3. Исследована точность восстановления плотности распределения. Данное исследование показало, что восстановление данных ведет к снижению неопределённости в рассчитываемых показателях (параметры экспоненциального закона распределения), тем самым показывая необходимость учета пропущенных данных.

4. Выполнено сравнение результатов оценивания информации, которая была восстановлена различными ме-

тодами: методом повторной выборки, методом бутстреп и методом заполнения средним арифметическим. Метод заполнения средним показал, что его применение ведет к смещению оценки параметра закона распределения. В то время как методом повторной выборки и бутстреп методом мы получили результаты без смещения.

Библиографический список

1. Антонов А. В. Системный анализ. – М.: Высшая Школа, 2004. – 454 с.
2. Антонов А. В., Никулин М.С. Статистические модели в теории надёжности: Учебное пособие. – М.: Абрис, 2012. – 390 с.
3. Зангиева И. К. Решение проблем неполных данных массовых опросов // Российская социология завтрашнего дня. – 2008. – С. 84-95.
4. Злоба Е., Яцкив И. Статистические методы восстановления пропущенных данных // Computer Modeling & New Technologies. – 2004. – Vol. 6. – P. 55-56.
5. Кокс Д. Р., Оукс Д. Анализ данных типа времени жизни. – М.: Финансы и статистика, 1988. – 191 с.
6. Литтл Р. А. Д., Рубин Д. Б. Статистический анализ данных с пропусками. – М.: Финансы и статистика, 1991. – 336 с.
7. Эфрон Б. Нетрадиционные методы многомерного статистического анализа. – М.: Финансы и статистика, 1988. – 261с.
8. Bischl B., Mersmann O., Trautmann H. Resampling methods in model validation // Algorithm Engineering Report. – August 2010. – № 7 – P. 14-31.
9. Meeker W. Q., Escobar A. Statistical Methods for Reliability Data. – New York: JOHN WILEY & SONS, Inc., 1998. – 701 p.

Сведения об авторе

Дмитрий А. Николаев – аспирант Обнинского института атомной энергетики НИЯУ МИФИ. 249040, Россия, Калужская область, г. Обнинск, Студгородок, д.1, e-mail: dafanday@gmail.com

Поступила 01.06.2016