


Малеев Е.А., Чепурко В.А.

КОРНЕВАЯ ОЦЕНКА ПЛОТНОСТИ РАСПРЕДЕЛЕНИЯ ПО НЕПОЛНЫМ ДАННЫМ

В статье предложены две модификации непараметрической корневой оценки плотности распределения в ситуации наличия неполных данных в виде группированных частот отказов. Первый (интегральный) метод связан с соответствующим изменением функции правдоподобия. Второй (resampling) метод восстановления отказов основан на итерационном восстановлении моментов отказов. Исследованы точности предложенных методов оценивания.

Ключевые слова: *пси-функция, метод квадратного корня, функция правдоподобия, псевдоотказы.*

Объем информации, поступающей на обработку с объектов атомных станций, как правило, ограничен. Приходится сталкиваться с информацией, в которой наряду с наработками отказавших объектов присутствуют наработки объектов, продолжающих работать, но наблюдения за функционированием которых были по различным причинам приостановлены. Кроме этого, часто приходится иметь дело с группированной информацией об отказах, в которой потеряна информация о наработках отказавших объектов, а известна лишь частота их появления. Информацию подобной неопределенности называют цензурированной.

Как известно, методы анализа статистической информации делятся на параметрические и непараметрические. Для анализа данных об отказах, поступающих с объектов атомных станций, рациональнее использовать непараметрические методы, не требующие, чтобы распределение вероятностей было описано каким-либо параметрическим законом распределения. [1]

Наиболее общей характеристикой, описывающей поведение одномерной случайной величины, является ее плотность распределения $f(t)$. Задача оценки плотности распределения наблюдаемой случайной величины по конечному числу ее реализаций при наличии неопределенностей является одной из ключевых задач статистического анализа, что и определяет актуальность настоящей статьи.

Известно множество методов оценивания плотности распределения полных и цензурированных данных: гистограммные, проекционные, ядерные, корневые оценки. Все методы имеют как достоинства, так и недостатки.

Метод гистограмм прост в реализации, однако не слишком нагляден, и гистограмма, построенная по малым выборкам, не позволяет сделать правильных выводов. Недостатком проекционной оценки является то, что на краях рассматриваемого интервала она может принимать отрицательные

значения, тогда как плотность по определению неотрицательна. Качество ядерной оценки сильно зависит от выбора «ядра».

Корневая оценка представляет собой квадрат разлагаемой по ортонормированному базису функции и заведомо задает плотность. Оценка хорошо изучена для полных данных. В статье рассматривается корневая оценка для данных, имеющих неопределенность в моменте реализации исследуемого признака, т.е. для цензурированных данных.

В статье корневой метод оценки плотности условно разделен на два метода: интегральный и итеративный.

Интегральный метод корневого оценивания – классический метод корневой оценки, где искомая плотность распределения $f_\xi(x)$ находится, как квадрат так называемой пси-функции:

$$f_\xi(x) = |\psi(x)|^2. \tag{1}$$

Пусть

$$\psi(x) = \sum_{i=1}^{m} c_i \varphi_i(x),$$

где $\{\varphi_i(x)\}$ – ортонормированная система, $\{c_i\}$ – коэффициенты разложения, подлежащие оценке [2, 3].

В дальнейшем предполагается, что функции $\varphi_i(x)$, $\psi(x)$ и коэффициенты c_i действительны. Из условия нормировки $\int f_\xi(x)dx = 1$ следует равенство

$$\sum_{i,j=1}^{m} c_i c_j \int \varphi_i(x)\varphi_j(x)dx = \sum_{i=1}^{m} c_i^2 = 1. \tag{2}$$

Следовательно, необходимо оценить *m-1* независимых коэффициентов. Для их оценки используется метод максимального правдоподобия.

Если выборка повторная – $\vec{\xi} = (\xi_1, ..., \xi_p)$, то функция правдоподобия (ФП) имеет вид:

$$L_n(\vec{c}) = \prod_{k=1}^{p} \hat{f}_\xi(\xi_k) = \prod_{k=1}^{p} \left(\sum_{i=1}^{m} c_i \varphi_i(\xi_k) \right)^2. \tag{3}$$

Логарифмическая функция правдоподобия (ЛФП):

$$l_n(\vec{c}) = \ln L_n(\vec{c}) = \sum_{k=1}^{p} \ln \hat{f}_\xi(\xi_k) = \sum_{k=1}^{p} \ln \left(\sum_{i=1}^{m} c_i \varphi_i(\xi_k) \right)^2 = \sum_{k=1}^{p} \ln \sum_{i=1}^{m} \sum_{j=1}^{m} c_i c_j \varphi_i(\xi_k)\varphi_j(\xi_k).$$

Ее частные производные:

$$\frac{\partial l_n(\vec{c})}{\partial c_i} = \sum_{k=1}^{p} \frac{\partial}{\partial c_i} \ln \hat{f}_\xi(\xi_k) = \sum_{k=1}^{p} \frac{1}{\hat{f}_\xi(\xi_k)} \frac{\partial}{\partial c_i} \hat{f}_\xi(\xi_k) = \sum_{k=1}^{p} \frac{1}{\hat{f}_\xi(\xi_k)} \frac{\partial}{\partial c_i} \left(\sum_{j=1}^{m} c_j \varphi_j(\xi_k) \right)^2 =$$

$$= \sum_{k=1}^{p} \frac{2\varphi_i(\xi_k) \left(\sum_{j=1}^{m} c_j \varphi_j(\xi_k) \right)}{\hat{f}_\xi(\xi_k)} = 2 \sum_{k=1}^{p} \sum_{j=1}^{m} \frac{\varphi_j(\xi_k)\varphi_i(\xi_k)}{\hat{f}_\xi(\xi_k)} c_j .$$

Возникает оптимизационная задача

$$L_n(\vec{c}) = \prod_{i=k}^{p} \hat{f}_\xi(\xi_k) = \prod_{i=k}^{p} \left(\sum_{i=1}^{m} c_i \varphi_i(\xi_k) \right)^2 \to \max_{\vec{c}}$$

с ограничением типа равенства: $\sum_{i=1}^{m} c_i^2 = 1$. Коэффициенты c_i подбираются таким образом, чтобы ФП была максимальна, при этом их сумма квадратов равна 1.

Для нахождения максимального значения логарифмической функции правдоподобия $l_n(\vec{c})$ с учетом ограничения (2) формируется функция Лагранжа:

$$L(\vec{c}) = l_n(\vec{c}) + \lambda \left(1 - \sum_{i=1}^{m} c_i^2 \right).$$

Производная функции Лагранжа приравнивается к нулю:

$$\frac{\partial}{\partial c_i} L(\vec{c}) = 2 \sum_{k=1}^{p} \sum_{j=1}^{m} \frac{\varphi_j(\xi_k)\varphi_i(\xi_k)}{\hat{f}_\xi(\xi_k)} c_j - 2\lambda c_i = 0 . \qquad (4)$$

Умножая обе части (4) на c_i и суммируя по i получим $\lambda = p$. Подставим это в (4):

$$c_i = \frac{1}{p} \sum_{k=1}^{p} \varphi_i(\xi_k) \frac{\sum_{j=1}^{m} c_j \varphi_j(\xi_k)}{\hat{f}_\xi(\xi_k)} = \frac{1}{p} \sum_{k=1}^{p} \varphi_i(\xi_k) \left(\sum_{j=1}^{m} c_j \varphi_j(\xi_k) \right)^{-1} .$$

Далее для нахождения коэффициентов используются итерационные численные методы.

В случае если имеются цензурированные данные, в качестве интервалов для построения оценки выступают элементы массива интервалов $\vec{LR} = \left[(l_1, r_1); (l_2, r_2); ...; (l_s, r_s) \right]$. Величина скачка эмпирической функции распределения в начале каждого интервала пропорциональна количеству элементов выборки (случайному числу отказов), попавших на данный интервал $\vec{\nu} = (\nu_1, \nu_2, ..., \nu_s)$.

Функция правдоподобия (3) для такого рода данных примет вид:

$$L_n(\vec{c}) = \prod_{k=1}^{p} \hat{f}_\xi(\xi_k) \times \prod_{m=1}^{s} \left(\hat{F}_\xi(r_m) - \hat{F}_\xi(l_m)\right)^{\nu_m},$$

т.е. к функции правдоподобия полных данных добавится сомножитель $\prod_{m=1}^{s} \left(\hat{F}_\xi(r_m) - \hat{F}_\xi(l_m)\right)^{\nu_m}$, отвечающий за цензурированные данные.

Частные производные логарифмической функции правдоподобия будут равны следующим суммам:

$$\frac{\partial l_n(\vec{c})}{\partial c_i} = 2\sum_{k=1}^{p}\sum_{j=1}^{m} \frac{\varphi_j(\xi_k)\varphi_i(\xi_k)}{\hat{f}_\xi(\xi_k)} c_j + \sum_{m=1}^{s} \nu_m \frac{\partial \ln\left(\hat{F}_\xi(r_m) - \hat{F}_\xi(l_m)\right)}{\partial c_i}.$$

Рассмотрим отдельно

$$\frac{\partial \ln\left(\hat{F}_\xi(r_m) - \hat{F}_\xi(l_m)\right)}{\partial c_i}. \tag{5}$$

Если оценка плотности распределения

$$\hat{f}_\xi(x) = \left(\sum_{i=1}^{m} c_i \varphi_i(x)\right)^2, \tag{6}$$

то целесообразно взять в качестве оценки функции распределения интеграл:

$$\hat{F}_\xi(x) = \int_{-\infty}^{x} \hat{f}_\xi(u)du. \tag{7}$$

В дальнейших выкладках предположим, что плотность распределения имеет носитель-отрезок [0,1]. Для плотностей с другим носителем распределения можно сделать необходимое линейное преобразование случайной величины, отображающее множество значений случайной величины в отрезок [0,1] или использовать иные ортонормированные базисы.

Как известно, $\varphi_k(x) = \sqrt{2}\sin(k\pi x)$, $k = 1,2,...$ – ортонормированный базис Фурье на отрезке *[0,1]*. Найдем функцию распределения (7), используя разложение (6).

$$\hat{F}_\xi(x) = \sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j \int_0^x \varphi_i(u)\varphi_j(u)du = \sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j 2\int_0^x \sin(i\pi u)\sin(j\pi u)du =$$

$$= x - \frac{1}{2\sqrt{2}\pi}\left[\sum_{i=1}^{m} c_i^2 \frac{\varphi_{2i}(x)}{i} + \sum_{i=1}^{m}\sum_{j=1, j\neq i}^{m} c_i c_j \left(\frac{\varphi_{i+j}(x)}{i+j} - \frac{\varphi_{i-j}(x)}{i-j}\right)\right].$$

Частные производные функции распределения равны:

$$\frac{\partial \hat{F}_\xi(x)}{\partial c_i} = \frac{1}{\sqrt{2}\pi}\left[2\sum_{j=1, j\neq i}^{m} c_j\left(\frac{\varphi_{i-j}(x)}{i-j} - \frac{\varphi_{i+j}(x)}{i+j}\right) - \frac{c_i\varphi_{2i}(x)}{i}\right] =$$

$$= \frac{\sqrt{2}}{\pi}\left[\sum_{j=1, j\neq i}^{m}\frac{c_j\varphi_{i-j}(x)}{i-j} - \sum_{j=1}^{m}\frac{c_j\varphi_{i+j}(x)}{i+j}\right].$$

После подстановки полученных результатов (частных производных) в выражение (5) получим следующие уравнения:

$$\frac{\partial \ln\left(\hat{F}_\xi(r_m) - \hat{F}_\xi(l_m)\right)}{\partial c_i} = \frac{\sqrt{2}}{\pi\left(\hat{F}_\xi(r_m) - \hat{F}_\xi(l_m)\right)}\left[\sum_{j=1, j\neq i}^{m}\frac{c_j\left(\varphi_{i-j}(r_m) - \varphi_{i-j}(l_m)\right)}{i-j} - \right.$$

$$\left. - \sum_{j=1}^{m}\frac{c_j\left(\varphi_{i+j}(r_m) - \varphi_{i+j}(l_m)\right)}{i+j}\right].$$

$$\text{Тогда } \frac{\partial l_n(\vec{c})}{\partial c_i} = 2\sum_{k=1}^{p}\sum_{j=1}^{m}\frac{\varphi_j(\xi_k)\varphi_i(\xi_k)}{\hat{f}_\xi(\xi_k)}c_j + \sum_{m=1}^{s}\frac{\sqrt{2}\nu_m}{\pi\left(\hat{F}_\xi(r_m) - \hat{F}_\xi(l_m)\right)}\times$$

$$\times\left[\sum_{j=1, j\neq i}^{m}\frac{c_j\left(\varphi_{i-j}(r_m) - \varphi_{i-j}(l_m)\right)}{i-j} - \sum_{j=1}^{m}\frac{c_j\left(\varphi_{i+j}(r_m) - \varphi_{i+j}(l_m)\right)}{i+j}\right].$$

Необходимое условие экстремума, как и для полных данных, сводится к условиям равенства нулю частных производных функции Лагранжа:

$$\frac{\partial}{\partial c_i}L(\vec{c}) = 2\sum_{k=1}^{p}\sum_{j=1}^{m}\frac{\varphi_j(\xi_k)\varphi_i(\xi_k)}{\hat{f}_\xi(\xi_k)}c_j - 2\lambda c_i + \sum_{m=1}^{s}\frac{\sqrt{2}\nu_m}{\pi\left(\hat{F}_\xi(l_m) - \hat{F}_\xi(r_m)\right)}\times$$

$$\times\left[\sum_{j=1, j\neq i}^{m}\frac{c_j\left(\varphi_{i-j}(r_m) - \varphi_{i-j}(l_m)\right)}{i-j} - \sum_{j=1}^{m}\frac{c_j\left(\varphi_{i+j}(r_m) - \varphi_{i+j}(l_m)\right)}{i+j}\right] = 0. \qquad (8)$$

Умножим обе части (8) на c_i, просуммируем по i, получим уравнения для λ:

$$2(p-\lambda)+\sum_{m=1}^{s}\frac{\sqrt{2}\nu_m}{\pi\left(\hat{F}_\xi(r_m)-\hat{F}_\xi(l_m)\right)}\times\sum_{i=1}^{m}c_i\times$$

$$\times\left[\sum_{j=1,j\neq i}^{m}\frac{c_j\left(\varphi_{i-j}(r_m)-\varphi_{i-j}(l_m)\right)}{i-j}-\sum_{j=1}^{m}\frac{c_j\left(\varphi_{i+j}(r_m)-\varphi_{i+j}(l_m)\right)}{i+j}\right]=0.$$

Решаем:

$$\lambda=p+\sum_{m=1}^{s}\frac{\nu_m}{\pi\sqrt{2}\left(\hat{F}_\xi(r_m)-\hat{F}_\xi(l_m)\right)}\times\sum_{i=1}^{m}c_i\times\left[\sum_{j=1,j\neq i}^{m}\frac{c_j\left(\varphi_{i-j}(r_m)-\varphi_{i-j}(l_m)\right)}{i-j}-\right.$$

$$\left.-\sum_{j=1}^{m}\frac{c_j\left(\varphi_{i+j}(r_m)-\varphi_{i+j}(l_m)\right)}{i+j}\right].$$

Итерационный процесс для нахождения коэффициентов разложения:

$$c_i^{l+1}=\alpha c_i^l+\frac{1-\alpha}{2\lambda}\left\{2\sum_{k=1}^{p}\sum_{j=1}^{m}\frac{\varphi_j(\xi_k)\varphi_i(\xi_k)}{\hat{f}_\xi^l(\xi_k)}c_j^l+\sum_{m=1}^{s}\frac{\sqrt{2}\nu_m}{\pi\left(\hat{F}_\xi^{\,l}(r_m)-\hat{F}_\xi^{\,l}(l_m)\right)}\times\right.$$

$$\left.\times\left[\sum_{j=1,j\neq i}^{m}\frac{c_j^{\,l}\left(\varphi_{i-j}(r_m)-\varphi_{i-j}(l_m)\right)}{i-j}-\sum_{j=1}^{m}\frac{c_j^{\,l}\left(\varphi_{i+j}(r_m)-\varphi_{i+j}(l_m)\right)}{i+j}\right]\right\}.$$

Теперь приведем формулы для корневой оценки на произвольном отрезке локализации. В произвольном случае (случайная величина распределена на отрезке *[a, b]*) базис будет иметь вид:

$$\varphi_k(x)=\sqrt{\frac{2}{b-a}}\sin\left(k\pi\frac{x-a}{b-a}\right).$$

В качестве *[a, b]* могут быть взяты первая и последняя порядковые статистики, т.е. наименьшее и наибольшее значение.

Итерационный процесс для нахождения коэффициентов разложения в этом случае приведет к следующей схеме:

$$c_i^{l+1}=\alpha c_i^l+\frac{1-\alpha}{2\lambda}\left\{2\sum_{k=1}^{p}\sum_{j=1}^{m}\frac{\varphi_j(\xi_k)\varphi_i(\xi_k)}{\hat{f}_\xi^l(\xi_k)}c_j^l+\sum_{m=1}^{s}\frac{\sqrt{2(b-a)}\nu_m}{\pi\left(\hat{F}_\xi^{\,l}(r_m)-\hat{F}_\xi^{\,l}(l_m)\right)}\times\right.$$

$$\left.\times\left[\sum_{j=1,j\neq i}^{m}\frac{c_j^{\,l}\left(\varphi_{i-j}(r_m)-\varphi_{i-j}(l_m)\right)}{i-j}-\sum_{j=1}^{m}\frac{c_j^{\,l}\left(\varphi_{i+j}(r_m)-\varphi_{i+j}(l_m)\right)}{i+j}\right]\right.$$

Следующая модификация метода корневого оценивания по цензурированным данным основана на методе восстановления псевдоотказов. В западной литературе эта процедура носит название Resampling-method. Принцип моделирования псевдонаблюдений опирается на известное свойство монотонной функции распределения – случайная величина $F_\xi(\xi)$ является равномерно распределенной на отрезке *[0;1]*. Для каждого источника с цензурированными данными производится поиск значения функции распределения в точках границ интервалов цензурирования, затем в каждом из построенных интервалов функций моделируется некоторое число точек, равное числу отказов на данном интервале. С помощью интерполяции производится обратное отображение смоделированных «псевдоотказов» на ось наработок (рис. 1). [4]

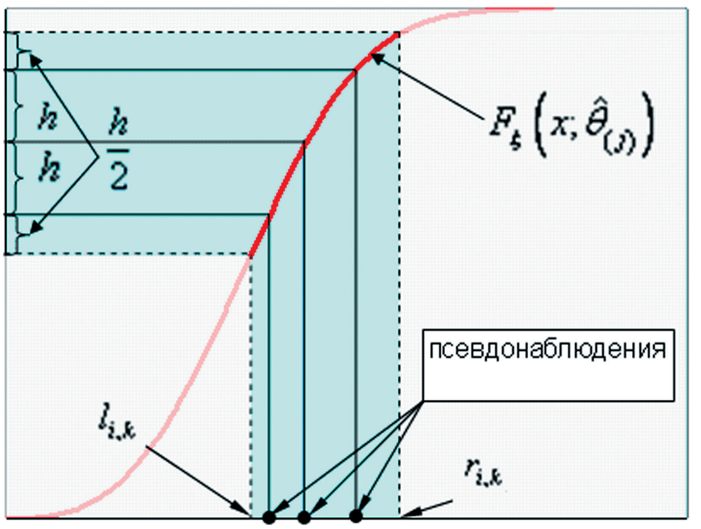

Рис. 1. Моделирование псевдонаблюдений на участке неопределенности

На следующем шаге восстановленные отказы от каждого источника с группированной информацией собираются в один массив, далее с ними работаем, как с полными данными, по ним строится корневая (интегральная) оценка по формуле (1). Новая функция распределения находится как интеграл от этой оценки, по новой функции опять восстанавливаются отказы, и так до тех пор, пока итерационный процесс не сойдется.

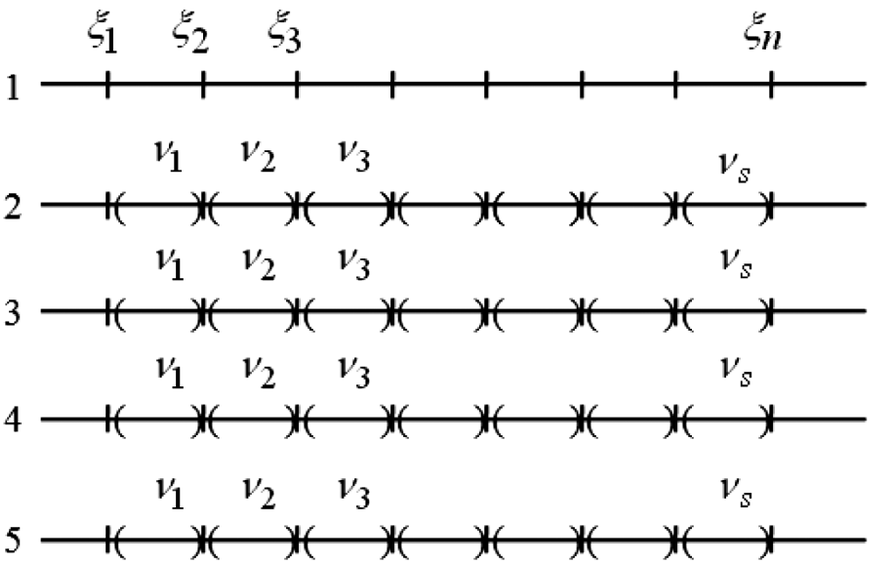

Рис. 2. Моделирование цензурированных данных

Отказы восстанавливались по функции $F_{ср}$, представляющей собой среднее значение эмпирических функций распределения источников с цензурированными данными и функции $F_{пр}$, взятой как интеграл от корневой оценки плотности полных данных.

Анализ оценок. Оценки плотности распределения исследовались на примере закона распределения Вейбулла с параметром формы $\alpha = 2$ и параметром масштаба $\lambda = 2$. Моделирование случайных величин производилось с помощью метода обратных функций. К полученным выборкам применялась корневая оценка плотности распределения, для проверки соответствия оценки истинной плотности. Исследования проводились для пяти источников информации: одного с полной информацией и четырех с цензурированными данными. Схема моделирования цензурирования изображена на рис. 2.

Принятое число наблюдений n для каждого источника – 100. Длина интервала цензурирования по умолчанию – 0,1.

Корневые оценки цензурированной информации интегральным методом изображены на рис. 3. Графики представлены для числа гармоник m, равного 2, 4 и 6.

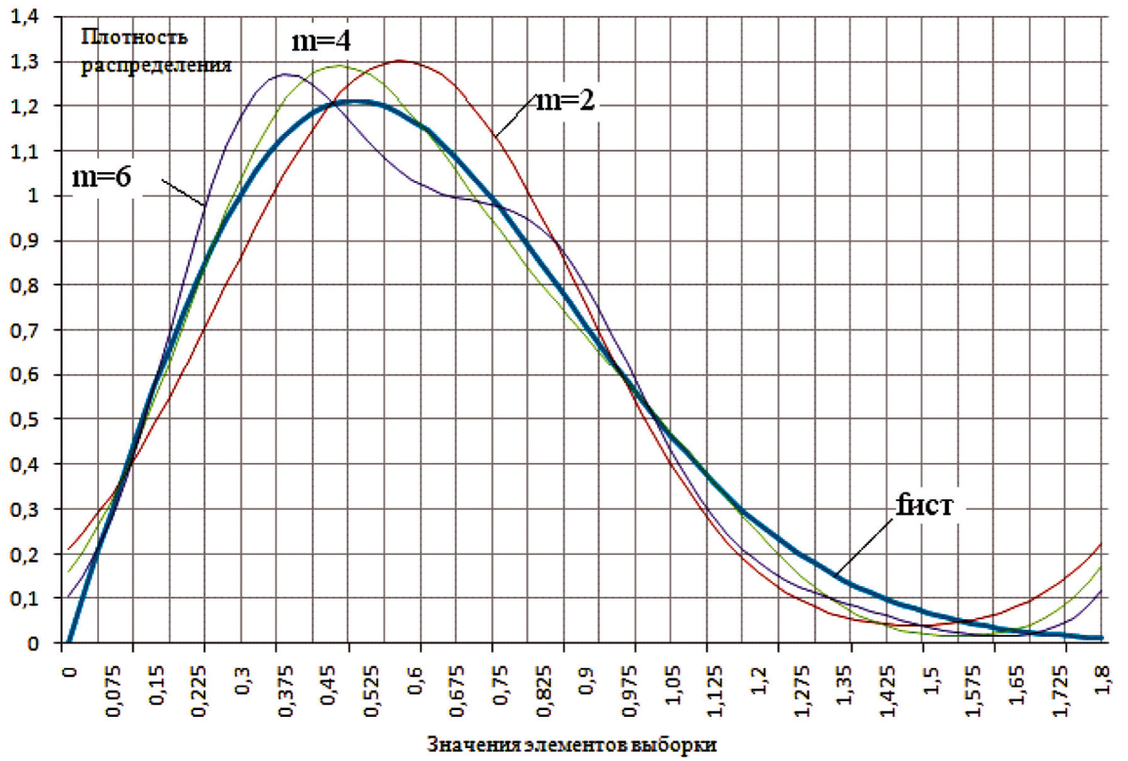

Рис. 3. Корневая оценка плотности распределения цензурированных данных для разного числа гармоник

Качество оценивания зависит от числа гармоник и от длины интервала группирования данных. В ходе исследования выяснилось, что чем меньше интервал группирования, тем точнее оценка.

Графики суммарной ошибки оценивания для разного числа гармоник представлены на рис. 4.

Существует число гармоник, при котором достигается оптимальная оценка плотности. В данном случае $m = 4$ – оптимальное число гармоник.

Корневая оценка плотности методом восстановления отказов изображена на рис. 5, число гармоник $m = 3$.

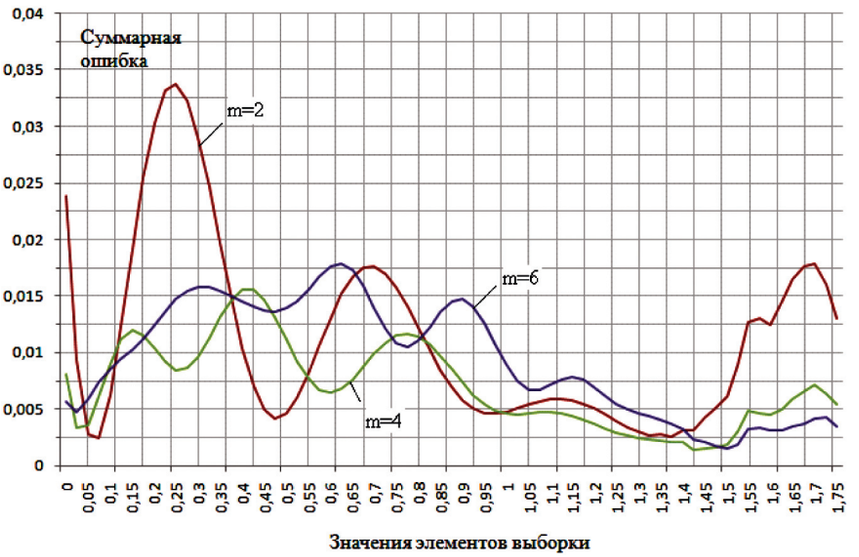

Рис. 4. Суммарная ошибка корневой оценки плотности распределения цензурированных данных для разного числа гармоник

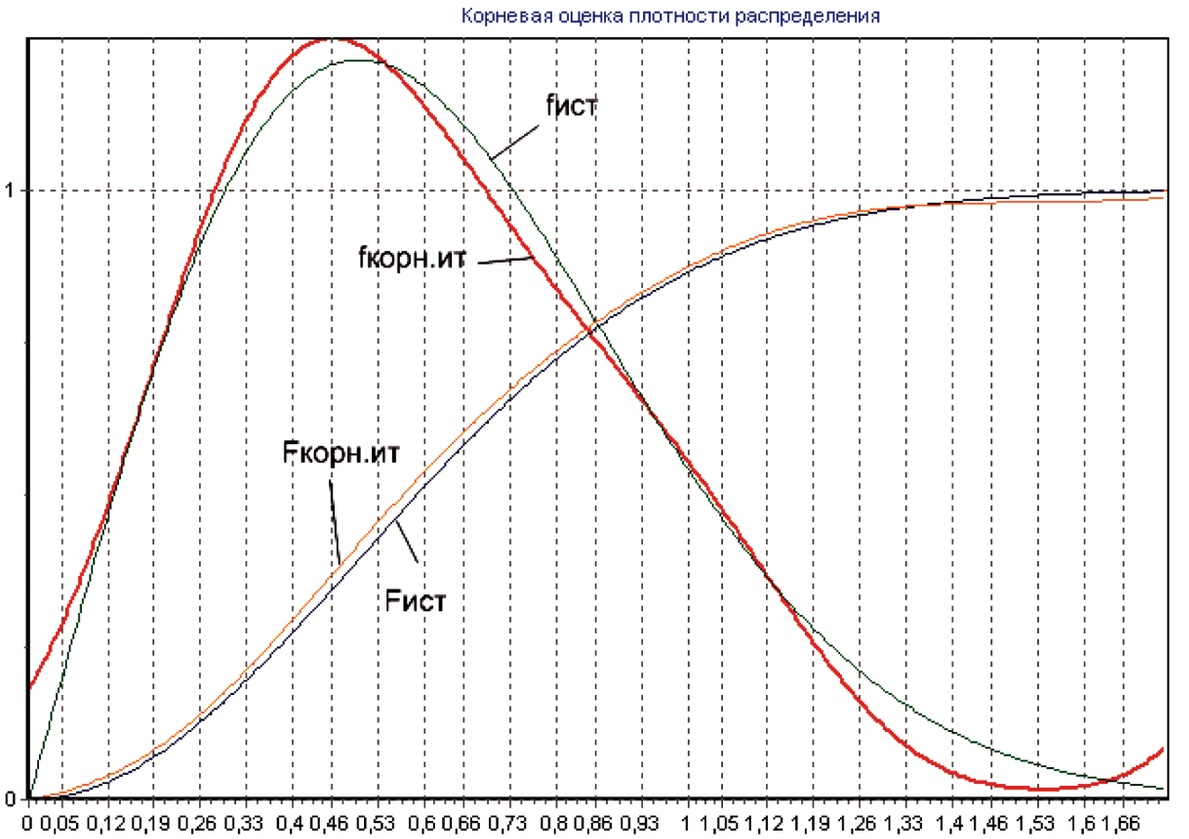

Рис. 5. Корневая оценка плотности методом псевдоотказов, с учетом эмпирических функций распределения цензурированной информации

На рисунке *fист* и *Fист* – истинные значения плотности и функции распределения, *fкорн.ит* и *Fкорн.ит* – их оценки методом восстановления отказов.

Длина интервала цензурирования сильно влияет на качество оценки. Оценка тем точнее, чем меньше длина интервала цензурирования.

Качество оценивания итеративным методом уступает качеству интегрального метода, однако

его несомненное преимущество состоит в том, что фактически он работает с полными данными, нет необходимости в применении сложных формул. Метод итеративного восстановления отказов более прост и удобен в применении.

Литература

1. **Антонов А.В., Чепурко В.А.** Построение непараметрической плотности распределения на основании цензурированной информации. Надежность. – М.: Издательский дом «Технология», 2005, №2. – с.3.

2. **Богданов Ю.И.** Основная задача статистического анализа данных: корневой подход. – М: МИЭТ, 2002. – 96с.: ил.

3. **Крянев А.В., Лукин Г.В.** Математические методы обработки неопределенных данных. – М.: ФИЗМАТЛИТ, 2003. – 216 с.

4. **Ершов А.Н., Чепурко В.А.** Итерационная оценка параметров закона распределения случайной величины при наличии цензурированных данных. Диагностика и прогнозирование состояния сложных систем: сборник научных трудов № 18 каф. АСУ.– Обнинск: ИАТЭ, 2009.– с.14–22.