Надежные оценки эмпирических распределений в условиях малых выборок

Б. С. Добронец

Сибирский федеральный университет

BDobronets@yandex.ru

О. А. Попова

Сибирский федеральный университет

OlgaArc@yandex.ru

Аннотация. **В работе рассматривается проблема построения надежных оценок эмпирических распределений в условиях малых выборок. Для оценки неопределенностей вероятностных оценок предлагается использовать новый подход, основанный на применении вычислительного вероятностного анализа и понятия распределение второго порядка. Для численных операций над распределениями второго порядка применяются вероятностные расширения их параметризованных представлений. Приводятся примеры использования данного подхода для оценки надежности технических объектов в условиях малых выборок.**

Ключевые слова: неточные вероятности, распределения второго порядка, вычислительный вероятностный анализ, оценки надежности

I. Введение

Оценка функций распределения по эмпирическим данным всегда содержит неточности. Особенно это актуально в условиях недостаточного объема информации. Это приводит к неопределенности вероятностных оценок. Существуют различные подходы оценки неточной вероятности, такие как вероятности второго порядка [1].

Если у нас есть интервальная оценка вероятности P, тогда мы можем оценить границы $P_1 > P > P_2$. Одним из распространенных методов работы с неопределенными вероятностями является использование интервальных функций $[F_1(x), F_2(x)]$ для оценки неизвестной функции распределения $F(x)$. Этот подход получил название Probability Boxes (P-box) [2]. Развитие этого подхода привело к понятию гистограммы второго порядка, в этом случае вместо интервалов для оценки $F(x)$ используется гистограмма [4,5]. В работе [3] показано использование неопределенных вероятностей в инженерных расчетах. Неизвестная функция плотности вероятности в этом подходе заменяется семейством распределений. В вычислительном вероятностном анализе для аппроксимации вероятностей второго порядка используются кусочно-полиномиальные модели [4, 5]. Подобные модели получили название распределений второго порядка.

Вероятности второго порядка с успехом используются в различных областях, включая принятие решений [8,9]. Среди монографий, посвященных методам работы с неопределенностями, следует выделить работы О.И. Ужга-Реброва [10].

II. Распределения второго порядка

Для определения распределения второго порядка будем использовать понятия случайный процесс и случайное поле [4, 5].

В вероятностном пространстве $(\Omega,\mathcal{F},\mathbb{P})$ случайный процесс представляет собой набор случайных величин

$$\{a(x,\omega), x \in D, \omega \in \Omega\}.$$

Термин «случайное поле» обычно относится к случайному процессу, принимающему значения в евклидовом пространстве R^n. Случайное поле можно представить двумя способами:

- для фиксированного $x \in D$, $a(x,\cdot)$ является случайной величиной в Ω;
- для фиксированного $\omega \in \Omega$, $a(\cdot,\omega)$ является реализацией случайного поля в D.

Таким образом, определим *распределение второго порядка* $\boldsymbol{f}^{(2)}$ как случайное поле $f(x,\omega), x \in D, \omega \in \Omega$,заданное на $D \subset R$, где $(\Omega,\mathcal{F},\mathbb{P})$ – вероятностное пространство. Обладает следующими свойствами: для фиксированного $\omega \in \Omega$, $f(\cdot,\omega)$ является функцией распределения $\boldsymbol{f}_\omega$.

Для распределений второго порядка можно определить арифметические операции. Пусть $\boldsymbol{f}^{(2)}, \boldsymbol{g}^{(2)}$ – распределения второго порядка, $(\Omega_f,\mathcal{F}_f,\mathbb{P}_f)$, $(\Omega_g,\mathcal{F}_g,\mathbb{P}_g)$ – соответственно их вероятностные пространства. Тогда результат операции $\boldsymbol{f}^{(2)} * \boldsymbol{g}^{(2)}$, $* \in \{+,-,\cdot,/\}$ распределение второго порядка $\boldsymbol{F}^{(2)}$:

$$F(\cdot,\omega_*) = \{\boldsymbol{f}(\cdot,\omega_f) * \boldsymbol{g}(\cdot,\omega_g) \mid (\omega_f,\omega_g) \in \Omega_*\},$$

где $(\Omega_*,\mathcal{F}_*,\mathbb{P}_*)$ есть вероятностное пространство с $\Omega_* = \Omega_f \times \Omega_g$. Далее операции над распределениями второго порядка используются для оценок надежности.

III. Надежные оценки функций распределений

В разделе рассмотрено использование распределений второго порядка для надежных оценок функций распределений.

Пусть $(\xi_1,\ldots,\xi_n)$ – выборка случайной величины X с функцией распределения $F(t), t \in [a,b]$. Эмпирическая

функция распределения определяется следующим образом

$$F_n(t) = \frac{m_t}{n},$$

где m_t число элементов $\xi_i < t$.

Пусть $z_i = F(x_i), i = 1,...,n$. Заметим что $z_i, i = 1,...,n$ – равномерно распределенные случайные величины на [0,1]. Если $z_1 \le z_2 \le ... \le z_n$, тогда z_k–k-я порядковая статистика, математическое ожидание которой [5]

$$\mathsf{E}[z_k] = k/(n+1).$$

Далее будем использовать точки $(x_i, i/(n+1))$ для построения аппроксимации функции распределения $F(t)$.

Пусть $\omega = \{a = \xi_0 < \xi_1 < \xi_2 < ... < \xi_n < b = x_{n+1}\}$ – сетка. Построим на ω кусочно-линейную функцию s

$$s(x_i) = i/(n+1), i = 1,...,n, s(a) = 0, s(b) = 1.$$

Заметим, если бы мы могли вместо математических ожиданий $i/(n+1)$ использовать точные значения z_i, тогда погрешность кусочно-линейной функции $s(x)$ на сеткеω удовлетворяла бы оценке

$$\| F - s \| \le Kh^2 \| F^{(2)} \|.$$

Таким образом, даже при относительно небольших значениях n, построенные оценки достаточно хорошо аппроксимируют функцию распределения F. Относительно z_i известно, что они образуют порядковые статистики.

Плотность вероятности k-й порядковой статистики

$$p_k(z) = \frac{n!}{(n-k)!(k-1)!} z^{k-1}(1-z)^{n-k}, z \in [0,1].$$

Совместная плотность вероятности вектора (z_j, z_k) выражается следующим образом

$$p_{j,k}(z_j, z_k) = \frac{n!}{(j-1)!(k-j-1)!(n-k)!} z_j^{j-1}(z_k - z_j)^{k-j-1}(1-z_k)^{n-k},$$

$$j < k, 0 \le z_j \le z_k \le 1.$$

Каждому случайному вектору $(z_1, z_2, ..., z_n)$ соответствует кусочно-линейная функция s. Перебирая все возможные случайные векторы $(z_1, z_2, ..., z_n)$, получаем все множество кусочно-линейных функций $\{s\}$. Заметим, что $\{s\}$ содержит кусочно-линейную функцию интерполирующую распределение F. Таким образом, используя для значения в узле ξ_k плотность вероятности k-й порядковой статистики, множество $\{s\}$ можно представить в виде случайной кусочно-линейной функции L. Соответственно, L – надежная оценка эмпирической функции распределения.

Рассмотрим пример построения кусочно-полиномиального представления для распределения второго порядка. Пусть, x_1, x_2,..., x_n, n = 9 – выборка случайной величины X с функцией распределения $F(t), t \in [0,2]$. Далее $z_i = F(x_i), i = 1,...,n$. Заметим, что $z_i, i = 1,...,n$ – равномерно распределенные случайные величины на [0,1]. Если $z_1 \le z_2 \le ... \le z_n$, тогда z_k – k-я порядковая статистика и математическое ожидание $[\mathsf{E}z_k] = k/(n+1)$.

На рис. 1 представлена функция распределения второго порядка аппроксимирующая распределения кусочно-линейных интерполяций распределений Ирвина–Холла n=3, построенная на выборке случайной величины размерности 9. Голубые линии – плотности вероятности случайной кусочно-линейной функции. Красная линия – точная функция распределения. Зеленые линии – границы 95 % доверительной области.

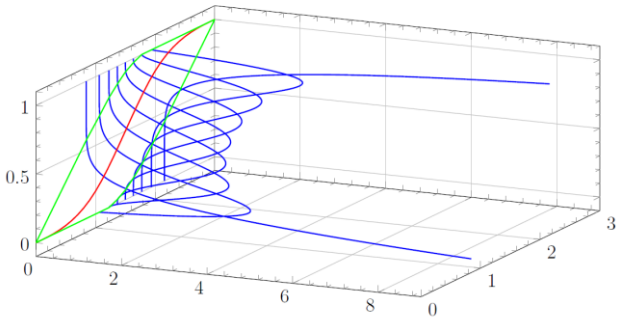

Рис. 1. Распределение второго порядка, аппроксимирующее распределение Ирвина–Холла третей степени

В математической статистике скорость сходимости эмпирической функции распределения определяется на основе теоремы Колмогорова. Пусть $\Xi = x_1, x_2,..., x_n$ есть вещественные случайные величины с функцией распределения F, F_n есть эмпирическая функция распределения, построенная на Ξ. Тогда

$$\sqrt{n} \sup_{x \in R} | F - F_n | \to K, \text{ при } n \to \infty,$$

где K – случайная величина имеющая распределение Колмогорова [2].

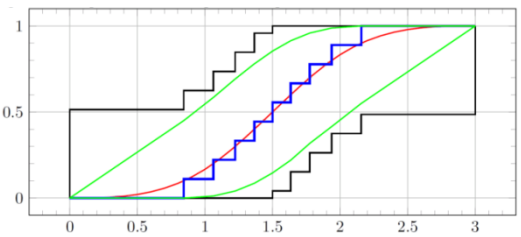

Рис. 2. Сравнение границ Колмогорова–Смирнова и надежных оценок

Красная линия – точная функция распределения. Черные линии – границы Колмогорова–Смирнова. Зеленые линии – границы 95 % доверительной области Голубая линия – эмпирическая функция распределения.

На основе этой теоремы построится интервальная функция распределения (P-box), содержащая функцию распределения F с вероятностью γ для $n \to \infty$:

$$F(x) \in F_n(x) + [-\Delta, \Delta],$$

где $\Delta = k_\gamma / \sqrt{n}$ и k_γ определяется как решение уравнения $K(k_\gamma) = \gamma$ [2].

Функция плотности вероятности есть производная от функции распределения

$$f(x) = F'(x),$$

следовательно, производная от функции распределения второго порядка будет функцией плотности вероятности второго порядка.

Производная от кусочно-линейной функции – кусочно-постоянная функция. На рис. 3 приведен пример кусочно-постоянной функции – функции плотности вероятности второго порядка (надежной оценке ф.п.в.).

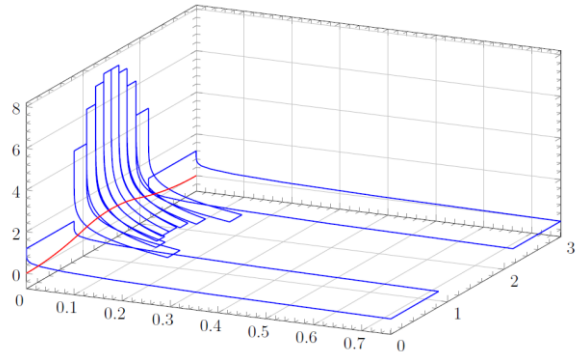

Рис. 3. Надежная оценка функции плотности вероятности

Красная линия – точная функция плотности вероятности. Синие линии – функции плотности вероятности надежной оценки, построенной на выборке случайной величины размерности 9.

IV. Параметризация распределений второго порядка

Рассмотрим построение моделей для представления распределений второго порядка. Предварительно рассмотрим задачу оценки функции среднего значения $m(x)$ и ковариационной функции $\mathrm{Cov}_a(s,t)$ случайной функции $a(x,\omega)$ которое было получено из экспериментальных данных. Имеем

$$m(x) := E[a(x,\cdot)],$$

$$\mathrm{Cov}_a(s,t) := E[(a(s,\cdot) - m(s))(a(t,\omega) - m(t))], \forall t, s$$

Теперь предположим, что мы можем наблюдать M независимых реализаций $a(x,\omega)$, записанных как $a_1(x), \ldots, a_M(x)$. Пусть наблюдаемая функция выборочного среднего $m_M^*(x)$ и выборочная ковариационная функция $S_M^*(s,t)$ для $a_1(x), \ldots, a_M(x)$ имеет вид

$$m_M^*(x) = \frac{1}{M}\sum_{i=1}^{M} a_i(x),$$

$$S_M^*(s,t) = \frac{1}{M}\sum_{i=1}^{M} (a_i(s) - m_M^*(s))(a_i(t) - m_M^*(t)).$$

Тогда мы видим, что $m_M^*(x)$ и $S_M(s,t) = [M/(M-1)]S_M^*(s,t)$ – несмещенные и непротиворечивые оценки из $m(x)$ и $\mathrm{Cov}_a(s,t)$ соответственно.

Согласно [5] случайное поле $a(x,\omega)$ может быть аппроксимировано усеченным разложением Karhunen–Loéve, имеющим форму

$$a(x,\omega) \approx a(x,\omega)_N = m(x) + \sum_{n=1}^{N} \sqrt{\lambda_n} b_n(x) \xi_n(\omega),$$

где λ_n и $b_n(x)$ для $n = 1, \ldots, N$ – собственные значения и соответствующие собственные функции для ковариационной функции, и $y_n(\omega)$ для $n = 1, \ldots, N_a$ обозначают некоррелированные вещественные случайные величины.

Разложение KL также известно как правильные ортогональные разложения (POD) и функциональный анализ главных компонентов (PCA).

Сходимость $a(x,\omega) \to a(x,\omega)_N$ гарантируется следующей теоремой [5].

Теорема [5]. Пусть область $D \subset R^d$ замкнута, пусть μ – строго положительная борелевская мера на D, пусть $Cov_a(x,x')$ – непрерывная функция на $D \times D$, которая является симметричной:

$$Cov_a(x,x') = Cov_a(x',x), \forall x, x' \in D,$$

неотрицательно определенной:

$$\int_D Cov_a(x,x')v(x)v(x')dxdx' \geq 0, \forall v(x),$$

и интегрируемой с квадратом.

Тогда

$$\lim_{N\to\infty} \max_{(x,x')\in D\times D} | Cov_a(x,x') - \sum_{n=1}^{N} \sqrt{\lambda_n} b_n(x) b_n(x') | = 0.$$

Более того, ошибка монотонно уменьшается с ростом числа слагаемых в разложении.

Другой подход основан на использовании методов случайной интерполяции и в частности случайных сплайнов [6]. Далее будем использовать регрессионные эрмитовы кубические сплайны S и точки (x_i, z_i) для построения аппроксимации функции распределения. Функция штрафа при этом подбирается таким образом, что бы, сплайн был монотонным. Таким образом, распределение второго порядка S можно представить в параметризованном виде

$$S(x, f, m) = fv(x-1) + mw(x-1) + v(x-2).$$

Для нахождения функции плотности вероятности распределения второго порядка, продифференцируем *S*.

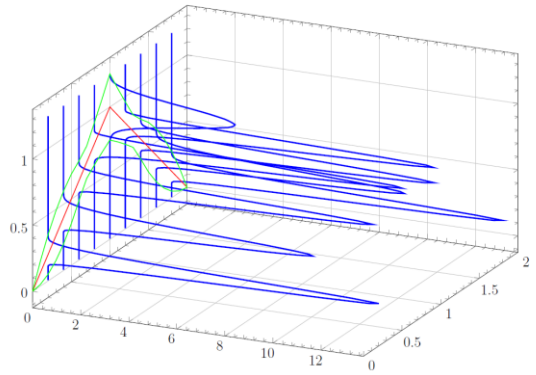

Рис. 4. Функция плотности вероятности распределения второго порядка

На рис. 4 представлена функция плотности вероятности второго порядка. Красная линия – математическое ожидание, зеленые линии – верхние и нижние границы 5 % и 95 % квантилей, синие линии – функции плотности вероятности при фиксированных значениях аргумента.

В силу того, что распределения второго порядка мы представляем в параметризованном виде, для операций над распределениями второго порядка воспользуемся техникой вероятностных расширений [7].

Таким образом, арифметические операции можно представить как вероятностные расширения соответствующих функций от случайных параметров [7].

V. Достоверные оценки надежности

В качестве примера использования распределений второго порядка рассмотрим построение достоверных оценок показателей надежности оборудования в условиях малых выборок.

Вероятность безотказной работы $P(t)$ – это вероятность того, что в течение указанного времени работы не произойдет отказа. Время работы – это продолжительность, или объем работы. Частота отказов – это мера отказов за единицу времени. Частота отказов зависит от распределения отказов, которое представляет собой совокупную функцию распределения, которая описывает вероятность отказов до момента времени t

$$\lambda(t) = -\frac{P'(t)}{P(t)}.$$

Пусть $(\xi_1, \xi_2, \ldots, \xi_n)$ – статистика отказов, полученная опытным путем.

Тогда

$$-\ln(z_i) = \int_0^{\xi_i} \lambda(\xi) d\xi,$$

где $z_i = P(\xi_i)$. Для нахождения $\lambda(t)$ будем использовать технику описанную выше. Представив $\lambda(t)$ в виде случайной кусочно-полиномиальной функции, например, случайного эрмитового сплайна, можно вычислить случайную функцию $P(t)$ для некоторого момента t. Таким образом, для фиксированного t функция $P(t)$ есть функция плотности вероятности. Можем, например, оценить риск того, что вероятность $P(t) > p_1$ или $P(t) < p_2$.

Применение кусочно-полиномиальных моделей распределений второго порядка позволяет строить надежные оценки эмпирических функций распределений. Рассмотренные примеры численных операций над распределениями второго порядка в задачах построения оценок надежности оборудования в условиях малых выборок подтверждают этот вывод. Дальнейшее использование распределений второго порядка может быть направлено на оценки рисков, принятие решений и стохастическое моделирование в условиях эпистемической неопределенности.

Список литературы

[1] Augustin T., Coolen F., Cooman G., Troffaes M. Introduction to Imprecise Probabilities – John Wiley & Sons, 2014.

[2] Ferson S., Kreinovich V., Ginzburg L., Myers D.S., and Sentz K. Constructing Probability Boxes and Dempster–Shafer Structures. Sandia National Laboratories, SAND2002-4015,2003.

[3] Swiler L.P., Giunta A.A. Aleatory and Epistemic Uncertainty Quantification for Engineering Applications. Sandia Technical Report, SAND2007-2670C, 2007.

[4] Добронец Б.С., Попова О.А. Распределения второго порядка: построение, операции, приложения // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2022. № 61. С. 61-68.

[5] Добронец Б.С., Попова О.А. Вычислительный вероятностный анализ: модели и методы. Красноярск: Сибирский федеральный университет, Институт космических и информационных технологий, 2020. 236 с.

[6] Попова О.А. Применение численного вероятностного анализа в задачах интерполяции // Вычислительные технологии. 2017. Т. 22. № 2. С. 99-114.

[7] Добронец Б.С., Попова О.А. Вычислительные аспекты вероятностных расширений // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2019. № 47. С. 41-48. DOI: https://doi.org/10.25728/ubs.2020.84.6

[8] Sahlin N., Goldsmith R. The role of second-order probabilities in decision making //Advances in Psychology. Vol. 14, 1983, pp 455-467.

[9] Utkin L., Augustin T. Decision Making with Imprecise Second-Order Probabilities. Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications, Lugano, Switzerland, July 14-17, 2003 pp 547-561.

[10] Ужга-Ребров О.И. Управление неопределенностями. Часть 4. Комбинирование неопределённостей. Rezekne, 2014. 414 с.