Некоторые задачи доверительного оценивания параметров распределений экстремальных значений

Михайлов Дмитрий Андреевич, гр.622

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель: д.ф.-м.н., д. Ермаков М.С.
Рецензент: к.ф.-м.н., с.н.с. Проурзин В. А.

Предсказание редких событий

Часто редкие события имеют распределения с тяжелыми хвостами. Хвост распределения F(x) при степенном убывании:

$$\overline{F}(x) = 1 - F(x) = C \cdot x^{-\alpha}, \qquad (1)$$

где α — параметр формы. Для его оценки часто используют оценку Ципфа:

Определение

$$\hat{\alpha_Z} = \frac{\sum_{i=1}^{k} \log(\frac{n}{i}) \cdot \log X_{(n-i+1)}}{\sum_{i=1}^{k} \log^2(\frac{n}{i})}, \qquad (2)$$

где k — число элементов, по которым считается хвост, n — размер выборки, $X_{(i)}$ — i-я порядковая статистика выборки.

Повышение точности оценивания границ доверительного интервала. Метод существенной выборки

Задача для получения доверительных оценок с использованием метода существенной выборки формулируется следующим образом. Пусть P_0 — теоретическое распределение наших данных, а $\hat{P}_n$ — моделируемое эмпирическое распределение, $T(P)$ — оцениваемый функционал.

Вероятность уклонения истинного значения статистики $T(P_0)$ от ее оценки $T(\hat{P}_n)$ не более, чем на некоторое значение $b \in \mathbb{R}$:

$$\omega = \mathbf{P}(T(\hat{P}_n) - T(P_0) > b). \tag{3}$$

Метод существенной выборки. Процедура моделирования

Для получения оценки этой вероятности выбирается абсолютно непрерывная относительно P_0 вероятностная мера Q, по которой моделируется K независимых выборок:

$$Y_1^{(k)}, Y_2^{(k)}, \ldots, Y_n^{(k)} \sim Q_n^{(k)} \tag{4}$$

где $k \in [1, K]$.
Несмещенная оценка вероятности $\hat{\omega}_n$:

$$\hat{\omega}_n = \frac{1}{K} \sum_{k=1}^{K} \chi(T(\hat{Q}_n^{(k)}) - T(P_0) > b_n) \prod_{j=1}^{n} q_n^{-1}(Y_j^{(k)}), \tag{5}$$

где $\hat{Q}_n^{(k)}$ — эмпирическое распределение выборки $Y_j^{(k)}$, $q_n = \frac{dP_0}{dQ_n}$, а $\chi(C) = 1$, если условие C соблюдается, и $\chi(C) = 0$ в обратном случае.

Метод существенной выборки. Дисперсия оценки

Дисперсия такой оценки будет иметь вид:

$$\mathbf{D}(\hat{\omega}_n) = U_n - \omega_n^2, \tag{6}$$

где $\omega_n = \mathbf{E}(\hat{\omega}_n)$, а U_n:

$$U_n = \mathbf{E}_{Q_n}\left[\chi(T(\hat{Q}_n^{(1)}) - T(P_0) > b_n)\prod_{i=1}^{n} q_n^{-2}(Y_i^{(1)})\right]. \tag{7}$$

Критерием оптимальности Q является асимптотическая эффективность (в смысле логарифмической асимптотики), то есть:

$$\overline{\lim_{n\to\infty}}\frac{\log U_n}{\log \omega_n^2} = 1. \tag{8}$$

Метод существенной выборки. Асимптотическая эффективность

При соблюдении некоторых ограничений (Ermakov, 2007), процедура существенной выборки, основанная следующей вероятностной мере будет асимптотически эффективна:

$$q_n(x) = \lambda_n + b_n \cdot h(x) \cdot \chi\left(h(x) > -\frac{\delta}{b_n}\right), \tag{9}$$

где $\lambda_n \in \mathbb{R}$, $\delta \in [0,1]$ — константы нормализации, $h(x)$ — функция влияния функционала $T(P)$.
Доверительный интервал оценки:

$$b_n = \frac{N_{0,1}^{-1}(\gamma)}{\sqrt{n}\sigma(F)}, \tag{10}$$

где $\gamma \in [0,1]$ — уровень доверия.
В работе оптимальная мера $q_n(x) = g(x) = p(x)(1 + b_n h(x))$, $p(x)$ — базовое распределение.

Цель и задачи

Цель: определение границ доверительных интервалов для оценки Ципфа на основе статистического моделирования методом существенной выборки.

Задачи:

1. найти явный вид функции влияния $h(x)$;
2. разработать алгоритм моделирования случайных величин с плотностью распределения $g(x)$;
3. осуществить моделирование и найти границы оценки малых вероятностей;
4. исследовать зависимость оценки от параметров, которые задаются перед моделированием.

Метод существенной выборки. Общий вид оцениваемого функционала и его функции влияния

$$T(F_y) = \frac{\sum_{i=1}^{k} \log(\frac{n}{i}) \cdot Y_{(n-i+1)}}{\sum_{i=1}^{k} \log^2(\frac{n}{i})}, \tag{11}$$

где $Y = \log(X)$, $Y \sim \mathrm{Exp}(\alpha)$.

Общий вид таких функционалов (Serfling, 1980):

$$T(F) = \int\limits_0^1 F^{-1}(t)J(t)\,\mathrm{d}t + \sum_{j=1}^{l} a_j F^{-1}(p_j). \tag{12}$$

где $J(t)$ — весовая функция, p_j — уровни квантилей $F^{-1}(p_j)$, a_j – соответствующие им веса.

$$h(x) = -\int\limits_{-\infty}^{+\infty} [\chi(x \leq y) - F(y)]J(F(y))dy + \sum_{j=1}^{l} a_j \frac{p_j - \chi(x \leq F^{-1}(p_j))}{f(F^{-1}(p_{ij}))}$$

Метод существенной выборки. Итоговый вид функции влияния

При $x_1 = F^{-1}(\beta)$, где β — доля выборки, которая не участвует в оценке, получаем:

$$h(x) = \begin{cases} -h_0, & \text{если } x \le x_1; \\ \frac{\alpha(x^2 - x_1^2)}{2 \cdot \beta_{const}} - h_0 & \text{иначе}, \end{cases} \tag{13}$$

где:

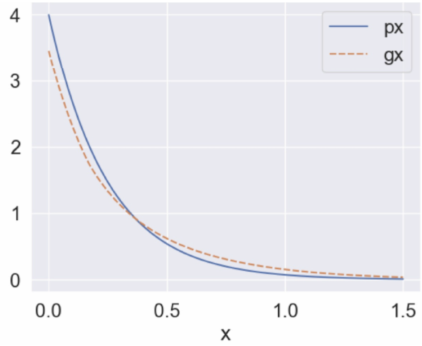
$$\beta_{const} = (1-\beta)\,\log^2(1-\beta) + (2\,\beta - 2)\,\log(1-\beta) - 2\,\beta + 2, \tag{14}$$

$$h_0 = \frac{(\alpha x_1 + 1)(1-\beta)}{\alpha \cdot \beta_{const}}. \tag{15}$$

Метод существенной выборки. Моделируемая плотность

Подставим [13] в определение $g(x)$:

$$g(x) = p(x) \cdot \left(1 + b_n \cdot \left[\frac{\alpha(x^2 - x_1^2) \cdot \chi(x > x_1)}{2 \cdot \beta_{const}} - h_0\right]\right). \quad (16)$$



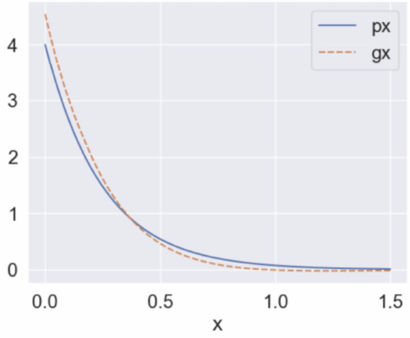


Рис. 1: График $g(x)$ и $p(x)$ при $b_n > 0$ (слева) и $b_n < 0$ (справа)

Моделируемая плотность при $b_n > 0$ и $x \leq x_1$

Будем называть такую область A. Плотность $g(x)$ принимает следующий вид:

$$g(x) = p(x) \cdot \left(1 - b_n \cdot h_0\right).$$

Распределение в этом случае экспоненциальное, умноженное на константу $1 - b_n \cdot h_0$. Его можно смоделировать с помощью метода обратной функции.

Моделируемая плотность при $b_n > 0$ и $x > x_1$

Будем называть такую область B. Плотность $g(x)$ принимает следующий вид:

$$g(x) = p(x) \cdot \left(1 + b_n \cdot \left[\frac{\alpha(x^2 - x_1^2)}{2 \cdot \beta_{const}} - h_0\right]\right).$$

Такое распределение уже нельзя будет промоделировать предыдущим методом из-за наличия x в показателе степени и в квадратной функции. Метод композиции также нельзя применять в данном случае, так как $-h_0$ является отрицательным коэффициентом.

Приведение моделируемой плотности к случаю смеси двух распределений

$$g(x) = \alpha e^{-\alpha x}\left(1 - \frac{\alpha \cdot b_n \cdot x_1^2}{2 \cdot \beta_{const}} - b_n \cdot h_0\right) + \frac{\alpha^2 \cdot b_n \cdot x^2 \cdot e^{-\alpha x}}{2 \cdot \beta_{const}}. \quad (17)$$

Нам необходимо произвести моделирование случайной величины $\xi \sim g(x)$, которая является суммой двух других случайных величин:

$$\xi = \theta + \eta. \quad (18)$$

По первой части уравнения [17] заметим, что $\theta \sim \left(1 - \frac{\alpha \cdot b_n \cdot x_1^2}{2 \cdot \beta_{const}} - b_n \cdot h_0\right) \cdot \mathrm{Exp}(\alpha)$. Вторая часть с помощью алгебраических преобразований сводится к гамма-распределению: $\eta \sim \frac{\Gamma(3) \cdot b_n}{2 \cdot \alpha \cdot \beta_{const}} \cdot \Gamma(3, \alpha)$.

Моделируемая плотность при $b_n > 0$

Оценка при заданных условиях:

$$\hat{\omega}_n = \frac{1}{K} \sum_{i=1}^{K} \left[\chi\big(T(\hat{Q}_i) - T(P_0)\big) > b_n\right] \cdot \prod_{j=1}^{n} \frac{1}{1 + b_n \cdot h(Y_j^{(i)})}$$

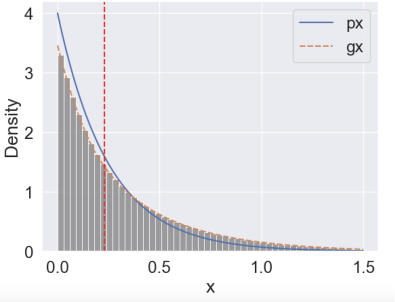


Рис. 2: Результат моделирования $g(x)$ при $b_n > 0$

Моделируемая плотность при $b_n < 0$ и $x \leq x_1$

Будем называть такую область C. Плотность $g(x)$ принимает следующий вид:

$$g(x) = p(x) \cdot \Big(1 + b_n \cdot h_0\Big).$$

$g(x)$ в этом случае представляет собой экспоненциальное распределение, умноженное на константу $1 + b_n \cdot h_0$. Его можно смоделировать с помощью метода обратной функции.

Моделируемая плотность при $b_n < 0$ и $x > x_1$

Будем называть такую область D. Плотность $g(x)$ принимает следующий вид:

$$g(x) = p(x) \cdot \left(1 - b_n \cdot \left[\frac{\alpha(x^2 - x_1^2)}{2 \cdot \beta_{const}} - h_0\right]\right).$$

В данном случае воспользуемся методом мажорант. Пусть:

$$D_D = \frac{\int\limits_{x_1}^{\infty} p(x)dx}{\int\limits_{x_1}^{\infty} g(x)dx}.$$

Тогда алгоритм моделирования следующий:

1. Моделируем $\xi \sim \frac{p(x)}{|D_D|}$;
2. Моделируем $\theta \sim U(0, p(\xi))$;
3. Если $\theta > g(\xi)$, начинаем итерацию заново, иначе $\xi \sim g(x)$;

Моделируемая плотность при $b_n < 0$

При $b_n < 0$ формула оценки немного меняется:

$$\hat{\omega}_n = \frac{1}{K}\sum_{i=1}^{K}\left[\chi\left(T(\hat{Q}_i) - T(P_0)\right) < -b_n\right] \cdot \prod_{j=1}^{n} \frac{1}{1 - b_n \cdot h(Y_j^{(i)})} \quad (19)$$

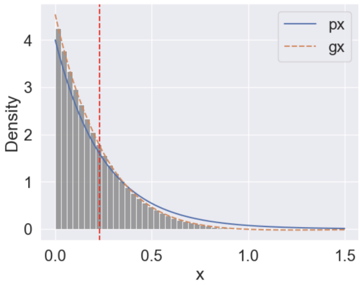


Рис. 3: Результат моделирования $g(x)$ при $b_n < 0$

Зависимость оценки вероятности уклонения $\hat{\omega}_n$ от уровня значимости $1-\gamma$

Мы моделируем 2 оценки — для $b_n > 0$ и для $b_n < 0$.

$$b_n = \frac{N_{0,1}^{-1}(\gamma)}{\sqrt{n}\sigma(F)}, \gamma \in [0, 1].$$

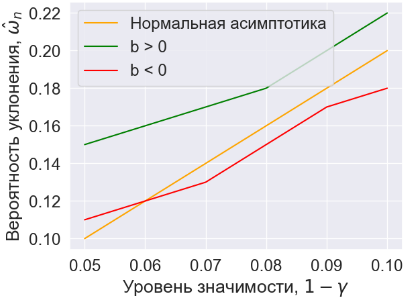


Рис. 4: Зависимость $\hat{\omega}_n$ от уровня значимости. Параметры моделирования: $n = 1000, K = 50, \beta = 0.8, \alpha = 4$

Влияние величины уклонения b_n на оценку вероятности уклонения $\hat{\omega}_n$

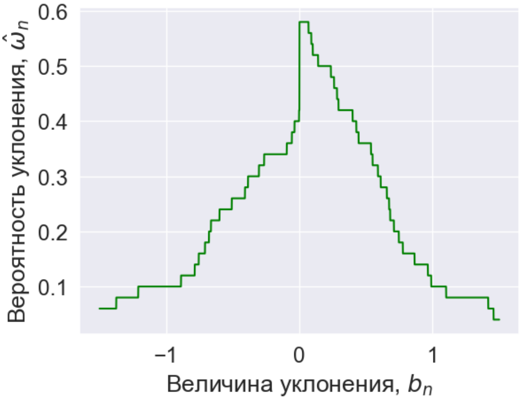


Рис. 5: Зависимость $\hat{\omega}_n$ от b_n. Параметры моделирования: $n = 1000, K = 50, \beta = 0.8, \alpha = 4$

Сравнение эффективности

Прямая оценка вероятности уклонения:

$$\hat{\omega}_D = \frac{1}{K}\sum_{i=1}^{K}\chi\left[T(\hat{P}_n) - T(\hat{P}_0) > b_n\right],$$

где $P = \mathrm{Exp}(\alpha)$. Дисперсия оценки:

$$\mathbf{D}\hat{\omega}_D = \sqrt{\frac{\hat{\omega}_D \cdot (1 - \hat{\omega}_D)}{K}}$$

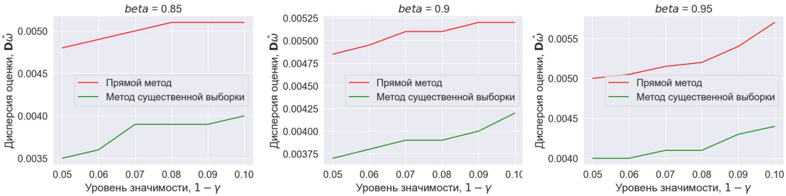


Рис. 6: Зависимость дисперсии оценки от уровня значимости, Параметры моделирования: $n = 1000, K = 50, \alpha = 4$.

Заключение

Полученные результаты:

1. Найден общий вид вспомогательной плотности распределения $g(x)$ для оценки Ципфа;
2. Произведено моделирование выборки из распределения $g(x)$ с помощью различных статистических процедур;
3. Был применен метод существенной выборки для оценивания доверительных интервалов в задачах оценки тяжести хвоста распределений;
4. Исследовано влияние параметров вспомогательного распределения на оценку вероятности уклонения.