Санкт-Петербургский государственный университет

Прикладная математика и информатика

Статистическое моделирование

Старков Артём Константинович

Доверительное оценивание статистик экстремальных значений

Отчет о научно-исследовательской работе

Научный руководитель:

д. ф.-м. н., профессор М. С. Ермаков

Рецензент:

д. ф.-м. н., профессор Г. Л. Шевляков

Санкт-Петербург

2018

Оглавление

Введение

События, имеющие малую вероятность, встречаются во многих отраслях человеческой деятельности, таких как строительство, страхование, менеджмент, финансы и др. При этом часто возникают ситуации, когда наблюдения, описывающиеся этими событиями, принимают существенно большие значения, чем остальные наблюдения в выборке. Это порождает задачи оценки тяжести хвоста распределения.

Существует множество работ, вводящих различные характеристики для оценки показателя степени при степенном убывании распределения на хвосте, такие как оценка Хилла [1], момент-оценки Dekkers [2], Pickands estimator [3] и другие. В статистике экстремальных значений показатель степени убывания хвоста распределения называется параметром формы (shape). Наиболее изученной и широко используемой оценкой этого параметра является оценка Хилла [1].

Оценка Хилла строится на основе наблюдения редких событий, и в условиях ограниченности размера выборки это особенно подчеркивает необходимость построения доверительного интервала вследствие ее неустойчивости. Построение доверительного интервала имеет малый уровень значимости, и его вычисление может быть достаточно ресурсоемким. В таких случаях могут использоваться различные методы оптимизации для построения доверительных интервалов. Одним из таких методов является метод существенной выборки. Он позволяет уменьшить дисперсию случайной величины при моделировании методом Монте-Карло, путем замены исходной вероятностной меры на смещенное распределение, выделяющее области с более высокой вероятностью для данного конкретного алгоритма.

Данная работа посвящена построению доверительных интервалов для оценки Хилла путем статистического моделирования по методу существенной выборки. Работа разделена на теоретическую и практическую части. В теоретической части будет рассмотрен оцениваемый параметр формы предельного распределения с использованием статистики оценки Хилла; описаны возможности использования метода существенной выборки в задачах оценки малых вероятностей и приведен общий вид вероятностной меры, на основе которой осуществляется эффективное моделирование методом существенной выборки.

В практической части будут продемонстрированы расчеты, связанные с поиском

эффективной меры моделирования, описан алгоритм моделирования случайной величины, распределенной по этой мере. Будут приведены результаты моделирования доверительных интервалов и оценена эффективность примененного метода в сравнении с прямой оценкой.

Глава 1

Теоретическая часть

1.1. Предельные распределения

В вопросах оценивания вероятностей больших уклонений весьма важную роль играет вопрос о распределении максимумов. Например, пусть дана последовательность независимых одинаково распределенных случайных величин $(X_i)_{i=1}^n$. Будем называть M_n их выборочным максимумом:

$$M_1 = X_1,\ M_n = \max(X_1, ..., X_n).$$

Одним из результатов изучения распределения M_n является теория предельных распределений. Важнейший результат этой теории заключен в теореме, представленной ниже.

Теорема 1 (Фишера-Типпета-Гнеденко [4, стр. 121]). *Пусть $(X_i)_{i=1}^n$ — последовательность независимых, одинаково распределенных случайных величин, M_n — их выборочный максимум $M_n = \max(X_1, ..., X_n)$. Если существуют такие $c_n > 0, d_n \in \mathbb{R}$ и некоторая плотность распределения H, что выполняется*

$$\frac{(M_n - d_n)}{c_n} \xrightarrow{p} H,$$

тогда H принадлежит одному из семейств предельных распределений:

Фреше: $\Phi_\alpha(x) = e^{-x^{-\alpha}}, \quad x > 0, \alpha > 0;$

Вейбулл: $\Psi_\alpha(x) = e^{-(-x)^\alpha}, \quad x \leq 0, \alpha > 0;$

Гумбель: $\Lambda(x) = e^{-e^{-x}}, \quad x \in \mathbb{R}.$

Очевидно, что вид предельного распределения напрямую зависит от исходного распределения X_i.

Определение 1. *Будем говорить, что если некоторая последовательность $(X_i)_{i=1}^n$ с функцией распределения $F(x)$ имеет предельное распределение H, то она принадлежит его области максимального притяжения (maximum domain attraction), и записывать:*

$$F \in \mathsf{MDA}(H).$$

В данной работе рассматривается предельное распределение Фреше. Оно соответствует семействам распределений Парето, Коши, Булла и др.

Распределение Фреше имеет параметры shape $\alpha > 0$, scale $s > 0$ и location $m < x$:

$$F(x) = e^{-(\frac{x-m}{s})^{-\alpha}},$$

$$p(x) = \frac{\alpha}{s}\left(\frac{x-m}{s}\right)^{-1-\alpha} e^{-(\frac{x-m}{s})^{-\alpha}}.$$

Теорема 2 ([4, стр. 131]). *Функция распределения $F(x)$ принадлежит области максимального притяжения распределения Фреше $F \in \mathsf{MDA}(\Phi_\alpha)$ для $\alpha > 0$, если и только если $\bar{F}(x) = x^{-\alpha}L(x)$, для некоторой медленно меняющейся функции L, где $\bar{F}(x) = 1 - F(x)$ — распределение хвоста $F(x)$.*

На основании этого факта Хиллом была получена оценка для α, совпадающая с оценкой максимального правдоподобия [1]:

$$\hat{\alpha}_H = \left(\frac{1}{k}\sum_{i=1}^{k} \ln X_{i:n} - \ln X_{k:n}\right)^{-1}, \tag{1.1}$$

где $k = k(n) \to \infty$. Это оценка т.н. «нижнего хвоста» (*lower tail estimate*), в той же статье делается оценка и «верхнего хвоста» (*upper tail estimate*), т. е. для распределения хвоста на интервале $[\beta, 1)$. Она идентична (1.1) и делается путем замены $X = Y^{-1}$; соответственно с учетом смены направления упорядочивания вариационного ряда получаем

$$Y_{i:n} = X_{n-i+1:n},$$

$$\left(\frac{1}{n-m+1}\sum_{i=m}^{n} \ln X_{i:n} - \ln X_{m:n}\right)^{-1}, \tag{1.2}$$

где $m = n - k + 1 : m = [\beta n]$.

Для многих прикладных задач параметр α имеет большое значение. Он используется в качестве индикатора тяжести хвоста распределения. Кроме оценки Хилла, существует большое количество других статистик, оценивающих тяжесть хвоста, например оценка Пиканда [3], qq-estimator [5], оценка гармонического момента [6], оценка для зависимых гетерогенных даных [7], а так же различные вариации на основе оценки (1.1) (Resnick 1997 [8], Gomes and Martins 2001 [9] и др.).

В условиях ограниченности выборки возникает задача определения качества полученной оценки. Так как оценка строится в условиях $k = k(n) \to \infty$, то любая попытка

определения качества оценки затруднительна ввиду необходимости получения большого количества реализаций $n \to \infty$ при том, что действительно полезных из них будет относительно немного (в зависимости от параметра β). В решении этой проблемы в задачах статистического моделирования может помочь асимптотическая эффективность метода существенной выборки.

1.2. Метод существенной выборки

Определение 2. *Метод существенной выборки (выборка по значимости, importance sampling) – общая техника оценивания свойств частичных распределений, при которой осуществляется замена вероятностной меры.*

Пусть поставлена задача вычисления следующей вероятности:

$$\omega = P(T(\hat{P}_n) - T(P_0) > b), b \in R, \tag{1.3}$$

где P_0 – теоретическое распределение; $\hat{P}_n$ – эмпирическая функция распределения; $T(P)$ – некоторый функционал. Такая задача соответствует подсчету вероятности уклонения оценки статитики, построенной по выборке $T(\hat{P}_n)$ от ее истинного значения $T(P_0)$. Согласно методу, необходимо выбрать меру Q, такую, что Q абсолютно непрерывна к P_0. Моделируются K независимых выборок с распределением Q:

$$Y_1^{(k)}, Y_2^{(k)}, ..., Y_n^{(k)}, 1 \le k \le K.$$

Оценка вероятности (1.3):

$$\hat{\omega}_n = \frac{1}{K} \sum_{k=1}^{K} \chi\big(T(\hat{Q}_n^{(k)}) - T(P_0) > b_n\big) \prod_{j=1}^{n} q_n^{-1}(Y_j^{(k)}), \tag{1.4}$$

где $\hat{Q}_n^{(k)}$ – эмпирическое распределение выборки $Y_j^{(k)}$, $q_n = \frac{dQ_n}{dP_0}$.

Дисперсия оценки (1.4) имеет вид

$$V(Q_n) = \mathsf{Var}[\hat{\omega}_n] = U_n - \omega_n^2, \tag{1.5}$$

где ω_n – математическое ожидание (1.4),

$$U_n = \mathsf{E}_{Q_n} \left[\chi\big(T(\hat{Q}_n^{(1)}) - T(P) > b_n\big) \prod_{j=1}^{n} q_n^{-2}(Y_j^{(1)}) \right],$$

Так как $U_n \omega_n$ малые величины, то естественно рассматривать асимптотическую эффективность процедуры существенной выборки.

Определение 3. *Процедура называется асимптотически эффективной (в смысле логарифмической асимптотики), если*

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

Основной трудностью при использовании метода существенной выборки является выбор меры Q_n, при котором процедура асимптотически эффективна или близка к такой.

Предпосылкой для решения задачи о выборе меры Q_n стала работа математика И. Н. Санова [10]. Им была исследована задача вычисления вероятности

$$P(\hat{P}_n \in \Omega), \tag{1.6}$$

где $\hat{P}_n$ – эмпирическая функция распределения, построенная по выборке $X_i \sim P_0, 1 \leq i \leq n$; Ω принадлежит пространству мер. Он исследовал асимптотическое поведение вероятности (1.6) и получил следующие результаты:

$$\begin{aligned} \overline{\lim_{n\to\infty}}\ \tfrac{1}{n}\ln P(\hat{P}_n \in \Omega) &\leq -\inf_{H\in \mathfrak{int}(\Omega)} K(H, P_0), \\ \underline{\lim_{n\to\infty}}\ \tfrac{1}{n}\ln P(\hat{P}_n \in \Omega) &\geq -\inf_{H\in \mathfrak{cl}(\Omega)} K(H, P_0), \end{aligned} \tag{1.7}$$

где $\mathfrak{int}(\Omega)$ и $\mathfrak{cl}(\Omega)$ – соответственно замыкание и внутренность Ω в специальной топологии со слабой сходимостью,

$$K(H, P_0) = \begin{cases} \int \ln \frac{dH}{dP_0} dH, & \text{если } H \ll P_0; \\ \infty, & \text{иначе.} \end{cases} \tag{1.8}$$

На основе этих результатов в дальнейшем была развита теория вероятности больших уклонений, окончательно формализованная С. Р. Варадханом. В рамках этой теории неравенства типа (1.7) стали называться принципом больших уклонений, а функционалы типа (1.8) – функционалами действия. Они являются в некотором смысле мерой близости распределений H и P_0, схожей с расстоянием Кульбака-Лейблера.

Рассмотрим задачу (1.3) в зоне больших уклонений:

$$P(T(\hat{P}_n) - T(P_0) \in \Omega), \Omega \in \mathbb{R}.$$

На практике для такой задачи нахождение H часто является неразрешимой задачей. В зоне умеренных уклонений и действия ЦПТ для задачи в виде

$$P(T(\hat{P}_n) - T(P_0) > b_n), \tag{1.9}$$

где $nb_n^2 \to \infty, b_n \to 0, b_n > 0$, действует нормальная асимптотика, в смысле логарифма аппроксимации. Здесь задача нахождения H разрешима для очень широкого класса функционалов T.

М. Джонсоном была построена асимптотически эффективная процедура метода существенной выборки в зоне нормальной аппроксимации. На зону умеренных уклонений его результаты были перенесены М. С. Ермаковым в статье [11]. В статье рассматривается задача (1.9) в случае, когда статистика $T(\hat{P}_n)$ имеет функцию влияния g и верны предположения, рассматриваемые ниже.

Введем следующие множества:

1. множество функций Φ такое, что для любой $f \in \Phi$ $\mathsf{E}f(x) = 0$ и
$$\lim_{n\to\infty} \frac{1}{nb_n^2} \log\left(nP_0(|f(X)| > nb_n)\right) = -\infty;$$
2. множество Λ_Φ всех мер $Q \in \Lambda$ таких, что для любой $f \in \Phi$ $\int\limits_\Omega |f| dQ < \infty$;
3. множество $\Lambda_{0\Phi}$ всех зарядов $G = P - R; P, R \in \Lambda$.

Тогда предположим, что $g \in \Phi$ и существует полунорма $N \in \Lambda_{0\Phi}$, такая, что для любого $Q \in \Lambda_\Phi$
$$|T(Q) - T(P_0) - \int\limits_\Omega g dQ| \le \omega\left(N(Q - P_0),\ \int\limits_\Omega g dQ,\ T(Q) - T(P_0)\right)$$
с функцией $\omega\colon \mathbb{R}^3 \to \mathbb{R}_+$, такой, что
$$\lim_{t_1,t_2,t_3\to 0} \frac{\omega(t_1,t_2,t_3)}{t_1+t_2+t_3} = 0.$$

М. С. Ермаковым была доказана следующая теорема.

Теорема 3 ([11]). *В предположениях, изложенных выше, рассмотрим процедуру существенной выборки, основанную на вероятностной мере Q_n с плотностью*
$$q_{1n}(x) = \lambda_n + b_n h(x)\chi\left(h(x) > -\frac{\delta}{b_n}\right)$$
или
$$q_{2n}(x) = c_n e^{b_n h(x)}\chi\left(h(x) < \frac{\delta}{b_n}\right),$$
где λ_n, c_n – константы нормализации, $\delta \in [0,1]$ и $\mathsf{E}h(x) = 0, \mathsf{E}|h(x)| < \infty, \mathsf{E}h^2(x) < \infty$. Тогда процедура существенной выборки асимптотически эффективна, если $h = \frac{g}{\sigma_g^2}$

1.3. Постановка задачи

Пусть $(X_i)_{i=1}^n$ — независимые случайные величины с функцией распределения $F \in \mathsf{MDA}(\Phi_\alpha)$. Упорядочим выборку по возрастанию, получим вариационный ряд:

$$X_{1:n} \leq X_{2:n} \leq ... \leq X_{n:n}.$$

Распределение хвоста $\bar{F}(x) = 1 - F(x)$ принадлежит семейству предельных распределений Фреше Φ_α, в котором неопределенным остается параметр формы (shape) α. Имеем для него оценку Хилла:

$$\hat{\alpha}_H = \left(\frac{1}{n-m+1} \sum_{i=m}^{n} \ln X_{i:n} - \ln X_{m:n} \right)^{-1}, \tag{1.10}$$

где $m = [\beta n]$.

Рассмотрим задачу оценки вероятности уклонения оцененного значения от истинного, представленной в формуле (1.3), где в качестве функционала $T(F)$ будет выступать оценка $\hat{\alpha}_H$. Так как сама оценка строится в условиях $m = m(n) \to \infty$, то обычные оценки становятся неэффективными. Для решения данной проблемы будет использована асимптотическая эффективность метода существенной выборки.

По теореме 3 будем использовать вероятностную меру Q_n вида

$$g(x) = p(x)(1 + b_n h(x)), \tag{1.11}$$

где $h(x)$ — функция влияния оцениваемого функционала, b_n – константа для определения ширины доверительного интервала,

$$b_n = \pm \frac{N_{(0,1)}^{-1}(\gamma)}{\sqrt{n}\sigma}, \tag{1.12}$$

$$\sigma^2(F) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} J(F(x))J(F(y))\big(F(x \wedge y) - F(x)F(y)\big) \tag{1.13}$$

— асимптотическая дисперсия функционала $T(F)$.

Целью данной работы является построение доверительного интервала вероятности вида (1.3) для функционала оценки Хилла. Для ее достижения необходимо:

1. определить внешний вид функционала $T(P)$, его функцию влияния и дисперсию (1.13);

2. определить и промоделировать вспомогательное распределение Q_n (1.11);

3. на основе модельных данных построить доверительный интервал для вероятности (1.4);

4. оценить вычислительную эффективность данной оценки по сравнению с прямым методом.

Глава 2

Моделирование

2.1. Оцениваемый функционал

В данной работе рассматривается оценка Хилла верхнего хвоста предельного распределения (1.10). Используя теорему 2, и при допущении $L(x) = const$ сделаем замену:

$$F(x) = 1 - Cx^{-\alpha}|_{\ln x = y} = 1 - Ce^{-\alpha y} = exp(y; \alpha). \tag{2.1}$$

Исходя из этого, получаем, что распределение исходной случайной величины становится экспоненциальным с параметром α. Кроме этого, для удобства будем оценивать не $\hat{\alpha}_H$, а $\hat{\alpha}_H^{-1}$. С учетом данных замечаний, функционал оценки принимает вид:

$$T(F_y) = \hat{\alpha}_H^{-1} = \frac{1}{n-m+1}\sum_{i=m}^{n} Y_{i:n} - Y_{m:n} \tag{2.2}$$

Общий вид таких функционалов для L-оценок [12, стр. 263]:

$$T(F) = \int_0^1 F^{-1}(t)J(t)dt + \sum_{j=1}^{l} a_j F^{-1}(p_j), \tag{2.3}$$

где $J(t)$ — некоторая интегрируемая весовая функция, p_j — уровни квантилей $F^{-1}(p_j)$, a_j – соответствующие им веса. Там же предложен общий вид функции влияния:

$$IC(x; T, F) = -\int_{-\infty}^{\infty} [\chi(x \le y) - F(y)]J(F(y))dy + \sum_{j=1}^{l} a_j \frac{p_j - \chi(x \le F^{-1}(p_j))}{f(F^{-1}(p_j))}.$$

Перепишем (2.2) с учетом формы (2.3):

$$T(F_y) = \frac{1}{n-m+1}\sum_{i=m}^{n} Y_{i:n} - Y_{m:n} = \frac{1}{n-m+1}\sum_{i=m}^{n} F_y^{-1}(\frac{i}{n}) - F_Y^{-1}(\beta) =$$

$$= \frac{1}{1-\beta}\int_{\beta}^{1} F_y^{-1}(t)dt - F_y^{-1}(\beta), \tag{2.4}$$

$$J(t) = \frac{\chi(t > \beta)}{1-\beta}. \tag{2.5}$$

Соответственно функция влияния принимает вид:

$$h(x) = IC(x; T, F_y) = \frac{1}{\beta - 1}\int_{F_y^{-1}(\beta)}^{\infty} (\chi(x \le t) - F_y(t))dt - \frac{\beta - \chi(x \le F_y^{-1}(\beta))}{f_y(F_y^{-1}(\beta))};$$

Обозначив $F_y^{-1}(\beta) = x_1$, получаем:

$$h(x) = \begin{cases} \frac{1}{\beta-1}\int\limits_{x_1}^{\infty}(1-F_y(t))dt - \frac{\beta-1}{f_y(x_1)}, & \text{если } x \le x_1; \\ \frac{1}{1-\beta}\int\limits_{x_1}^{x} F_y(t)dt + \frac{1}{\beta-1}\int\limits_{x}^{\infty}(1-F_y(t))dt - \frac{\beta}{f_y(x_1)}, & \text{иначе.} \end{cases} \tag{2.6}$$

Используя (2.1), подставим функцию распределения как экспоненциальную с параметром α:

$$h(x) = \begin{cases} \frac{1}{\alpha(\beta-1)}\big(e^{-\alpha x_1} + \beta - 1\big), & \text{если } x \le x_1; \\ \frac{1}{\alpha(1-\beta)}\big(\alpha(x-x_1) - e^{-\alpha x_1} - \beta\big), & \text{иначе.} \end{cases}$$

Заметим, что

$$e^{-\alpha x_1} = e^{-\alpha F^{-1}(\beta)} = e^{-\alpha\left(-\frac{1}{\alpha}\ln(1-\beta)\right)} = e^{\ln(1-\beta)} = 1-\beta.$$

Также обозначим $x_2 = h^{-1}(0) = \frac{1}{\alpha} + x_1$. С учетом этого получаем итоговую функцию для $h(x)$:

$$h(x) = \frac{x-x_2}{1-\beta}\chi(x > x_1).$$

Функция распределения (1.11) принимает вид:

$$g(x) = p(x)(1 + b_n h(x)) = \alpha e^{-\alpha x}\left(1 + b_n\frac{x-x_2}{1-\beta}\chi(x > x_1)\right). \tag{2.7}$$

Для моделирования требуется знать интеграл функции (2.7) на произвольном промежтке $[a,b] \in [x_1, \infty)$:

$$\int\limits_a^b g(x)dx = \int\limits_a^b \alpha e^{-\alpha x}(1 + b_n\frac{x-x_2}{1-\beta})dx =$$

$$= (e^{-\alpha a} - e^{-\alpha b})(1 - \frac{b_n x_2}{1-\beta}) - \frac{b_n}{\alpha(1-\beta)}\left(e^{-\alpha b}(\alpha b + 1) - e^{-\alpha a}(\alpha a + 1)\right). \tag{2.8}$$

Вывод дисперсии функционала оценки осуществляется по формуле (1.13). Он приведен в приложении А. Итоговая дисперсия (А.1):

$$\sigma^2(F) = \frac{\beta+1}{\alpha^2(1-\beta)}$$

При использовании $b_n < 0$ изменяется функция $g(x)$, потому для нее строится распределение, отличное от случая $b_n > 0$.

2.2. Случай $b_n > 0$

Решается задача вычисления оценки (1.4) в исходой форме:

$$\hat{\omega}_n = \frac{1}{k}\sum_{i=1}^{k} \chi\big(T(\hat{Q}_n^{(i)}) - T(P_0) > b_n\big) \prod_{j=1}^{n} \frac{1}{1 + b_n h(Y_j^{(i)})}, \tag{2.9}$$

Разобьем область моделирования на 3 части:

$$0 \le A < x_1 \le B \le C < \infty.$$

В каждой области функция имеет свои особенности. Внешний вид плотности распределения показан на рисунке 2.1.

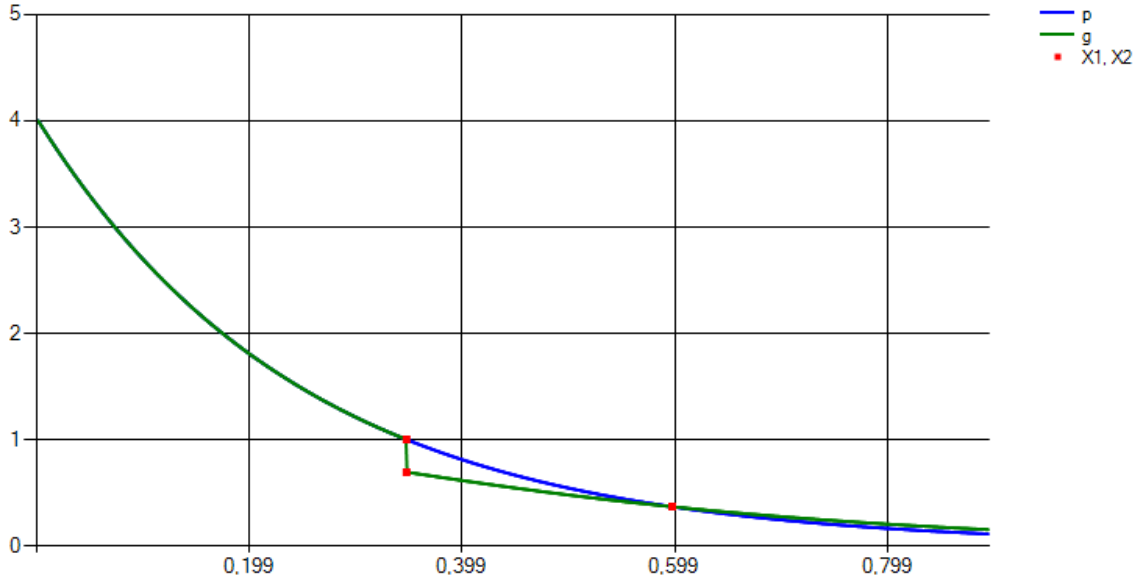

Рис. 2.1. Распределение $g(x), p(x)$ и точки x_1, x_2

Область $A = [0, x_1)$

Область A не отличается от исходного показательного распределения. Моделирование проводится методом обратной функции. Интеграл $\int_A g(x)dx = \beta$.

Область $B = [x_1, x_2)$

Моделирование проходит по алгоритму метода мажорант [13]:

Алгоритм 1 (Метод мажорант)**.** *Пусть*

$$D_S = \frac{\int_S p(x)dx}{\int_S g(x)dx},$$

где $g(x)$ *— моделируемая плотность,* $p(x)$ *— некоторая другая плотность, такая, что* $\forall x \in S\ p(x) \geq g(x)$*. Тогда:*

1. *получим реализацию* $\xi \sim \frac{p(X)}{|D|}$*;*
2. *получим реализацию* $\eta_{s+1} \sim U(0, p(\xi))$*;*
3. *если* $\eta_{s+1} > g(\xi)$*, переходим к пункту 1; иначе* ξ *— искомая случайная величина,* $\xi \sim g(x)$*.*

Эффективность метода равна $\frac{1}{|D_S|}$*.*

Эффективность метода:

$$\frac{1}{|D_S|} = 1 - \frac{b_n}{\alpha(e-1)(1-\beta)}.$$

Область $C = [x_2, \infty)$

В области используем метод композиции [13].

Алгоритм 2 (Метод композиции)**.** *Пусть искомая плотность* $g(x)$ *представлена суммой плотностей* $p_i(x)$ *с некоторыми неотрицательными весами* a_i*:*

$$g(x) = \sum_{i=1}^{M} a_i p_i(x). \tag{2.10}$$

Тогда метод выражается следующим алгоритмом:

1. *получаем реализацию* $\xi \sim U(0, \sum_i a_i)$*;*
2. *выбираем индекс* $s = \arg\max(a_i : a_i < \xi)$*;*
3. *искомая случайная величина —* $\eta \sim p_s(x)$*.*

Представим функцию в виде:

$$g(x) = p(x)(1 + b_n h(x)) = \alpha e^{-\alpha x} + \frac{b_n(x - x_2)}{1-\beta}\alpha e^{-\alpha x} = p(x) + p_1(x).$$

Заменим во второй части $y = x - x_2$:

$$p_1(y) = \frac{b_n y}{1-\beta}\alpha e^{-\alpha(y+x_2)} = \frac{b_n}{\alpha(1-\beta)}\alpha^2 y e^{-\alpha y}e^{-\alpha x_2} = \frac{b_n}{\alpha e}\alpha^2 y e^{-\alpha y}.$$

Теперь $p_1(y)$ выражается гамма-распределением с параметрами $shape = 2, scale = \frac{1}{\alpha}$. Получаем форму (2.10) с весами $a_1 = 1, a_2 = \frac{b_n}{\alpha e}$ и распределениями

$$p_0(x) = p(x) = exp(x, \alpha),$$
$$p_1(x) = \gamma(x, 2, \frac{1}{\alpha}).$$

Общее распределение $g(x), b_n > 0$

Моделирование проводится при помощи алгоритма 2. В качестве весов берутся площади плотности $g(x)$ в соответствующих областях. Пример результата моделирования распределения в виде гистограммы с наложенными функциями $g(x), p(x)$ представлен на рисунке 2.2.

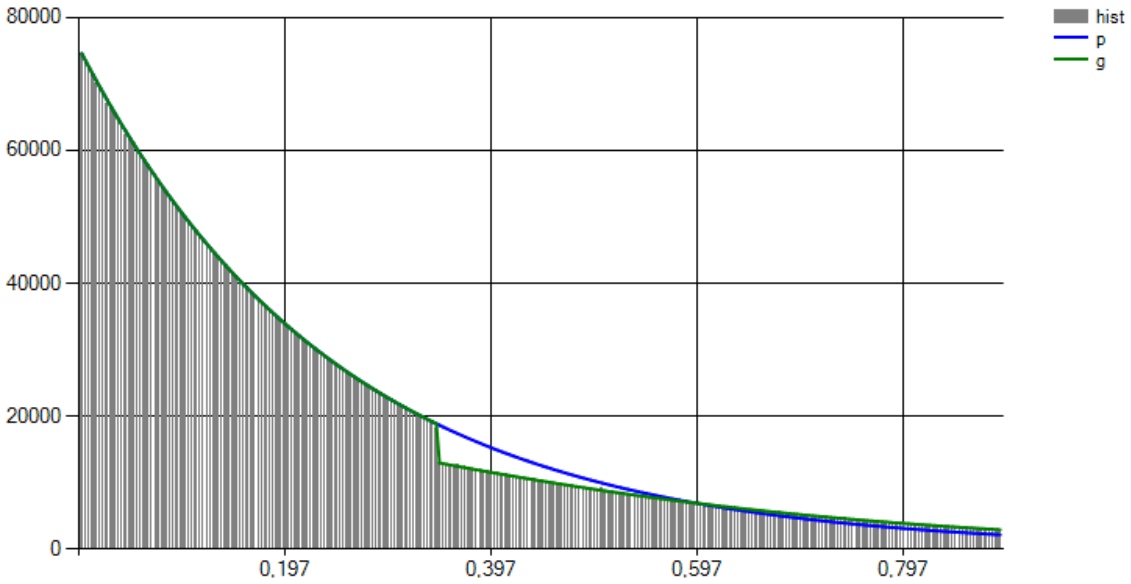

Рис. 2.2. Распределение $g(x), p(x)$

2.3. Случай $b_n < 0$

Для данного случая изменен вид функции $g(x) = p(x)(1 - b_n h(x))$, также в задаче поиска оценки (1.4) изменяется индикатор:

$$\hat{\omega}_n = \frac{1}{k}\sum_{i=1}^{k} \chi\big(T(\hat{Q}_n^{(i)}) - T(P_0) < -b_n\big) \prod_{j=1}^{n} \frac{1}{1 - b_n h(Y_j^{(i)})}, \tag{2.11}$$

Как и в предыдущем разделе, разобьем область моделирования на 3 части. Интегралы соответствующих зон получаются из формул для случая $b_n > 0$ при замене знака при b_n на противоположный. Внешний вид плотности распределения показан на рисунке 2.3.

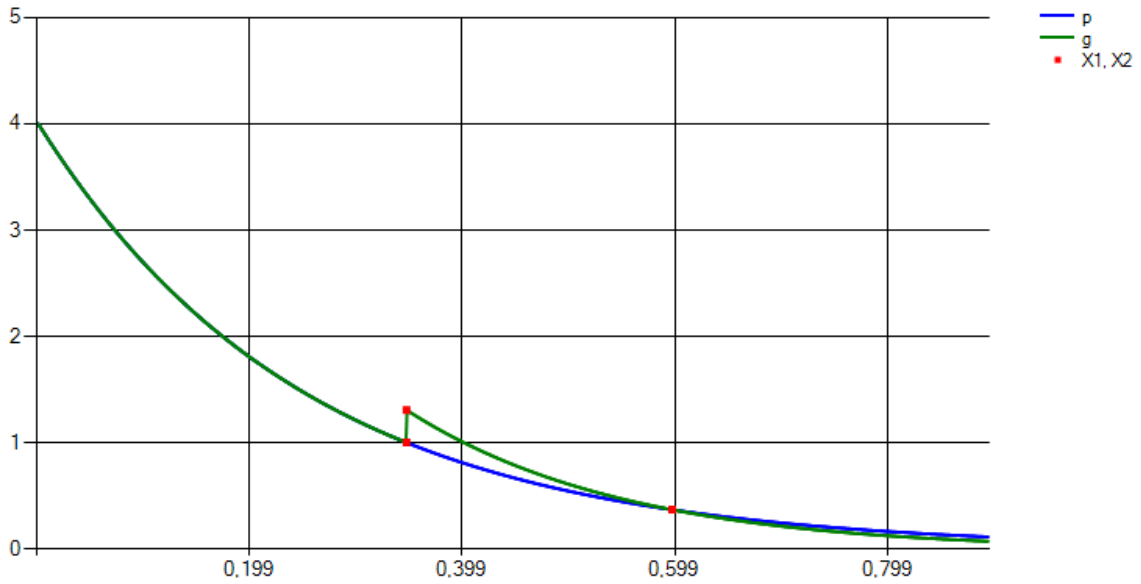

Рис. 2.3. Распределение $g(x), p(x)$ и точки x_1, x_2

Область $A = [0, x_1)$

Также, как и в первой части, область A не отличается от исходного показательного распределения. Моделирование проводится методом обратной функции.

Область $B = [x_1, x_2)$

Необходимо сгенерировать случайную величину с плотностью

$$g(x) = p(x)(1 - b_n h(x)) = \alpha e^{-\alpha x}(1 - b_n \frac{x - x_2}{1 - \beta}) =$$

$$= \alpha e^{-\alpha x}(1 + b_n \frac{x_2 - x}{1 - \beta}) \tag{2.12}$$

Будем проводить моделирование при помощи алгоритма 1. Заметим, что $x_1 \le x \le x_2$, а значит $0 \le x_2 - x \le x_2 - x_1$. Так как $(1 + b_n \frac{x_2 - x}{1 - \beta})$ линейная убывающая функция,

можем рассмотреть ее максимум на интересующем нас промежутке. Он достигается на $\min x = x_1$. Тогда мажорирующая функция имеет вид

$$g_m(x) = \alpha e^{-\alpha x}(1 + b_n \frac{x_2 - x_1}{1 - \beta}).$$

$$\int\limits_B g_m(x) = \frac{e-1}{e}(1 - \beta + \frac{b_n}{\alpha}),$$

Эффективость:

$$\frac{1}{|D_S|} = \frac{1 - \beta + \frac{b_n}{\alpha(e-1)}}{1 - \beta + \frac{b_n}{\alpha}}.$$

Область $C = [x_2, \infty)$

В области C также используем метод мажорант 1. В данной области $g(x)$ успешно мажорируется функцией $p(x)$. Эффективность:

$$\frac{1}{|D_S|} = \frac{1 - \beta + \frac{b_n}{\alpha}}{1 - \beta} = 1 + \frac{b_n}{\alpha(1 - \beta)}.$$

Общее распределение $g(x), b_n < 0$

Моделирование проводится при помощи алгоритма 2. В качестве весов берутся площади плотности $g(x)$ в соответствующих областях. Результат моделирования распределения в виде гистограммы с наложенными функциями $g(x), p(x)$ представлен на рисунке 2.4.

2.3.1. Построение доверительного интервала

При построении доверительного интервала используется соответствующая формула оценки вероятности (2.9) или (2.11). Дисперсия оценки вычисляется при помощи формулы (1.5).

Доверительный интервал с уровнем доверия $\bar{\omega}(\gamma)$ имеет следующий вид:

$$\bar{\alpha}^{-1} = \hat{\alpha}_H^{-1} \pm b_n(\gamma), \tag{2.13}$$

$$\bar{\omega}(\gamma) = \hat{\omega}_n(\gamma) \pm \sigma(\hat{\omega}_n(\gamma)). \tag{2.14}$$

Для оценки полученных результатов необходимо получить другую оценку, которую можно было бы принять за эталонную. Для этого используется подсчет статистики (2.9) в следующем виде:

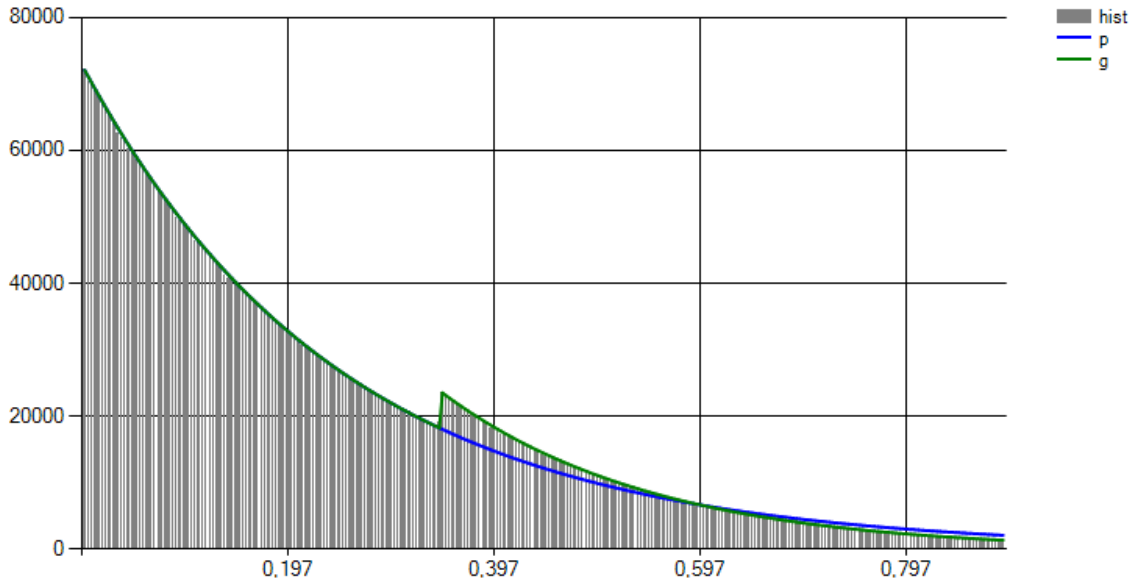

Рис. 2.4. Распределение $g(x), p(x)$

$$\hat{\omega}_D = \frac{1}{k}\sum_{i=1}^{k} \chi\big(T(\hat{Q}^{(i)}) - T(P_0) > b_n\big), \tag{2.15}$$

где $T(\hat{Q}^{(i)})$ строится по формуле (2.2) с $Y_i \sim exp(x;\alpha)$. Дисперсия оценки:

$$D\hat{\omega}_D = \sqrt{\frac{\hat{\omega}_D(1-\hat{\omega}_D)}{k}}.$$

Оценка (2.15) является прямой оценкой для исходного функционала.

2.4. Результаты

Для генерации случайных величин с требуемой веротяностной мерой, и построения на их основе доверительных интервалов, была написана программа, включающая в себя расчетный модуль и простой пользовательский интерфейс.

Расчетный модуль для генерации распределения и построения доверительных интервалов реализован на языке C#, платформа .NET 4.0. Для генерации сложных распределений (например, гамма-распределения), используется библиотека Accord.NET. Пользовательский интерфейс программы реализован на базе набора библиотек WinForms в составе платформы .NET. Программа позволяет строить графики зависимости γ от $\hat{\omega}_n$ и гистограммы распределения $g(x)$ для случаев $b_n > 0$ и $b_n < 0$. Поддержана прямая оценка и оценка при помощи метода существенной выборки.

Создание прототипа программы велось в среде разработки R Studio с использованием интерпретируемого языка R. Ввиду большой вычислительной сложности скриптовый прототип был заменен приложением, что в совокупности с возможностями языка C# позволило увеличить быстродействие приложения примерно на 3 порядка.

В качестве демонстрации результатов моделирования представлены примеры построения доверительных интервалов для случаев $b_n > 0$ и $b_n < 0$ (рисунок 2.5, таблица 2.1). Выбор β делается исходя из стандартной практики в данной области, ее значение выбирают в зависимости от объема выборки в диапазоне от 0.8 до 0.9.

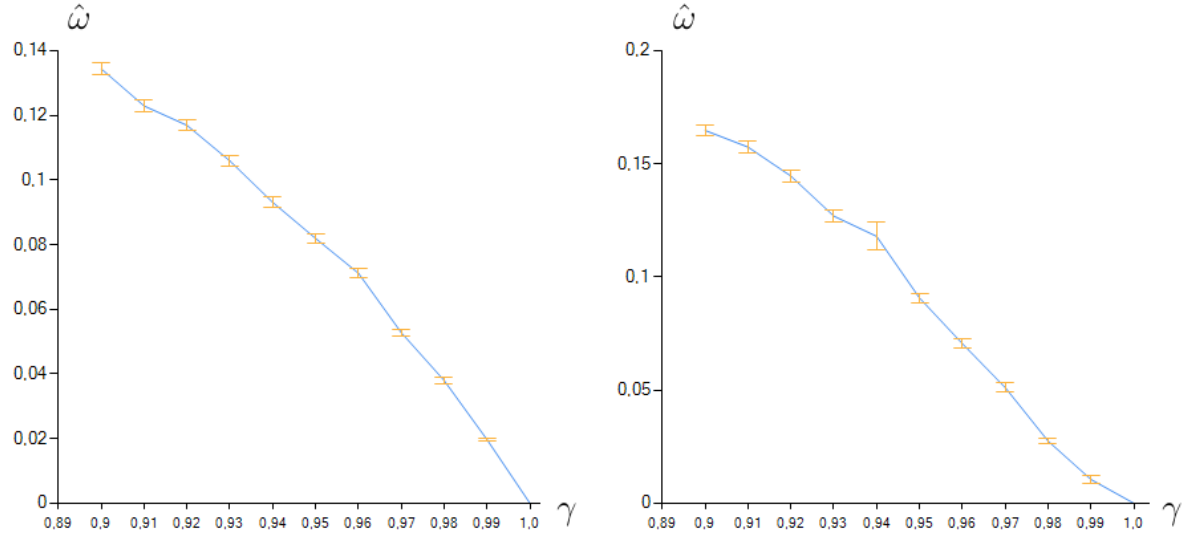

Рис. 2.5. Результаты моделирования вероятности уклонения для $k = 10000, n = 1000, \alpha = 2, \eta = 0.85; b_n > 0$ (слева) и $b_n < 0$ (справа)

По таблице 2.1 можно построить доверительные интервалы по формулам (2.13) и (2.14). Здесь $\sigma = \sigma(\hat{\omega}_n(\gamma))$, $\hat{\omega} = \hat{\omega}_n(\gamma)$. Значение $\hat{\alpha}_H^{-1}$ оценивается отдельно для произвольной выборки, $b_n(\gamma)$ рассчитывается по формуле (1.12).

При построении доверительного интервала фиксируется α, затем последовательно выбирается квантиль γ из заданного интервала с указанным шагом. Для построения по формуле (1.12) вычисляется b_n, затем моделируется k выборок $Y^{(i)} \sim g(x)$ размера n. В качестве $T(\hat{Q}_n^{(i)})$ берется оценка функционала (2.2) по выборке $Y^{(i)}$, $T(P_0) = \alpha^{-1}$. Оценивание производится отдельно для случая $b_n > 0$ и $b_n < 0$, так как распределение для метода существенной выборки у них различно.

Сравнение результатов проводилось для метода существенной выборки и для прямой оценки (2.15). Сравнивались отклонения для оцененных значений вероятностей $\hat{\omega}$.

Таблица 2.1. Результаты моделирования вероятности уклонения для $k = 10000, n = 1000, \alpha = 2, \eta = 0.85$

γ	$\hat{\omega}, b_n > 0$	σ	$\hat{\omega}, b_n < 0$	σ
0.90	0.13434	0.00194	0.16471	0.00250
0.91	0.12288	0.00186	0.15738	0.00266
0.92	0.11689	0.00178	0.14455	0.00280
0.93	0.10589	0.00170	0.12688	0.00262
0.94	0.09306	0.00156	0.11794	0.00605
0.95	0.08190	0.00144	0.09071	0.00204
0.96	0.07113	0.00142	0.07052	0.00196
0.97	0.05274	0.00115	0.05106	0.00211
0.98	0.03797	0.00092	0.02747	0.00110
0.99	0.01966	0.00059	0.01049	0.00186
1.00	0.00000	0.00000	0.00000	0.00000

Сгенерированы оценки с параметрами $\alpha = \{1, 2, 3, 5, 10\}; \beta = \{0.80..0.95\}$ с шагом по $0.5; \gamma = \{0.8..0.98\}$ с шагом по 0.2. Доверительные интервалы строились на 1000 выборках по 1000 элементов. В таблице Б.1 представлены полученные среднеквадратичные отклонения в зависимости от перечисленных параметров, с усреднением по параметру γ.

Заключение

В работе исследована задача построения доверительного интервала для вероятности уклонения параметра формы (*shape*), полученного при помощи оценки Хилла. Задача решалась на основе асимптотической эффективной процедуры существенной выборки для $b_n > 0$ и $b_n < 0$.

Найден явный вид асимптотически эффективной меры моделирования в методе существенной выборки для этой задачи. Предложен алгоритм моделирования случайных величин, соответстующих данной вероятностной мере. Осуществлено моделирование методом существенной выборки. Представлены примеры построения доверительных интервалов, получены результаты сравнения дисперсий данного метода и прямой оценки.

Представляет интерес провести моделирование доверительных интервалов для других офенок параметра формы shape и таким образом осуществить сравнение качества этих оценок.

Список литературы

1. Hill B. M. A simple general approach to inference about the tail of a distribution // Annals of Statistics, 13. 1975. P. 331–341.
2. Dekkers A. L. M., Einmahl J. H. J., Haan L. De. A Moment Estimator for the Index of an Extreme-Value Distribution // Ann. Statist. 1989. 12. Т. 17, № 4. С. 1833–1855. URL: https://doi.org/10.1214/aos/1176347397.
3. III James Pickands. Statistical Inference Using Extreme Order Statistics // Ann. Statist. 1975. 01. Т. 3, № 1. С. 119–131. URL: https://doi.org/10.1214/aos/1176343003.
4. Embrects P. Modelling Extremal Events for Insurance and Finance. New York: Springer, 1997. 655 p.
5. Kratz Marie, I. Resnick Sidney. The QQ-estimator and heavy tails. 1995. 04. Т. 12.
6. Henry J. B. A harmonic moment tail index estimator. 2009. С. 141–162.
7. Hill Jonathan B. ON TAIL INDEX ESTIMATION FOR DEPENDENT, HETEROGENEOUS DATA // Econometric Theory. 2010. Т. 26, № 5. С. 1398–1436. URL: http://www.jstor.org/stable/40800887.
8. Resnick Sidney, Stărică Catalin. Tail index estimation for dependent data // Ann. Appl. Probab. 1998. 11. Т. 8, № 4. С. 1156–1183. URL: https://doi.org/10.1214/aoap/1028903376.
9. Gomes Maria, Martins M. Generalizations of the Hill estimator – asymptotic versus finite sample behaviour. 2001. 02. Т. 93. С. 161–180.
10. Санов И. Н. О вероятности больших отклонений случайных величин // Матем. сб. 1957. Т. 42(84). С. 11–44. URL: http://www.ams.org/mathscinet-getitem?mr=88087.
11. Ermakov M. S. Importance sampling for simulatins of moderate deviation probailities of statistics // Statistics and Decisions, 25. 2007. P. 265–284.
12. Serfling R. J. Approximation Theorems of Mathematical Statistics. New York: Wiley, 1980. 371 p.
13. Ермаков С. М. Метод Монте-Карло в вычислительной математике. Санкт-Петербург, 2009. 192 с.

Приложение А

Дополнительные расчеты

А.1. Дисперсия оценки

По (1.13) имеем:

$$\sigma^2(F) = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} J(F(x))J(F(y))\big(F(x\wedge y) - F(x)F(y)\big).$$

В (2.5) была выражена функция $J(t)$:

$$J(t) = \frac{\gamma(t > \beta)}{1-\beta},$$

$$J(F(t)) = \frac{\gamma(F(t) > \beta)}{1-\beta}.$$

Зафиксируем y и частично выразим (1.13):

$$\int\limits_{-\infty}^{\infty} \big(F(\min(x,y)) - F(x)F(y)\big)dx =$$

$$= \int\limits_{-\infty}^{y} F(x)dx + \int\limits_{y}^{\infty} F(y)dx - \int\limits_{-\infty}^{\infty} F(y)F(x)dx =$$

$$= \int\limits_{-\infty}^{y} F(x)dx + \int\limits_{y}^{\infty} F(y)dx - \int\limits_{-\infty}^{y} F(y)F(x)dx - \int\limits_{y}^{\infty} F(y)F(x)dx =$$

$$= \int\limits_{-\infty}^{y} F(x)dx + F(y)\left(\int\limits_{y}^{\infty} dx - \int\limits_{-\infty}^{y} F(x)dx - \int\limits_{y}^{\infty} F(x)dx\right) =$$

$$= \int\limits_{-\infty}^{y} F(x)dx + F(y)\left(\int\limits_{y}^{\infty}(1-F(x))dx - \int\limits_{-\infty}^{y} F(x)dx\right) =$$

$$= (1-F(y))\int\limits_{-\infty}^{y} F(x)dx + F(y)\int\limits_{y}^{\infty}(1-F(x))dx.$$

Получаем 2 интеграла:

$$\int\limits_{-\infty}^{\infty} dy \int\limits_{-\infty}^{y} dx(1-F(y))F(x) + \int\limits_{-\infty}^{\infty} dy \int\limits_{y}^{\infty} dxF(y)(1-F(x)),$$

которые можно выразить друг через друга путем замены пределов интегрирования. Получаем итоговую формулу:

$$\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}\big(F(\min(x,y))-F(x)F(y)\big)dx = 2\int\limits_{-\infty}^{\infty}(1-F(y))\int\limits_{-\infty}^{y}F(x)dxdy.$$

Используя ее и подставляя (2.5) в (1.13), получаем:

$$\sigma^2(F) = \frac{2}{(1-\beta)^2}\int\limits_{x_1}^{\infty}(1-F(y))\int\limits_{x_1}^{y}F(x)dxdy =$$

$$= \frac{2}{(1-\beta)^2}\int\limits_{x_1}^{\infty}e^{-\alpha y}\int\limits_{x_1}^{y}(1-e^{-\alpha x})dxdy =$$

$$= \frac{2}{(1-\beta)^2}\int\limits_{x_1}^{\infty}e^{-\alpha y}(y-x_1+\frac{1}{\alpha}(e^{-\alpha y}-e^{-\alpha x_1}))dy =$$

$$= \frac{2}{(1-\beta)^2}\int\limits_{x_1}^{\infty}e^{-\alpha y}(y-x_1+\frac{1}{\alpha}e^{-\alpha y}-\frac{1-\beta}{\alpha})dy =$$

$$= \frac{2}{(1-\beta)^2}\int\limits_{x_1}^{\infty}e^{-\alpha y}(-\frac{1-\beta}{\alpha}-x_1)+ye^{-\alpha y}+\frac{1}{\alpha}e^{-2\alpha y}dy =$$

$$= \frac{2}{(1-\beta)^2}\left(-\frac{1}{\alpha}e^{-\alpha y}(-\frac{1-\beta}{\alpha}-x_1)-\frac{1}{\alpha^2}e^{-\alpha y}(\alpha y+1)-\frac{1}{2\alpha^2}e^{-2\alpha y}\right)|_{x_1}^{\infty} =$$

$$= \frac{2}{(1-\beta)^2}\left(\frac{1}{\alpha}e^{-\alpha x_1}(-\frac{1-\beta}{\alpha}-x_1)+\frac{1}{\alpha^2}e^{-\alpha x_1}(\alpha x_1+1)+\frac{1}{2\alpha^2}e^{-2\alpha x_1}\right) =$$

$$= \frac{2}{(1-\beta)^2}\left(\frac{1-\beta}{\alpha}(-\frac{1-\beta}{\alpha}-x_1)+\frac{1-\beta}{\alpha^2}(\alpha x_1+1)+\frac{(1-\beta)^2}{2\alpha^2}\right) =$$

$$= \frac{2}{(1-\beta)^2}\frac{1-\beta}{\alpha}\left((-\frac{1-\beta}{\alpha}-x_1)+\frac{1}{\alpha}(\alpha x_1+1)+\frac{(1-\beta)}{2\alpha}\right) =$$

$$= \frac{2}{(1-\beta)^2}\frac{1-\beta}{\alpha}\frac{\beta+1}{2\alpha} = \frac{\beta+1}{\alpha^2(1-\beta)}$$

Итоговая формула для дисперсии

$$\sigma^2(F) = \frac{\beta+1}{\alpha^2(1-\beta)}. \tag{A.1}$$

Приложение Б

Результаты

Б.1. Сравнение метода существенной выборки и прямой оценки

Таблица Б.1. Отклонения $\sigma(\hat{\omega})$ прямой оценки и оценки метода существенной выборки в зависимости от параметров α, β; 1000 выборок по 1000 релизаций на значение.

Alpha	*Beta*	Прямой метод	Сущ. выборки	p.value	Отношение дисперсий
1	0.80	3.091e-04	2.616e-04	4.129e-09	1.18
	0.85	3.096e-04	2.660e-04	1.940e-08	1.16
	0.90	3.098e-04	2.706e-04	7.025e-07	1.14
	0.95	3.074e-04	2.748e-04	2.029e-04	1.11
2	0.80	2.763e-04	2.141e-04	1.392e-11	1.29
	0.85	2.882e-04	2.277e-04	2.172e-13	1.26
	0.90	2.969e-04	2.422e-04	1.816e-17	1.22
	0.95	3.018e-04	2.577e-04	2.873e-11	1.17
3	0.80	2.152e-04	1.679e-04	9.177e-08	1.28
	0.85	2.403e-04	1.866e-04	9.145e-09	1.28
	0.90	2.668e-04	2.084e-04	1.182e-10	1.28
	0.95	2.880e-04	2.363e-04	8.180e-15	1.21
5	0.80	1.167e-04	1.006e-04	3.401e-05	1.16
	0.85	1.423e-04	1.189e-04	9.090e-06	1.19
	0.90	1.806e-04	1.460e-04	1.597e-06	1.23
	0.95	2.376e-04	1.881e-04	5.882e-09	1.26
10	0.80	4.120e-05	3.923e-05	0.103832	1.05
	0.85	5.117e-05	4.817e-05	0.026655	1.06
	0.90	6.994e-05	6.418e-05	0.013681	1.08
	0.95	1.149e-04	9.795e-05	3.868e-05	1.17