Московский физико-технический институт (национальный исследовательский университет) (МФТИ)

Физтех-школа радиотехники и компьютерных технологий
Кафедра интегрированных киберсистем

Галкин Всеволод Александрович

Методы анализа экстремальных событий

Домашняя работа по практическим задачам

1 курс магистратуры, группа М01-006б

Преподаватель
_____ Н. М. Маркович
«___» _____ 2020 г.

Москва, 2020 г.

1. Задача №1

Постановка:
Сгенерировать выборку из распределения Фреше, которое имеет функцию распределения

$$F(x) = \exp -(\gamma x)^{\frac{1}{\gamma}} 1\{x > 0\} \tag{1}$$

и параметром $\gamma = 1.5$. Размер выборки $n = 100$.

Решение: Для сэмплирования из распределения Фреше, воспользуемся **методом обратной функции**, для этого

1. сэмплируем из равномерного распределения $U[0,\, 1]$;
2. находим значения обратной функции к функции распределения Фреше

$$X_i = \frac{1}{\gamma} \left(-\log U_i\right)^{-\gamma}.$$

По теореме об обратной функции к функции распределения, случайная величина X распределена по закону Фреше (1).

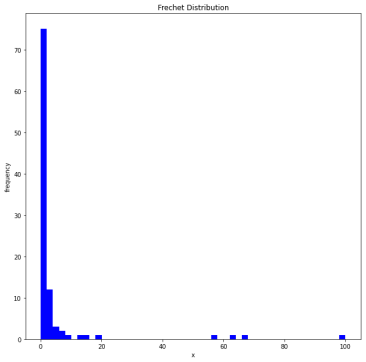


Рис. 1.1. Распределение сгенерированных данных из распределение Фреше

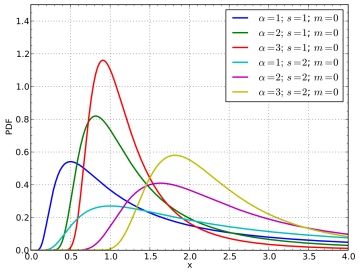


Рис. 1.2. Функция плотности вероятности распределения Фреше. Графики взяты из Википедии

Выводы:
Полученная гистограмма частоты появления случайной величины из распределения Фреше "соответствует"теоретической плотности (см. рис. 1.2). Для качественного оценивания необходимо проверить тот или иной **критерий согласия**, однако этот эксперимент выходит за рамки данной работы.

2. Задача №2

Постановка: Вычислить статистику

$$R_n(p) = \frac{M_n(p)}{S_n(p)},\ n \geq 1,\ p > 0,$$

где

$$M_n(p) = \max\left(|X_1|^p,\, \ldots,\, |X_n|^p\right),\ S_n(p) = |X_1|^p + \ldots + |X_n|^p.$$

В представленных формулах $X_{i\,i=1}^{n}$ независимая выборка из некоторого распределения.

Требуется построить зависимость статистики $R_n(p)$ при фиксированной выборке (размер n большой) и значении параметра p от объема m выбранной подвыборки ($m < n$) и сделать выводы относительно количества конечных моментов распределения.

Решение: Для реализации поставленной задачи и построения графика зависимости сгенерируем выборки из распределения Фреше размером $n = 50\,000$, для параметра m зададим равномерную сетку из диапазона $[10,\, n]$ с числом узлов 500 для различных значений параметра $p = \{1, 2, 3, 4\}$. Графики зависимости представлен на 2.1.

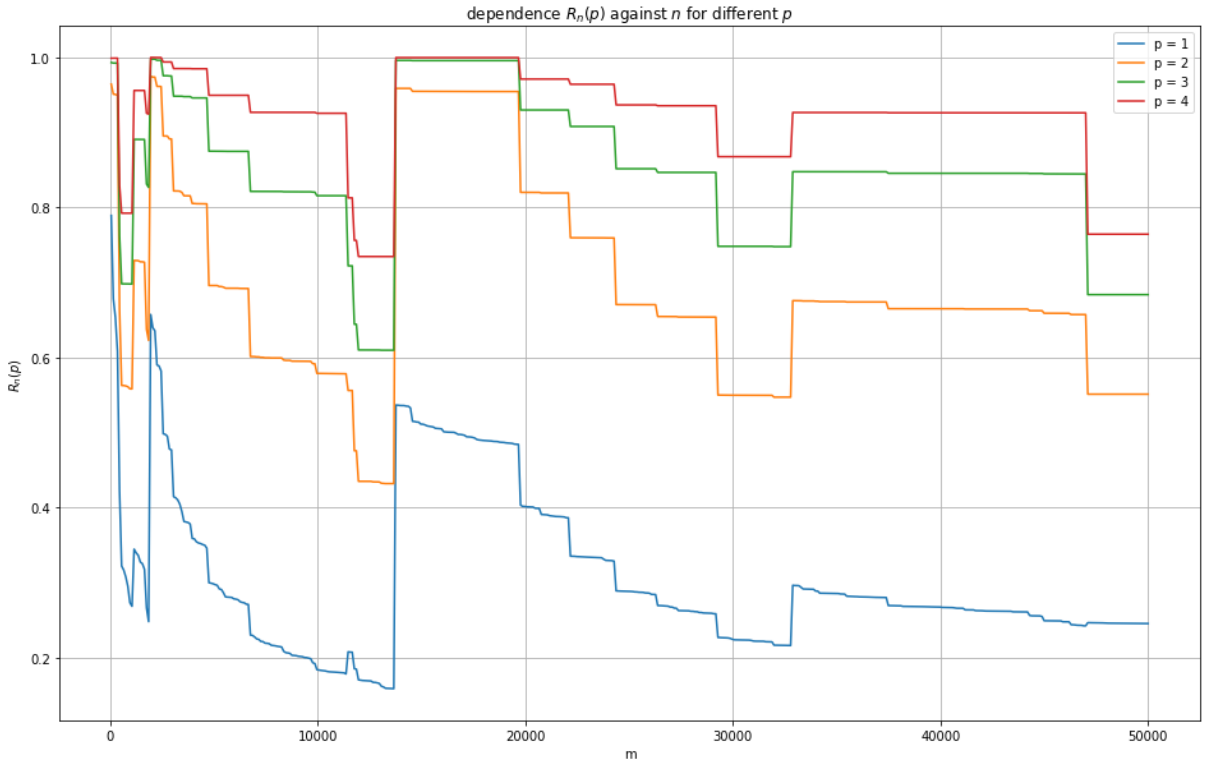


Рис. 2.1. Графики зависимости статистики $R_n(p)$ от объема выборки при различных значениях параметра p

Вывод: Для распределения $F(x)$ момент степени p существует, если оценка $R_n(p)$ сходится к 0 при увеличении размера выборки n. И момент степени p не существует, если расхождение оценки $R_n(p)$ с нулем велико.

Из представленных графиков можно сделать вывод, что у распределения Фреше со значением параметра $\gamma = 1.5$ **не существует даже первого момента**.

3. Задача №3

В данном разделе проведем построение графика QQ-plot, целью которого является исследование принадлежности выборки к распределению. Сам график разброса имеет следующий вид:

$$\{\left(X_k,\, F^{-1}\left(\frac{n-k+1}{n+1}\right)\right) : k = 1,\, \ldots,\, n\},$$

где $X_{(1)} \geq \ldots \geq X_{(n)}$, а F^{-1} – обратная функции к некоторой функции распределения $F(x)$.

Статистическая выборка X распределена по закону с функцией $F(x)$, если QQ-plot близок к линейной функции. Так же следует отметить, что на форму данного графика влияют аномальные выбросы, поэтому при построении нужно учитывать отсев экстремальных данных.

На данном этапе работы проведем следующий эксперимент. Сгенерируем выборку из распределения Фреше размером 1000 экземпляров и проведем построение QQ-plot графика с отсевом 0, 25, 100, 500 максимальных реализаций.

Результаты моделирования приведены на рис. 3.1. Можно заметить, что при увеличении числа отброшенных максимальных экземпляров график становится похожим на экспоненциальный, а не на линейный (**не понимаю почему так происходит, наоборот должен быть более линейным**).

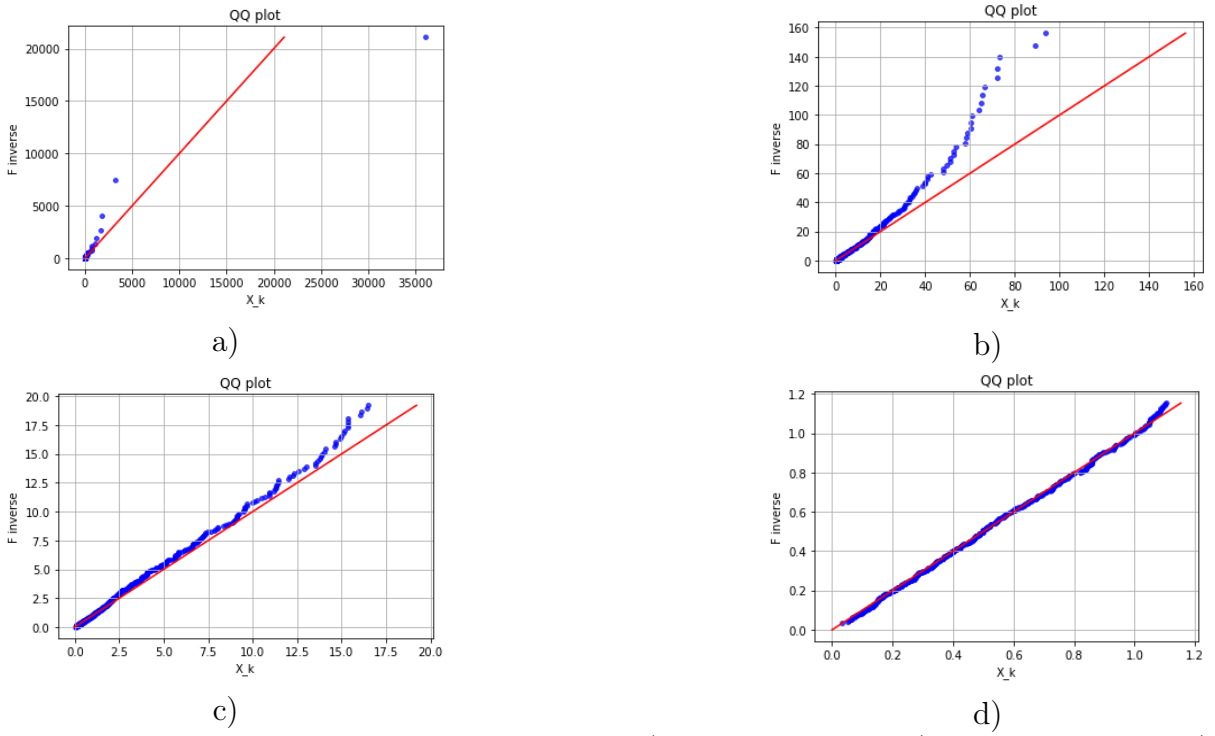


Рис. 3.1. QQ-plot с отсевом максимальных: a) 0 экземпляров, b) 25 экземпляров, c) 100 экземпляров, d) 500 экземпляров.

Можно заметить, что разброс точек при исключении большего числа экстремальных экземпляров выборки стремится к линейной функции. Из этого следует вывод, что выборка получена из распределения Фреше.

Так же следует отметить, что QQ-plot не обязан располагаться в квадрате $[0,\,1]\times[0,\,1]$ при условии отсутствия дополнительной нормировки разброса точек. График разброса QQ-plot строится по принципу

$$\left\{\left(X_k,\,F^{-1}\left(\frac{n-k+1}{n+1}\right)\right):k=1,\,\ldots,\,n\right\}$$

и это график не обязан располагаться внутри единичного квадрата по следующим причинам:

1. Используемая статистическая выборка $X=\{X_k\}_{k=1}^n$ может не принадлежать диапазону $[0,\,1]$;
2. Область определения функции распределения $F(x)$, а следовательно и область значения функции $F^{-1}(x)$ не обязана быть из диапазона $[0,\,1]$.

В силу этих двух причин QQ-plot в общем случае не принадлежит единичному квадрату.

4. Задача №4

Имея статистическую выборку $X^n = X_1, \ldots, X_n$. В данном задании необходимо получить функцию the empirical mean excess, которая имеет следующий вид:

$$e_n(u) = \frac{\sum_{i=1}^{n}(X_i - u)I\{X_i > u\}}{\sum_{i=1}^{n} I\{X_i > u\}},$$

где I – индикаторная функция.

Возьмем выборку, представленную в разделе 1 из распределения Фреше, размер выборки $n = 1000$ экземпляров. Выборка независима и экземпляры одинаково распределены. Для построения функции $e_n(u)$ зададим аргумент u из диапазона [0, 1000] с размером шага 2. Результаты представлены на графике 4.1.

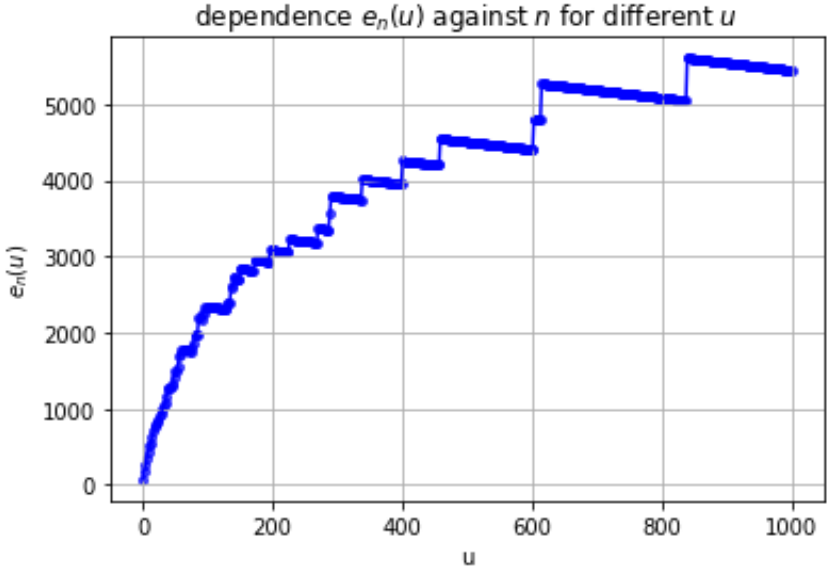


Рис. 4.1. Функция $e_n(u)$

Для анализа результатов воспользуемся следующим свойством, если при больших значениях аргумента u:

1. $e_n(u)$ уходит на бесконечность, то распределение с тяжелыми хвостами;
2. $e_n(u)$ линейная функция, то распределение Парето;
3. $e_n(u)$ постоянная, то экспоненциальное распределение с интенсивность λ;
4. $e_n(u)$ стремится к нулю, то распределение с легкими хвостами.

На представленном графике 4.1 функция $e_n(u)$ уходит на бесконечность, что свидетельствует о том, что распределение с тяжелыми хвостами.

5. Задача №5

Для исследования свойств полученной выборки и оцениванию тяжести его хвоста, получим следующие статистические оценки:

1. Hill estimator;

$$\gamma^H_{n,\,k} = \frac{1}{k}\sum_{i=1}^{k}\sum_{i=1}^{k}\log X_{(n-i+1)} - \log X_{(n-k)},\ k = 1,\,\ldots,\,n-1.$$

2. Ratio estimator;

$$a_n = a_n(x_n) = \sum_{i=1}^{n}\log\frac{X_i}{x_n} I\{X_i > x_n\} / \sum_{i=1}^{n} I\{X_i > x_n\},\ X_{(1)} < x_n < X_{x(n)}.$$

3. Moment estimator;

$$\gamma^M_{n,k} = \gamma^H_{n,\,k} + 1 - 0.5\left(1 - (\gamma^H_{n,\,k})^2 / S_{n,k}\right),$$

где

$$S_{n,k} = \frac{1}{k}\sum_{i=1}^{k}\left(\log X_{(x-i+1)} - \log X_{(n-k)}\right)^2.$$

4. UH estimator;

$$\gamma^{UH}_{n,k} = \frac{1}{k}\sum_{i=1}^{k}\log UH_i - \log UH_{k+1},\ UH_i = X_{(n-i)}\gamma^H_{n,\,k}.$$

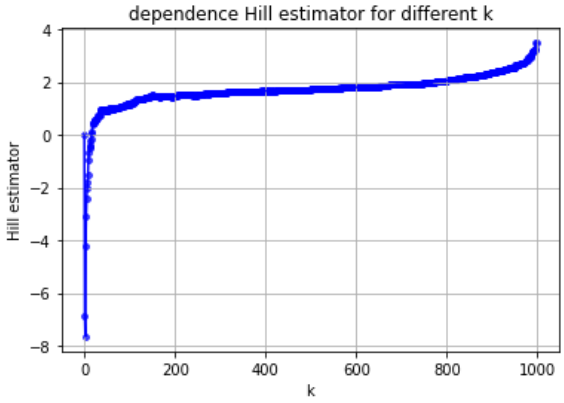
5. Pickands estimator;

$$\gamma^P_{k,n} = \frac{1}{\log 2}\log\frac{X_{(n-k+1)} - X_{(n-2k+1)}}{X_{(n-2k+1)} - X_{(n-4k+1)}},\ k \le \frac{n}{4}.$$

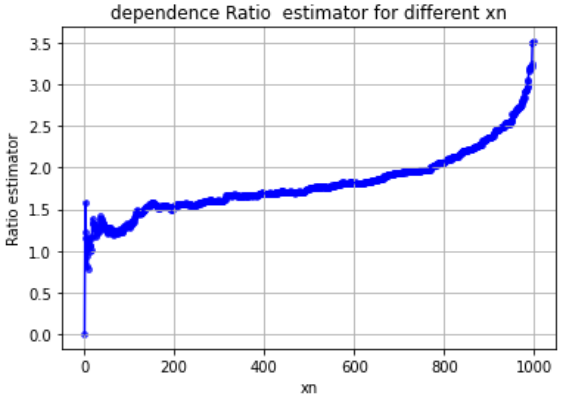
Полученные оценки приведены на рис. 5.1.

При проведении экспериментов использовалась выборка из распределения Фреше объемом 1 000 экземпляров.

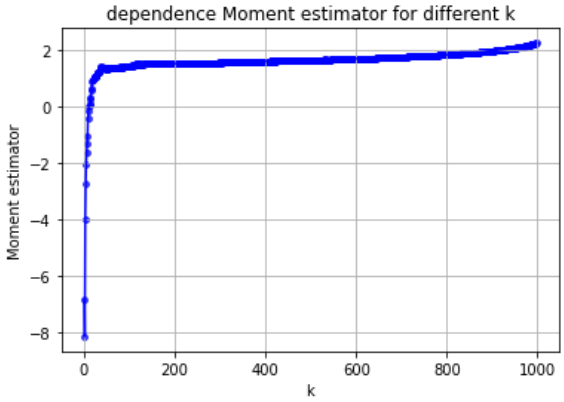
Следует отметить, что поведение оценки Хилла демонстрирует стабильность при $k \in [25,\,1000]$. Положительность оценки Хилла указывает на наличие тяжелых хвостов. Оценка Хилла стабилизируется в окрестности значения 2, следовательно из критерия $\beta < \frac{1}{\gamma}$ можно сделать вывод о конечности моментов. В нашем случае $\beta < \frac{1}{2}$, следовательно распределение имеет только нулевой конечный момент. Следует отметить, что и другие статистические оценки параметра γ стабилизируются в окрестности значения 2.



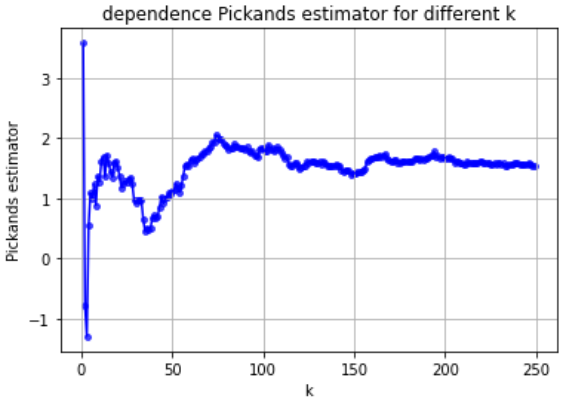
a)



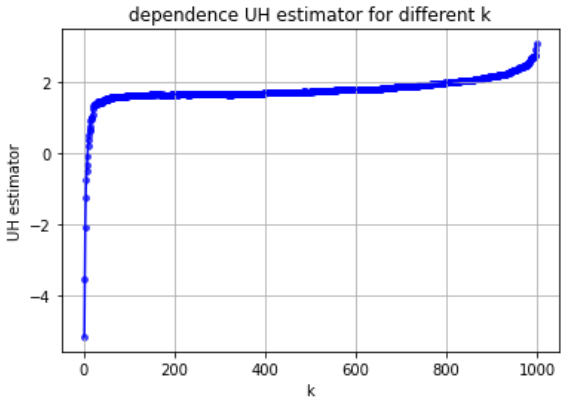
b)



c)



d)



e)

Рис. 5.1. Построенные оценки: a) Hill estimator, b) Ratio estimator, c) Moment estimator, d) Pickands estimator e) UH estimator.

6. Задача №7

В данной задаче требуется сгенерировать выборку из распределения с регулярноменяющимися хвостами, а именно:

1. $1 - F(x) = P\{X > x\} = x^{-\frac{1}{\gamma}} l(x)$, где $l(x) = 1$, 2 и $\gamma = 0.5$.

2. Распределение Вейбула $1 - F(x) = e^{-cx^{1/\gamma}}$, где $(c, \gamma) = \{(1, 2), (2, 3)\}$.

Вычислить оценки Хилла для распределений, изучить влияние функции $l(x)$ на оценку и сравнить полученные оценки с действительными значениями параметра γ.

Для создания статистических выборок воспользуемся методом обратной функции, для этого к функциями распределения

1. $1 - x^{-\frac{1}{\gamma}} l(x)$;

2. $1 - e^{-cx^{1/\gamma}}$

найдем обратные функции.

Такими являются функции (с учетом того, что $l(x) = \tilde{C} = \text{const}$):

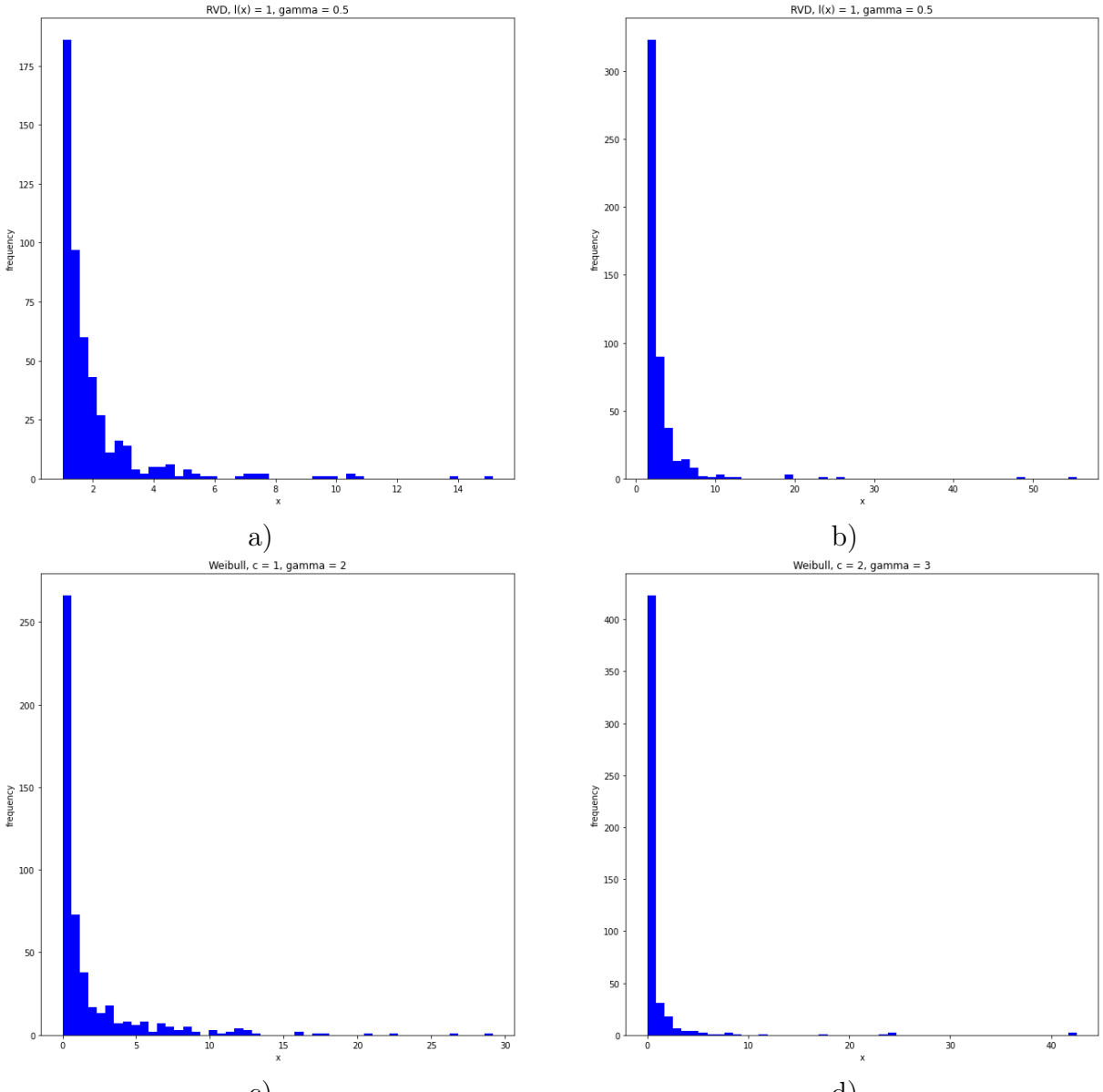
1. Для первой функции распределения

$$x = \left(\frac{\tilde{C}}{1 - y} \right)^{\gamma}.$$

2. Для второй функции распределения

$$x = \left(-\frac{1}{c} \log (1 - y) \right)^{\gamma}.$$

Для создания выборки, вместо y подставим выборку, распределенную по равномерному закону из диапазона $U[0, 1)$.



a) b)

c) d)

Рис. 6.1. Построенные распределения: a) RVD $l(x) = 1$, $\gamma = 0.5$, b) RVD $l(x) = 2$, $\gamma = 0.5$, c) Weibull $c = 1$, $\gamma = 2$, d) Weibull $c = 2$, $\gamma = 3$.

Построим оценки Хилла для полученных распределений.

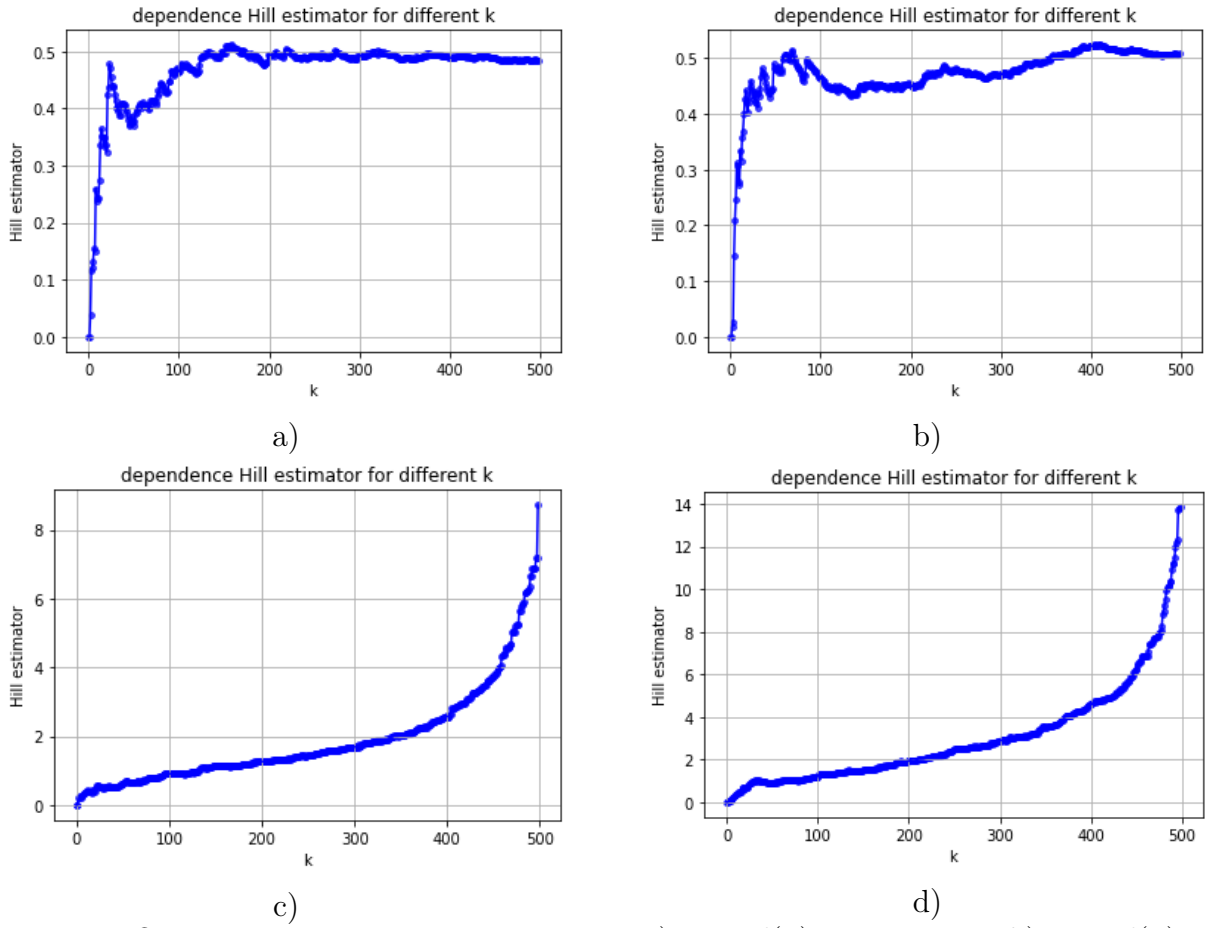


Рис. 6.2. Оценки Хилла для распределений: a) RVD $l(x) = 1$, $\gamma = 0.5$, b) RVD $l(x) = 2$, $\gamma = 0.5$, c) Weibull $c = 1$, $\gamma = 2$, d) Weibull $c = 2$, $\gamma = 3$.

Из представленных графиков можно заметить следующую зависимость, чем больше хвостовая функция (см. графики для выборки из RVD) тем медленнее сходится оценка Хилла, а для выборки из распределения Вейбула оценка при увеличении параметра k и вовсе расходится.

7. Задача №8

Пусть имеется статистическая выборка $X^n = \{X_1, \ldots, X_n\}$, необходимо разделить ее на l групп $V_1, \ldots, V_l$. В каждой группе содержится m независимых, одинаково распределенных экземпляров.

Требуется вычислить групповую оценку хвостового индекса

$$z_l = \frac{1}{l}\sum_{i=1}^{l} k_{lj} = \frac{\tilde{\alpha}}{\tilde{\alpha}+1} = \frac{1}{1+\tilde{\gamma}} \to \tilde{\gamma} = \frac{1}{z_l} - 1,$$

где

$$k_{li} = \frac{M_{li}^{(2)}}{M_{li}^{(1)}},\ M_{li}^{(1)} = \max\{X_j : X_j \in V_i\},$$

а $M_{li}^{(2)}$ второй наибольший элемент из группы V_i.

Необходимо построить график $\{(m,\ \frac{1}{z_m} - 1)\}$, где $m = 10,\ 11,\ \ldots$ вместе с доверительным интервалом. Параметр n взять из $\{150,\ 500,\ 1000\}$.

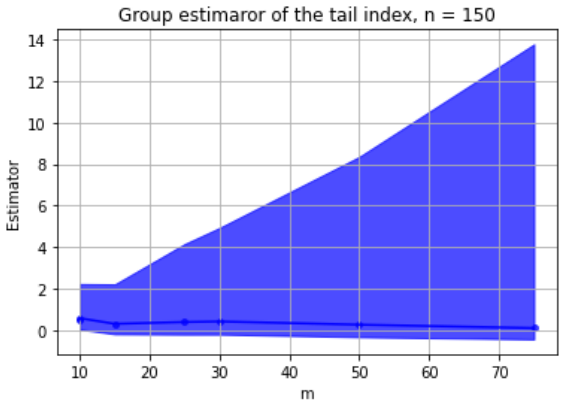
Так же требуется построить доверительный интервал для оценки параметра γ с уровнем $\alpha = 0.95$. Границы доверительного интервала вычисляются как

$$\gamma_{1,\,2} = \left(\overline{k} - \frac{\pm 1.96\sqrt{A}}{l}\right)^{-1} - 1,$$

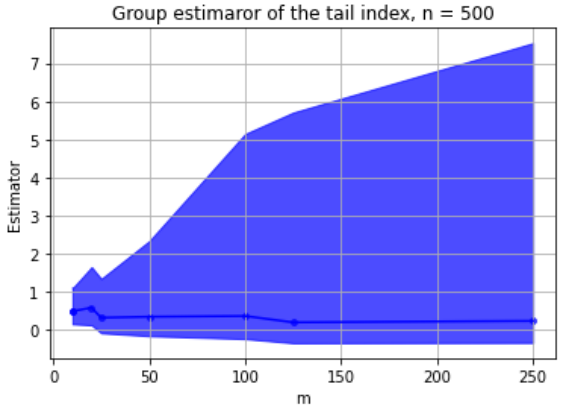
где $\overline{k} = \frac{1}{l}\sum_{i=1}^{l} k_{li}$, $A = \sum_{i=1}^{l} k_{li}^2 - \overline{k}$.

Для реализации эксперимента выберем размер выборок и их количество таким образом, чтобы строго выполнялось условие $n = lm$, где n- общий размер выборки, m – количество экземпляров в группе, а l– количество групп. В качестве базовой выборки возьмем статистическую выборку из распределения с регулярноменяющимся хвостом $F(x) = 1 - x^{-\frac{1}{\gamma}} l(x)$ с параметрами $l(x) = 1$, $\gamma = 0.5$.

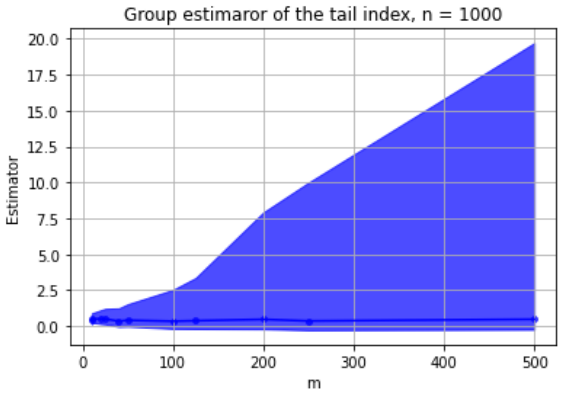
Результаты моделирования для выборок различного объема приведены на 7.1.



a)



b)



c)

Рис. 7.1. Групповые оценки Хилла для выборки из распределения Фреше объемом: a) $n = 150$ b) $n = 500$, c) $n = 1000$.

Для начала оценим возможные знаки величины A. По определению $A = \sum_{i=1}^{l} k_{li}^2 - \overline{k}$, где $k_{li} = \frac{M_{li}^{(2)}}{M_{li}^{(1)}}$. Предположим, что элементы выборки положительные и меньше 1. Отсюда величина k_{il} так же будет положительной и меньше чем 1. В таком случае сумма квадратов чисел меньших 1 будет меньше, чем просто сумма чисел. Таким образом, возможно такая ситуация, что величина A будет отрицательной. Для вычисления границ доверительного интервала величина A входит под знаком квадратного корня, но может иметь отрицательное значение. Это ведет к возникновению комплексных корней, что неестественно для границ доверительных интервалов.

Для распределения с регулярноменяющимся хвостом справедлива сходимость оценки при условии $l = m$. Данное соотношение и наблюдается на представленных результатах, оценка сходится к истинной и доверительный интервал сужается.