Методы современной прикладной статистики
1. Введение

Родионов Игорь Владимирович
vecsell@gmail.com

ШАД

Отчетность по курсу

- Будет 11 (одно сдвоенное) домашних заданий, в каждом из них будет $\leq$ 6 задач (в том числе $\leq$2 теоретических) с общей суммой в 12 баллов.
- В конце семестра будет практический экзамен, за который можно получить максимум 50 баллов.
- Оценка за курс: $\min([(A+B)/12], 10)$, где A – количество баллов за ДЗ, B – оценка за экзамен, $[\cdot]$ – округление к ближайшему целому.

Замечания

- Экзамен является необязательным.
- ДЗ отсылается семинаристу (Anytask).
- ДЗ будет выкладываться на странице курса. Там же можно будет найти программу и презентации лекций.
- Дедлайн по сдаче ДЗ – в 16.30 в четверг, т.е. на выполнение ДЗ дается неделя.
- Если студент попадается на списывании, то в первый раз у него аннулируется одно ДЗ, во второй раз – аннулируются все ДЗ.

Рассмотрим основную задачу математической статистики: по данным/наблюдениям сделать выводы о распределении этих данных. Эту задачу можно решать в рамках одной из трех парадигм

1. параметрической;
2. непараметрической;
3. семипараметрической;

и в рамках одного из 2 подходов: частотного или байесовского.

Параметрическая парадигма

В рамках параметрической парадигмы мы считаем, что множество распределений, из которых мы выбираем подходящее, параметризовано: $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$.

Примеры задач и методов:
1) Точечное оценивание;
2) Доверительное оценивание параметра;
3) Линейная регрессия.

Замечание.

Точное распределение, которому подчиняются наши данные, найти невозможно. Мы можем лишь подобрать модель, с какой-то степенью точности приближающую наши данные.

Семипараметрический подход

В рамках семипараметрического (полупараметрического) подхода считаем, что на семействе распределений $\mathcal{P}$ задан параметр, который хорошо описывает данное семейство. Например, семейство распределений типа Вейбулла с ф.р.

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

где $l(x) : \lim_{t\to\infty} \frac{l(tx)}{l(t)} = 1$ – медленно меняющаяся функция, хорошо описывается параметром γ.

Непараметрический подход

В рамках непараметрической парадигмы мы считаем, что множество распределений, из которых мы выбираем подходящее, невозможно параметризовать с помощью конечномерного параметра.

Примеры задач и методов:
1) Ядерные оценки плотности;
2) Непараметрическая регрессия;
3) Ранговые критерии.

В рамках байесовского подхода мы считаем, что на множестве распределений, из которых мы выбираем подходящее, задана некая вероятностная мера Q (априорное распределение).

Априорное распределение выбирается исходя из имеющихся на текущий момент данных о неизвестном распределении. Если данных нет, априорное распределение полагается неинформативным.

Пусть X – наблюдение из распределения $P \in \mathcal{P}$. Пусть семейство распределений $\mathcal{P}$ параметризовано, $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, и $q(\theta)$ – априорная плотность на Θ. Формат вывода в рамках байесовского подхода таков:

$$q(\theta|X) = \frac{p(X|\theta)q(\theta)}{\int_\Theta p(X|\theta)q(\theta)d\theta},$$

где $p(X|\theta)$ – функция правдоподобия наблюдения X, а $q(\theta|X)$ – апостериорная плотность.

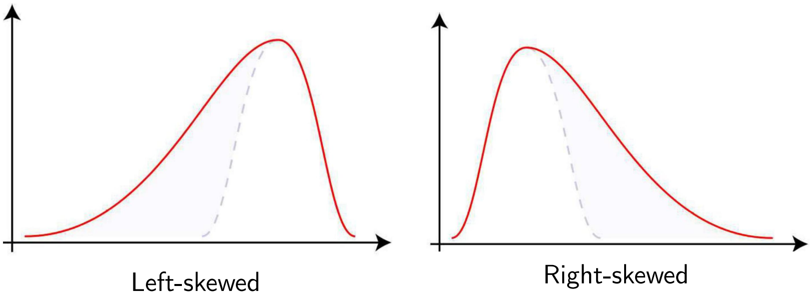
Характеристики распределений

Пусть случайная величина X имеет функцию распределения F. Тогда

- Математическое ожидание: $EX = \int_{\mathbb{R}} x dF(x)$;
- Дисперсия: $DX = E(X - EX)^2 = EX^2 - (EX)^2$;
- Квантиль уровня α : $u_\alpha = \inf\{x : F(x) \geq \alpha\}$;
- Медиана: квантиль уровня 1/2.

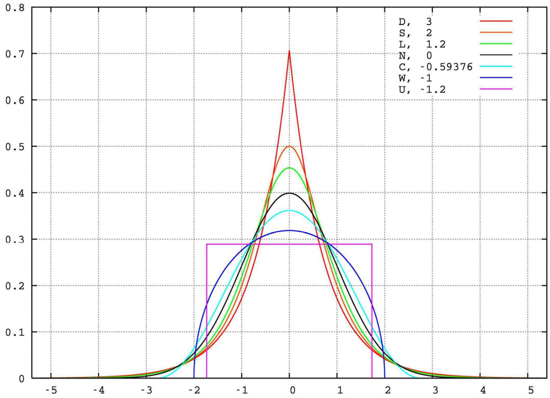
- Коэффициент асимметрии (skewness)

$$Sk = \frac{E(X - EX)^3}{(DX)^{3/2}}$$



- Коэффициент эксцесса (excess, без вычитания 3 – kurtosis)

$$K = \frac{E(X - EX)^4}{(DX)^2} - 3$$



Нормальное распределение

$X \sim N(a, \sigma^2)$, если плотность

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-a)^2}{2\sigma^2}\right\}.$$

Свойства:
1) Является слабым пределом суммы независимых (слабозависимых) случайных величин;
2) $EX = a$; $DX = \sigma^2$;
3) Обозначим $\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$, $\Phi(x) = \int_{-\infty}^{x} \varphi(y)dy$, тогда $F_X(x) = \Phi\left(\frac{x-a}{\sigma}\right)$.

Распределение хи-квадрат

$X \sim \chi^2_n$, если плотность

$$p_X(x) = \frac{x^{n/2-1}}{2^{n/2}\Gamma(n/2)} \exp(-x/2).$$

Свойства:
1) Если $\{\xi_i\}_{i=1}^n$ – нез. сл.в., $\forall i \; \xi_i \sim N(0,1)$, то $\sum_{i=1}^n \xi_i^2 \sim \chi^2_n$.
2) $EX = n$, $DX = 2n$.

Распределение Стьюдента

Пусть $X \sim N(0,1)$, $Y \sim \chi^2_n$, X и Y независимы, тогда случайная величина

$$Z \stackrel{d}{=} \frac{X}{\sqrt{Y/n}}$$

будет иметь распределение Стьюдента с n степенями свободы, $Z \sim St(n)$ (также пишут $Z \sim T_n$).

Свойства:
1) При $n \to \infty$ $T_n \Rightarrow N(0,1)$.
2) При $n = 1$ Z имеет стандартное распределение Коши.

Распределение Фишера

Пусть $X \sim \chi^2_n$, $Y \sim \chi^2_m$, X и Y независимы, тогда случайная величина

$$Z \stackrel{d}{=} \frac{X/n}{Y/m}$$

будет иметь распределение Фишера с n и m степенями свободы, $Z \sim F(n, m)$.

Свойства:

1) При $n, m \to \infty$ $F(n, m) \Rightarrow 1$.

2) При $n = 1$ $\sqrt{Z} \sim T_m$.

Другие распределения

1. $X \sim Exp(\alpha)$, если $p(x) = \alpha e^{-\alpha x}$, $x > 0$;
2. $X \sim U[a, b]$, если $p(x) = \frac{1}{b-a} I\{x \in [a, b]\}$;
3. $X \sim Cauchy(\theta)$, если $p(x) = \frac{\theta}{\pi(x^2+\theta^2)}$;
4. $X \sim Gamma(\alpha, \gamma)$, если $p(x) = \frac{x^{\alpha-1}\gamma^{\alpha}}{\Gamma(\alpha)} e^{-\gamma x}$, $x > 0$;
5. $X \sim Beta(\alpha, \beta)$, если $p(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{(\alpha,\beta)} I(x \in [0, 1])$;
6. $X \sim Pareto(\alpha, \kappa)$, если $p(x) = \frac{\alpha\kappa^{\alpha}}{x^{\alpha+1}} I(x > \kappa)$.

Другие распределения

1. $X \sim Bern(p)$, если $P(X=0)=p$, $P(X=1)=1-p$, $0<p<1$;
2. $X \sim Bin(n,p)$, если $P(X=k)=C_n^k p^k (1-p)^{n-k}$, $0<p<1$, $k \in \{0\dots n\}$;
3. $X \sim Pois(\lambda)$, если $P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$, $\lambda>0$, $k \in \mathbb{Z}_+$;
4. $X \sim Geom(p)$, если $P(x=k)=p(1-p)^{k-1}$, $k \in \mathbb{N}$, $0<p<1$.

Статистики

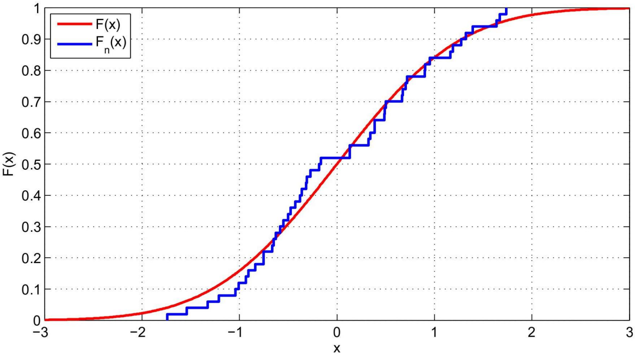
Пусть $X = (X_1, \ldots, X_n)$ – выборка из распределения P. Статистикой $T(X)$ называют любую измеримую функцию выборки. Примеры:

- Выборочные характеристики: $\overline{g(X)} = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$;
  $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ – выборочное среднее,
  $\overline{X^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ – выборочный k-тый момент;
- Выборочные центральные моменты:
  $\overline{X_c^k} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$; $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ – выборочная дисперсия;

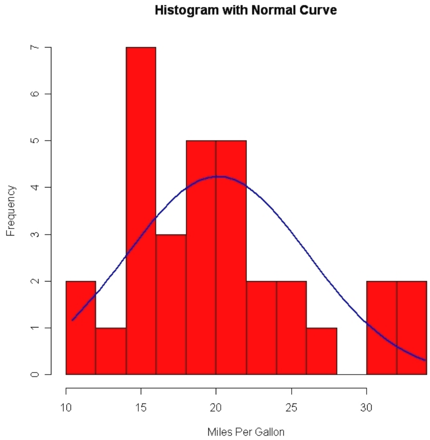
- Порядковые статистики $X_{(1)} = \min(X_1, \ldots, X_n)$,
  $X_{(2)} = \min(X_1, \ldots, X_n \backslash X_{(1)}), \ldots$
  $X_{(n)} = \max(X_1, \ldots, X_n)$;
- Вариационный ряд: $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$;
- Ранг элемента выборки (в вариационном ряду): $R(X_i) = r$, если $X_i = X_{(r)}$;
- Выборочная α-квантиль: $\widehat{Z}_\alpha = X_{([n\alpha])}$;
- Выборочная медиана:

$$\widehat{\mu} = \begin{cases} X_{(k+1)}, & \text{если } n = 2k+1; \\ \frac{1}{2}(X_{(k)} + X_{(k+1)}), & \text{если } n = 2k; \end{cases}$$

- $F_n(x) = \sum_{i=1}^{n} I(X_i \leq x)$ – эмпирическая функция распределения.



- Гистограмма распределения:
$p_n(x) = \sum_{i,j} I(x \in B_j) I(X_i \in B_j)$, $\{B_j\}_{j=1}^{k}$ – разбиение.



Свойства точечных оценок

Рассмотрим параметрическую модель: $F \in \{F_\theta, \theta \in \Theta\}$. Статистики, имеющие значения в Θ, называются оценками параметра θ. Свойства:

- Несмещенность: $E_\theta \widehat{\theta} = \theta \; \forall \theta \in \Theta$;
- Состоятельность: $\widehat{\theta}_n \xrightarrow{P_\theta} \theta$ при $n \to \infty \; \forall \theta \in \Theta$;
- С/к-оптимальность: $D_\theta \widehat{\theta} = \min\limits_{\overline{\theta}: E_\theta \overline{\theta} = \theta} D_\theta \overline{\theta} \; \forall \theta \in \Theta$;
- Асимптотическая нормальность: $\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d_\theta} N(0, \sigma^2(\theta))$ при $n \to \infty \; \forall \theta \in \Theta$;
- Робастность: устойчивость $\widehat{\theta}$ относительно отклонений распределения X от модельного семейства и выбросов, содержащихся в выборке.

- **Метод моментов.**
1) Пусть $\theta = (\theta_1, \ldots, \theta_k)$, выберем k пробных функций $g_1(x), \ldots, g_k(x)$. Положим $m_i(\theta) = E_\theta g(X_1)$.
2) Составим систему уравнений

$$\begin{cases} \mathrm{m}_1(\theta) = \overline{g_1(X)}; \\ \ldots \\ \mathrm{m}_k(\theta) = \overline{g_k(X)}; \end{cases}$$

где $\overline{g_i(X)}$ – выборочные характеристики.
3) Решение этой системы относительно $(\theta_1, \ldots, \theta_k)$ и будет оценкой θ по методу моментов.

Пример.
Пусть $(X_1, \ldots, X_n) \sim N(a, \sigma^2)$, найдем оценки a и σ^2 методом моментов. Выберем $g_1(x) = x$, $g_2(x) = x^2$, тогда $Eg_1(X_1) = EX_1 = a$; $Eg_2(X_1) = \sigma^2 + a^2$. Составим систему:

$$\begin{cases} a = \overline{X}; \\ \sigma^2 + a^2 = \overline{X^2}; \end{cases}$$

откуда $\widehat{a} = \overline{X}$ и $\widehat{\sigma^2} = \overline{X^2} - (\overline{X})^2$.

Свойство. Обозначим $m(\theta) = (m_1(\theta), \ldots, m_k(\theta))$. Если m^{-1} непрерывна, то оценка методом моментов состоятельна, а если непрерывно дифференцируема, то и асимптотически нормальна.

- **Метод максимального правдоподобия.**
  Функция праводоподобия выборки

$$f(X_1, \ldots, X_n; \theta) = \prod_{i=1}^{n} p_\theta(X_i),$$

где $p_\theta(x)$ – обобщенная плотность X_1. Тогда оценка методом максимального правдоподобия (ОМП)

$$\widehat{\theta} = \arg\max_\theta f(X_1, \ldots, X_n; \theta).$$

Пример.
Пусть $(X_1, \ldots, X_n) \sim Exp(\alpha)$, найдем оценку α методом максимального правдоподобия. Имеем

$$f(X, \alpha) = f(X_1, \ldots, X_n; \alpha) = \prod_{i=1}^{n} \alpha e^{-\alpha X_i} = \alpha^n e^{-\alpha \sum_i X_i}.$$

Продифференцируем по α логарифмическую функцию правдоподобия, чтобы найти точку максимума:

$$L(X, \alpha) = \ln f(X, \alpha) = n \ln \alpha - \alpha \sum_i X_i;$$

$$\frac{\partial}{\partial \alpha} L(X, \alpha) = \frac{n}{\alpha} - \sum_i X_i = 0,$$

откуда ОМП $\widehat{\alpha} = 1/\overline{X}$.

Свойство.
При некоторых условиях регулярности (в частности, решение уравнения правдоподобия должно быть единственным) оценка максимального правдоподобия является состоятельной, а при некоторых дополнительных условиях выполнено

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d_\theta} N(0, (i(\theta))^{-1}), \; n \to \infty,$$

где $i(\theta) = E_\theta \left(\frac{d}{d\theta} \ln p_\theta(X_1) \right)^2$ – информация Фишера одного элемента выборки.

- **Метод выборочной квантили.**
При некоторых условиях на плотность распределения
$$\sqrt{n}(\widehat{Z}_\alpha - u_\alpha) \xrightarrow{d} N\left(0, \frac{\alpha(1-\alpha)}{p^2(u_\alpha)}\right), \quad n \to \infty;$$

- **Метод спейсингов.**
Положим $D_i(\theta) = F_\theta(X_{(i+1)}) - F_\theta(X_{(i)})$, $i = 0, \dots, n$, $X_{(0)} = -\infty$, $X_{(n+1)} = +\infty$. Тогда
$$\widehat{\theta} = \arg\max_\theta \sum_{i=0}^{n} \ln D_i(\theta).$$

- **М-оценки.**
$$\widehat{\theta} = \arg\min_\theta \rho(X, \theta),$$
где $\rho(X, \theta)$ – некая функция потерь.

Доверительные интервалы

Пара статистик $(T_1(X), T_2(X))$ называется доверительным интервалом для параметра θ уровня доверия $1-\alpha$, если $\forall \theta \in \Theta$

$$P_\theta(T_1(X) < \theta < T_2(X)) = 1 - \alpha.$$

Последовательность пар статистик $(T_{1,n}(X), T_{2,n}(X))$ называется асимптотическим доверительным интервалом для параметра θ уровня доверия $1-\alpha$, если $\forall \theta \in \Theta$

$$P_\theta(T_{1,n}(X) < \theta < T_{2,n}(X)) \to 1 - \alpha, \;\; n \to \infty.$$

Пример.

Пусть $\widehat{\theta}_n$ – асимптотически нормальная оценка параметра θ,

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d_\theta} N(0, \sigma^2(\theta)), \ \ n \to \infty,$$

тогда при $n \to \infty$

$$P_\theta\left(\widehat{\theta}_n - u_{1-\frac{\alpha}{2}} \frac{\sigma(\widehat{\theta}_n)}{\sqrt{n}} < \theta < \widehat{\theta}_n + u_{1-\frac{\alpha}{2}} \frac{\sigma(\widehat{\theta}_n)}{\sqrt{n}}\right) \to 1 - \alpha,$$

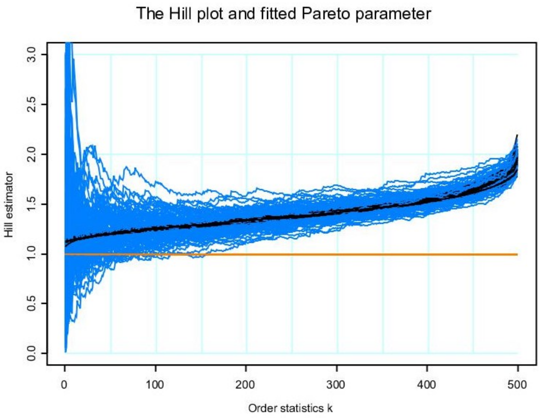
где $u_{1-\frac{\alpha}{2}}$ – $(1 - \frac{\alpha}{2})$-квантиль $N(0, 1)$.

Ключевой вопрос для практика: как ведет себя оценка при конечных значениях n?

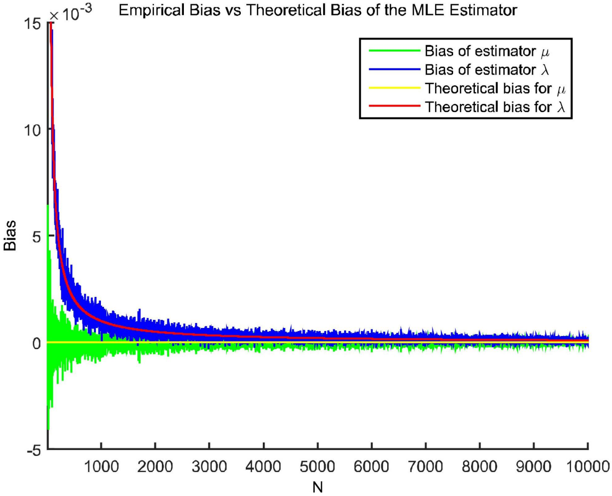
- Если оценка состоятельна, каково её смещение при конечных n?
- Если смещение её невелико, каков разброс её значений (т.е. велика ли её дисперсия)?
- Какой объем выборки нужен, чтобы оценить величину с заданной точностью?

Пояснения

- Смещение: $Bias(\widehat{\theta}) = E_\theta \widehat{\theta} - \theta$.
- Эмпирическое смещение: $Bias(\widehat{\theta}, \theta) = \frac{1}{k} \sum_{i=1}^{k} \widehat{\theta}_i - \theta$, если у нас есть несколько выборок и известно θ.
- Квадратичное смещение: $MSE(\widehat{\theta}) = E_\theta(\widehat{\theta} - \theta)^2$, $RMSE = \sqrt{MSE}$.
- Эмпирическое квадратичное смещение $MSE(\widehat{\theta}) = \frac{1}{k} \sum_{i=1}^{k} (\widehat{\theta}_i - \theta)^2$.
- Bias-variance trade-off: $MSE(\widehat{\theta}) = D_\theta \widehat{\theta} + (Bias(\widehat{\theta}))^2$.



На графике – 100 траекторий значений оценки Хилла индекса экстремального значения $\widehat{\gamma}_H = \frac{1}{k}\sum_{i=0}^{k-1} \ln X_{(n-i)} - \ln X_{(n-k)}$, $n = 500$, $k \in \{1, \ldots, 500\}$. Черным выделена медиана значений статистик.



На графике – смещение (bias) и эмпирическое смещение (empirical bias) $\widehat{\theta}_n - \theta$ двух оценок методом максимального правдоподобия.

Моделирование

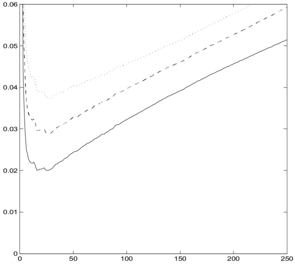
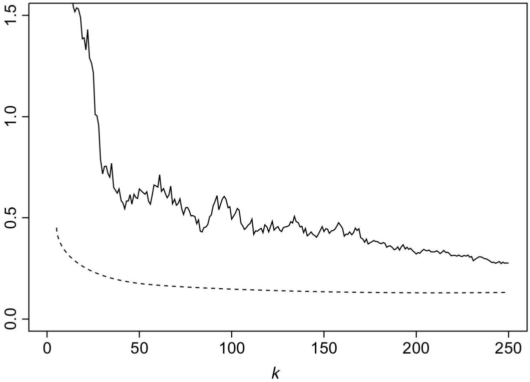


Figure 1: Empirical squared bias as a function of k obtained with $\widehat{\theta}_n$ computed on 500 samples of size 500 from $F_{1/2,\tau,\rho}$. Up: $\rho = -1/2$, down: $\rho = -1/4$, solid line: $\tau = 1$, dashed line: $\tau = 1/2$, dotted line: $\tau = 0$.

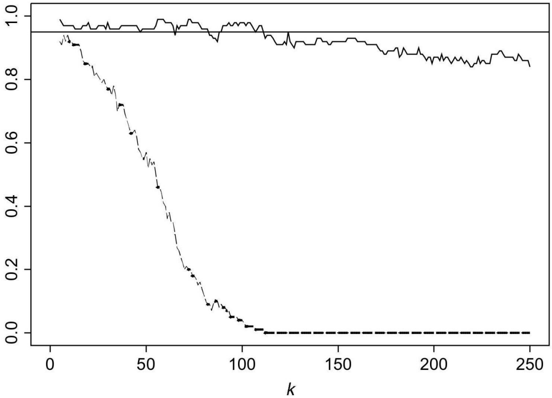
На графике – эмпирическое квадратичное смещение (empirical MSE) значений оценки вейбулловского индекса для трех разных распределений, $n = 500$, $k \in \{1, \ldots, 250\}$, количество выборок – 500.

TAIL ESTIMATION UNDER PARETO-TYPE MODELS

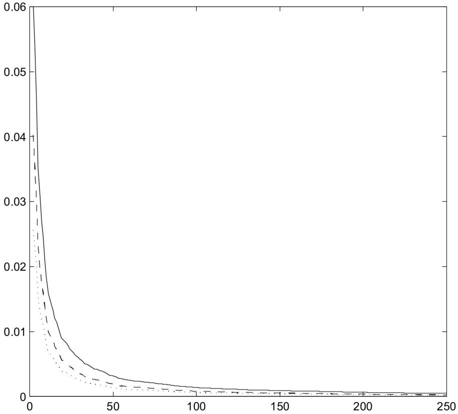


На графике – медиана эмпирического стандартного отклонения $(\widehat{\theta}-\theta)^2$ значений оценки Хилла и ОМП индекса экстремального значения, $n=500$, $k\in\{5,\dots,250\}$.

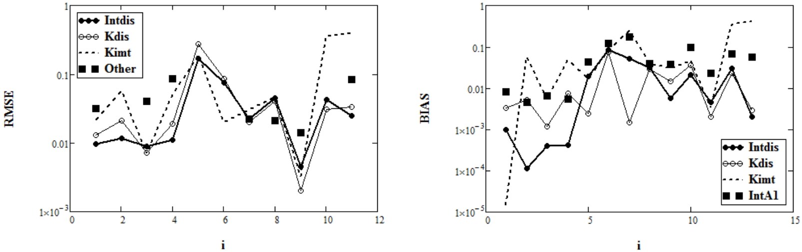
Моделирование



На графике – вероятность покрытия доверительным интервалом, построенным с помощью оценки Хилла и ОМП экстремального индекса, истинного значения параметра, $n = 500$, $k \in \{5, \dots, 250\}$.



На графике – выборочная дисперсия значений оценки вейбулловского индекса для трех разных распределений, $n = 500$, $k \in \{1, \ldots, 250\}$, количество выборок – 500.



The best RMSE and Bias for the intervals estimator ('Intdis') and K-gaps ('Kdis') estimators with threshold u selected by discrepancy equations (14) and (16); the K-gaps estimator with u selected by the test IMT ('Kimt') and the intervals estimator with the "plateau-finding"algorithm A1 to select u ('IntA1') against the number of processes related to the columns labels in Tables 1 and 2.

Finita!