Санкт-Петербургский политехнический университет Петра Великого
Институт прикладной математики и механики
Кафедра прикладной математики

Робастное оценивание параметров многомерного нормального закона

Диссертация на соискание ученой степени магистра

Выполнила студентка гр. 63601/3, А.А. Шагал
Научный руководитель, д.ф.-м.н., Г. Л. Шевляков

2017 г.

Постановка задачи

- Сравнительный анализ различных робастных оценок параметра положения для одномерных и многомерных распределений.
- Анализ ухудшения качества оценок с ростом размерности пространства для многомерных распределений.

Робастность — устойчивость статистических выводов к отклонениям от принятых моделей распределений данных, в частности, к грубым ошибкам.

M-оценки

$$x_1, x_2, \dots x_n \in X \subseteq R^1, \qquad f(x-\theta), \theta \in R^1$$

M-оценкой $\hat{\theta}$ параметра положения назовём решение с ледующей задачи:

$$\sum_{i=1}^{n} \psi\left(x_i - \hat{\theta}\right) = 0,$$

где ψ — оценочная функция.

$$V(\psi, f) = \frac{\int_{-\infty}^{+\infty} \psi^2(x) f(x) dx}{\left(\int_{-\infty}^{+\infty} \psi'(x) f(x) dx\right)^2} \qquad eff(\psi, f) = \frac{V(-f'/f, f)}{V(\psi, f)} = \frac{1}{I(f) V(\psi, f)},$$

где $I(f)$ — информация Фишера.

Одномерный случай: распределения

- Обобщённое нормальное распределение:

$$f(x;\theta,p) = \frac{p}{2\Gamma(1/p)}\left[\frac{\Gamma(3/p)}{\Gamma(1/p)}\right]^{1/2} exp\left\{-\frac{|x-\theta|^p}{\left(\frac{\Gamma(1/p)}{\Gamma(3/p)}\right)^{p/2}}\right\}, \mathrm{p} \geq 1.$$

- Распределение Стьюдента:

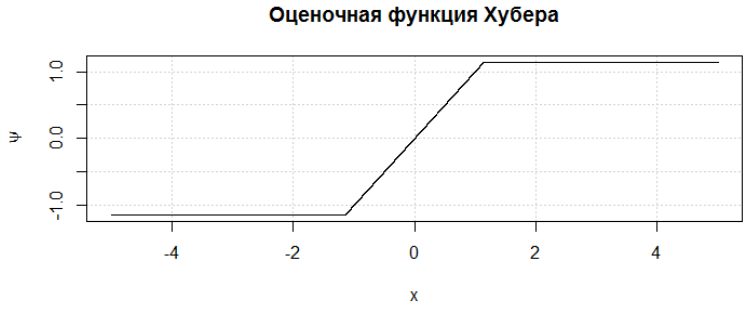
$$f(x;\theta,n) = \frac{1}{\sqrt{\pi n}}\frac{\Gamma((n+1)/2)}{\Gamma(n/2)}\left(1+\frac{(x-\theta)^2}{n}\right)^{-\frac{n+1}{2}}, \qquad n > 0.$$

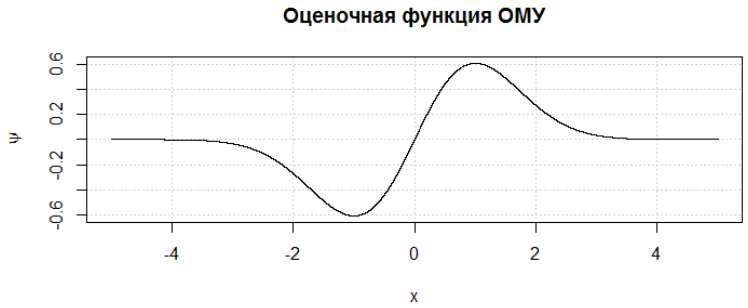
- Модель смеси Тьюки: $f(x) = (1-\varepsilon)\varphi(x;\theta,\sigma_0^2) + \varepsilon\varphi(x;\theta,\sigma_1^2)$,

где $0 \leq \varepsilon \leq 0.5, \sigma_0^2 \ll \sigma_1^2$.

Одномерный случай: оценки параметра положения

- Выборочные среднее и медиана;
- полусумма минимума и максимума;
- оценка Хьюбера;
- оценка Хампеля;
- двухэтапная оценка;
- оценки Мешалкина-Шурыгина:
  - оценка максимальной устойчивости;
  - радикальная оценка.





Одномерный случай: критерии качества оценок

- Эффективность оценки:

$$eff(\hat{\theta}).$$

- Смещение:

$$b(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

- Средний квадрат ошибки:

$$MSE(\hat{\theta}) = D(\hat{\theta}) + E^2(\hat{\theta}).$$

Одномерный случай: моделирование

Параметры генерации:

Размер выборки (*n*)	5, 10, 20, 40, 60, 100
Число повторений генерации (*m*)	1000, 10000
Параметр формы обобщённого нормального распределения (*p*)	1, 1.5, 2, 5, 100
Число степеней свободы распределения Стьюдента (*df*)	1, 1.5, 2, 5, 100
Доля засоряющих наблюдений в модели Тьюки (ε)	0.01, 0.05, 0.1, 0.2, 0.5
Параметр масштаба в модели Тьюки (σ_0, σ_1)	$\sigma_0 = 1, \sigma_1 = 3$

Одномерный случай: результаты

- Обобщённое нормальное распределение:

p	Mean	Med	MinMax	Huber	Hampel	OMU	RAD	TwoStep
1	0,020	0,011	0,81	0,014	0,014	0,015	0,013	0,015
1,5	0,013	0,014	0,21	0,012	0,013	0,014	0,013	0,013
2	0,0099	0,015	0,094	0,012	0,012	0,015	0,012	0,010
5	0,0060	0,015	0,010	0,0072	0,0066	0,21	0,50	0,0061
100	0,0035	0,010	0,00030	0,0036	0,0035	0,56	0,92	0,0035

- Распределение Стьюдента:

df	Mean	Med	MinMax	Huber	Hampel	OMU	RAD	TwoStep
1	4583	0,025	11310370	0,039	4579	30	18,0	0,026
1,5	12	0,022	31044	0,022	12	0,32	0,14	0,020
2	0,26	0,020	483	0,018	0,16	0,034	0,021	0,019
5	0,016	0,017	1,3	0,013	0,014	0,016	0,014	0,014
100	0,010	0,015	0,098	0,011	0,011	0,014	0,012	0,011

- Модель смеси Тьюки:

eps	Mean	Med	MinMax	Huber	Hampel	OMU	RAD	TwoStep
0,01	0,011	0,016	0,47	0,011	0,011	0,011	0,012	0,011
0,05	0,014	0,016	1,3	0,012	0,012	0,014	0,013	0,011
0,1	0,018	0,018	1,4	0,013	0,013	0,018	0,015	0,012
0,2	0,026	0,021	1,3	0,016	0,016	0,026	0,018	0,015
0,5	0,049	0,035	0,9	0,030	0,031	0,051	0,033	0,037

Многомерный случай: распределения

- Нормальное распределение

$$f(x;\theta,C) = \frac{1}{\sqrt{(2\pi)^k \mathrm{C}}} \mathrm{e}^{-\frac{(x-\theta)^T \mathrm{C}^{-1}(x-\theta)}{2}},$$

$$\theta \epsilon R^k, \qquad \mathrm{C} \epsilon R^{k*k};$$

- распределение Коши;
- распределение Лапласа;
- равномерное распределение;
- модель смеси Тьюки.

Многомерный случай: оценки параметра положения

- Покомпонентные среднее и медиана.
- Многомерная выборочная медиана:

$$\theta_m = \frac{\sum x_i/\sqrt{q_i}}{\sum 1/\sqrt{q_i}}, \sqrt{q_i} = \sqrt{(x_i - \theta)^T C^{-1}(x_i - \theta)}.$$

- *M*-оценки Маронны-Хьюбера:

$$\frac{1}{n}\sum v_1(q_i)(x_i - \theta) = 0,$$

$$\frac{1}{n}\sum (v_2(q_i)(x_i - \theta)(x_i - \theta)^T - v_3(q_i)C) = 0\,,$$

$$v_1(y) = \frac{\psi_1(y)}{y}, \qquad v_2(y) = \frac{\psi_2(y)}{y}, \qquad v_3 = 1,$$

$$\psi_1(y) = \psi_H(y,k), \psi_2(y) = \psi_H(y,k^2),\ \ \psi_H = \min(k, \max(y, -k)), k = 1.14\,.$$

Многомерный случай: оценки параметра положения

- Оценки Мешалкина-Шурыгина:

$$\sum (x_i - \theta) e^{-\lambda q_i/2} = 0,$$

$$\sum \left((x_i - \theta)(x_i - \theta)^T - (1 + \eta)^{-1} C \right) e^{-\lambda q_i/2} = 0,$$

$$\lambda, \eta > 0:$$

  - оценка максимальной устойчивости $(\lambda = \eta = 0.5)$;
  - радикальная оценка $(\lambda = \eta = 1)$.

- Двухэтапная оценка:
  - отбраковка выбросов в данных;
  - применение классического метода оценивания к «очищенным» данным.

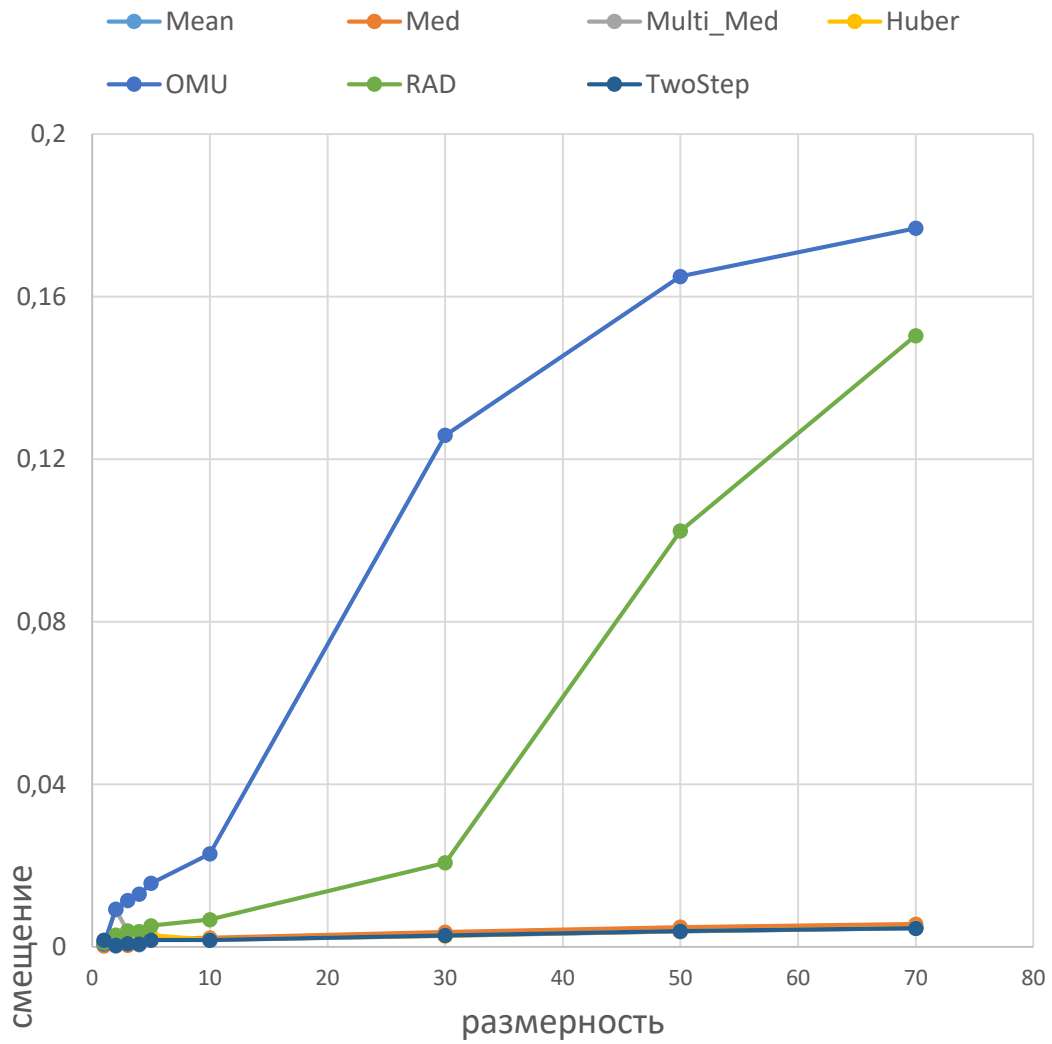
Многомерный случай: моделирование

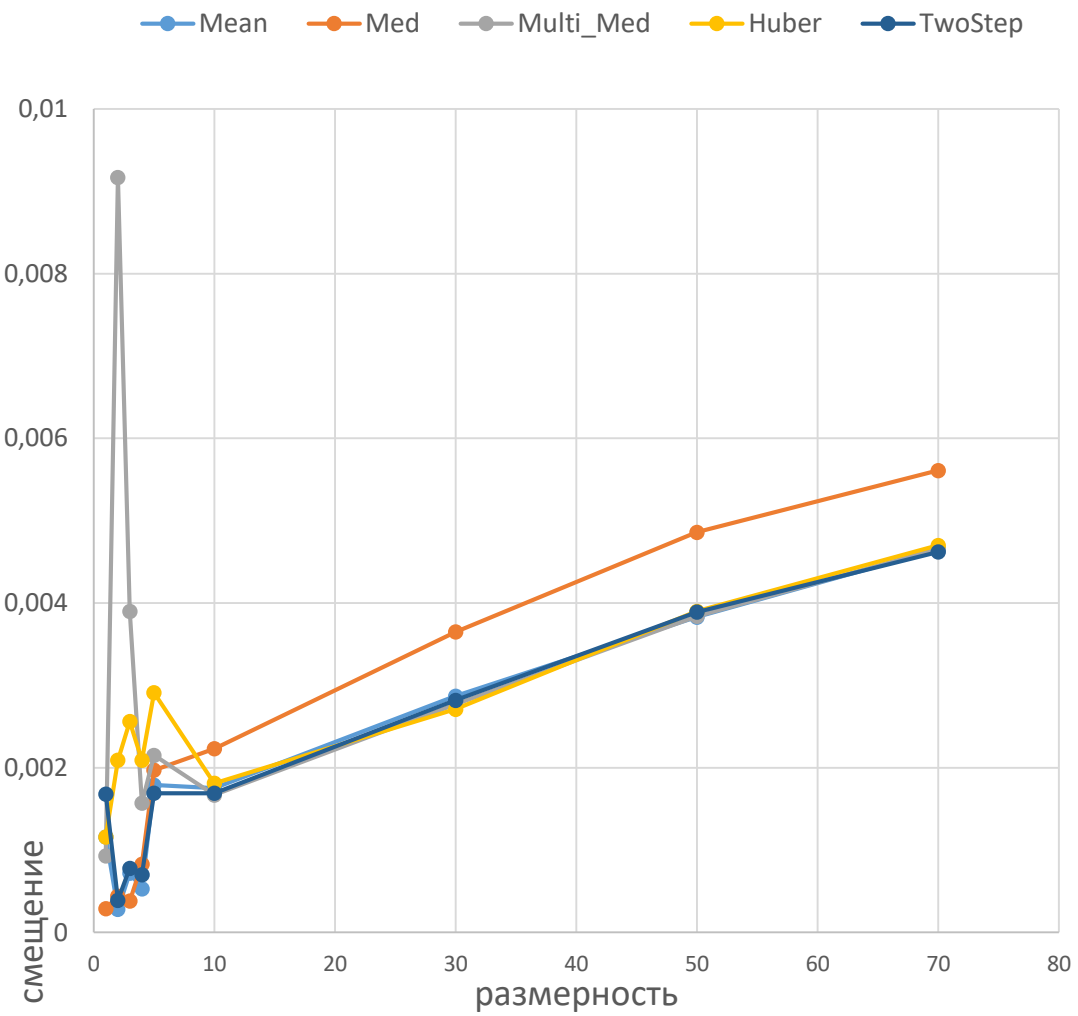
Параметры генерации

Размер выборки (n)	5, 50, 100, 300
Число повторений генерации (m)	10000
Размерность пространства (p)	1, 2, 3, 4, 5, 10, 30, 50, 70
Доля засоряющих наблюдений в модели Тьюки (ε)	0.1

Критерии качества оценки: смещение, средний квадрат ошибки.

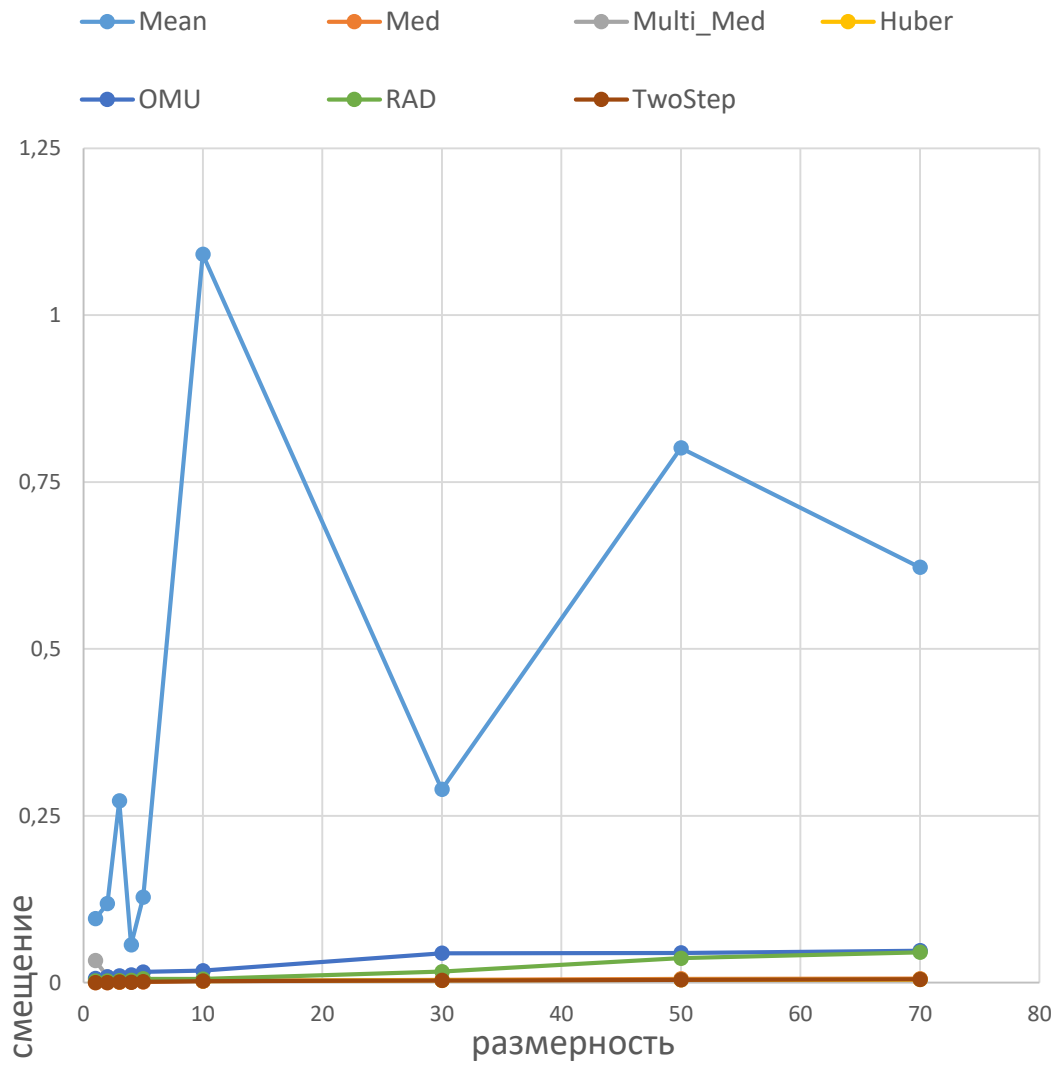
Многомерный случай: результаты

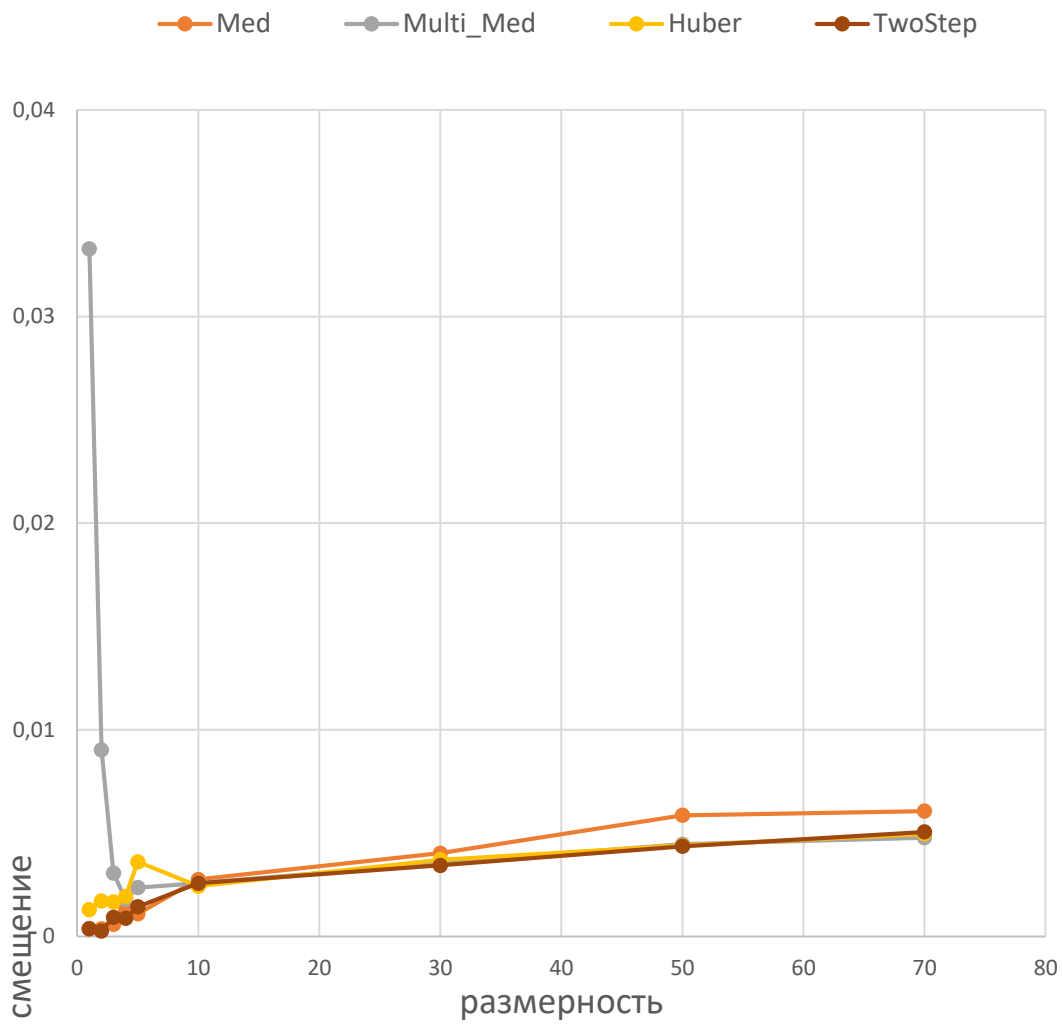




Нормальное распределение

Многомерный случай: результаты





Модель Тьюки

Заключение

- Произведён сравнительный анализ оценок параметра положения для одномерных распределений.
- На основе модельных данных построена зависимость качества оценки от размерности пространства для многомерных распределений.
- Выявлены лучшие оценки для различных классов распределений.
- Сформулированы направления будущих исследований:
  - асимптотика ухудшения качества оценок;
  - эффективность оценок;
  - оценки ковариационной матрицы.

Спасибо за внимание!

М-оценки (1)

$$x_1, x_2, \dots x_n \in X \subseteq R^1, \qquad f(x-\theta), \theta \in R^1$$

М-оценкой $\hat{\theta}_n$параметра положения назовём решение следующей задачи на минимизацию:

$$\hat{\theta}_n = argmin_\theta \left(\sum_{i=1}^{n} \rho(x_i, \theta) \right),$$

$\rho(x, \theta)$ — функция контраста.

- $\rho(x, \theta) = -\log\big(f(x, \theta)\big)$ — оценка максимума правдоподобия;
- $\rho(x, \theta) = (x - \theta)^2$ — метод наименьших квадратов.

М-оценки (2)

ψ — оценочная функция: $\psi(x,\theta) = c(\theta)\frac{\partial}{\partial\theta}\rho(x,\theta)$

$$\sum_{i=1}^{n}\psi\left(x_i - \hat{\theta}\right) = 0 \Rightarrow \sum_{i=1}^{n}(x_i - \theta)\xi(x_i - \theta) = 0$$

$$\hat{\theta}_k = \frac{\sum_{i=1}^{n} x_i \xi(x_i - \hat{\theta}_{k-1})}{\sum_{i=1}^{n} \xi(x_i - \hat{\theta}_{k-1})}$$

Тогда асимптотическая дисперсия оценки $\hat{\theta}$:

$$V(\psi, f) = \frac{\int_{-\infty}^{+\infty}\psi^2 f dx}{\left(\int_{-\infty}^{+\infty}\psi' f dx\right)^2}$$

Минимаксный подход Хьюбера

$$V(\psi^*, f) \leq V(\psi^*, f^*) = \inf_{\psi \in \Psi} \sup_{f \in F} V(\psi, f)$$

- Наихудшей функцией f^* будет являться функция, минимизирующая информацию Фишера:

$$f^* = \arg\min_{f \in F} \int \left(\frac{f'}{f}\right)^2 f dx$$

Минимаксная оценочная функция:

$$\psi^* = -\frac{f^{*\prime}}{f^*}$$

Для класса ε-загрязнённого нормального распределения:

$$\psi_H^* = \max(-k, \min(x, k)), \qquad k = 1.14$$

Робастный подход Хампеля

Функция влияния для М-оценки:

$$IF(x;\psi,f)=\frac{\psi(x)}{\int \psi' f dx}$$

$$\psi_{a,b,r}(x)=\begin{cases} x, 0\le |x|\le a \\ a\,sign(x), a\le |x|\le b \\ a\dfrac{r-|x|}{r-b}\,sign(x), b\le |x|\le r \\ 0, |x|\ge r \end{cases}, \qquad 0<a\le b<r<\infty$$

$$a=1.31, b=2.039, r=4$$

Оценки Шурыгина

- Оценка максимальной устойчивости:

$$\psi_*(x,\theta) = c\left(\frac{\partial}{\partial\theta} f(x,\theta) + \beta f(x,\theta)\right)$$

- Радикальная оценка:

$$\psi(x,\theta) = c\left(\frac{\partial}{\partial\theta} \ln f(x,\theta) + \beta\right)\sqrt{f(x,\theta)}$$

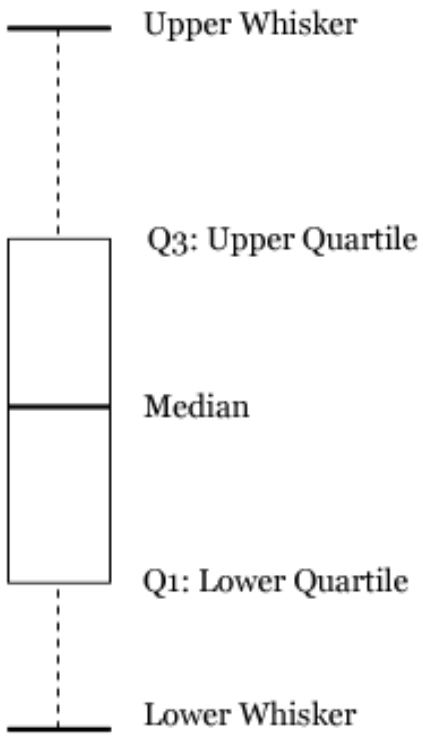
Двухэтапная оценка

- Отбраковка части выборки

$$x_u = \max\{UQ + 1.5 * IQR, x_{(n)}\}$$

$$x_l = \min\{LQ - 1.5 * IQR, x_{(1)}\}$$

$$IQR = UQ - LQ$$



- Получение оценки на основе оставшихся значений

Одномерный случай. Методы оценки

- выборочное среднее

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- выборочная медиана

$$\hat{\theta} = \begin{cases} \dfrac{x_{(k)} + x_{(k+1)}}{2}, n = 2k \\ x_{(k+1)}, \qquad n = 2k + 1 \end{cases}$$

- полусумма минимума и максимума

$$\hat{\theta} = \frac{x_{(1)} + x_{(n)}}{2}$$