Министерство образования Республики Беларусь

Учреждение образования
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Кафедра информационных технологий
автоматизированных систем

В. С. Муха

Статистические методы обработки данных

Часть 3

Учебно-методическое пособия для студентов специальности
"Автоматизированные системы обработки информации"

Минск 2007

11 РОБАСТНОСТЬ СТАТИСТИЧЕСКИХ ПРОЦЕДУР

11.1 Понятие робастности статистических процедур

Статистические выводы (оценки) лишь отчасти основываются на наблюдениях. Важную основу этих выводов составляют исходные предположения о математической модели изучаемого явления (модельные предположения). Например, в задаче точечного оценивания параметров распределений мы предполагали, что известна плотность вероятности генеральной совокупности с точностью до оцениваемого параметра. В задаче байесовского точечного оценивания мы предполагали известной также плотность вероятности оцениваемого параметра. Обычно от такого рода предположений не требуется абсолютной точности. При формулировке таких предположений обычно исходят из принципа, что малая ошибка в математической модели не должна приводить к существенной ошибке окончательного статистического вывода. Вместе с тем в последние десятилетия обнаружено, что некоторые статистические процедуры весьма чувствительны к довольно малым отклонениям от модельных предположений. В связи с этим появились новые статистические процедуры – робастные процедуры (от английского слова robust – крепкий, дюжий). Под робастностью статической процедуры понимается нечувствительность к малым отклонениям от модельных предположений, то есть нечувствительность к искажениям постулируемых моделей.

11.2 Искажения Тьюки-Хьюбера

Наиболее часто рассматриваются отклонения от модельных предположений типа Тьюки-Хьюбера [23]. Считается, что искаженная модель (генеральная совокупность) описывается плотностью вероятности вида

$$f_\xi(x) = (1-\varepsilon) f_\eta(x) + \varepsilon f_\omega(x), \qquad (11.1)$$

где $f_\xi(x)$ – плотность вероятности искаженной модели, $f_\eta(x)$ – плотность вероятности основного (гипотетического, предполагаемого) распределения, $f_\omega(x)$ – плотность вероятности искажений (засорений, загрязнений), $\varepsilon > 0$ – малая величина, такая, что $0 \le \varepsilon \le \varepsilon_+$. Обычно $\varepsilon_+ = 0{,}5$.

Смысл модели искажений типа Тьюки-Хьюбера состоит в том, что выборочные значения из распределения $f_\xi(x)$ состоят из "хороших" и "плохих" наблюдений. "Хорошие" наблюдения появляются с вероятностью $1-\varepsilon$, а "плохие" – с вероятностью ε. Обычно предполагается, что распределение $f_\eta(x)$ и распределение искажений $f_\omega(x)$ имеют одно и то же среднее значение, т.е.

$$E(\eta) = a_\eta = E(\omega) = a_\omega, \tag{11.2}$$

а дисперсия распределения искажений в q раз больше дисперсии основного распределения, т.е.

$$\sigma_\omega^2 = q\sigma_\eta^2,\ \sigma_\eta^2 = D(\eta),\ \sigma_\omega^2 = D(\omega). \tag{11.3}$$

В соответствии со сказанным "плохие" наблюдения имеют гораздо большую дисперсию по сравнению с "хорошими" и поэтому называются выбросами, а модель Тьюки-Хьюбера (11.1) – (11.3) – моделью искажений типа выбросов.

11.3 Числовые характеристики модели Тьюки-Хьюбера

Пусть имеется искаженная модель (11.1). Найдем числовые характеристики распределения $f_\xi(x)$. Для математического ожидания получим

$$a_\xi = E(\xi) = \int_{-\infty}^{\infty} x f_\xi(x)dx =$$

$$= (1-\varepsilon)\int_{-\infty}^{\infty} x f_\eta(x)dx + \varepsilon\int_{-\infty}^{\infty} x f_\omega(x)dx = (1-\varepsilon)a_\eta + \varepsilon a_\omega,$$

где $a_\eta = E(\eta),\ a_\omega = E(\omega)$. Таким образом,

$$a_\xi = (1-\varepsilon)a_\eta + \varepsilon a_\omega.$$

Для начального момента k-го порядка $\nu_\xi^{(k)}$ получим

$$\nu_\xi^{(k)} = E(\xi^k) = \int_{-\infty}^{\infty} x^k f_\xi(x)dx = (1-\varepsilon)\int_{-\infty}^{\infty} x^k f_\eta(x)dx + \varepsilon\int_{-\infty}^{\infty} x^k f_\omega(x)dx,$$

$$\nu_\xi^{(k)} = (1-\varepsilon)\nu_\eta^{(k)} + \varepsilon\nu_\omega^{(k)},$$

где $\nu_\eta^{(k)} = E(\eta^k)$, $\nu_\omega^{(k)} = E(\omega^k)$. Данные формулы справедливы для любых распределений $f_\eta(x)$, $f_\omega(x)$.

Если выполняется условие (11.2), то искаженная модель будет иметь то же математическое ожидание, что и основная модель: $a_\xi = a_\eta$. Кроме того, в этом случае получается простая формула для центральных моментов искаженной модели:

$$\mu_\xi^{(k)} = E((\xi - E(\xi))^k) = \int_{-\infty}^{\infty}(x - E(\xi))^k f_\xi(x)dx =$$

$$= (1-\varepsilon)\int_{-\infty}^{\infty}(x - E(\xi))^k f_\eta(x)dx + \varepsilon\int_{-\infty}^{\infty}(x - E(\xi))^k f_\omega(x)dx =$$

$$= (1-\varepsilon)\int_{-\infty}^{\infty}(x - E(\eta))^k f_\eta(x)dx + \varepsilon\int_{-\infty}^{\infty}(x - E(\omega))^k f_\omega(x)dx =$$

$$= (1-\varepsilon)\mu_\eta^{(k)} + \varepsilon\,\mu_\omega^{(k)}.$$

Таким образом, при $a_\xi = a_\eta$

$$\mu_\xi^{(k)} = (1-\varepsilon)\mu_\eta^{(k)} + \varepsilon\,\mu_\omega^{(k)}.$$

В частности, обозначая $\mu_\xi^{(2)} = \sigma_\xi^2$, $\mu_\eta^{(2)} = \sigma_\eta^2$, $\mu_\omega^{(2)} = \sigma_\omega^2$, будем иметь

$$\sigma_\xi^2 = (1-\varepsilon)\sigma_\eta^2 + \varepsilon\,\sigma_\omega^2.$$

Если выполняется также и условие (11.3), то мы получим $\sigma_\xi^2 = (1-\varepsilon)\sigma_\eta^2 + q\varepsilon\sigma_\eta^2$, или

$$\sigma_\xi^2 = (1-\varepsilon+q\varepsilon)\sigma_\eta^2.$$

Итак, мы получили числовые характеристики модели Тьюки-Хьюбера с произвольными распределениями, обладающими свойствами (11.2), (11.3).

11.4 Гауссовская модель Тьюки-Хьюбера

Если в модели Тьюки-Хьюбера распределения $f_\eta(x)$ и $f_\omega(x)$ гауссовские, то можно получить выражение для центрального момента 4-го порядка $\mu_\xi^{(4)}$. Действительно, для нормального распределения $\mu_\eta^{(4)} = 3\sigma_\eta^4$, $\mu_\omega^{(4)} = 3\sigma_\omega^4 = 3q^2\sigma_\eta^4$, и

$$\mu_\xi^{(4)} = (1-\varepsilon)3\sigma_\eta^4 + \varepsilon 3q^2\sigma_\eta^4 = 3(1-\varepsilon+q^2\varepsilon)\sigma_\eta^4.$$

В итоге для гауссовской модели Тьюки-Хьюбера с условиями (11.2), (11.3) мы имеем соотношения

$$a_\xi = a_\eta,$$

$$\sigma_\xi^2 = (1-\varepsilon+q\varepsilon)\sigma_\eta^2,$$

$$\mu_\xi^{(4)} = 3(1-\varepsilon+q^2\varepsilon)\sigma_\eta^4.$$

11.5 Равномерная модель Тьюки-Хьюбера

Рассмотрим модель Тьюки-Хьюбера (11.1) – (11.3) в предположении, что распределения $f_\eta(x)$ и $f_\omega(x)$ равномерны в промежутках $(a_1,b_1),(a_2,b_2)$ соответственно, т.е.

$$f_\xi(x) = (1-\varepsilon)U(a_1,b_1) + \varepsilon U(a_2,b_2).$$

Найдём такие a_2 и b_2, чтобы выполнялись условия (11.2), (11.3). Для этого необходимо решить систему уравнений

$$\frac{b_1+a_1}{2} = \frac{b_2+a_2}{2},$$

$$\frac{(b_2-a_2)^2}{12} = q\frac{(b_1-a_1)^2}{12}.$$

Отсюда получаем систему уравнений

$$b_1 + a_1 = b_2 + a_2,$$

$$b_2 - a_2 = \sqrt{q}(b_1 - a_1),$$

из которой находим

$$b_2 = \frac{(1+\sqrt{q})b_1 + (1-\sqrt{q})a_1}{2},$$

$$a_2 = \frac{(1-\sqrt{q})b_1 + (1+\sqrt{q})a_1}{2}.$$

С целью упрощения выкладок рассмотрим распределение, симметричное относительно нуля, т. е. рассмотрим случай $a_1 = -b_1$, $b_1 = 1$. В этом случае

$$a_2 = -\sqrt{q}\ ,\ b_2 = \sqrt{q}\ .$$

Итак, будем рассматривать модель

$$f_\xi(x) = (1-\varepsilon)U(-1,1) + \varepsilon U(-\sqrt{q},\sqrt{q}).$$

Плотность вероятности этой модели имеет вид

$$f_\xi(x) = (1-\varepsilon)\begin{cases}\frac{1}{2}, & x \in (-1,1),\\ 0, & x \notin (-1,1),\end{cases} + \varepsilon\begin{cases}\frac{1}{2\sqrt{q}}, & x \in (-\sqrt{q},\sqrt{q}),\\ 0, & x \notin (-\sqrt{q},\sqrt{q}).\end{cases}$$

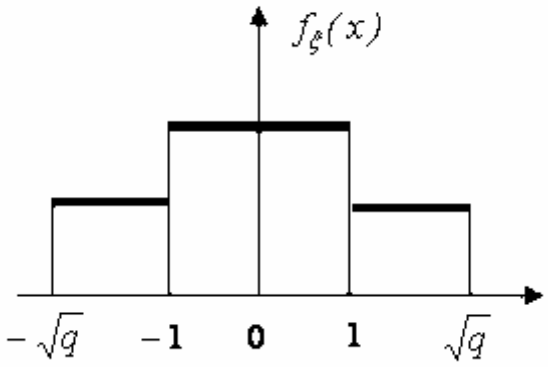

Рис. 11.1. Плотность вероятности равномерной модели Тьюки-Хьюбера

При $q > 1$ эта функция представлена на рисунке 11.1. Перепишем выражение для $f_\xi(x)$ в виде

$$f_\xi(x)=\begin{cases}\dfrac{\varepsilon}{2\sqrt{q}}, & x\in(-\sqrt{q},-1),\\ \dfrac{1-\varepsilon}{2}+\dfrac{\varepsilon}{2\sqrt{q}}, & x\in(-1,1),\\ \dfrac{\varepsilon}{2\sqrt{q}}, & x\in(1,\sqrt{q}),\\ 0, & x\in(-\infty,-\sqrt{q})\cup(\sqrt{q},+\infty),\end{cases}$$

или

$$f_\xi(x)=\begin{cases}f_1(x)=\dfrac{\varepsilon}{2\sqrt{q}}, & x\in(-\sqrt{q},-1),\\ f_2(x)=\dfrac{\sqrt{q}+\varepsilon(1-\sqrt{q})}{2\sqrt{q}}, & x\in(-1,1),\\ f_3(x)=\dfrac{\varepsilon}{2\sqrt{q}}, & x\in(1,\sqrt{q}),\\ 0, & x\in(-\infty,-\sqrt{q})\cup(\sqrt{q},+\infty).\end{cases}$$

Найдем также функцию распределения данной модели. Для $x\in(-\infty,-\sqrt{q})$ $F_\xi(x)=0$. Для $x\in(-\sqrt{q},-1)$ получаем

$$F_\xi(x)=F_1(x)=\int_{-\sqrt{q}}^{x}f_1(x)dx=\frac{\varepsilon}{2\sqrt{q}}(x+\sqrt{q}),\ F_1(-1)=\frac{-\varepsilon(1-\sqrt{q})}{2\sqrt{q}}.$$

Для $x\in(-1,1)$

$$F_\xi(x)=F_2(x)=F_1(-1)+\int_{-1}^{x}f_2(z)dz=\frac{\varepsilon(-1+\sqrt{q})}{2\sqrt{q}}+\frac{[\sqrt{q}+\varepsilon(1-\sqrt{q})](x+1)}{2\sqrt{q}},$$

$$F_2(1)=\frac{\varepsilon(\sqrt{q}-1)}{2\sqrt{q}}+\frac{2\sqrt{q}+2\varepsilon(1-\sqrt{q})}{2\sqrt{q}}=\frac{2\sqrt{q}+\varepsilon(1-\sqrt{q})}{2\sqrt{q}}.$$

Для $x\in(1,\sqrt{q})$

$$F_\xi(x)=F_3(x)=F_2(1)+\int_{1}^{x}f_3(z)dz=\frac{2\sqrt{q}+\varepsilon(1-\sqrt{q})}{2\sqrt{q}}+\frac{\varepsilon(x-1)}{2\sqrt{q}}.$$

Для $x\in(\sqrt{q},+\infty)$

$$F_\xi(x) = F_3(\sqrt{q}) = \frac{2\sqrt{q} + \varepsilon(1-\sqrt{q}) - \varepsilon(1-\sqrt{q})}{2\sqrt{q}} = 1.$$

11.6 Характеристики робастности на основе риска

Пусть $d \in D$ – принимаемое решение относительно состояния $s \in S$ некоторой системы. Качество решения будем характеризовать риском $r_\varepsilon = E(w(s,d))$, где $w(s,d)$ – функция потерь, и усреднение ведется по искаженной модели (11.1).

Значение риска при $\varepsilon = 0$ (при отсутствии искажений) обозначим r_0. Рассмотрим следующие характеристики робастности решающего правила [20].

1. Гарантированный риск прогнозирования

$$r_+ = \max_{0 \le \varepsilon \le \varepsilon_+} r_\varepsilon .$$

В любой задаче желательно, чтобы риск был как можно меньше. Гарантированный риск r_+ – это такое значение риска, больше которого мы не получим при любых допустимых значениях искажений.

2. Коэффициент неустойчивости решающего правила

$$k_+ = \frac{r_+ - r_0}{r_0}.$$

3. δ-допустимый уровень искажений $\varepsilon^*(\delta) = \max\{\varepsilon : k_+ \le \delta\}$ представляет собой максимальное значение ε, при котором $k_+ \le \delta$.

Для любого решающего правила вычисляются указанные характеристики, которые затем сравниваются с аналогичными характеристиками другого решающего правила, и на основании проведенного сравнения устанавливают, какое решающее правило является наиболее предпочтительным с точки зрения робастности.

Пример 11.1. Найдем характеристики робастности выборочного среднего $\bar{x}$ для модели Тьюки-Хьюбера (11.1) – (11.3) как оценки математического

ожидания a_η гипотетического (основного) распределения в предположении, что $a_\xi = a_\eta$, $\sigma_\omega^2 = q\sigma_\eta^2$.

В данном примере предполагается, что имеется выборка $x_1,...,x_n$ из искаженного распределения $f_\xi(x)$, и по ней получена оценка $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, которая считается оценкой генерального среднего a_η. Найдем характеристики робастности этой оценки: $r_{\bar{x}}$ – риск оценки, $r_{\bar{x}+}$ – гарантированный риск, $k_{\bar{x}+}$ – коэффициент неустойчивости, δ-допустимый уровень искажений. В данном случае мы имеем $a_\xi = a_\eta$, $\sigma_\xi^2 = (1-\varepsilon+q\varepsilon)\sigma_\eta^2$. Найдем риск оценки $r_{\bar{x}} = E((\bar{x} - a_\eta)^2)$, где усреднение ведется по искаженной модели. Поскольку

$$E(\bar{x}) = E(\frac{1}{n}\sum_{i=1}^{n} x_i) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = a_\eta ,$$

то

$$r_{\bar{x}} = D(\bar{x}) = \frac{\sigma_\xi^2}{n} = (1-\varepsilon+q\varepsilon)\frac{\sigma_\eta^2}{n}.$$

Полагая $\varepsilon = 0$, получаем

$$r_{\bar{x},0} = \frac{\sigma_\eta^2}{n}.$$

Полагая $\varepsilon = \varepsilon_+$, получаем

$$r_{\bar{x},+} = D(\bar{x}) = (1-\varepsilon_+ + q\varepsilon_+)\frac{\sigma_\eta^2}{n}.$$

Коэффициент неустойчивости

$$k_{\bar{x},+} = \frac{r_{\bar{x}+} - r_{\bar{x},0}}{r_{\bar{x},0}} = (q-1)\varepsilon_+ .$$

Дельта-допустимый уровень искажений $\varepsilon^*(\delta)$ определяется как максимальное значение ε, при котором

$$k_{\bar{x},+} = (q-1)\varepsilon_+ \le \delta .$$

Получаем

$$\varepsilon^*(\delta) = \frac{\delta}{q-1}.$$

При $q = 3$, $\varepsilon_+ = 0{,}5$, $\delta = 0{,}5$ находим

$$k_{\bar{x},+} = 1,\ \varepsilon^*(\delta) = 0{,}25.$$

При таких значениях характеристик робастности оценка считается не робастной. В таблице 11.1 приведены значения риска $r_{\bar{x}} = r_{\bar{x}}(\varepsilon)$ при $n = 50$, $\sigma_\eta^2 = 1$, $q = 3$.

Таблица 11.1

ε	0	0,05	0,1	0,15	0,2	0,25	0,3	0,35	0,4	0,45	0,5
$r_{\bar{x}}$	0,02	0,022	0,024	0,026	0,028	0,03	0,032	0,034	0,036	0,038	0,04

Видно, что при изменении уровня искажений ε от нуля до 0,5 риск возрастает в 2 раза. При 0,5-допустимом уровне искажений $\varepsilon^*(0{,}5) = 0{,}25$ риск возрастает в 1,5 раза. Риск является монотонной (линейной) функцией уровня искажений ε. Таким образом, выборочное среднее $\bar{x}$ считается не робастной оценкой. Более устойчивой к выбросам является такая оценка, как выборочная медиана. К сожалению, найти аналогичные характеристики медианы для большинства распределений в модели Тьюки-Хьюбера не представляется возможным. Вместе с тем выводы относительно выборочного среднего справедливы для любых распределений в модели Тьюки-Хьюбера.

Пример 11.2. Найдем характеристики робастности выборочного среднего и выборочной медианы как оценки математического ожидания равномерной модель Тьюки-Хьюбера.

Ввиду сложности расчетов мы сможем получить характеристики робастности выборочной медианы лишь для выборки объема $n = 3$. Характеристики робастности выборочного среднего были получены нами в

примере 11.1, где было отмечено, что они справедливы для любых моделей Тьюки-Хьюбера. Приведем их.

Риск выборочного среднего

$$r_{\bar{x}}(\varepsilon) = (1 - \varepsilon + \varepsilon q)\frac{\sigma_\eta^2}{n}.$$

Риск без искажений

$$r_{\bar{x},0} = \frac{\sigma_\eta^2}{n}.$$

Гарантированный риск

$$r_{\bar{x},+}(\varepsilon) = (1 - \varepsilon_+ + \varepsilon_+ q)\frac{\sigma_\eta^2}{n}.$$

Коэффициент неустойчивости риска

$$k_{\bar{x},+} = (\sqrt{q} - 1)\varepsilon_+.$$

Поскольку мы рассматриваем выборку из равномерного распределения $f_\eta(x) = U(-1,1)$, то имеем $a_\eta = E(\eta) = 0$, $\sigma_\eta^2 = \frac{1}{3}$.

Перейдем к выборочной медиане $x_{(2)}$. Риск для этой оценки имеет вид

$$r_{x_{(2)}}(\varepsilon) = E((x_{(2)} - a_\eta)^2) = E(x_{(2)}^2) - 2a_\eta E(x_{(2)}) + a_\eta^2 = E(x_{(2)}^2).$$

Мы видим, что риск представляет собой начальный момент второго порядка данной статистики, который рассчитывается по формуле

$$E(x_{(2)}^2) = \int_{-\infty}^{\infty} x^2 f_{3,2}(x)dx,$$

где $f_{3,2}(x)$ – плотность вероятности статистики $x_{(2)}$ (второй порядковой статистики при объеме выборки $n = 3$). Как известно из раздела 1.9,

$$f_{3,2}(x) = 2F_\xi(x)[1 - F_\xi(x)]f_\xi(x).$$

Используя выражения $f_\xi(x)$, $F_\xi(x)$, полученные для равномерной модели Тьюки-Хьюбера в разделе 11.5, будем иметь

$$E(x_{(2)}^2) = \int_{-\infty}^{\infty} x^2 f_{3,2}(x)dx = 2\int_{-\sqrt{q}}^{1} x^2 F_1(x)[1-F_1(x)]f_1(x)dx +$$

$$+2\int_{-1}^{1} x^2 F_2(x)[1-F_2(x)]f_2(x)dx + 2\int_{1}^{\sqrt{q}} x^2 F_3(x)[1-F_3(x)]f_3(x)dx.$$

Вычислив эти интегралы, получим

$$r_{x_{(2)}}(\varepsilon) = \frac{-1}{60a^2}(-3\varepsilon^3 + 3\varepsilon^2 - 16\varepsilon^2 a + 8\varepsilon^3 a + \varepsilon^3 a^4 - 5\varepsilon^2 a^4 - 4a^2 +$$

$$+ 8a\varepsilon - 8a^2\varepsilon + 18a^2\varepsilon^2 - 6a^2\varepsilon^2),$$

где $a = \sqrt{q}$. Мы видим, что риск представляет собой многочлен третьей степени переменной ε. В таблице 11.2 приведены значения рисков $r_{\bar{x}}(\varepsilon)$, $r_{\bar{x}_{(2)}}(\varepsilon)$ для выборочного среднего и выборочной медианы в зависимости от уровня искажений ε. Мы видим, что риск для выборочной медианы, также как и для выборочного среднего, является монотонной функцией уровня искажений ε. По приведенным выше формулам или по таблице 11.2 определяем характеристики робастности выборочного среднего

$$r_{\bar{x},+} = 0{,}222,\ r_{\bar{x},0} = 0{,}111,\ k_{\bar{x},+} = \frac{r_{\bar{x}+} - r_{\bar{x}0}}{r_{\bar{x}0}} = 1.$$

Таблица 11.2

ε	0,0	0,05	0,1	0,2	0,3	0,35	0,4	0,45	0,5
$r_{\bar{x}}(\varepsilon)$	0,111	0,122	0,133	0,156	0,175	0,190	0,2	0,211	0,222
$r_{\bar{x}(2)}(\varepsilon)$	0,067	0,07	0,073	0,081	0,091	0,097	0,103	0,109	0,115

По таблице 11.2 находим характеристики робастности выборочной медианы

$$r_{\bar{x}_{(2)},+} = 0{,}115,\ r_{\bar{x}_{(2)},0} = 0{,}067,\ k_{x_{(2)},+} = 0{,}72.$$

Так как $k_{\bar{x}_{(2)},+} < k_{\bar{x},+}$, то выборочная медиана менее чувствительна к большим выбросам.

Пример 11.3. Найдем характеристики робастности выборочной дисперсии $\bar{s}^2$ как оценки генеральной дисперсии σ_η^2 нормальной модели Тьюки-Хьюбера.

Будем считать, что имеем нормальную модель Тьюки-Хьюбера (11.1) – (11.3). Выборочная дисперсия и ее риск определяются выражениями

$$\bar{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 ,$$

$$r_{\bar{s}^2} = E((\bar{s}^2 - \sigma_\eta^2)^2).$$

Введем обозначения $x_i - a_\eta = \overset{\circ}{x}_i$, $\bar{x} - a_\eta = \frac{1}{n}\sum \overset{\circ}{x}_i = \overset{\circ}{\bar{x}}$, и прибавим и вычтем в скобках выражения для $\bar{s}^2$ величину a_η. Получим

$$\bar{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(\overset{\circ}{x}_i - \overset{\circ}{\bar{x}})^2 = \frac{1}{n}\sum_{i=1}^{n}\overset{\circ}{x}_i^2 - (\frac{1}{n}\sum_{i=1}^{n}\overset{\circ}{x}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\overset{\circ}{x}_i^2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\overset{\circ}{x}_i\,\overset{\circ}{x}_j =$$

$$= \frac{n-1}{n^2}\sum_{i}^{n}\overset{\circ}{x}_i^2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\overset{\circ}{x}_i\,\overset{\circ}{x}_j ,$$

$$E(\bar{s}^2) = \frac{n-1}{n^2}\sum_{i=1}^{n}E(\overset{\circ}{x}_i^2) = \frac{n-1}{n}\sigma_\xi^2 ,$$

$$E(\bar{s}^2) = \frac{n-1}{n}(1-\varepsilon+q\varepsilon)\sigma_\eta^2 .$$

Найдём начальный момент второго порядка. Поскольку

$$(\bar{s}^2)^2 = \frac{(n-1)^2}{n^4}\sum_{i=1}^{n}\sum_{j=1}^{n}\overset{\circ}{x}_i^2\,\overset{\circ}{x}_j^2 - \frac{2(n-1)}{n^4}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{\substack{l=1\\ l\neq j}}^{n}\overset{\circ}{x}_i^2\,\overset{\circ}{x}_j\,\overset{\circ}{x}_l + \frac{1}{n^4}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{l=1}^{n}\sum_{\substack{m=1\\ m\neq l}}^{n}\overset{\circ}{x}_i\,\overset{\circ}{x}_j\,\overset{\circ}{x}_l\,\overset{\circ}{x}_m ,$$

то

$$E((\bar{s}^2)^2) = \frac{(n-1)^2}{n^4}[n\mu_\xi^{(4)} + (n^2-n)\sigma_\xi^4] + \frac{2n(n-1)}{n^4}\sigma_\xi^4 =$$

$$= \frac{(n-1)^2}{n^3}\mu_\xi^{(4)} + \frac{(n-1)^3}{n^3}\sigma_\xi^4 + \frac{2(n-1)}{n^3}\sigma_\xi^4 =$$

$$= \frac{3(n-1)^2}{n^3}(1-\varepsilon+q^2\varepsilon)\sigma_\eta^4 + \frac{(n-1)^3}{n^3}(1-\varepsilon+q\varepsilon)^2\sigma_\eta^4 + \frac{2(n-1)}{n^3}(1-\varepsilon+q\varepsilon)^2\sigma_\eta^4 .$$

Риск выборочной дисперсии будет равен

$$r_{\bar{s}^2}(\varepsilon) = E((\bar{s}^2 - \sigma_\eta^2)^2) = E((\bar{s}^2)^2) - 2\sigma_\eta^2 E(\bar{s}^2) + \sigma_\eta^4 =$$

$$= \frac{2n-1}{n^2}\sigma_\eta^4 + \frac{\sigma_\eta^4}{n^3}\{3(n-1)^2(q^2\varepsilon-\varepsilon) + [(n-1)^3 + 2(n-1)](q\varepsilon-\varepsilon)^2 +$$

$$+ [2(n-1)^3 + 4(n-1) - 2n^2(n-1)](q\varepsilon-\varepsilon)\}.$$

Риск без искажений получим при $\varepsilon = 0$:

$$r_{\bar{s}^2,0} = \frac{2n-1}{n^2}\sigma_\eta^4.$$

Из выражения для риска видно, что это квадратичная функция уровня искажений ε. Можно убедиться (см. таблицу 11.3), что, как и в случае с выборочным средним, риск $r_{\bar{s}^2}(\varepsilon)$ является монотонно возрастающей функцией уровня искажений ε. Это позволяет найти гарантированный риск $r_{\bar{s}^2+}$ как

$$r_{\bar{s}^2,+} = \max_{0\le\varepsilon\le\varepsilon_+} r_{\bar{s}^2}(\varepsilon) = r_{\bar{s}^2}(\varepsilon_+),$$

то есть путем подстановки в выражение риска $r_{\bar{s}^2}(\varepsilon)$ значения $\varepsilon = \varepsilon_+$. Коэффициент неустойчивости риска определяется формулой

$$\kappa_{\bar{s}^2,+} = \frac{r_{\bar{s}^2}(\varepsilon_+) - r_{\bar{s}^2}(0)}{r_{\bar{s}^2}(0)}.$$

В таблице 11.3 приведены значения риска $r_{\bar{s}^2}(\varepsilon)$ в зависимости от уровня искажений ε при $q = 3$, $\sigma_\eta^2 = 1$, $n = 50$.

Таблица 11.3

ε	0	0,05	0,1	0,15	0,2	0,25	0,3	0,35	0,4	0,45	0,5
$r_{\bar{s}^2}$	0,04	0,07	0,11	0,17	0,25	0,25	0,47	0,619	0,77	0,94	1,14

Из таблицы 11.3 получаем, что $r_{\bar{s}^2,0} = 0{,}04$, $r_{\bar{s}^2,+} = 1{,}14$, так что коэффициент неустойчивости риска $k_{\bar{s}^2,+} = \frac{1{,}14 - 0{,}04}{0{,}04} = 27{,}5$. Это очень большое значение,

свидетельствующее о том, что выборочная дисперсия не является робастной оценкой дисперсии нормальной генеральной совокупности. Из таблицы 11.3 видно, что при $\varepsilon = 0{,}05$ риск возрастает почти в 2 раза. Это значит, что достаточно пяти "плохих" наблюдений из ста, чтобы риск увеличился в 2 раза. Более устойчивой оценкой разброса случайной величины относительно ее среднего значения является среднее абсолютное отклонение

$$v = \frac{1}{n}\sum_{i=1}^{n} | x_i - \bar{x} | .$$

Однако следует иметь в виду, что v оценивает не дисперсию σ^2, как $\bar{s}^2$, а величину $\sigma\sqrt{2/n}$.

В заключение отметим, что поскольку получение выражений риска для большинства оценок связано с аналитическими трудностями, то сравнение характеристик робастности для различных оценок часто выполняется с помощью компьютерного статистического моделирования.

11.7 Характеристики робастности на основе функции влияния

В теории робастности некоторые обозначения отличаются от тех, которые были приняты нами в начале данного пособия. Характеристики распределений определяются как функционалы и обозначаются как $T = T(F)$, где $F = F(y)$ – функция распределения. Оценки характеристик распределений по выборке объема n обозначаются как T_n, эмпирическая функция распределения, полученная по выборке объема n, обозначается $F_n(y)$.

Функционалом называется отображение $T : \Phi \to R$ множества Φ функций распределения в множество действительных чисел R. Это отображение каждой функции распределения $F(y) \in \Phi$ ставит в соответствии действительное число $T \in R$. Любая рассматриваемая функция распределения $F(y)$ является точкой пространства Φ, то есть аргументом функционала. Функционалами, например, являются моменты распределения, в частности, начальные моменты

$$\nu^{(k)} = \int_{-\infty}^{\infty} y^k dF(y),\ k = 0,1,2,\ldots .$$

В данном случае формула начального момента записана в виде интеграла Стилтьеса. С помощью интеграла Стилтьеса объединяют случаи непрерывной и дискретной случайных величин. Если случайная величина непрерывная, т. е. существует производная $f(y) = F'(y)$, называемая плотностью вероятности, то $dF(y) = F'(y)dy = f(y)dy$ представляет собой дифференциал функции распределения, и интеграл Стилтьеса представляет собой интеграл Римана

$$\nu^{(k)} = \int_{-\infty}^{\infty} y^k f(y)dy .$$

Для случайной дискретной величины функция распределения $F(y)$ представляет собой ступенчатую функцию. В этом случае интеграл Стилтьеса представляет собой сумму

$$\nu^{(k)} = \sum_{i=1}^{n} y_i^k p_i ,$$

где p_i – вероятности возможных значений случайной дискретной величины.

Оценка T_n функционала T также является функционалом $T_n = T(F_n)$, где F_n – эмпирическая функция распределения. Например, выборочные начальные моменты определяются выражением

$$\overline{\nu}^{(k)} = \int_{-\infty}^{\infty} y^k dF_n(y) = \frac{1}{n}\sum_{i=1}^{n} y_i^k .$$

В теории робастности на основе функций влияния анализируются генеральные характеристики, и по результатам этого анализа судят о свойствах соответствующих оценок.

В основе рассматриваемого подхода лежит понятие функции влияния, введенной в рассмотрение Хампелем [19].

Функцией влияния функционала $T(F)$ называется функция $IF(x;T,F)$ (influence function), определяемая выражением

$$IF(x;T,F)=\lim_{t\to 0}\frac{T((1-t)F(y)+th(y-x))-t(F(y))}{t}, \qquad (11.4)$$

где $h(z)$ – функция Хевисайда (Heviside)

$$h(z)=\begin{cases}1, & z\geq 0,\\ 0, & z<0,\end{cases}$$

а $h(y-x)$ – смещенная в точку x функция Хевисайда

$$h(y-x)=\begin{cases}1, & y\geq x,\\ 0, & y<x.\end{cases}$$

Функция влияния (11.4) определяет чувствительность функционала к скачкообразному изменению функции распределения в некоторой точке x. Она представляет собой следующую производную:

$$IF(x;T,F)=\frac{d}{dt}T((1-t)F(y)+th(y-x))\Big|_{t=0}.$$

Это значит, что в функционале $T(F)$ мы заменяем функцию распределения $F(y)$ функцией $F_t(y)=(1-t)F(y)+th(y-x)$ и находим производную

$$\frac{d}{dt}F_t(y)\Big|_{t=0}.$$

Заметим, что функция $F_t(y)$ представляет собой модель Тьюки-Хьюбера при скачкообразной функции распределения выбросов.

На основании функции влияния вводятся следующие характеристики робастности.

1. Чувствительность к большой ошибке функционала T в точке F

$$\gamma^*=\sup_x\left|IF(x;T,F)\right|$$

2. Чувствительность к малым искажениям или к локальному сдвигу

$$\lambda^*=\sup_{y\neq x}\frac{\left|IF(y;T,F)-IF(x;T,F)\right|}{\left|y-x\right|}$$

Эта характеристика определяет чувствительность функционала T к небольшим ошибкам выборочных данных, которые могут появиться, например, в результате округления. В этом случае значение в точке x как бы убирается, и

вместо него вводится значение в точке y так, что мы получаем разность $IF(y;T,F)-IF(x;T,F)$. Эффект от малых изменений наблюдений называется "ерзаньем", а λ^* является характеристикой чувствительности к "ерзанью".

3. Точка удаления ρ^* представляет собой наименьшее значение аргумента функции влияния, при котором она обращается в ноль. Для симметричных распределений

$$\rho^* = inf\left(r>0 : IF(x;T,F)=0, \forall x>r\right).$$

Если величина γ^* для некоторого функционала T конечна, то функционал называется B-робастным, а если бесконечна – не B-робастным. Характеристика λ^* является менее важной по сравнению с γ^*. Если для функционала T ρ^* конечно, то это значит, что все наблюдения больше ρ^* в оценке удаляются.

Положительной особенностью функции влияния является то, что с ее помощью можно вычислять асимптотическую дисперсию оценки. Дисперсия функционала T определяется выражением

$$V(T,F)=\int\left[IF(y;T,F)\right]^2 dF(y),$$

а асимптотическая дисперсия оценки T_n – выражением

$$V(T_n)=\frac{1}{n}V(T,F).$$

11.8 Анализ робастности выборочного среднего

Пусть $a=E(\eta)$ – математическое ожидание распределения $F(x)$:

$$a=T(F)=\int y dF(y).$$

Оценкой математического ожидания является функционал T_n

$$T_n=\bar{x}=\int y dF_n(y),$$

где $F_n(y)$ – эмпирическая функция распределения. Нашей задачей в данном разделе является исследование робастности оценки T_n на основе функции влияния. Получим выражение функции влияния для функционала $T(F) = \int y dF(y)$. Для этого найдем предел

$$IF(x;T,F) = \lim_{t\to 0} \frac{T((1-t)F(y) + th(y-x)) - T(F(y))}{t} =$$

$$= \lim_{t\to 0} \frac{1}{t}\left[\int y d((1-t)F(y) + th(x))(y) - \int y dF(y)\right] =$$

$$= \lim_{t\to 0} \frac{1}{t}\left[(1-t)\int y dF(y) + t\int y dh(y) - \int y dF(y)\right] =$$

$$= \lim_{t\to 0} \frac{1}{t}[(1-t)a + tx - a] = x - a.$$

Таким образом, $IF(x;T,F) = x - a$. Эта функция представляет собой прямую, проходящую под углом 45° через точку $x = a$ (рис. 11.2, функция 1). Так как функция влияния не ограничена, то чувствительность к большой ошибке $\gamma^* = +\infty$. Это значит, что данный функционал не является B – робастным, то есть выборочное среднее $\bar{x}$ не является B-робастной оценкой математического ожидания. Чувствительность к локальному сдвигу для данного функционала $\lambda^* = 1$, что свидетельствует о малой чувствительности выборочного среднего к "ёрзанью". Однако этот положительный факт служит слабым утешением, поскольку эта характеристика важна гораздо меньше, чем γ^*. Точка удаления функционала $\rho^* = +\infty$. Это значит, что никакие наблюдения в оценке не удаляются (имеются в виду резко выделяющиеся наблюдения – выбросы). Таким образом, выборочное среднее достаточно приемлемо, например, при ошибках округления, но совершенно не приемлемо в ситуации, когда возможно появление резко выделяющихся больших наблюдений. Полученная функция влияния позволяет найти асимптотическую дисперсию выборочного среднего

$$V(\bar{x}) = \frac{1}{n}\int [IF(y;T,F)]^2 dF(y) = \frac{1}{n}\int (y-a)^2 dF(y) = \frac{\sigma^2}{n}$$

(впрочем, известную нам из раздела 1.8).

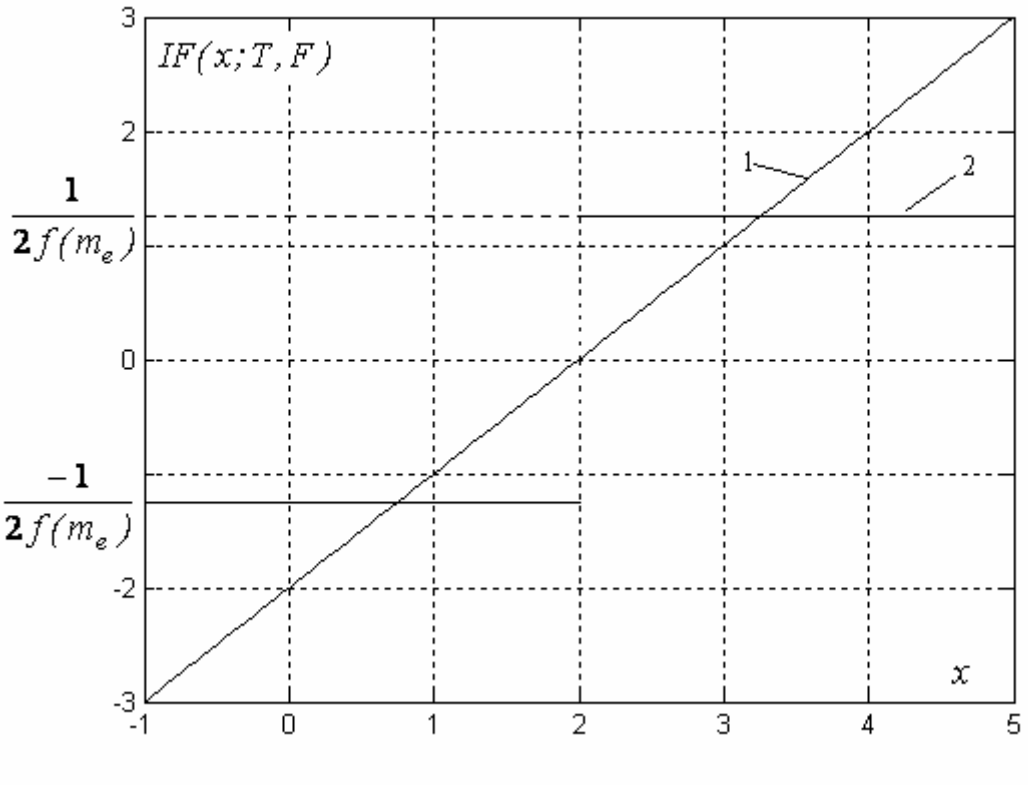

Рис.11.2. Функции влияния среднего значения и медианы

11.9 Получение функции влияния медианы

Медиана m_e распределения $F(y)$ определяется как решение относительно y уравнения $F(y) = 1/2$ и определяется с помощью равенства

$$m_e = med(F(y)) = F^{-1}(1/2).$$

Найдем функцию влияния этой характеристики. Для этого рассмотрим случаи $x \geq m_e$ и $x < m_e$, где x – точка, в которой расположена функция Хевисайда функции влияния. Первый случай означает, что искажающее наблюдение (выброс) расположен справа от медианы, второй – слева. Первый случай представлен на рис. 11.3.

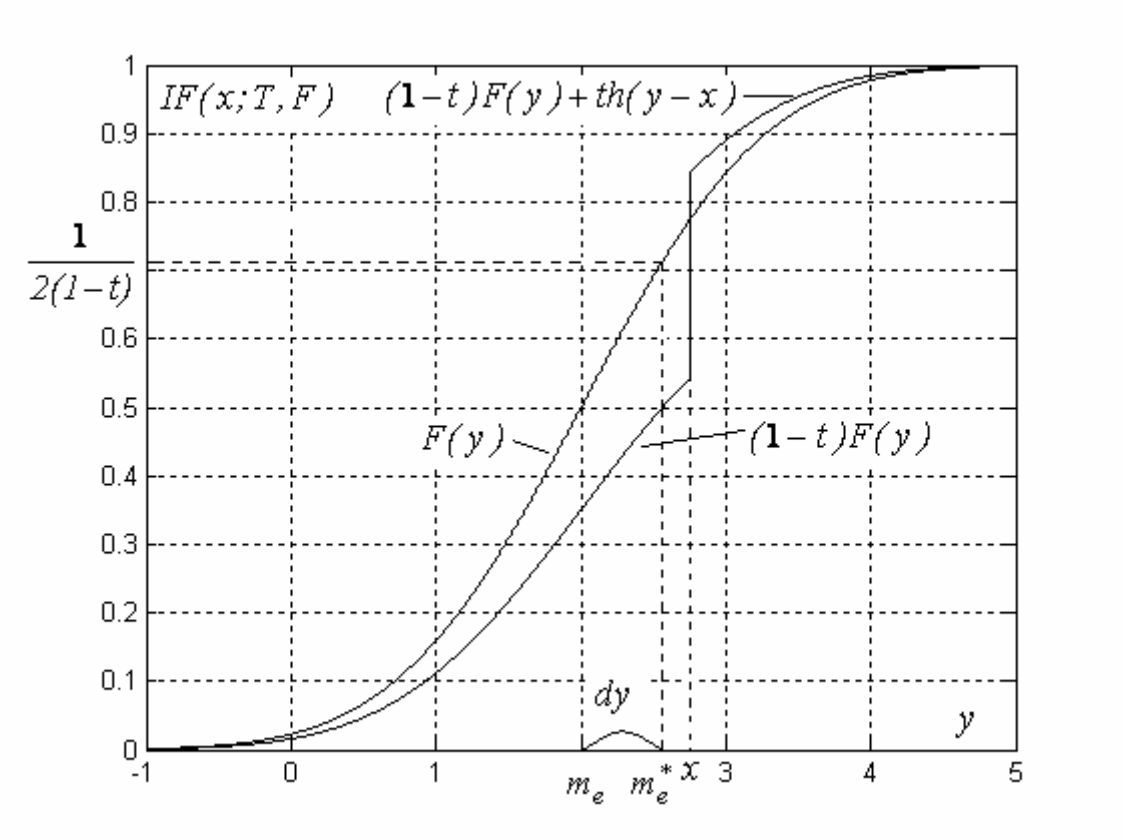

Рис. 11.3. К расчету функции влияния медианы

Найдем функцию влияния для первого случая $x \geq m_e$

$$IF(x;T,F) = \lim_{t\to 0}\frac{1}{t}[T((1-t)F(y) + th(y-x)) - T(F(y))] =$$

$$= \lim_{t\to 0}\frac{1}{t}[med((1-t)F(y) + th(y-x)) - med(F(y))].$$

В первом случае имеем $med((1-t)F(y) + th(y-x)) = med((1-t)F(y))$. Обозначим $med((1-t)F(y)) = m_e^*$. Тогда получим (см. рис. 11.3)

$$IF(x;T,F) = \lim_{t\to 0}\frac{1}{t}(m_e^* - m_e) = \lim_{t\to 0}\frac{1}{t}dy.$$

Поскольку $(1-t)F(m_e^*) = \frac{1}{2}$ и, следовательно, $F(m_e^*) = \frac{1}{2(1-t)}$, а также $F(m_e) = \frac{1}{2}$, то

$$P(Y \in (m_e, m_e + dy)) = F(m_e^*) - F(m_e) = \frac{1}{2(1-t)} - \frac{1}{2} = \frac{t}{2(1-t)}.$$

С другой стороны,

$$P(Y \in (m_e, m_e + dy)) = f(m_e)dy,$$

где $f(x) = F'(x)$ – плотность вероятности. В таком случае

$$f(m_e)\,dy = \frac{t}{2(1-t)},$$

откуда $dy = \dfrac{t}{2(1-t)\,f(m_e)}$. В результате для функции влияния получим выражение

$$IF(x;T,F) = \lim_{t\to 0}\frac{1}{t}dy = \lim_{t\to 0}\frac{1}{t}\frac{t}{2(1-t)f(m_e)} = \frac{1}{2f(m_e)}.$$

Найдем теперь функцию влияния для второго случая $x < m_e$.

$$IF(x;T,F) = \lim_{t\to 0}\frac{1}{t}[T((1-t)F(y) + th(y-x)) - T(F(y))] =$$

$$= \lim_{t\to 0}\frac{1}{t}[med((1-t)F(y) + th(y-x)) - med(F(y))].$$

При $x < m_e$ имеем $med((1-t)F(y) + th(y-x)) = med((1-t)F(y) + t)$. Обозначим $med((1-t)F(y) + t) = m_e^*$. Тогда получим (см. рис. 11.4)

$$IF(x;T,F) = \lim_{t\to 0}\frac{1}{t}(m_e^* - m_e) = \lim_{t\to 0}\frac{1}{t}(-dy).$$

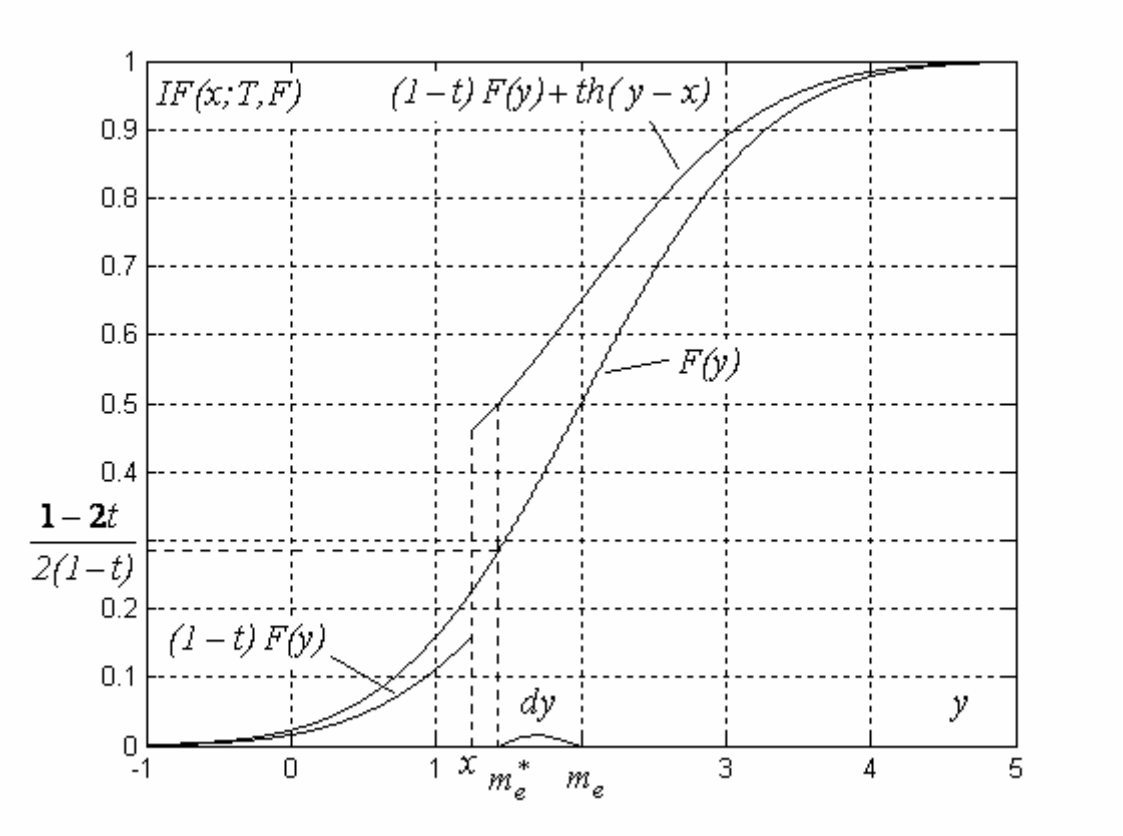

Рис. 11.4. К расчету функции влияния медианы

Поскольку $(1-t)F(m_e^*)+t=\frac{1}{2}$, откуда $F(m_e^*)=\frac{1-2t}{2(1-t)}$, а также $F(m_e)=\frac{1}{2}$, то

$$P(Y\in(m_e,m_e+dy))=F(m_e)-F(m_e^*)=\frac{1}{2}-\frac{1-2t}{2(1-t)}=\frac{t}{2(1-t)}.$$

Так как

$$P(Y\in(m_e,m_e+dy))=f(m_e)dy,$$

где $f(x)=F'(x)$ – плотность вероятности, то

$$f(m_e)\,dy=\frac{t}{2(1-t)},$$

откуда $dy=\frac{t}{2(1-t)\,f(m_e)}$. В результате для функции влияния получим выражение

$$IF(x;T,F)=\lim_{t\to 0}\frac{1}{t}(-dy)=\lim_{t\to 0}\frac{1}{t}\left(\frac{-t}{2(1-t)f(m_e)}\right)=\frac{-1}{2f(m_e)}.$$

Объединяя эти два случая, получим окончательное выражение для функции влияния

$$IF(x;T,F)=\begin{cases}\dfrac{1}{2f(m_e)}, & x\ge m_e\\ \dfrac{-1}{2f(m_e)}, & x<m_e\end{cases}=\frac{sign(x-m_e)}{2f(m_e)},$$

где

$$sign(x)=\begin{cases}1, & x>0,\\ -1, & x<0.\end{cases}$$

Полученная функция влияния приведена на рис. 11.2, функция 2.

11.10 Анализ робастности выборочной медианы

Поскольку функция влияния медианы (рис. 11.2, функция 2) ограничена, то чувствительность к большей ошибке $\gamma^*=\frac{1}{2f(m_e)}$. Это значит, что выборочная медиана является B-робастной оценкой математического ожидания. Для нормального закона чувствительность к большой ошибке имеет значение $\gamma^*=\frac{1}{2\sigma\sqrt{2\pi}}$. Чувствительность к локальному сдвигу для данного функционала $\lambda^*=\infty$, поскольку функция влияния претерпевает скачок в точке $x=m_e$. Это свидетельствует о том, что медиана чувствительна к эффекту "ерзанья", то есть к малым ошибкам в данных. Однако этот недостаток не очень важный. Точка удаления функционала $\rho^*=+\infty$. Это значит, что медиана не удаляет резко выделяющиеся наблюдения. Если в выборку добавить одно резко выделяющееся наблюдение, то медиана все-таки потянется в сторону этого наблюдения. Асимптотическая дисперсия выборочной медианы определяется выражением

$$V(\hat{m}_e)=\frac{1}{n}\int\left[IF(y;T,F)\right]^2dF(y)=\frac{1}{n}\int\frac{sign^2(y-m_e)}{4f^2(m_e)}dF(y)=\frac{1}{4nf^2(m_e)}.$$

Например, для нормального распределения $f(m_e) = \frac{1}{\sigma\sqrt{2\pi}}$ и

$$V(\hat{m}_e) = \frac{\pi}{2}\frac{\sigma^2}{n},$$

что больше в $\pi / 2$ раз, чем для выборочного среднего.

11.11 М-оценки

Одним из методов получения робастных оценок является метод получения так называемых М-оценок. М-оценки являются обобщением максимально правдоподобных оценок, то есть МП-оценок.

Как известно, МП-оценка параметра T максимизирует функцию правдоподобия $\prod_{i=1}^{n} f(x_i, T)$, где $f(x,T)$ – плотность вероятности генеральной совокупности. Обычно задачу на максимум функции правдоподобия сводят к следующей задаче на минимум:

$$\sum_{i=1}^{n} - ln\, f(x_i, T) \to \min_T . \qquad (11.5)$$

Обобщением задачи (11.5) является задача

$$\sum_{i=1}^{n} \rho(x_i, T) \to \min_T . \qquad (11.6)$$

Если функция ρ в (11.6) имеет производную $\psi(x,T) = \frac{\partial}{\partial T}\rho(x,T)$, то задача (11.6) сведется к решению уравнения

$$\sum_{i=1}^{n} \psi(x_i, T) = 0 . \qquad (11.7)$$

Оценки, получаемые как решение задачи (11.6) (или (11.7)), называются М-оценками.

Понятно, что МП-оценки являются частным случаем М-оценок при $\rho(x,T) = ln\, f(x,T)$.

Для М-оценок можно получить выражение функции влияния следующим образом. Аналогом уравнения (11.7) является следующее интегральное уравнение относительного функционала T

$$\int \psi(x, T(F))dF(x) = 0. \qquad (11.8)$$

Уравнение (11.8) позволяет получить функцию влияния. Для этого нужно в (11.8) заменить $F(x)$ на

$$F_t(y) = (1-t)F(y) + th(y-x) \qquad (11.9)$$

и взять от нее производную по t при $t = 0$. Поскольку функционал $T(F)$ не задан в явном виде, то воспользуемся правилами неявного дифференцирования. Подставим (11.9) в (11.8),

$$\int \psi(y, T(F_t(y)))dF_t(y) = 0,$$

и выполним дифференцирование под интегралом:

$$\int \frac{\partial}{\partial T}\psi(y,T)\frac{d}{dt}T(F_t)dF_t(y) + \int \psi(y, T(F_t))dF_t(y) = 0.$$

При $t = 0$ получим

$$\int \frac{\partial}{\partial T}\psi(y,T)\frac{d}{dt}T(F)dF(y) + \psi(x, T(F)) = 0.$$

откуда находим функцию влияния

$$IF(x; T, F) = \frac{\psi(x,T)}{-\int \frac{\partial}{\partial T}\psi(y,T)dF(y)}, \qquad (11.10)$$

и дисперсию оценки

$$V(T,F) = \frac{\int \psi^2(x, T(F))dF(x)}{n\left[\int \frac{\partial}{\partial \theta}\psi(y,T)dF(y)\right]^2}.$$

Мы видим, что функция влияния пропорциональна функции ψ, которую мы можем выбирать. Выбрав ψ такой, чтобы она была ограниченной, мы получим B-робастную оценку.

11.12 М-оценки параметра сдвига

Если функция распределения генеральной совокупности $F(x,\theta)$ может быть представлена в виде $F(x-\theta)$, то параметр θ называется параметром сдвига. Например, математическое определение нормального распределения является параметром сдвига.

При получении М-оценки параметра сдвига рекомендуется функцию ψ выбирать в виде $\psi(x,\theta)=\psi(x-\theta)$, причем для симметричного распределения ее следует выбирать несимметричной: $\psi(-x)=-\psi(x)$.

Для параметра сдвига функция влияния приобретает следующий вид:

$$IF(x;T,F)=\frac{\psi(x,T)}{\int\psi'(x,T)dF(x)}$$

где $\psi'(x)$ – производная по x. Дисперсия оценки определяется выражением

$$V(T,F)=\frac{\int\psi^2(x)dF(x)}{n\left[\int\psi'(x)dF(x)\right]^2}.$$

Посмотрим, какой вид имеет функция ψ для МП-оценок. В этом случае $\rho=ln\,f(x,T)$, $\psi=\frac{\partial}{\partial T}\rho(T)=-\frac{f'(x)}{f(x)}$. В частности, для нормального распределения $f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-x^2}{2}}$, $f'(x)=-x\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-x^2}{2}}$, и $\psi=x$. При такой функции ψ функция влияния параметра сдвига неограниченна, в связи с чем МП-оценка параметра сдвига не В-робастна.

Хьюбер предложил для получения робастной оценки параметра сдвига выбирать функцию ψ вида $\psi=x\,min(1,\frac{b}{|x|})$, $b>0$. Вид этой функции приведен на рис. 11.5.

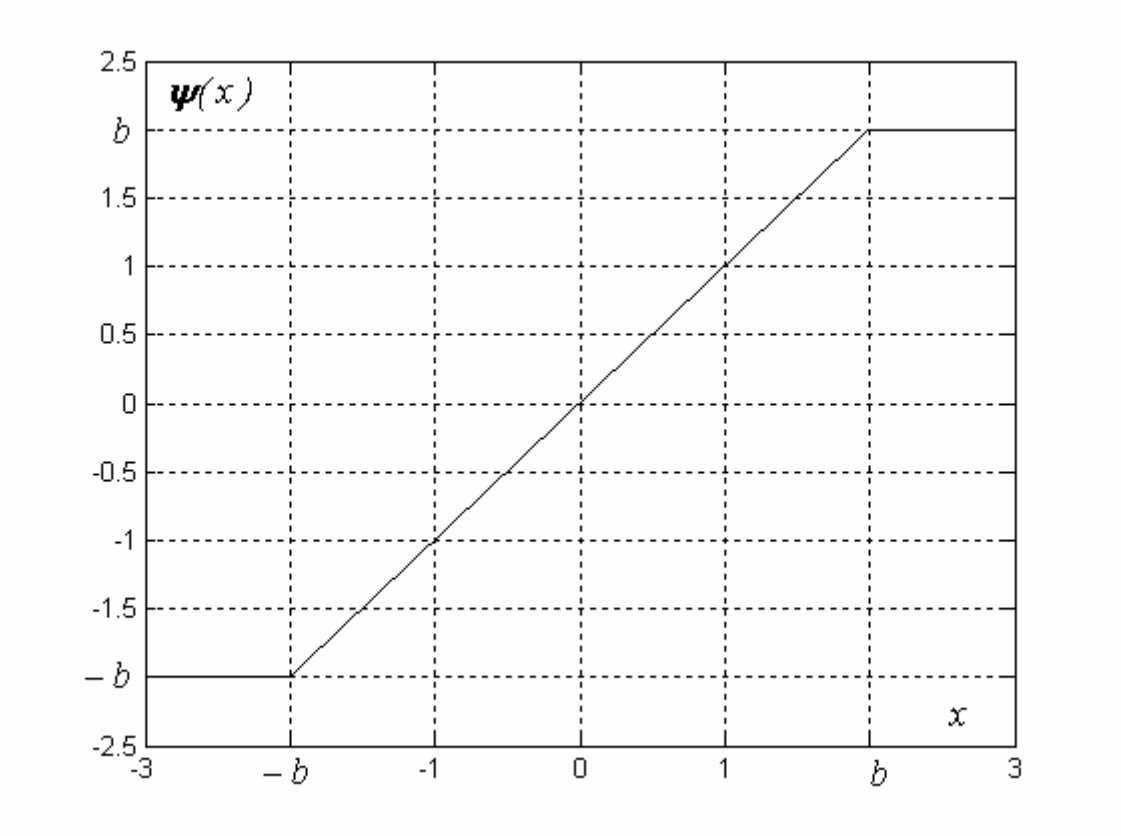

Рис. 11.5. Функция для получения робастной оценки параметра сдвига

11.13 М-оценки параметра масштаба

Параметр θ называется параметром масштаба, если плотность вероятности генеральной совокупности $f(x,\theta)$ может быть представлена в виде $f(x,\theta)=\frac{1}{\theta}f(\frac{x}{\theta})$. Например, дисперсия σ^2 нормального распределения является параметром масштаба.

Для получения робастной М-оценки параметра масштаба необходимо надлежащим образом выбрать функцию ψ в формуле (11.10)). Прежде всего ее следует выбирать в виде $\psi=\frac{1}{\theta}\psi(\frac{x}{\theta})$. Если распределение генеральной совокупности симметрично, то функцию следует выбирать симметричной: $\psi(-\frac{x}{\theta})=\psi(\frac{x}{\theta})$. Для того чтобы представить себе вид такой функции ψ,

посмотрим, какова она в случае известных нам МП-оценок. Для МП-оценок функция ψ представляет собой производную от логарифма плотности вероятности: $\psi = \frac{\partial}{\partial\theta}\ln f(x,\theta)$. Для параметра масштаба получаем

$$\psi = \frac{\partial}{\partial\theta} ln(\frac{1}{\theta} f(\frac{x}{\theta})) = \frac{\partial}{\partial\theta}\left[- ln\theta + ln f(\frac{x}{\theta})\right] = -\frac{1}{\theta} - \frac{f'(x)x}{\theta^2 f(\frac{x}{\theta})}.$$

Обычно функцию ψ записывают при $\theta = 1$, поскольку в этом случае она определяется полностью. В итоге для МП-оценок функция ψ имеет вид

$$\psi = \psi(x) = -x\frac{f'(x)}{f(x)} - 1. \qquad (11.11)$$

В частности, для нормального распределения

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},\ f'(x) = -x\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

и по формуле (11.11) получаем $\psi(x) = x^2 - 1$. Вид этой функции приведен на рис. 11.6. Эта функция неограничена, неограниченной будет и соответствующая функция влияния. Следовательно, максимально правдоподобная оценка параметра масштаба (дисперсии) нормального распределения не B-робастна.

Для получения B-робастной М-оценки параметра масштаба можно взять для любого распределения следующую функцию

$$\psi(x) = \left[-x\frac{f'(x)}{f(x)} - 1 - a\right]_{-b}^{b},$$

где $b > 0$, a – любое действительное число, $[...]_{-b}^{b}$ означает ограничение функции на уровне b. Для нормального распределения эта функция имеет вид

$$\psi(x) = \left[x^2 - 1 - a\right]_{-b}^{b}.$$

Для больших значений b эта функция имеет вид, представленный на рис. 11.7. В этом случае она ограничена сверху. Для малых значений b эта функция имеет вид, представленный на рис. 11.8. В этом случае она ограничена сверху и

снизу. Выбрав такую функцию, мы получим B-робастную М-оценку параметра масштаба.

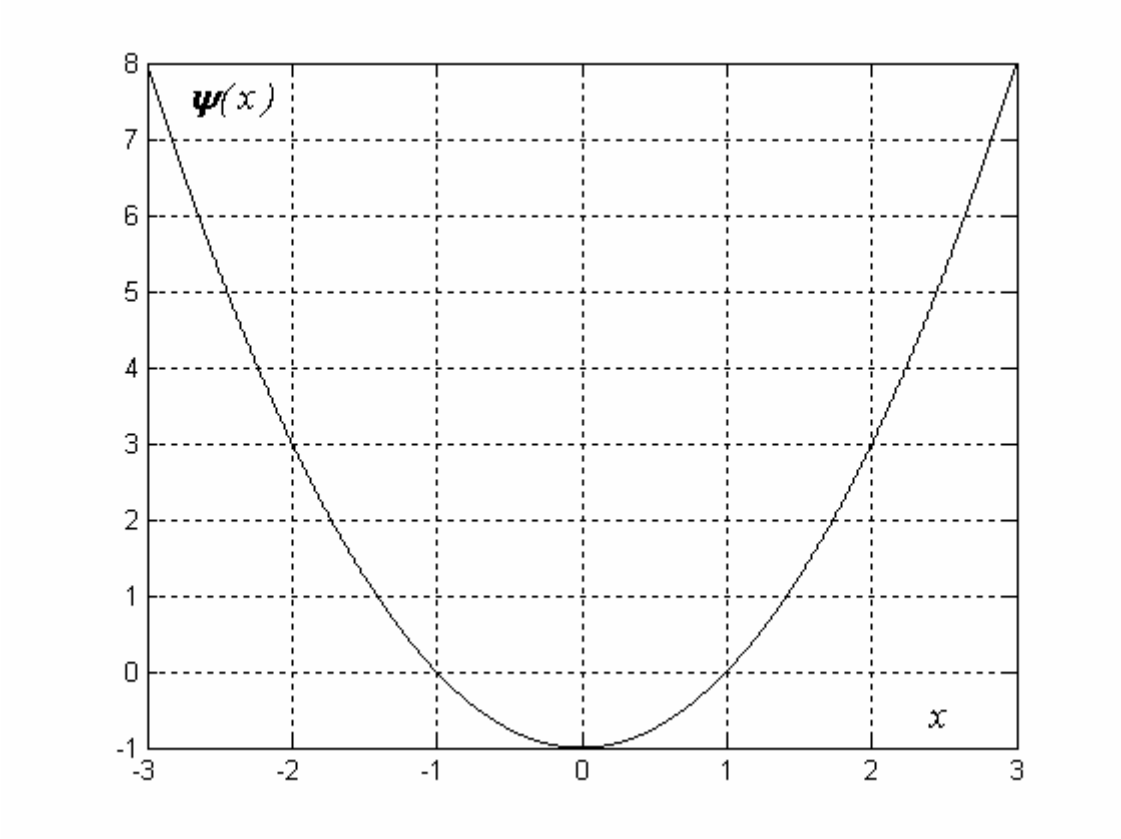

Рис.11.6.

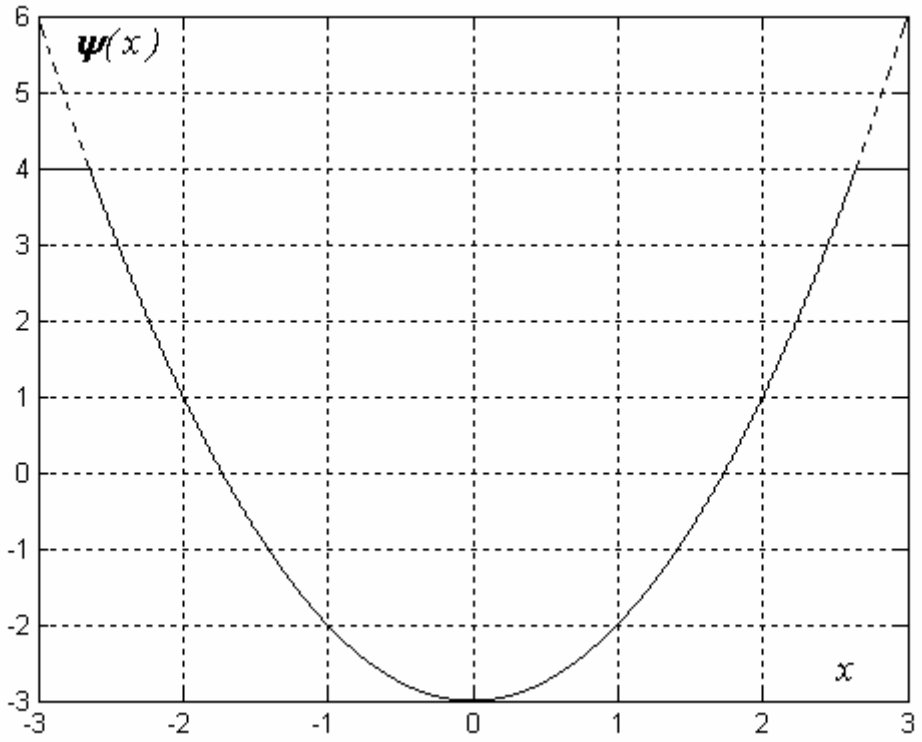

Рис. 11.7.

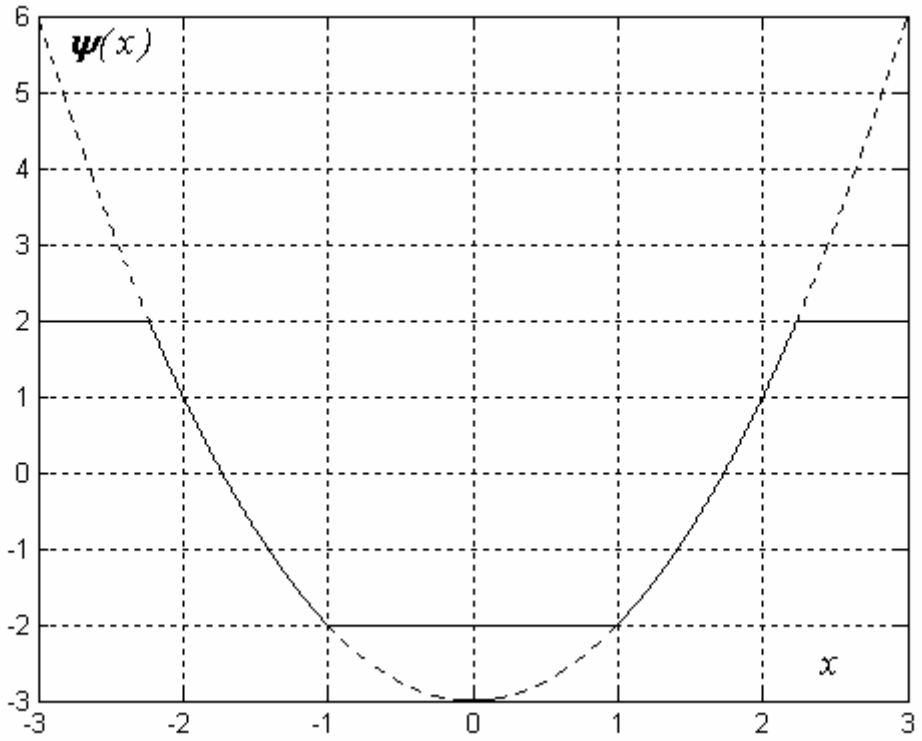

Рис. 11.8.

11.14 Усеченные и винзорированные оценки

Пусть $x_{(i)}$ – порядковая статистика и

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_{(i)} \; - \qquad (11.12)$$

выборочное среднее. Если отбросить в выражении (11.12) $100\alpha\,\%$ статистик справа и слева, то мы получим так называемую α-усеченную оценку, которая записывается в виде

$$\bar{x}' = \frac{1}{n - [\alpha n]} \sum_{i=[\alpha n]+1}^{n-[\alpha n]} x_{(i)} ,$$

где $[\alpha n]$ – целая часть αn. Такая оценка обладает большей устойчивостью к выбросам по сравнению с выборочным средним. Другой прием получения устойчивых к выбросам оценок состоит в том, что $[\alpha n]$ порядковых статистик мы не отбрасываем, а переносим в ближайшую оставляемую точку. В результате получаем оценку вида

$$\bar{x}'' = \frac{1}{n} \sum_{i=[\alpha n]+2}^{n-[\alpha n]-1} x_{(i)} + ([\alpha n] + 1)(x_{([\alpha n]+2)} + x_{(n-[\alpha n]-1)}) .$$

Это так называемая α-винзорированная оценка (по имени ученого Винзора, впервые ее предложившего).

12 СТОХАСТИЧЕСКАЯ АППРОКСИМАЦИЯ

Термином стохастическая аппроксимация объединены асимптотические процедуры (алгоритмы), предназначенные для отыскания решения в условиях помех [7, 24].

12.1 Процедура Роббинса-Монро

В теории численных методов известен метод простой итерации для нахождения корня уравнения

$$f(x) = 0 \tag{12.1}$$

в промежутке $(a,b) \subseteq R$. В соответствии с этим методом обе части (12.1) умножаются на $(-\lambda)$, к обеим частям прибавляется x, в результате чего уравнение (12.1) представляется в виде $x = \varphi(x)$, где $\varphi(x) = x - \lambda f(x)$, и организуются последовательные приближения по формуле

$$x_{n+1} = \varphi(x_n),\ n = 1,2,\ldots . \tag{12.2}$$

Доказано, что процедура (12.2) сходится к корню x^* уравнения (12.1), если $|\varphi'(x)| < 1$ для любых x на всех итерациях.

Процедура Роббинса-Монро является обобщением метода простой итерации на случай наличия ошибок в измерениях. Предполагается, что $f(x)$ − функция регрессии некоторой случайной величины ξ на x, то есть $f(x) = E(\xi / x)$. Требуется найти нуль функции регрессии, то есть корень уравнения

$$f(x) = E(\xi / x) = 0$$

по наблюдениям $y_i = f(x_i) + e_i$ случайной величины ξ, где e_i − ошибки наблюдений, представляющие собой независимые случайные величины с нулевым математическим ожиданием.

Процедура Роббинса-Монро для решения этой задачи формулируется в виде следующей теоремы.

Теорема 12.1. Пусть случайная величина ξ определяется выражением $\xi = f(x) + \eta$, где x – неслучайная переменная, η – случайная величина с $E(\eta) = 0$, и $y_n = f(x_n) + e_n$, $n = 1,2,...$, – наблюдения случайной величины ξ. Процедура

$$x_{n+1} = x_n - a_n y_n,$$

где $a_1, a_2, ...$ – последовательность положительных чисел, сходится по вероятности к корню x^* уравнения $f(x) = 0$ при выполнении условий

$$\sum_{n=1}^{\infty} a_n = \infty, \tag{12.3}$$

$$\sum_{i=n}^{\infty} a_n^2 < \infty, \tag{12.4}$$

$$D(e_n) < \infty.$$

Условие (12.4) означает, что $a_n \underset{n\to\infty}{\to} 0$, то есть коэффициенты должны уменьшаться с возрастанием n. Условие (12.3) означает, что убывание коэффициентов должно быть не очень сильным. Условиям (12.3), (12.4) удовлетворяет, например, следующая последовательность:

$$a_n = \frac{c}{n^\gamma}, \ \frac{1}{2} < \gamma \le 1.$$

В частности, при $c = 1$, $\gamma = 1$ получаем гармонический ряд $a_n = \frac{1}{n}$, который обеспечивает сходимость процедуры Роббинса-Монро.

12.2 Процедура Кифера-Вольфовица

Процедура Кифера-Вольфовица предназначена для поиска максимума функции регрессии. Она является обобщением процедуры поиска максимума детерминированной функции одной переменной.

Пусть в некотором промежутке $(a,b) \subseteq R$ детерминированная функция $f(x)$ имеет максимум x^*. Для поиска максимума можно организовать следующий итерационный процесс:

$$x_{n+1} = x_n + \frac{a_n}{c_n}(f(x_n + c_n) - f(x_n - c_n)),\ n = 1,2,\dots,$$

где a_n, c_n − последовательности положительных чисел. При некоторых условиях на коэффициенты a_n, c_n и на функцию $f(x)$ данный процесс сходится к корню уравнения, то есть $x_n \underset{n\to\infty}{\to} x^*$.

Процедура Кифера-Вольфовица является обобщением данной процедуры на случай, когда $f(x)$ − функция регрессии.

Теорема 12.2. Пусть $f(x)$ − функция регрессии величины ξ на x, то есть $f(x) = E(\xi / x)$, $\xi = f(x) + \eta$, где η − случайная величина с $E(\eta) = 0$ и

$$y(x_i) = f(x_i) + \varepsilon_i,\ i = 1,2,\dots\ ,-$$

наблюдения случайной величины ξ. Процедура вида

$$x_{n+1} = x_n + \frac{a_n}{c_n}(y(x_n + c_n) - y(x_n - c_n))$$

сходится по вероятности к максимуму функции регрессии $f(x)$, если выполняются условия для коэффициентов

$$\sum_{i=1}^{\infty} a_i = \infty,$$

$$\sum_{i=1}^{\infty} a_i c_i < \infty,$$

$$\lim_{i\to\infty} c_i = 0,$$

$$\sum_{i=1}^{\infty}\left(\frac{a_i}{c_i}\right)^2 < \infty,$$

и функция $f(x)$ удовлетворяет условиям:

1) $f(x)$ – одноэкстремальная, то есть строго возрастает при $x < x^*$, и строго убывает при $x > x^*$;

2) $f(x)$ имеет ограниченный рост, то есть существуют коэффициенты $k_1 > 0$ и $k_2 > 0$ такие, что как только $|x_1 - x_2| < k_1$, то $|f(x_1) - f(x_2)| < k_2$;

3) $f(x)$ имеет не слишком большую производную в окрестности x^*, то есть существуют коэффициенты $k_3 > 0$ и $k_4 > 0$ такие, что как только $|x_1 - x^*| + |x_2 - x^*| < k_3$, то $|f(x_1) - f(x_2)| < k_4 |x_1 - x_2|$;

4) $f(x)$ не слишком сильно убывает вдали от точки экстремума, то есть $\forall \varepsilon > 0 \quad \exists \delta(\varepsilon) > 0$ такое, что как только $|x - x^*| > \varepsilon$, то $\inf\limits_{0<h<\varepsilon/2} \dfrac{|f(x+h) - f(x)|}{h} > \delta(\varepsilon)$.

Достоинством метода стохастической аппроксимации является его простота. Недостатком является медленная сходимость к решению и сложные условия сходимости (в частности, трудности с выбором коэффициентов a_n, c_n в процедуре Кифера-Вольфовица).

ЛИТЕРАТУРА

1. Андерсон, Т. Статистический анализ временных рядов / Т. Андерсон.– М.: Мир, 1976. – 756 с.

2. Большев, Л.Н. Таблицы математической статистики / Л.Н. Большев, Н.В. Смирнов. – М.: Наука, 1983. – 416 с.

3. Боровков, А.А. Теория вероятностей / А.А. Боровков. – М.: Наука, 1986. – 432 с.

4. Боровков, А.А. Математическая статистика / А.А. Боровков. – М.: Наука, 1984. – 472 с.

5. Ван-дер-Варден, Б. Математическая статистика / Б. Ван-дер-Варден. – М.: ИЛ, 1960. – 434 с.

6. Вероятность и математическая статистика. Энциклопедия. – М.: БРЭ, 1999. – 912 с.

7. Вазан, М. Стохастическая аппроксимация / М. Вазан. – М.; Мир, 1972. – 296 с.

8. Вучков, И.Н. Прикладной линейный регрессионный анализ / И.Н. Вучков, Л. Бояджиева, Е. Солаков. – М.: Финансы и статистика, 1987. – 238 с.

9. Гаусс, Ф.К. Избранные геодезические сочинения. Т. 1. Способ наименьших квадратов / Ф.К. Гаусс. – М.: Изд-во геодезич. лит.-ры, 1957. – 157 с.

10. Гельфанд, И.М. Вариационное исчисление / И.М. Гельфанд, С.В. Фомин. – М.: Физматгиз, 1961. – 228 с.

11. Де Грот, М. Оптимальные статистические решения / М. Де Гроот. – М.: Мир, 1974. – 496 с.

12. Крамер, Г. Математические методы статистики / Г. Крамер. – М.: Мир, 1975. – 648с.

13. Левин, Б.Р. Теоретические основы статистической радиотехники. Книга вторая / Б.Р. Левин. – М.:Сов. радио,1968. – 504 с.

14.Муха, В.С. Теория вероятностей: Учебное пособие для студентов технических специальностей высших учебных заведений / В.С. Муха. – Мн.: БГУИР, 2001. – 167 с.

15.Рао, С.Р. Линейные статистические методы и их применение / С.Р. Рао. – М.: Наука, 1968. – 548 с.

16.Смирнов, Н.В. Курс теории вероятностей и математической статистики для технических приложений / Н.В. Смирнов, И.В. Дунин-Барковский. – М.: Наука, 1969. – 512 с.

17.Стрейц, В. Метод пространства состояний в теории дискретных линейных систем управления / В. Стрейц. – М.: Наука, 1985. – 296 с.

18.Уилкс, С. Математическая статистика / С. Уилкс. – М.: Наука, 1967. – 632 с.

19.Хампель, Ф. Робастность в статистике: Подход на основе функций влияния / Ф. Хампель, Э. Рончетти, П. Рауссеу, В. Штаэль. – М.: Мир, 1989. – 512 с

20.Харин, Ю.С. Робастность в статистическом распознавании образов / Ю.С. Харин. – Мн.: Университетское, 1992. – 232 с.

21.Харин, Ю.С. Математическая и прикладная статистика / Ю.С. Харин, Е.Е. Жук. – Мн.: БГУ, 2005. – 279 с.

22.Химмельблау, Д. Анализ процессов статистическими методами / Д. Химмельблау. – М.: Мир, 1973. – 960 с.

23.Хьюбер, П. Робастность в статистике / П. Хьюбер. – М.: Мир, 1984. – 304 с.

24.Юдин, Д.Б. Математические методы управления в условиях неполной информации / Д.Б. Юдин. – М.: Сов. радио, 1974. – 400 с.

ОГЛАВЛЕНИЕ