**данным имеет определенные достоинства, но не лишен и недостатков. В связи с этим мы сочли целесообразным опубликовать наряду с самой статьей высказывания ряда других авторов по данной статье и в связи с ней. Указанные заметки решено поместить в полном объеме, чтобы дать возможность оппонентам обсудить не только данную статью, но и в целом проблему построения зависимостей при нарушении стандартных предпосылок регрессионного анализа. Содержание статьи обсуждалось на семинаре «Многомерный статистический анализ и вероятностное моделирование реальных процессов», где в общем получило положительную оценку.**

УДК 519.24

©*А. П. ВОЩИНИН, А. Ф. БОЧКОВ, Г. Р. СОТИРОВ*

# МЕТОД АНАЛИЗА ДАННЫХ ПРИ ИНТЕРВАЛЬНОЙ НЕСТАТИСТИЧЕСКОЙ ОШИБКЕ

**Московский энергетический институт,
Софийский Высший машинно-электротехнический институт**

В прикладных исследованиях важное место занимает проблема построения моделей объектов по данным эксперимента, содержащего ошибки, когда объект описывается уравнением

$$y(x)=y_0(x)+e(x), \qquad (1)$$

где $x$ — вектор входных переменных; $y(x)$ — измеренное значение выходной переменной; $y_0(x)$ — неизвестное истинное значение выходной переменной; $e(x)$ — ошибка эксперимента.

Наиболее изучена принятая в математической статистике модель случайной ошибки измерения $e(x)$. Однако при проведении экспериментов в условиях заводских лабораторий часто нет оснований постулировать некоторое вероятностное распределение ошибки, например, нормальное. В частности, подобная ситуация имеет место, если ошибка эксперимента определяется в основном метрологическими погрешностями измерительных приборов и методов анализа [1—3].

Для метрологов естественной является модель нестатистической ошибки измерения $e(x)$, ограниченной известным предельным значением $\varepsilon(x)$, когда для любого $x$

$$|e(x)|\leqslant \varepsilon(x), \qquad (2)$$

а величина $\varepsilon(x)$ определяется, например, классом точности измерительного прибора.

В данной работе описан метод интервального анализа данных при нестатистической ошибке эксперимента.

## Исходные гипотезы

$H_1$: Объект описывается уравнением (1), $y_0(x)$ является линейно-параметризованной функцией известного вида

$$y_0(x)=\varphi^{\tau}(x)\beta, \qquad (3)$$

где $\varphi^{\tau}(x)=(\varphi_1(x),...,\varphi_m(x))^{\tau}$ — вектор известных базисных функций; $\beta=(\beta_1...\beta_m)^{\tau}$ — вектор неизвестных параметров.

$H_2$: Экспериментом называется совокупность

$$x_i, [y_i^-;\ y_i^+],\ i=1...N, \qquad (4)$$

т. е. в каждом $i$-том опыте известен интервал $[y_i^-;\ y_i^+]$* возможных значений выходной переменной. Предполагается, что

$$y_i^- \leqslant y_0(x_i) \leqslant y_i^+,\ \forall_i=1...N. \qquad (5)$$

Эксперимент может содержать точные измерения, когда $y_i^-=y_i^+$ и параллельные опыты. Предполагается, что $N\geqslant m$.

* Если заданы точечные измерения $y_i$ и известна предельная ошибка $\varepsilon_i$, то $y_i^-=y_i-\varepsilon_i$, $y_i^+=y_i+\varepsilon_i$.

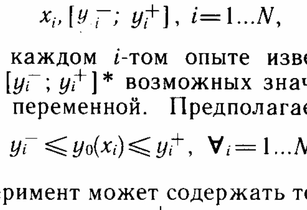


Рис. 1. Линейная модель, адекватная интервальным измерениям

$H_3$: Адекватной моделью объекта является любая функция заданного вида $\hat{y}(x)=\varphi^{\tau}(x)b$, удовлетворяющая условиям

$$y_i^- \leqslant \varphi^{\tau}(x_i)\, b \leqslant y_i^+,\ \forall\ i=1...N, \qquad (6)$$

т. е. проходящая через все интервальные измерения (рис. 1).

**Область возможных значений параметров**

Рассмотрим систему (6), которая при подстановке в нее данных эксперимента является системой $N$ линейных неравенств относительно $m$ неизвестных $b_1...b_m$. Допустим, что система (6) совместна и обозначим множество ее решений

$$\Omega_b = \{b \in R^m : y_i^- \leqslant \varphi^{\tau}(x_i) b \leqslant y_i^+,\ i=1\ ...\ N\}. \qquad (7)$$

Очевидно, выбрав любой вектор $b \in \Omega_b$, можно получить уравнение $y(x)=\varphi^{\tau}(x)b$, которое в соответствии с гипотезой $H_3$ является адекватной моделью объекта. Следовательно, $\Omega_b$ определяет множество оценок $b$ параметров адекватных моделей вида (6). В этом смысле $\Omega_b$ является аналогом доверительной области в регрессионном анализе.

Анализируя выражения (3), (5) и (7), можно прийти к выводу, что $\Omega_b$ одновременно является множеством возможных значений истинных параметров $\beta$, т. е.

$$\Omega_b = \Omega_\beta = \{\beta \in R^m : y_i^- \leqslant \varphi^{\tau}(x_i)\beta \leqslant y_i^+,\ i=1\ ...\ N\}. \qquad (8)$$

В связи с этим в дальнейшем не будем делать различия между $\Omega_b$ и $\Omega_\beta$ и обозначим их одной буквой $\Omega$. Однако необходимо различать фиксированную точечную оценку $b \in \Omega$ и неизвестный вектор параметров $\beta \in \Omega$.

Рассмотрим некоторые свойства множества $\Omega$:

1) Множество $\Omega$ является выпуклым многогранником в пространстве параметров (рис. 2, *б*).

Если обозначить

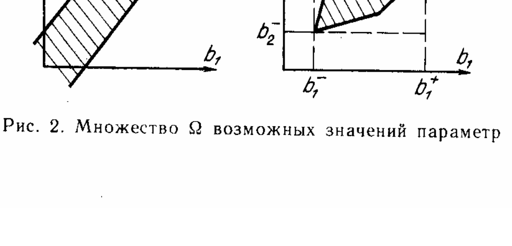


Рис. 2. Множество $\Omega$ возможных значений параметр

$$\mathrm{F} = \begin{pmatrix} \varphi_1(x_1) \ldots \varphi_m(x_1) \\ \vdots \quad\quad \vdots \\ \varphi_1(x_N) \ldots \varphi_m(x_N) \end{pmatrix},$$

$$\rho(\Omega) = \max_{b_i,\, b_j \in \Omega} |b_i - b_j|, \qquad (9)$$

где соответственно $\mathrm{F}(N \times m)$ — матрица значений базисных функций; $\varrho(\Omega)$ — диаметр множества $\Omega$, то справедливы следующие утверждения:

2) если $\mathrm{rang}\,\mathrm{F}=m$, то $\varrho(\Omega) \leqslant M$, т. е. $\Omega$ — ограниченное множество;

3) если $\mathrm{rang}\ \mathrm{F} < m$, то $\rho(\Omega) \to \infty$, т. е. $\Omega$ — неограниченное множество, (рис. 2, *а*);

4) если $\rho(\Omega)=0$, и множество $\Omega=\{b\}$ — включает единственный элемент, то $b=\beta$.

Отметим, что в интервальном анализе последняя ситуация возможна и при конечном числе опытов, чего в принципе не может быть при статистической ошибке;

5) с ростом числа опытов множество $\Omega$ стягивается в точку, совпадающую с истинным вектором $\beta$, т. е. при

$$N \to \infty,\ \rho(\Omega) \to 0,\ \Omega \to \beta. \qquad (10)$$

Это свойство — аналог состоятельности в математической статистике — вытекает из того факта, что с ростом числа опытов растет число условий в выражении (7) и, следовательно, уменьшается размер области, где они одновременно выполняются. (Свойство выполняется при достаточно «хорошо организованном» эксперименте, когда он не пополняется лишь за счет линейно зависимых строк матрицы F).

Для более наглядного описания множества $\Omega$ может оказаться полезной его «аппроксимация» описанной прямоугольной призмой

$$\Pi = \{b \in R^m : b_i^- \leqslant b_i \leqslant b_i^+,\ i=1...m\}, \qquad (11)$$

где границы изменения параметров

$$b_i^- = \min_{b \in \Omega} b_i,\quad b_i^+ = \max_{b \in \Omega} b_i \qquad (12)$$

могут быть найдены как решение соответствующих задач линейного программирования.

Очевидно, при этом выполняется условие $\Omega \subset \Pi$ (рис. 2, *а*).

Анализируя приведенные свойства, можно заключить, что множество $\Omega$ является аналогом доверительной области на параметры модели в регрессионном анализе.

Однако заметим, что, если добавление или исключение $j$-того неравенства в системе (7) не приводит к изменению множества $\Omega$, то $j$-тое наблюдение в интервальном анализе является неинформативным.

## Точечные оценки коэффициентов

Экспериментатору часто необходимо иметь лишь точечную оценку $b$ параметров $\beta$, обладающую некоторыми оптимальными свойствами.

Учитывая, что в рамках постулируемых гипотез никаких вероятностных суждений об ошибке не делается, все точечные оценки $b \in \Omega$ можно считать в известном смысле равноценными.

Ошибку $\Delta b$ произвольной точечной оценки $b \in \Omega$ естественно определить как максимально возможное расстояние между $b \in \Omega$ и неизвестным вектором $\beta \in \Omega$, т. е.

$$\Delta b = \max_{\beta \in \Omega} |b - \beta| = \max_{\beta \in \Omega} \sqrt{\sum_{i=1}^{m} (b_i - \beta_i)^2}\,. \tag{13}$$

С использованием выражения (13) могут быть определены наилучшие в некотором смысле точечные оценки, в частности:

минимаксная

$$b_1 = \min_{b \in \Omega} \max_{\beta \in \Omega} |b - \beta|, \tag{14}$$

наиболее точная в среднем

$$b_2 = \min_{b \in \Omega} \frac{1}{k} \sum_{i=1}^{k} |b_i^* - b|. \tag{15}$$

В формуле (15) $b_i^*$ означает угловые точки множества $\Omega$ и поэтому оценка $b_2$ близка к центру тяжести этого множества.

Очевидно, при любом способе оценивания ошибка $\Delta b$ не превышает величины $\rho(\Omega)$.

Для вычисления оценок $b_1$ и $b_2$ могут быть использованы пакеты программ квадратичного и линейного программирования соответственно.

Возникает вопрос: можно ли использовать обычные МНК-оценки

$$b_{\text{МНК}} = (F^{\text{т}}F)^{-1}F^{\text{т}}Y, \tag{16}$$

вычисленные по средним интервальных измерений, в качестве точечных оценок?

В общем случае ответ отрицательный, так как даже при совместности системы (7) возможна ситуация, когда $b_{\text{МНК}} \notin \Omega$, т. е. модель с МНК-оценками не обеспечивает выполнение условий (6) и, следовательно, является неадекватной интервальным данным эксперимента.

## Интервальная модель выходной переменной

Как уже отмечалось выше, в качестве адекватной модели для выходной переменной может быть выбрана любая функция

$$\hat{y}(x) = \varphi^{\text{т}}(x)b, \quad b \in \Omega. \tag{17}$$

Границы коридора ошибок модели (17) определяются из условия

$$\hat{y}^-(x) = \min_{b \in \Omega} \varphi^{\text{т}}(x)b; \quad \hat{y}^+(x) = \max_{b \in \Omega} \varphi^{\text{т}}(x)b. \tag{18}$$

Это позволяет записать интервальную модель для выходной переменной

$$[y(x)] = [\hat{y}^-(x) \leqslant y_0(x) \leqslant \hat{y}^+(x)], \tag{19}$$

которая определяет границы возможного измерения истинного значения выходной переменной при фиксированном аргументе $x$.

Особенно просто интервальная модель (19) записывается, если вместо множества $\Omega$ используется его аппроксимация призмой (11)

$$[y(x)] = \varphi_1(x)[b_1] + \ldots + \varphi_m(x)[b_m], \tag{20}$$

где $[b_i] = [b_i^-;\ b_i^+]$.

При этом границы коридора имеют вид, показанный на рис. 3. Если $\varphi_i(x) \geqslant 0$, $\forall x$, то

$$\begin{aligned} \hat{y}^-(x) &= \varphi_1(x)b_1^- + \ldots + \varphi_m(x)b_m^-; \\ \hat{y}^+(x) &= \varphi_1(x)b_1^+ + \ldots + \varphi_m(x)b_m^+. \end{aligned} \tag{21}$$

В качестве характеристики точности прогноза при фиксированном $x$ можно взять ширину коридора ошибок

$$\Delta y = \hat{y}^+(x) - \hat{y}^-(x). \tag{22}$$

Наглядной и легко вычисляемой обобщенной характеристикой точности является средняя

$$\bar{\Delta y} = \frac{1}{N} \sum_{i=1}^{N} \Delta y(x_i) \tag{23}$$

или максимальная

$$\max \Delta y = \max_i \Delta y(x_i)$$

ширина коридора, вычисленная по интервальным измерениям.

Легко установить, что при $N \to \infty$ $\Delta y(x_i) \to 0$, т. е.

$$\hat{y}(x) \to y_0(x). \tag{24}$$

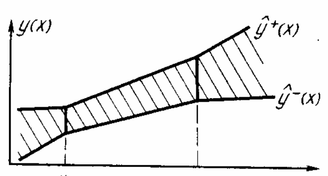


Рис. 3. Границы коридора ошибок интервальной модели

## Проверка значимости коэффициентов. Наилучшая интервальная модель

Учитывая, что $\Omega$ является множеством возможных значений истинных параметров $\beta$, проверка различных гипотез относительно коэффициентов $\beta_i$ оказывается при интервальном анализе очень простой.

Например, если выдвинута гипотеза $H_0{:}\beta=0$, то она принимается, если $0\in\Omega$. В этом случае границы интервалов возможных значений параметров $b_i$ имеют разные знаки, т. е.

$$\operatorname{sign} b_i^- \neq \operatorname{sign} b_i^+, \quad \forall_i = 1...m, \qquad (25)$$

следовательно, все коэффициенты являются незначимыми.

Наоборот, знаки всех коэффициентов $b_i$ достаточно точно идентифицированы (т. е. все коэффициенты значимы), если

$$\operatorname{sign} b_i^- = \operatorname{sign} b_i^+, \quad \forall_i = 1...m. \qquad (26)$$

Если же условие (25) имеет место только для отдельного коэффициента $b_i$, то его можно считать незначимым и исключить из модели.

При этом необходимо отметить парадоксальное на первый взгляд обстоятельство — простая модель является более точной, чем сложная.

Действительно, пусть имеются две модели

$$y_1 = \varphi^{\mathrm{T}}(x) b_m, \qquad (27)$$
$$y_2 = \varphi^{\mathrm{T}}(x) b_m + \psi^{\mathrm{T}}(x) b_k,$$

где $b_m$ и $b_k$ — векторы $(m\times 1)$ и $(k\times 1)$ соответственно, и пусть по результатам одного и того же эксперимента определены множества $\Omega_1$ и $\Omega_2$ возможного значения их параметров. $(\Omega_1\neq\varnothing,\ \Omega_2\neq\varnothing)$. Тогда справедливо следующее выражение

$$\bar{\rho}(\Omega_1)\leqslant\rho(\Omega_2), \quad \max\Delta y_1\leqslant\max\Delta y_2, \qquad (28)$$

т. е. размер множества $\Omega$ и ошибка прогноза более простой интервальной модели $y_1$ по крайней мере не выше, чем у более сложной модели $y_2$.

Эта ситуация иллюстрируется на рис. 4, где изображены интервальные модели $y_1=b_1$ и $y_2=b_1+b_2x$.

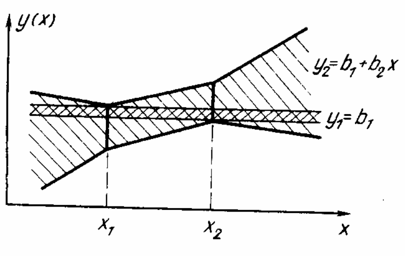


Рис. 4. Две модели, адекватные интервальным измерениям

Приведенное свойство позволяет ввести понятие наилучшей интервальной модели как наиболее простой (с наименьшим числом коэффициентов или наиболее простой структуры) функции, проходящей через все интервальные измерения.

*Пример.* Рассмотрим процедуру анализа интервальных данных на примере исследования котлоагрегата типа БК-3-210, широко используемого на ТЭЦ для выработки пара высокого давления.

Основной эксплуатационной характеристикой агрегата является зависимость $y=f(x)$ производительности агрегата $y$, измеряемой как расход пара на выходе агрегата (т/ч) от расхода топлива $x$ на входе агрегата.

В результате испытаний котлоагрегата после проведения регламентных работ были получены данные трех измерений, причем расход пара измерялся с 5 %-ной относительной ошибкой (табл. 1 и рис. 5).

Котлоагрегаты типа БК-3-210 хорошо изучены энергетиками и известно, что зависимость $y_0=f(x)$ является достаточно гладкой неубывающей функцией.

В связи с этим было решено искать модель агрегата в классе линейных функций $y_0=\beta_1+\beta_2x$ с неотрицательной первой производной. Эти условия приводят к следующей системе неравенств для неизвестных параметров $\beta_1$, $\beta_2$:

$$122\leqslant\beta_1+12\beta_2\leqslant 134;$$
$$171\leqslant\beta_1+16\beta_2\leqslant 189;$$
$$230\leqslant\beta_1+20\beta_2\leqslant 241;$$
$$0\leqslant\beta_2.$$

Множество возможных значений параметров $\beta_1$, $\beta_2$ изображено на рис. 6.

Угловые точки многогранника имеют координаты

$\beta^{(1)}=(6{,}5;\ 10{,}625)$; $\beta^{(2)}=(-26{,}5;\ 13{,}375)$; $\beta^{(3)}=(-56{,}5;\ 14{,}875)$; $\beta^{(4)}=(-25;\ 12{,}25)$; $\beta^{(5)}=(-21;\ 12)$.

Экстремальными угловыми точками, определяющими размер множества $\Omega$, являются точки $\beta^{(1)}$ и $\beta^{(2)}$. При этом $\rho(\Omega)=49{,}8$. В качестве точечных оценок можно взять любую из приведенных ниже оценок:

*Таблица 1. Результаты измерений*

| $i$ | $x_i$ | $y_i$ | $\Delta x_i$ | $[y_i^-;\ y_i^+]$ |
|---|---|---|---|---|
| 1 | 12 | 128 | 6 | [122; 134] |
| 2 | 16 | 180 | 9 | [171; 189] |
| 3 | 20 | 230 | 11 | [219; 241] |

Примечание. $y_i$ — измеренное значение производительности агрегата; $\Delta x_i\approx y_i\cdot 0{,}05$; $y_i^-=y_i-\Delta x_i$; $y_i^+=y_i+\Delta x_i$.

$$b_1 = \frac{\beta^{(1)}+\beta^{(3)}}{2} = (-25;\ 12{,}75);$$

$$b_2 = \frac{\sum_{i=1}^{5}\beta^{(i)}}{5} = (-24{,}5;\ 12{,}625);$$

$b_{\text{МНК}} = (-24{,}66;\ 12{,}75)^{\text{т}}$.

Заметим, что в рассматриваемом случае все три оценки практически совпадают.

Описанный вокруг множества $\Omega$ прямоугольник (рис. 6) задается координатами

$$\Pi = \{b \in R^2{:}-56{,}5 \leqslant b_1 \leqslant 6{,}5;\ 10{,}625 \leqslant b_2 \leqslant 14{,}875\}.$$

Анализируя структуру множества $\Omega$, можно заметить, что в угловых точках $\beta^{(1)}$ и $\beta^{(3)}$ параметр $b_1$ имеет разные знаки и, следо-

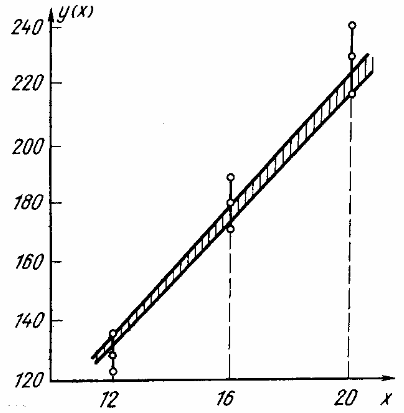


Рис. 5. Интервальная модель котлоагрегата

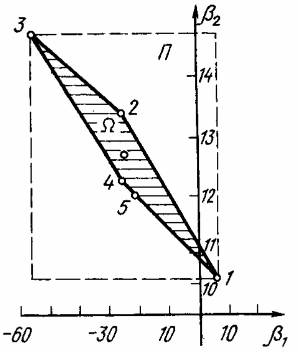


Рис. 6. Множество возможных значений параметров котлоагрегата

вательно, правомерным является проверка и принятие гипотезы $H_0{:}\beta_1 = 0$.

Используя более простую структуру модели: $y_0 = \beta_2 x$ без коэффициента $\beta_1$ и данные эксперимента, получаем следующую систему неравенств:

$$122 \leqslant 12\beta_2 \leqslant 134;$$
$$171 \leqslant 16\beta_2 \leqslant 189;$$
$$219 \leqslant 20\beta_2 \leqslant 241;$$
$$\beta_2 \geqslant 0,$$

решением которой является интервал возможных значений параметра $\beta_2$ $[\beta_2] = [10{,}95;\ 11{,}15]$.

Очевидно, знак параметра $\beta_2$ надежно идентифицирован и соответствует требуемому условию $\beta_2 \geqslant 0$, а максимальная ошибка оценивания сравнительно невелика: $\max \Delta b_2 = 0{,}2$.

В качестве точечной оценки коэффициента $\beta_2$ целесообразно взять оценку $b_2 = 11{,}05$, соответствующую середине интервала.

Относительная ошибка оценивания параметра $\beta_2$ с использованием этой точечной оценки составляет 1 %, что в 5 раз меньше относительной ошибки измерения выхода.

Заметим, что МНК — оценка параметра $\beta_2$, вычисленная по средним $y_i$ и равная $b_{\text{МНК}} = 11{,}27$, не попадает в интервал $[10{,}95;\ 11{,}15]$ возможных значений параметра $\beta_2$ и является в смысле сделанных допущений недопустимой.

Интервальные значения прогноза выходной переменной в точках эксперимента составляют: $[y(12)] = [131{,}4;\ 134]$, $[y(16)] = [175{,}2;\ 178{,}6]$; $[y(20)] = [219;\ 223]$. Коридор ошибок прогноза изображен на рис. 6. Средняя ошибка прогноза составляет 3,3 т/ч, т. е. около 2 % от средней производительности агрегата.

Таким образом по результатам трех интервальных измерений построена достаточно точная и удовлетворяющая физическим ограничениям модель агрегата в виде $[y(x)] = [10{,}95;\ 11{,}15]\,x$.

В распоряжении исследователей имелись результаты испытаний двух агрегатов этого же типа (табл. 2).

Было решено проверить гипотезу о совпадении их эксплуатационных характеристик

$$H_0{:}f_1(x) = f_2(x) = f_3(x).$$

Очевидно, гипотезу $H_0$ можно принять, если существует решение системы

$$y_i^- \leqslant \beta_2(x) \leqslant y_i^+,$$
$$\beta_2 \geqslant 0$$

составленной по результатам девяти измерений, сделанных на всех трех агрегатах.

Несложные расчеты показывают, что в данном случае решение системь существует и задается интервалом $\beta_2 = [10{,}95;\ 11]$.

Таким образом, можно принять гипотезу о совпадении эксплуатационных характеристик однотипных агрегатов.

Кроме того, значение параметра $\beta_2$ определено с очень высокой точностью (0,5 %).

Приведенный пример показывает, что

*Таблица 2. Результаты испытаний двух агрегатов*

| № агрегата | $i$ | $x_i$ | $y_i$ | $\Delta x_i$ | $y_i^-$ | $y_i^+$ |
|---|---|---|---|---|---|---|
| 2 | 1 | 12 | 130 | 7 | 123; | 137 |
| | 2 | 17 | 178 | 9 | 169; | 187 |
| | 3 | 20 | 210 | 11 | 199; | 221 |
| 3 | 1 | 13 | 140 | 7 | 133; | 147 |
| | 2 | 17 | 184 | 9 | 173; | 193 |
| | 3 | 20 | 220 | 11 | 209; | 231 |

предлагаемый метод анализа интервальных данных достаточно хорошо согласуется с физическими представлениями технологов, позволяет естественно и просто учесть всю априорную информацию об ошибке и структуре модели и получить достаточно точные модели даже при небольшом числе измерений.

На наш взгляд, предложенный метод интервального анализа данных базируется на достаточно естественных предположениях, которые хорошо согласуются с опытом и представлениями прикладников. Результаты интервального анализа носят ясный и легко интерпретируемый характер и позволяют с неожиданной стороны взглянуть на сложные для содержательного толкования понятия регрессионного анализа (состоятельность оценок, адекватность модели, значимость коэффициентов, наилучшая модель, информативные наблюдения и т. п.).

Вычислительная процедура метода основана на хорошо разработанных программах линейного и квадратичного программирования, на решении систем линейных неравенств, что позволяет использовать программы линейного программирования.

В принципе метод может быть использован при построении и нелинейных по параметрам моделей, в частности, оценивание параметров отношения полиномов может быть также сведено к линейному случаю.

Предлагаемый метод, который лежит в общем русле активно развиваемых в последнее время методов интервальной арифметики и алгебры, в известном смысле можно рассматривать как альтернативу регрессионному анализу [4—6].

Заключая отметим, что в советско-болгарской лаборатории «Оптимизация производственных систем», сотрудниками которой являются авторы статьи, разработано программное обеспечение метода интервального анализа данных и интервальной оптимизации для персональной ЭВМ типа IBM PC/XT.

## ЛИТЕРАТУРА


1. *Бакан Г. М.* / Автоматика. 1980. № 2. С. 13—21.
2. *Вощинин А. П.* / Заводская лаборатория. 1987. Т. 54. № 7. С. 15—18.
3. *Вощинин А. П.* Метод оптимизации объектов по интервальным моделям целевой функции.— М.: Изд. МЭИ, 1987.— С. 109.
4. *Ицкович Э. Л., Соркин Л. В.* / Автоматика и телемеханика. 1981. № 8. С. 112—110.
5. *Кунцевич В. М., Лычак М. М., Никитенко А. С.* / Кибернетика. 1988. № 1. С. 17—52.
6. *Шокин Ю. И.* Интервальный анализ.— Новосибирск: Наука, 1981.— С. 284.


*Поступила 23.12.1988 г.*

## I

© *В. П. БОРОДЮК*

Содержательный уровень результатов, полученных при обработке данных тем или иным формальным методом, зависит от уровня априорной информации, определяющей границы правомерности использования метода. Реально априорная информация выступает в виде предпосылок (допущений), сопровождающих данный метод. При одинаковом объеме априорной информации предпочтительнее будет метод, обеспечивающий более высокий содержательный уровень результатов. Попытаемся с этих позиций сопоставить предложенный авторами приведенной выше публикации интервальный метод с традиционным методом регрессионного анализа с МНК-оценками, относительно которого сами авторы рассматривают свою процедуру то как несопоставимую (принципиально другую), то как выигрышно альтернативную; т. е. сопоставим предпосылки и содержательный уровень результатов РА и предложенного в статье интервального метода.

Авторы предполагают известным класс прибора для измерения $y(x)$ или предельное значение ошибки $|e_{(x)}|\leqslant\varepsilon_{(x)}$. Обе эти величины через метрологические соотношения несложно пересчитываются в значение стандартного отклонения ошибки $\sigma_{e(x)}$. Следовательно, как и в РА, предполагается заданным условие $\sigma^2_{e(x)}=\text{const}$, либо известно $\sigma^2_{e(x)}$ для всех $x$. Условие симметричности интервалов $[y_i^-;\ y_i^+]$ предполагает отсутствие систематической составляющей измерений или ее постоянство, что равносильно предпосылкам $M[e(x)]=0$ или $M[e(x)]=\mu_i=\text{const}$, как и в регрессионном анализе. При допущении независимости ошибок указанные предпосылки образуют условия Гаусса — Маркова, при их выполнении МНК-оценки в регрессионном анализе являются наилучшими (в смысле дисперсионного критерия точности) и состоятельными в классе линейных несмещенных оценок. Авторы указывают лишь одно свойство оценок $b$, а именно при $N\to\infty$: $\rho(\Omega)\to 0$ и $\Omega\to\beta$ (10)* (аналог состоятельности оценок в регрессионном анализе) и молчаливо предполагают их несмещенность, поскольку, если принять $M[b]\neq\beta$, то условие $\Omega\to\beta$ не будет выполнено (как в регрессионном анализе). Оптимальные свойства оценок в смысле точности авторы закладывают в саму процедуру их получения (15) или (16). По структуре операторов это нелинейные процедуры оценивания, и в принципе на этом пути можно получить оценки, которые в условиях Гаусса — Маркова будут иметь меньшую дисперсию, чем МНК — оценки [1, 2]. Однако процедуры (15) или (16) в общем случае на практике реализовать невозможно, равно как невозможно в общем случае вычислить величину $\rho(\Omega)$. Поэтому вопрос о точности получаемых оценок $b$ остается открытым. Точностные характеристики оценок, полученных в примере, нельзя считать убедительными, поскольку для получения модели $\hat{y}=b_2x$ привлекалась дополнительная априорная информация в виде $\beta_2\geqslant 0$. По этой же причине неправомерно сравнение $b_2$ с МНК — оценкой этого параметра.

Говоря об адекватности модели, авторы молчаливо предполагают существование одной единственной адекватной в смысле набора базисных функций $\varphi_j(x)$ и изначально постулируемой (как в регрессионном анализе) структуры $y^0(x)=\varphi^{\tau}(x)\beta$, так как только в этом случае коридор ошибок $\Delta y(x)=\hat{y}^+(x)-\hat{y}^-(x)$ при фиксированном $x_i$ и $N\to\infty$ будет стягиваться в точку $y(x_i)=\varphi^{\tau}(x_i)\beta$. Это обстоятельство служит еще одним подтверждением неявного присутствия в работе предпосылки о несмещенности оценок $b$.

Теперь попытаемся сопоставить проверку гипотезы $H_0$: $\beta_j=0$ против альтернативной $H_1$:$\beta_j\neq 0$, т. е. сравним работоспособность процедур выявления значимых параметров $\beta_j$ для регрессионного анализа и интервального метода. В регрессионном анализе для этого необходимо ввести предпосылку нормальности распределения ошибки $e(x)$. Интервальный подход такой предпосылки не требует; процедура проверки гипотезы основана на использовании известных интервалов $[y_i^-;\ y_i^+]$. В условиях несравнимых предпосылок представляется разумным сопоставить результаты процедур по их чувствительности (устойчивости) к нарушению своих предпосылок, что наиболее надежно в рассматриваемом случае можно было бы сделать с помощью имитационного моделирования. Поскольку такое исследование авторами не сделано, ограничимся следующими замечаниями.

Как показано в работе [4], критерий отбора значимых коэффициентов в виде отношения $t_j=|b_j|/\hat{\sigma}_{b_j}$ оказался наиболее устойчивым (в смысле потери мощности) к отклонению от нормальности среди других критериев (последовательное включение, последовательное исключение, шаговый метод и др.), показав высокую устойчивость показателя мощности при существенном искажении нормальности. Что касается интервального подхода, то размер и конфигурация области $\Omega_b$ могут оказаться весьма чувствительными к вариации задания ширины интервалов $[y_i^-;\ y_i^+]$, поскольку границы области $\Omega_b$ определяются каждым из таких интервалов. Кроме того, очевидно, что в интервальном методе (как и в регрессионном анализе) для конечных результатов небезразличен порядок проверки на значимость коэффициентов $\beta_j$, так как при удалении незначимого $\beta_j$ конфигурация области $\Omega_b$ будет существенно и непредсказуемо меняться. Однако в отличие от хорошей проработки этого вопроса в регрессионном анализе, авторы не приводят никаких рекомендаций по определению порядка проверки коэффициентов $\beta_j$. Другую важную с точки зрения практики гипотезу $H_0$:$\beta=0$ об отсутствии зависимости между $y$ и всей совокупностью переменных $x_j$, видимо, просто нельзя сформировать и интерпретировать в рамках интервального подхода.

Несколько дополнительных суждений.

Вызывает возражение замечание авторов о «парадоксальном обстоятельстве: простая модель является не менее (а часто и более) точной, чем сложная» (с. 79). Никакого парадокса здесь нет. В регрессионном анализе показано [3, 4], что наибольшей точностью обладает модель, по структуре совпадающая с истинной зависимостью (адекватная). Среди неадекватных моделей модель с «упрощенной» структурой обеспечивает более точные результаты, чем с «усложненной».

Нельзя согласиться с утверждением, что «принципиальным моментом интервального анализа является наличие множества

* См. обсуждаемую статью.

равноценных моделей, одинаково «хорошо» согласующихся с экспериментальными данными». Такая ситуация является типичной для случая, когда по фиксированному набору данных определяются не оценки параметров зависимости $y^0(x)=\varphi^{\tau}(x)\beta$ (адекватная модель), а восстанавливается зависимость $\hat{y}=\varphi^{\tau}(x)b$, наилучшим образом аппроксимирующая экспериментальный материал. Этот эффект неоднократно описан в литературе по регрессионному анализу [3, 5].

На фоне практического совпадения предпосылок регрессионного анализа и интервального подхода и отсутствия безусловных преимуществ оценок $b$ перед МНК нельзя не упомянуть о работе Кощеева [2], в которой получены конструктивные результаты по улучшению МНК-оценок $b_{\text{МНК}}$ за счет привлечения дополнительной априорной информации о диапазоне возможного изменения значений $y$ в виде $c \leqslant y(x) \leqslant d$, что концептуально весьма близко рассматриваемому в статье подходу и требует подробного сравнительного анализа.

Заостренно критическая позиция автора этой заметки обусловлена стремлением глубже вскрыть содержательный смысл результатов, представленных в статье А. П. Вощинина с соавторами и не совпадает с его безоговорочно положительным отношением к пионерской попытке привлечь интервальный подход к решению важной прикладной задачи.

ЛИТЕРАТУРА

1. *Вапник В. Н.* Восстановление зависимостей по эмпирическим данным.— М.: Наука, 1979.— 447 с.

2. *Кощеев В. А.* / В кн.: Статистические методы теории управления.— М.: Наука, 1978.— 315 с.

3. *Себер Дж.* Линейный регрессионный анализ.— М.: Мир, 1980.— 456 с.

4. *Архиреева И. Н., Бородюк В. П., Полвонов Ф. Ю.* / Заводская лаборатория. 1987. Т. 53. № 10. С. 94—99.

5. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ.— М.: Финансы и статистика, 1987. Вып. I.— 241 с.

## II

© *Е. З. ДЕМИДЕНКО*

Хорошо известно, что классический подход исследования экспериментальных зависимостей обладает рядом недостатков, являющихся следствием принятия системы гипотез и предпосылок, необходимых для проведения соответствующего анализа. Одна из таких гипотез — постулирование случайной природы отклонений $e(x)$.

В статье А. П. Вощинина, А. Ф. Бочкова и Г. Р. Сотирова предлагается отойти от указанного предположения и считать отклонения детерминированными. Теперь вместо априорного распределения авторы предполагают известными интервалы изменения этих ошибок, которые связываются исключительно с метрологическими погрешностями зависимой переменной $y_0(x)$. В этом случае моделей, удовлетворяющих этому ограничению, может как не существовать, так и наоборот, быть бесконечно много (в рамках предлагаемого подхода это свойство является фундаментом для построения доверительных интервалов коэффициентов модели). Образно при $m=2$ рассматриваемый подход можно представить себе как попытку уложить палку посреди $N$ пар гвоздей, вбитых в пол на определенном расстоянии. Если таких пар очень много или они достаточно криволинейно расположены, то ни при одном положении палка не может уместиться между гвоздями. Если же найдется хотя бы одно удовлетворительное положение, то весьма вероятно, что покачиванием между зажатыми гвоздями можно менять исходное положение палки.

Применение данного подхода неявно ведет к следующим двум важным посылкам:

а) используемая модель абсолютно точна (если бы измерения были бы точны, то модель «прошла» бы через все экспериментальные точки);

б) измеряется только зависимая переменная и считается, что независимые переменные измерены абсолютно точно.

На этих двух моментах хочется остановиться более подробно. Сразу следует заметить, что обе посылки малореалистичны. Вряд ли можно найти достаточное количество примеров, где вид оцениваемой зависимости и набор определяющих факторов известен абсолютно точно. Это иллюстрируют сами авторы на своем примере с исследованием котлоагрегата: видимо они не будут настаивать на том, что выбранная линейная зависимость является абсолютно точной. Вместе с тем в классической модели отклонения могут содержать как ошибки измерения, так и ошибки спецификации (неправильно выбранная форма связи, отсутствие необходимого фактора и т. д.). Как следствие получаем следующий парадокс: чем точнее произведены измерения, тем вероятнее в рассматриваемом подходе отсутствие решения.

Реалистично предположить, что независимая переменная тоже измеряется с ошибкой (в действительности, в предлагаемом примере трудно представить, что расход пара измеряется с ошибкой, а расход топлива абсолютно точно). Здесь есть резервы для дальнейшего теоретического анализа предлагаемого интервального подхода. Заметим лишь, что подобной однобокостью страдает и классический регрессионный анализ. В частности, последствия влияния ошибок в независимых переменных на точность оценки МНК исследуются в моей статье «О понятии устойчивости в регрессионном анализе» (Заводская лаборатория. 1989. № 5).

© Э. К. ЛЕЦКИЙ

# III

По-видимому, не вызовет возражений соображение о том, что чем больше сведений имеется об объекте, тем более эффективно можно построить процедуры его экспериментального исследования. Класс задач, рассматриваемый в статье Вощинина А. П., Бочкова А. Ф. и Сотирова Г. Р., включает проблемы построения многофакторных моделей. При этом, как известно (см., например, [1]), необходимо определять совокупность факторов, включаемых в модель, выбирать вид модели, планировать эксперимент, вычислять оценки неизвестных параметров многофакторной модели, проводить статистический анализ результатов (проверять адекватность модели, значимость коэффициентов, строить доверительные области и т. д.). Важнейшим обстоятельством, по-существу определяющим методы решения перечисленных выше задач, является вид модели наблюдений.

Под моделью наблюдений [2] понимается совокупность характеристик наблюдаемых в эксперименте величин (т. е. результатов измерений). Пусть, например, известно, что связь между вектором факторов $x$ и откликом $y$ описывается параметрической функцией $\eta(\beta, x)$, а результат наблюдения

$$\tilde{y}=\eta(\beta, x)+\varepsilon, \qquad (1)$$

где $\varepsilon$ — погрешность наблюдений.

Выражение (1) вместе с описанием свойств величины $\varepsilon$ представляет собой модель наблюдений. Можно утверждать, что даже с точки зрения структуры выражение (1) не охватывает всех возможных моделей наблюдений. Действительно, легко обобщить (1) следующим образом:

$$\tilde{y}=R\eta(\beta+\varepsilon_\beta,\ x+\varepsilon_x)+\sum_{i=1}^{\eta}\varepsilon_i, \qquad (2)$$

где $\varepsilon_\beta$ — «мешающие» составляющие вектора параметров $\beta$; $\varepsilon_x$ — погрешность измерений факторов $x$; $\varepsilon_i$ — составляющие погрешности наблюдений; $R$ — некоторое преобразование выхода.

Модели наблюдений, отличающиеся по структуре от выражения (1), на практике встречаются довольно часто, хотя планируя эксперимент или обрабатывая данные наблюдений, внимание на это не обращают. При этом, конечно, возникают потери в точности результатов и, следовательно, в достоверности выводов.

Рассмотрим простой пример: путь переменная $y$ может измеряться с помощью $m$ приборов, каждый из которых дает некоторую систематическую погрешность $\varepsilon_i$, $i=1, 2, \ldots, m$. Тогда составляющую $\varepsilon$ модели наблюдений (1) целесообразно разбить на две части

$$\tilde{y}_i=\eta(\beta,\ x)+\varepsilon_i+\varepsilon^*, \qquad (3)$$

где $i$ — индекс прибора, использованного в данном наблюдении.

Необходимость выделения составляющей $\varepsilon_i$ из общей погрешности $\varepsilon$ определяется тем, что никакими способами, кроме как дублированием наблюдений с помощью других приборов, уменьшить влияние $\varepsilon_i$ на точность результатов нельзя, в то время как для «борьбы» с $\varepsilon^*$ могут оказаться достаточными повторные измерения одним и тем же прибором. Не выделив $\varepsilon_i$, мы существенно исказим точностные показатели результатов эксперимента (например оценок параметров $\beta$), если в эксперименте использовался только один прибор.

Наличие «мешающей» составляющей у $\beta$ также легко объяснимо. Дело в том, что проводя исследования какого-либо образца объекта (например, двигателя, вагона, колеса, оси и т. п.), выводы мы хотим, как правило, распространить на все образцы объекта (т. е. например, на все двигатели данной марки и данного производителя). Но между образцами одного объекта могут существовать различия, что и учитывает составляющая $\varepsilon_\beta$ [3].

Речь до сих пор шла о структуре модели наблюдений. Различными могут быть и предположения о свойствах погрешностей $\varepsilon$. Здесь представляются наиболее важными следующие модели: вероятностная, детерминированная, смешанная. Таким образом, можно говорить о наличии спектра моделей наблюдений, отличающихся как структурой, так и свойствами погрешностей наблюдений.

Наилучших результатов экспериментального исследования можно ожидать, когда модель наблюдений, используемая при обработке данных и планировании исследования, соответствует истинному положению дел. Исследователь должен иметь возможность альтернативного выбора одной (или ряда) из множества моделей наблюдений; для каждой модели необходимо при этом иметь научно-методические разработки и соответствующие программные средства для проведения расчетов на ЭВМ. Однако достаточно полно в настоящее время разработаны методы обработки и анализа данных при построении многофакторных зависимостей лишь для моделей наблюдений вида (1) с вероятностной погрешностью $\varepsilon$ (при предположении о нормальности закона ее распределения). Небольшое число работ (в том числе и опубликованная выше) посвящено методам обработки данных для детерминированных $\varepsilon$, характеризуемых только интервалом значений. Практически не изучены смешанные модели, где погрешность включает случайную и неслучайную составляющие. В этой ситуации исследователь вынужден использовать аппарат вероятностных ошибок, который наиболее полно разработан и подкреплен программными средствами для проведения расчетов на ЭВМ.

К задачам альтернативного подхода к выбору моделей наблюдений при построении многофакторных зависимостей относятся: систематизация моделей наблюдений, развитие методов решения задач всех этапов экспериментального исследования (выбор факторов, планирование, обработка, интерпретация) для различных видов моделей наблюдений, создание инструментальных средств (пакетов программ для ЭВМ). И еще одна важная проблема — это развитие методов анализа данных, предназначенных для использования при выборе одной (или ряда) из совокупности альтернативных моделей наблюдений.

Естественна постановка задачи оптимального планирования эксперимента при построении многофакторных моделей при интервальной нестатистической ошибке. В качестве критерия оптимальности плана может быть выбран, например, диаметр области $\Omega$ допустимых значений коэффициентов модели (или объем этой области, как это сделано в работах [1, 4]; в этом случае можно найти аналогию с критерием $D$-оптимальности, используемым при вероятностной погрешности наблюдений). Подобная задача уже исследовалась применительно к проблеме построения разделяющей поверхности в нестатистических задачах классификации [4], которая, как оказывается и с точки зрения оценивания параметров (в обоих случаях необходимо решать систему неравенств), и с точки зрения планирования эксперимента, практически совпадает с соответствующими задачами построения многофакторных моделей при интервальной нестатистической ошибке. По-видимому, в этой задаче построение априорных планов окажется невозможным, а выходом явится последовательная процедура выбора условий эксперимента, позволяющая на каждом шаге максимально (в заданном смысле) повысить точность результатов (например, уменьшить диаметр области $\Omega$).

Еще одно соображение, касающееся применения моделей с интервальными ошибками. Есть круг задач, заключающихся в поиске аппроксимирующих выражений для сложных (иногда неявно заданных) функций. Часто их решают с применением метода наименьших квадратов, испытывая при этом большие затруднения при суждении об адекватности полученных выражений. Аппарат интервальных ошибок позволяет подойти к этой задаче так: 1) ввести интервал ошибок, принимая за адекватную любую модель, лежащую внутри допустимого коридора ошибок; 2) задав многофакторную модель простейшего вида (например, линейную по факторам), отыскать (решая систему неравенств) значения коэффициентов, при которых все расчетные значения находятся внутри коридора; 3) если в пункте 2 решение отсутствует, то перейти к более сложной модели (например квадратичной); если решение в пункте 2 получено, то его можно рассматривать как адекватную аппроксимацию сложной функции.

Следует иметь в виду и такую непростую ситуацию, в которой может оказаться исследователь, когда у него нет оснований для того, чтобы отдать предпочтение, например, модели с вероятностной погрешностью по сравнению с моделью с интервальной погрешностью. Нужно оставить исследователю возможность обойтись при этом без дополнительных экспериментов, направленных на уточнение свойств погрешностей наблюдений. Для этого следует провести расчеты и выполнить анализ исходя из двух различных моделей, сопоставить результаты и в случае их близости не рассматривать вопрос об истинном виде модели наблюдений. Так можно действовать, если полностью решена инструментальная проблема, т. е. когда в распоряжении исследователя имеется набор обрабатывающих программ, ориентированных на разные виды моделей наблюдений.

ЛИТЕРАТУРА

1. *Хартман К., Лецкий Э. К., Шефер В.* Планирование эксперимента в исследовании технологических процессов.— М.: Мир, 1977.— 552 с.

2. РД 50-353-82 Методические указания. Планирование исследовательских испытаний. Основные положения.— М.: Изд-во стандартов, 1983.— 32 с.

3. *Лецкий Э. К.* Заводская лаборатория. 1986. Т. 52. № 7. С. 53—55.

4. *Лецкий Э. К., Фомин Г. А.* Изв. АН СССР. Техническая кибернетика. 1972. № 1. С. 122—126.

# IV

© *А. И. ОРЛОВ*

**В** теории и практике статистических методов назревают перемены. Один из симптомов — усиление внимания к непосредственному учету факта наличия погрешностей наблюдений в статистических процедурах. Чтобы выявить значение перемен, целесообразно встать на общую точку зрения, рассмотреть теорию и практику прикладной статистики за последние сто лет. Эти методологические рассмотрения, как показано ниже, позволяют придти к вполне конкретным алгоритмам обработки статистических данных.

До начала XX века вероятностные модели не играли большой роли при обработке данных. Статистические таблицы «говорили сами за себя», в руках профессионалов — весьма о многом. Однако постепенно накопились вопросы, для ответов на которые понадобилось перейти на уровень вероятностных моделей. Так, на первый план вышла вероятностная статистика. Одновременно была заложена основа конфликта между двумя направлениями — вероятностным и детерминированным, который выступает сейчас в основном в виде конфликта между экономической наукой «статистикой», продолжающей использовать методологию XIX века, и прикладной математической статистикой.

Переход от описания данных к математическим моделям в статистике, как и в других областях научно-практической деятельности, имеет две стороны: во-первых, математическое вероятностное моделирование обеспечивает более глубокое проникновение в суть явлений, а потому получение более существенных выводов; во-вторых, выступает на первый план проблема обоснования правильного выбора модели. Дело осложняется тем, что для решения одной и той же практической задачи могут использоваться различные математические модели, а в одной и той же модели могут быть полезны разнообразные методы. На примере проверки однородности двух независимых выборок это продемонстрировано в работе [1].

Зачем же понадобились вероятностные модели в статистике? Например, чтобы ответить на вопрос о значимости (существенности) различия долей дефектных изделий в двух цехах, чтобы указать доверительный интервал для числа дефектных единиц продукции в партии, исходя из результатов контроля выборки, и т. д. Вопросы соотношения вероятностных и детерминированных статистических методов заслуживают специального исследования (частично это сделано в работах [2—4]), не менее подробного, чем рассмотрение общих вопросов моделирования в науке и технике [5].

В первой трети XX века были заложены основы современной математической статистики, прежде всего параметрической. Новые методы, разработанные К. Пирсоном, Р. Фишером, А. Чупровым и другими, дали исследователям различных специальностей инструмент для решения стоящих перед ними задач. Так, сравнение групп с помощью критерия Стъюдента стало общепринятой процедурой в медицинских научных исследованиях. Как известно статистикам, критерий Стъюдента основан на предположении нормальности распределения результатов измерений и равенства дисперсий для двух совокупностей. Однако прикладники «забывают» о необходимости обоснования этого предположения, что может привести к ошибочным выводам [6].

Еще в 1927 г. С. Н. Бернштейн хорошо видел недостатки параметрической статистики [7]. Использование параметрических семейств оправдано лишь в случае, когда параметрический вид распределения результатов измерений вытекает из некоторой системы аксиом [8]. Подобную систему аксиом можно указать лишь весьма редко. Поэтому область применения параметрической статистики должна быть весьма ограниченной. Однако параметрические методы, в частности, основанные на нормальном распределении, применяются достаточно часто, особенно недостаточно квалифицированными специалистами, и являются препятствием на пути распространения более прогрессивных — непараметрических, робастных и других — методов [9].

Хорошо известно, что большинство реальных распределений не может быть адекватно описано нормальным законом или каким-либо еще параметрическим семейством. Почему же столь упорно применяются статистические методы, не соответствующие реальности?

Основная часть вины лежит на оторванных от жизни преподавателях статистики, любящих рассказывать о математических задачах параметрической статистики и не задумывающихся о том, имеет ли эта красивая математика какое-либо отношение к реальности. Невнимание к методологическим вопросам становится тормозом на пути внедрения современных статистических методов [10].

Справедливости ради надо сказать, что некоторые параметрические методы оказываются достаточно разумными и в непараметрической ситуации [6]. Их надо воспринимать как непараметрические и преподавать именно так.

Большая методологическая нагрузка лежит на подходе [11—14], развиваемом с целью непосредственного учета погрешностей исходных данных в статистических процедурах. Речь идет об отказе от парадигмы, сложившейся в первой трети XX века при формулировке основных математических моделей статистики, согласно которой результаты измерений, вошедшие в выборку, считались известными абсолютно точно.

Очевидно, любой результат измерения может содержать погрешность. Это касается и количественных, и качественных признаков, и любых нечисловых данных [3]. В классической математической статистике факт наличия погрешностей игнорировался. Оправданием служили малый объем типичных выборок и примитивность (с нынешней точки зрения) вычислительных средств. Фактически предполагалось, что влияние погрешностей измерения результатов наблюдений существенно меньше, чем влияние разброса результатов от наблюдения к наблюдению. В математической теории, игнорирующей погрешности результатов наблюдений, появились такие понятия, как несмещенность, состоятельность, асимптотическая эффективность, результаты, например, об асимптотической эффективности оценок максимального правдоподобия и т. д.

Стоило обратить внимание, что погрешностями измерения результатов наблюдений нельзя пренебрегать — и вся эта теория оказалась не имеющей отношения к реальности, т. е. внутриматематической. В частности, обычно используемые оценки не являются состоятельными, нецелесообразно увеличивать объем выборки сверх некоторого порога, упорядоченность алгоритмов по качеству отличается от классической, например, оценки максимального правдоподобия могут потерять свое преимущество перед оценками метода моментов, как это продемонстрировано в работе [14] на примере оценивания параметров гамма-распределения.

Итак, необходима разработка теории, алгоритмов, программ взамен классических, внедрение их в преподавание и практическую деятельность. Очевидно, сопротивление будет не меньше, чем при замене параметрической статистики на непараметрическую.

Погрешности измерений результатов наблюдений можно описывать разными способами. Анализ погрешностей реальных средств измерений показал [15], что в подавляющем большинстве случаев они не являются нормальными, а также имеют ненулевую систематическую ошибку. С другой стороны, в паспорте средства измерения указывают границы для его абсолютной или относительной погрешности. Поэтому в качестве базовой модели в реалистической статистике естественно принять, что

$$y_i = z_i + \varepsilon_i,\ i = 1, 2, \ldots, n, \qquad (1)$$

где $y_i$ — результат наблюдения, имеющийся у статистика; $z_i$ — истинный результат наблюдения; $\varepsilon_i$ — погрешность измерения результата наблюдения, $i = 1, 2, \ldots, n$. Заданы либо ограничения на абсолютную погрешность

$$|\varepsilon_i| \leqslant \Delta,\ i = 1, 2, \ldots, n, \qquad (2)$$

либо ограничения на относительную погрешность

$$|\varepsilon_i| \leqslant \delta |z_i|\ \ i = 1, 2, \ldots, n. \qquad (3)$$

Конкретные значения $\Delta$ и $\delta$ берутся из паспорта средства измерения.

По искаженному погрешностями набору $\{y_1, y_2, \ldots, y_n\}$ необходимо сделать статистические выводы о наборе $\{z_1, z_2, \ldots, z_n\}$.

Продемонстрируем подход реалистической статистики на примере доверительного математического ожидания и дисперсии.

На основе предположения о том, что $z_1, z_2, \ldots, z_n$ — независимые одинаково распределенные случайные величины, имеющие первые четыре момента, требуется по $y_1, y_2, \ldots, y_n$ оценить $E(z_1)$и $\sigma^2 = D(z_1)$.

Доверительный интервал для $E(z_1)$ построим, исходя из

$$\bar{y} = \frac{y_1 + \ldots + y_n}{n} = \frac{z_1 + z_2 + \ldots + z_n}{n} + \frac{\varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_n}{n} = \bar{z} + \bar{\varepsilon}. \qquad (4)$$

Пусть справедливы ограничения (2) на $\varepsilon_i$. Что можно сказать о $\bar{\varepsilon}$? Из выражения (2) следует, что $|\bar{\varepsilon}| \leqslant \Delta$, но усилить это неравенство в общем случае нельзя. Таким образом,

$$\bar{y} - \Delta \leqslant \bar{z} \leqslant \bar{y} + \Delta. \qquad (5)$$

В силу центральной предельной теоремы при больших объемах выборок вероятность неравенства

$$\bar{z} - u(\gamma)\frac{\sigma}{\sqrt{n}} \leqslant E(z_1) \leqslant \bar{z} + u(\gamma)\frac{\sigma}{\sqrt{n}} \qquad (6)$$

близка к $\gamma$, где $u(\gamma)$ — квантиль стандартного нормального распределения порядка $(1+\gamma)/2$.

При малых $\Delta$ это утверждение остается справедливым, если заменить $\sigma$ в неравенстве (6) на $S$, где

$$S^2 = \frac{1}{n-1} \sum_{1 \leqslant i \leqslant n} (y_i - \bar{y})^2. \qquad (7)$$

Из последних соотношений вытекает, что асимптотический доверительный интервал для $E(z_1)$, соответствующий доверительной вероятности $\gamma$, имеет вид

$$\left[\bar{y} - \Delta - u(\gamma)\frac{S}{\sqrt{n}};\ \bar{y} + \Delta + u(\gamma)\frac{S}{\sqrt{n}}\right]. \qquad (8)$$

Его полуширина равна

$$\Delta + u(\gamma)\frac{S}{\sqrt{n}} \qquad (9)$$

и не стремится к 0 при росте объема выборки. Следовательно, среднее арифметическое не является состоятельной оценкой математического ожидания.

Невозможность состоятельного оценивания известна давно. Еще Огюст Курно в книге, изданной в Париже в 1843 г., отмечал, что асимптотически точное оценивание невозможно в принципе из-за наличия погрешностей результатов наблюдений [16, с. 206—207, 220—221]. Об этом же на примере оценивания математического ожидания писали Б. В. Гнеденко и А. Я. Хинчин [17, с. 120—121], не приводя, однако, математических обоснований. Известно это и в метрологии [15]. По мнению П. Е. Эльясберга, «Состоятельность статистических оценок можно с полным основанием считать одним из «мифов XX века» [18].

Подход реалистической статистики [14] позволяет перейти от обоснованной критики к алгоритмам статистических расчетов, типа расчета доверительного интервала согласно формуле (8). Из соотношения (9) вытекает, что не имеет смысла безгранично увеличивать объем выборки, поскольку полуширина доверительного интервала не может быть меньше $\Delta$. Где остановиться? Согласно «принципу уравнения погрешностей» [19], следует приравнять случайную погрешность и погрешность, связанную с неточным измерением результатов наблюдений:

$$\frac{\sigma}{\sqrt{n}} = \Delta, \qquad (10)$$

откуда находим рациональный объем выборки

$$n_{\text{рац}} = \left(\frac{\sigma}{\Delta}\right)^2. \qquad (11)$$

Следовательно, если абсолютная погрешность составляет 0,1 от $\sigma$, то более 100 наблюдений проводить нецелесообразно.

Перейдем к доверительному оцениванию дисперсии на основе выборочной дисперсии $S^2$ имеющихся у статистика результатов наблюдений (см. формулу (7)). Положим

$$S_1^2 = \frac{1}{n-1} \sum_{1 \leqslant i \leqslant n} (z_i - \bar{z})^2, \qquad (12)$$

где $S_1^2$ — выборочная дисперсия истинных результатов наблюдений. Тогда с точностью до членов более высокого порядка по $\varepsilon_i$ имеем

$$S^2 = S_1^2 + \frac{2}{n-1} \sum_{1 \leqslant i \leqslant n} (z_i - \bar{z})\varepsilon_i. \qquad (13)$$

Пусть заданы ограничения на абсолютную погрешность (2). Последний член в равенстве (13) будет максимальным, если

$$\varepsilon_i = \begin{cases} \Delta, & z_i - \bar{z} \geqslant 0, \\ -\Delta, & z_i - \bar{z} < 0. \end{cases} \qquad (14)$$

Итак, погрешность статистической процедуры, вызванная погрешностями исходных статистических данных, равна

$$\frac{2\Delta}{n-1} \sum_{1 \leqslant i \leqslant n} |z_i - \bar{z}| \approx 2\Delta E|z_1 - E(z_1)|. \qquad (15)$$

Известно, что выборочная дисперсия асимптотически нормальна; при большом объеме выборки распределение случайной величины

$$(S_1^2-\sigma^2)\sqrt{n} \qquad (16)$$

близко к нормальному с математическим ожиданием 0 и дисперсией, равной $D(z_1^2)$ (см. например, [20, с. 419]). Следовательно, доверительный интервал для $\sigma^2$, соответствующий доверительной вероятности $\gamma$, имеет вид

$$[S^2-A;\ S^2+A], \qquad (17)$$

где

$$A=\frac{u(\gamma)}{\sqrt{n-1}\sqrt{n}}\times$$
$$\times\sqrt{\sum_{1\leqslant i\leqslant n}\left(y_i^2-\frac{1}{n}\sum_{1\leqslant j\leqslant n}y_j^2\right)^2}+$$
$$+\frac{2\Delta}{n-1}\sum_{1\leqslant i\leqslant n}|y_i-\bar{y}_i|, \qquad (18)$$

причем $u(\gamma)$ — то же, что и в формуле (6).

Заменяя выборочные характеристики теоретическими, получаем полудлину доверительного интервала в виде

$$\frac{u(\gamma)}{\sqrt{n}}\sqrt{D(z_1^2)}+2\Delta E|z_1-E(z_1)|, \qquad (19)$$

откуда находим рациональный объем выборки:

$$n_{\text{рац}}=\frac{D(z_1^2)}{4\Delta^2(E|z_1-E(z_1)|)^2}. \qquad (20)$$

Если $z_1$ имеет нормальное распределение с математическим ожиданием 0 и дисперсией 1, то выражение (20) принимает вид

$$n_{\text{рац}}=\frac{1}{\pi\Delta^2}. \qquad (21)$$

Следовательно, если абсолютная погрешность составляет 0,1 от $\sigma$, то более 32 наблюдений для оценивания $\sigma^2$ проводить нецелесообразно.

Для нахождения рационального объема выборки по реальным данным теоретические моменты в формулах (11) и (20) следует заменить на выборочные. Строгие доказательства всех сформулированных выше утверждений проводятся методами, развитыми в работе [14] для частного случая гамма-распределения.

## ЛИТЕРАТУРА


1. *Орлов А. И., Фомин В. Н.* / Надежность и контроль качества. 1988. № 12. С. 3—9.
2. *Орлов А. И.* / Доклады МОИП 1984. Общая биология: Цитогенетический и математический подходы к изучению биосистем. М.: Наука, 1986. С. 179—182.
3. *Орлов А. И.* В кн.: Анализ нечисловой информации в социологических исследованиях.— М.: Наука, 1985. С. 58—92.
4. *Тутубалин В. Н.* Границы применимости: Вероятностно-статистические методы и их возможности.— М.: Знание, 1977.— 64 с.
5. *Неуймин Я. Г.* Модели в науке и технике: История, теория, практика.— Л.: Наука, 1984.— 190 с.
6. *Орлов А. И.* / Вестник АМН СССР. 1987. № 2. С. 88—94.
7. *Бернштейн С. Н.* Собрание сочинений. Т. IV.— М.: Наука, 1964. С. 217—232.
8. *Орлов А. И.* / Надежность и контроль качества. 1986. № 11. С. 29—34.
9. *Гнеденко Б. В., Орлов А. И.* / Заводская лаборатория. 1988. Т. 54. № 1. С. 1—4.
10. *Комаров Д. М., Орлов А. И.* В кн.: Вопросы применения экспертных систем.— Минск: НПО «Центрсистем», 1988. С. 151—160.
11. *Вощинин А. П.* Метод оптимизации объектов по интервальным моделям целевой функции.— М.: Изд. МЭИ, 1987.— 48 с.
12. *Кузнецов В. П.* Интервально-статистические модели.— М.: Радио и связь, 1989.— 454 с.
13. *Лейфер Л. А.* В кн.: Статистика. Вероятность. Экономика.— М.: Наука, 1985. С. 349—354.
14. *Орлов А. И.* В кн.: Статистические методы оценивания и проверки гипотез. Межвузовский сборник научных трудов.— Пермь: Пермский ун-т, 1988. С. 88—97.
15. *Новицкий П. В., Зограф И. А.* Оценка погрешностей результатов измерений.— Л.: Энергоатомиздат, 1985.— 248 с.
16. *Курно Ог.* Основы теории шансов и вероятностей. / Пер. с франц. изд.: Париж, 1843.— М.: Наука, 1970.— 384 с.
17. *Гнеденко Б. В., Хинчин А. Я.* Элементарное введение в теорию вероятностей / Изд. 7-е.— Наука, 1970.— 168 с.
18. *Эльясберг П. Е.* Измерительная информация. Сколько ее нужно, как ее обрабатывать? — М.: Наука, 1983.— 208 с.
19. *Орлов А. И.* Устойчивость в социально-экономических моделях.— М.: Наука, 1979.— 296 с.
20. *Боровков А. А.* Математическая статистика.— М.: Наука, 1984.— 472 с.

Очень часто графическое представление результатов наблюдений реализаций какого-либо процесса имеет вид, показанный на рис. 1. Наблюденные значения вытянуты вдоль некой (неизвестной) кривой и значительно «размазаны», так что в итоге получили целую полосу наблюдений. Такая же картина возникнет и тогда, когда мы строим

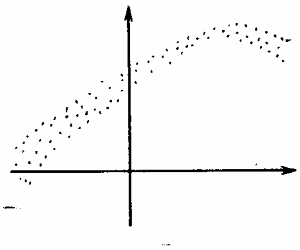

Рис. 1. Пример реализации нескольких процессов, протекающих в одинаковых условиях

реализации нескольких экспериментов, повторяемых в одних и тех же условиях.

Можно предположить, что изучаемый процесс описывается формулой

$$y(t)=f(t),$$

а наблюдаем на самом деле

$$\xi_t(\omega)=f(t)+\eta_t(\omega), \quad t\in T\subset R^1,$$

где $\xi_t(\omega)$ — измеренное значение выходной переменной; $f(t)$ — тренд, неизвестное истинное значение выходной переменной; $\eta_t(\omega)$ — случайная ошибка эксперимента. Предполагается, что ее математическое ожидание равно нулю и известна ее корреляционная функция $\mathscr{C}(t, s)$.

Зададим функции $\varphi_1(t), ..., \varphi_n(t)$.

В том случае, когда из каких-либо априорных рассуждений предполагают, что функция $f(t)$ в точности является алгебраической суммой известных функций, т. е. имеет место точное равенство

$$f(t)=\alpha_1\varphi_1(t)+\alpha_2\varphi_2(t)+...+\alpha_n\varphi_n(t), \quad (1)$$

можно решать задачу оценивания коэффициентов регрессии $\alpha_1, ..., \alpha_n$.

Если же мы не имеем веских оснований для предположения, что неизвестная функция точно представима в форме (1), то и регрессионную задачу следует решать осторожно, особенно в момент интерпретации полученных результатов.

Представим задачу в более общей форме: для этого обратимся сначала к примеру.

Довольно часто в практических инженерных задачах можно предположить, что неизвестная функция $f(t)$ представима в виде разложения в ряд Тейлора (в окрестности точки $t_0$):

$$f(t)=f(t_0)+\frac{f'(t_0)}{1!}(t-t_0)+...+$$
$$+\frac{f^{(n)}(t_0)}{n!}(t-t_0)^n+Q(t),$$

т. е. раскладывается по степеням полинома с известной погрешностью $Q(t)$.

Если предположить только, что неизвестная функция $f(t)$ непрерывна, то (по теореме Вейерштрасса) ее можно с любой наперед заданной степенью точности приблизить полиномами.

Существуют и другие приближения функций. Для нас важно только одно: вместо того, чтобы считать, что мы точно «угадали» разложение неизвестной функции, предполагать, что погрешность аппроксимации неизвестной функции с помощью заданных функций не выходит за определенные пределы. А именно:

— пусть задана система функций $\varphi_1(t), ..., \varphi_n(t)$;
— пусть задана допустимая точность аппроксимации $M(t)$ (всюду неотрицательна);
— относительно функции $f$ теперь будем предполагать, что ее можно представить с

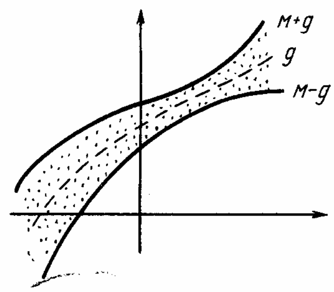


Рис. 2. Полоса приближений для неизвестного тренда $g(t)$

помощью функций $\varphi_1, ..., \varphi_n$ с заданной степенью точности $M$:

$$|f(t)-g(t)|\leqslant M(t), \quad (2)$$

где обозначено

$$g(t)=\vartheta_1\varphi_1(t)+\vartheta_2\varphi_2(t)+...+\vartheta_n\varphi_n(t). \quad (3)$$

Класс функций $f$, допускающий такое представление, будем обозначать $\Phi_T(\varphi, M)$.

*Замечание.* Точная постановка задачи дана в работах[1, 2].

[1] *Легостаева И. Л., Благовещенский Ю. Н.* / ДАН СССР. 1982. Т. 264. № 4. С. 791—794.

[2] *Легостаева И. Л.* Минимаксное оценивание тренда случайного процесса. Автореф. дис. канд. физ.-мат. наук. Вильнюс. 1986.— 16 с.

Соотношения (2) и (3) описывают любые функции $f$, лежащие в интервале от $g(t)-M(t)$ до $g(t)+M(t)$ (см. рис. 2).

Задание функции $M(t)$ определяет тип оценки решаемой задачи: 1) $M(t)=0$ почти всюду на $T$ — регрессионная задача («угадали» точное представление для неизвестной функции); 2) $M(t)>0$ почти всюду на $T$ — интервальная оценка; 3) $M(t_0)=0$, $M(t)>0$ при $t\neq t_0$ — точечная оценка;

$$4)\ M(t)=\begin{cases} \varepsilon, & t\in[a,b]\subset T, \\ \varepsilon+U(t), & t\in T\setminus[a,b],\ U(t)\geqslant 0, \end{cases}$$

$\varepsilon$ — равномерная оценка на интервале $[a, b]$.

После того, как мы пришли к выводу, что оцениваемая функция $f$ принадлежит классу $\Phi_T(\varphi, M)$, т. е. для нее существует представление (2), (3), возникает задача оценивания функции $g(t)$, или, что то же самое, коэффициентов $\vartheta_1, ..., \vartheta_n$. По ряду причин оценить все эти коэффициенты в совокупности технически не удается[2]. Можно оценить их порознь, но тогда встает задача обоснования возможности их одновременного использования. Мы будем оценивать любую наперед заданную линейную комбинацию неизвестных параметров

$$\beta_0=\sum_{i=1}^{n}\vartheta_j c_j,\ c_1, ..., c_n \text{ — заданы.}$$

В качестве оценки $\beta$ для $\beta_0$ возьмем линейный функционал

$$\beta=\int_T l(s)\xi_s(\omega)\,ds. \tag{4}$$

Это — случайная величина, $\beta=\beta(\omega)$. Вычислим ее среднее (считаем, что все необходимые допущения теоремы Фубини выполнены)

$$E\beta=E\int_T l(s)\xi_s(\omega)\,ds=\int_T l(s)E[g(s)+r(s)+$$

$$+\eta_t(\omega)]ds=\int_T l(s)g(s)\,ds+\int_T l(s)r(s)\,ds+0=$$

$$=\sum_{j=1}^{n}\vartheta_j\int_T l(s)\varphi_j(s)\,ds+\int_T l(s)r(s)\,ds.$$

Мы хотим, чтобы оценки были несмещенными ($E\beta=\beta_0$) при точных регрессионных моделях (т. е. когда $r(s)=0$ почти всюду):

$$E\beta=\sum_{j=1}^{n}c_j\vartheta_j.$$

В то же время,

$$E\beta=\sum_{j=1}^{n}\vartheta_j\int_T l(s)\varphi_j(s)\,ds.$$

Сопоставляя эти формулы, получаем

$$\sum_{j=1}^{n}\vartheta_j\Big[c_j-\int_T l(s)\varphi_j(s)\,ds\Big]=0.$$

Это условие должно выполняться по крайней мере для всех $c_i=\varphi_i(t_m)$, $t_m\in T$ (в этом случае $\beta_0=\sum_1^n\vartheta_i c_i=\sum_1^n\vartheta_i\varphi_i(t_m)=g(t_m)$, т. е. это оценка для функции $f$ в точке $t_m$ с точностью $M(t_m)$), что приводит к условию

$$\int_T l(s)\varphi_j(s)\,ds=c_j,\quad j=1, ..., n. \tag{5}$$

Качество оценки будет определяться функцией риска

$$R_{\beta_0}(\beta, \beta_0)=E_{\beta_0}(\beta-\beta_0)^2,$$

где $\beta_0$ — неизвестное «истинное» значение. Вычисления дают[2]

$$R_{\beta_0}(l, f)=\Big[\int_T l(s)r(s)\,ds\Big]^2+$$

$$+\int_T\int_T \mathscr{C}(t, s)l(s)l(t)\,ds\,dt,$$

где $\mathscr{C}(t, s)$ — известная корреляционная функция процесса $\eta_t(\omega)$. Здесь мы ввели запись функции риска как функционала от $l$ и $f$, так как $\beta$ зависит от $l$ и $\xi(f)$ и в дальнейшем нам это будет удобно.

Обозначим

$$R_{\beta_0}(l)=\sup_{f\in\Phi_T(\varphi, M)}R_{\beta_0}(l, f).$$

*Определение.* Оценка $\hat\beta=\beta(\hat l, \xi)$, удовлетворяющая условиям (4) и (5), называется линейной минимаксной оценкой, если для любой другой оценки $\beta=\beta(l, \xi)$ с теми же условиями выполнено неравенство

$$R_{\beta_0}(\hat l)\leqslant R_{\beta_0}(l).$$

Следовательно, задача сводится к отысканию $\hat l$, удовлетворяющей условию (5), при которой выполнено равенство

$$R_{\beta_0}(\hat l)=\inf_l R_{\beta_0}(l), \tag{6}$$

где инфинум берется по всем функциям $l$, удовлетворяющим условию (5). Функцию $\hat l$ будем называть минимаксной весовой функцией.

Равенство в формуле (6) может быть недостижимо, тогда минимаксная весовая функция не существует. В работах[1, 2] сформулированы условия, при выполнении которых минимаксная весовая функция существует и единственна; приведена система уравнений, решение которой $(\hat l, \hat\alpha, T_1)$ определяет эту функцию

$$\int_T\mathscr{C}(t, s)\hat l(s)\,ds+M(t)\operatorname{sign}\hat l(t)\int_T M(s)|\hat l(s)|\,ds=$$

$$= \sum_{j=1}^{n} \hat{\alpha}_j \varphi_j(t), \quad t \in T_1;$$

$$\int_T \hat{l}(s) \varphi_j(s)\,\mathrm{d}s = c_j, \quad j = 1, \ldots, n;$$

$$T_1 = \{u \in T:\ |\hat{l}(u)| > 0\},$$

$$M(t) \int_T M(s)\,|l(s)|\,\mathrm{d}s \geqslant \Big| \sum_{j=1}^{n} \hat{\alpha}_j \varphi_j(t) - $$

$$- \int_T \mathscr{C}(t, s) \hat{l}(s)\,\mathrm{d}s \Big|, \quad t \in T \setminus T_1.$$

Эта система зависит только от исходной корреляционной функции и заданной системы функций $M$, $\varphi$, что позволяет находить минимаксные оценки для более широкого класса шумов, чем это делалось раньше.

Пусть наблюдения имеют вид, представленный на рис. 3. Предположим, что модель наблюдений есть $\xi_t(\omega) = f(t) + \eta_t(\omega)$ при $t \in R^1$. Нас интересует $\varepsilon$-равномерная на отрезке $[-a, a]$ минимаксная оценка тренда. Ее можно, например, задать функцией

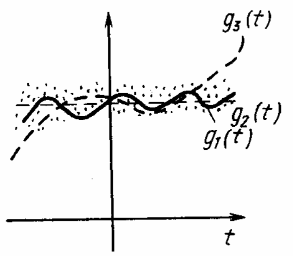


Рис. 3. Допустимые варианты функций $g(t)$

$$M(t) = \begin{cases} \varepsilon \text{ при } t \in [-a, a], \\ \varepsilon a^{-1} |t| \text{ при } t \in R / [-a, a]. \end{cases}$$

Можно предложить различные варианты для функций $\varphi$, например:

$$f(t) = \vartheta + r_1(t),$$

где $|r_1(t)| \leqslant M(t)$,

или

$$f(t) = \vartheta_1 \sin wt + \vartheta_2 \cos wt + r_2(t),$$

где $|r_2(t)| \leqslant M(t)$.

Разумно остановиться на простейшей модели, которая приведена первой, полагая $c = 1$. В этом случае $|f(t) - \vartheta| \leqslant M(t)$, откуда

$$\vartheta - M(t) \leqslant f(t) \leqslant \vartheta + M(t),$$

что легко представить на графике (рис. 4). В результате вычислений найдем оценку для $\vartheta$; тем самым будет определена полоса, в которой с заданной степенью точности $M(t)$ лежит неизвестная функция $f(t)$. Теперь в качестве функции $f(t)$ можно взять любую, не выходящую за установленные рамки, при этом больше ничего не надо оценивать. При использовании регрессионной модели для каждого изменения представления неизвестной функции надо заново пересчитывать все коэффициенты.

При гауссовском белом шуме с корреляционной функцией $\mathscr{C}(t, s) = d^2 \delta(t - s)$, где

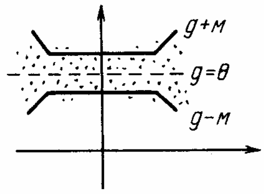


Рис. 4. $\varepsilon$-равномерное приближение наблюдений

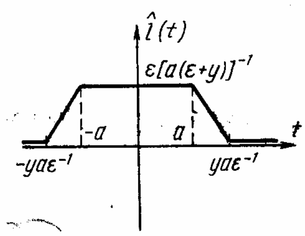


Рис. 5. Минимаксная $\varepsilon$-равномерная оценка тренда $g(t) = \theta$

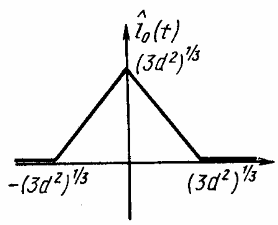


Рис. 6. Точечная минимаксная оценка тренда $g(t)$ в точке $t = 0$

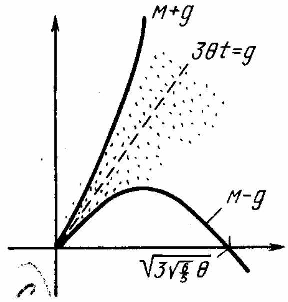


Рис. 7. Границы аппроксимации для винеровского процесса с трендом $g(t) = 3\theta t$ и точностью $M(t) = \sqrt{5/6}\, t^3$

$\delta$ — функция Дирака, получаем следующую оценку:

$$\hat{\vartheta}=\int_{-ya\varepsilon^{-1}}^{ya\varepsilon^{-1}} l(s)\xi_s \mathrm{d}s,$$

где $y$ — единственный корень уравнения $y^3+3y\varepsilon^2-4\varepsilon^3-3d^2\varepsilon a^{-1}=0$. Весовая функция показана на рис. 5.

Если положить $a=\varepsilon$ и устремить $\varepsilon$ к нулю, $\varepsilon\downarrow 0$, тогда $M(t)=|t|$ и минимаксная весовая функция

$\hat{l}_0(t)=(3d^2)^{1/3}\ \max\{0;\ (3d^2)^{1/3}-|t|\}$ (см. рис. 6). Это пример точечной оценки в нуле ($M(0)=0$).

В заключение приведем пример минимаксной оценки для винеровского шума с корреляционной функцией $\mathscr{E}(t, s)=\min(t, s)$ при $t, s\in R^+$ (рис. 7). Рассмотрим однопараметрическое представление $f(t)=\vartheta\varphi(t)+r(t)$ с функциями $\varphi(t)=3t$ и $M(t)=\sqrt{5/6}t^3$. В итоге получим:

$$\hat{l}(t)=\begin{cases} t, & t\in[0, 1], \\ 0, & t\geqslant 1, \end{cases} \quad \text{и} \quad \hat{\beta}=\hat{\vartheta}=\int_0^1 t\xi_t \mathrm{d}t.$$

Таким образом, приведена одна из возможных постановок задач интервального оценивания детерминированного тренда в модели с аддитивным случайным шумом.

## VI

© В. П. КУЗНЕЦОВ

Цель настоящего сообщения — обратить внимание на общий интервальный подход к проблемам описания неопределенности, случайности, незнания, излагаемый в книге Кузнецова В. П. «Интервальные статистические модели», которая готовится в издательстве «Радио и связь» к выходу в 1990 г.

Современные вероятностные построения имеют чрезмерную детализацию: для задания распределения вероятностей требуется знать точно вероятности огромного числа событий (сигма — алгебры). Большинство же встречающихся на практике задач как раз являются чрезвычайно бедными в этом отношении. Использование семейств распределений вероятностей все более сложной конфигурации (параметрических, непараметрических, засоренных) трудностей не снимает, а наоборот, усугубляет изначальную громоздкость вероятностных моделей, а отсюда и статистических методов. Поэтому неотлагательной представляется потребность в математическом аппарате, основанном на отрывочных, частичных, размытых, упрощенных описаниях свойств случайного явления, который (в отличие от интервального анализа и теории нечетких множеств) стоял бы на статистической платформе. Такое аппарат предлагается интервальными статистическими моделями.

В основу интервальных моделей положено понятие интервального статистического среднего. Для пояснения рассмотрим пример. Рост произвольного мужчины есть случайная величина. Можно иметь следующие сведения о ней: 1) диапазон изменения, для этого достаточно взять с запасом интервал от 100 до 250 см; 2) точное среднее, например, 176 см. Последняя в отличие от первой — статистическая характеристика, но совершенно точно вряд ли кому известная. Экспертам значительно удобнее, не боясь ошибиться, указать интервал, например, 170—180 см. В виде интервалов подобная оценка получается и по опытным данным, если наложить требования на ее надежность. В зависимости от качества эксперта, объема экспериментального материала, требуемой надежности получаются разные интервалы как для выбранной, так и для других характеристик рассматриваемой случайной величины: среднего квадрата роста, интервальной вероятности (средней частоте) превышения порога, например, 190 см.

Таким образом:

— интервальные средние занимают промежуточное положение между полным незнанием среднего и полным знанием, отражая частичное знание и придавая сведениям надежность;

— каждой числовой функции на исходах, т. е. образованной случайной величине, соответствует свое интервальное среднее, а так как образовать функций можно сколь угодно много, то также много будет интервальных средних для одного явления;

— интервальные средние существуют самостоятельно независимо от вероятностей, которые являются лишь составной частью средних.

Эти соображения легли в основу структуры интервальных моделей. На пространстве элементарных событий $X$ образованы числовые функции $f(x)$ (случайные величины), называемые признаками явления. Рассматриваются всевозможные признаки $\forall f$. На части из них $\{g(x),\ g\in G\}$, составляющих набор $G$, названный первичным, исходно задаются первичные интервальные средние

$(Mg, \tilde{M}g)$, $g \in G$. Эти средние однозначно распространяются по фиксированному алгоритму продолжения на все $\forall f$, формируя совокупность $(\underline{M}f, \overline{M}f)$, $\forall f$. В эту совокупность входят как часть и первичные средние с той лишь разницей, что волнистые линии заменяются на прямые $(\underline{M}g, \overline{M}g)$, символизируя тем самым пересчет в возможно более узкие (точные) интервалы. В образованной совокупности средние согласованы в смысле заложенных нами с самого начала аксиом (вместо аксиом теории вероятностей).

Собственно, на описанной схеме базируются классические модели теории вероятностей. В качестве первичных берутся индикаторные признаки алгебры (сигма — алгебры) событий, тогда средние превращаются в вероятности и они задаются исходно точными (т. е. с совпадающими границами). Интегрированием по вероятностному распределению (как следствие общего алгоритма продолжения) они распространяются на измеримые признаки $f$, придавая им точные средние $Mf$, и далее — на неизмеримые $(\underline{M}f, \overline{M}f)$, где станут уже интервальными. Для семейств распределений вероятностей все средние, вычисляемые как минимумы и максимумы по семействам, будут интервальными, формируя интервальные модели.

Итак, вероятностные модели и их семейства являются частными случаями наших моделей. К интервальным моделям приводит задание моделей моментами, отдельными корреляционными свойствами и некоторыми вероятностями в точном или интервальном их понимании, составляющими первичные наборы (играя роль гиперповерхностей, окаймляющих поверхности семейств), а также субъективные (относительные) вероятности и модели интервального анализа. Таким образом, интервальные модели образуют единый аппарат для охвата разнообразных областей исследований и наук.

Особенности наших моделей в том, что чем меньше первичных данных, тем проще получаются модели. Самая простая (голая) модель образуется, когда никаких данных о явлении (кроме возможных исходов) нет. Модели вкладываются друг в друга по мере уточнения данных и роста их числа. Определены понятия объединения (расширения семейств, роста неопределенности) и пересечения (добавления новых сведений) моделей.

В терминах средних на базе интервальных моделей ставятся и решаются аналогичные задачи, что и для классической теории вероятностей. Это определение совместных, частных, условных моделей, определение зависимости и независимости, расчеты интервалов средних после различных преобразований, задание случайных последовательностей, процессов, свои предельные и допредельные теоремы и т. д.

На интервальных моделях закладывается новый общий подход к математической статистике. Исходные данные о наблюдениях и об их связи с интересующими нас полезными параметрами оформляются в виде совместной интервальной модели. Вырабатываются критерии качества (вводится риск) правил принятия решений (оценки параметров, проверки гипотез) и ищутся оптимальные правила. Весомую часть в процедуре синтеза составляют теоремы о достаточности, облегчающие поиск правил. Утверждается, что структура ожидаемых оптимальных правил, внешний вид зависят целиком от первичных признаков, формирующих совместную модель, останется лишь найти коэффициенты, при которых минимизируется риск. Задача синтеза сводится к параметрической минимизации линейной формы при ограничениях на коэффициенты, причем чем меньше первичных признаков, тем меньше будет коэффициентов.

Внешние свойства интервальных моделей в виде симметрии или инвариантности к преобразованиям порождают подобные же свойства у оптимальных правил, упрощая тем самым их исходную структуру и уменьшая число варьируемых коэффициентов.

Важный момент состоит в введении «настроения» пессимизма и оптимизма в синтез правил, ориентируя их соответственно на верхнюю и нижнюю границы интервального риска, или на нечто промежуточное. Методы максимального правдоподобия, как оказывается, являются итогом режима крайнего оптимизма. Синтез на идеализированном материале точных распределений вероятностей соответствует режиму полуоптимизма. Нас же всего более интересует режим пессимизма: только при нем возможна достаточная редукция.

При синтезе доверительных оценок, обладающих заданной надежностью, используются специальные теоремы о достаточности, согласно которым оптимальное правило является усеченной снизу комбинацией первичных признаков модели. Это не будет, в общем, доверительный интервал, как обычно для распределений вероятностей. Так, если речь идет об оценивании параметра сдвига из аддитивной смеси с шумом, о котором известна только среднестатистическая мощность, то форма доверительной оценки будет параболической по наследству от вида признака, незримо присутствующего в слове «мощность». Размывание интервальной модели увеличивает расплывчатость оценки, тогда как сбор дополнительных данных (сужение модели) уменьшает ее. Но не всегда так. Например, дополнительное задание среднего значения шума не влияет на вид и свойства упомянутой только что оптимальной оценки параметра сдвига. Корреляция же отсчетов является тем направлением получения знаний, которое сильно сказывается на улучшении оценки.

Нечто подобное происходит и с правилами проверки гипотез. Размытость интервальных моделей ведет к рандомизированности правил, и не на границе критической области, а повсеместно, давая тем самым осторожные решения и оставляя последнее слово за человеком. Для различения нескольких гипотез синтез сводится к геометрической задаче оптимального расположения поверхностей в пространстве наблюдений, разделяющих сферы влияния гипотез. Формы поверхностей определяются видом первичных признаков.

В разрабатываемые методы синтеза включаются как частные случаи известные. Так, для точных вероятностей синтез связывается с отношением правдоподобия. Определенные возможности дает представление моделей в виде семейств распределений вероятностей. Тогда синтез сводится к поиску наименее благоприятных экземпляров, то, что постулируется робастной теорией. Этот путь довольно громоздок и ориентируется на специальный класс моделей, описываемых окрестностями распределений.

Наконец, мы подошли к последнему важному моменту интервальной теории. Какую же выгоду помимо упрощений сулит использование интервальных моделей? Казалось бы, дополнительная размытость моделей, отход от точных распределений вероятностей ведет к ухудшению свойств оптимальных правил. Дело в том, что размытость есть единственное и универсальное средство гарантировать надежность моделей. Истинные свойства правил принятия решений складываются из двух частей: первая определяется надежностью моделей, и вторая — это расчетные свойства по выбранной модели. Это сложение дает критерий, позволяющий искать и находить истинно оптимальные правила, для этого в процедуру синтеза добавляется регулировка ширины (надежности) модели. Примером такого правила является доверительная оценка параметра сдвига, адаптирующаяся к неизвестной дисперсии шума (предположение распределений вероятностей в нашем аппарате принципиально не требуется).

Предлагаемая нами теория открывает дорогу к новым исследованиям как по обоснованию условий оптимальности имеющихся квазиоптимальных и эвристических алгоритмов, так и получению новых, истинно оптимальных правил принятия решений, соответствующих реальным данным. Эту теорию следует рассматривать как перспективную математическую оболочку, готовую для наполнения в духе содержащихся в обсуждаемой книге примеров новыми фактами применительно к разным областям деятельности, и в первую очередь, переводу статистических выводов на рельсы реальных данных в обход распределений вероятностей как крайне идеальных и громоздких моделей.

## *В ЛАБОРАТОРИЯХ. ОБМЕН ОПЫТОМ*

УДК 546.22:543.258

© *М. С. ПОПОВА, С. В. ТУРОВСКАЯ, Л. В. РЫБИНЦЕВА*

# КУЛОНОМЕТРИЧЕСКИЙ МЕТОД ОПРЕДЕЛЕНИЯ МАССОВОЙ ДОЛИ СЕРЫ В КОКСЕ

**Коммунарский металлургический комбинат**

Для определения серы в коксе при контроле качества сырья применяется гравиметрический метод (ГОСТ 8606—72). Метод трудоемкий и по длительности не удовлетворяет требованиям производства.

Разработан и аттестован кулонометрический метод определения серы в коксе на экспресс-анализаторе АС-7012 в диапазоне концентраций от 0,9 до 2 %.

Проверена устойчивость поглотительного