

УДК 622.012

ОПРЕДЕЛЕНИЕ НЕОБХОДИМОГО ОБЪЕМА ВЫБОРОЧНЫХ ДАННЫХ ДЛЯ СОЗДАНИЯ ШКАЛЫ С ДИНАМИЧЕСКИМИ ОГРАНИЧЕНИЯМИ

С. В. ПРОКОПЧИНА

*профессор, доктор технических наук, Финансовый университет
при Правительстве Российской Федерации,*

Москва, Россия

E-mail: svprokopchina@mail.ru

Рассмотрены вопросы создания шкал с динамическими ограничениями с требуемой точностью и надежностью измерений интегральных показателей сложных систем. Приведены формульные зависимости для определения объема экспериментальных данных, обеспечивающего достижение требуемой точности измерений. Шкалы рекомендованы к применению в условиях разнотипной, разновременной и неточной, неполной, нечеткой информации.

Ключевые слова: шкала с динамическими ограничениями, интегральные показатели, определение объема экспериментальных данных

Investment, financial and management analysis

DETERMINE THE AMOUNT OF SAMPLE DATA NEEDED TO CREATE A SCALE WITH DYNAMIC CONSTRAINTS

S. V. PROKOPCHINA

*Professor, Doctor of Technical Sciences, Financial University
under the Government of the Russian Federation,*

Moscow, Russia

E-mail: svprokopchina@mail.ru

The problems of creating scales with dynamic constraints with the required accuracy and reliability of measurements of integrated indicators of complex systems are considered. Formula dependences for determination of the volume of experimental data providing achievement of the required accuracy of measurements are given. Scales are recommended for use in conditions of different types, different times and inaccurate, incomplete, fuzzy information.

Keywords: scale with dynamic constraints, integral indicators, determination of the volume of experimental data.

Необходимый объем выборочных данных для создания шкалы с динамическими ограничениями на этапе построения разделяющих границ реперов можно определить, используя аппарат теории статистической проверки гипотез. Основными понятиями этой теории являются понятия ошибок 1(α) и 2(β) рода, которые определяют соответствие вероятности отклонения гипотезы в случае ее справедливости и принятия, когда она не верна. Нулевая гипотеза – это гипотеза, адекватность которой проверя-

ется, а все другие гипотезы, которые могут иметь место при данных условиях, являются альтернативными. Для конкретизации решений рассмотрим пример определения объема выборочных данных при построении двумерной шкалы с динамическими ограничениями при измерении показателей асимметрии и эксцесса по выборочным данным, приведенный в [1].

Пусть H_0 – нулевая гипотеза о виде плотности вероятности исследуемого процесса, которой соответствуют теоретические значения коэф-

коэффициентов асимметрии A_i и эксцесса E_i . Тогда все остальные эталонные распределения будут альтернативными по отношению к H_i . Учитывая зависимость признаков коэффициентов асимметрии и эксцесса, выражение для ошибки 1-го рода можно записать в виде:

$$\alpha_i = \alpha_{A_i} \cdot \alpha_{E_i/A_i}, \quad (1)$$

которое при нормальном распределении выборочных коэффициентов асимметрии и эксцесса для больших объемов выборки имеет вид:

- для ошибки первого рода при различении значений коэффициентов асимметрии:

$$\alpha_{A_i} = 1 - \int_{x_i}^{x_{i+1}} f[(x); A_i, \sigma_{\bar{A}_i}] dx = \quad (2)$$

$$= \int_{-\infty}^{x_i} f[x; A_i, \sigma_{\bar{A}_i}] dx - \int_{-\infty}^{x_{i+1}} f[(x); A_i, \sigma_{\bar{A}_i}] dx$$

- для ошибки первого рода при различении значений коэффициентов эксцесса при условии $A = A_i$:

$$\begin{aligned} \alpha_{E_i/A_i} &= 1 - \int_{y_i}^{y_{i+1}} f[x; E_i, \sigma_{\bar{E}_i}] dx = \\ &= 1 - \int_{-\infty}^{y_{i+1}} f[(x; E_i, \sigma_{\bar{E}_i})/A_i] dx + \\ &+ \int_{-\infty}^{y_i} f[(x; E_i, \sigma_{\bar{E}_i})/A_i] dx. \end{aligned} \quad (3)$$

В формулах (2), (3) $f(\cdot)$ – плотность вероятности нормального распределения с параметрами $A_i, \sigma_{\bar{A}_i}; E_i, \sigma_{\bar{E}_i}, x_i, x_{i+1}, y_i, y_{i+1}$ – значения разделяющих границ области, соответствующей H_i -й нулевой гипотезе (см. рис. 1).

Аналогичным образом записываются выражения для ошибок второго рода:

- при различении гипотез для H_i -й нулевой гипотезы:

$$\beta_i = \beta_{A_i} \cdot \beta_{E_i/A_i}; \quad (4)$$

- при различении значений коэффициентов асимметрии:

$$\begin{aligned} \beta_{A_i} &= \sum_{j=1}^{i-1} \left[1 - \int_{-\infty}^x f(x; A_j, \sigma_{\bar{A}_j}) dx \right] + \\ &+ \sum_{j=i+1}^{n_1} \int_{-\infty}^{x_{i+1}} f(x; A_j, \sigma_{\bar{A}_j}) dx, \end{aligned} \quad (5)$$

где n_1 – общее количество гипотез с различными значениями асимметрии;

- при различении значений коэффициентов эксцесса при условии $A = A_i$:

$$\begin{aligned} \beta_{E_i/A_i} &= \sum_{j=1}^{i-1} \left\{ 1 - \int_{-\infty}^{y_i} f[(y; E_j, \sigma_{\bar{E}_j})/A_i] dy \right\} + \\ &+ \sum_{j=i+1}^{n_2} \int_{-\infty}^{y_{i+1}} f[(y; E_j, \sigma_{\bar{E}_j})/A_i] dy, \end{aligned} \quad (6)$$

где n_2 – количество эталонных гипотез с асимметрией, равной A_i .

Существует тесная связь ошибки второго рода с мощностью критерия ($\mathcal{D}$), которая может быть записана в виде:

$$\mathcal{D} = 1 - \beta. \quad (7)$$

Мощность критерия $\mathcal{D}$ отражает величину вероятности отклонения неверной альтернативной гипотезы при определенной нулевой гипотезе. По аналогии с (1), (4) можно записать:

$$\mathcal{D}_i = \mathcal{D}_{A_i} \cdot \mathcal{D}_{E_i/A_i}; \quad (8)$$

- мощность различения значений коэффициентов асимметрии:

$$\begin{aligned} \mathcal{D}_{E_i/A_i} &= \sum_{j=1}^{i-1} \left\{ 1 - \int_{-\infty}^{y_i} f[(y; E_j, \sigma_{\bar{E}_j})/A_i] dy \right\} + \\ &+ \sum_{j=i+1}^{n_2} \int_{-\infty}^{y_{i+1}} f[(y; E_j, \sigma_{\bar{E}_j})/A_i] dy, \end{aligned} \quad (9)$$

- мощность различения значений коэффициентов эксцесса при условии $A = A_i$:

$$\begin{aligned} \mathcal{D}_{E_i/A_i} &= \sum_{j=1}^{i-1} \int_{-\infty}^{y_i} f[(y; E_j, \sigma_{\bar{E}_j})/A_i] dy + \\ &+ \sum_{j=i+1}^{n_2} \left\{ 1 - \int_{-\infty}^{y_{i+1}} f[(y; E_j, \sigma_{\bar{E}_j})/A_i] dy \right\}. \end{aligned} \quad (10)$$

Определим необходимый объем выборочных данных для обеспечения требуемой мощности различения значений асимметрии $\mathcal{D}_{A_i}$ и эксцесса $\mathcal{D}_{E_i}$ для H_i -й нулевой гипотезы. Пусть $\mathcal{D}_{S_i}$ определяет требуемую величину мощности различения для i -го признака ($S = 1$ при A , $S = 2$ при E). С учетом нормального распределения признаков в соответствии с (7) можно записать:

$$\mathcal{D}_{S_i} = 1 - \beta_{S_i} = F(\mu_i; m_i, \sigma_i), \quad (11)$$

где $F(\cdot)$ – интегральная функция нормального закона распределения с параметрами m_i, σ_i , определяемыми соответствующими значениями выборочного распределения оценки i -го признака при нулевой гипотезе H_i . Если принять с целью упрощения

дальнейшего вывода, что наиболее существенные вклады в уровень ошибки 2 рода вносят близлежащие альтернативы H_{i-1} , H_{i+1} (рис. 1) и их вклады примерно равны, то из (11) можно записать:

$$\frac{|q_{iS} - m_{(i+1)S}|}{\sigma_{(i+1)S}} = \mu_{1Si}, \quad (12)$$

где величина $\mu_{1Si} = F^{-1}(\mu_{1Si})$ является аргументом функции $F(\cdot)$.

С учетом (5), (6) и сделанного допущения о величинах вкладов альтернативных гипотез в уровень ошибки 2-го рода:

$$F_i(\mu_{1Si}) = 1/2 F(\mu_{Si}, m_i, \sigma_i). \quad (13)$$

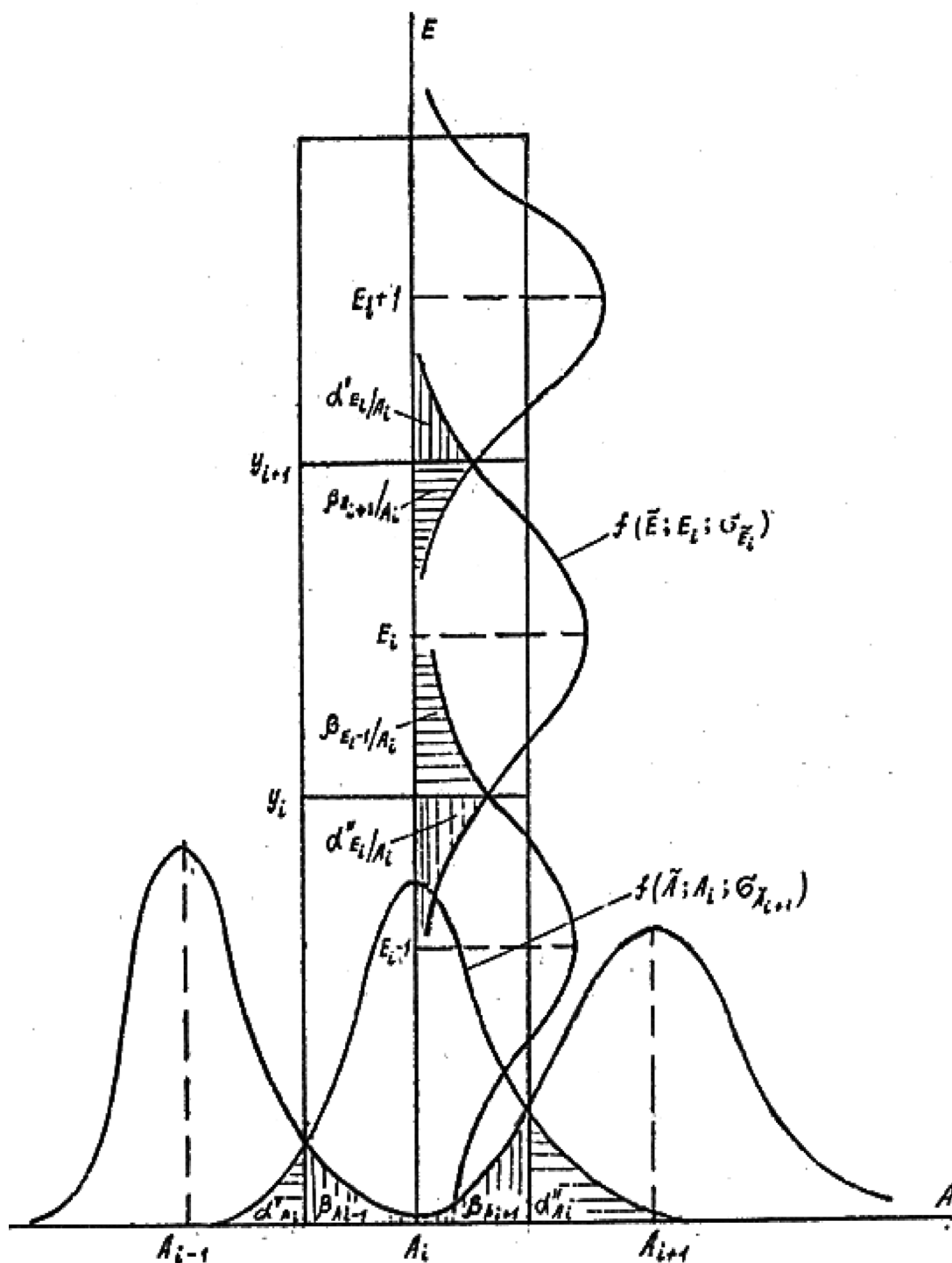


Рис. 1. Вероятности ошибок 1 и 2 рода на этапе выдвижения гипотезы о виде ПВ

В (12) q_{is} – значение i -й разделяющей границы i -го признака. Допущение оказывается тем более справедливым, чем ближе априорное распределение вероятностей появления близлежащих альтернативных гипотез к равновероятному и чем меньше различия в величинах дисперсий признаков, соответствующих этим гипотезам.

В связи с тем, что среднеквадратические отклонения оценок признаков обратно пропорциональны корню квадратному из объема выборки, можно записать:

$$\sigma_{Si} = \frac{K_{Si}}{\sqrt{N_{Si}}}. \quad (14)$$

В формуле (14) примем K_{Si} , N_{Si} – коэффициент пропорциональности и необходимый объем выборочных данных для обеспечения требуемой мощности различения для 1-го признака и при H_0 -й нулевой гипотезе.

Подставляя в (12) формулу для значений разделяющих границ (5) или (6) и учитывая (14), получаем уравнение:

$$ax^2 + bx + c = 0, \quad (15)$$

где $x = \sqrt{N_{Si}}$ и коэффициенты которого равны:

$$\left. \begin{aligned} a &= (m_{is} - m_{(i+1)s})^2; \\ b &= 2(m_{(i+1)s} - m_{is})\mu_{1Si}K_{(i+1)s}; \\ c &= K_{is}^2 2 \ln \frac{P_{(i+1)s}K_{is}}{P_{is}K_{(i+1)s}} - \mu_{1Si}^2 (K_{is}^2 - K_{(i+1)s}^2) \end{aligned} \right\} \quad (16)$$

При решении этого уравнения необходимый объем выборки N_{Si} может быть найден по следующей формуле

$$N_{Si} = \frac{\mu_{1Si}^2 \left(K_{(i+1)s} + K_{is} \sqrt{1 - \frac{2}{\mu_{1Si}^2} \ln \frac{P_{(i+1)s}K_{is}}{P_{is}K_{(i+1)s}}} \right)^2}{(m_{is} - m_{(i+1)s})^2}. \quad (17)$$

Из (17) видно, что при выбранном количестве эталонных классов распределений, т.е. при определенном расстоянии между теоретическими значениями признаков, соответствующих i -му и $i+1$ -му эталонным классам, и установленных априорных вероятностях появления H_i -й и H_{i+1} -й

эталонных гипотез, необходимый объем выборочных данных увеличивается с увеличением требований к мощности критерия, определяемой значением μ_{1Si} . Причем, из (17) минимальному значению N_{Si} будет соответствовать значение μ_{1Si} , равное $2 \ln \frac{P_{(i+1)s}K_{is}}{K_{(i+1)s}P_{is}}$. Условием существования N_{Si} будет условие неотрицательности подкоренного выражения в (17), т.е.

$$\mu_{1Si}^2 \geq 2 \ln \frac{P_{(i+1)s}K_{is}}{K_{(i+1)s}P_{is}}. \quad (18)$$

Из (18) видно, что если значение логарифма принимает отрицательное значение, то ограничений на значение μ_{1Si} не будет, т.е. уровень ошибок 2-го рода может быть установлен сколь угодно большим, и в соответствии с (7) мощность сколь угодно малой. Таким образом, величина объема выборочных данных может быть уменьшена с уменьшением значения μ_{1Si} , т.е. с увеличением допустимого уровня ошибок 2-го рода. В том случае, если логарифм в (18) принимает положительное значение, на минимально допустимое значение накладываются ограничения снизу, определяемые условием (18). В связи с этим, устанавливаются ограничения снизу и на величину необходимого объема выборочных данных N_{Si} .

Перепишем формулы для вычисления значений разделяющих границ (5), (6) в общем виде (см. формулу (19)).

При положительном значении логарифма в (19) с увеличением значения μ_{1Si}^2 и соответствующим увеличением N_{Si} , значение разделяющей границы q_{Si} сдвигается слева направо, в соответствии с (19). Таким образом, задаваясь значением μ_{1Si} , можно изменять значение разделяющих границ в соответствии с требуемой мощностью для каждого двух соседних классов по каждому из признаков.

Определим теперь необходимый объем выборочных данных для обеспечения требуемого уровня ошибок 1-го рода. Пусть допустимый уровень ошибок 1-го рода равен $F(\mu'_{2Si})$, $F(\cdot)$ – интегральная функция нормального распределения; индексы имеют смысл, принятый для (12). Тогда, принимая во внимание симметричность распределения оценок признаков и с учетом (3), (4) можно записать:

$$q_{is} = \frac{m_{(i+1)s}K_{is}^2 - m_{is}K_{(i+1)s}^2 - K_{(i+1)s}K_{is} \sqrt{(m_{(i+1)s} - m_{is})^2 + \frac{2(K_{is}^2 - K_{(i+1)s}^2)}{N_{Si}^2} \ln \frac{K_{is}P_{(i+1)s}}{K_{(i+1)s}P_{is}}}}{K_{is}^2 - K_{(i+1)s}^2}. \quad (19)$$

$$\alpha_{Si} = 2F(\mu_{2Si}) = F(\mu'_{2Si}), \quad (20)$$

Откуда

$$\mu_{2Si} = \frac{|qiS - miS|}{\sigma_{iS}}. \quad (21)$$

Подставляя в (21) выражение для вычисления значений разделяющих границ (5), получаем уравнение вида (15) со следующими коэффициентами:

$$\left. \begin{aligned} a &= (m_{(i+1)S} - m_{iS})^2; \\ b &= 2K_{iS}(m_{(i+1)S} - m_{iS})\mu_{2Si}; \\ c &= (K_{iS}^2 - K_{(i+1)S}^2)\mu_{2Si}^2 - 2K_{(i+1)S}^2 \ln \frac{P_{(i+1)S}K_{iS}}{P_{iS}K_{(i+1)S}} \end{aligned} \right\}. \quad (22)$$

Формула для определения необходимого объема выборки в этом случае имеет следующий вид:

$$N_{Si} = \frac{\mu_{2Si}^2 \left(K_{iS} + K_{(i+1)S} \sqrt{\mu_{2Si}^2 + 2 \ln \frac{P_{(i+1)S}K_{iS}}{P_{iS}K_{(i+1)S}}} \right)^2}{(m_{(i+1)S} - m_{iS})^2}. \quad (23)$$

В этом случае в соответствии с (23) увеличение μ_{2Si} (снижение уровня ошибок 1 рода) влечет за собой увеличение величины необходимого объема выборочных данных. Минимальный объем выборочных данных соответствует величине μ_{1Si} , равной $-2 \ln \frac{P_{(i+1)S}K_{iS}}{P_{iS}K_{(i+1)S}}$.

$$(\mu_{2Si})_{\text{пр}}^2 \geq -2 \ln \frac{P_{(i+1)S}K_{iS}}{P_{iS}K_{(i+1)S}}. \quad (24)$$

При положительном значении логарифма в (24) на величину μ_{2Si} не накладывается никаких ограничений и задавать μ_{2Si} можно сколь угодно малым. Это будет соответствовать большому уровню ошибок 1-го рода и малому объему выборочных данных. Если же значение логарифма в (24) отрицательно, то значение $(\mu_{2Si})_{\text{пр}}^2$ не может быть меньше удвоенного значения логарифма, взятого с обратным знаком. Таким образом, на минимально необходимый объем выборочных данных, требуемый для построения разделяющих границ с заданным уровнем ошибки 1-го рода, накладывается ограничение снизу.

При отрицательном значении логарифма в (19) уменьшение допустимого уровня ошибок 1-го рода и соответствующее этому увеличение объема выборочных данных N_{Si} ведет к смещению значения разделяющей границы справа налево. Таким образом, увеличивая уровень ошибок 1-го рода и выбирая необходимый при этом объем выборки по (23), можно изменять положение разделяющей границы q_{Si} на оси значений признаков.

Покажем теперь, что условия (24), (18) есть условия существования разделяющей границы q_{Si} для 1-го и $i+1$ -го репера данной двумерной шкалы при $P_{(i+1)S} \sigma_{iS} \neq P_{iS} \sigma_{(i+1)S}$.

В случае $P_{(i+1)S} \sigma_{iS} > P_{iS} \sigma_{(i+1)S}$ логарифм в (18) принимает положительное значение, т.е. есть ограничения на величину μ_{1Si} . Принимая во внимание (12) и (3) после преобразований получаем следующую формулу для вычисления разделяющих границ по i -му признаку для i -го эталонного класса в зависимости от мощности, определяемой величиной μ_{1Si} .

$$q_{Si} = 2 \left[m_{iS} + \sigma_{iS} \sqrt{\mu_{1Si} - 2 \ln \frac{P_{(i+1)S} \sigma_{iS}}{P_{iS} \sigma_{(i+1)S}}} \right]. \quad (25)$$

Из (25) видно, что условием существования q_{Si} будет условие не отрицательности подкоренного выражения, что совпадает с (19).

При $P_{(i+1)S} \sigma_{iS} < P_{iS} \sigma_{(i+1)S}$ логарифм в (24) принимает отрицательное значение, и существует ограничение снизу на уровень ошибки 1-го рода. Тогда, учитывая (21), после преобразований получаем аналитическую зависимость для вычисления значений разделяющей границы q_{Si} при минимально допустимом уровне ошибки 1-го рода:

$$q_{Si} = 2 \left[m_{(i+1)S} + \sigma_{(i+1)S} \sqrt{\mu_{2Si}^2 + 2 \ln \frac{P_{(i+1)S} \sigma_{iS}}{P_{iS} \sigma_{(i+1)S}}} \right]. \quad (26)$$

Тогда условие существования q_{Si} заключается в условии неотрицательности значения подкоренного выражения в (26), совпадающего с условием существования N_{Si} (24) в этом случае.

Как уже отмечалось в настоящем параграфе, снижение требований к мощности критерия или увеличение допустимого уровня ошибок 1 рода ведет к уменьшению величины необходимого в этих случаях объема выборочных данных N_{Si} , рассчитываемого по формулам (17), (23). однако, уменьшение μ_{1Si} или уменьшение μ_{2Si} может привести к тому, что при прочих равных условиях, значение разделяющей границы сместится влево или вправо (см. рис. 1) настолько, что значение разделяющей границы будет лежать вне интервала $m_{iS} + \Delta m_{iS}$, где $\Delta m_{iS} = m_{(i+1)S} - m_{iS}$. Правильное построение разделяющих границ классов реперов шкалы заключается в таком разбиении диапазона шкалы, при котором каждому классу соответствует своя область. Таким образом, встает вопрос определения минимально необходимого объема выборочных данных для обеспечения

правильного построения разделяющих границ. Для этого при $P_{(i+1)S} \sigma_{iS} > P_{iS} \sigma_{(i+1)S}$, необходимо выполнение следующих двух условий: соблюдения условия (19) и условия того, что значение разделяющей границы q_{Si} должно лежать в диапазоне $m_{iS} + \Delta m_{iS}$, т.е.

$$q_{Si} \leq - \sqrt{2 \ln \frac{P_{(i+1)S} \sigma_{iS}}{P_{iS} \sigma_{(i+1)S}} \sigma_{(i+1)S}^2 + m_{(i+1)S}} \left\{ \begin{array}{l} \frac{q_{Si} - m_{iS}}{\sigma_{iS}} \geq 0 \end{array} \right. \quad (27)$$

Выполнение равенств в (7) соответствует минимально допустимому объему выборочных данных для правильного построения разделяющих границ между i -м и $i+1$ -м эталонными классами по S -му признаку. Формулу для вычисления этого объема данных можно получить из (27), используя (2'):

$$N_{min iS} = \frac{2 \ln \left[\frac{P_{(i+1)S} K_{iS}}{P_{iS} K_{(i+1)S}} \right]}{(m_{(i+1)S} - m_{iS})^2}. \quad (28)$$

При увеличении $N > N_{min Si}$ значение разделяющей границы сдвигается внутрь интервала $m_{iS} + \Delta m_{iS}$.

В случае, когда $P_{(i+1)S} \sigma_{iS} < P_{iS} \sigma_{(i+1)S}$ для обеспечения правильности построения разделяющих границ, рассуждая аналогично первому случаю, можно также записать два условия:

$$q_{Si} \geq \sqrt{-2 \ln \frac{P_{(i+1)S} \sigma_{iS}}{P_{iS} \sigma_{(i+1)S}} \sigma_{iS}^2 + m_{iS}} \left\{ \begin{array}{l} \frac{q_{Si} - m_{(i+1)S}}{\sigma_{(i+1)S}} \geq 0 \end{array} \right. \quad (29)$$

Случай выполнения равенства в системе неравенств (29) соответствует минимально допустимому объему выборочных данных, требуемому для правильного построения разделяющих границ между 1-м и $i+1$ -м эталонными классами по S -му признаку. Формула для его вычисления может быть найдена в результате некоторых простых преобразований (29):

$$N_{min iS} = \frac{-2 \ln \left[\frac{P_{(i+1)S} K_{iS}}{P_{iS} K_{(i+1)S}} \right] K_{iS}^2}{(m_{(i+1)S} - m_{iS})^2}. \quad (30)$$

С увеличением объема выборки выше $N_{min iS}$ значение разделяющей границы q_{Si} сдвигается вглубь интервала $m_{iS} + \Delta m_{iS}$.

Минимально допустимый объем выборочных данных для правильного построения разделяю-

щих границ при выбранном количестве эталонных классов реперов шкалы и априорных вероятностях появления эталонных гипотез находится как максимальный по всем признакам (коэффициент асимметрии и эксцесса) и среди всех классов

$$N_{min} = \max_{i=1, \dots, n_1} S=1, 2 (N_{min iS}), \quad (31)$$

Таким образом, если выбирать объем выборочных данных N_1 в соответствии с условием:

$$N_1 \geq N_{min}, \quad (32)$$

то правильность построения разделяющих границ будет обеспечена.

Список литературы

1. Прокопчина С. В. Организация измерительных процессов в условиях неопределенности. Регуляризирующий байесовский подход // СПб. Сборник докладов Международной конференции по мягким вычислениям и измерениям SCM-98, 22–26 июня 1998, т. 1.
2. Прокопчина С. В. Диссертация на соискание ученой степени кандидата технических наук «Разработка метода аппроксимации плотности вероятности случайной величины и случайного процесса типовыми распределениями», Л., ЛЭТИ, 1976 г.
3. Прокопчина С. В. Основные принципы решения экономических задач в условиях информационной неопределенности // Научные труды Вольного экономического общества России. 2010. Т. 144.

References

1. Prokopchina S. V. (1998) Organization of measuring processes under uncertainty. Regularizing Bayesian approach. SPb. Collection of reports of the international conference on soft computing and measurements SCM-98, June 22–26, vol. 1.
2. Prokopchina S. V. (1976) Thesis for the degree of candidate of technical Sciences "Development of the method of approximation of probability density of a random variable and a random process by typical distributions", L., LETI.
3. Prokopchina S. V. (2010) Basic principles of solving economic problems in conditions of uncertainty. Proceedings of the Free economic society of Russia. Vol. 144.