**ВЛИЯНИЕ ОШИБОК ПЕРВОГО И ВТОРОГО РОДА
ПРИ ОПРЕДЕЛЕНИИ УЧАСТКОВ ЛОКАЛЬНОЙ СТАЦИОНАРНОСТИ
НА ДИСПЕРСИЮ ОШИБКИ ОЦЕНИВАНИЯ
НЕСТАЦИОНАРНЫХ СЛУЧАЙНЫХ ДАННЫХ**

К.А. Индюхов, К.А. Батенков

Академия ФСО России, Орёл, Российская Федерация

Показано, что при оценке нестационарных данных дисперсия ошибки оценивания будет зависеть как от объемов анализируемых подвыборок, так и от законов распределения, которым подчинены исследуемые случайные величины. Предложен подход к определению участков локальной стационарности. Проведен анализ влияния возможных ошибок при определении однородности анализируемых данных на общую дисперсию ошибки оценивания для случаев, когда исследуемые подвыборки распределены нормально. Также определены условия, при которых заведомо неоднородные подвыборки лучше объединять в один участок локальной стационарности с точки зрения уменьшения дисперсии ошибки оценивания.

Ключевые слова: ядерные методы оценки, участки локальной стационарности, общая дисперсия ошибки оценивания, ошибки первого и второго рода, гипотезы.

**THE INFLUENCE OF ERRORS OF THE FIRST AND SECOND KIND
IN DETERMINING LOCAL STATIONARY SITES ON DISPERSION
OF ERROR OF EVALUATION OF NON-STATIONARY RANDOM DATA**

K.A. Indukhov, K.A. Batenkov

Academy FGS of Russia, Orel, Russian Federation

It is shown that when evaluating non-stationary data, the variance of the estimation error will depend both on the volumes of the analyzed subsamples and on the distribution laws to which the studied random variables are obeyed. An approach to the definition of local stationarity sites is proposed. The analysis of the impact of possible errors in the determination of the homogeneity of the analyzed data for the total variance of the estimation error for the cases when the investigated sub-samples are normally distributed. The conditions under which obviously heterogeneous subsamples are better combined into one section of local stationarity are also determined from the point of view of reducing the variance of the error.

Keywords: kernel assessment methods, local stationarity areas, integral standard error of estimation, errors of the first and second kind, hypotheses.

Статистические методы анализа очень часто сталкиваются с проблемой, когда требуется доказать однородность рассматриваемых данных. Так же встречаются ситуации, когда требуется неоднородные в целом данные разделить на участки, на которых поведение анализируемых величин подчиняется определенным законам распределения. В общем виде задача определения стационарности формируется таким образом, что гипотеза $H_0: F_{k,j}(x) = F_{(k+i),j}(x)$ (называемая нулевой гипотезой) противопоставляется конкурирующей гипотезе $H_1: F_{k,j}(x) \neq F_{(k+i),j}(x)$, где $F_{k,j}(x)$ – эмпирическая функция распределения, построенная по подвыборке x_k размера j , а $F_{k+1,j}(x)$ – эмпирическая функция распределения, построенная по подвыборке x_{k+1} размера j . В данной работе предлагается использовать двухвыборочный критерий Смирнова, вероятные исходы, при использовании которого представлены в таблице.

Далее данные разбитые на участки локальной стационарности предлагается оценивать известными методами. Первоначально возникает желание, используя знания о природе анализируемых величин, выдвинуть предположение о виде их закона распределения и воспользоваться параметрическими методами оценки. В данной работе изначально предлагается использовать непараметрические методы оценки, поскольку априорные данные о форме распределения анализируемой последовательности отсутствуют. Среди таковых выделяются ядерные методы оценки, общий механизм применения которых довольно похож на построение гистограмм. Отличием в этом случае будет являться применение ядерных функций взамен индикаторной (1).

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K_{ker}(z) \quad (1)$$

где n – число реализаций случайной величины X ; h – ширина окна непараметрической оценки; X_i – i -е значение случайной величины X ; $\hat{f}(x)$ – непараметрическое оценивание функции распределения; $K_{ker}(z)$ – ядерная функция.

Результат работы такого метода можно оценить с помощью общей дисперсии ошибки оценивания [1]:

$$d[\hat{f}(x)] = \frac{\Phi_0}{nh} + \frac{h^4}{4} k_2^2 \Phi_1$$

где $\Phi_0 = \int_{\Omega_z} K_{ker}^2(z) dz$; $\Phi_1 = \int_{\Omega_x} [f''(x)]^2 dx$; $k_2 = \int_{\Omega_z} z^2 K_{ker}(z) dz \neq 0$; Ω_z – область определения подынтегральной функции z ; Ω_x – область определения подынтегральной функции x .

Вероятные исходы при использовании двухвыборочного критерия Смирнова

Верная гипотеза	H_0		H_1	
Принимаемое решение	H_0	H_1	H_0	H_1
Вероятность ситуации	$1-\alpha$	α	β	$1-\beta$
Общая дисперсия ошибки оценивания	$d_{0/0}$	$d_{1/0}$	$d_{0/1}$	$d_{1/1}$
Предполагаемая функция плотности распределения по результатам работы двухвыборочного критерия Смирнова	f_0	f_0/f_1	f_0	f_0/f_1
Реальная функция плотности распределения	f_0	f_0	f_0/f_1	f_0/f_1

Первый вариант развития событий наиболее вероятен, и его результатом является состоятельная оценка f_0 с определённой общей дисперсией ошибки оценивания $d_{0/0}$. Математическое выражение для вероятности такого события представлено в (2). Значение этой величины еще называют оперативной характеристикой критерия.

$$P\{f_0|f_0\} = 1 - \alpha. \quad (2)$$

Ситуация, когда верная гипотеза об однородности отклоняется, в дальнейшем приводит к уменьшению объемов анализируемых выборок, что, в свою очередь, приводит к снижению точности оценки. Вероятность такого события, представленного в выражении (3), называется уровнем значимости критерия α .

$$P\{f_0/f_1|f_0\} = \alpha. \quad (3)$$

Другой вариант, без ошибки, заданный в выражении (4), когда отвергается неправильная гипотеза, является со статистической точки зрения значимым событием и определяет мощность критерия.

$$P\{f_0/f_1|f_0/f_1\} = 1 - \beta. \quad (4)$$

Противоположная ситуация – выражение (5), когда принимается за правильную на самом деле неверная гипотеза, будет вносить ошибку в оценивание получившейся выборки.

$$P\{f_0|f_0/f_1\} = \beta. \quad (5)$$

Теперь вернемся к рассмотрению вариантов развития событий при использовании двухвыборочного критерия Смирнова для определения стационарности анализируемых

данных (табл. 1) и найдем общую дисперсию ошибки оценивания для них. В общем виде формулу для определения d можно преобразовать к следующему виду [2]:

$$d(f, j) = \frac{5}{4} C(K_k) \left\{ \int_{\Omega_x} f''^2(x) dx \right\}^{\frac{1}{5}} j^{-\frac{4}{5}}, \quad (6)$$

$$C(K_k) = k_2^{2/5} \left\{ \int_{\Omega_x} K_k(x)^2 dx \right\}^{4/5}.$$

где $C(K_k)$ – константа, значение которой зависит только от выбора вида ядерной функции.

В работе используется ядро Епанечникова, поэтому $C(K_k)$ будем считать константой и далее записывать C . Теперь для ситуации из формулы (2) и используя выражение (6), получаем

$$d_{0/0} = d(f_0, J), \quad (7)$$

где $J = j_0 + j_1$.

Для варианта с ошибкой первого рода из выражения (3) и предполагая, что противодействие импульсной помехе будет проводиться по последнему участку локальной стационарности, имеем

$$d_{1/0} = d(f_0, j_1). \quad (8)$$

Сравнивая общие дисперсии ошибок оценивания, полученные в выражениях (7), (8) и учитывая вышеописанную логику, получаем следующее:

$$d_{1/0} = \frac{(j_0 + j_1)^{\frac{4}{5}}}{j_1^{\frac{4}{5}}} d_{0/0} \quad (9)$$

Таким образом, общая дисперсия ошибки оценивания в случае ошибки первого рода будет зависеть от объемов анализируемых выборок. Используя выражение (9), можно подбирать такие j_0 и j_1 , чтобы минимизировать $d_{1/0}$.

Для случая из выражения (4), мы также будем считать общую дисперсию ошибки по последнему участку, с той лишь разницей, что теперь подынтегральная функция будет уже f_1 :

$$d_{1/1} = d(f_1, j_1) \quad (10)$$

Для выражения (5) общая дисперсия ошибки будет вычисляться по всей выборке. Однако необходимо учесть, что на самом деле интегрирование необходимо проводить по двум функциям, а также учесть неравноценный вклад каждой из них. Последнее связано с выбором объемов подвыборок.

$$f_{0,1}(x) = P_0 f_0(x) + P_1 f_1(x),$$

$$f_{0,1}''^2(x) = \{P_0 f_0''(x) + P_1 f_1''(x)\}^2, \quad (11)$$

где $f_{0,1}(x)$ – функция плотности распределения шума при ошибке второго рода.

$$P_0 = j_0/J,$$

$$P_1 = j_1/J.$$

Учитывая формулу (11) в выражении (6), получаем:

$$d_{0/1} = d(f_{0,1}, J) = \frac{5}{4} C \left\{ \int_{\Omega_x} (P_0 f_0''(x) + P_1 f_1''(x))^2 dx \right\}^{\frac{1}{5}} J^{-\frac{4}{5}} =$$

$$= \frac{5}{4} C \left\{ \int_{\Omega_x} (P_0 f_0''(x))^2 dx + 2 \int_{\Omega_x} P_0 f_0''(x) P_1 f_1''(x) dx + \int_{\Omega_x} (P_1 f_1''(x))^2 dx \right\}^{\frac{1}{5}} J^{-\frac{4}{5}} =$$

$$= \frac{5}{4} C J^{-\frac{6}{5}} \left\{ j_0^2 \int_{\Omega_x} f_0''^2(x) dx + 2 j_0 j_1 \int_{\Omega_x} f_0''(x) f_1''(x) dx + j_1^2 \int_{\Omega_x} f_1''^2(x) dx \right\}^{\frac{1}{5}} \quad (12)$$

Теперь мы можем найти отношение общих дисперсий ошибок при ошибке второго рода и правильном разделении выборки:

$$\frac{d_{0/1}}{d_{1/1}} = \frac{d(f_{0,1}, J)}{d(f_1, j_1)} = \frac{J^{-\frac{6}{5}} \left\{ j_0^2 \int_{\Omega_x} f_0''^2(x) dx + 2 j_0 j_1 \int_{\Omega_x} f_0''(x) f_1''(x) dx + j_1^2 \int_{\Omega_x} f_1''^2(x) dx \right\}^{\frac{1}{5}}}{\left\{ \int_{\Omega_x} f_1''^2(x) dx \right\}^{\frac{1}{5}} j_1^{-\frac{4}{5}}} =$$

$$= \frac{j_0^{\frac{4}{5}}}{j_1^{\frac{6}{5}}} \left\{ \frac{j_0^2 \int_{\Omega_x} f_0''^2(x) dx + 2j_0 j_1 \int_{\Omega_x} f_0''(x) f_1''(x) dx + j_1^2 \int_{\Omega_x} f_1''^2(x) dx}{\int_{\Omega_x} f_1''^2(x) dx} \right\}^{\frac{1}{5}}. \quad (13)$$

Данное соотношение зависит как от объемов подвыборок, так и от функций плотностей распределений. Анализируя формулу (13), очевидно, что отношение будет зависеть от выбора соотношения объемов подвыборок.

На рисунке отображен график зависимости $\frac{d_{0/1}}{d_{1/1}}$ от j_0 , при том что общий объем выборки J оставался неизменным, а f_0 и f_1 – функции плотности распределения нормального закона со следующими параметрами $N_0(0,1)$ и $N_1(0,\sigma_1^2)$, где σ_1^2 изменяется от 1 до 10. Видно, что после определенного значения объема подвыборки для некоторых значений среднеквадратического отклонения общая дисперсия ошибки для безошибочного случая оказывается намного больше, чем для случая с ошибкой второго рода. Можно также отметить, что для σ_1^2 от 5 до 7, график зависимости пересекает единичное значение дважды, при j_0 меньше 500 и больше 4500. На этих участках $d_{0/1}$ будет меньше, чем $d_{1/1}$, но при σ_1^2 больше 7 эта особенность пропадает. При изменении дисперсии нормального закона, по которому распределена вторая подвыборка, от 2 до 4 общая дисперсия ошибки оценивания для ошибочного случая будет меньше чем для случая правильного разделения. Таким образом, если первая подвыборка распределена по $N_0(0,1)$, а вторая по $N_2(0,2)$, $N_3(0,3)$, $N_4(0,4)$, то при таком подходе к оцениванию всей выборки лучше не разделять ее, а объединять в один участок стационарности.

СПИСОК ЛИТЕРАТУРЫ

1. Джеффри Расин. Непараметрическая эконометрия: вводный курс // Квантиль. 2008. № 4. С. 7–56.
2. Silverman B.W. Density estimation for statistics and data analysis. Springer-Science+Business Media, B.V., 1986. 175 с.