ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 14 за 2013 г.

Орлов Ю.Н.

Оптимальное разбиение гистограммы для оценивания выборочной плотности функции распределения нестационарного временного ряда

Рекомендуемая форма библиографической ссылки: Орлов Ю.Н. Оптимальное разбиение гистограммы для оценивания выборочной плотности функции распределения нестационарного временного ряда // Препринты ИПМ им. М.В.Келдыша. 2013. № 14. 26 с. URL: http://library.keldysh.ru/preprint.asp?id=2013-14

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

Ю.Н. Орлов

Оптимальное разбиение гистограммы для оценивания выборочной плотности функции распределения нестационарного временного ряда

Москва — 2013

Орлов Ю.Н.

Оптимальное разбиение гистограммы для оценивания выборочной плотности функции распределения нестационарного временного ряда

В работе исследуются распределения расстояний между эмпирическими плотностями распределений вероятностей для нестационарных временных рядов. Предложен метод нахождения оптимальной мелкости равномерного разбиения гистограммы, согласованный с точностью оценки эмпирического распределения. Оценки проведены новым непараметрическим методом.

Ключевые слова: оптимальное разбиение гистограммы, нестационарные распределения, временные ряды

Orlov Yu.N.

Optimal histogram interval for non-stationary time-series distribution function density estimation

The properties of distributions of the distances between two empirical distribution function densities for non-stationary time-series are investigated. The optimal choice of histogram interval is suggested on the basis of self-agreement accuracy of empirical probability. The estimation is given by a new non-parametric method.

Key words: optimal histogram class interval, non-stationary distributions, time series

Работа выполнена при поддержке гранта РФФИ, проект № 11-01-00444

Содержание

1. Введение

В настоящей работе изучается проблема аппроксимации выборочной плотности функции распределения (далее ВПФР) временного ряда, который в общем случае не является стационарным. Оценка вероятностей проводится методом гистограмм.

Очевидно, вид гистограммы зависит от того, как построены классовые интервалы принадлежности случайной величины. Даже в сильно упрощающем случае равномерного разбиения до сих пор нет удовлетворительного способа построения «правильного» разбиения области гистограммы, представляющей эмпирическую плотность вероятности. «Правильность» разбиения подразумевает, что в стационарном случае ошибка аппроксимации предположительно непрерывной плотности функции распределения кусочно-постоянной функцией минимальна. Трудность состоит в том, что оцениваемая плотность неизвестна, поэтому число интервалов сильно сказывается на виде распределения частот конечной выборки. С одной стороны, при фиксированной длине выборки укрупнение интервалов разбиения ведет к уточнению эмпирической вероятности попадания в них, но слишком сильно сглаживает изучаемое распределение. С другой стороны, измельчение интервалов делает вид распределения неоправданно изрезанным в силу малого количества данных, случайно попадающих в каждый из интервалов. Следовательно, задача построения гистограммы эквивалентна задаче оптимальной фильтрации шума в зависимости от объема данных.

Задача оптимального разбиения гистограммы довольно популярна среди специалистов по математической статистике, но нет единого устоявшегося мнения относительно ее решения, поскольку ответы носят полуэмпирический характер, либо требуют априорного знания закона распределения вероятностей, либо вообще имеют характер бытовых советов [1] «группировать так, чтобы было не меньше 6 и не больше 20 интервалов». В классических же курсах по математической статистике эта проблема не обсуждается, видимо, как не решенная до конца или, может быть, как не существенная. Последнее, как будет показано далее, не лишено оснований, хотя для нестационарных рядов эта задача вполне актуальна в связи с определенными требованиями к точности распознавания момента, когда распределение значимо изменилось.

В работах по статистике довольно часто упоминается эмпирическое правило Стерджеса [2], предложенное еще в 1926 г. Оно основано не на статистике, а на том формальном наблюдении, что если фактическое число попаданий k_i значений случайной величины в i-ый классовый интервал равно биноминальному коэффициенту

$$k_i = C_{n-1}^{i-1},$$

то, поскольку сумма по всем n интервалам разбиения должна быть равна количеству N элементов в выборке

$$\sum_{i=1}^{n} k_i = \sum_{i=1}^{n} C_{n-1}^{i-1} = 2^{n-1} = N ,$$

то соответствующее число классовых интервалов n зависит от длины выборки N по логарифмическому закону:

$$n = 1 + \log_2 N . \qquad (1)$$

Это, конечно, слишком идеалистическое требование на группировку данных, не имеющее отношения к реальной ситуации.

В 1942 г. Манн и Вальд [3] дали степенную оценку оптимального числа интервалов для применения критерия хи-квадрат:

$$n = 4\left(\frac{3}{4}(N-1)^2\right)^{1/5} . \qquad (2)$$

Из (2) следует, что при больших N число интервалов растет как $N^{2/5}$. Заметим, что оценки (1) и (2) не зависят от вида распределения и уже потому не являются удовлетворительными.

В 1950 г. Смирнов [4] показал, что уклонение гистограммы от графика неизвестной плотности вероятности убывает как $N^{-1/3}$. Скотт в 1979 г. [5] уточнил эту оценку для аппроксимации дифференцируемой плотности при больших длинах выборки и получил оптимальное число интервалов в виде

$$n = \left(\frac{N}{6}\int_{-\infty}^{+\infty} f'^2(x)dx\right)^{1/3} . \qquad (3)$$

В сборнике работ [6] была приведена оценка оптимального разбиения для аппроксимации дважды дифференцируемой плотности $f(x)$: число интервалов разбиения равно

$$n = \left(\frac{N\left|f''(x)\right|_{\max}}{4 f(x)_{\max}}\right)^{1/5} . \qquad (4)$$

Формулы (3) и (4) обладают вырожденностью в отношении равномерного распределения, для которого производная плотности равна нулю. Это означает, что их нельзя применить к плоским трапециевидным распределениям.

В книге [7] без ссылок и вывода приводится эмпирическая формула, пригодная, по мнению автора, «для широкого класса распределений»:

$$n = \frac{1+\kappa}{6} N^{2/5}, \qquad (5)$$

где κ есть эксцесс распределения. Формула эта на самом деле взята из известной книги Новицкого и Зографа [8], где она также приведена без доказательства, но со ссылкой на первоисточник (ныне труднодоступный) – автореферат кандидатской диссертации И.У. Алексеевой из Уфимского авиационного института на соискание ученой степени к.т.н. в 1975 г.

Таким образом, формулы (3-5) различаются по виду степенного закона, а вместе с (2) они отличаются от функционального закона (1).

Существует также подход, основанный на максимизации информационной энтропии. В работе [9] получена конструктивная оценка на число равных интервалов разбиения гистограммы:

$$n_{opt} = \arg\max_{n} \left(\sum_{i=1}^{n} k_i \ln k_i + N \ln n - n \right). \qquad (6)$$

Однако не очевидно, какое вообще отношение имеет информационная энтропия к задаче о разбиении, т.е. отношение-то имеет в силу статистики размещения данных по промежуткам, но не ясно, что дает ее максимизация.

Мы видим, что предлагаемые формулы не только различны, но и не особенно хорошо обоснованы, т.е. обоснования приводятся косвенные по отношению к решаемой задаче. Наиболее адекватной представляется формула (5), но и она является полуэмпирической.

Тем не менее, после перечисления и, частично, анализа большого количества результатов по оптимальному разбиению гистограмм Новицкий и Зограф приходят в [8] к следующим выводам: «оптимальное число интервалов разбиения, безусловно, существует; оно зависит от эксцесса сильнее, чем от объема выборки; зависимость $n = AN^{2/5}$, где A – теоретически вычисляемый коэффициент, в дальнейших уточнениях, по-видимому, не нуждается». Несмотря на последнее оптимистичное заявление, ситуация с оптимальным разбиением продолжает обсуждаться в издаваемой и в настоящее время обширной литературе по математической статистике.

Например, в справочнике [10] приведены «рекомендации» (именно в такой формулировке) выбирать число интервалов гистограммы в соответствии с уже упомянутыми работами [2-8], затем дана практическая рекомендация «выбирать как можно большее n, но не превышающее $N/5$», а в заключение читатель отсылается к ГОСТу «Прикладная статистика» [11] как источнику, содержащему «исчерпывающие рекомендации по методологии выбора числа классовых интервалов». В упомянутом ГОСТе, к сожалению, вновь приведены отнюдь не рекомендации – в каких случаях какое разбиение выбирать, а набор несовпадающих между собой результатов работ [2-8] и ряда других, менее известных, после чего приводится таблица рекомендованных ВНИИМ им. Д.И. Менделеева оценок [12], упоминаемых и в [8] (она короткая: если $N = 40 \div 100$, то $n = 7 \div 9$; если $N = 100 \div 500$, то $n = 8 \div 12$; если $N = 500 \div 1000$, то $n = 10 \div 16$; наконец, если $N = 1000 \div 10000$, то $n = 12 \div 22$). Перечень этих эмпирических правил завершается знакомым уже инженерным советом «выбирать как можно большее n, но не превышающее $N/(5 \div 10)$», т.е. авторы ГОСТа как будто не доверяют сложным формулам, а предлагают обходиться практическим здравым смыслом.

Не претендуя на завершение решения этой задачи, в настоящей работе предлагается еще один способ выбора оптимального равномерного разбиения гистограммы. Ситуация осложняется еще и тем, что для представляющих практический интерес нестационарных временных рядов не существует генеральной совокупности, которую якобы приближает гистограмма, и теоретически ей не от чего «уклоняться».

Отличием предлагаемого подхода от всех вышеперечисленных является введение согласованного уровня значимости, на котором принимается решение об аппроксимации теоретического распределения выборочным: нельзя априори требовать определенной значимости независимо от вида распределения, поскольку им (видом) определяется, в частности, различие между выборочными плотностями распределений независимых выборок равных длин. Желательно, чтобы уровень значимости, теоретически предъявляемый к точности решения задачи, имел бы реальное практическое выражение в виде доли ошибок, допускаемых при многократном повторении соответствующего случайного эксперимента. Заметим здесь, что для функции распределения (не плотности, а интегральной функции) непараметрический критерий Колмогорова-Смирнова [13] не зависит от способа разбиения гистограммы, но на оценку плотности этот результат, увы, не переносится.

На практике измеряемые величины всегда дискретны, поэтому можно было бы ввести естественное разбиение гистограммы в соответствии с точностью измерения. Например, если рассматривается временной ряд цен на некоторый финансовый инструмент и цены выражены в рублях с точностью до сотых, то естественным разбиением будет разбиение с шагом по цене, равным 0,01. Более мелкий шаг, естественно, делать бессмысленно. Однако для практического анализа такое разбиение будет неудобно, поскольку размах цен за период, например, в месяц может доходить до десятков единиц, что будет означать наличие тысяч классовых интервалов. В таком случае дискретная по существу случайная величина естественным образом начинает трактоваться как непрерывная, а приближенный график ее «истинного» распределения будет представлять собой гистограмму в объединенных классовых интервалах. Именно для такого случая задача оптимального разбиения гистограммы имеет практический смысл.

Для непрерывных же распределений ситуация в некотором роде парадоксальная: почти всегда оказывается, что способ разбиения не имеет определяющего значения (см. далее п. 3). Если, например, аппроксимируется непрерывное распределение, близкое к равномерному, то уменьшение числа интервалов (вплоть до одного!) актуально повышает точность аппроксимации. Если же анализируется другой предельный случай распределения с узким максимумом, то все промежутки гистограммы вне этого максимума не играют роли, а максимум так или иначе куда-нибудь попадет – либо в широкий промежуток, либо в набор узких, и гистограмма будет вполне адекватно отражать вид распределения.

В настоящей работе не анализируются варианты неравномерного разбиения области гистограммы, а оценивается оптимальная мелкость равномерного разбиения. Цель оптимизации – не наилучшая точность приближения генеральной плотности, а наиболее адекватное описание эволюции выборочной плотности функции распределения нестационарного временного ряда.

2. Оптимальное разбиение гистограммы

Рассмотрим последовательность из N значений случайной величины ξ, которые попали в определенные классовые интервалы, число которых n. Элементы этой последовательности обозначим x_j, $j = 1, 2, ..., N$. Пусть из них значение x_i встретилось k_i раз. Тогда выборочной плотностью функции распределения, оцениваемой по гистограмме по заданной выборке длины N, называется совокупность $f_N(i,t)$ величин

$$f_N(i,t) = \frac{k_i}{N}, \quad i = 1, 2, ..., n. \tag{7}$$

Время t в аргументе $f_N(i,t)$ указывает на текущий момент времени, от которого назад отсчитывается выборка длины N.

Согласно теореме Леви-Линдеберга [13] отклонение выборочного среднего значения $\overline{\xi}(N)$ стационарной случайной величины, определяемого по выборке длины N, от генерального среднего распределено асимптотически нормально с нулевым средним и стремящейся к нулю дисперсией σ^2 / N, где σ^2 есть дисперсия этой величины по гипотетической генеральной совокупности. Рассмотрим в качестве такой случайной величины саму эмпирическую частоту $f(j) = f_N(j)$ попадания в j-ый классовый интервал. В условиях, когда генеральная дисперсия не известна, а оценивается только по выборочной дисперсии s_N^2, следует рассматривать *t*-статистику Стьюдента, которая записывается в виде [13]

$$t = \sqrt{N-1}\,\frac{\left| f_N(j) - f^*(j) \right|}{s_N(j)}, \tag{8}$$

где $f^*(j)$ есть гипотетическая генеральная вероятность, если бы процесс был стационарным, а

$$s_N^2(j) = f_N(j) \cdot \left(1 - f_N(j)\right). \tag{9}$$

Тогда на уровне значимости ε выражение $\left| f_N(j) - f^*(j) \right|$ не превосходит величины

$$t_{1-\varepsilon/2}(N-1)s_N(j)/\sqrt{N-1},$$

где $t_\alpha(N-1)$ есть α-квантиль распределения Стьюдента с $N-1$ степенью свободы. При больших N можно считать $N-1 \approx N$ и число степеней свободы в квантиле распределения Стьюдента взять для простоты бесконечным (тогда это распределение совпадает с нормальным).

Следующее требование является ключевым. Поскольку уровень значимости не должен быть точнее уровня неопределенности в позиционировании доверительного интервала $\left| f_N(j) - f^*(j) \right|$, то естественно потребовать выполнения условия

$$\sum_{j=1}^{n} \left| f_N(j) - f^*(j) \right| \le \varepsilon . \tag{10}$$

Поэтому, если будет выполнено условие

$$\left| f_N(j) - f^*(j) \right| \le \varepsilon f^*(j), \tag{11}$$

то при условии $t_{1-\varepsilon/2} \dfrac{s_N(j)}{\sqrt{N}} \le \varepsilon f^*(j)$ будет достигнут требуемый уровень значимости для статистики Стьюдента. Однако если некоторые вероятности в результате выбранного разбиения сами оказались малы, много меньше ε, то нет необходимости требовать, чтобы и они были оценены с той же точностью. Поэтому уместно для каждой вероятности выбрать свою точность аппроксимации ε_j и считать, что требуемый в целом уровень значимости определяется средневзвешенной по разбиению точностью, так что

$$t_{1-\varepsilon/2} = \frac{1}{\Sigma_N(n)} \sum_{j=1}^{n} s_N(j) t_{1-\varepsilon_j/2}, \tag{12}$$

где сумма, определяющая влияние мелкости разбиения гистограммы на точность оценки эмпирических вероятностей, равна

$$\Sigma_N(n) = \sum_{i=1}^{n} s_N(j) = \sum_{i=1}^{n} \sqrt{f_N(i)\left(1 - f_N(i)\right)} . \tag{13}$$

Тогда из (11, 12) получаем, что

$$\sum_{j=1}^{n} t_{1-\varepsilon_j/2} \frac{s_N(j)}{\sqrt{N}} = t_{1-\varepsilon/2} \frac{\Sigma_N(n)}{\sqrt{N}} \le \varepsilon \sum_{j=1}^{n} f^*(j) = \varepsilon ,$$

откуда на уровне значимости ε следует оценка

$$\frac{t_{1-\varepsilon/2}}{\varepsilon} \le \frac{\sqrt{N}}{\Sigma_N(n)} . \tag{14}$$

При заданной точности ε и способе разбиения гистограммы формула (14) для знака равенства дает оценку на минимальную длину выборки, при которой эта точность достигается в среднем. Поскольку функция $t_{1-\varepsilon}$ табулирована (см. [13, 14]), то функция $t_{1-\varepsilon}/\varepsilon$ известна. Некоторые ее значения приведены в

табл. 1. Функция $t_{1-\varepsilon}/\varepsilon$ монотонно убывает с ростом ε, поэтому к ней существует обратная, значение которой и дает верхнюю оценку точности определения эмпирических вероятностей по заданному разбиению гистограммы. Обозначим для краткости

$$\varphi(\varepsilon)=\frac{t_{1-\varepsilon}}{\varepsilon},\quad \psi=\varphi^{-1},\quad z\equiv z(N,n)=\frac{\sqrt{N}}{\Sigma_N(n)}. \tag{15}$$

Тогда точность оценки ВПФР определяется формулой

$$\varepsilon=2\psi(2z). \tag{16}$$

Табл. 1. Зависимость функции $t_{1-\varepsilon}/\varepsilon$ от ε

ε	$t_{1-\varepsilon}$	$t_{1-\varepsilon}/\varepsilon$
0,0005	3,291	6582,0
0,0025	2,807	1122,8
0,0050	2,576	515,2
0,0075	2,432	324,3
0,010	2,326	232,6
0,015	2,170	144,7
0,020	2,054	102,7
0,025	1,960	78,4
0,05	1,645	32,9
0,10	1,282	12,8
0,15	1,036	6,9
0,20	0,842	4,2
0,30	0,524	1,7
0,40	0,253	0,6

Сумма (13) выражает качество приближения плотности гистограммой, поскольку, чем меньше сумма, тем выше точность оценки ВПФР, т.е. тем меньше число ε. С увеличением числа интервалов сумма (13) возрастает, что означает снижение точности оценки ВПФР. В стационарном случае эта сумма представляет собой некий эффективный функционал учета особенностей графика функции плотности при аппроксимации плотности генерального распределения.

При $\varepsilon>0{,}01$ имеет место аппроксимация [14] квантиля нормального распределения, с которым при больших N совпадает квантиль распределения Стьюдента:

$$t_{1-\varepsilon/2}=\sqrt{-\frac{\pi}{2}\ln\left(1-(1-\varepsilon)^2\right)}. \tag{17}$$

Относительная ошибка аппроксимации (17) составляет 0,037. Графики функции $\varphi(\varepsilon)$ и ее аппроксимации по формуле (17) показаны на рис. 1. Видно,

что практически во всем диапазоне изменения точности, кроме самой прецизионной ($\varepsilon < 0{,}01$), формула (17) может считаться достаточно точной.

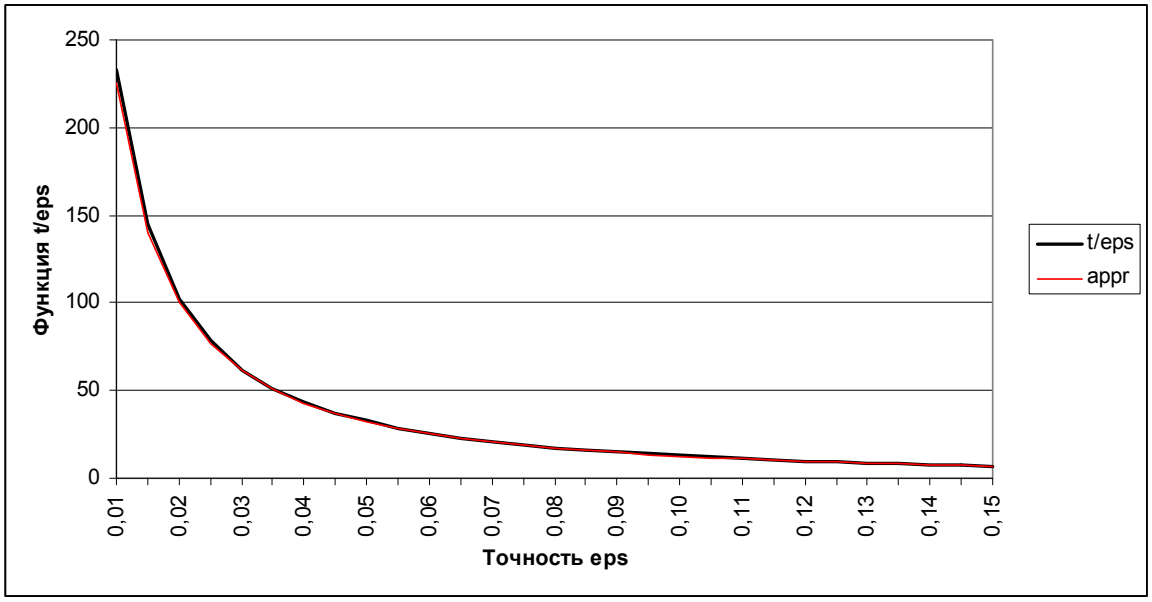

Рис. 1. Вид функции $\varphi(\varepsilon)$ из (15)

Заметим теперь, что если вероятности оценены с точностью ε, то, как легко показать, с той же точностью оценивается и среднее значение. Поэтому знание фактических величин с большей точностью будет излишним (не вообще, а в смысле превышения точности). С другой стороны, если выбрано разбиение на n классовых интервалов, то среднее значение по выборке отличается от среднего значения по гистограмме на величину $1/n$.

Исходя из этих качественных соображений, можно сделать вывод, что на практике не имеет смысла выбирать мелкость разбиения гистограммы, превышающую точность оценки вероятностей (по этой самой гистограмме), если только вопрос не касается аппроксимации графика известной функции. Для нестационарных временных рядов такой функции нет, поэтому критерий минимального уклонения от генеральной плотности не состоятелен. Так как число интервалов желательно выбрать наибольшим из возможных, то оптимальным равномерным разбиением гистограммы будем тогда считать ближайшее натуральное число n (классовых интервалов) к решению уравнения, согласующего уровень значимости (16) и мелкость разбиения:

$$\frac{1}{n} = 2\psi\left(\frac{2\sqrt{N}}{\Sigma_N(n)}\right). \qquad (18)$$

Может показаться, что статистический шум имеет другую природу и проявляется безотносительно метода укрупнения классовых интервалов, как, например, для случайного эксперимента по бросанию кубика с шестью исходами, где предлагаемая трактовка неуместна. Однако следует принять во внимание, что шесть исходов – это наше субъективное видение результатов эксперимента. Если грани окрашены каждая в один из двух цветов и, по

мнению наблюдателя, значение имеет только цвет грани, то и распределение вероятностей будет другим, отличным от «шестигранного» эксперимента. Поэтому хотя бы отчасти шум имеет субъективную составляющую.

Решение уравнения (18) может не существовать в натуральных числах, но оно всегда существует и единственно относительно точности ε, так как с уменьшением ε число разбиений $n = 1 + [1/\varepsilon]$ не убывает, как и аргумент функции ψ в правой части (18), причем сама функция ψ при этом монотонно возрастает, а области значений обеих этих непрерывных функций ε и $\psi(\varepsilon)$ совпадают. Оптимально число разбиений n_{opt} располагается между n (некоторым произвольно выбранным числом разбиений) и $1 + [1/\varepsilon(n)]$, или наоборот, в зависимости от того, какое из этих чисел меньше ($\varepsilon(n)$ определяется в (16)). Изменение числа интервалов в любую сторону приводит к противоположному изменению точности ε, так что с каждой такой итерацией n_{opt} заключено во все меньшем целочисленном промежутке. В результате за конечное число шагов находится разбиение, согласованное с точностью аппроксимации эмпирических вероятностей.

Применим эту методику к меняющейся во времени ВПФР. Тогда величина (13) является еще и функцией времени. Естественно, нет смысла на каждом временном шаге менять разбиение гистограммы, так как тогда будет несколько затруднительно сравнивать между собой распределения, меняющиеся во времени. Поэтому следует использовать не локально-оптимальное разбиение гистограммы, а ввести некоторое характерное разбиение в соответствии с характерным же уровнем точности эмпирических вероятностей. Последний строится следующим образом.

Пусть в результате произвольно выбранного разбиения гистограммы на n классовых интервалов получен временной ряд точностей оценки ВПФР в соответствии с (16). Построим плотность $\nu(\varepsilon, n)$ распределения этих точностей по всему доступному для анализа массиву данных и определим согласованный уровень точности $\varepsilon^*(n)$, определяемый как квантиль распределения $\nu(\varepsilon, n)$, равный собственному уровню значимости, т.е.

$$\int_0^{\varepsilon^*} \nu(\varepsilon, n) d\varepsilon = 1 - \varepsilon^*. \qquad (19)$$

Определим отвечающее этому согласованному уровню точности число промежутков разбиения $n^* = [1/\varepsilon^*(n)]$. Если оказалось, что $n^* > n$, число интервалов увеличивается и процедура повторяется, строится новая плотность $\nu(\varepsilon, n)$ и т.д. Если же $n^* < n$, число интервалов уменьшается. Таким образом, процесс нахождения оптимального разбиения будет иметь итерационный характер. Наилучшее целочисленное приближение к решению уравнения (18) и

даст оптимальную мелкость равномерного разбиения гистограммы. Эта мелкость в конечном итоге зависит от длины выборки и от суммы (13).

3. Сравнение с другими методами разбиения

Применим описанный алгоритм сначала для аппроксимации стационарных распределений, когда результаты численного моделирования могут быть сравнены с теоретическими, поскольку сумма (13) вычисляется тогда в явном виде независимо от реализации временного ряда. В стационарном случае наибольшее значение суммы (13), как и любого функционала от плотности распределения при условии сохранения нормировки, достигается на равномерном распределении, и тогда $\Sigma_N(n) = \sqrt{n-1}$. Еще раз подчеркнем, что оптимальность понимается не в смысле минимального отклонения от генеральной плотности (для этой цели достаточно разбить весь промежуток на один-единственный интервал), а в смысле согласования точности попадания числа в тот или иной классовый интервал в зависимости от ширины шага гистограммы, с точностью оценки вероятности такого попадания.

Используя приближение (17), можно для равномерного распределения получить оптимальную точность ε его аппроксимации по гистограмме в зависимости от длины N выборки как обращение функции $N(\varepsilon)$ из равенства

$$N = -\frac{\pi \ln\left(1-(1-\varepsilon)^2\right)}{\varepsilon^3}. \qquad (20)$$

Интересно сравнить получаемые результаты с другим предельным случаем – узким распределением типа «колокола», для чего возьмем нормальное распределение с центром в точке 0,5 и стандартным отклонением, равным $\sigma = 0{,}1$. Тогда на отрезке [0;1] содержится почти все распределение, так как интеграл от плотности по этому отрезку равен

$$erf\left(1/(2\sigma\sqrt{2})\right) \approx erf\,(3{,}54) \approx 0{,}9995.$$

Для полноты картины следует рассмотреть и промежуточный случай, когда плотность неравномерно «размазана» по всему отрезку [0;1]. Для этой цели возьмем стационарный ряд хаотической автономной динамической системы, задаваемой логистическим отображением

$$x(t+1) = 4x(t)\cdot\left(1-x(t)\right). \qquad (21)$$

Как известно из [15], динамически-инвариантная мера этого отображения дается формулой

$$d\mu = \frac{dx}{\pi\sqrt{x(1-x)}}. \qquad (22)$$

Оптимальные равномерные разбиения отрезка [0;1] в зависимости от длины выборки, отвечающие равномерному и нормальному распределениям, а также для плотности динамически-инвариантной меры (22) даны на рис. 2.

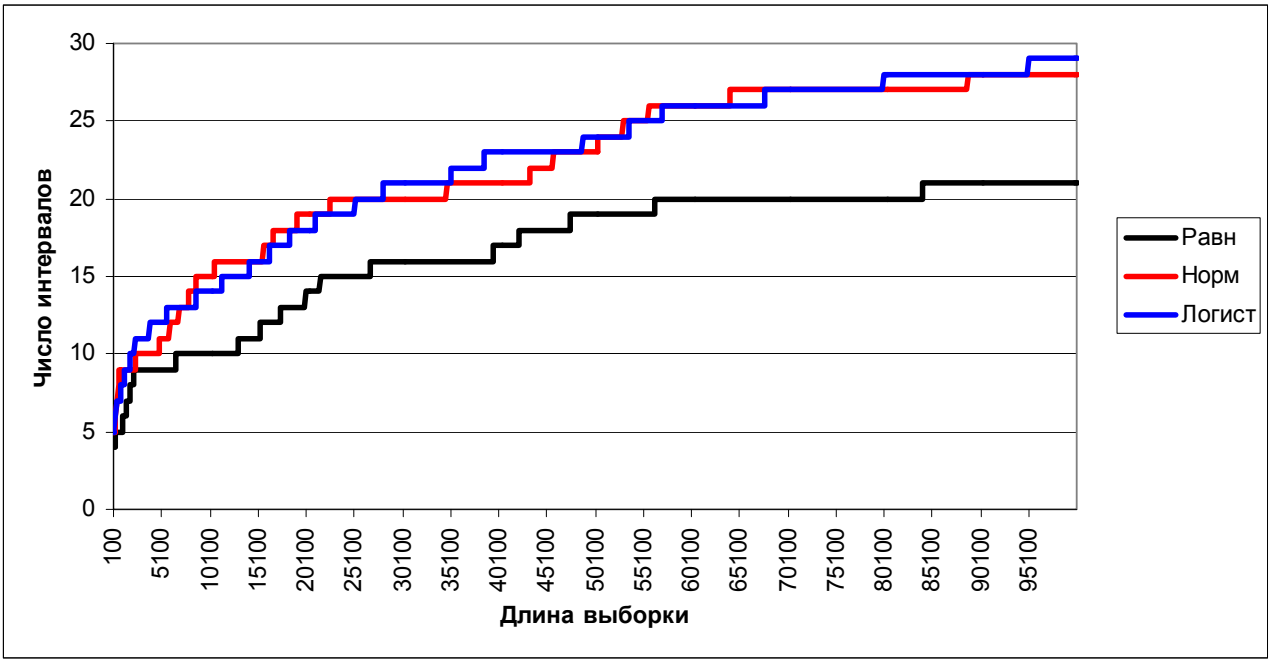

Рис. 2. Оптимальное число интервалов в зависимости от длины выборки для некоторых стационарных распределений

Как и следовало ожидать, равномерное распределение требует меньше промежутков разбиения, чем нормальное, а логистическое отображение чуть больше, чем нормальное. Качественно это связано с тем, что в тех областях, где генеральное распределение имеет исчезающее малую плотность, число интервалов и их ширина не имеют значения. Для распределений с большой дисперсией (а не эксцессом, в отличие от (5)) число интервалов может быть малым. Впрочем, на выборках относительно небольших длин (до 1000) точности приближения всех распределений близки.

На выборках средних 5-10 тыс. различия в разбиениях уже заметны. Может показаться, что выборка в 10 тыс. данных достаточно представительна, в [8], например, выборка в 1000 данных считается длинной. На самом деле это не так, поскольку наилучшая достижимая точность оценки выборочных вероятностей даже для выборки в 10 тыс. данных из нормального распределения составляет всего лишь 0,05. Для точности же 0,01 требуется более миллиона данных. Если раньше такое количество данных в практическом эксперименте по измерению, например, числа отказов электротехнического прибора получить было невозможно даже при использовании для этой цели всех выпущенных приборов, то в настоящее время существует область деятельности, поставляющая данные десятками, а то и сотнями миллионов: это биржевые временные ряды, когда в сутки по отдельному финансовому инструменту проходит порядка 100 тыс. сделок. В результате статистический анализ можно было бы провести с очень высокой точностью, если бы не нестационарность изучаемых процессов. Еще раз подчеркнем, что используемый в данной работе подход согласования числа интервалов гистограммы и точности оценки

эмпирических вероятностей годится только для рядов случайных величин, значения которых требуют агрегирования.

Сравним точность приближения вышеприведенных непрерывных распределений по выборке в 10 тыс. точек гистограммой, разбитой по методике (16-19) на соответствующее количество промежутков, с вариантами, следующими из формул (1-5). Числа разбиений приведены в табл. 2.

Табл. 2. Число промежутков разбиения по разным методикам

Формула методики	Равн	Норм	Логист
(1)	13	13	13
(2)	150	150	150
(3)	-	63	-
(4)	-	12	-
(5)	19	27	33
(18)	**10**	**15**	**14**

Для сравнения точности аппроксимаций были сгенерированы 2000 выборок из равномерного, нормального и логистического распределений на отрезке $[0;1]$, и для каждого из них вычислена интегральная ошибка отклонения

$$\varepsilon_k = \int_0^1 \left| f_{appr,k}(x) - f(x) \right| dx, \quad k = 1,...,2000. \qquad (23)$$

После этого для каждого метода была построена эмпирическая интегральная функция распределения ошибок по этим 2000 экспериментов.

Оказалось, что для аппроксимации нормального распределения с узким максимумом, как в данном случае, все методы дают близкие значения ошибок, среднюю ошибку (0,17) и стандартное отклонение ошибки (0,03). Различия проявляются в третьем знаке после запятой. Определенная позитивная особенность предложенного метода (16-18) в том, что вероятность малых ошибок в нем наибольшая: до значения $\varepsilon = 0{,}15$ интегральное распределение ошибки идет выше распределений, отвечающих другим методам. Впрочем, преимущество оказалось небольшим и проявилось во втором знаке после запятой.

Для равномерного распределения ситуация, очевидно, вырожденная: чем меньше число интервалов, тем точнее приближается плотность вероятности. Укажем, что, как и должно быть по определению согласованной точности аппроксимации, вероятность ошибки, меньшей 0,05, по оптимальному разбиению (18) равна 0,95, а для разбиения по формуле (5) она равна 0,91.

Такое же наблюдение относится и к плотности меры (22). Все разбиения, кроме близкого варианта по формуле (1), дают худшую аппроксимацию. Отметим также, что распределения отклонений (23) не имеют нормального вида, а больше похожи на гамма-распределения.

4. Сравнение с критерием Колмогорова-Смирнова

Подход, развиваемый в настоящей работе, использует согласованные уровни точности и значимости (16, 19). Представляется, что он дает наиболее корректные оценки мелкости разбиения в непараметрической статистике. Полезно сравнить получающиеся результаты с классическим непараметрическим критерием Колмогорова-Смирнова.

Напомним (см., напр., [13]), что мера уклонения интегральной выборочной функции распределения (ВФР) $F_N(x)$ от теоретического распределения $F(x)$ дается следующим утверждением. Если ряд стационарный, то при $N \to \infty$ разность $F_N(x) - F(x)$ распределена асимптотически нормально с параметрами $\mu = 0,\ \sigma^2 = F(x)\big(1 - F(x)\big)/N$. Величина уклонения ВФР от асимптотики неравномерна по x. Для получения равномерных оценок уклонения рассматривается статистика Колмогорова $D_T = \sup_x \left|F_N(x) - F(x)\right|$. Распределение этой статистики дается теоремой Колмогорова. Если теоретическое распределение $F(x)$ генеральной совокупности непрерывно, то ВФР статистики $\sqrt{N} D_N$ сходится при $N \to \infty$ к функции Колмогорова $K(z)$:

$$\lim_{N\to\infty} P\left\{0 < \sqrt{N} \sup_x \left|F_N(x) - F(x)\right| < z\right\} = K(z) = \sum_{k=-\infty}^{\infty} (-1)^k \exp\left(-2k^2 z^2\right). \qquad (24)$$

Обычно распределение генеральной совокупности не бывает известно. Тогда, в предположении, что оно существует, для изучения вопроса принадлежности двух ВФР одному и тому же распределению генеральной совокупности применяется статистика Смирнова $S_N = \sup_x \left|F_{1,N}(x) - F_{2,N}(x)\right|$, где для простоты рассмотрены выборки равных объемов. Для этой статистики справедливо следующее утверждение. Пусть проводятся две независимых серии испытаний по составлению выборок из некоторой генеральной совокупности. Тогда ВФР статистики $\sqrt{N/2} S_N$ сходится при $N \to \infty$ к функции Колмогорова в смысле формулы (24). Это означает, что если для двух ВФР было найдено значение S_N и вычислена величина $z = \sqrt{\frac{N}{2}} S_N$, то величина $1 - K(z)$ приближенно считается равной вероятности того, что $\sqrt{\frac{N}{2}} S_N \geq z$. Если величина $1 - K(z)$ мала, то осуществилось маловероятное событие, несовместимое с понятием случайности, и эти выборки следует считать различными. Некоторая неопределенность вывода состоит в том, что надо априори задать желаемый уровень малости критерия $1 - K(z)$. Какую вероятность считать достаточной для того, чтобы признать выборки одинаковыми?

Поставим аналогичный предыдущему статистический эксперимент по генерации выборок из нормального распределения на [0;1] и построим функцию распределения расстояний (24) между выборками. На рис. 3 показан временной ряд расстояний между ВФР двух выборок длины $N = 1000$ данных, а также и ряд расстояний между соответствующими ВПФР. Отличие между этими рядами состоит в том, что в первом случае берется норма в пространстве непрерывных функций (супремум модуля разности), а во втором – в пространстве суммируемых гистограмм (сумма модулей разностей). Как и должно быть, норма в C не превосходит нормы в $L1$.

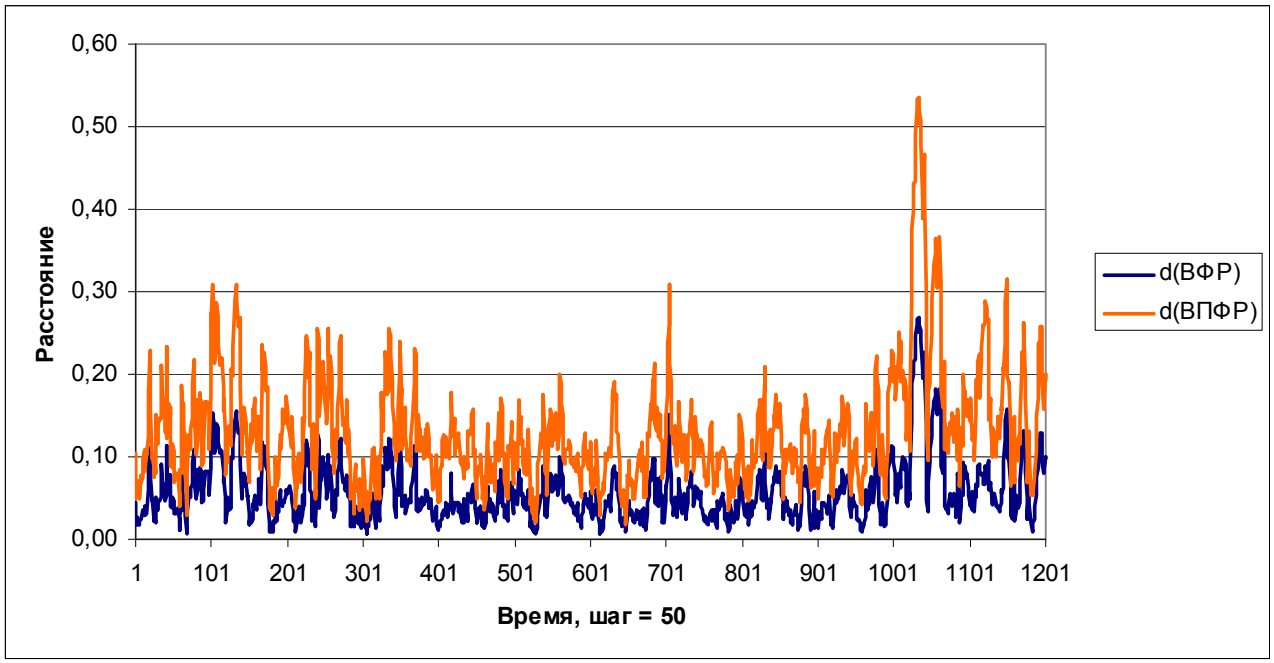

Рис. 3. Временные ряды расстояний между независимыми выборочными распределениями из нормальной совокупности

Распределения расстояний, временные ряды которых показаны на рис. 4, даны на двух графиках рис. 4-5.

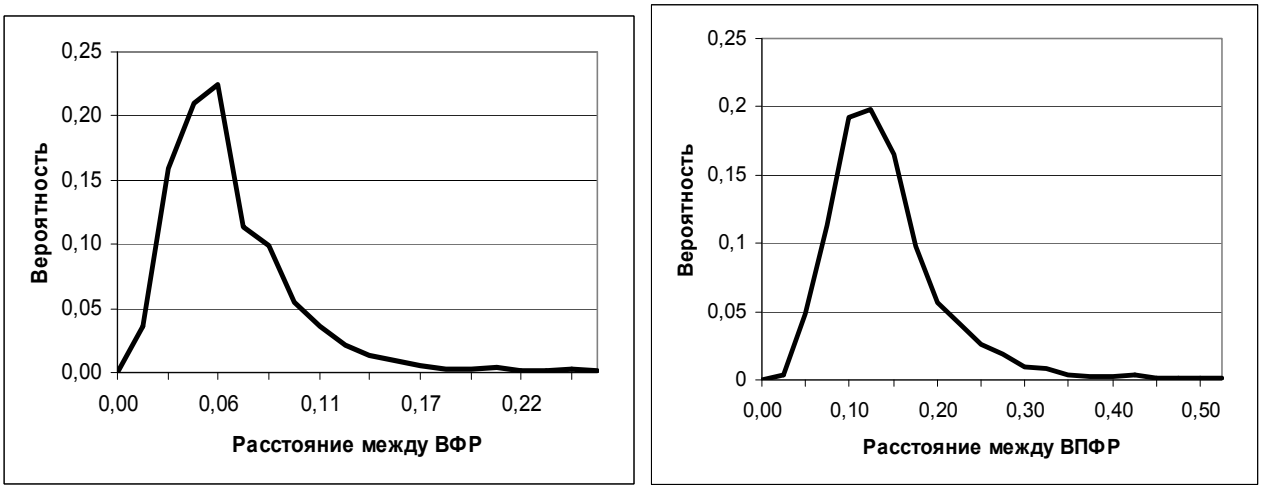

Рис. 4-5. Распределения расстояний между стационарными выборочными распределениями

Поскольку все выборки в этом примере взяты из одного распределения, то ошибкой на уровне значимости ε будет непризнание выборок «одинаковыми», если расстояние между ними оказалось больше ε. Смысл же уровня значимости таков, что по итогам 100 экспериментов в среднем в 100ε случаях будет сделан ошибочный вывод. Но из рис. 3, в частности, следует, что при сравнении ВФР для выборок в 1000 данных бессмысленно задавать уровень значимости менее 0,01, так как тогда почти всегда придется признавать выборки различными, а на самом деле это не так. Следовательно, при анализе выборок определенной длины неправильно задавать априори желаемый уровень значимости, так как для заданной длины выборки N лишь при одном значении $\varepsilon = \varepsilon_0(N)$ вероятность превышения значения ε_0 равна достоверности используемого для этой цели критерия, и это значение находится из уравнения

$$1 - \varepsilon_0 = K\left(\sqrt{\frac{N}{2}}\varepsilon_0\right). \qquad (25)$$

Решение уравнения (25) единственно, поскольку правая часть монотонно возрастает от нуля до единицы, а левая монотонно убывает от единицы до нуля. График решения $\varepsilon_0(N)$ приведен на рис. 6, а некоторые значения – в табл. 3.

Непосредственно по эмпирическому графику распределения расстояний между независимыми ВФР по длинам 1000 данных находим, что вероятность превышения расстояния, равного уровню значимости, равна примерно 0,11, т.е. интеграл от этого распределения от 0 до 0,11 равен примерно 0,89. С другой стороны, численное решение уравнения (25) для $N = 1000$ дает значение $\varepsilon_0(1000) \approx 0{,}06$ (см. табл. 3), а не 0,11. Это заметное расхождение между эмпирическими и теоретическими данными обусловлено мелкостью (точнее, крупностью) разбиения гистограммы. Хотя сам критерий (24) не связан ни с какой гистограммой, но он применяется к непрерывной функции распределения, а точность, с которой мы эту непрерывность реализуем, как раз и равна шагу гистограммы, который, как это следует из рис. 2, равен для выборок 1000 точек примерно 0,1. Поэтому критерий Колмогорова-Смирнова на практике редко дает достоверный результат, так как он требует мелкости разбиения, по крайней мере в два раза превосходящей оптимальную. Основная причина ошибки по этому критерию – не статистическая (мало данных), а своего рода техническая, связанная с низкой точностью аппроксимации непрерывного распределения.

На достаточно высоком уровне достоверности (выше, чем 0,99) численно найденное решение уравнения (25) при больших объемах выборок аппроксимируется эмпирической зависимостью

$$\varepsilon_0(N) = 0{,}06N^{-0{,}45}. \qquad (26)$$

Получаемая отсюда функция $N(\varepsilon)$ всегда идет ниже, чем аппроксимация (20), т.е. равная точность ε достигается в случае (25) на большем объеме данных.

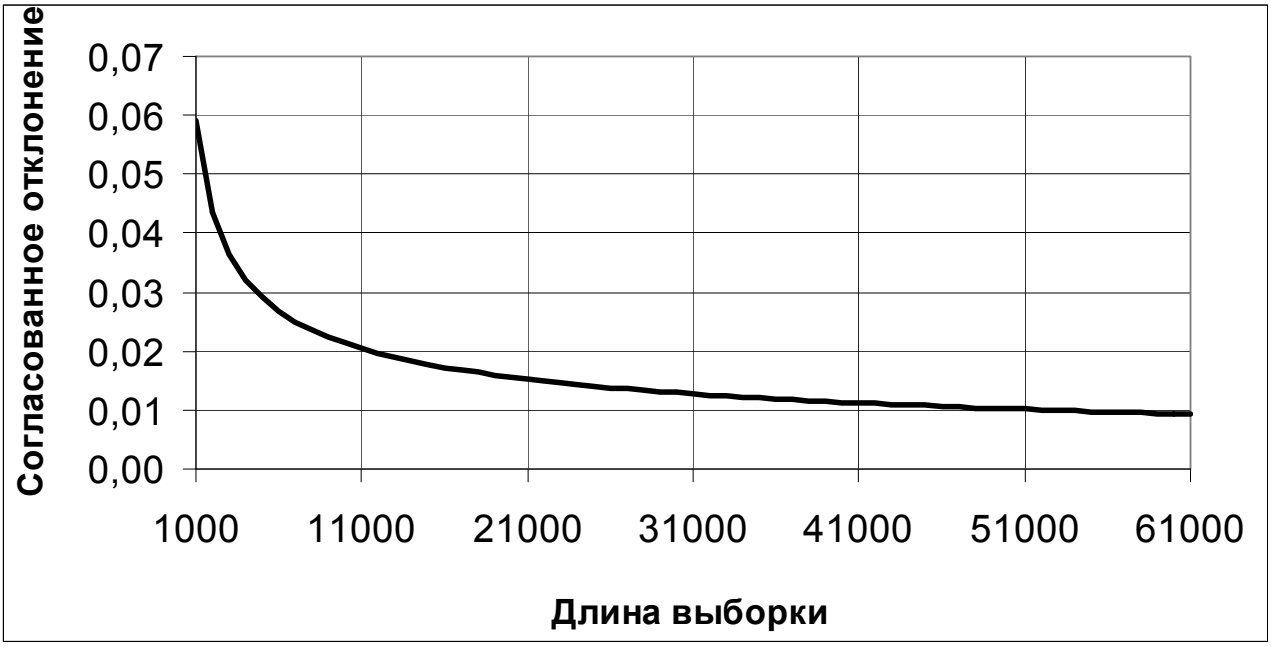

Рис. 6. Согласованный уровень значимости
для критерия Колмогорова-Смирнова

Табл. 3. Значения согласованного уровня значимости критерия (24)
(расчет автора)

Длина выборки N	Уровень значимости $\varepsilon_0(N)$
100	0,18
200	0,14
300	0,10
400	0,09
500	0,08
1000	0,05926
2000	0,04364
3000	0,03641
4000	0,03220
5000	0,02910
10000	0,02135
50000	0,01023

С дальнейшим увеличением N согласованный уровень значимости меняется очень слабо. Например, для $N = 100$ тыс. $\varepsilon_0(N) = 0{,}009$.

Поскольку на практике временной ряд изначально не рассматривается как выборка из непрерывного распределения, то кластеризация его значений в классовые интервалы естественно приводит к анализу расстояний между выборочными плотностями, а не интегральными распределениями.

5. Распределение расстояний между нестационарными ВПФР

Рассмотрим теперь построение оптимального разбиения по описанной методике в нестационарном случае на примере биржевых рядов. Были взяты ряды индекса RTS по часовым ценам закрытия и тиковые данные фьючерсов: GC на золото и CL на нефть. Число интервалов в зависимости от длины выборки приведено на рис. 7. На рис. 8 оптимальное число интервалов вычислено для выборки достаточно большой длины в 100 тыс. данных (фьючерсы CL на нефть).

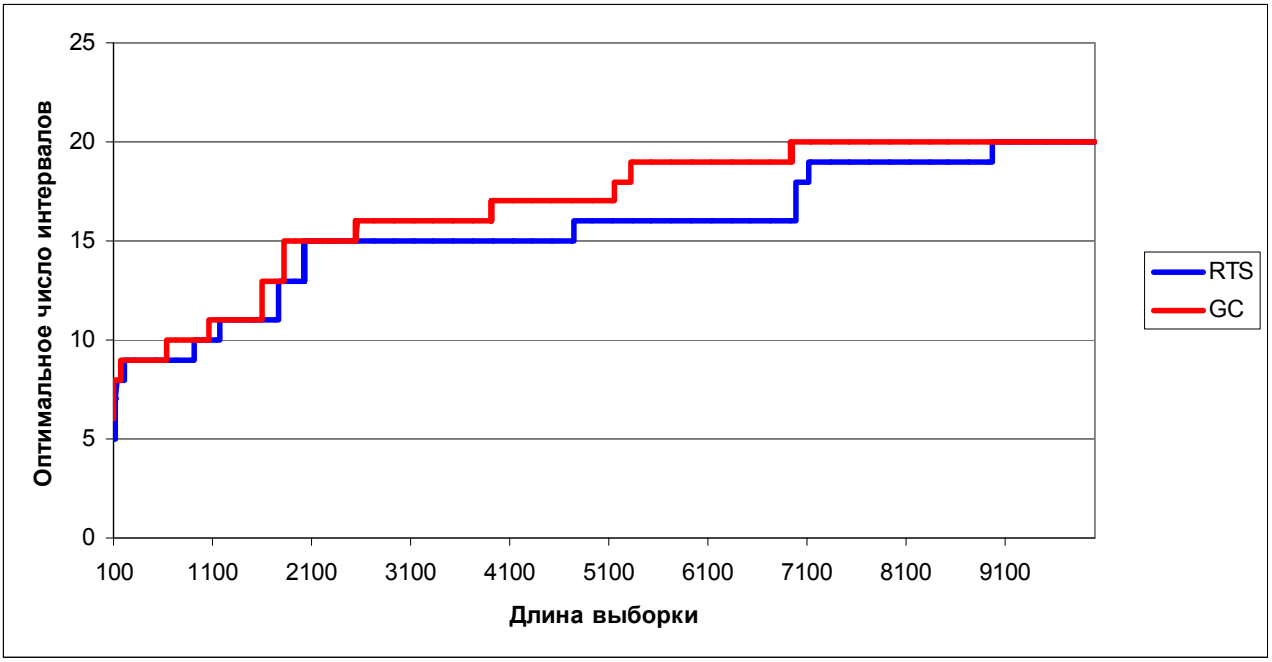

Рис. 7. Оптимальное число интервалов в зависимости от длины выборки

Различие в рассмотренных временных рядах проявляется, в частности, в том, что существует диапазон длин выборок в 3-7 тыс., где отличия в мелкости разбиения принципиальны. Разбиение на число интервалов, большее 20, для нестационарных рядов, как и для стационарных, оказывается актуальным при длинах выборки, больших 30 тыс. Последнее важно для тиковых рядов, характерное число данных в которых за достаточно короткий промежуток времени в одни сутки достигает 100 тыс. (рис. 8).

Заметим, что в рассмотренных примерах полученные величины оптимального числа интервалов для небольших выборок 100-500 данных довольно близки ко всем вышеприведенным оценкам (1-5), включая рекомендации ВНИИМ [10]. Но, в отличие от последних, вместо просто рекомендованного диапазона мы имеем конкретные значения в зависимости от количества данных для каждого из временных рядов. Впрочем, значения суммы (13) для многих реальных временных рядов при одной и той же мелкости разбиения варьируются не особенно значительно, и потому оптимальное разбиение при небольшом числе данных (до нескольких тысяч) оказывается

примерно одинаковым независимо от ряда. В этом смысле рекомендации [10] имеют практическое значение даже в большей степени, чем теоретические оценки (1-5). Однако все эти оценки вместе взятые заметно отклоняются от фактического оптимального разбиения на большом диапазоне изменения длины выборки, хотя наиболее близким к нему оказывается результат Смирнова [3], когда число интервалов растет как корень третьей степени из длины выборки.

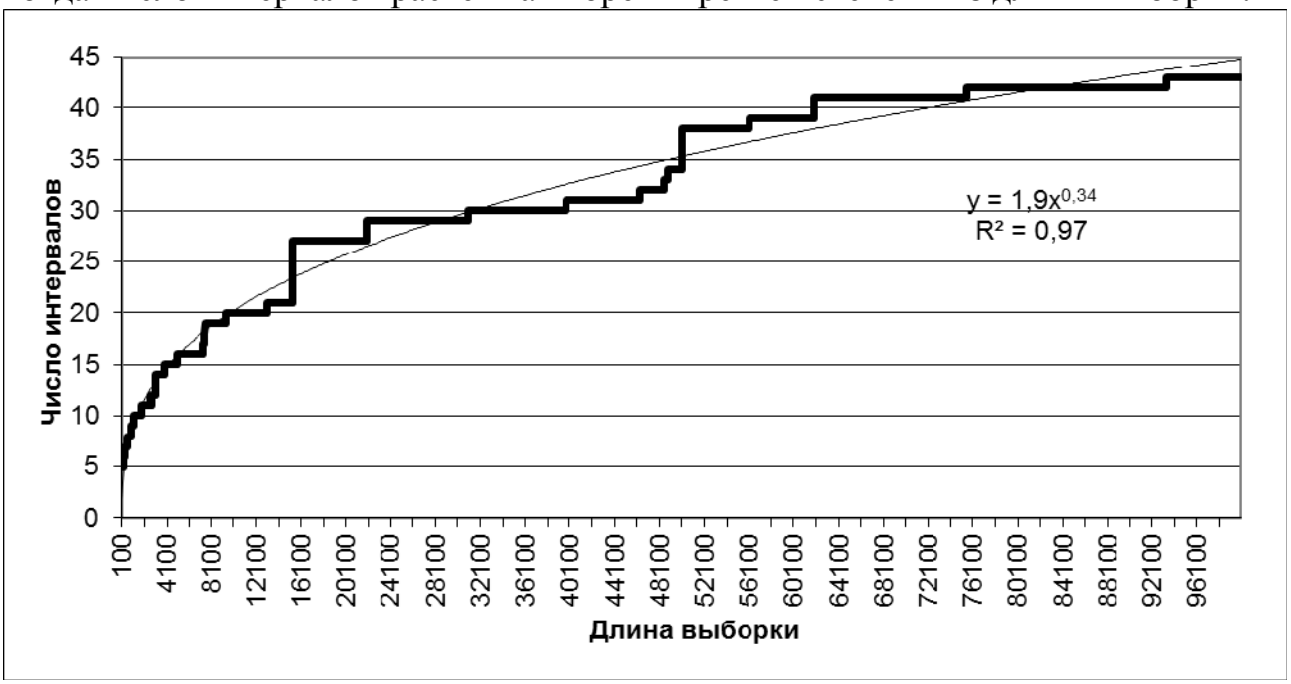

Рис. 8. Оптимальное число интервалов в зависимости от длины выборки для тикового ряда фьючерсов CL и его степенная аппроксимация

Определив оптимальное разбиение гистограммы, представляющей ВПФР в текущий момент времени, введем затем расстояние в пространстве суммируемых функций и определим расстояние между двумя ВПФР, построенными по выборкам длины N, смещенным одна относительно другой на τ шагов:

$$\rho(N,\tau;t) = \left\| f_N(x,t) - f_N(x,t+\tau) \right\| = \sum_{i=1}^{n} \left| f_N(i,t) - f_N(i,t+\tau) \right|. \qquad (27)$$

Величину смещения τ будем варьировать от 1 до N. Возможные значения расстояний между выборками образуют дискретный набор $\tau+1$ чисел: $\rho \in \{0; 2/N; 4/N; ...; 2\tau/N\}$. Обозначим через $G_{N,\tau}(k)$ вероятность того, что расстояние ρ между выборками при сдвиге на τ равно $2k/N$. Величина

$$F_{N,\tau}(\rho) = \sum_{l=0}^{k(\rho)} G_{N,\tau}(l), \quad k(\rho) = [N\rho/2] \qquad (28)$$

представляет эмпирическую вероятность того, что расстояние между распределениями не больше ρ. Наибольшее возможное расстояние между выборками равно двум, когда мера пересечения носителей соответствующих

распределений равна нулю.

Определим согласованный уровень стационарности (далее СУС) $\rho^*(N,\tau)$ так, что вероятность его превышения равна уровню значимости критерия, т.е.

$$F_{N,\tau}\left(\rho^*\right)=1-\rho^*\frac{N}{2\tau}. \qquad (29)$$

Равенство (29) определяет функцию $\rho^*(N,\tau)$, которая обладает тем свойством, что при проведении достаточно большого числа экспериментов по вычислению расстояний между двумя выборочными распределениями длины N, сдвинутыми на окно τ, в доле $\rho^* N/(2\tau)$ случаев будет наблюдаться превышение расстояния, равного ρ^*. Тем самым ρ^* можно трактовать как характерное расстояние между распределениями на уровне значимости, не превосходящем этого расстояния. Пример зависимости СУС $\rho^*(N)\equiv\rho^*(N,N)$ от длины выборки для тикового ряда CL и для часовых цен закрытия индекса RTS при сдвиге на ширину окна выборки приведен на рис. 9.

Для многих нестационарных рядов характерно наличие статистически значимых локальных минимумов согласованного уровня стационарности в зависимости от длины «встык-выборок», тогда как для стационарного ряда наблюдается монотонная зависимость со слабой флуктуацией, исчезающей после перехода к усредненному графику. Аргументы этих локальных минимумов можно трактовать как типовые для данного ряда промежутки времени, на которых происходит смена режима работы наблюдаемой системы.

Так, для ряда CL характерными размерами выборок являются 500, 1100, 2600 данных, а для ряда «часовиков» RTS – 200 и 1100.

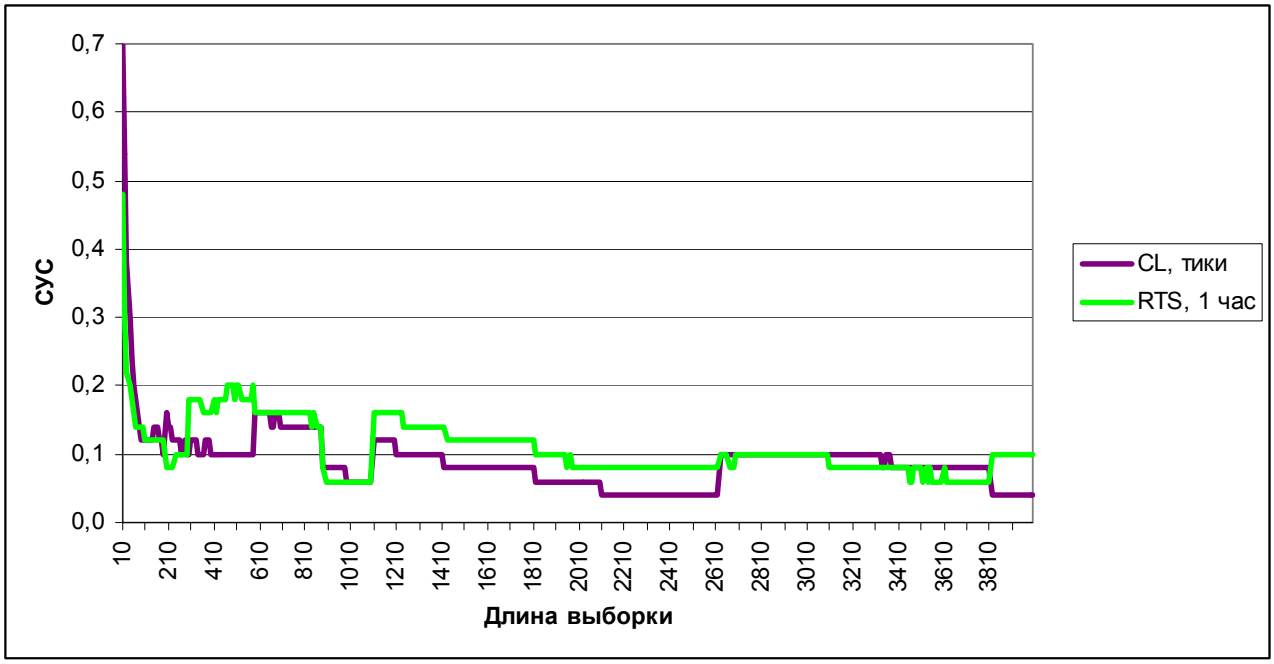

Рис. 9. Зависимость СУС от длины выборки

Стационарный ряд также имеет СУС $\rho_0(N,\tau)$, определяемый генеральным распределением $G^0_{N,\tau}(k)$, которое, в принципе, может быть вычислено по стационарным вероятностям $f_N(i)$ через суммы их степеней. На правом рис. 5 приведено такое стационарное выборочное распределение для нормальной совокупности при $\tau = N$ (независимые встык-выборки). Уровень стационарности для него равен 0,22, что, как и должно быть с точки зрения качественных соображений, в два раза превосходит точность, с которой в этом примере оценивались эмпирические вероятности при оптимальном разбиении.

Определим теперь индекс нестационарности $J(N,\tau)$ ряда, положив его равным отношению доли расстояний, не превосходящих фактический СУС, построенный по имеющимся эмпирическим данным, к уровню шума, за который естественно принять долю расстояний между ВПФР, не превосходящих $2\varepsilon^*\tau/N$ при заданной длине выборки N и сдвиге τ:

$$J(N,\tau) = \frac{F_{N,\tau}\left(\rho^*\right)}{F_{N,\tau}\left(2\varepsilon^*\tau/N\right)}. \qquad (30)$$

Если $J(N,\tau) \le 1$, ряд считается стационарным, а если $J(N,\tau) > 1$, то ряд нестационарный. Полезно протестировать описанную методику на стационарных временных рядах, для которых индекс нестационарности не должен в идеале превосходить единицы. Результаты компьютерного эксперимента по вычислению индекса $J(N,N)$ для стационарной нормальной совокупности в сравнении с нестационарным рядом фьючерсов CL приведены на рис. 10.

Рис. 10. Индекс нестационарности $J(N,N)$

Как и должно быть для любого стационарного ряда, индикатор (28) слабо флуктуирует вблизи единицы.

Существенно, что стационарные и нестационарные ряды при малых выборках (до нескольких тысяч) неотличимы в силу низкой точности оценки эмпирических вероятностей, и лишь с увеличением длины выборки, когда $\rho^* > 2\varepsilon^*$, такое разделение достоверно. Так, нестационарный тиковый ряд CL начинает отделяться от стационарного (нормальное распределение) по индикатору (30) с длин выборок порядка 8 тыс., но значимо нестационарность начинает проявляться после 16 тыс. данных. Рассмотрим фрагмент временного ряда, образованного приростами дистинктивных тиков ряда CL (рис. 11).

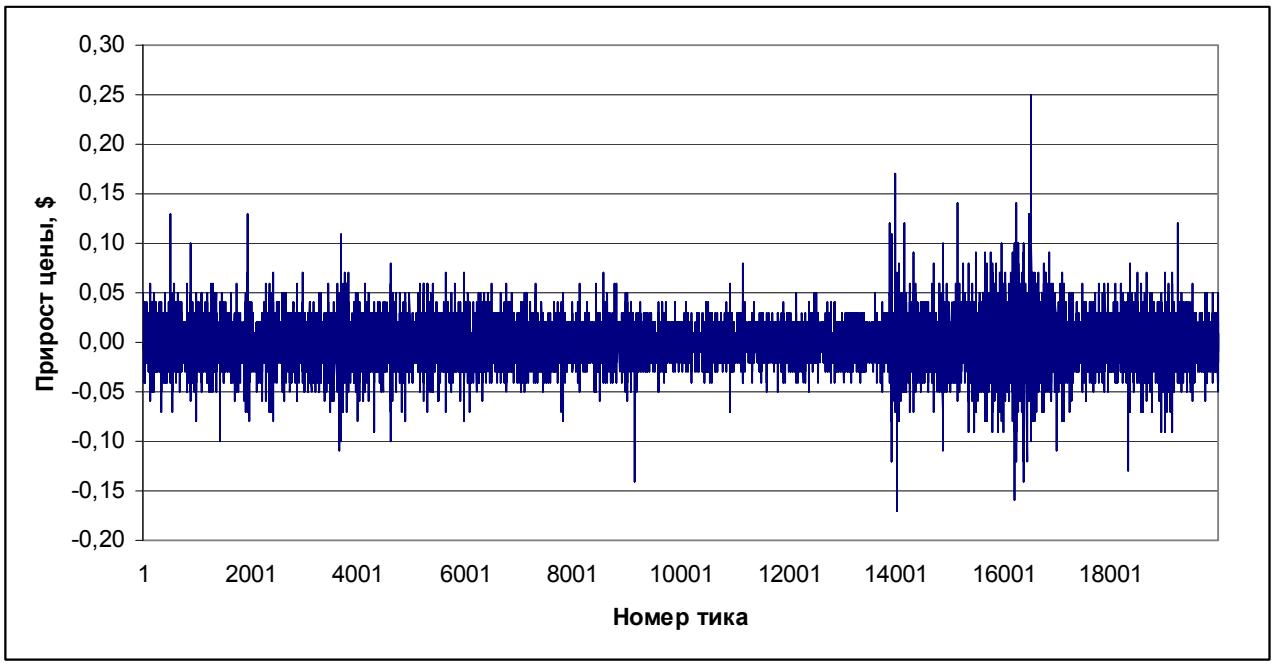

Рис. 11. Временной ряд первых разностей по тикам CL

Распределение ряда первых разностей показано на гистограмме рис. 12. Оптимальное разбиение гистограммы для выборки длиной 20 тыс. состоит из 27 классовых интервалов, но от нуля отличны лишь 16. Оптимальность мелкости состоит здесь в том, что вероятности в центральной части оцениваются с наилучшей достоверностью. Как видно из рис. 11 и рис. 8, если длина выборки превосходит 100 тыс., достигается естественная мелкость разбиения с шагом в 1 цент, и далее кластеризация уже не проводится.

Для того чтобы достоверно отмечать изменения в выборочном распределении приростов, необходимо превышение отклонения между плотностями, сдвинутыми по времени на промежуток τ, величины точности, характерной для данного распределения и подходящего разбиения. Статистически лишь тиковый ряд имеет достаточно большое число событий за короткий промежуток времени, так что именно такие ряды могут быть исследованы методом функций распределения. Ряды с минутными и более

отсчетами требуют иных подходов, так как оценка эмпирических вероятностей слишком сильно флуктуирует на малых длинах выборки, а большие длины не актуальны для принятия решения в течение, скажем, суток.

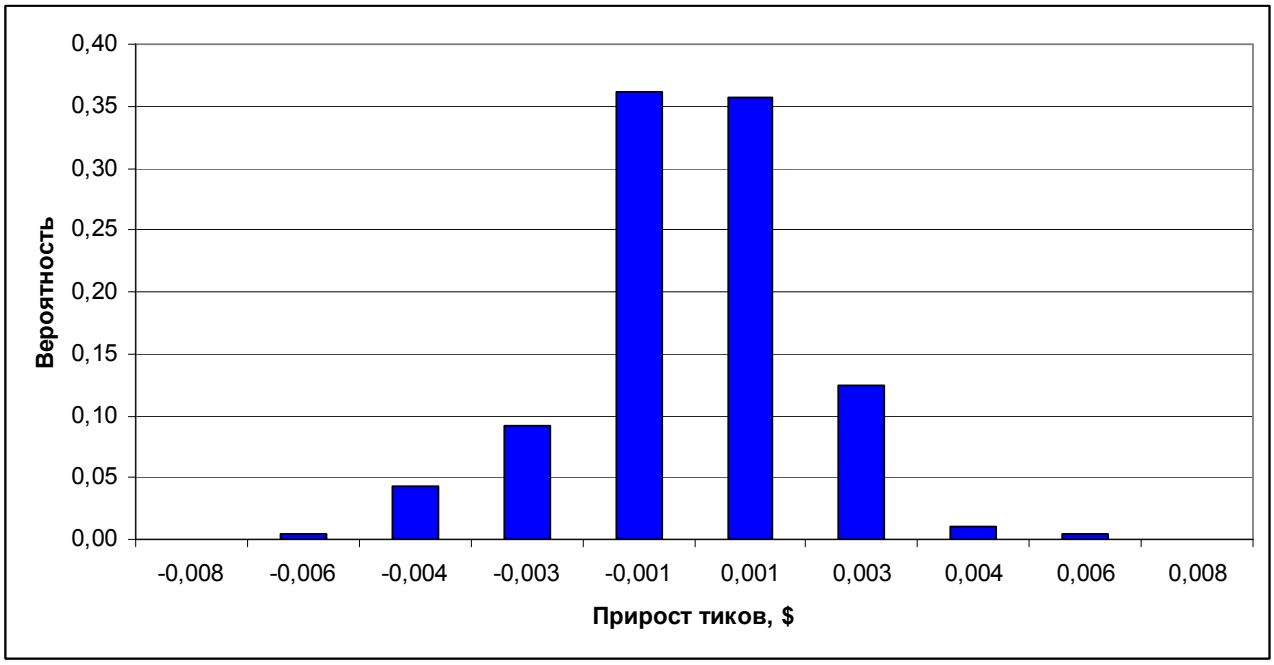

Рис. 12. ВПФР первых разностей, длина выборки 20 тыс.

Типичное свойство распределения тиковых приростов – длинные тонкие «хвосты», не заметные на графике, в которых сосредоточена малая часть нормировки: вне четырех центральных ячеек находится 0,07 событий, а вне восьми ячеек, видимых на гистограмме – всего 0,01. Тем не менее, «хвосты» эти принято называть толстыми, так как они не спадают экспоненциально с увеличением отклонения, а имеют примерно постоянную или степенным образом спадающую толщину, обрывающуюся нулем за пределами носителя распределения.

Рис. 13 иллюстрирует изменение расстояния (27) между встык-выборками во времени, а на рис. 14 показан типичный вид поверхности функционала расстояния (27) в некоторый момент времени как функции длины выборки и величины сдвига. Шаг по времени, а также длинам выборки и сдвига равен 100.

В целом расстояние между ВПФР с увеличением длины выборки при фиксированном окне сдвига уменьшается, а с увеличением окна сдвига при фиксированной длине выборки растет. Однако исходный временной ряд может обладать квазипериодической разладкой, проявляющейся не только локально по времени, но и в среднем на уровне СУС в немонотонном его поведении. На рис. 13 отчетливо виден минимум расстояния на длине выборки около 20 тыс. (200 точек с шагом 100) при любом сдвиге, а также локальный максимум на выборках, больших 20 тыс., при сдвиге порядка 5000. Последнее

свидетельствует о возможности идентифицировать разладку на меньших, чем встык-выборки, длинах.

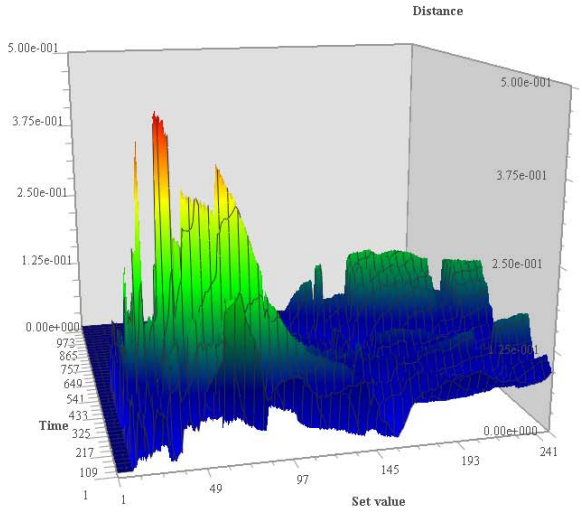

Рис. 13. Расстояние между встык-выборками в зависимости от длины выборки в различные моменты времени

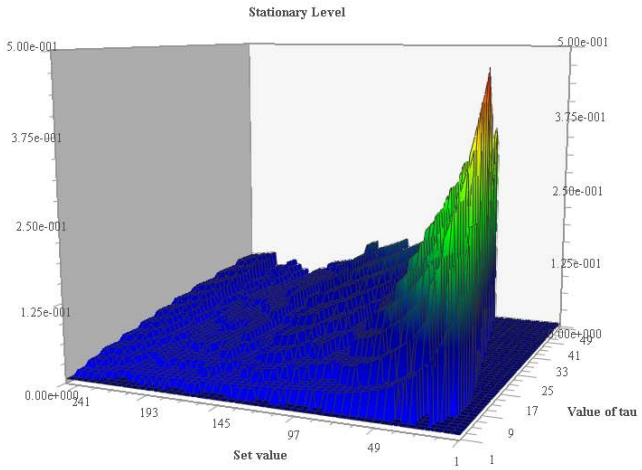

Рис. 14. СУС внахлест-выборок в зависимости от длин выборки и сдвига

Таким образом, в работе предложен метод согласованного разбиения гистограммы для анализа распределений нестационарных временных рядов и введен индикатор нестационарности. В дальнейшем предполагается изучить индикаторы разладки типа статистической добротности, введенной в [16], как функции длины выборки и величины сдвига при формировании «внахлест-выборок».

Литература

1. Шторм Р. Теория вероятностей. Математическая статистика. Статистический контроль качества. – М.: Мир, 1970. 368 с.
2. Sturges H.A. The choice of a class interval // JASA. 1926. V.21. P. 65-66.
3. Mann H.B., Wald A. On the choice of number of intervals in the application of the chi-square test // AMS. 1942. V. 18. P. 50-54.
4. Смирнов Н.В. О построении доверительной области для плотности распределения случайной величины // Доклады АН СССР. 1950. Т. 74. № 2. С. 189-192.
5. Scott D.W. On optimal and data-based histograms // Biometrika. 1979. V.66. P. 605-610.
6. Лившиц М.Е., Иванов-Муромский К.А., Заславский С.Я., Войтинский Е.Я., Лернер В.А., Ромм Б.И. Численные методы анализа случайных процессов. – М.: Наука, 1976. 128 с.
7. Булашев С.В. Статистика для трейдеров. – М.: Компания Спутник+. – 2003. 245 с.
8. Новицкий П.В., Зограф И.А. Оценка погрешностей результатов измерений. – Л.: Энергоатомиздат, 1991. 304 с.
9. Taylor C. Akaike's information criterion and the histogram // Biometrika. 1987. V.74. P. 636-639.
10. Кобзарь А.И. Прикладная математическая статистика. – М.: Физматлит, 2006. 816 с.
11. ГОСТ Р 50.1.033-2001 Прикладная статистика / Постановление Госстандарта России от 14 декабря 2001 г. №525-ст.
12. Бурдун Г.Д., Марков Б.Н. Основы метрологии. – М.: Изд-во стандартов, 1985. 120 с.
13. Королюк В.С., Портенко Н.И., Скороход А.В., Турбин А.Ф. Справочник по теории вероятностей и математической статистике. – М.: Наука, 1985. – 640 с.
14. Абрамовиц М., Стиган И.М. Справочник по специальным функциям. – М.: Наука, 1979.
15. Аникин В.М., Голубенцев А.Ф. Аналитические модели детерминированного хаоса. – М.: Физматлит, 2007. 326 с.
16. Орлов Ю.Н., Шагов Д.О. Индикативные статистики для нестационарных временных рядов // Препринты ИПМ им. М.В. Келдыша РАН. 2011. № 53. 20 с. URL: http://library.keldysh.ru/preprint.asp?id=2011-53