Юрий Николаевич Орлов
Институт прикладной математики им. М.В. Келдыша РАН
Кафедра высшей математики МФТИ

Практические задачи анализа нестационарных временных рядов

Некоторые сведения из теории стационарных временных рядов

Сходимость выборочной ФР

Т1. (Гливенко) Эмпирическое выборочное распределение $F_T(x)$ случайной стационарной величины равномерно по x сходится по вероятности к распределению генеральной совокупности $F(x)$:

$$P\left\{\lim_{T\to\infty}\sup_x\left|F_T(x)-F(x)\right|=0\right\}=1$$

Т2. (Колмогоров) Пусть генеральное распределение $F(x)$ непрерывно. Тогда статистика

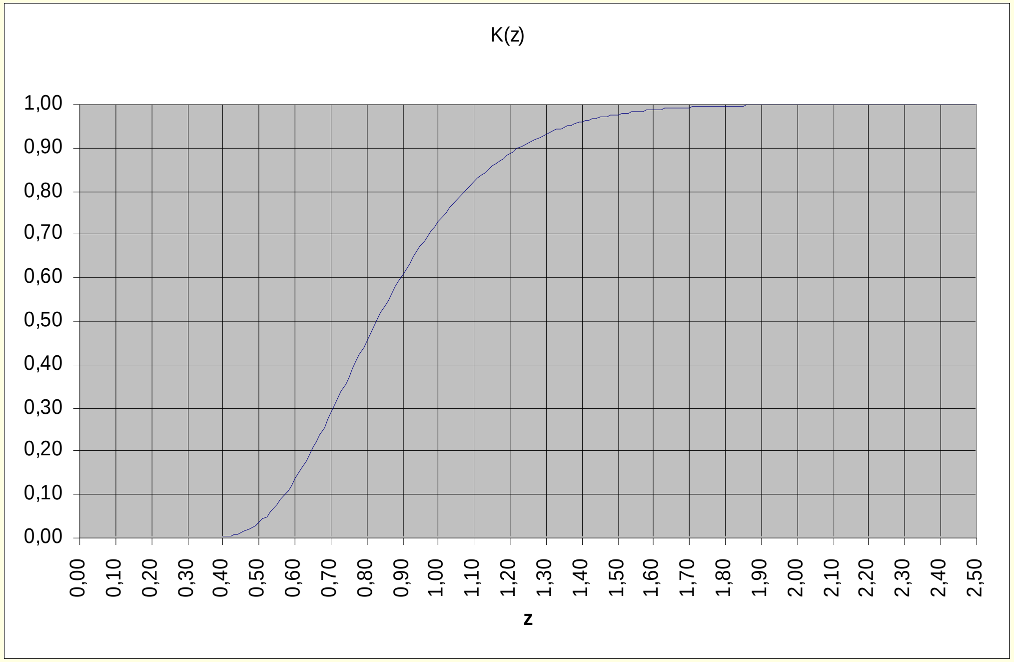
$$\sqrt{T}\sup_x\left|F_T(x)-F(x)\right|$$

супремума модуля отклонения ВФР от генеральной ФР сходится по вероятности при $T\to\infty$ к функции Колмогорова:

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

Критерий Колмогорова-Смирнова

Пусть случайные величины имеют стационарное распределение и являются независимыми. Тогда вероятность того, что две выборки объема n различаются между собой в норме C менее, чем на ε, равна $K\left(\varepsilon\sqrt{n/2}\right)$



$$S_n = \sup_x \left| F_{1,n}(x) - F_{2,n}(x) \right|$$

$$\lim_{n\to\infty} P\left\{0 < \sqrt{\frac{n}{2}} S_n < z\right\} = K(z) = \sum_{k=-\infty}^{\infty} (-1)^k \exp\left(-2k^2 z^2\right)$$

ЦПТ для сумм

- Т3. (Леви-Линдеберг) Пусть $\{\xi_n\}$ есть последовательность независимых одинаково распределенных случайных величин, имеющих конечные матожидание μ и дисперсию σ^2 .

  Тогда распределение величин

$$z_n = \frac{1}{\sigma\sqrt{n}}\left(\sum_{k=1}^{n}\xi_k - n\mu\right)$$

является асимптотически нормальным с нулевым средним и единичной дисперсией.

ЦПТ для выборочных моментов

- Т4. (Гофдинг) Пусть $\{x_1,\ldots,x_n\}$ есть выборка из распределения $F(x)$, которое имеет конечные моменты

$$\mu_r = \int x^r dF(x).$$

Тогда выборочные моменты

$$m_r(n) = \frac{1}{n}\sum_{k=1}^{n}(x_k)^r$$

являются несмещенными состоятельными оценками генеральных моментов. Их распределения асимптотически нормальны с параметрами

$$M_r = \mu_r, \quad D_r = \frac{\mu_{2r} - \mu_r^2}{n}.$$

Анализ стационарного процесса

- Т5. (Разложение Вольда) Всякий стационарный случайный процесс может быть единственным образом представлен в виде некоррелированной суммы детерминированного процесса и преобразования фильтрации некоторого процесса с независимыми приращениями (белого шума)

$$\eta(t) = m(t) + \int_{-\infty}^{\infty} h(t-s)\varepsilon(s)ds$$

или, для дискретных процессов (временных рядов)

$$\eta(t) = m(t) + \sum_{k=0}^{\infty} h_k \varepsilon(t-k), \quad \sum_{k=0}^{\infty} |h_k| < \infty.$$

Матожидание стационарного ряда

- Эргодическая теорема (Биргхоф – Хинчин) Пусть случайный процесс стационарный, его генеральное распределение $F(x)$ имеет конечный первый момент μ и дисперсию, а автокорреляционная функция такова, что

$$B(\tau) = \langle (x(t) - \mu)(x(t+\tau) - \mu) \rangle \to 0, \quad \tau \to \infty.$$

Тогда почти наверное

$$\lim_{T \to \infty} \frac{1}{T} \int_0^T x(t)dt = \int x dF(x)$$

Трудности анализа нестационарного ряда

- В последовательные моменты времени наблюдаются значения **разных случайных процессов**, поэтому отсутствует понятие **генеральной совокупности**, и связанные с ней предельные теоремы и **статистические критерии**, строго говоря, **не применимы**

- **Эмпирические оценки** вероятностей попадания значений в заданный интервал с увеличением объема выборки **не сходятся** к генеральной совокупности ни в слабом смысле, ни по норме

- Если все же требуется оценить эмпирическую вероятность, то по **выборке какого объема** это следует делать, и с **какой точностью** такая оценка будет выполняться на заданном **горизонте** прогноза?

Поиск оптимального объема выборки для анализа нестационарного временного ряда

Причины ошибки прогнозирования

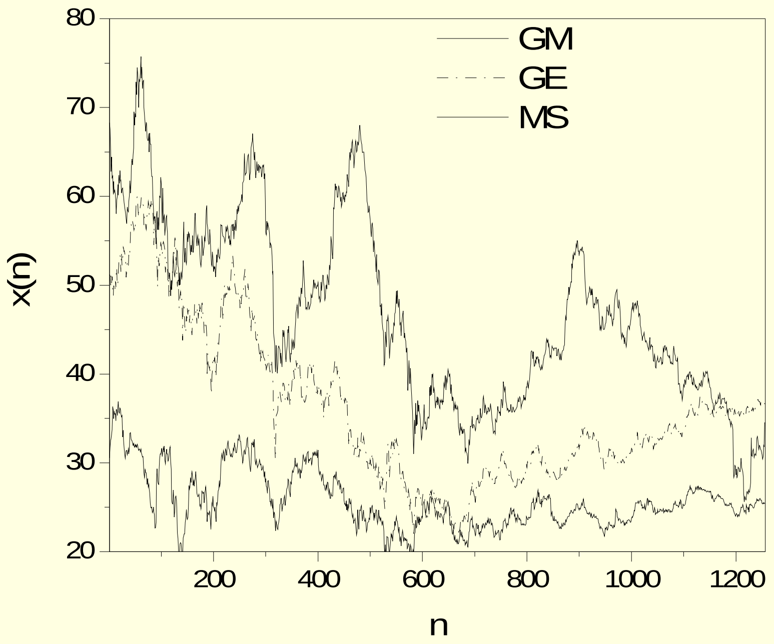
$$\left\langle \Psi\left[x(t)\right]\right\rangle_{\Delta} = \frac{1}{T}\int_{t_0}^{t_0+T} \Psi\left[x(t)\right] dt$$

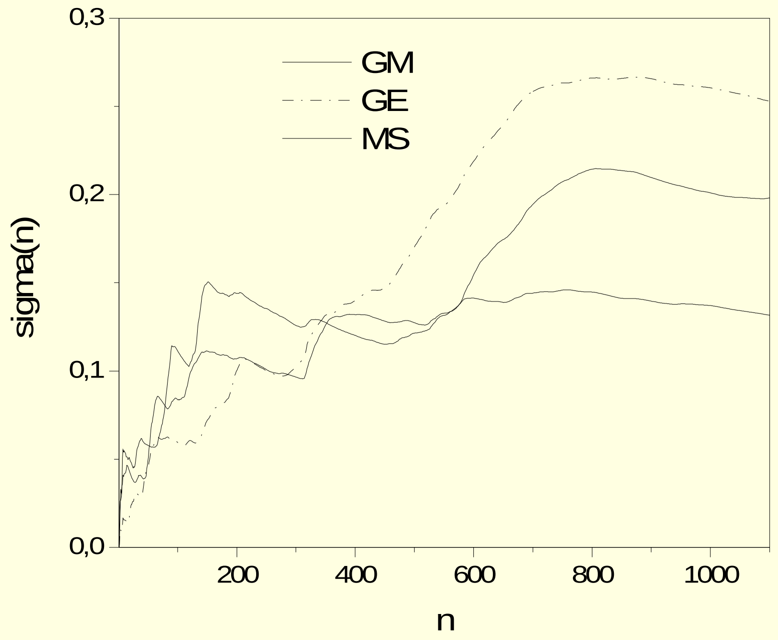
Конечность промежутка Т – нерепрезентативность выборки; для уменьшения ошибки следует увеличивать Т.

Нестационарность процесса, т.е. изменение статистических свойств на промежутке Т; для уменьшения ошибки следует уменьшать Т.

Следовательно, существует оптимальный объем выборки, на котором суммарная ошибка прогноза минимальна.

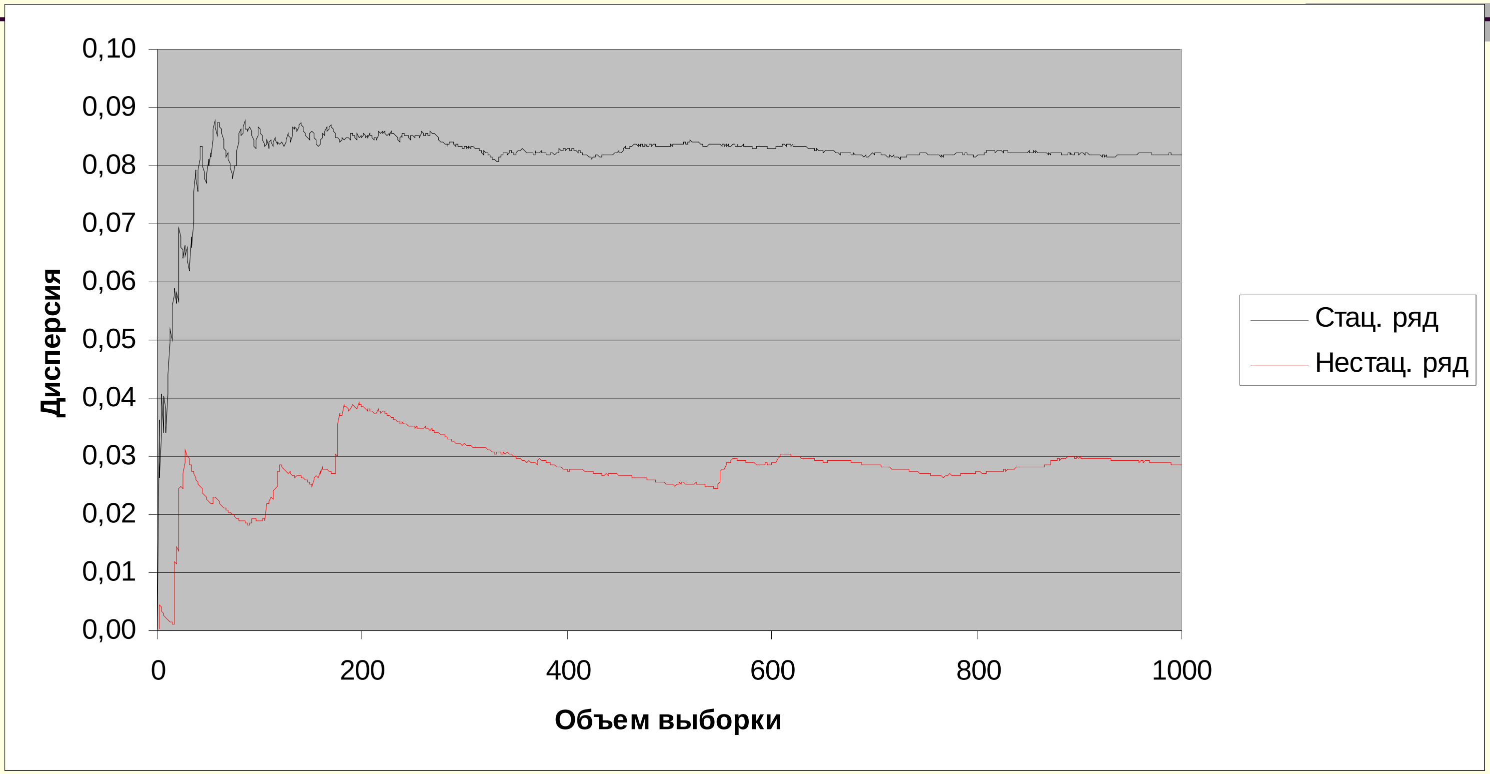
Характерный пример





Типичное поведение относительной дисперсии как функции объема выборки нестационарного ряда: при увеличении n нет стабилизации к значению по генеральной совокупности, и даже при достаточно больших объемах выборки наблюдается рост дисперсии

Дисперсия: stat vs non-stat



Для равномерно ограниченного ряда наблюдается кажущаяся стабилизация дисперсии, но ее нестационарные колебания сопоставимы по амплитуде с самой дисперсией, а не с шумом

Кажущаяся стабилизация дисперсии

Выборочная дисперсия, построенная по объему 1000 данных,

за 1000 шагов успевает измениться более чем на 30%, т.е. на самом деле она **не стабилизируется** к генеральному значению

Оценка сверху ошибки прогноза

Ошибка прогнозирования ВПФР в силу нестационарности:

$$\varepsilon(t) = \int_0^1 \left| \hat{f}(x,t) - f(x,t) \right| dx$$

Ошибка прогнозирования среднего значения временного ряда:

$$\left| \hat{\bar{x}} - \bar{x} \right| = \left| \int_0^1 x\hat{f}(x)dx - \int_0^1 xf(x)dx \right| \le \int_0^1 \left| x \cdot (\hat{f} - f) \right| dx = \int_0^1 |x| \left| \hat{f} - f \right| dx \le \varepsilon$$

Ошибка прогнозирования собственно временного ряда:

$$\delta^2 = \int_0^1 \left(x - \hat{\bar{x}} \right)^2 f(x,t)dx = \int_0^1 \left(x - \hat{\bar{x}} + \bar{x} - \bar{x} \right)^2 f(x,t)dx = \sigma^2 + (\bar{x} - \hat{\bar{x}})^2 \le \sigma^2 + \varepsilon^2$$

Горизонтный ряд

Расстояние между двумя ВПФР:

$$\rho\left(f_1, f_2\right) = \left\| f_{T_1}(x,t_1) - f_{T_2}(x,t_2) \right\| = \int \left| f_{T_1}(x,t_1) - f_{T_2}(x,t_2) \right| dx$$

Функционал близости между двумя ВПФР:

$$V(T,\tau;t) \equiv \rho\left(f_T(x,t+\tau),\ f_T(x,t)\right) = \int \left| f_T(x,t+\tau) - f_T(x,t) \right| dx$$

Горизонтным рядом $h(t)$ называется такой минимальный объем выборки $h\left(t,\tau;\varepsilon\right)$, что при всех $T \geq h(t,\tau;\varepsilon)$ выполнено условие $V(T,\tau;t) \leq \varepsilon$

Равномерная по времени оценка на горизонтный ряд:

$$0 \leq V(T,\tau;t) \leq \min\left(2\tau / T;\ 2\right)$$

Горизонтный ряд и оптимальный объем выборки

- Если ВПФР ряда x(t) является ε-стационарной, то ВПФР горизонтного ряда $h(t,\tau;\varepsilon)$ также ε-стационарна.
- Относительная дисперсия распределения горизонтного ряда есть величина порядка $O(\varepsilon)$.
- Минимально достаточным объемом выборки для прогноза ВПФР на горизонт τ с точностью ε называется величина

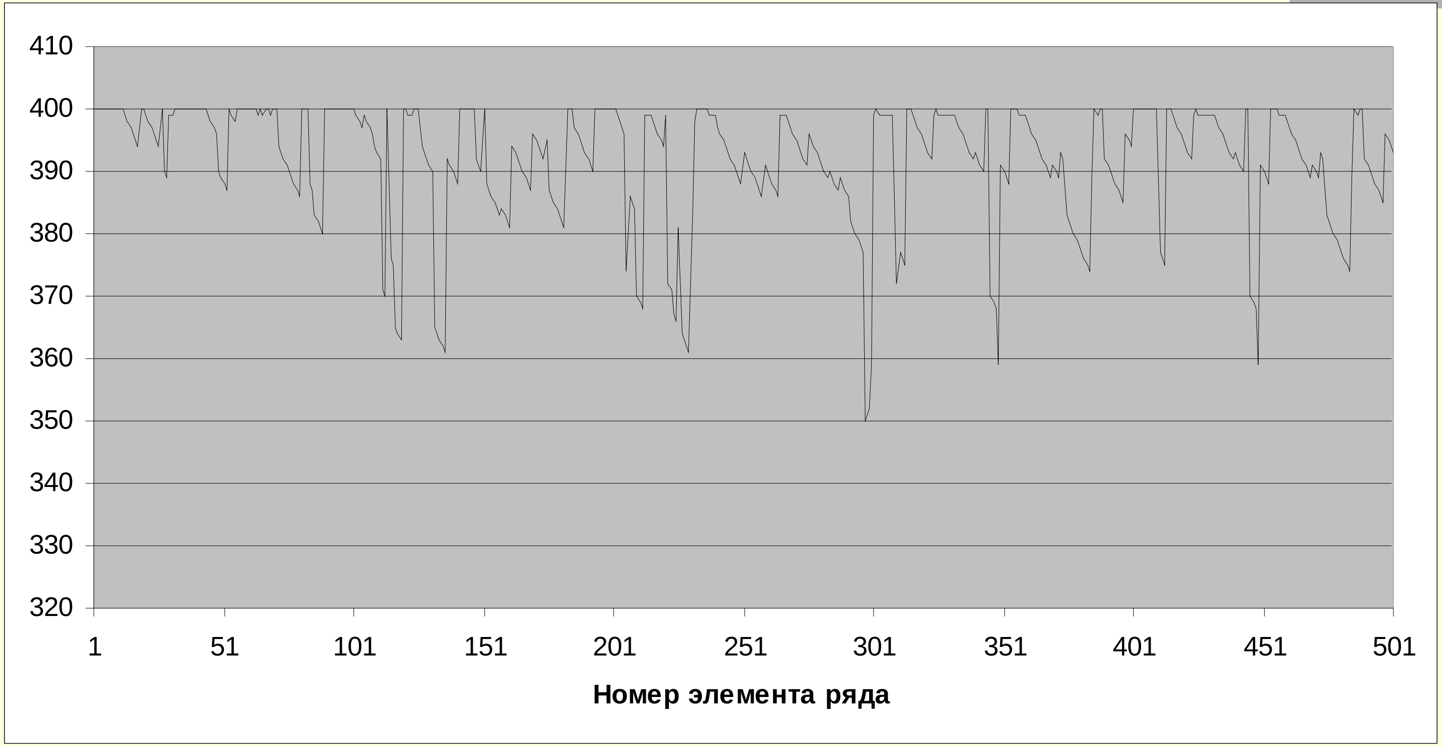
$$h(\tau,\varepsilon) = \max\{ h(t,\tau;\varepsilon) \}.$$

- Оптимальным объемом T для прогнозирования ряда называется величина, доставляющая минимум функционала оценки ошибки прогноза с учетом нестационарности ВПФР:

$$\sigma(h,t)^2 + V^2(h,\tau;t) \to \min$$

Горизонтный ряд стац. процесса

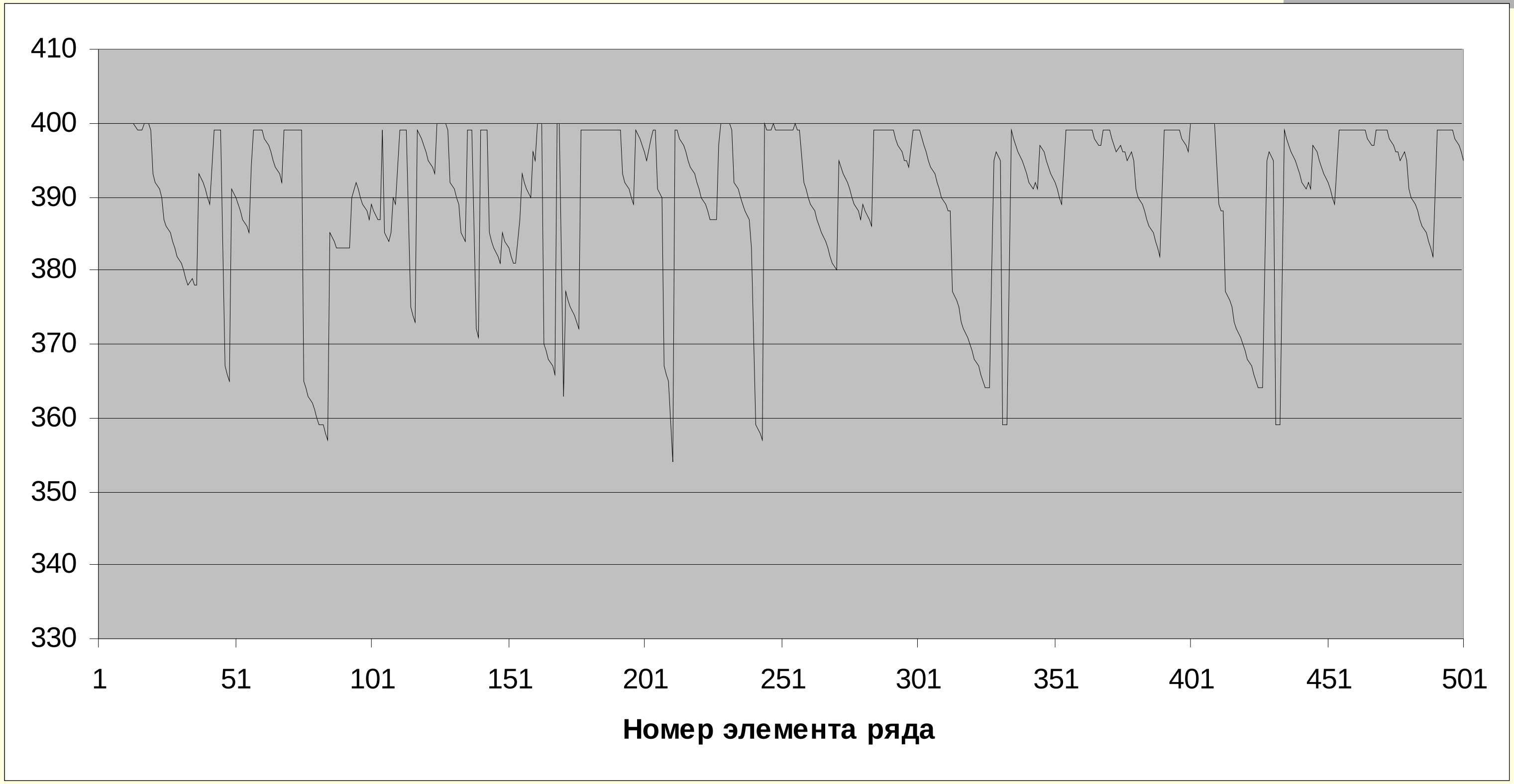
$$h(t,\tau;\varepsilon)=\min\left\{T:V(T,\tau;t)<\varepsilon\right\}<2\tau/\varepsilon$$



Горизонтный ряд равномерного распределения при сдвиге на 10 шагов и точности 0,05

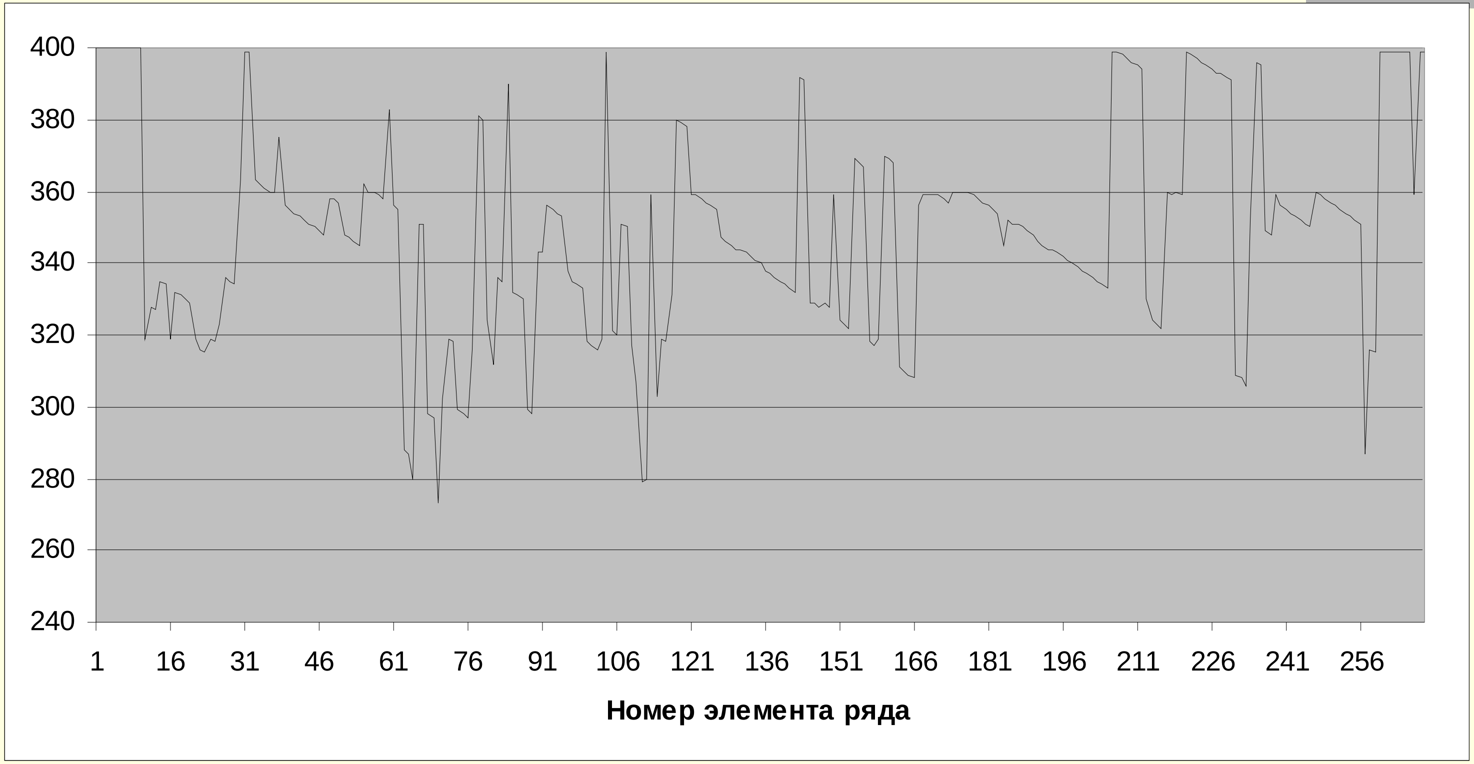
Горизонтный ряд для ХДС

$$x_{n+1} = 1 - 2x_n^2, \quad x_0 \in [-1; 1]$$



Горизонтный ряд логистической динамической системы
при сдвиге на 10 шагов и точности 0,05

Горизонтный ряд нестац. процесса



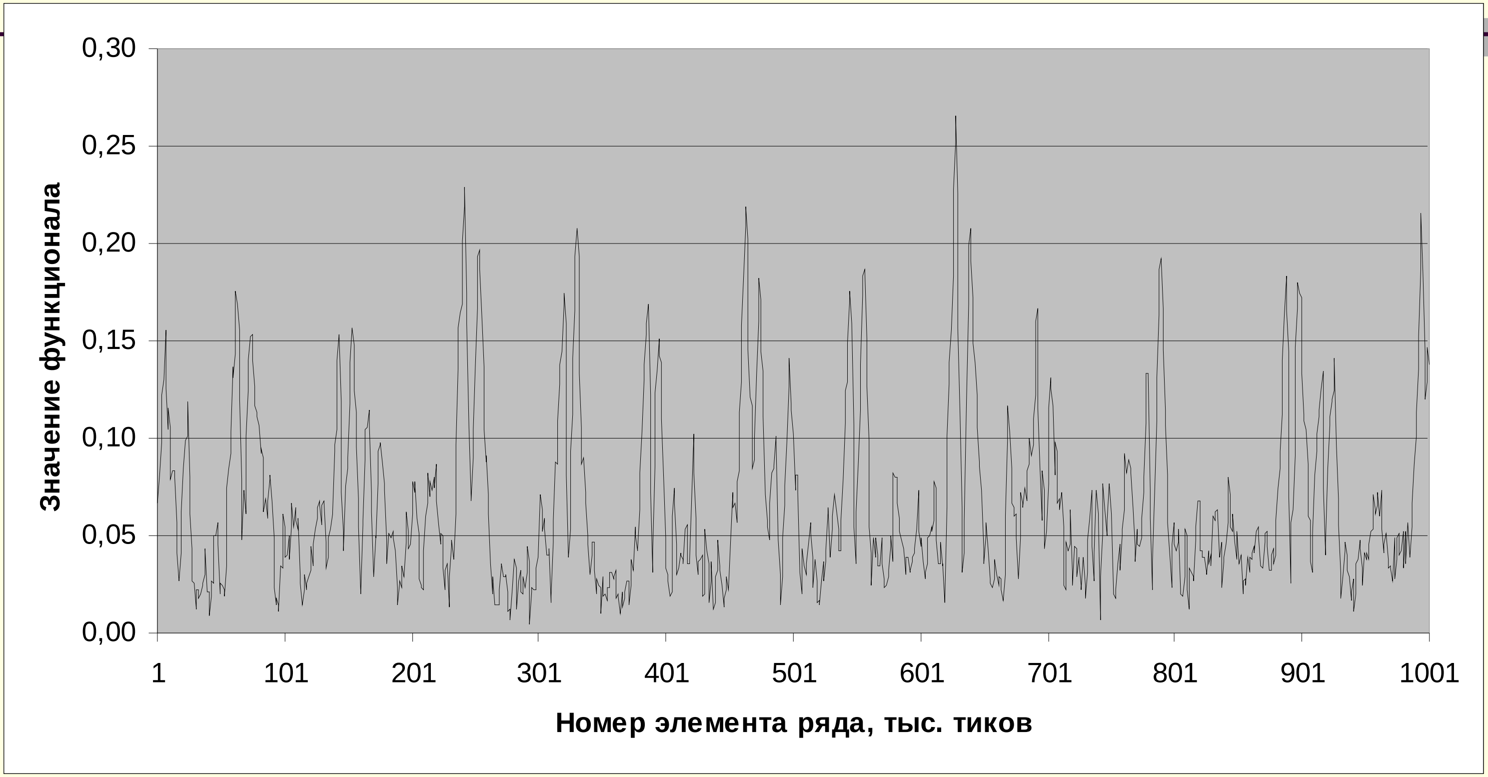
Горизонтный ряд для курса евро/доллар при сдвиге на 10 шагов и точности 0,05

Распределения горизонтных рядов

БШ: полочка перед максимумом в последней точке
ХДС: максимум в предпоследней точке, утолщенный хвост
CLO: максимум в промежуточной точке, немонотонность

Нахождение момента разладки и определение уровня нестационарности ряда

Ряд значений разности двух ВПФР встык в норме L1



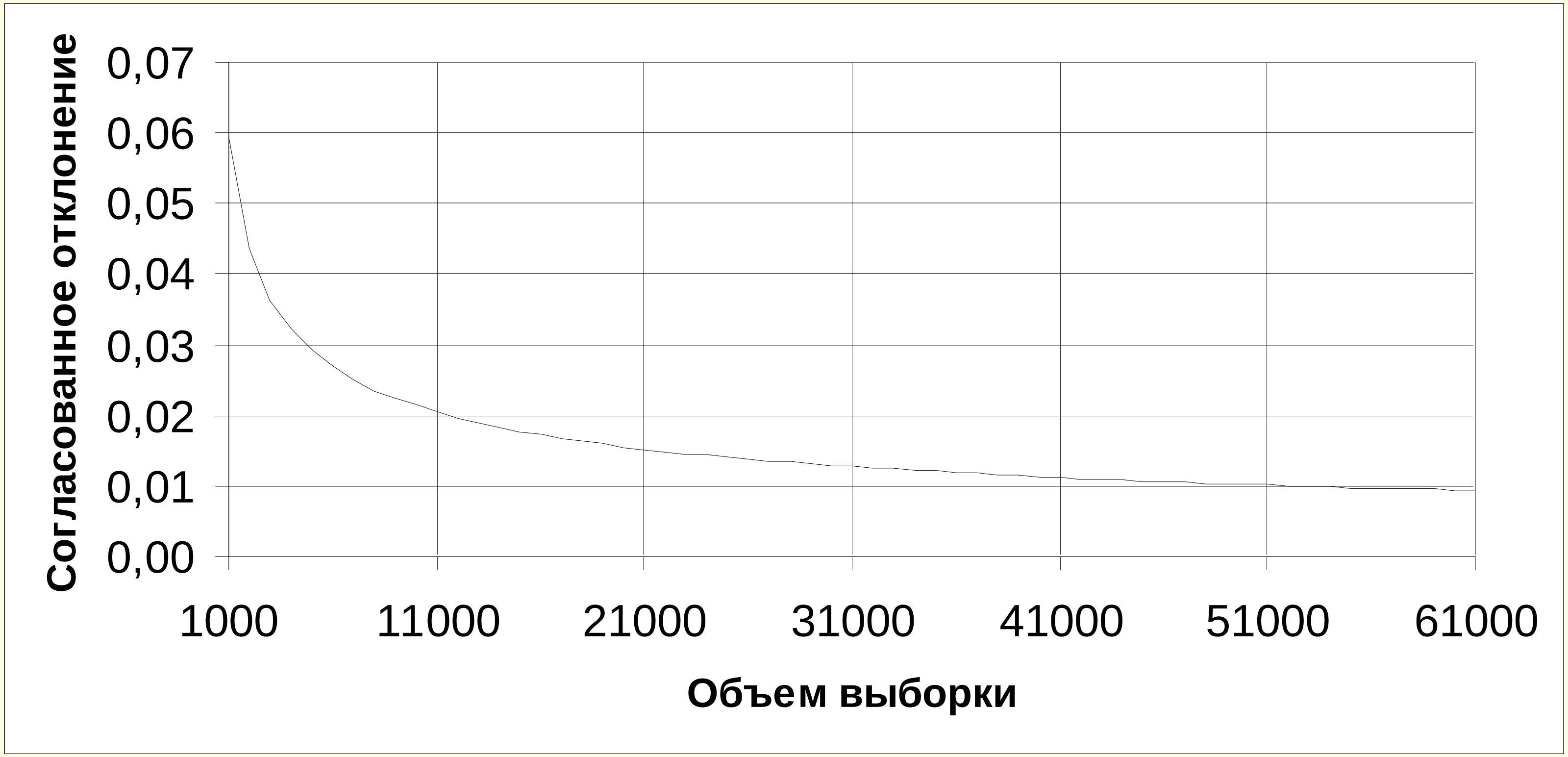
Скользящий ряд расстояний между выборками объемом 1000 тиков.

Как определить, какое отклонение достаточно велико?

Каково типовое отклонение в стационарном случае?

Каким объемом выборки сканировать тиковый ряд?

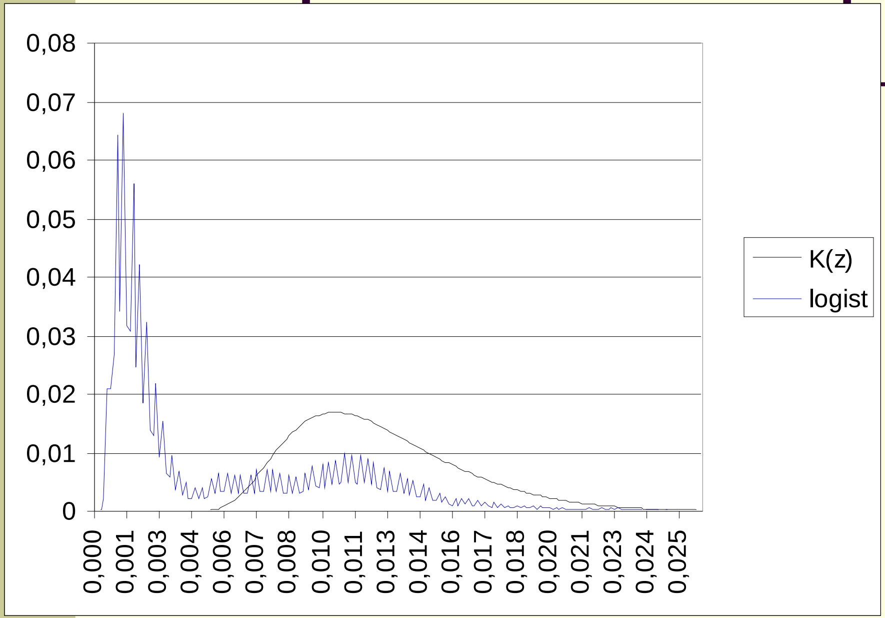
Доля стационарно объясняемых отклонений в зависимости от объема n

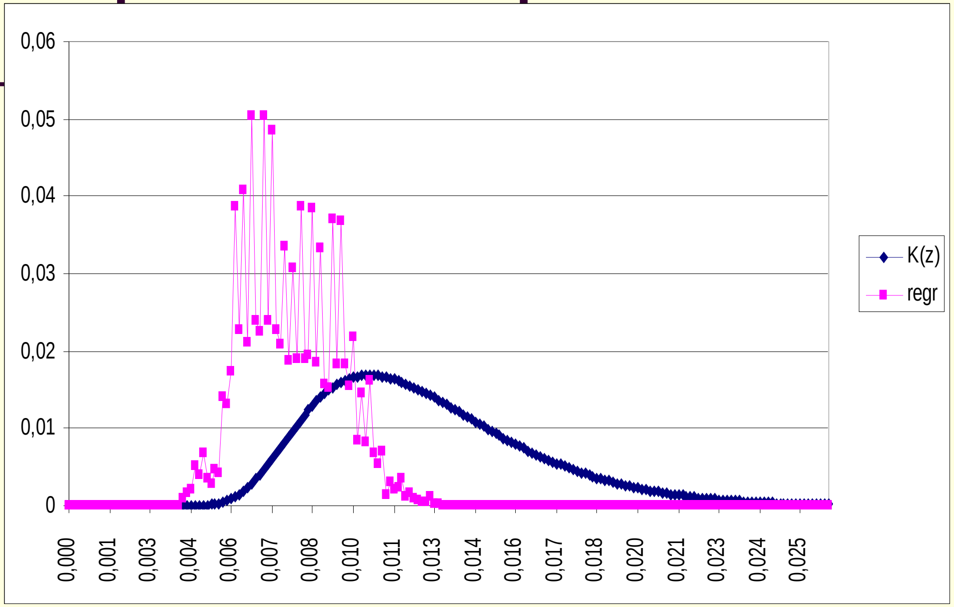


Уровень стационарности: доля расстояний, превышающих этот уровень, равна уровню значимости критерия

$$1 - \frac{V_0}{2} = K\left(\sqrt{\frac{n}{2}} V_0\right)$$

Зависимые данные – сдвиг влево по сравнению с критерием Смирнова

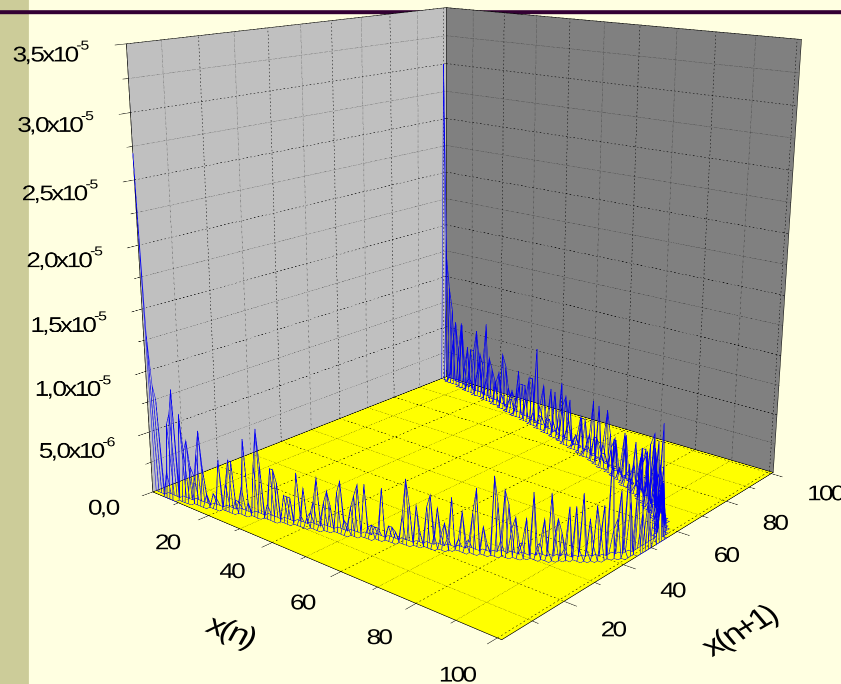


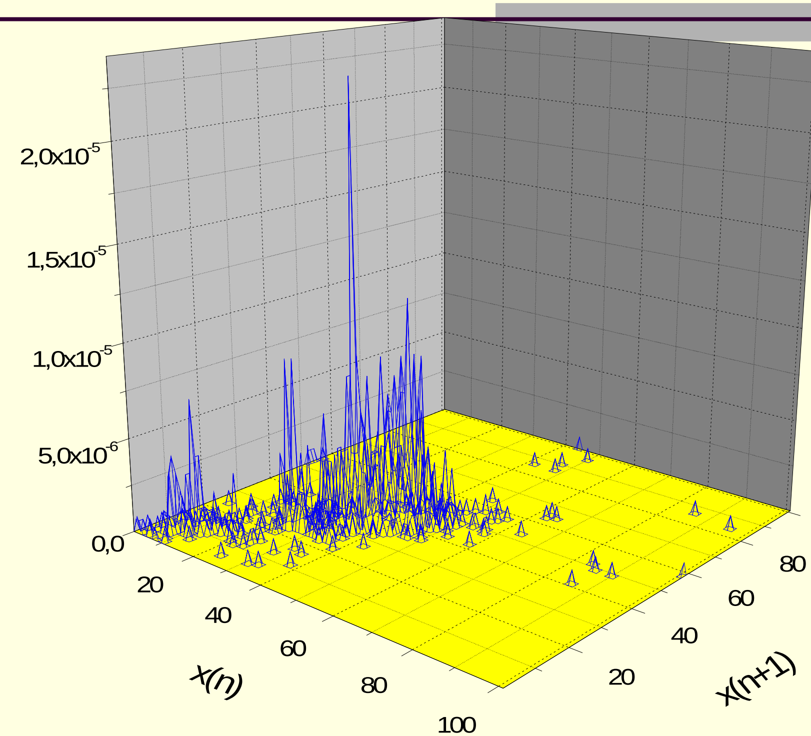


Распределение расстояний между выборками в 10 тыс. данных для логистической ДС $x_{n+1} = 1 - 2x_n^2$ по 1 млн экспериментов (слева) и для линейно зависимых величин (справа).

Классический критерий дает значение уровня стационарности 0,021, а на практике этот уровень оказался равен 0,014 для ХДС и 0,011 для регрессионной зависимости между выборками.

Анализ носителя совместного распределения для определения зависимости величин





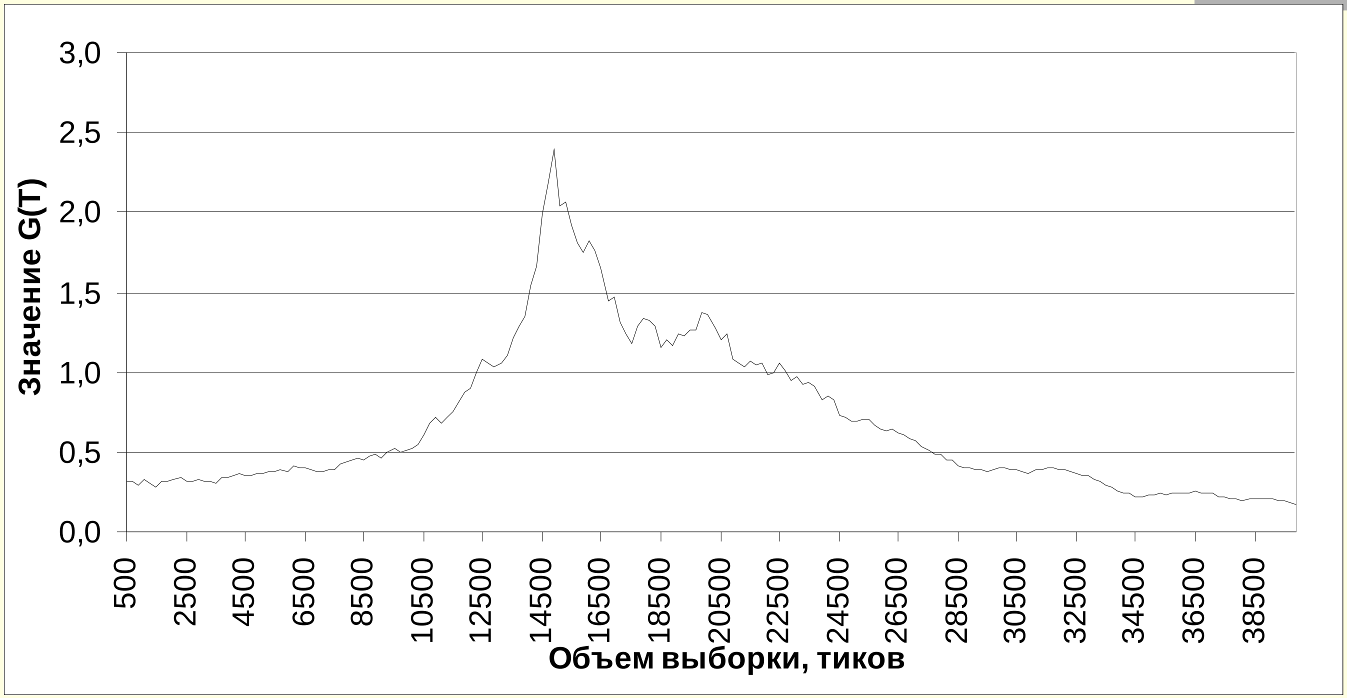
Приближенная функциональная зависимость между элементами ряда (функциональная корреляция) может быть найдена из носителя распределения. Слева – логистическая ХДС $x_{n+1} = 1 - 2x_n^2$, справа – ряд остатков от средненедельного часового профиля цен на электроэнергию

Для нестационарных выборок – сдвиг уровня стационарности вправо

Уровень нестационарности: доля значений, превышающих уровень, равна самому этому уровню

$$\int_0^{V_S(T)} g_T(V)\,dV = 1 - \frac{V_S(T)}{2}$$

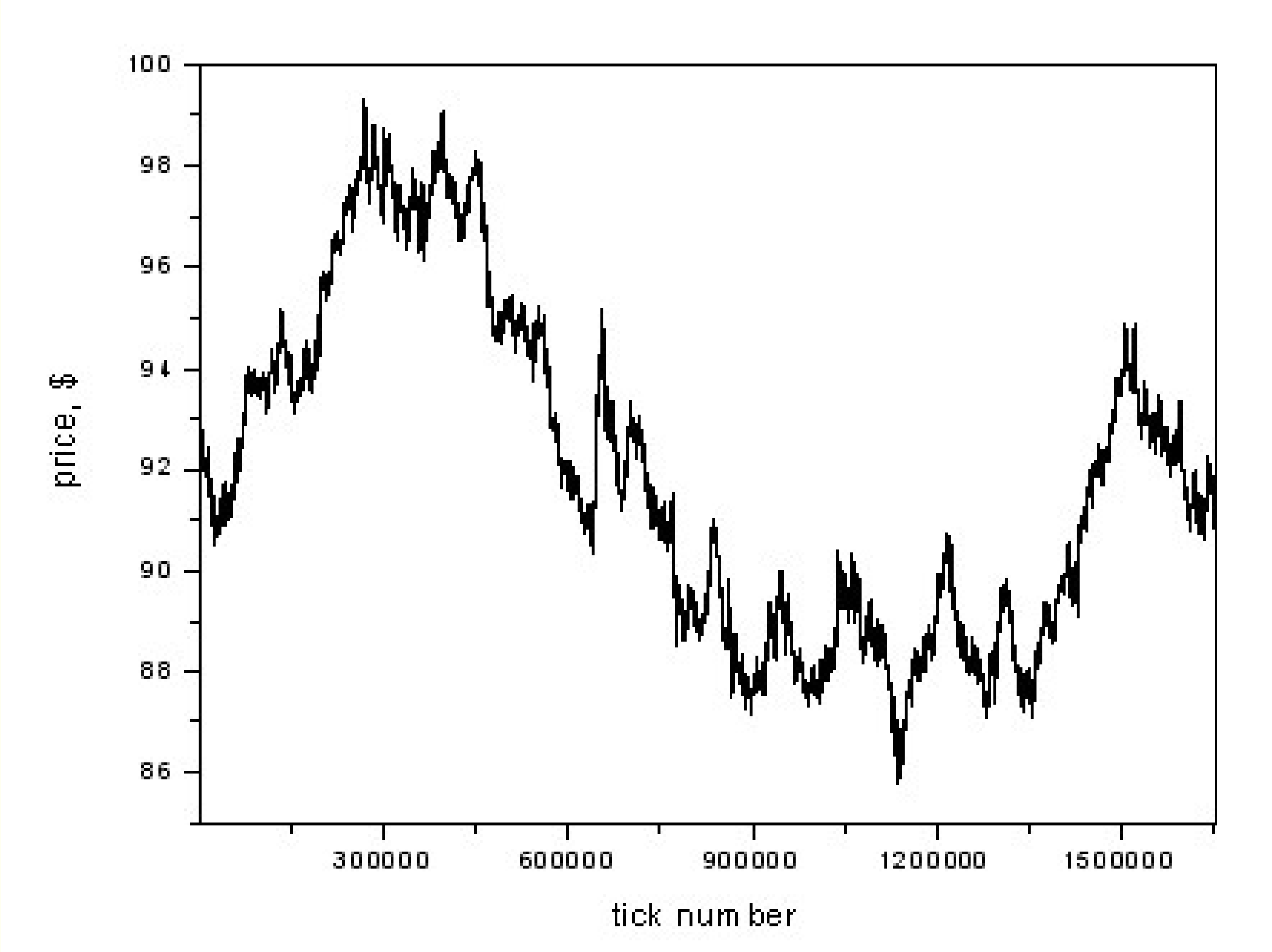
Статистическая добротность ряда



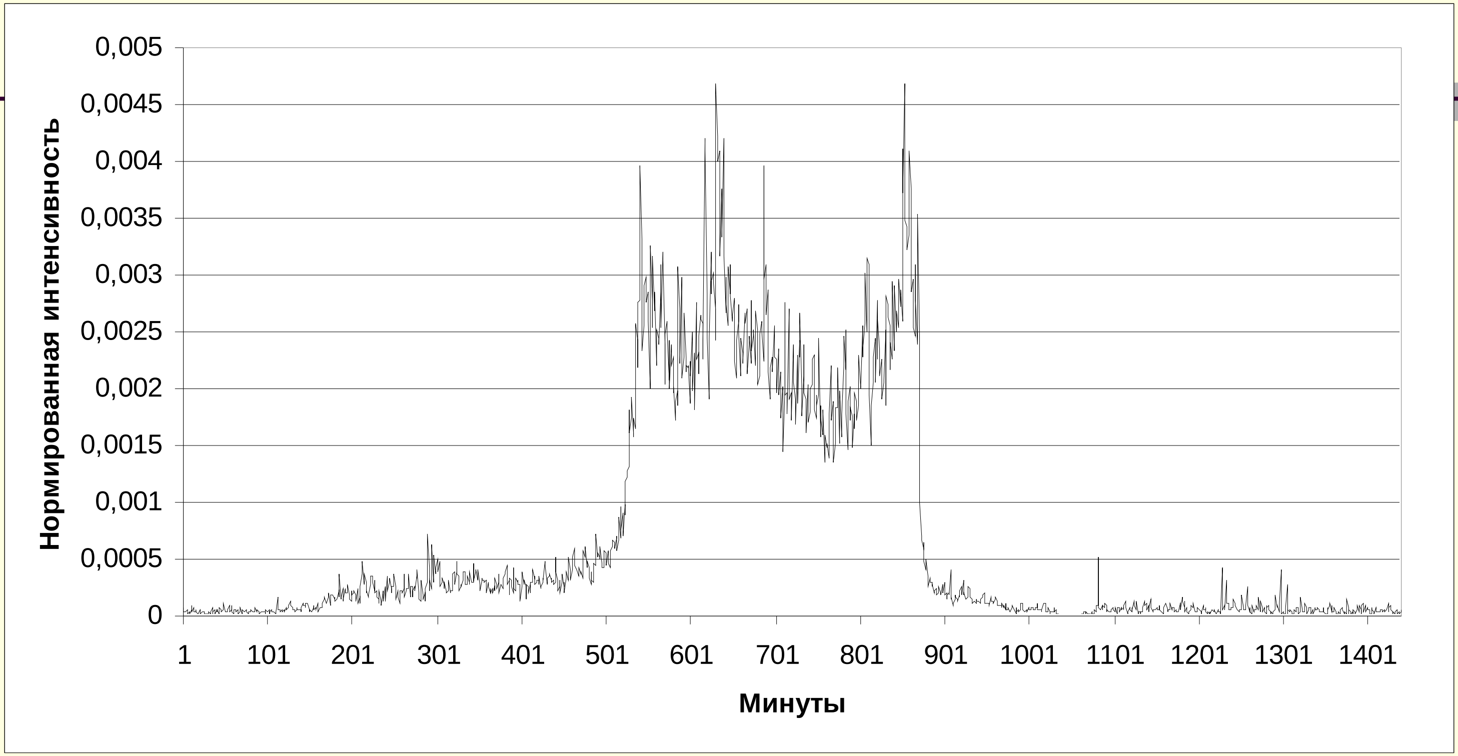
$$r(n) = \frac{V_s(n)}{V_{\max}(n)} \qquad Q(n) = \frac{V_s(n)}{V_0(n)} \qquad G(n) = Q(n)\left(1 - r(n)\right)$$

Идентификация текущей ситуации

Тиковый ряд за месяц, oil contract



Среднесуточный профиль активности



Параметр потока μ есть среднее число событий за единицу времени. Вероятность того, что за период $\Delta_t(\tau) = [t; t+\tau]$ произойдет ровно k событий, равна

$$p_k(t, t+\tau) = \frac{(\Lambda(t,\tau))^k}{k!} \exp(-\Lambda(t,\tau)), \quad \Lambda(t,\tau) = \tau\mu(t, t+\tau)$$

Эталонные распределения трендов

Нулевой прирост цены исключен из ряда.

Существует связь формы эталона с интенсивностью потока событий: чем выше интенсивность, тем выше максимумы.

Выводы

- Для нестационарного ряда всегда существует **оптимальный объем выборки**, максимизирующий достоверность анализа
- Оптимальный объем связан с горизонтом прогноза и его точностью; в первом приближении можно использовать среднее значение горизонтного ряда как минимально достаточный объем выборки
- Максимальные значения горизонтного ряда отвечают максимальной хаотизации, а минимальные – консолидации; промежутки времени между ними – период релаксации системы
- Индикатор статистической добротности показывает, выборкой какого объема лучше всего сканировать исходный ряд; критерием разладки является превышение уровня нестационарности

Кинетический подход к анализу нестационарных временных рядов

Фазовое пространство системы, ассоциированной с временным рядом

- По бесконечному набору значений x(t) можно построить ряды производных $\dot{x}(t) = x(t+1) - x(t)$, $\ddot{x}(t) = \dot{x}(t+1) - \dot{x}(t), \ldots$ и ввести фазовое пространство с мерой

$$d\Gamma = F_{\infty}\left(x, \dot{x}, \ddot{x}, \ldots, x^{(k)}, \ldots; t\right) \prod_{k=0}^{\infty} dx^{(k)}$$

- Частичные многомерные ПФР определяются затем по формуле

$$f_n\left(x, \dot{x}, \ldots, x^{(n)}; t\right) = \int F_{\infty}\left(x, \dot{x}, \ddot{x}, \ldots, x^{(k)}, \ldots; t\right) \prod_{k=n+1}^{\infty} dx^{(k)} =$$

$$= \int f_{n+1}\left(x, \dot{x}, \ldots, x^{(n)}, x^{(n+1)}; t\right) dx^{(n+1)} .$$

Модель эволюции ВПФР

- Не предполагая, что ВПФР отвечает какой-либо дискретной динамической системе, построим оператор эволюции ВПФР, сохраняющий ее нормировку. Для этого рассмотрим совместное выборочное распределение $F_T(\xi, t)$ величин

$$x(t),\ \dot{x}(t) = x(t+1) - x(t),\ \ddot{x}(t) = \dot{x}(t+1) - \dot{x}(t), \ldots$$

так что $$f_T(x,t) = \int F_T(x, \dot{x}, t)\,d\dot{x}$$

- Формальное уравнение эволюции ВПФР, полученное на основе теоремы Лиувилля, имеет вид

$$\frac{\partial F_T(\xi,t)}{\partial t} + div_\xi\left(\dot{\xi} F_T(\xi,t)\right) = 0, \quad \xi = (x, \dot{x}, \ldots), \quad \dot{\xi} = (\dot{x}, \ddot{x}, \ldots)$$

- Поскольку объем выборки T конечен, то число компонент ξ должно быть ограниченным. Обрыв цепочки позволяет получить модельное уравнение эволюции ВПФР с локально меняющейся скоростью.

Эмпирическая скорость

- Рассмотрим совместное выборочное распределение случайной величины и ее приращений, т.е. положим $\xi = (x, \dot{x})$. Введем среднюю локальную скорость $u_T(x,t)$ согласно равенству

$$u_T(x,t) f_T(x,t) = \int \dot{x} F_T(x,\dot{x},t) d\dot{x}$$

- Тогда уравнение эволюции ВПФР («эмпирическое» уравнение Лиувилля, записанное относительно эмпирической скорости) имеет вид

$$\frac{\partial f_T(x,t)}{\partial t} + \frac{\partial \left(u_T(x,t) f_T(x,t)\right)}{\partial x} = 0$$

- Аналогично вводится среднее локальное ускорение (и т.д.)

$$w_T(x,\dot{x},t) F_T(x,\dot{x},t) = \int \ddot{x} \Phi_T(x,\dot{x},\ddot{x},t) d\ddot{x}$$

Замыкание кинетической модели

Дополним уравнение Лиувилля уравнением эволюции для скорости:

$$\left(\frac{\partial u}{\partial t} - u\frac{\partial u}{\partial x}\right) f - u^2 \frac{\partial f}{\partial x} = -\frac{\partial (ef)}{\partial x} + Wf$$

$$e(x,t) f(x,t) = \int \dot{x}^2 F(x,\dot{x},t) d\dot{x}, \quad W(x,t) f(x,t) = \int w(x,\dot{x},t) F(x,\dot{x},t) d\dot{x}$$

Уравнение для u(x,t) требуется вследствие того, что, по построению, скорость известна в предыдущий момент по сравнению с ВПФР:

$$u(i+1,t) = \frac{u(i,t) f(i,t) - f(i,t) + f(i,t+1)}{f(i+1,t)}$$

Если известны e(x,t) и W(x,t), то система замкнется. В противном случае добавляются уравнения для этих величин, которые зависят от моментов высших порядков и распределений более высокой размерности. Обрыв цепочки на каком-нибудь порядке приводит к замкнутым моделям эволюции.

СПАСИБО ЗА ВНИМАНИЕ