Орлов Ю.Н.
22.10.2018

Оптимальное равномерное разбиение гистограммы на классовые интервалы

1. Неточность моделирования
2. Уровень стационарности и число интервалов
3. Оптимальное разбиение гистограммы
4. Сравнение нестационарных распределений

Неточность моделирования

ε_1 неточность модели (дифф. ур.) по сравнению с реальностью

ε_2 неточность измерительного инструмента этой реальности

ε_3 неточность вычислительного инструмента (компьютер)

ε_4 погрешность дискретного алгоритма

$$\frac{\partial f(x,t)}{\partial t} + \frac{\partial}{\partial x} u(x,t) f(x,t) = 0 \qquad \frac{f(i,t+1) - f(i,t)}{\tau} = \frac{u(i,t) f(i,t) - u(i+1,t) f(i+1,t)}{h}$$

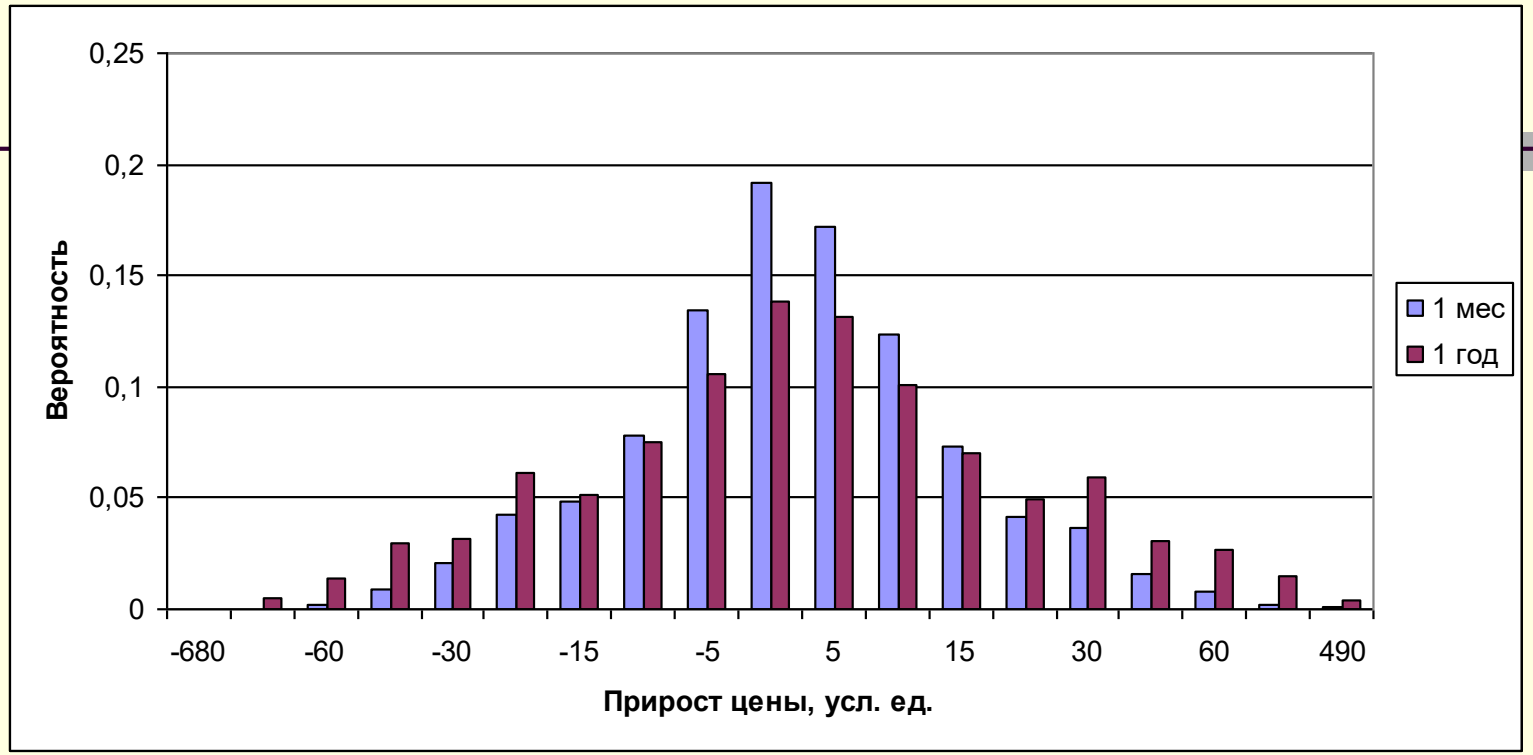
$$\left| x(t_n) - \tilde{x}(t_n) \right| \le \varepsilon \qquad \frac{dx(t_n)}{dt} = \frac{x(t_{n+1}) - x(t_n)}{\tau} + O(\tau) \approx \frac{\tilde{x}(t_{n+1}) - \tilde{x}(t_n)}{\tau} + O(\tau)$$

погрешность аппроксимации производной $O(\tau) + 2\varepsilon / \tau$

погрешности равны, если $\tau = O\left(\sqrt{\varepsilon}\right)$

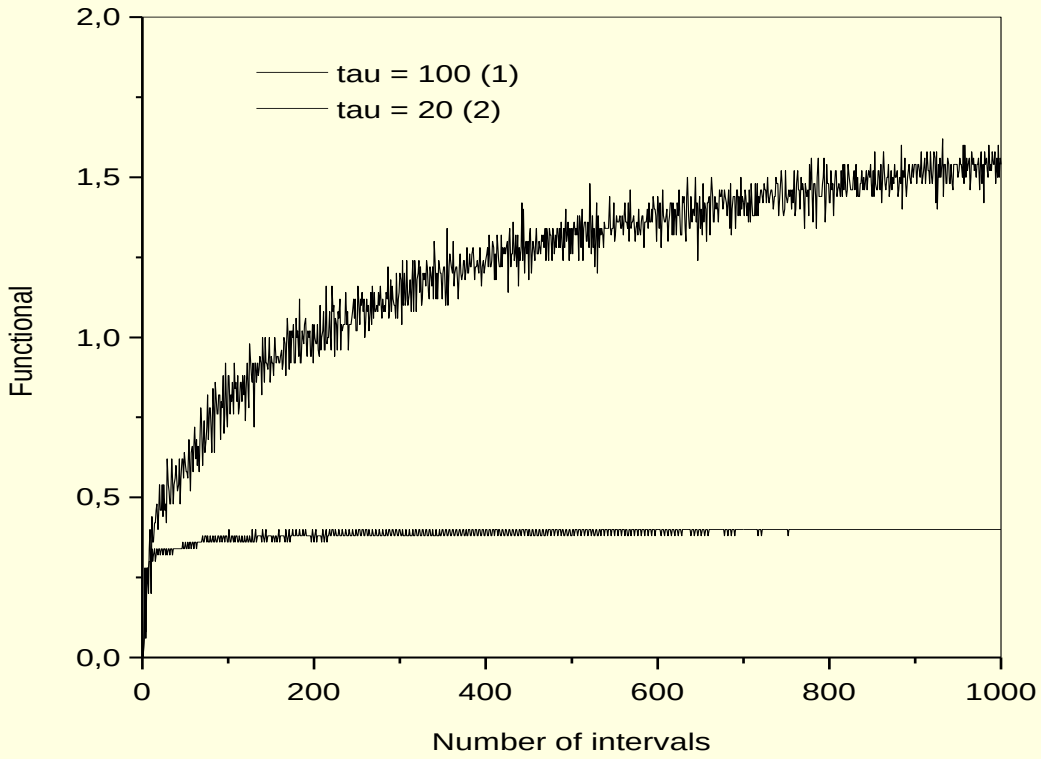
ε_5 неточность приближения решения дифф.ур. гистограммой

Вопросы при построении ВПФР



1. С какой мелкостью разбивать область гистограммы?
2. По выборке какой длины строить гистограмму?
3. На каком горизонте по времени распределение стационарно?
4. С какой точностью оценены вероятности?
5. Как выбрать индикатор и уровень разладки?

Зависимость расстояния в норме L1 от числа классовых интервалов, выборка=100 данных



$$\rho(N;tau;t) = \left\| f_N(x,t) - f_N(x,t+tau) \right\|$$

Оптимальное число n интервалов разбиения

- **Sturges, 1926** $n = 1 + \log_2 N$
- **Mann, Wald, 1942** $n = 4\left(\frac{3}{4}(N-1)^2\right)^{1/5}$
- **Смирнов, 1950** $n = AN^{1/3}$
- **Алексеева, 1975** $n = \frac{1+\kappa}{6} N^{2/5}$
- **Лернер, 1976** $n = \left(\frac{N\left|f''(x)\right|_{\max}}{4f(x)_{\max}}\right)^{1/5}$
- **Scott, 1979** $n = \left(\frac{N}{6}\int_{-\infty}^{+\infty} f'^2(x)dx\right)^{1/3}$

Точность оценки вероятности по эмпирической частоте

Пусть выбрано равномерное разбиение на n интервалов.

Тогда эмпирическая вероятность по выборке длины *N* равна

$$f_N(j) = \frac{k_j}{N}, \quad j = 1, 2, ..., n$$

Дисперсия выборочной вероятности равна

$$s_N^2(j) = f_N(j) \cdot \left(1 - f_N(j)\right)$$

Если каждая вероятность оценивается с точностью ε, то выборочное распределение ε-стационарно:

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

Статистика Стьюдента

Если изучаемое распределение стационарно, то статистика

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

имеет распределение Стьюдента $\varphi(t, N-1)$ с N-1 степенями свободы:

$$\varphi(t;N) = \frac{\Gamma\left(\frac{N+1}{2}\right)}{\sqrt{\pi N}\,\Gamma(N/2)}\left(1+\frac{t^2}{N}\right)^{-\frac{N+1}{2}}$$

При больших *N* распределение Стьюдента сходится к нормальному:

$$\lim_{N\to+\infty}\varphi(t;N) = \frac{1}{\sqrt{2\pi}}\exp\left(-t^2/2\right)$$

Доверительный интервал оценки вероятности

Интервальная оценка частоты попадания в j-ый классовый промежуток на уровне значимости α имеет вид

$$\left| f^*(j) - f_N(j) \right| \le t_{1-\alpha/2}(N-1)\frac{s_N(j)}{\sqrt{N}}$$

где $t_\gamma(N)$ есть γ- квантиль распределения $\varphi(t;N)$ (приближенно квантиль нормального распределения).

Поскольку уровень значимости равен принимаемой ошибке в оценке частоты, то $\alpha = \varepsilon$. Тогда интегральная неопределенность в распределении заведомо не превосходит ε, если

$$t_{1-\varepsilon/2}\frac{s_N(j)}{\sqrt{N}} \le \varepsilon f^*(j)$$

Согласованная точность оценки

- Если $\left| f_N(j) - f^*(j) \right| \le \varepsilon f^*(j)$, то выполнено неравенство

$$t_{1-\varepsilon/2} \frac{s_N(j)}{\sqrt{N}} \le \varepsilon f^*(j)$$

Введем средневзвешенную точность через взвешенный квантиль:

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

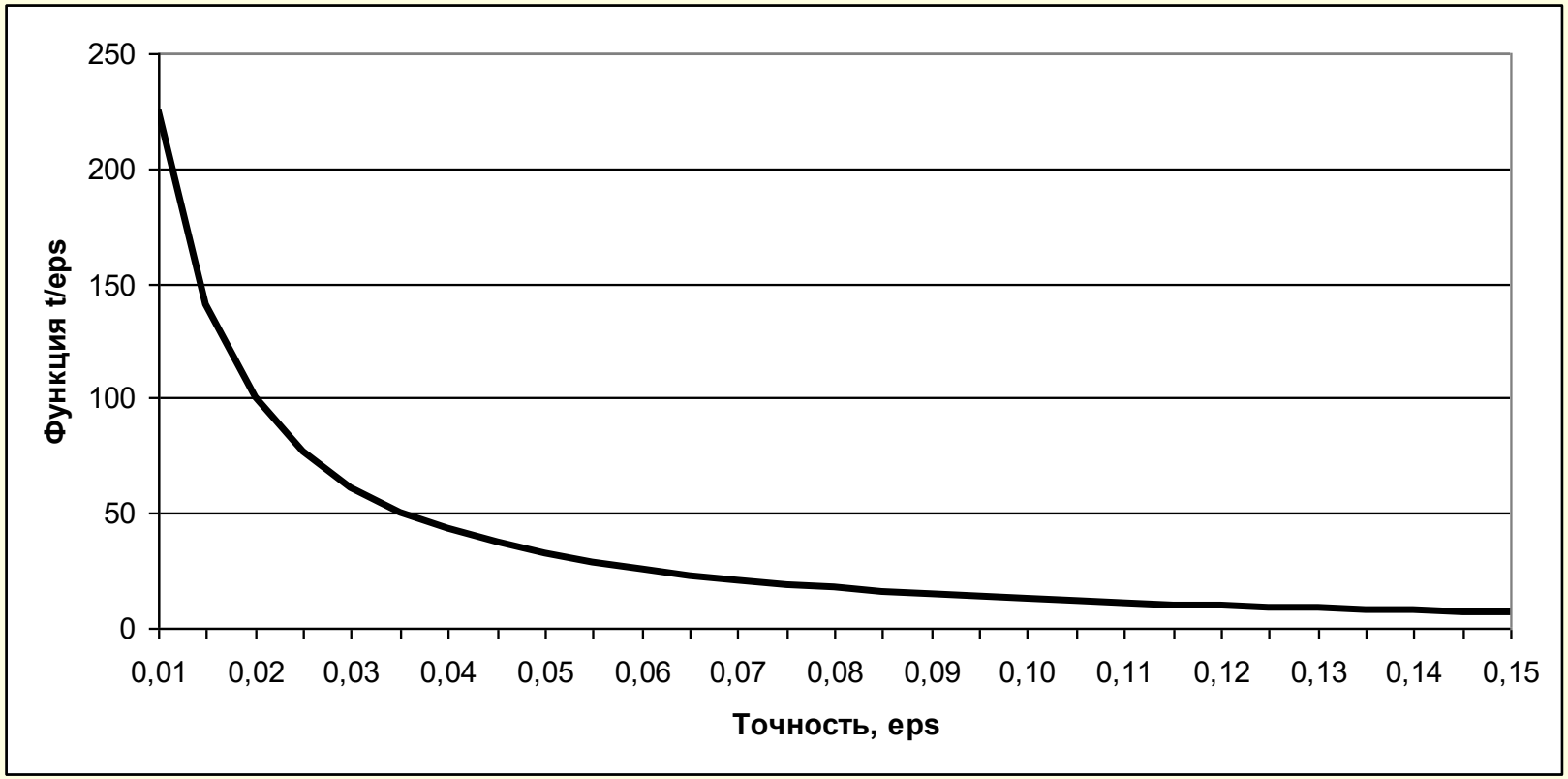
$$\Sigma_N(n) = \sum_{j=1}^{n} s_N(j) = \sum_{i=1}^{n} \sqrt{f_N(j)\left(1 - f_N(j)\right)}$$

Тогда точность оценки распределения вероятностей есть

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

Вид функции точности

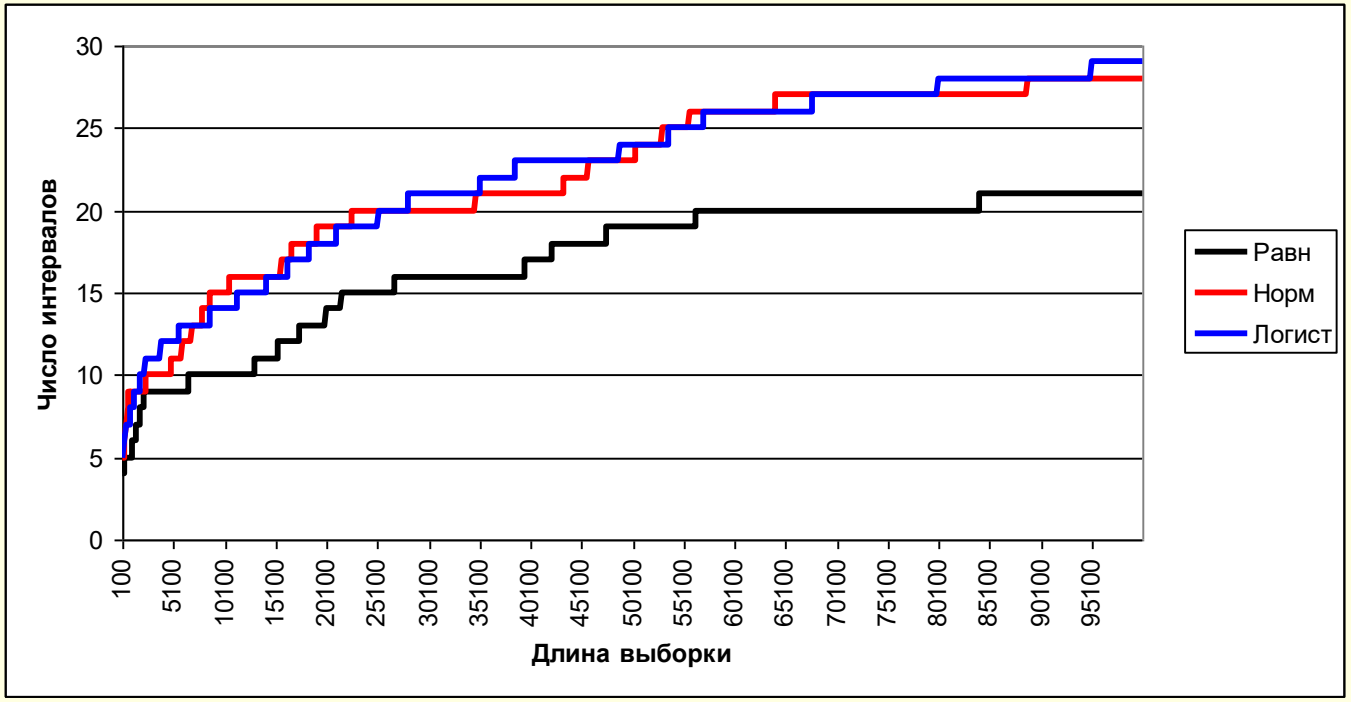
$$\varphi(\varepsilon) = \frac{t_{1-\varepsilon}}{\varepsilon}$$



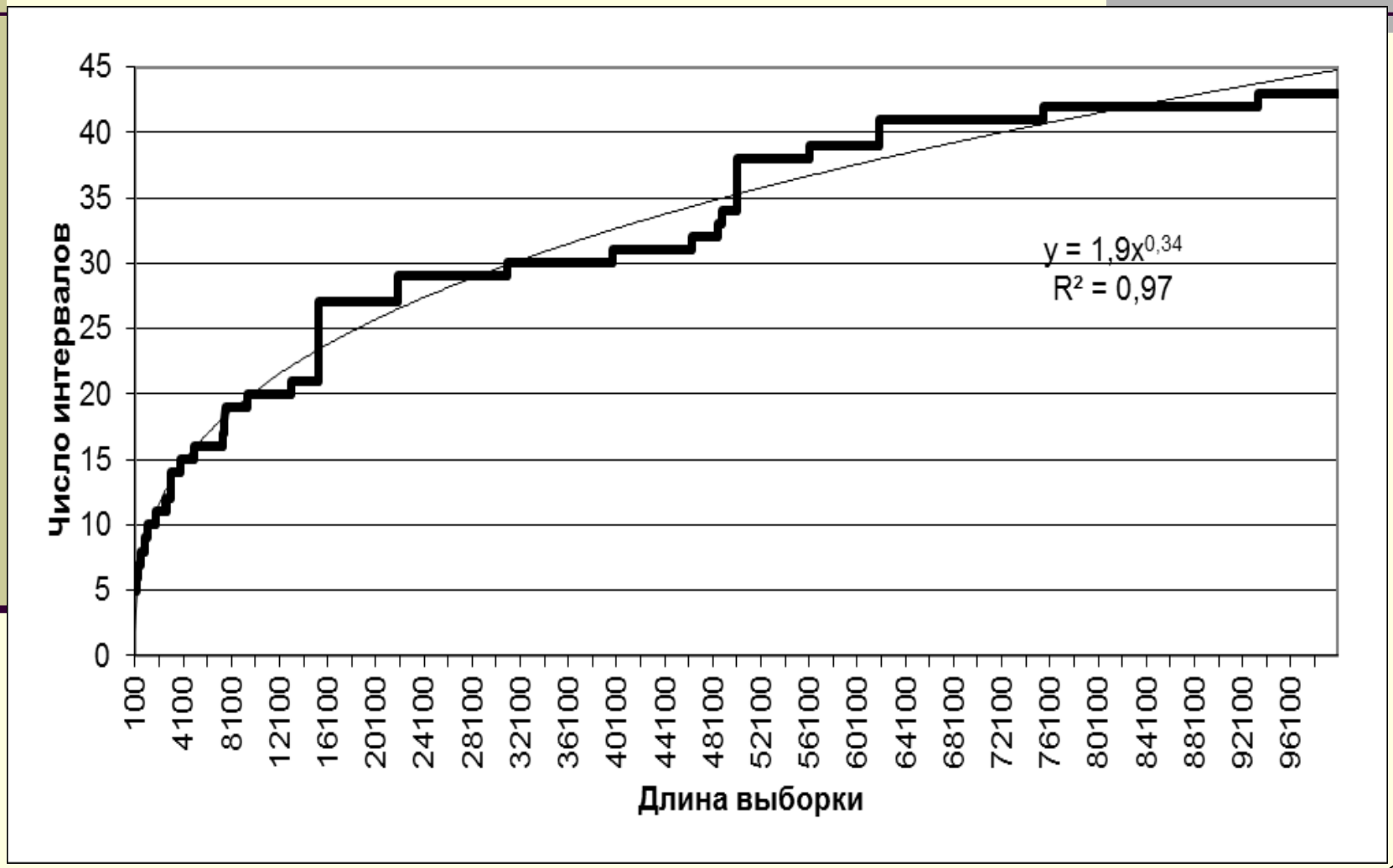
Выбор числа интервалов

- Точность оценки распределения совпадает с точностью измерения случайной величины. Тогда

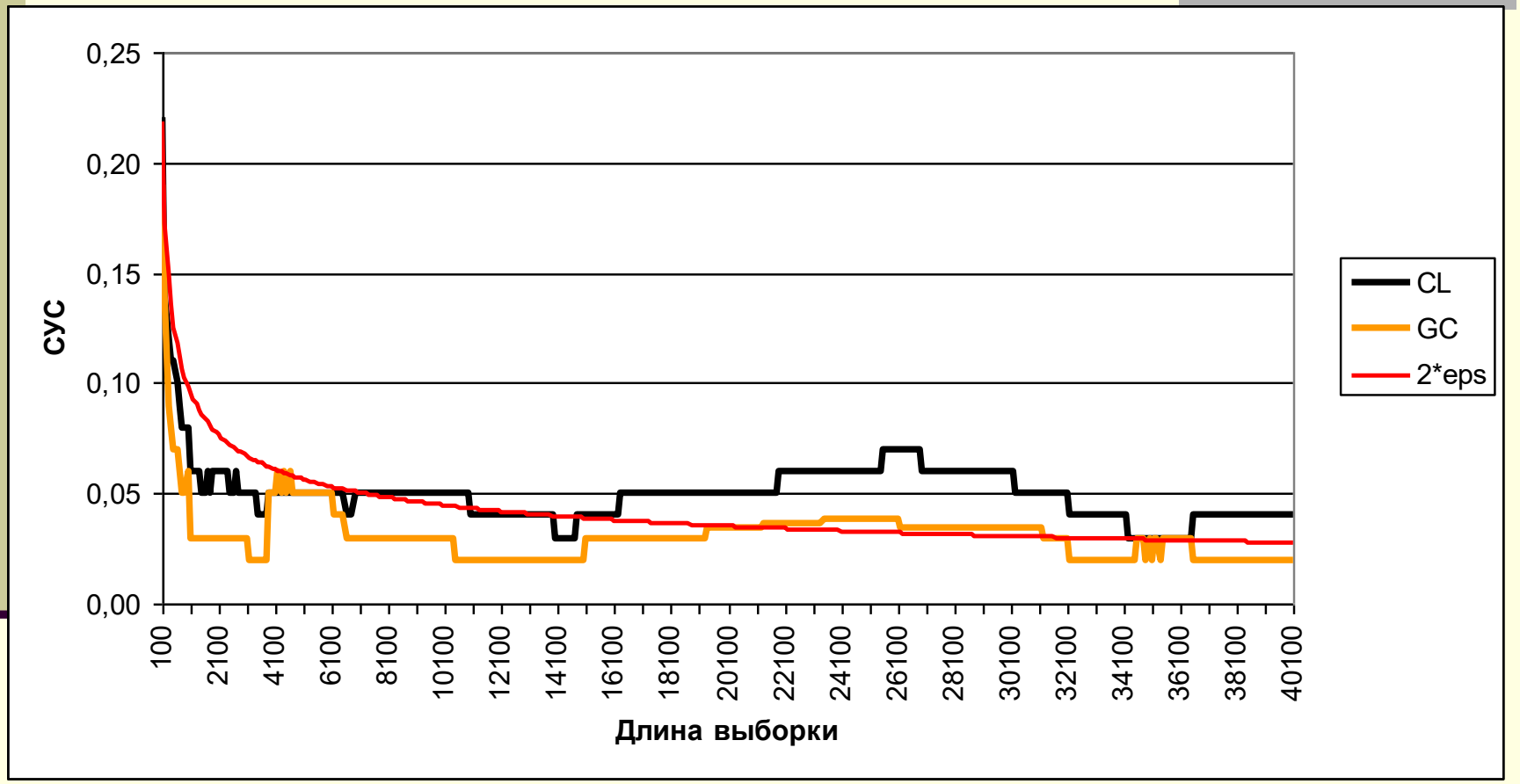
$$\varphi(\varepsilon) = \frac{t_{1-\varepsilon}}{\varepsilon}, \quad \psi = \varphi^{-1}, \qquad \varepsilon = \frac{1}{n} = 2\psi\left(\frac{2\sqrt{N}}{\Sigma_N(n)}\right)$$



Асимптотика числа интервалов



Сравнение распределений



СПАСИБО ЗА ВНИМАНИЕ