СТАТИСТИКА БОЛЬШИХ ДАННЫХ

Лекция 14. Задачи непараметрической статистики. Часть 2

Модели «дрейфа» распределения данных

Модель последовательного изменения формы распределений в виде смеси

Рассмотрим серию выборок, где начальная выборка порождается распределением $F_0(x)$, конечная выборка порождается распределением $F_1(x)$, а промежуточные выборки – смесью этих распределений:

$$(1-\varepsilon)F_0(x)+\varepsilon F_1(x),$$

где ε изменяется от 0 до 1 с заданным шагом.

Данная модель широко применяется для моделирования бимодальных распределений, часто встречающихся в описании природных процессов, например, времени между извержениями определенных гейзеров, в экономике при наличии некоторой сезонности в спросе на продукт, и т.п.

Важным преимуществом этих моделей является простота в вычислении характеристик распределений. Весовой коэффициент $\boldsymbol{\varepsilon}$ можно рассматривать как вероятность того, что случайная величина порождается распределением F_1.

Модели «дрейфа» распределения данных

Далее рассмотрено несколько смесей, в которых распределение $F_0(x)$ соответствует $\mathcal{N}(0,1)$ – стандартное нормальное распределение. Используемые варианты смесей:

1. $F_1(x)$ соответствует нормальному распределению $\mathcal{N}(3,1)$.
2. $F_1(x)$ соответствует распределению Коши $C(3,1)$.
3. $F_1(x)$ соответствует равномерному распределению $U\left(-\sqrt{3},\sqrt{3}\right)$

Для исследования робастности добавлялись импульсные помехи с амплитудой 10 стандартных отклонений в 5% данных.

Модели «дрейфа» распределения данных

Модель последовательного изменение обобщенного гауссовского распределения:

Плотность распределения:

$$\beta = \beta_0 + k\Delta\beta;\ \beta_0 = 2, \alpha = \sqrt{2}, \mu = 0.$$

$$f(x;\ \mu, \alpha, \beta) = \frac{\beta}{2\alpha\Gamma\left(\frac{1}{\beta}\right)} e^{-\left(\frac{|x-\mu|}{\alpha}\right)^{\beta}}$$

Задается закон постепенного изменения параметра β с шагом $\Delta\beta$. Рассматривается 2 варианта: переход от нормального распределения $(\beta = 2)$ к распределению Лапласа $(\beta = 1)$, и переход от нормального к практически равномерному распределению $(\beta = 10)$.

Модели «дрейфа» распределения данных

Модель последовательного изменения числа степеней свободы в распределении Стьюдента

Плотность распределения:

$$t(x;n)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)\left(1+\frac{x^2}{n}\right)^{\frac{n+1}{2}}}$$

$$n=n_0+\Delta n;,n_0=10,n_1=1;$$

Задается закон постепенного изменения числа степеней свободы n с шагом Δn. При $n=10$ имеем приближенно нормальное распределение, при $n=1$ – распределение Коши.

Методология исследования

- Методом Монте Карло исследуется поведение приведенных выше непараметрических статистических критериев при работе с тремя перечисленными моделями. В каждом эксперименте проводится $N = 1000$ повторений.
- Исследуется влияние размера выборки. Рассматриваются выборки из $\{10, 20, 40, 80, 100, 200, 300, 900\}$ точек.
- Величина *m* везде равна 50.

Во всех моделях исходно предполагается, что данные распределены нормально (нулевая гипотеза).

Методология исследования

Сравниваются четыре критерия при решении задачи проверки простой гипотезы H_0: $F(x) = F_0(x)$. Для всех критериев уровень значимости $\alpha = 0.05$.

Показатели чувствительности

1. Для модели смесей

$$maxEps = \max_{\varepsilon}\{P(\varepsilon) > 0.5\}$$

2. Для модели обобщенного гауссовского распределения

$$maxBeta = \max_{\beta}\{P(\beta) > 0.5\}$$

3. для модели распределения Стьюдента

$$maxDF = \max_{n}\{P(n) > 0.5\}$$

где $P(a) = \frac{l}{N}$ – вероятность принятия гипотезы H_0, l – количество не отвергнутых тестов при параметре a.

Результаты моделирования

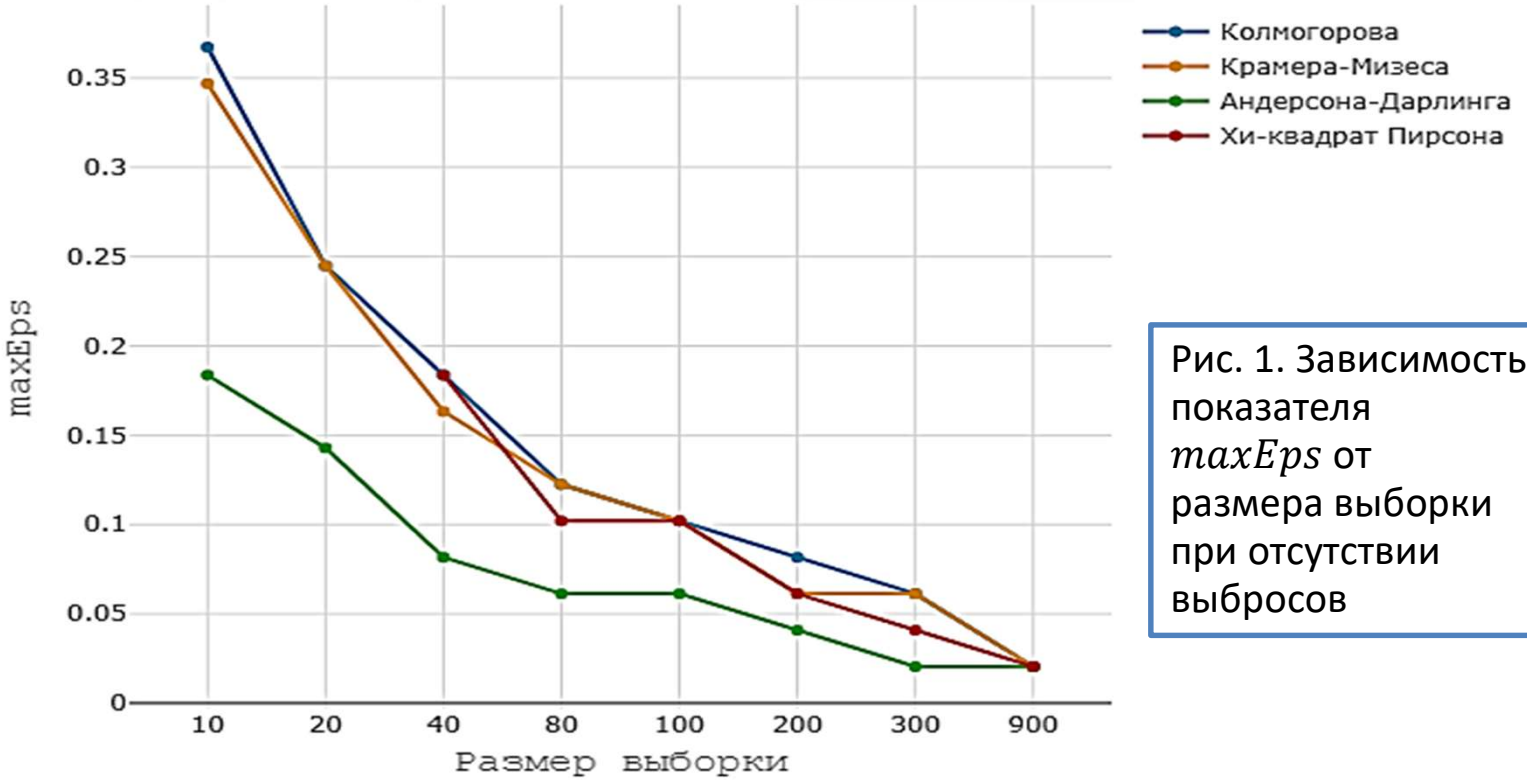


Рис. 1. Зависимость показателя $maxEps$ от размера выборки при отсутствии выбросов

Результаты моделирования

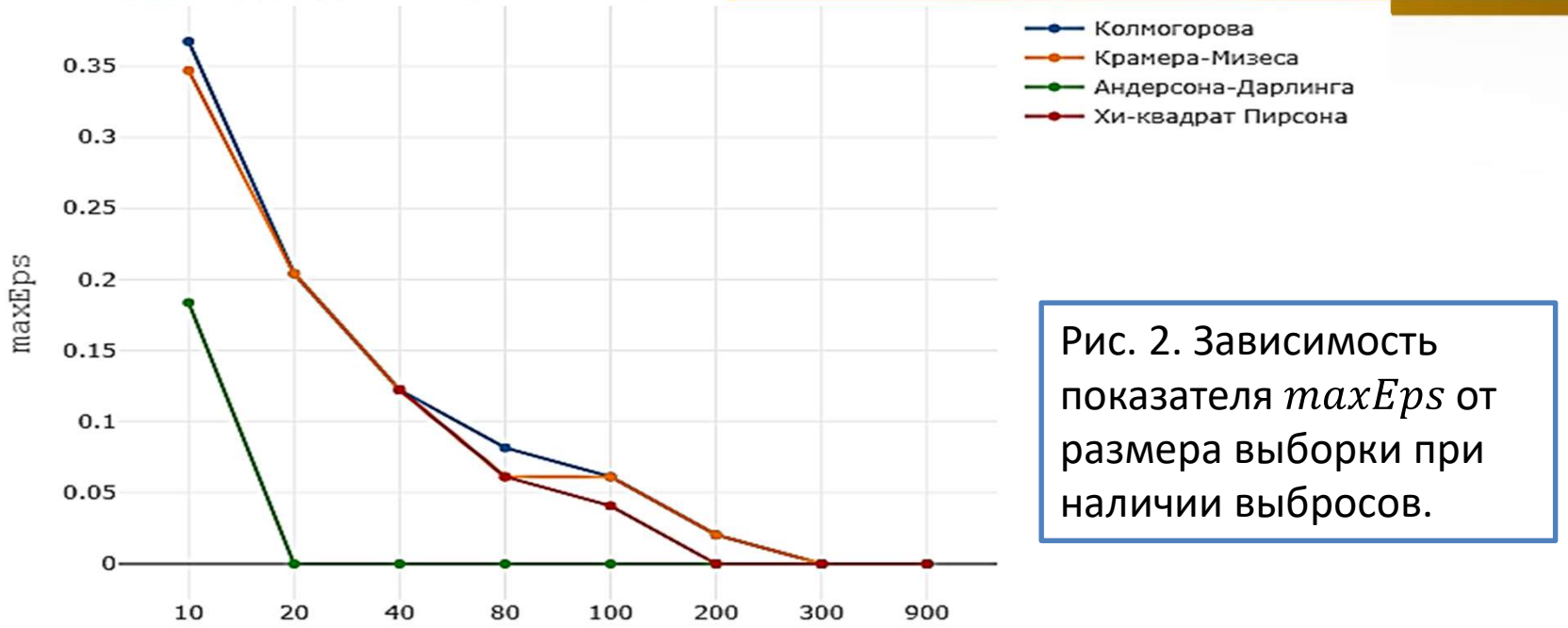


Рис. 2. Зависимость показателя $maxEps$ от размера выборки при наличии выбросов.

Наиболее чувствительным себя показал критерий Андерсона-Дарлинга, особенно сильно это заметно при наличии выбросов.

Результаты моделирования

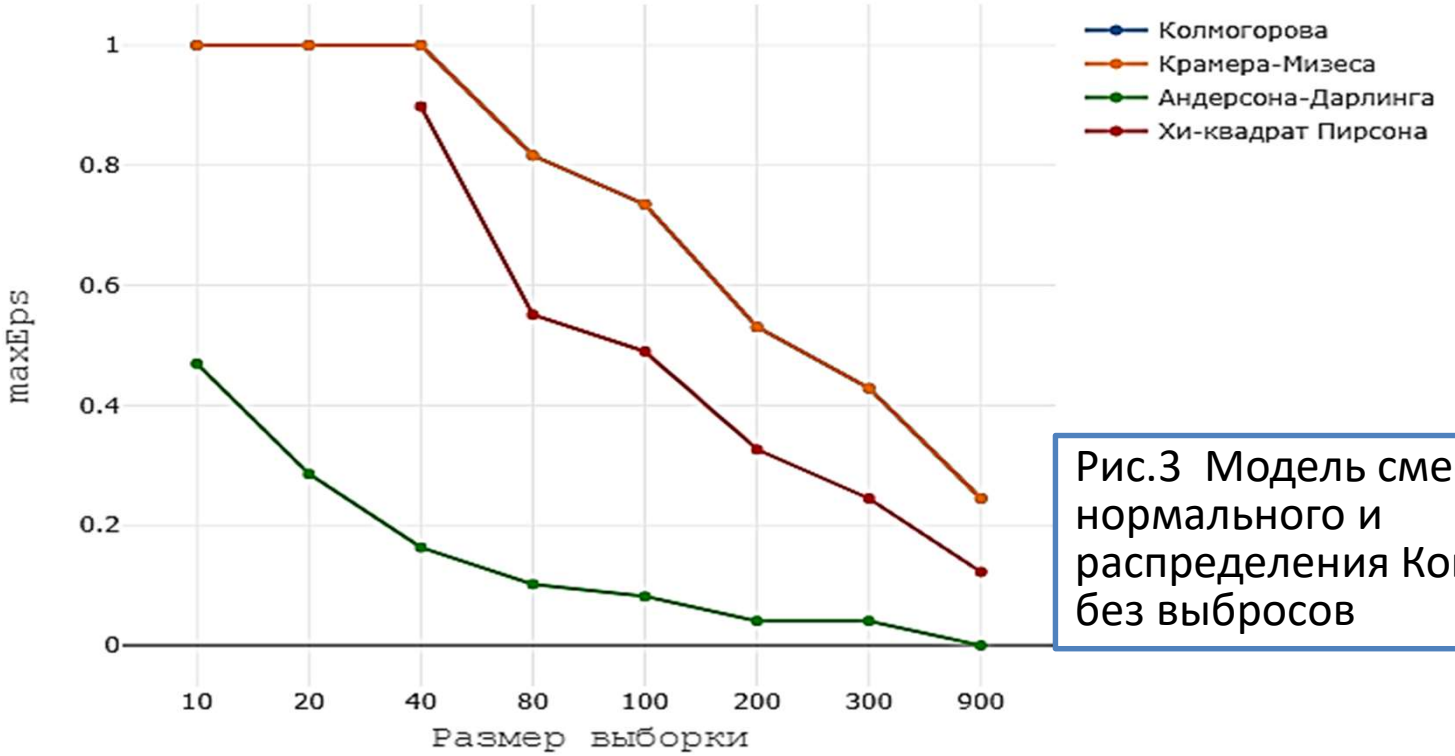


Рис.3 Модель смеси нормального и распределения Коши без выбросов

Результаты моделирования

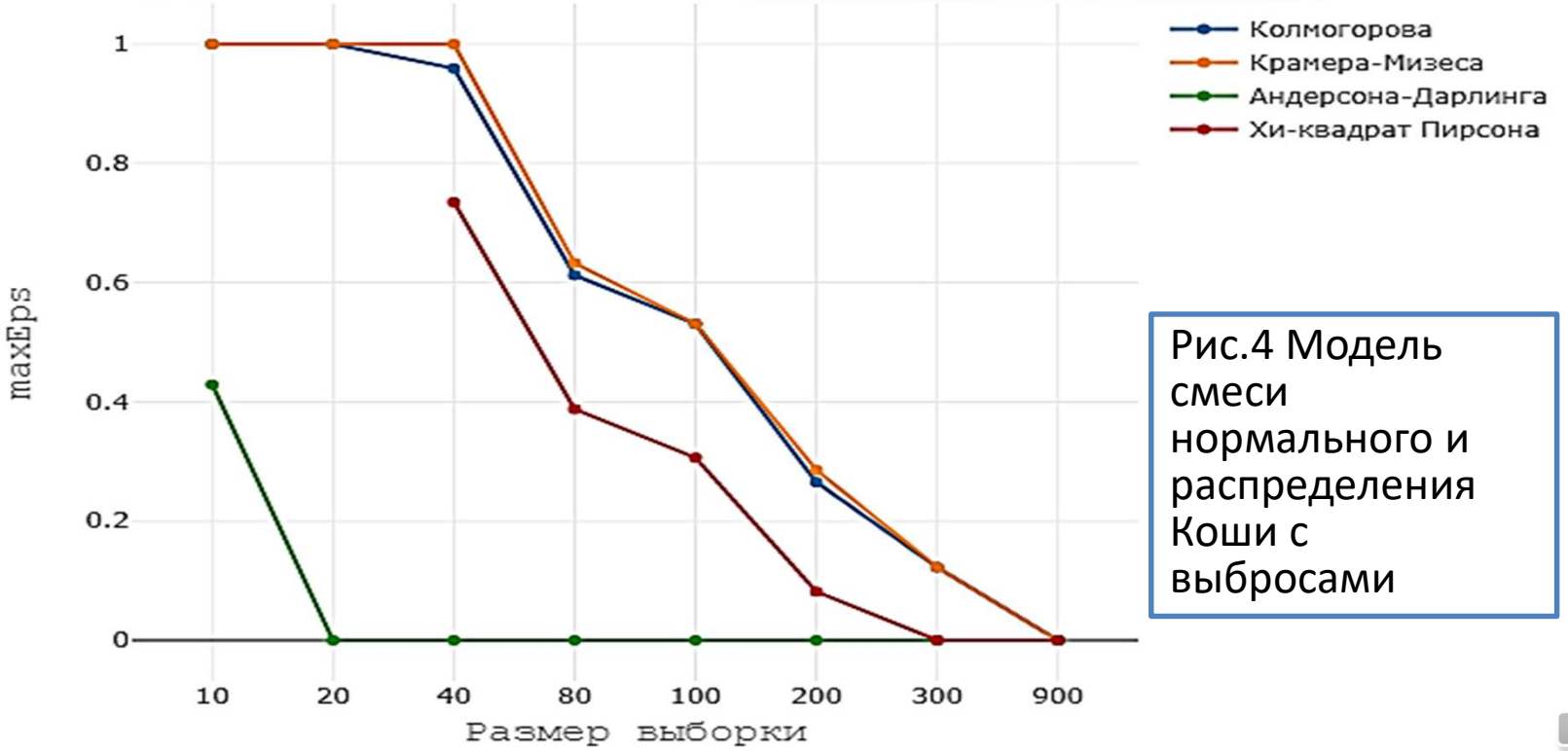


Рис.4 Модель смеси нормального и распределения Коши с выбросами

Результаты моделирования

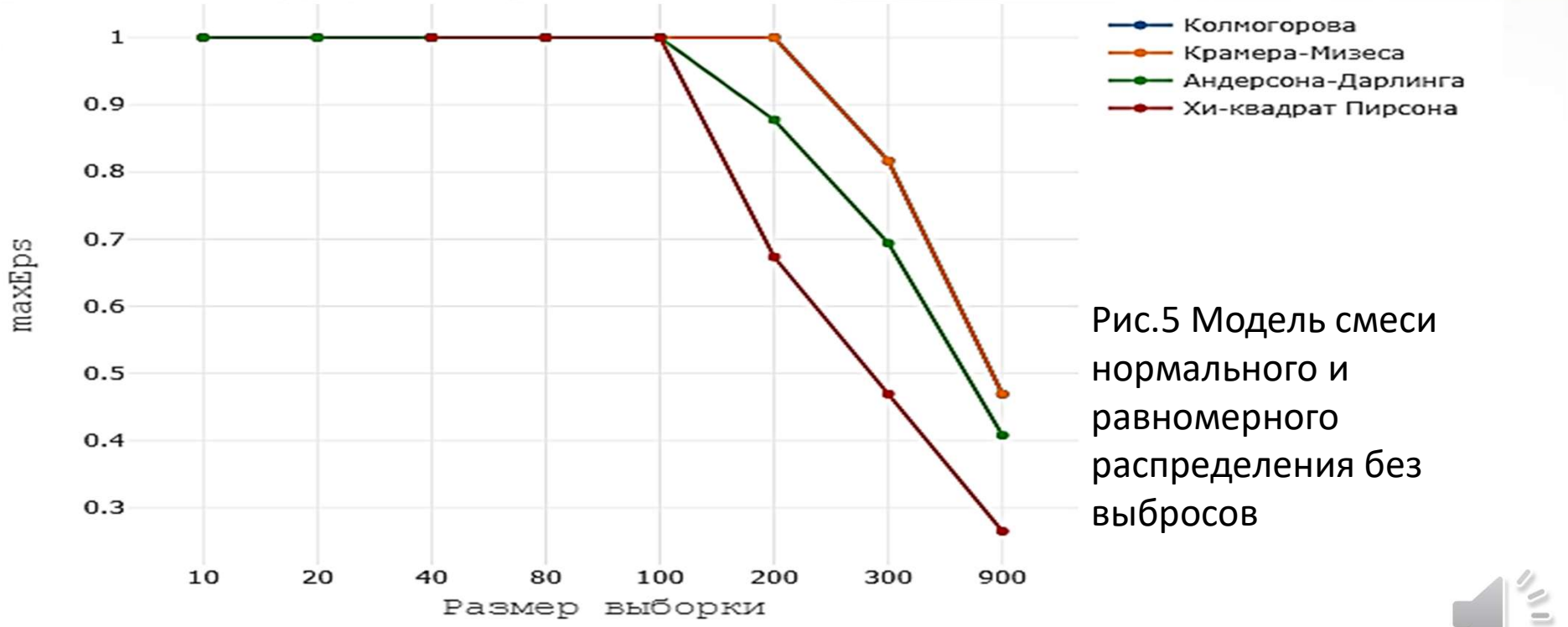


Рис.5 Модель смеси нормального и равномерного распределения без выбросов

Результаты моделирования

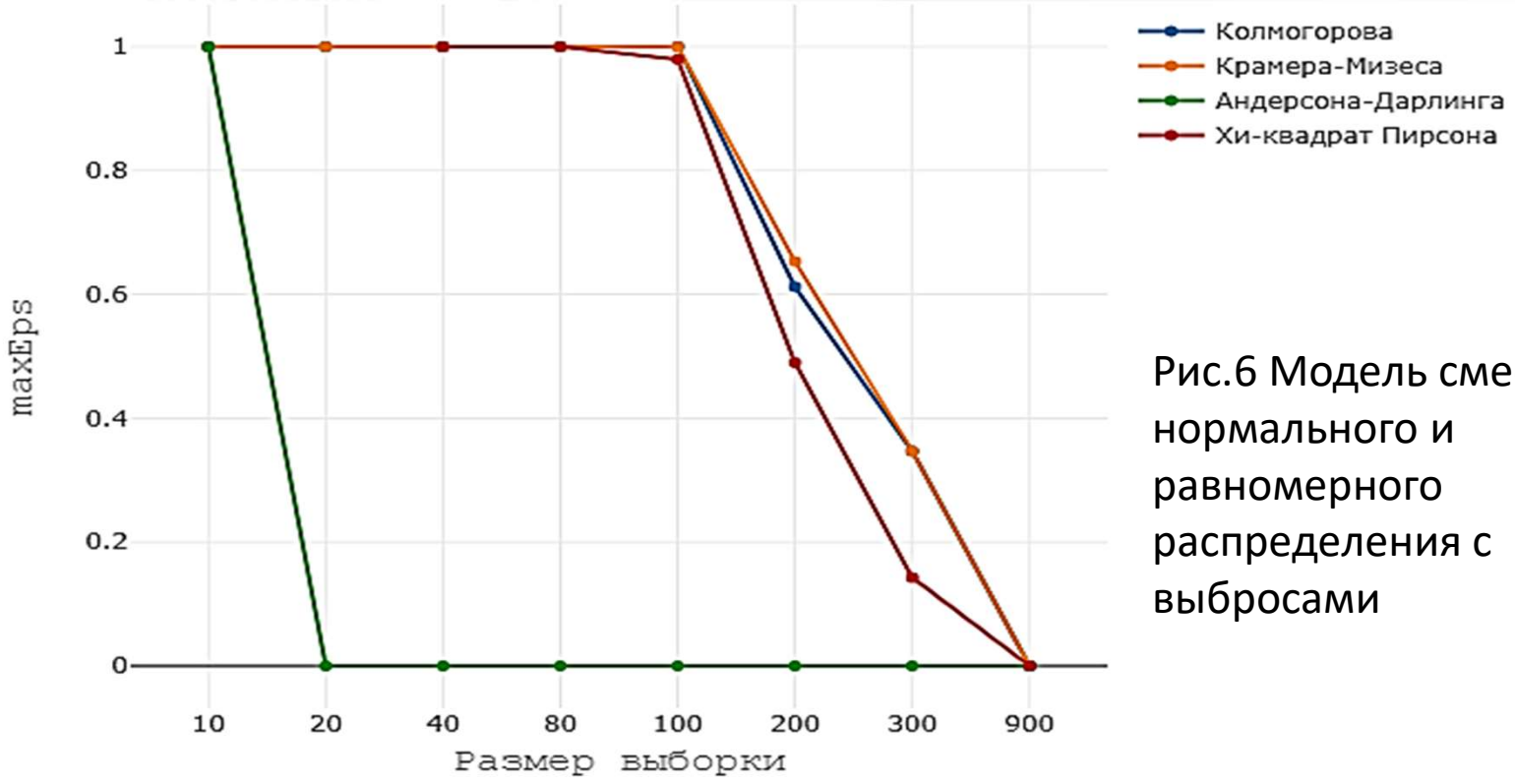


Рис.6 Модель смеси нормального и равномерного распределения с выбросами

Результаты моделирования

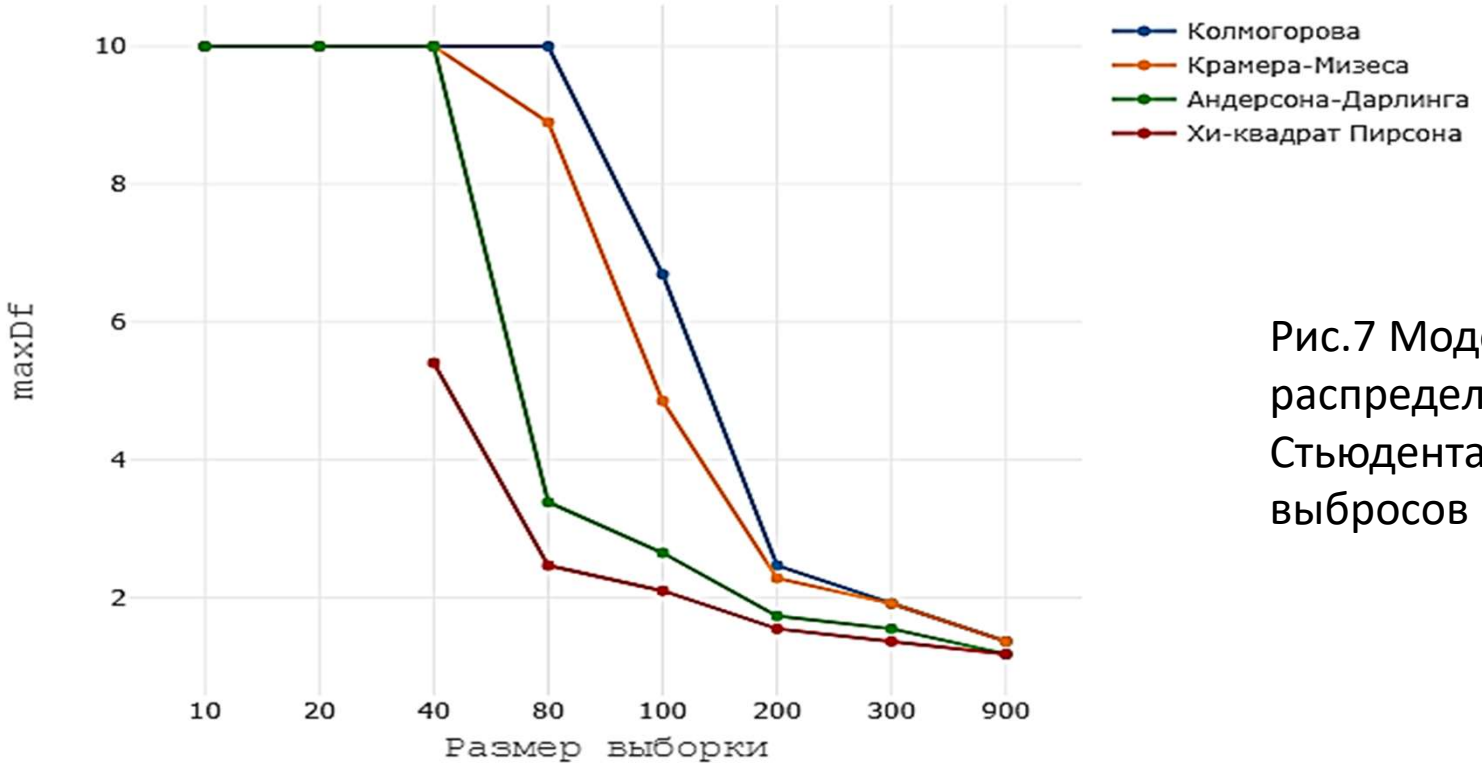


Рис.7 Модель распределения Стьюдента без выбросов

Результаты моделирования

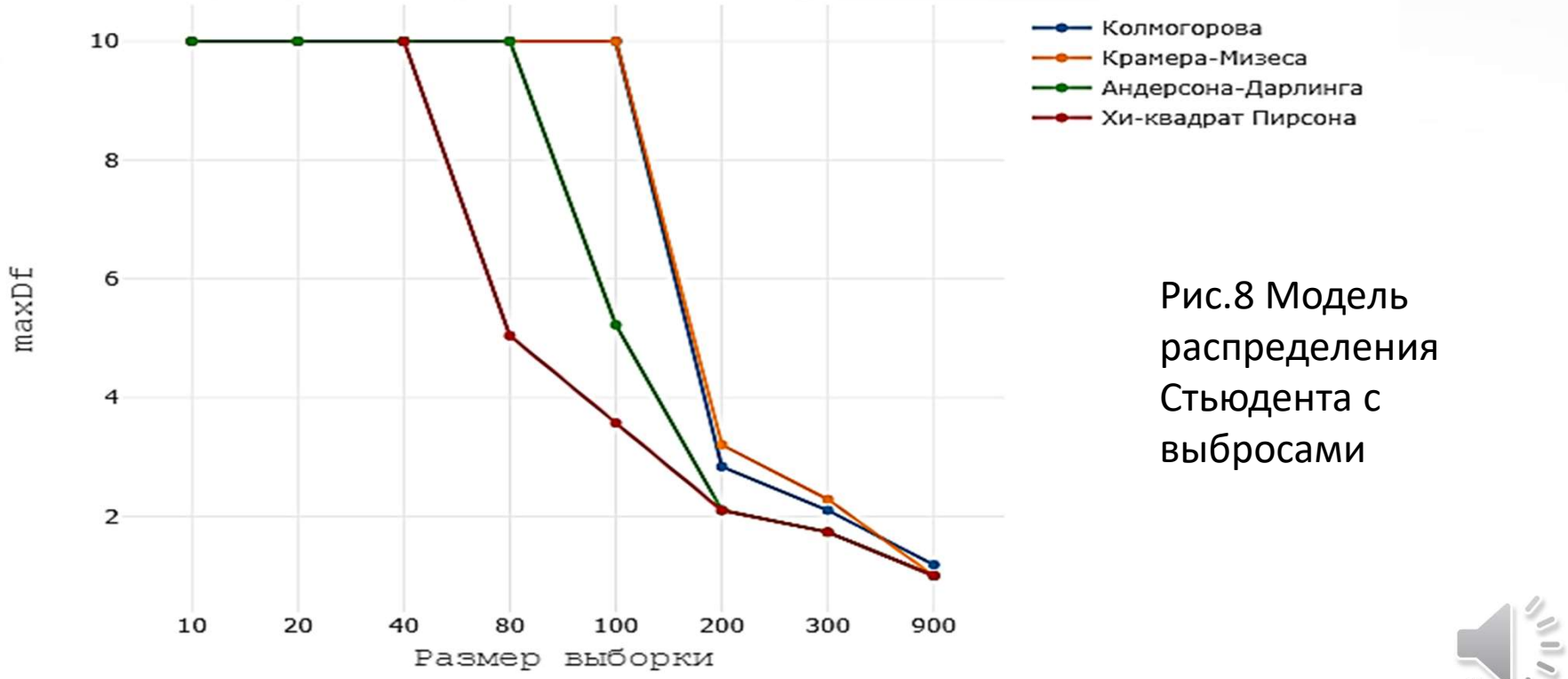


Рис.8 Модель распределения Стьюдента с выбросами

Результаты моделирования

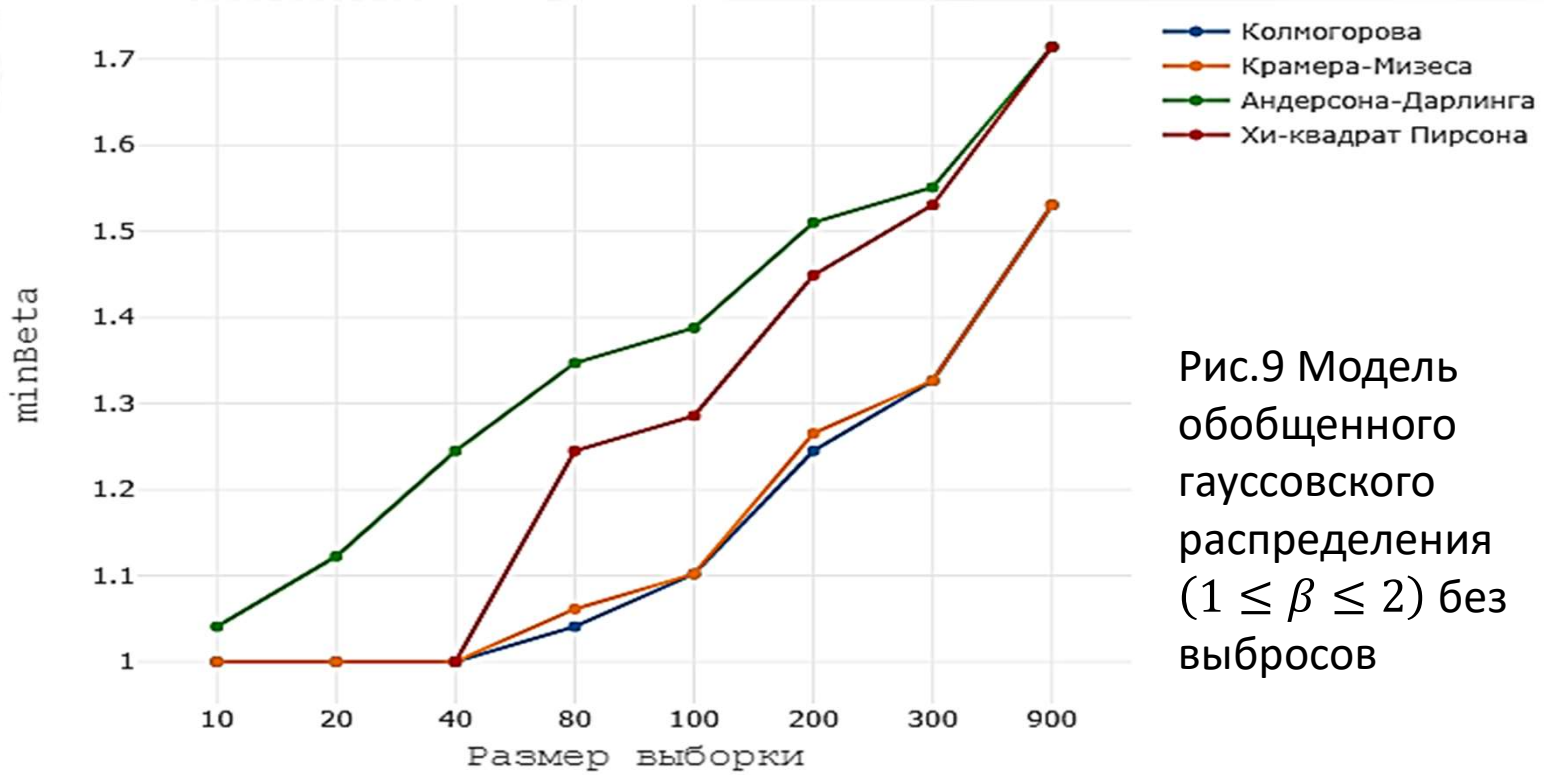


Рис.9 Модель обобщенного гауссовского распределения $(1 \leq \beta \leq 2)$ без выбросов

Результаты моделирования

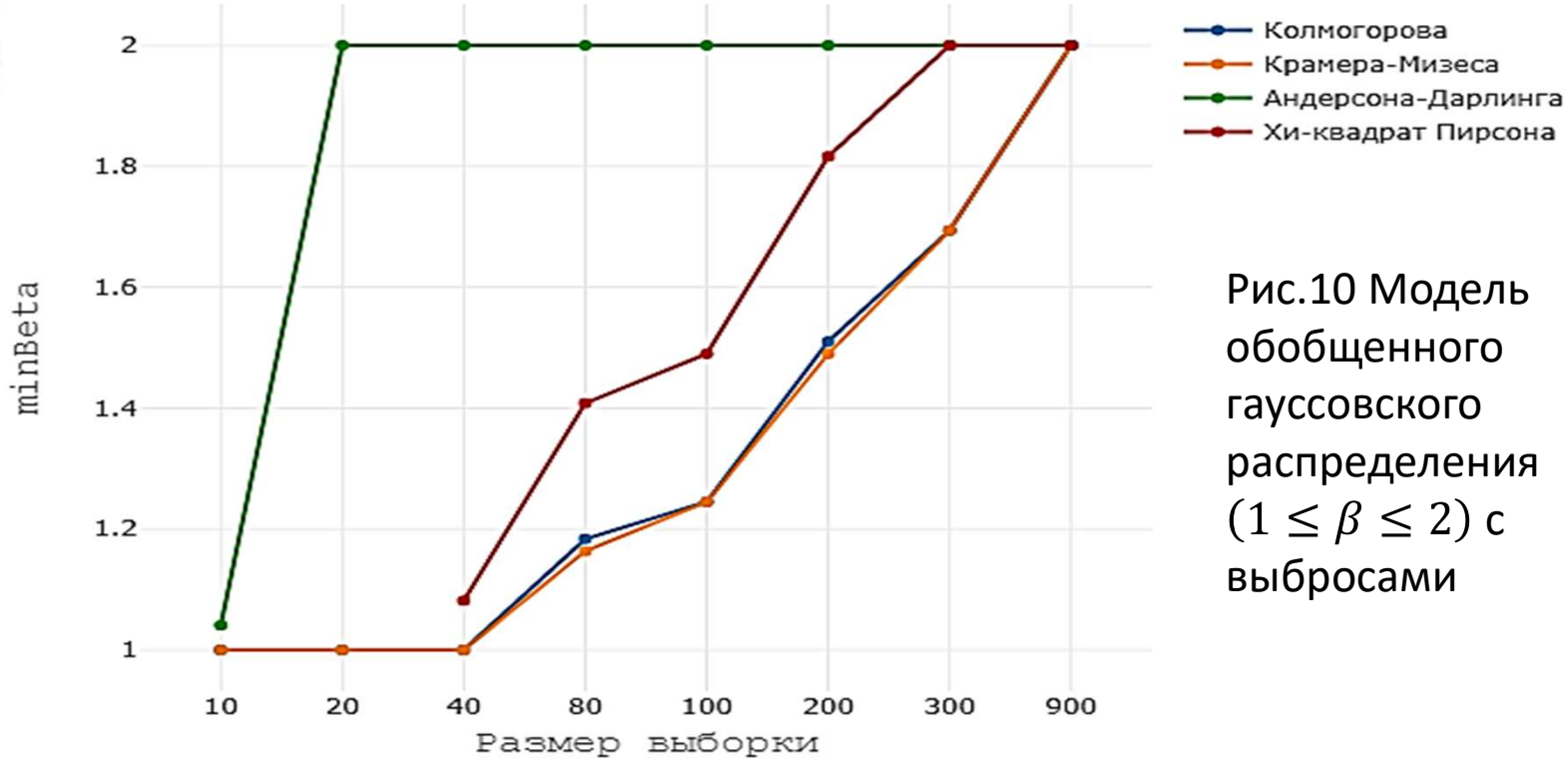
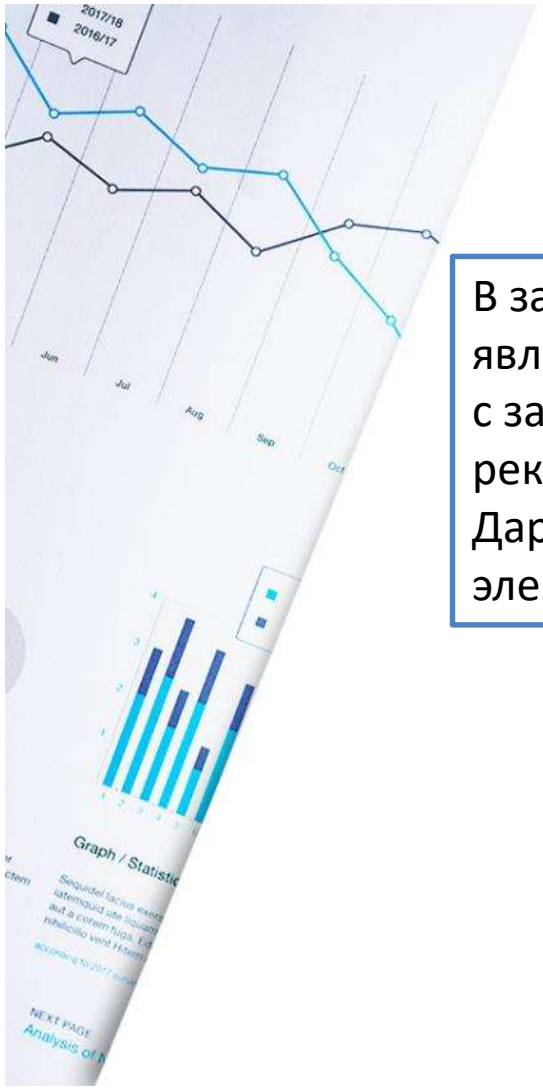


Рис.10 Модель обобщенного гауссовского распределения $(1 \leq \beta \leq 2)$ с выбросами

Рекомендации

Модель данных	Наиболее чувствительный	Наименее чувствительный
Смесь двух нормальных распределений	Андерсона-Дарлинга	Колмогорова
Смесь нормального распределения и Коши	Андерсона-Дарлинга	Колмогорова
Смесь нормального и равномерного распределения	Хи-квадрат Пирсона	Крамера-Мизеса
Распределение Стьюдента	Хи-квадрат Пирсона	Колмогорова
Обобщённое гауссовское распределение уменьш.	Андерсона-Дарлинга	Колмогорова
Обобщённое гауссовское распределение увелич.	Хи-квадрат Пирсона	Крамера-Мизеса



Выводы

В задаче проверки гипотезы согласия наилучшим выбором является критерий хи-квадрат. Однако если мы имеем дело с задачами повышенного требования к чувствительности, то рекомендуется использовать критерий Андерсона-Дарлинга. Рекомендуемый размер выборки не менее 200 элементов.