БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ
Кафедра математического моделирования и анализа данных

Журавский Вячеслав Сергеевич

Использование ИТ в оценке параметров бинарной выборки

Научный руководитель
Лобач Виктор Иванович

Введение

На практике очень часто возникает потребность оценки каких-либо статистических параметров на основе полученных данных.

На современном этапе существует ряд программных средств от языков программирования до готовых программных пакетов, которые позволяют получить такие статистические оценки.

Программное обеспечение

Математические пакеты:

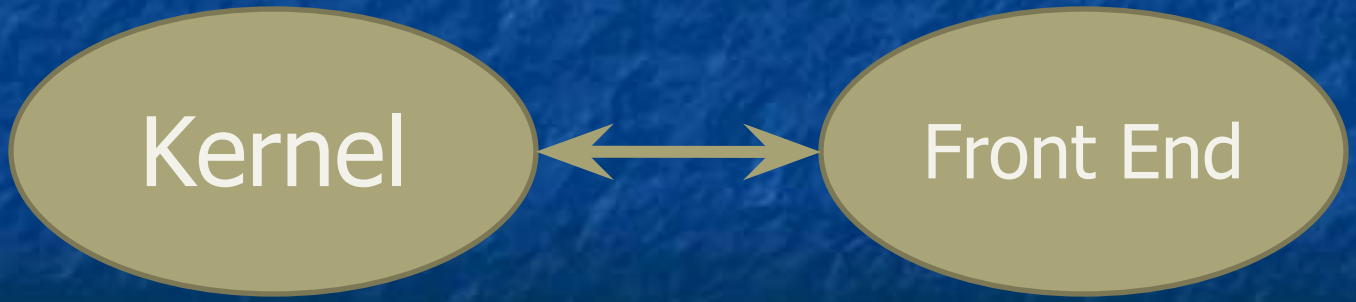
- Matlab
- Mathematica
- Mathcad
- Maple

Статистические пакеты

- Statistica
- EViews

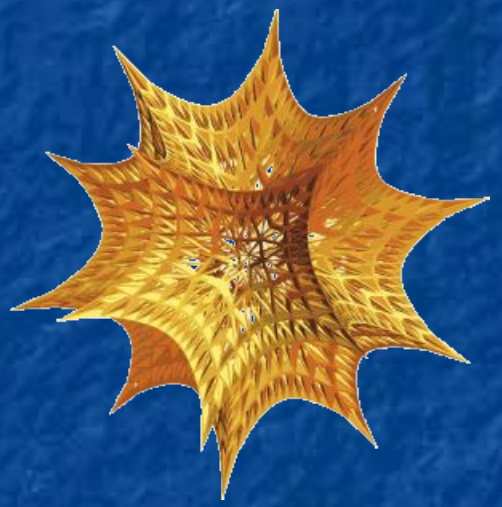
Пакет Mathematica

Mathematica относится к системам компьютерной алгебры. Центральное место в системах класса Mathematica занимает машинно-независимое ядро математических операций — Kernel. Для ориентации системы на конкретную машинную платформу служит программный интерфейсный процессор Front End.

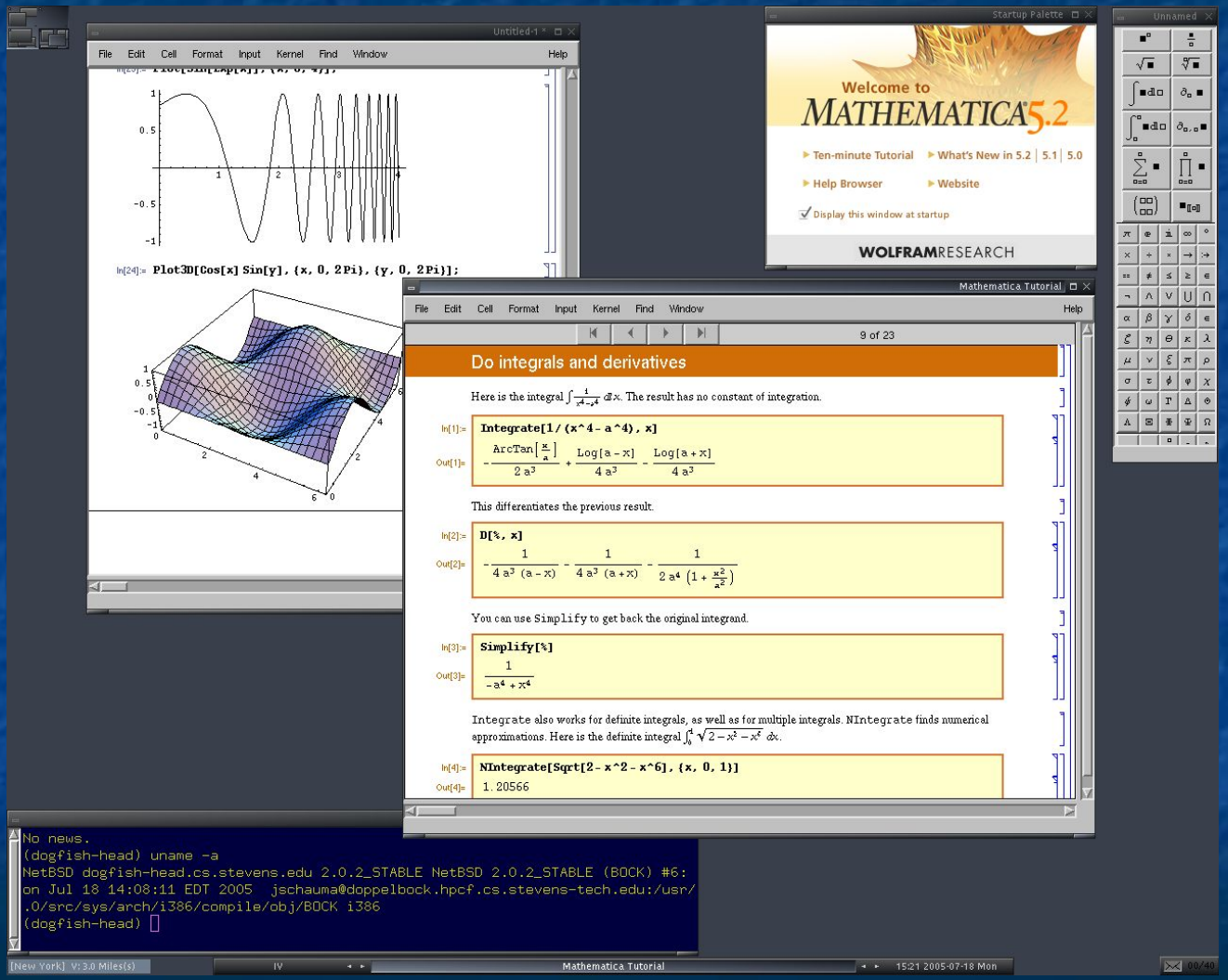


Возможности Mathemetica

- Аналитические преобразования
- Численные расчёты
- Теория чисел
- Линейная алгебра
- Графика и звук



Интерфейс Mathemetica



STATISTICA

STATISTICA — пакет для всестороннего статистического анализа, разработанный компанией StatSoft. В пакете STATISTICA реализованы процедуры для анализа данных (data analysis), управления данными (data management), добычи данных (data mining), визуализации данных (data visualization).

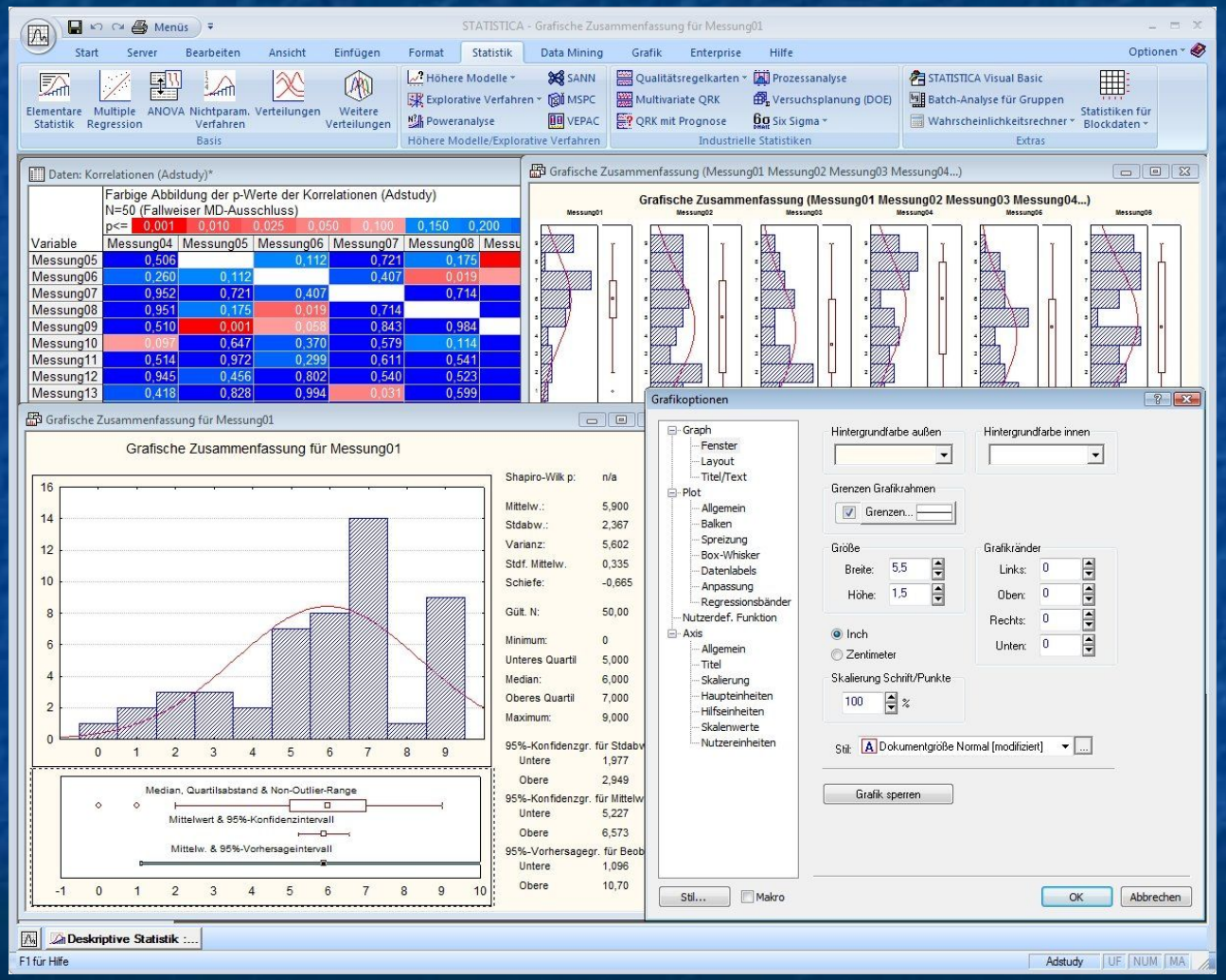
Программа вычисляет практически все используемые описательные статистики общего характера. Практически все описательные статистики и графики могут быть построены для данных, категоризованных (сгруппированных) по значениям одной или нескольких группирующих переменных.

Пакет STATISTICA имеет модульную структуру. Каждый модуль содержит уникальные процедуры и методы анализа данных.

Модули STATISTICA

Base — включает в себя обширный выбор основных статистик, широкий набор методов для разведочного анализа.
Advanced Linear/Non-Linear Models — предлагает широкий спектр линейных и нелинейных средств моделирования, регрессионный анализ, анализ компонент дисперсий, анализ временных рядов и т. д.
Multivariate Exploratory Techniques — многомерные разведочные технологии анализа *STATISTICA* предоставляет широкий выбор разведочных технологий
QC — Контроль качества — предоставляет широкий спектр аналитических методов управления качеством, а также контрольные карты презентационного качества.
Neural Networks — (отдельный модуль) единственный в мире программный продукт для нейросетевых исследований, полностью переведенный на русский язык
Data Miner — интеллектуальный анализ данных

Интерфейс STATISTICA



EM-алгоритм

EM-алгоритм — очень общий итеративный алгоритм для МП-оценивания в задачах с неполными данными.

В EM-алгоритме формализована относительно старая идея обработки неполных данных:

- заполнение пропусков оценками пропущенных значений
- оценивание параметров
- повторное оценивание пропущенных значений и параметров

и так далее до сходимости процесса.

EM-алгоритм для смесей

Первая выборка имеет стандартное нормальное распределение. Вторая имеет нормальное распределение с математическим ожиданием a и дисперсией σ.

$$p(x) = \lambda \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} + (1-\lambda)\frac{e^{-\frac{(x-a)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$

$$l(x,a,\sigma) = \sum_{i=1}^{N} \ln\left(\lambda \frac{e^{-\frac{x_i^2}{2}}}{\sqrt{2\pi}} + (1-\lambda)\frac{e^{-\frac{(x_i-a)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}\right)$$

Результаты моделирования

Параметры: $a = 2$, $\sigma = 0{,}5$, $\lambda = 0{,}5$. Вероятность пропуска: $p_{mis}=0{,}4$. Объем выборки $N=1000$.

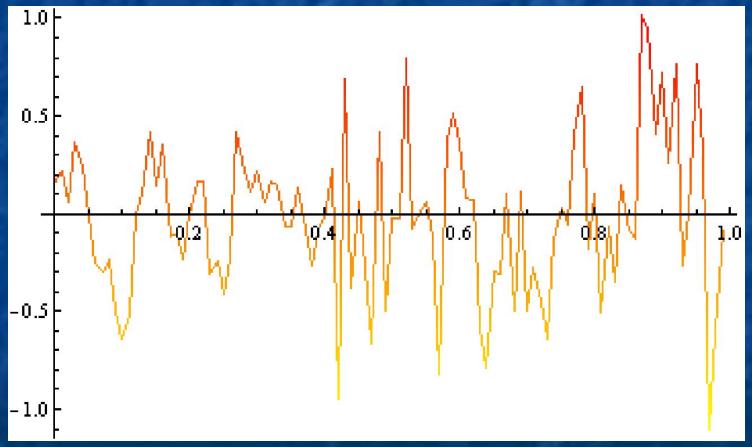
С использованием метода максимального правдоподобия и EM-алгоритма были получены следующие оценки.

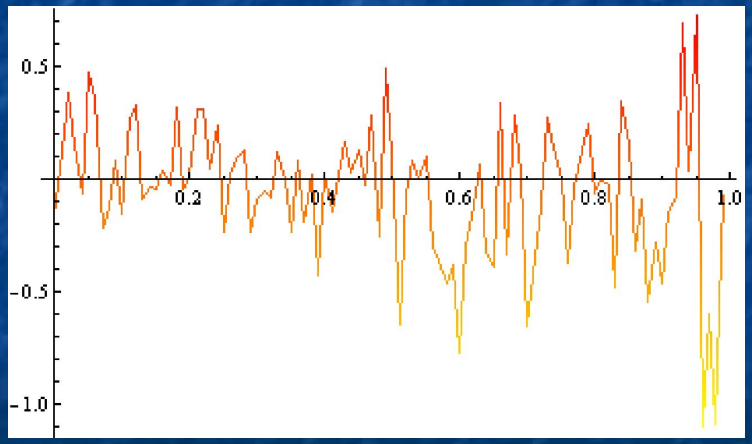
ММП	**EM-алгоритм**
$\hat{a} = 2.0312$	$\hat{a} = 2.01901$
$\hat{\sigma} = 0.504299$	$\hat{\sigma} = 0.500224$

Зависимость от количества пропусков

Объем выборки $N = 400$. Вероятность p_{mis} изменялась от 0 до 0,99 с шагом 0,01.

a ***σ***





Заключение

- Были изучены возможности различных математических пакетов.
- В качестве основного инструмента при выполнении поставленной задачи был выбран пакет Mathematica.
- ММП дает неплохие оценки, однако для оценки параметров выборок с пропусками следует использовать EM-алгоритм.
- с увеличением количества пропусков точность EM-алгоритма падает, однако не так сильно, как ожидалось.

Список использованных источников

1. В. Дьяконов. Mathematica 5/6/7. Полное руководство. Минск, 2009
2. В. Дьяконов. Mathematica 5.1/5.2/6 в математических и научно-технических расчетах. Минск, 2008
3. А. Халафян. Statistica 6. Статистический анализ данных. Минск, 2008
4. В. Боровиков. STATISTICA. Искусство анализа данных на компьютере. Минск, 2003

Спасибо за внимание!