

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

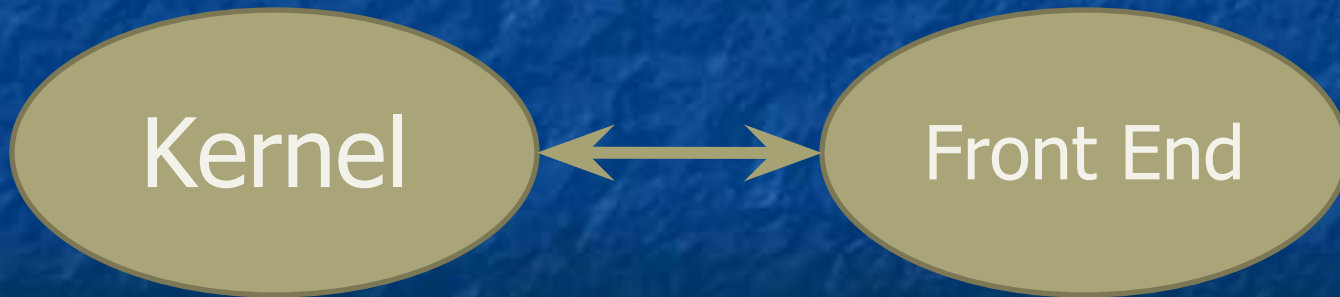
Статистические пакеты

- Statistica
- EViews



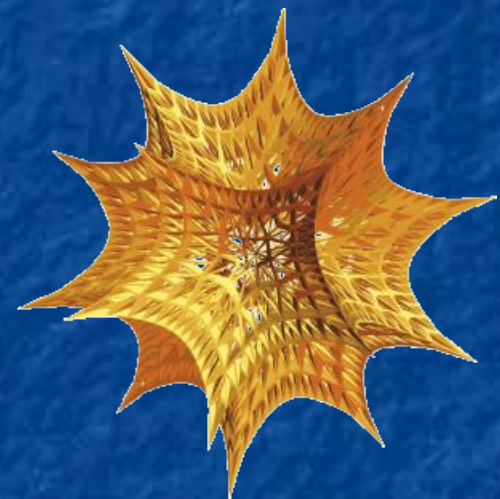
Пакет Mathematica

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 Mathematica

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



Интерфейс Mathematica

The screenshot displays the Mathematica 5.2 desktop environment. The main window shows a 3D plot of a surface defined by the equation $z = \cos(x) \sin(y)$ over the domain $x \in [0, 2\pi]$ and $y \in [0, 2\pi]$. The plot is rendered with a blue and purple color scheme. Below the plot, the input cell contains the command: `In[24]:= Plot3D[Cos[x] Sin[y], {x, 0, 2Pi}, {y, 0, 2Pi}]`.

A "Startup Palette" window is open, displaying the Mathematica 5.2 logo and navigation links: "Ten-minute Tutorial", "What's New in 5.2 | 5.1 | 5.0", "Help Browser", and "Website". A checkbox for "Display this window at startup" is checked. The Wolfram Research logo is at the bottom.

A "Mathematica Tutorial" window is open, showing a section titled "Do integrals and derivatives". The text reads: "Here is the integral $\int \frac{1}{x^4 - a^4} dx$. The result has no constant of integration." The input cell contains: `In[1]:= Integrate[1/(x^4 - a^4), x]`. The output cell shows the result: `Out[1]:= $-\frac{\text{ArcTan}\left[\frac{x}{a}\right]}{2a^3} + \frac{\text{Log}[a-x]}{4a^3} - \frac{\text{Log}[a+x]}{4a^3}$` . The text continues: "This differentiates the previous result." The input cell contains: `In[2]:= D[%, x]`. The output cell shows the derivative: `Out[2]:= $-\frac{1}{4a^3(a-x)} - \frac{1}{4a^3(a+x)} - \frac{1}{2a^4\left(1 + \frac{x^2}{a^2}\right)}$` . The text then says: "You can use Simplify to get back the original integrand." The input cell contains: `In[3]:= Simplify[%]`. The output cell shows the simplified result: `Out[3]:= $\frac{1}{-a^4 + x^4}$` . The text then states: "Integrate also works for definite integrals, as well as for multiple integrals. NIntegrate finds numerical approximations. Here is the definite integral $\int_0^1 \sqrt{2-x^2-x^6} dx$." The input cell contains: `In[4]:= NIntegrate[Sqrt[2-x^2-x^6], {x, 0, 1}]`. The output cell shows the numerical result: `Out[4]:= 1.20566`.

A terminal window at the bottom left shows the following text:

```
No news.  
(dogfish-head) uname -a  
NetBSD dogfish-head.cs.stevens.edu 2.0.2_STABLE NetBSD 2.0.2_STABLE (BOCK) #6:  
on Jul 18 14:08:11 EDT 2005 jschauma@doppelbock.hpccf.cs.stevens-tech.edu:/usr/  
.0/src/sys/arch/i386/compile/obj/BOCK i386  
(dogfish-head) █
```

The system tray at the bottom shows the location "[New York] V: 3.0 Miles(s)", the user "IV", the window title "Mathematica Tutorial", and the date and time "15:21 2005-07-18 Mon".



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

STATISTICA - Grafische Zusammenfassung für Messung01

Start Server Bearbeiten Ansicht Einfügen Format Statistik Data Mining Grafik Enterprise Hilfe Optionen

Elementare Statistik Multiple Regression ANOVA Basis Nichtparam. Verfahren Verteilungen Weitere Verteilungen Höhere Modelle Explorative Verfahren Höhere Modelle/Explorative Verfahren SANN MSPC VEPAC Qualitätsregelkarten Multivariate QRK QRK mit Prognose Prozessanalyse Versuchsplanung (DOE) Six Sigma STATISTICA Visual Basic Batch-Analyse für Gruppen Wahrscheinlichkeitsrechner Statistiken für Blockdaten Extras

Daten: Korrelationen (Adstudy)*

Farbige Abbildung der p-Werte der Korrelationen (Adstudy)
N=50 (Fallweiser MD-Ausschluss)

Variable	Messung04	Messung05	Messung06	Messung07	Messung08	Messung09
Messung05	0.506	0.112	0.050	0.100	0.175	0.175
Messung06	0.260	0.112	0.407	0.407	0.019	0.019
Messung07	0.952	0.721	0.407	0.714	0.714	0.714
Messung08	0.951	0.175	0.019	0.714	0.714	0.714
Messung09	0.510	0.001	0.058	0.843	0.984	0.984
Messung10	0.097	0.647	0.370	0.579	0.114	0.114
Messung11	0.514	0.972	0.299	0.611	0.541	0.541
Messung12	0.945	0.456	0.802	0.540	0.523	0.523
Messung13	0.418	0.828	0.994	0.031	0.599	0.599

Grafische Zusammenfassung (Messung01 Messung02 Messung03 Messung04...)

Grafische Zusammenfassung für Messung01

Shapiro-Wilk p: n/a

Mittelw.: 5,900
Stdabw.: 2,367
Varianz: 5,602
Stdf. Mittelw.: 0,335
Schiefe: -0,665
Gült. N: 50,00
Minimum: 0
Unteres Quartil: 5,000
Median: 6,000
Oberes Quartil: 7,000
Maximum: 9,000

95%-Konfidenzgr. für Stdabw.
Untere: 1,977
Obere: 2,949

95%-Konfidenzgr. für Mittelw.
Untere: 5,227
Obere: 6,573

95%-Vorhersagegr. für Beob.
Untere: 1,096
Obere: 10,70

Median, Quartilsabstand & Non-Outlier-Range
Mittelwert & 95%-Konfidenzintervall
Mittelw. & 95%-Vorhersageintervall

Grafikoptionen

- Graph
 - Fenster
 - Layout
 - Titel/Text
- Plot
 - Allgemein
 - Balken
 - Spreizung
 - Box-Whisker
 - Datenlabels
 - Anpassung
 - Regressionsbänder
 - Nutzerdef. Funktion
- Axis
 - Allgemein
 - Titel
 - Skalierung
 - Haupteinheiten
 - Hilfseinheiten
 - Skalenwerte
 - Nutzerseinheiten

Hintergrundfarbe außen: [Dropdown]
Hintergrundfarbe innen: [Dropdown]

Grenzen Grafikrahmen: Grenzen... [Dropdown]

Größe: Breite: 5,5; Höhe: 1,5

Grafikränder: Links: 0; Ober: 0; Rechts: 0; Unter: 0

Skalierung Schritt/Punkte: 100 %

Stil: [Dropdown: A Dokumentgröße Normal (modifiziert)]

Grafik sperren

Stil... Makro OK Abbrechen

Deskriptive Statistik ...

F1 für Hilfe Adstudy [UF] [NUM] [MA]



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

$$\hat{a} = 2.0312$$

$$\hat{\sigma} = 0.504299$$

EM-алгоритм

$$\hat{a} = 2.01901$$

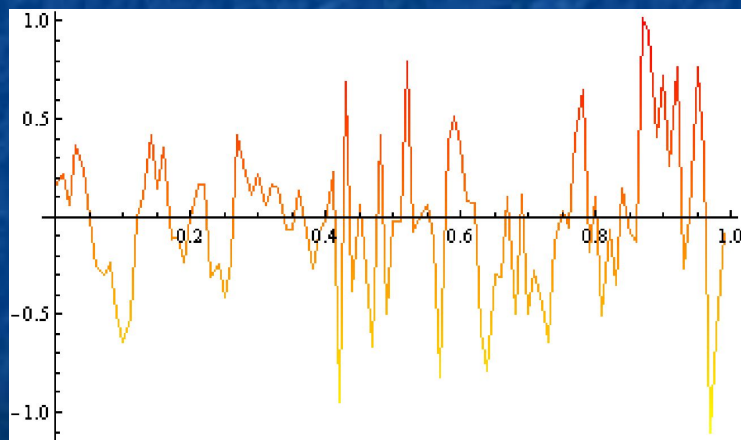
$$\hat{\sigma} = 0.500224$$



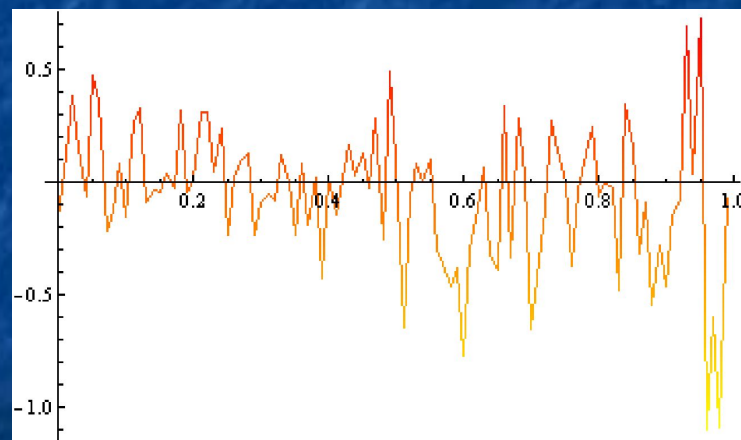
Зависимость от количества пропусков

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!

