



Methods & Models for Multivariate Data Analysis

Introduction

Ass. Prof. Anna Kalyuzhnaya, PhD

Instructional Staff

Lecturer



Anna V. Kalyuzhnaya, PhD

lectures, seminars, labs, coursework assessment, exam

kalyuzhnaya.ann@gmail.com

<http://en.hpc.ifmo.ru/staff/30/anna-kalyuzhnaya>

Office 310a, Birzhevaya line, 16

Office hours: 17:00-19:00 Tuesday, Friday

Assistants (workshops, labs)



Nikolay Nikitin, PhD



Alexander Hvatov, PhD



Irina Deeva, PhD student



Anna Bubnova, PhD student



Aim of the course

The purpose of this course is to provide students extended overview of the field of multivariate statistical methods and models.

This course provides a strong foundation for further study in the subject, as well as developing concepts which are relevant in a wide range of other subjects.

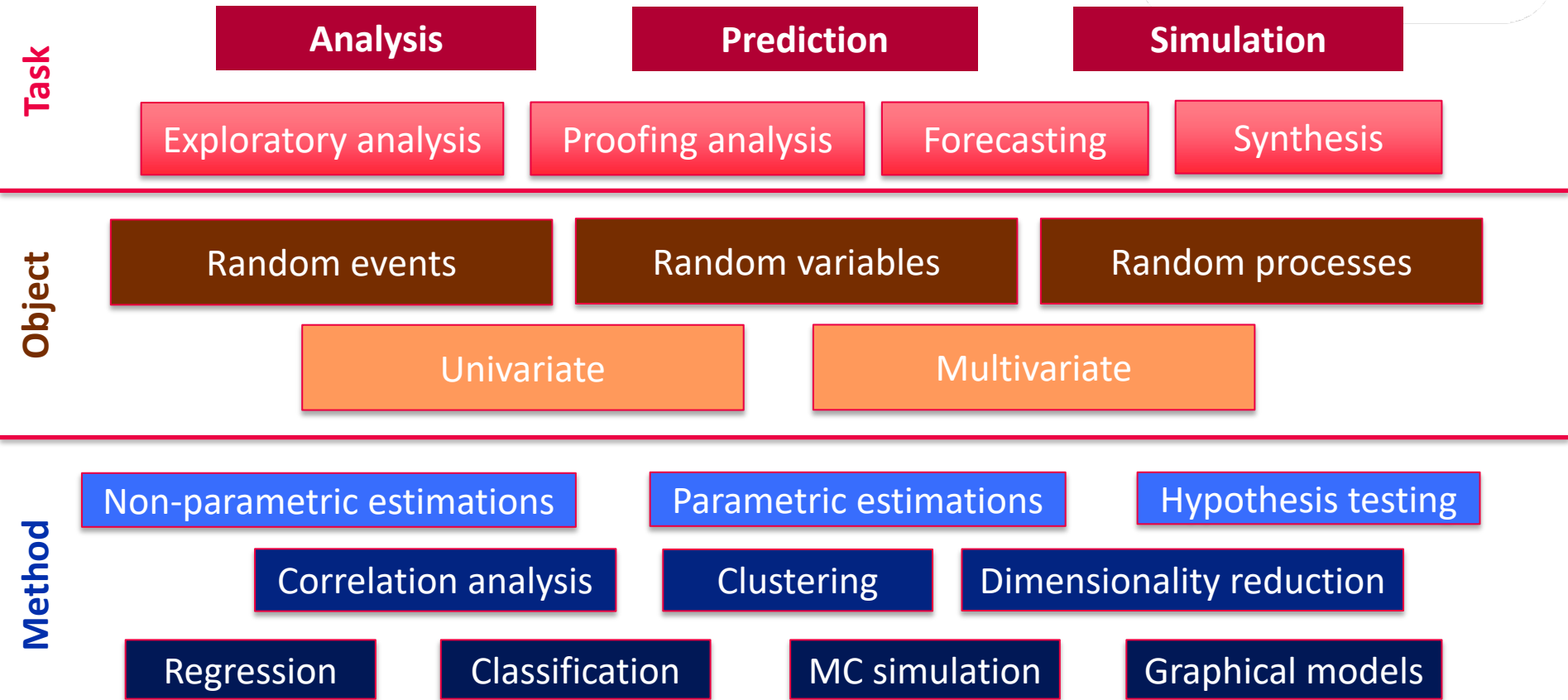


The course objectives

By the end of this course, students will be hopefully better able:

- to improve backgrounds in probability theory
- to understand on the fly and discuss the essence new statistical methods and techniques
- to understand the types of questions that the statistical methods and techniques addresses
- to develop skills in probabilistic modelling of random variables and random processes
- to develop skills in data analysis and ability of using these skills for solving real-life tasks
- to develop skills in statistical assessment of data quality and simulation results quality

Mind Map of the Course



Determinism vs. Uncertainty

Or what is understood as probability?

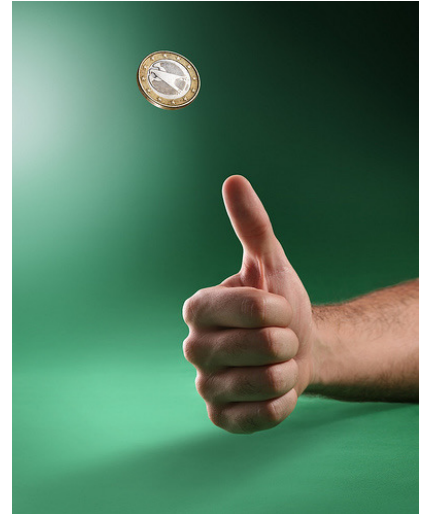
Probability definition

Probability (common definition) – quantitative assessment of a possibility of occurrence of an event at random experiment.

If the experiment is not random, then the possibility of occurrence of the event can be estimated by writing equations taking into account all influencing factors (**determinism**).

In real life, it is impossible to take all factors into account, that is, there is always **uncertainty** in the assessment.

Probability – an uncertainty measure.



What is the difference between probability theory and mathematical statistics courses?

Probability theory – section of mathematics, that investigate and create methods for description of stochastic phenomena via probabilistic models.

Mathematical statistics – section of applied mathematics, that investigate approaches for data analysis based on known probabilistic models or ways for identification of probabilistic models for real data.

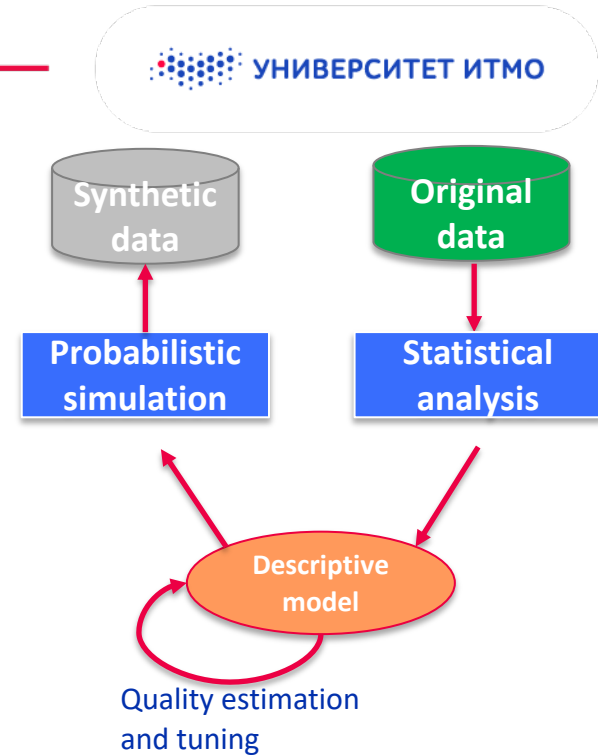
Probabilistic modelling & simulation

Modelling – construction / creation of descriptive model by means of probabilistic analysis

[Построение описательной модели (задача вероятностного анализа)]

Simulation –generative algorithm on the base of descriptive probabilistic models.

[Построение воспроизводящего алгоритма на основе описательной модели (задача вероятностного имитационного моделирования)]





Exploratory analysis includes the transformation of obtained data and means by which they can be visualized to identify the internal patterns that appear in this data.

Confirmatory analysis includes a toolkit of statistical methods of parameter estimation and hypothesis testing.

Basic concepts of multidimensional data

Multidimensional data – description of the same object in a variety of ways

Example:

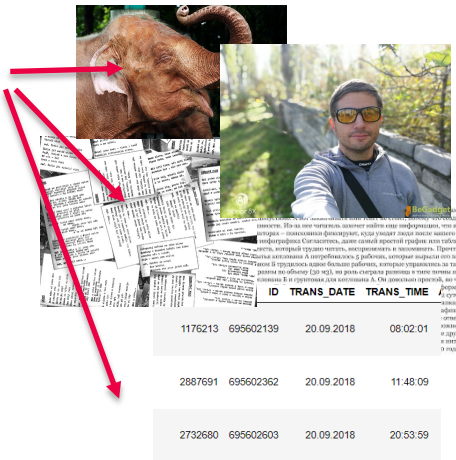
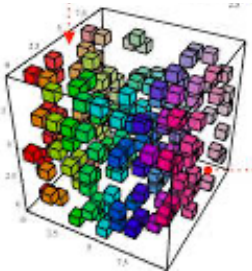
Obvious advantages:

Increasing of potentially useful and related information about object



Obvious (and not) disadvantages:

- rapidly growing volumes of data (dimension damnation);
- necessity of new methods of processing, analysis and forecasting (methods of machine learning);
- even larger volumes of data are necessary for effective ML models training (or rather increase in their density);
- rich information is usually put into different types of data, forming a composite object which hard to describe with basic probabilistic models.



ID	TRANS_DATE	TRANS_TIME
1176213	695602139	20.09.2018 08:02:01
2887691	695602362	20.09.2018 11:48:09
2732680	695602603	20.09.2018 20:53:59

Motivation: data are different

Nominal

Ordinal

Numerical

1. Names

2. Grades (ordered labels like beginner, intermediate, advanced)

3. Ranks (orders with 1 being the smallest or largest

2 the next smallest or largest, and so on)

4. Counted fractions (bound by 0 and 1)

5. Counts (non-negative integers)

6. Amounts (non-negative real numbers)

7. Balances (any real number)



УНИВЕРСИТЕТ ИТМО

Nominal



+7 (495) 500-55-50
8 (800) 555-55-50



Виталий Вячеславович А. Выход

Главная

Переводы и платежи

Карты

Пенсионные программы

Мои финансы



На этой странице Вы можете посмотреть структуру Ваших денежных средств на вкладах, картах и других продуктах, а также выполнить анализ расходов по разным категориям.

Расходы

Доступные средства

Календарь

Выбор периода

период 01/09/2015 по 16/01/2016

Все карты и наличные

Показать переводы

Показать снятие наличных

Показать комиссии



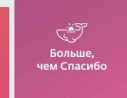
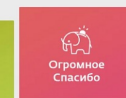
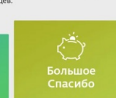
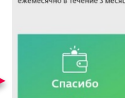
Всего списаний	1 235 823 руб.
39% Перевод с карты	486 970 руб.
22% Прочие расходы	274 881 руб.
15% Выдача наличных	187 373 руб.
13% Супермаркеты	162 657 руб.
5% Все для дома	61 991 руб.
5% Остальные категории списаний	61 951 руб.

Numerical

Ordinal

Уровни привилегий «Спасибо от Сбербанка»

Переходите на любой уровень привилегий, выполнив задание ежемесячно в течение 3 месяцев



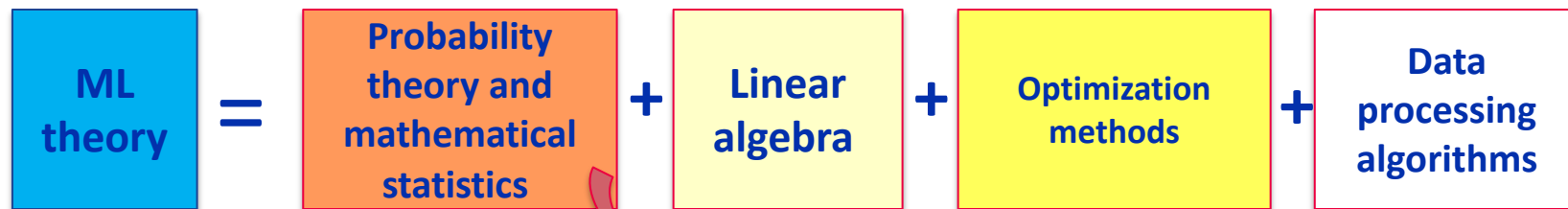
Бонусы от Партнеров

Бонусы от Сбербанка

Бонусы от Сбербанка

Бонусы от Сбербанка

What does ML consist of?



Let's start this semester...



- Случайное событие (вероятность P)
- Random event (probability P)

- Случайная величина (распределение $P(n)$ или $F(x)$)
- Random value (distribution $P(n)$ or $F(x)$)

- Многомерная случайная величина ($F(x, y, \dots)$)
- Multivariate random value ($F(x, y, \dots)$)

- Случайная последовательность, временной ряд
- Random sequence, time series

- Случайная функция или случайный процесс
- Stochastic function or stochastic process
- Случайная функция многих переменных
- Multidimensional stochastic function (or stochastic field)
- Многомерная случайная функция или поле
- Multivariate stochastic function or field
- Композитные вероятностные объекты (случайные графы, автоматы и пр.)
- Composite probabilistic objects (random graphs, automata etc.)

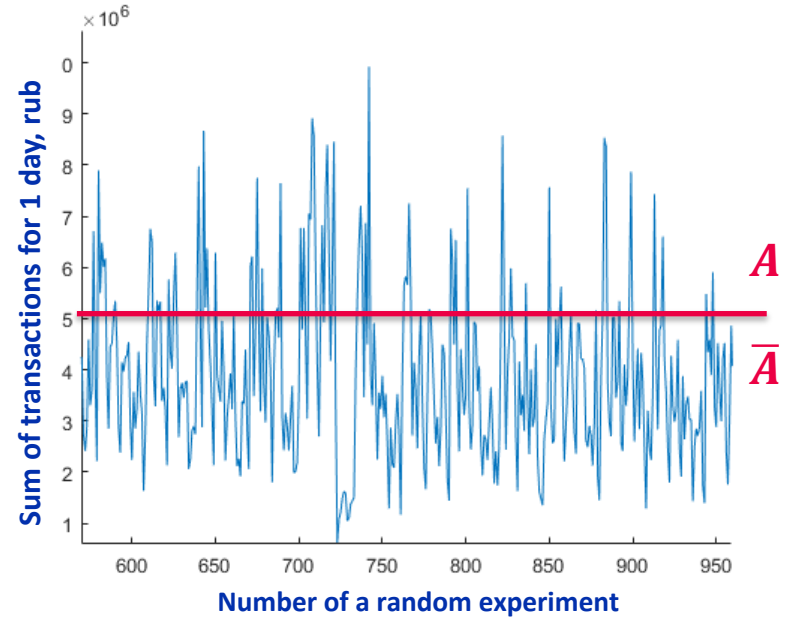
Basic concepts (1/2)

By means of what probabilistic abstractions it is possible to describe real data?

Random event – result of random experiment which will come with a certain probability.

Random variable – is a variable that takes on values from specified range.

Random process – variable that depends on not only it's position in a range but also on additional argument (usually, it is time).



Event A – sum > 5 million rubles

Event \bar{A} – sum < 5 млн

Basic concepts (1/2)

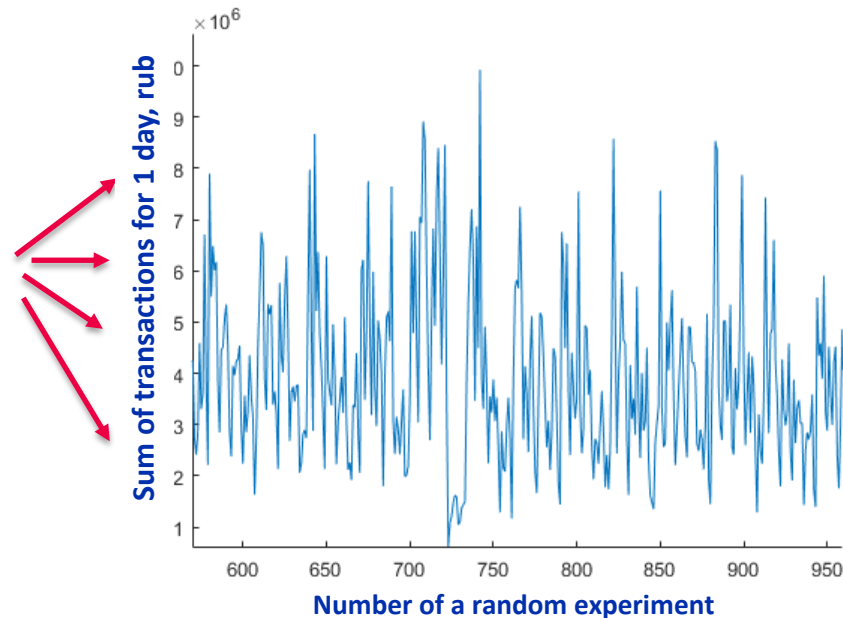
By means of what probabilistic abstractions it is possible to describe real data?

Random event – result of random experiment which will come with a certain probability.

Random variable – is a variable that takes on values from specified range.

Random process – variable that depends on not only it's position in a range but also on additional argument (usually, it is time).

Value of result of random experiment



Random value **X** – value of the sum of transactions in the experiment (1 day)

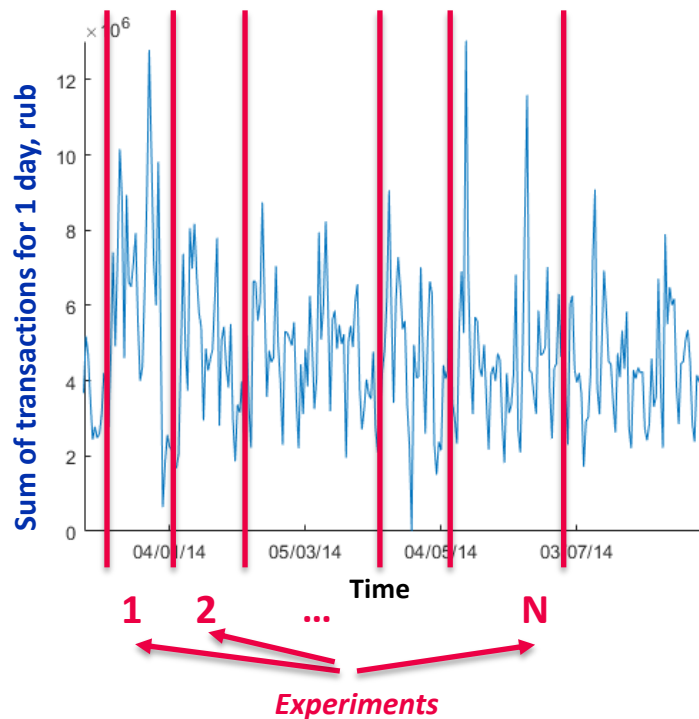
Basic concepts (1/2)

By means of what probabilistic abstractions it is possible to describe real data?

Random event – result of random experiment which will come with a certain probability.

Random variable – is a variable that takes on values from specified range.

Random process – variable that depends on not only it's position in a range but also on additional argument (usually, it is time).



Random process **R** describes results of series of experiments which values are connected in time (or space)

Thanks!

www.ifmo.ru

IT'sMO *re than a*
UNIVERSITY