СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ
SIBERIAN FEDERAL UNIVERSITY

Experimental data processing

Course Guide

Siberian Federal University

Experimental data processing

Course Guide

This course contributes to the requirements for the Degree of Candidate of Science in Computer Science.

Krasnoyarsk, 2020.

Contents

1. Course Description

This is a course, which contributes to the postgraduate program in 09.06.01 Information and computer Science

Courseperiod	1 semesters First semester: from October, the 1st to February, the 1st (18 weeks)
Studycredits	3 ECTS credits
Duration	108 hours
Languageofinstruction	English
Academicrequirements	– MSc degree in Computer Science or equivalent (transcript of records), – good command of English (certificate or other official document) **Prerequisites:** – **base knowledge ofapplied statistics; mathematical modeling ;information processes and systems;computer modelling;information theory,systems theory and systems analysis.**

1.1 Course overview

The course "Processing of experimental data" is intended for graduate students of engineering and other specialties, in which it is necessary to process and interpret the results of full-scale, simulation and other numerical experiments.

The content includes the following sections: data processing as a promising and important direction in data science; the main stages of the data processing process; types and models of data presentation and tasks of their processing; models and methods for Big Data Processing; numerical aggregation methods; Big time series and their processing and analysis; models and methods of uncertain data processing; Monte Carlo Method and Computational Probabilistic Analysis; reliable methods of data processing and analysis; knowledge discovery in data base (Data Mining technology and KDD), interactive simulation and visual technologies for data presentation and analysis.

The primary computing environment for the implementation of the investigated techniques, methods and algorithms is the software analytical platforms Deductor and Loginom.

The choice and use of software for studying the discipline is also oriented towards an individual approach depending on the wishes of the course participants, their scientific and practical interests and capabilities.

1.2 Special features

A characteristic feature of the course is the adaptation of its content to the tasks of a specific audience (that is, the number of certain sections of the course can be increased or decreased depending on the specific problems that postgraduates have when working on the dissertation material).

1.3 Course aims and objectives

Course Aims

The aim of the course is to study the theoretical foundations and the development the practical skills of work with the experimental data, as well as familiarity with modern computer information technologies of data processing, modeling, data analysis and extraction of knowledge and subsequent application to the solution of different research tasks respective areas of practical and scientific interest for post-graduate students

The course "The data processing" is intended for graduate students of technical and other fields where it is necessary to carry out the processing and interpretation of the results of full-scale, simulation, numerical and other types of experiments.

Course Objectives

- form a graduate student at the idea of the modern information technologies and computer processing of experimental data;
- introduce the main methods of computational mathematics, used for computer modeling and data processing;
- based on the study of a number of examples of applications solutions to form a graduate student at the skills of a scientific approach to the choice of methods and ways of working with the experimental data in the framework of the specific research problems;
- form at the post-graduate skills in the selection of his tasks adequate numerical methods of data processing and computing experiment;
- to acquaint graduate students with different data models and a variety of data processing tasks;
- to introduce the concepts and methods that take into account errors of direct and indirect measurement;
- to introduce the concept and methods of processing uncertain data;
- consider numerical methods for solving mathematical problems using simulation of random processes and events. Monte Carlo method;
- introduce modern information technologies of extraction and knowledge representation of these different volumes (technology Data Mining, Technology KDD, technology visual interactive simulation).

The main computing environment for the realization of the studied technologies, methods and algorithms is a software-analytical platform Deductor, software package SPSS. Selection and use of software tools for the study of the discipline also involves an individual approach, depending on the wishes of the course participants and their scientific and practical interests and opportunities.

1.4 Learning outcomes

By the end of the course, students will know theoretical foundations of methods for processing empirical information at all stages of data analysis in order to obtain the necessary knowledge about the object of research, as well as gain the necessary knowledge about modern computer systems and software analytical platforms.

By the end of the course, students will be able toapply the knowledge gained to analysis of professional information, highlight the main thing in it, structure, formalize and present in the form of analytical reviews and reports with substantiated conclusions and recommendations.

By the end of the course, students will havethe ability to data process and analyze the results of experiments, make the choice of optimal solutions, prepare and draft reviews, analytical reports and scientific publications; carry out computer analysis of data processing results; select data transformation models and modeling methods depending on their type and volume.

2. Course Lecturer, Contact Information

Olga A.Popova,
Ph.D. in Engineering, Docent, Associate Professor of the Department of Artificial Intelligence Systems,
School of Space and Information Technologies
Siberian Federal University
e-mail: OlgaArc@yandex.ru
Google Scholar page:
https://scholar.google.ru/citations?user=CyjzUW4AAAAJ&hl=ru
Additional information is available at:
http://ikit.sfu-kras.ru/e/115

Boris S. Dobronets,
Doctor of science in physics and mathematics (in computational modelling). Professorof the Department of Artificial Intelligence Systems, School of Space and Information Technologies Siberian Federal University

e-mail: BDobronets@yandex.ru
Google Scholar page:
https://scholar.google.ru/citations?user=ndB9WXsAAAAJ&hl=ru&oi=ao

Additional information is available at:
http://ikit.sfu-kras.ru/e/67

3. Prerequisites

A background in basic of data processing will help in faster and better understanding of every topic. Nevertheless, each part of the course includes a short introduction of methods that are required for its study. Therefore, a student without the denoted experience must be encouraged to make some additional efforts in education.

4. Course Outline

Week	Lectures	Seminars/ Assignments	Hours Lec/Lab/HA
1-2	Module 1 Topic 1.1. On the peculiarities of processing data of various types and volumes. The concept of the experimental data. Data types. Quantitative and qualitative data. The concept of measurement and measurement scale. Specifics of processing data depending on the amount of available information. The concept of big data. Application of modern data processing technologies to solve practical problems. Topic 1.2. Data uncertainties. Information uncertainty Information uncertainty, incomplete data, inaccurate data. Classification of uncertainty in the data. Models of uncertainty for the empirical data and classification of data	Creating a database on the basis of observations and data preprocessing. Primary analysis and study of the existing connections and relationships	4/8/14

	processing tasks. Topic 1.3. The main stages of data processing. Data cleansing, Data transformation, Data		
3-4	Module 2 Topic 2.1. Methods and models for processing large data volumes. Aggregation and Distributed Computing. Big Data Transformation. Models and methods. What is aggregation? Distributed Computing. Topic 2.2. Numerical methods for processing Big Time Series. Time series of distributions. About big time series. Practical tasks. What are time series of distributions? Topic 2.3. Statistical modeling methods and computational probabilistic analysis.	Models and numerical methods for solving practical problems on the basis of experimental data.	4/10/22
5-17	Topic 3.1. Data Mining. Text Mining. KDD technology. Topic 3.2. The technology of interactive visual modeling of multidimensional data.	Information processing technology, simulation and knowledge extraction from data.	6/10/30
18	**FinalCredit**		14/28/66

4.1 Course requirements

4.1.1 Web-pageofthecourse

Course materials and required reading materials are available on the course webpage “Experimental data processing”. The webpage is available through the SibFU E-learning portal www.e.sfu-kras.ru. You must be logged in to access this course. https://e.sfu-kras.ru/enrol/index.php?id=9071.

4.1.2 Required reading

List of training and methodological support for the independent work of graduate students in disciplines (modules)

1. Dobronets BS, Popova OA Numerical probabilistic uncertainty analysis of data: a monograph. - Krasnoyarsk: Sib. Feder. University Press, 2014. - 167 p.
2. Popova OA Data Management [Text]: ucheb method. Special allowance for students. 230201.65 "Information systems and technologies" / Sib. Fed. Univ; comp. OA Popova. - Krasnoyarsk: SFU, 2012. - 75 p. - Bibliography .: p. 75.

4.1.3 Course materials

The main book that will guide a student through the course is ***Numerical probabilisticanalysisof uncertainty data***book. It contains all of topics of this course according to the schedule. It will provide you with useful links at the end of each chapter that will help students to improve their understanding of the topics.

4.1.4 Required feedbacks

Students are free to contact the lecturer by email. The name of department and a number of a group should be written in the subject or in the beginning of the letter for convenience. More information on how to contact the lecturer can be found in «Lecturer information» section of this Guide.

Student’s Home or Lab Assignment reports must be attached as a separate pdf file. Student’s name and group number should be written on the first page of the file. Students send this report in electronic form only before the deadline.

4.2 Course Structure

LearningActivities	**Hours**
Lectures	14
Practicesessions / Seminars,	28
Self-studyAssignments	66
FinalCredit (includingpreparation)	42
Totalstudyhours	108

4.3 Time schedule of the course and course outline

№	*Theme*	*Week*	*Learning Activities*	*Hours*	*Home Assignment and Reading*
1	Module 1. Data processing as a promising and important area of Data Science.	1-4	**Lecture 1** ”On the peculiarities of data processing and overview various data types and data volumes”.	2	Course Book Chapter 1. Overview of data processing. Answer the questions on the topic in the e-course
			Lecture 2 ”Data uncertainties. Information uncertainty. The main stages of data processing”	2	Course Book Chapter 2. Overview of data uncertainty. Answer the questions on the topic in the e-course
			Lab 1 ”Creating a database on the basis of observations and data preprocessing”.	4	Select the task 1 from the collection of task book and follow in accordance with its content. See examples and directions on how to complete your task.
			Lab 2” Primary analysis and study of the existing connections and relationships”.	4	Select the task 2 from the collection of task book and follow in accordance with its content. See examples and directions on how to complete your task.
			Home	26	Watch presentations1, 2 and

			assignment 1		read additional resources posted on the course. Complete Labs 1 and 2. Make a report on the work done.
2	Module 2: Theoretical foundations of numerical modeling and analysis of empirical information	3-4	**Lecture 3** ”Methods and models for processing large data volumes. Aggregation and Distributed Computing. Big Data Transformation. Models and methods. What is aggregation? Distributed Computing.”	2	Course Book: Chapter 3. Overview of the methods and models for processing large data volumes . Answer the test questions on the topic in the e-course.
			Lecture4 ”Numerical methods for processing Big Time Series. Time series of distributions. About big time series. Practical tasks. What are time series of distributions. Statistical modeling methods and computational	2	Course Book: Chapter 4. Answer the questions on the topic in the e-course.

			probabilistic analysis.”		
			Lab3 ”Models and numerical methods for solving practical problems on the basis of experimental data”	10	Select the task3 from the collection of task book and follow in accordance with its content. See examples and directions on how to complete your task.
			Home assignment 2	10	Watch presentations 3-4 and read additional resources posted on the course. Complete Labs 3. Make a report on the work done.
3	Module 3. Technologies of the knowledge discovery in data base and software packages for data presentation, processing, modeling and analysis.	5-17	**Lecture**5 «Data Mining. Text Mining. KDD technology»	2	Course Book: Chapter 5. Answer the questions on the topic in the e-course.
			Lecture 6-7 «The technology of interactive visual modeling of multidimensional data»	4	Course Book: Chapter 6-7. Answer the questions on the topic in the e-course
			Lab 4 ”Information processing technology, simulation and knowledge extraction from data.”	10	

			Home assignment 3	30	Watch presentations 4 and read additional resources posted on the course. Complete Labs 4. Make a report on the work done.
5	Final credit			108	Prepare to final credit. Preparation for answering credit questions (available at e-courses and course book).

5. Assessment

Assessmentstrategy	Points, max	Evaluationcriteria
Tests	10	Test questions for lectures in the e-course
Lab works	40	Lab report
Individual Project	40	Individual database, models and methods, presentation and report about research results
Final credit	10	2 questions and a practical task that require preparatory reading and knowledge of the concepts explained

Grade policy for final assessment is:

A (credit) 61–100 points

B (failed work) <60 points

Program of the discipline in order to test the strength of assimilation of the material provides for a different form of control:

- Preliminary monitoring is necessary to determine the initial level of knowledge of students.

- Theme control determines the degree of assimilation by students of each section (themes in general), their ability to relate course material already acquired knowledge, to trace the development, the complexity of phenomena, concepts, main ideas.

- Form of control is test.

6. Attendance Policy

Students are expected to attend classes regularly. In case of missing an in-lab activity a student should perform additional work submitted to the instructor within a week after a class was missed.

Every topic involves an assignment. A written report on the assignment should be submitted within two weeks from the moment students received a list of problems. The final mark will rely on the same grading policy as for the final exam.

7. Required Course Participation

There are no special requirements for the course participation. The preferred type of report submission is the electronic one. Students can use the web-version of the course (link) for a better progress. All problems for solution could be found there together with text from the course book.

8. Facilities, Equipment and Software

Software:

MSOffice (MSWord, MSPowerPoint, MSExcel),

Adobe Acrobat, Adobe Flash Player or KM Player, Adobe Flash, Winamp.

Deductor, Loginom, SPSS

Laboratory equipment:

Control, testing and measuring equipment:

Annex 1 Example of Self-Study Assignment

The task:

There is a set of data from 100 values (points). 1) For this set of points, build a histogram and a frequency polygon. Use the following information: The boundaries of the data change (data variability) is [a,b].Step of mesh is h. Based on the constructed graphical models, it is necessary to draw a conclusion about the type of dependence of the initial data. 2) Express your assumptions (forecast) about the kind of functional dependence existing in the data. For example: linear dependence, quadratic dependence, cubic, and so on.

Solution:

For most of the work you do in this book, you will use a histogram to display the data. One advantage of a histogram is that it can readily display large data sets. A rule of thumb is to use a histogram when the data set consists of 100 values or more.

A histogram consists of contiguous (adjoining) boxes. It has both a horizontal axis and a vertical axis. The horizontal axis is labeled with what the data represents (for instance, distance from your home to school). The vertical axis is labeled either frequency or relative frequency (or percent frequency or probability). The graph will have the same shape with either label. The histogram (like the stemplot) can give you the shape of the data, the center, and the spread of the data.

The relative frequency is equal to the frequency for an observed value of the data divided by the total number of data values in the sample. (Remember, frequency is defined as the number of times an answer occurs.) If:

- f = frequency

•n = total number of data values (or the sum of the individual frequencies), and

•RF = relative frequency, For example, if three students in class of 40 students received from 90% to 100%, then $f = 3$, $n = 40$, then:

$$RF = \frac{f}{n} = \frac{3}{40} = 0.075.$$

7.5% of the students received 90–100%. 90–100% are quantitative measures.

To construct a histogram

First decide how many **bars** or **intervals**, also called classes, represent the data. Many histograms consist of five to 15 bars or classes for clarity. The number of bars needs to be chosen. Choose a starting point for the first interval to be less than the smallest data value. A **convenient starting point** is a lower value carried out to one more decimal place than the value with the most decimal places. For example, if the value with the most decimal places is 6.1 and this is the smallest value, a convenient starting point is 6.05 (6.1 – 0.05 = 6.05).

The following data are the heights (in inches to the nearest half inch) of 100 male semiprofessional soccer players. The heights are **continuous** data, since height is measured.

60; 60.5; 61; 61; 61.5 63.5; 63.5; 63.5 64; 64; 64; 64; 64; 64; 64; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5 66; 66; 66; 66; 66; 66; 66; 66; 66; 66; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5 68; 68; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69.5; 69.5; 69.5; 69.5; 69.5 70; 70; 70; 70; 70; 70; 70.5; 70.5; 70.5; 71; 71; 71 72; 72; 72; 72.5; 72.5; 73; 73.5 74

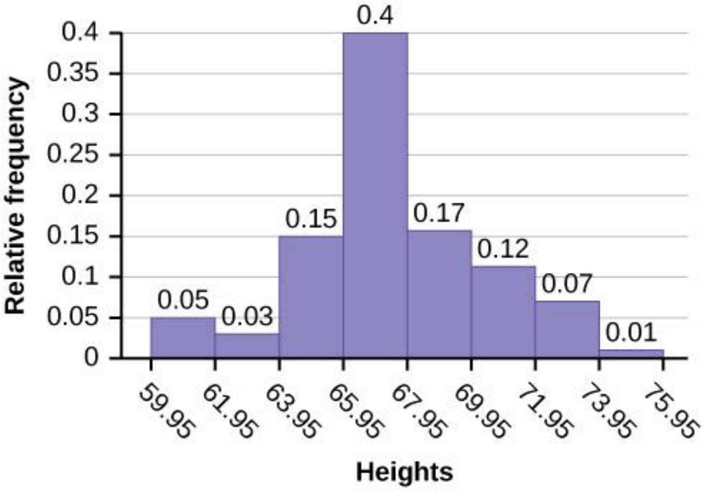
The smallest data value is 60. Since the data with the most decimal places has one decimal (for instance, 61.5), we want our starting point to have two decimal places. Since the numbers 0.5, 0.05, 0.005, etc. are convenient numbers, use 0.05 and subtract it from 60, the smallest value, for the convenient starting point. The starting point is, then, 60 – 0.05 = 59.95.

The largest value is 74, so 74 + 0.05 = 74.05 is the ending value.

Next, calculate the width of each bar or class interval. To calculate this width, subtract the starting point from the ending value and divide by the number of bars (you must choose the number of bars you desire). Suppose you choose eight bars. The boundaries are:

59.95; 61.95; 63.95; 65.95; 67.95; 69.95; 71.95; 73.95; 75.95.

The following histogram displays the heights on the x-axis and relative frequency on the y-axis.

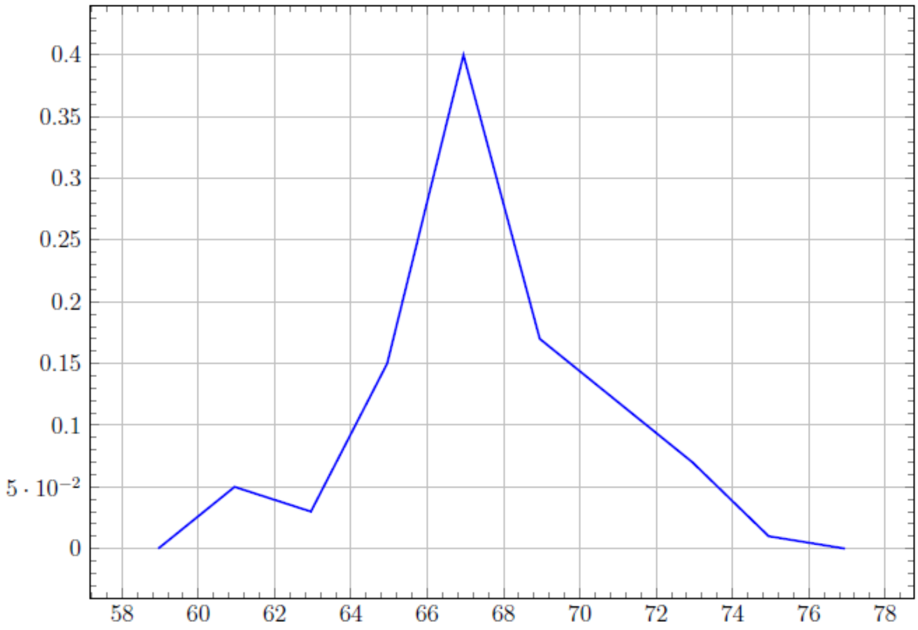


Frequency Polygons

Frequency polygons are analogous to line graphs, and just as line graphs make continuous data visually easy to interpret, so too do frequency polygons.

To construct a frequency polygon, first examine the data and decide on the number of intervals, or class intervals, to use on the x-axis and y-axis. After choosing the appropriate ranges, begin plotting the data points. After all the points are plotted, draw line segments to connect them.

A frequency polygon was constructed from the histogram.



As a conclusion, it can be argued that 1) the application of data transformation using a frequency polygon gives a more accurate picture for determining the functional dependence in the data than a histogram; 2) since the frequency polygon is a piecewise-linear model, and the histogram represents a piecewise-constant function, then the frequency polygon model is better suited for solving the problem of identifying the dependence in the data; 3) the graphs show that there is a non-linear relationship in the data; 4) as a forecast, you can make an assumption about a quadratic functional data model.

Annex 2 Example of Pre-Course Test Questions

1. What is the importance of information and intelligent technologies for the process of researching empirical data?
2. Expand the semantic meaning of the concepts of data, information, knowledge.
3. What do the concepts of measuring and measuring scales define?
4. List the main stages of data processing.
5. Expand the meaning of the concept of intelligent data processing.
6. What is undefined data? Give a classification of data uncertainties.
7. What models of uncertain data are used in the study of empirical information?
8. Data models and classification of processing tasks.
9. What is the purpose of the phase of data cleaning, transformation and aggregation.
10. What is processing, pre-processing and post-processing?
11. What are large time series?
12. What is Big Data?
13. The concept of incomplete, inaccurate and unreliable data.
14. Methods and models of data presentation.
15. What is data aggregation used for?
16. What is a large time series?
17. Features of building regression on big data.
18. Functional regression. Examples of practical assignments.
19. Expand the concept of statistical testing method. Monte Carlo method.
23. For what tasks is interpolation, extrapolation, smoothing, data applied?
24. Expand the content of concepts: KDD technology, visual interactive modeling technology.

25. Software-analytical platforms Deductor, Loginom. Stages of development, use and implementation in practice.
26. Determine the role and importance of using reliable methods and calculations for data processing.
27. Functional and symbolic data analysis. What are their practical and scientific achievements?

Annex 3 Outlines of Lab works

(List one. Title)

"SIBERIAN FEDERAL UNIVERSITY"
Institute of Space and Information Technologies
Department of Computer Science
Master's Programs "Digital Intelligent Control Systems"
Group number (group ID)

LABORATORY REPORT No. (Laboratory number)
Subject: (Subject of the task).

Tutor: (Name and surname of tutor / lecture).

Student: (Name and surname of the student).

Krasnoyarsk, 2020

(List two, etc. The progress)

Main aim: (Describe the aim of lab).

The task: (Describe the task of lab).

Solution: (short description (no more than 2-3 pages) of the problem solving process).

Annex A Data base

Annex B Diagram(s)

Annex 4 Example of Final Oral Exam Questions

1. Describe the main procedures for processing experimental data in the case of a large number of gaps.

2.What is aggregation? Give examples of the application of the big data aggregation procedure for practical tasks.

3. In accordance with Figure 1, describe the data processing algorithm and make a forecast.