COURSE UNIT (MODULE) DESCRIPTION

Course unit (module) title	Code
Introduction to Probability, Statistics, and Programming for Master Studies	

Lecturer(s)	Department(s) where the course unit (module) is delivered
Coordinator: assoc. prof. dr. Rūta Levulienė **Other(s):** assist. prof. dr. Andrius Buteikis, assoc. prof. dr. Jurgita Markevičiūtė, assist. prof. dr. Tomas Plankis, assoc. prof. dr. Viktor Skorniakov	Department of Statistical Analysis, Institute of Applied Mathematics, Faculty of Mathematics and Informatics

Study cycle	Type of the course unit (module)
Second	Selective

Mode of delivery	Period when the course unit (module) is delivered	Language(s) of instruction
Face-to-face	First Semester	English

Requirements for students	
Prerequisites:	**Additional requirements (if any):**

Course (module) volume in credits	Total student workload	Contact hours	Self-study hours
-	36	16	20

Purpose of the course unit (module): programme competencies to be developed		
The main goal of the module is to review the main concepts of probability and statistics and to explain important aspects of programming using R and Python software.		
Learning outcomes of the course unit (module)	**Teaching and learning methods**	**Assessment methods**
• To know probability and statistics concepts and models important for master studies in mathematics.	Lectures, case analysis.	-
• To be able to program in R and Python at the intermediate level.		

Content: breakdown of the topics	Contact hours						Self-study work: time and assignments	
	Lectures	Seminars	Exercises	Laboratory work	Internship/work placement	**Contact hours**	**Self-study hours**	**Assignments**
Part I. Introduction to probability for master studies.	4					**4**	**5**	
Part II. Introduction to statistics for master studies.	4					**4**	**5**	
Part III. Introduction to R for master studies.	4					**4**	**5**	
Part IV. Introduction to Python for master studies.	4					**4**	**5**	
Total	**16**					**16**	**20**	

DETAILS AND NOTES RELATED TO COURSE UNIT

Part I. Introduction to Probability for Master Studies

Coordinator: assoc. prof. dr. Viktor Skorniakov.

During the lectures, I plan to briefly walk through the main concepts along the way providing motivating examples designed to illustrate why the covered material is vital for Data Science and other master studies in mathematics. I also plan to devote some time to answering your questions.

Table 1 covers essential prerequisite material for the students entering the Data Science Master Study program. It is suggested at least to skim through the pointed subsections to get convinced that you feel comfortable with all the objects and concepts covered there. In doubt, take a deeper reading and accomplish the exercises posed at the end of [1], Chapter 6. In case you still feel unconfident, refer to [3], Chapters 2 – 5. The latter source covers all material presented in Table 1 in a more detailed fashion than [1] and [2]. Though being lengthy, it is written in a very accessible manner, appended by Python code, and full of well-designed exercises for self-practice. All books are available to download free of charge (just follow the link). Note that, in case of reference [2], you need to be logged on to the dedicated VPN since the subscription is paid and provided by Vilnius University.

Table 1. Suggested reading.

Theme	Suggested reading
Probability space	[1] – 6.1
Discrete and continuous random variables	[1] – 6.2; [2] – 2.2.1
Moments	[2] – 2.2.2 (also part of 2.2.3); [1] – 6.4.1 – 6.4.4
Joint, marginal, and conditional distributions	[2] – 2.2.3; [1] – 6.3
Independence	[1] – 6.4.5
Main probability distributions	[2] – 2.2.4; [1] – 6.5, 6.6 (except 6.6.2); [3] – 3.5.4, 4.5
Random variables' transformations	[1] – 6.7; [2] – 2.2.5
Information theory	[2] – 2.3

Literature

1. Deisenroth, Marc Peter, et al. *Mathematics for Machine Learning*. Cambridge University Press, 2020.
2. Jiang, Hui. *Machine Learning Fundamentals: A Concise Introduction*. Cambridge University Press, 2021.
3. Chan, Stanley H. *Introduction to Probability for Data Science*. Michigan Publishing, 2021.

Remark. [1], 6.8 provides additional sources for self-studying with comments regarding the extensiveness and the level of difficulty.

Part II. Introduction to Statistics for Master Studies

Coordinator: assoc. prof. dr. Jurgita Markevičiūtė

During the lectures, I intend to provide a concise overview of the main concepts while incorporating motivating examples that illustrate the importance of the covered material in the field of Data Science and Statistics. Additionally, I will allocate time to address any questions you may have.

Table 2 encompasses fundamental prerequisites for students entering the Data Science Master Study program. It is recommended to at least skim through the specified subsections to ensure that you are comfortable with the objects and concepts presented there. If you have any doubts, I encourage you to delve deeper into the reading and complete the exercises provided in reference [1]. If you still feel uncertain, you can refer to any introductory statistical book for Data Science, Social Sciences, or Engineering. Although it is a lengthy resource, it is written in an accessible manner, supplemented with Python or R codes, and includes well-designed exercises for self-practice. The recommended book can be downloaded free of charge by following the provided link.

Table 2. Suggested reading

Theme	Suggested reading
Exploratory data analysis	[1] – Ch. 1, Ch. 2
Correlation and Linear regression models	[1] – Ch. 3
Estimators and their properties	[1] – Ch. 12.1, 12.2, 14
Methods for estimators	[1] – Ch. 13, 15
Interval estimators	[1] – Ch. 16
Introduction to hypothesis testing	[1] – Ch. 17.1, 18.2, 18.3
Parametric hypothesis testing (t-test, Pearson correlation test, ANOVA tests)	[1] – Ch. 20.1 – 20.7, 22.2
Nonparametric hypothesis testing (Chi-squared, Mann-Whitney, Wilcoxon, Kruskal-Walis, Friedman, Spearman tests)	[1] – Ch. 19.3, 20.9, 22.5

Literature

1. Joseph C. Watkins. An Introduction to the Science of Statistics: From Theory to Implementation Preliminary Edition.

Note: For Lithuanian-speaking students you may use V. Čekanavičius & G. Murauskas Statistika ir jos taikymai I, II, and III.

Part III. Introduction to R for Master Studies

Coordinator: assist. prof. dr. Andrius Buteikis

During the lecture we will summarise the topics outlined in Table 1, which covers essential prerequisite material for students entering Master Study programs at Vilnius University, Faculty of Mathematics and Informatics. In preparation for the lecture, students should have R and RStudio installed and they should analyse the topics and the suggested reading material.
It is suggested to read through [1] and [6] for a quick rundown of the key components of R, while more advanced topics can be found in either [3] or [4]. Alternatively, various interactive and video tutorials (see the remarks section for a link to an interactive resource library within R itself) can be found online by searching using the topics below as keywords.

	Topic	Expected result	Reading
1.	R & RStudio setup	Ability to download, install, and setup R and RStudio on your laptop, basic navigation, and R package installation inside RStudio.	[1] – Ch. 1
2.	Data types and data structures (vectors, arrays, matrices, data.frames, lists)	Ability to create various objects in R, carry out basic mathematical operations, as well as extract data from these objects using indexes and the '$' operator.	[1] – Ch. 2 & 3, [4] – Ch. 13 to 20
3.	Control statements (if/else, switch)	Ability to use control statements to compare different R object values and carry out different calculations based on the comparison results.	[1] – Ch. 7.3 & 7.4, [6], [4] – Ch. 19.4
4.	Loops (for, while, apply, etc.)	Ability to write various loops to iterate, print, and modify different R objects.	[1] – Ch. 7.5, [6], [4] – Ch. 21
5.	Functions	Ability to write and debug own/custom functions.	[1] – Ch. 7.2, [6]
6.	Reading data (.txt, .csv, .xlsx)	Ability to read data files as R objects, ability to save data as R objects and text files.	[1] – Ch. 3.3, [6], [4] – Ch. 21
7.	R package collection 'tidyverse'	Ability to use the 'pipe' operator and familiarity with select packages from the 'tidyverse' R package collection. Additionally, familiarity with the 'data.table' is recommended.	[3] – Ch. 12 & 18, [4] – Ch. 4.4, [5]
8.	Data visualization	The ability to visualize data using 'ggplot2' and 'patchwork' libraries	[1] – Ch. 5 & [2], [4] – Ch. 10 to 12
9.	Parallel computing	Ability to vectorize and run calculations in parallel.	[7]

Table 1. Topics and suggested reading.

Literature

1. *Alex Douglas, et.al.* An Introduction to R*, 2023.*
2. Thomas Lin Pedersen, patchwork: The Composer of Plots , an R library, 2022.
3. Hadley Wickham, et. Al. R for Data Science, 1st Edition, 2017.
4. Hadley Wickham, et. Al. R for Data Science, 2nd Edition, 2023.
5. Hadley Wickham, Differences between the base R and magrittr pipes, 2023.
6. R for Statistics, Data Science, www.learn-r.org, 2022.
7. R parallel packages: doParallel, doSNOW and parallel.

Remarks

- An interactive tutorial inside RStudio is available with the swirl library.
- Note that while [4] is a newer edition of [3], it uses the newly introduced pipe operator in the base installation of R, while [3] uses a pipe operator, which is defined in a separate library. The differences between these operators are discussed in [5].
- When working with large datasets, computation time becomes quite important, and it might be challenging to select an appropriate library for data analysis. A good resource to compare the speed of different computations using various libraries in R and Python can be found at https://duckdblabs.github.io/db-benchmark/ .

Part IV. Introduction to Python for Master Studies

Coordinator: assist. prof. dr. Tomas Plankis

During the lectures, I plan to explain the main concepts of programming along the way providing some examples designed to illustrate why programming is vital for Data Science. I will also answer your questions.

Table 1 covers essential prerequisite material of Programming in Python for the students entering the Data Science Master Study program. It is useful to skim through the pointed sections and subsections to get comfortable with all the concepts covered there. For a deeper understanding, you can read more in [2]. In case you still feel unconfident or prefer reading books, you can find books online, for example [11]. If you feel like doing some exercises for self-practice, use [12]. At the moment everything is reachable online and free of charge.

Table 1. Suggested reading

Theme	Suggested reading
Data types, conditionals, loops	[1] – 3, 4.1 - 4.6, 5.1.3 – 5.1.4, 5.3 - 5.7, 7
Functions and modules	[1] – 4.7 - 4.8
Object-oriented programming	[1] – 9 (additional reading 8)
Overview of libraries for data analysis, main principles (numpy, scypy, sympy, matplotlib, pandas, Keras, Tensorflow, PyTorch)	[3-10] – don't bother reading everything 😊

Literature

1. https://docs.python.org/3/tutorial/index.html
2. https://docs.python.org/3/library/index.html
3. https://numpy.org/
4. https://scipy.org/
5. https://www.sympy.org/en/index.html
6. https://matplotlib.org/
7. https://pandas.pydata.org/
8. https://keras.io/
9. https://www.tensorflow.org/
10. https://pytorch.org/
11. https://wesmckinney.com/book/ (Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter 3rd Edition)
12. https://www.w3resource.com/python-exercises/