Syllabus: Probability and Statistics

2024/2025 Spring

Table of contents

Overview

Background and course rationale

This course introduces students to the discipline of probability statistics as a science of understanding and analyzing data. Themes include data exploratory analysis and inference.

Research projects teach the process of scientific discovery and synthesis and critical evaluation of research and statistical arguments.

In this course, students learn how to effectively make use of data in the face of uncertainty: how to collect data, how to analyze data, and how to use data to make inferences and conclusions about real world phenomena. Critiquing data-based claims and evaluating data-based decisions is at the core of this course. Throughout the course students acquire a conceptual understanding and mastery of statistical and quantitative reasoning tools in order to be able to make such critiques and evaluations.

In addition, students are presented with novel data sets and application examples on a daily basis, and they use these data to model outcomes and make inferences about unknown population characteristics. Students learn that the first step of any analysis is identifying the assumptions and conditions necessary to apply the statistical technique(s) required to answer the research question at hand. Students not only learn the mechanics of the quantitative analysis, but also how to interpret conclusions based on quantitative evidence in context of the data and the research questions as well as identifying limitations due to data collection and study design.

For the lab component of this course students prepare weekly lab reports presenting statistical analysis of real data. In addition, students complete two independent data analysis projects where they answer significant research questions via the analysis of real data using statistical inference.

Course aims to

The course learning objectives are as follows:

1. Recognize the importance of data collection, identify limitations in data collection methods, and determine how they affect the scope of inference.
2. Use statistical software to summarize data numerically and visually, and to perform data analysis.
3. Have a conceptual understanding of the unified nature of statistical inference.
4. Apply estimation and testing methods to analyze single variables or the relationship between two variables in order to understand natural phenomena and make data-based decisions.
5. Interpret results of data analysis in context of data collection method, research question, and assumptions.
6. Critique data-based claims and evaluate data-based decisions.
7. Complete research projects demonstrating mastery of statistical data analysis from exploratory analysis to inference to modeling.

Prerequisites

Basic knowledge of linear algebra and probability theory is welcome.

Learning outcomes

After successful completion of this course, students will have acquired the following skills and knowledge:

1. Data Literacy and Collection:
    - Develop proficiency in understanding and critically evaluating data sources.
    - Learn effective techniques for collecting and curating data, considering limitations and biases.
2. Exploratory Data Analysis (EDA):
    - Master the art of EDA: summarize data numerically and visually using appropriate statistical tools.
3. Statistical Inference:
    - Grasp the unified nature of statistical inference: estimation and hypothesis testing.
    - Apply these methods confidently to draw conclusions about population parameters.
4. Critical Thinking and Communication:
    - Critique data-based claims and evaluate decisions informed by statistical evidence.
    - Communicate findings clearly and effectively to diverse audiences, bridging the gap between data and insights.
5. Research Projects and Practical Skills:
    - Engage in hands-on data analysis through weekly lab reports and independent projects.
    - Demonstrate mastery of statistical techniques—from exploratory analysis to modeling—in real-world scenarios.

Learning materials

All books are **freely available online**.

- OpenIntro Statistics by David M Diez, Christopher D Barr, Mine Çetinkaya-Rundel
- Learning Statistics with R by Danielle Navarro

Course Faculty

Ihor Miroshnychenko, PhD in Economics

@araprof

@datamirosh

aranaur.rbind.io

aranaur

imiroshnychenko@kse.org.ua

ihormiroshnychenko

Career Highlights:

- Associate Professor, Department of Technology Management, FIT, KNU
- Associate Professor, Department of Mathematical Modeling and Statistics, , KNEU
- Lecturer, MBA, International Institute of Business

Research Interests:

- Data Science
- Statistics
- Data Analysis
- Machine Learning

Teaching Philosophy:

I believe that statistics is a powerful tool for understanding the world around us. My goal is to help students develop a deep appreciation for the role of data in decision-making and to equip them with the skills to analyze data effectively and communicate their findings clearly.

Personal Interests:

- Guitar playing

- Traveling
- Photography
- Gaming
- Reading
- Football

Interesting Facts:

- Big fan Middle-Earth universe.
- Engaged in historical reconstruction
- Metalhead

Course Structure

The goal of both the lectures and the labs is for them to be as interactive as possible. My role as instructor is to introduce you new tools and techniques, but it is up to you to take them and make use of them. A lot of what you do in this course will involve writing code, and coding is a skill that is best learned by doing. Therefore, as much as possible, you will be working on a variety of tasks and activities throughout each lecture and lab. Attendance will not be taken during class but you are expected to attend all lecture and lab sessions and meaningfully contribute to in-class exercises and discussion.

You are expected to bring a laptop to each class so that you can take part in the in-class exercises.

Next table contains an outline of the topics, content, and assignments for the semester. Note that this schedule will be updated as the semester progresses and the timeline of topics and assignments might be updated throughout the semester.

Grading Policy

The final course grade will be calculated as follows:

Category	Percentage
Attendance and participation	5%
Labs	50%
Projects	50%

WEEK	TOPIC	SUBTOPIC
1	**Welcome to Probability and Statistics! Probability Theory Recap**	The binomial distribution Laws of probability Bayesian statistics Related distributions Fundamentals of descriptive statistics
2	**Statistical criteria**	Statistical criterion Statistical criterion and R p-value Two-sided criterion
3	**Statistical power**	Power of the Statistical Criterion Minimum detectable effect (MDE) Confidence intervals
4	**Z -test**	Tests concerning the mean Fisher's Z-test
5	**Student's t-test**	t-test Confidence intervals t-test and unknown distribution Two-sample t-test. AB testing + one-sample test for proportion MDE for the t-test.
6	**Monte Carlo**	Checking the criteria Which test is better?
7	**Goodness-of-fit test**	Goodness-of-fit test. Kolmogorov's test Testing for normality. Shapiro-Wilk criterion. QQ-plot Testing the Kolmogorov criterion.
8	**Chi-square tes**	Pearson's Chi-square test Pearson's Chi-square test for independence Pearson's Chi-square test for homogeneity Yates correction
9	**Mann-Whitney test**	Mann-Whitney test Alternative hypothesis of the Mann-Whitney criterion One-sided Mann-Whitney criterion Does the Mann-Whitney work for other hypotheses?
10	**Bootstrap**	Bootstrap Two-sample bootstrap Types of bootstrap confidence intervals Bootstrap vs T-test
11	**Kolmogorov-Smirnov criterion**	Kolmogorov-Smirnov homogeneity criterion Sample size of the Kolmogorov-Smirnov criterion Independence of sampling elements of the Kolmogorov-Smirnov criterion Applicability of the Kolmogorov-Smirnov homogeneity criterion in practice

Attendance and participation

You are expected to be present at class meeting and actively participate in the discussion. While I might sometimes call on you during the class discussion, it is your responsibility to be an active participant without being called on.

If you attend at least 80% of the classes, you'll get all available points for this component.

Labs

In labs, you will apply the concepts discussed in lecture to various data analysis scenarios. Labs will focus on both computation and conceptualization. Lab assignments will be completed using Quarto and submitted as PDF/Word/HTML. While you may collaborate with others on lab assignments, your final solution should be your own.

Projects

The purpose of the project is to apply what you've learned throughout the semester to analyze an interesting data-driven research question. The project will be completed with in teams.

You cannot pass this course if you have not completed the project.

More information about the project will be provided during the semester.

Five tips for success

Your success on this course depends very much on you and the effort you put into it. The course has been organized so that the burden of learning is on you. I will help you be providing you with materials and answering questions and setting a pace, but for this to work you must do the following:

1. Complete all the preparation work before class.
2. Ask questions. As often as you can. In class, out of class. Ask me, ask your friends, ask the person sitting next to you. This will help you more than anything else. If you're not sure about the lab, ask. If the reading is confusing, ask.
3. Do the readings.
4. Do the lab. The earlier you start, the better. It's not enough to just mechanically plow through the exercises. You should ask yourself how these exercises relate to earlier material, and imagine how they might be changed.
5. Don't procrastinate. The content builds upon what was taught in previous weeks, so if something is confusing to you in Week 2, Week 3 will become more confusing, Week 4 even worse, etc.

Academic integrity warning

Academic integrity is highly valued by the KSE community and its vast majority of students. We have a zero-tolerance policy towards academic plagiarism, self-plagiarism, fabrication, falsification, cheating, deception, bribery and other types of academic integrity violations. Due to the highly competitive nature of the programmes, all students must be treated equally. Even a single case of violation of academic integrity is a serious misdemeanour that may lead to unjust redistribution of grades and, consequently, overall rankings, possible grants, fee reductions, and other merit-based awards. Therefore, penalties may vary from receiving zero points or a negative grade for an assignment to expulsion of the student from KSE, depending on the severity of the case and might include additional consequences such as deprivation of scholarships, financial assistance etc. All the rules and procedures regarding academic integrity are stated in the KSE Code of Academic Integrity. Before enrolling in this course, students must be aware of and abide by the KSE Code of Academic Integrity.

Academic honesty

TL;DR: Don't cheat!

Please abide by the following as you work on assignments in this course:

- **Collaboration**: Only work that is clearly assigned as team work should be completed collaboratively.
  - You may discuss lab assignments with other students; however, you may not directly share (or copy) code or write up with other students. For team assignments, you may collaborate freely within your team. You may discuss the assignment with other teams; however, you may not directly share (or copy) code or write up with another team. Unauthorized sharing (or copying) of the code or write up will be considered a violation for all students involved.
  - For the project, collaboration within teams is not only allowed, but expected. Communication between teams at a high level is also allowed however you may not share code or components of the project across teams.
- **Online resources**: I am well aware that a huge volume of code is available on the web to solve any number of problems. Unless I explicitly tell you not to use something, the course's policy is that you may make use of any online resources (e.g., StackOverflow) but you must explicitly cite where you obtained any code you directly use (or use as inspiration). Any recycled code that is discovered and is not explicitly cited will be treated as plagiarism.
- **Use of generative artificial intelligence (AI)**: You should treat generative AI, such as ChatGPT, the same as other online resources. There are two guiding principles that govern how you can use AI in this course:

1. Cognitive dimension: Working with AI should not reduce your ability to think clearly. We will practice using AI to facilitate—rather than hinder—learning.
2. Ethical dimension: Students using AI should be transparent about their use and make sure it aligns with academic integrity.

- AI tools for code: You may make use of the technology for coding examples on assignments; if you do so, you must explicitly cite where you obtained the code. Any recycled code that is discovered and is not explicitly cited will be treated as plagiarism. You may use these guidelines for citing AI-generated content.
- AI tools for narrative: Unless instructed otherwise, you may not use generative AI to write narrative on assignments. In general, you may use generative AI as a resource as you complete assignments but not to answer the exercises for you.

You are ultimately responsible for the work you turn in; it should reflect your understanding of the course content.

Late work & extensions

The due dates for assignments are there to help you keep up with the course material and to ensure the teaching team can provide feedback within a timely manner. We understand that things come up periodically that could make it difficult to submit an assignment by the deadline.

- Labs may be submitted up to 3 days late. There will be a 5% deduction for each 24-hour period the assignment is late.
- The late work policy for the project will be provided with the project instructions.