Welcome to DSCI 551: Descriptive Statistics and Probability for Data Science

Contents

Back to top

- License

This course introduces descriptive statistics and probability, including measures of location and spread, random variables, distributions, parameters, categorical variables, and uncertainty.

High-Level Goals

By the end of the course, students are expected to:

- Provide fundamental concepts in probability, including conditional, joint, and marginal distributions.
- Develop a statistical view of data coming from a probability distribution.

Learning Objectives

- Compute summary statistics, such as expected value and variance, of simple discrete and continuous probability distributions.
- Compare/contrast location summary statistics such as mean/median/mode/quantiles.
- Estimate summary statistics such as mean/median/variance from a plot of a distribution’s PDF or CDF.
- Identify common continuous distributions such as Gaussian/Poisson/uniform from a plot of a distribution’s PDF or CDF.
- Match common discrete distributions such as Bernoulli/binomial/multinomial to descriptions.
- Compare/contrast conditional, joint and marginal distributions.
- Explain the notion of “marginalizing out” a random variable.
- Identify independence between random variables from plots/tables of conditional/joint/marginal distributions.
- Connect conditional distributions to the notion of supervised learning.
- Explain the concept of maximum likelihood estimation.

- Identify the units of various quantities such as mean/variance/density for continuous distributions.
- Simulate sample generation from probability distributions, and interpret the results.

Teaching Team

Position	Name	Slack Handle	GHE Handle	Section
Lecture/Lab Instructor	Vincent Liu	`@Vincent Liu`	`@vliu07`	1
Lecture/Lab Instructor	Alexi Rodríguez-Arelis	`@Alexi`	`@alexrod6`	2
Teaching Assistant	Mohammad Mahdi Asmae	`@Mahdi Asmae`	`@masmae`	1
Teaching Assistant	Anne-Sophie Fratzscher	`@Anne-Sophie (TA)`	`@afratz`	1
Teaching Assistant	Cindy (Xiao Yu) Zhang	`@Cindy Xiao Yu Zhang (TA)`	`@szzxy`	1
Teaching Assistant	Ailar Mahdizadeh	`@Ailar Mahdizadeh`	`@ailarmz`	2
Teaching Assistant	Daniel Ramandi	`@Daniel Ramandi`	`@dani014`	2
Teaching Assistant	Armin Saadat Boroujeni	`@Armin Saadat`	`@arminsdt`	2
Teaching Assistant	Shaocheng Wu	`@shaochwu`	`@markutz`	2

Lecture Topics

This course occurs during **Block 1** in the 2024/25 school year. The course notes can be accessed **here**.

Lecture Topic/Notes	Required Readings	Optional Readings
Depicting Uncertainty	`lecture1` notes	Part 1: Core Probability
Parametric Families	`lecture2` notes	Part 2: Random Variables
Joint Probability	`lecture3` notes	Part 3: Probabilistic Models, Chapter 5.1, Covariance and correlation (video), How would you explain covariance ...
Conditional Probabilities	`lecture4` notes	Part 3: Probabilistic Models, Chapter 5.3
Continuous Distributions	`lecture5` notes	Chapter 4
Common Distribution Families and Conditioning	`lecture6` notes	Part 2: Random Variables
Maximum Likelihood Estimation	`lecture7` notes	Part 5: Machine Learning,Beyond Multiple Linear Regression, sections 2.1 to 2.4, Chapter 7.1 & 7.2
Simulation and Empirical Distributions	`lecture8` notes	Chapter 9: Applications to Computing

Cheat sheet

Here is a cheat sheet we created to summarize the main formulas and concepts covered in DSCI 551.

Deliverables

This is an **assignment-based course**. The following deliverables will determine your course grade:

Assessment	Weight
Lab Assignment 1	12%
Lab Assignment 2	12%
Lab Assignment 3	12%
Lab Assignment 4	12%
Quiz 1	25%
Quiz 2	25%
Lecture Attendance (iClicker)	2%

Lectures

Schedule

Refer to the MDS calendar.

Labs

Schedule

Refer to the MDS calendar.

Lab Topics and Due Dates

	Lab Topic	Due Date
1	Depicting Uncertainty and Parametric Families (Lectures 1 and 2)	Refer to the MDS calendar.
2	Joint and Conditional Probabilities (Lectures 3 and 4)	Refer to the MDS calendar.
3	Continuous Distribution Families (Lectures 5 and 6)	Refer to the MDS calendar.
4	Maximum Likelihood Estimation and Simulation (Lectures 7 and 8)	Refer to the MDS calendar.

Lab Grade Computation

Once lab grades are published on Gradescope, you will see your **raw lab mark** m. This **raw lab mark** m is the grand total of your granted marks throughout the whole lab assignment. Now, if we add up **all the marks (non-challenging and challenging)** in the handout corresponding to all `rubric={...}`, this sum is what we call the maximum raw lab mark m_{100} to get 100% as a percentage lab grade. On the other hand, if we add up **the non-challenging marks** in the handout found in `rubric={...}`, this sum is what we call the raw lab mark m_{95} to get a 95% as a percentage lab grade.

By the end of the block, **once all lab marking is finished on Gradescope**, your raw lab grades will be transferred to **Canvas**. Then, in your **Canvas gradebook**, you will see these raw lab grades (`raw lab1`, `raw lab2`, `raw lab3`, and `raw lab4`). Finally, for each of the four labs, you will also see your final lab grades (`lab1`, `lab1`, `lab3`, and `lab4`). Let g be the final lab grade of a specific lab **as a percentage**; it will be computed as follows:

- If $m > m_{95}$, then $g = 95 + \left(\frac{m - m_{95}}{m_{100} - m_{95}} \times 5 \right)$.
- If $m \leq m_{95}$, then $g = \left(\frac{m}{m_{95}} \right) \times 95$.

Quizzes

Refer to the MDS calendar.

Office Hours

Refer to the MDS calendar.

Communication

We will use **Slack** as the main communication channel.

If you have any questions regarding the course content, lectures, labs, autograders, or any other course-related matters, we kindly request that you avoid direct messaging (DM) the instructor or TAs. Instead, please post your question on the #551 channel. This approach not only enables our TAs to respond promptly but also benefits other students who might have similar questions.

> **Response time:** We will try our best to reply to your inquiries as soon as possible during the normal working hours (9AM-5PM Mon-Fri). If you send us a message outside of regular working hours, please expect a response on the next working day.

Use of LLMs

LLMs, such as ChatGPT, can be helpful tools if we use them responsibly. In this course, students are permitted to use these tools to gather more information, review concepts, or brainstorm, and students must cite these tools if they use them for assignment. Having said all this, it is **not** permitted to write any given assignment via copying and pasting AI-generated responses.

Resources

> **Note:** Some of these resources cover much more material than DSCI 551.

- Introduction to Probability for Data Science

- Course Reader CS109 Stanford by Chris Piech
- Chapter 3: Probability and Information Theory,from the Deep Learning Book by Goodfellow, I., Bengio, Y., and Courville, A. (2016)
- Probability & Statistics with Applications to Computing by Alex Tsun, Stanford
- JBstatistics (also on YouTube)
- Introduction to Probability, Statistics, and Random Processes
- Harvard STAT 110 course, YouTube videos
- Probability Cheatsheet
- Word problems for conditional probability

Policies

See the general MDS policies.

Attribution

The course is built upon previous years' materials developed by previous instructors.

License