STA 35C: Statistical Data Science III

Lecture 1: Introduction and Overview

Dogyoon Song

Fall 2026, UC Davis

Agenda

- Course overview
- Course logistics, syllabus, and calendar
- Introduction to probability: Intuition and a first example

What is statistical data science?

- **Statistics**: The study of collecting, analyzing, and drawing conclusions from data
- **Data science**:
  - Combines statistical thinking, computing, and subject-matter knowledge
  - Uses real-world data to answer questions and make predictions in various applications (science, engineering, business, sports, government, etc.)
- **STA 35 sequence**: Introductory data science courses from a statistical perspective, with an emphasis on computing with R
  - **STA 35A**: Introduction to statistics (probability, distributions, confidence intervals, hypothesis testing, etc.)
  - **STA 35B**: R programming + additional statistical methods (linear regression, ANOVA, permutation tests, etc.)
  - **STA 35C**: Fundamentals of **statistical learning** methods
    - understanding key ideas, how and when to apply them, and their limitations

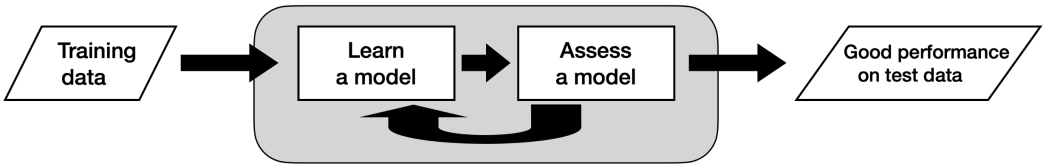
What is statistical learning?

- **Definition**: A set of tools for understanding data and making informed predictions
- **Examples**:
  - Identifying critical disease risk factors from large patient records
  - Predicting whether an event will occur (e.g., credit default, septic shock)
  - Classifying medical images or tissue cells (e.g., benign vs. malignant tumors)
  - Recognizing and localizing objects in images (e.g., for autonomous vehicles)
  - Evaluating impacts of new legislation or predicting unemployment rates
  - Modeling relationships among many variables to gain practical insights (e.g., which marketing strategies drive sales)
  - ...

Supervised vs. unsupervised learning

- **Supervised learning**
  - **Setup**: We have measurements of an outcome Y and predictors (features) X
  - **Goals**: Predict Y given X, or understand how Y is associated with X
  - **Examples**:
    - *Regression*: Forecast a product's sales next month
    - *Classification*: Predict if a customer will default on a loan
- **Unsupervised learning**
  - **Setup**: We only have features X, without a designated outcome variable
  - **Goals**: Discover hidden patterns or groupings in the data
  - **Examples**:
    - *Dimension reduction*: Summarize many features using fewer variables
    - *Clustering*: Cluster customers by purchasing behavior

Statistical learning and STA 35C



- **Core idea**: Fit a model from data, assess it, and refine it
  - Aim for good performance on ***new, unseen* data**
  - Rely on probability and statistical principles to handle uncertainty
- In **STA 35C**, you'll learn the fundamentals of these methods
  - When and how to use different supervised or unsupervised learning methods
  - How to assess and interpret models (cross-validation, model selection, ...)
- Our focus is on learning **first principles**; we **do not** cover advanced machine learning techniques (e.g., deep neural networks, large language models)

Course logistics

- **Instructor:** Dogyoon Song
  - Email: dgsong@ucdavis.edu
  - Office hours: TBA
- **TA:** Qiqi Liu
  - Email: qqiliu@ucdavis.edu
  - Office hours: TBA
- **Lectures:** Monday, Wednesday and Friday, 9:00–9:50 AM at Olson Hall 205
- **Lab/Discussions:** Thursday 11:00–11:50 AM at TLC 2212 (run by TA)
- **Online platforms**
  - Course webpage: Lecture notes, homework, supplementary materials, etc.
  - Canvas: Homework submission, solutions and grades, lab materials
  - Piazza: Announcements and discussion
  - Email: Private matters only (**do not** send messages on Canvas)

Course content & prerequisites

Course content:

- Probability basics
- Supervised learning
  - Fundamental concepts
  - Regression
  - Classification
- Model assessment, selection, and inference
  - Cross-validation and the bootstrap
  - Model selection and regularization
  - Simultaneous inference
- Nonlinear methods and smoothing
- Unsupervised learning
  - Dimension reduction and PCA
  - Clustering

Prerequisite(s):

- STA 035B (C- or better)
- MAT 016B, 017B, 019B or 021B (C- or better)

These requirements are strict

- If you don't meet prerequisites, please submit a petition ASAP
- Try "Homework 0" for self-assessment

Syllabus walkthrough

Open the Fall 2026 syllabus and follow along:

1. **Basic information (p. 1)**
   - Course staff, meeting times, and course platforms
   - Textbook: *An Introduction to Statistical Learning with Applications in R*, **2nd ed.** (2021)
2. **Course topics and assessment (pp. 2–3)**
   - Grading scheme, and exam policies
3. **Homework (p. 3)**
   - Submit a single PDF with your written solutions and code
4. **Course policies (p. 4)**
   - For homework, you can discuss and collaborate, but must write your own solutions and code
   - **AI assistance on graded work is prohibited**
5. **Getting help (p. 5)**
   - Piazza and office hours for course questions, email for private matters
   - If needed, submit SDC accommodation requests early

Grading

- **Homework**: 30%
  - 7 assignments (excluding HW 0), with the best six scores counting
  - Released on scheduled Fridays, due the following Thursday at 11:59 PM PT
    - Exception: HW 7 is due **Tue, Dec 1**
  - **No late homework accepted for any reason**
- **Midterm exams**: 30%
  - Two in-class midterms (**Mon, October 19 & Mon, November 9**)
  - Only the higher score counts (the lower is dropped)
  - **No make-up exams**; a missed midterm receives 0 and can be the dropped score
- **Final exam** (cumulative): 40%
  - **Wed, Dec 9, 8:00–10:00 AM**
- **Participation**: up to 3 **percentage points** extra credit
  - Up to 2 for lecture participation, plus up to 1 on the TA's recommendation

See syllabus for full details

Fall 2026 calendar at a glance

Date	Event
Wed, Sep 23	First lecture
Mon, Oct 19	**Midterm 1, in class**
Mon, Nov 9	**Midterm 2, in class**
Wed, Nov 11	**Veterans Day holiday: no lecture**
Wed, Nov 25	**No STA 35C lecture (course schedule)**
Thu–Fri, Nov 26–27	**Thanksgiving holiday: no discussion or lecture**
Tue, Dec 1	HW 7 due at 11:59 PM PT (exception)
Fri, Dec 4	Last lecture
Wed, Dec 9	**Final exam, 8:00–10:00 AM**

See the syllabus for the homework calendar (all times are Pacific Time)

Getting ready: Homework 0 and R

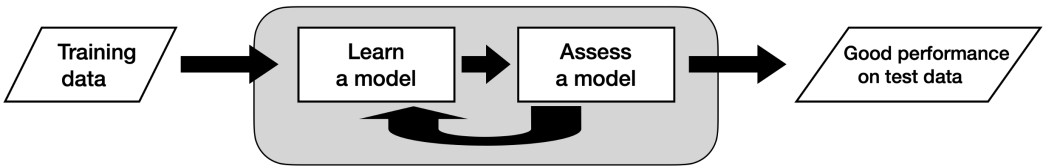
- **Homework 0: self-assessment**
  - Review key STA 35A/B material and check your familiarity with R
  - Not collected or graded. No solutions will be provided
  - Bring questions to discussion (**Thu, Sep 24**), or to office hours
- **Software access**:
  - You can install R and RStudio on your own computer, or see UC Davis Cloud Services
  - Before HW 1, check that you can run R code and render an R Markdown or Quarto document to PDF, showing code and output
  - If you encounter setup problems, let us know immediately so that we can help
- **Lab/discussion section**:
  - Thursdays, 11:00–11:50 AM, TLC 2212 where computers are available
  - You may also bring your own laptop (please charge it beforehand)
  - Lab/discussion and TA office hours are the best places to get help with R

Motivation: Why probability?

Recall the goals of statistical learning:

- Use predictors X to predict an outcome Y (learn a prediction rule $f : X \to Y$)
- Identify patterns in data X

Standard workflow:



Key challenge:

- We aim for good predictions or insights on ***new, unseen* data**
- How should we assess a model, given that training data $\neq$ test data?

How probability supports statistical learning

Probability helps us describe uncertainty and assess what we can learn from data, offering a formal language and methods

- **Issue/Need:** Training data $\neq$ test data
  - *Solution:* Assume training and test data are *randomly drawn* from same distribution
- **Issue/Need:** Uncertainty in prediction
  - *Solution:* Model Y as a *random variable*, and predict Y conditioned on X
- **Issue/Need:** Choosing among many models
  - *Solution:* Identify the "most likely" model given data, or update our belief about the "best model" based on observed data

We will discuss these aspects in more detail throughout the course

Probability in everyday examples

- Fair coin toss
  - Possible outcomes are Head or Tail; each has probability 0.5
- Fair six-sided die roll
  - Possible outcomes $\{1, 2, \ldots, 6\}$; each has probability $1/6$
- US childbirths[1]
  - About 51.2% of live births were recorded as male, and 48.8% as female
  - A birth selected uniformly from these has probability 0.512 of being recorded as male
- Commute time
  - Commute may take 30 minutes on average, but can vary with traffic, weather, etc.
- Subjective probabilistic beliefs
  - You may personally estimate the likelihood of a stock price rising or falling, based on your own analysis or expert opinions
  - This kind of probability reflects beliefs rather than strict long-run frequencies

[1] Source: CDC National Vital Statistics Reports, Births: Final Data for 2023

Toward formalizing probability: Sample space and events

- **Outcome:** a possible result of an experiment or trial
- **Sample space**: the set of all possible outcomes, often denoted by Ω
  - e.g., $\{H, T\}$, $\{1, 2, 3, 4, 5, 6\}$
- **Event:** a subset of Ω
  - e.g., for $\Omega = \{H, T\}$: $\emptyset$, $\{H\}$, $\{T\}$, $\{H, T\}$
  - e.g., for $\Omega = \{1, 2, 3, 4, 5, 6\}$: $\{6\}$, $\{1, 2\}$, $\{2, 4, 6\}$, ...
- **Probability**[2]**:** a map P that assigns a number in $[0, 1]$ to each event such that
  - $P(\Omega) = 1$;
  - For disjoint events $A_1, A_2, \ldots$, $P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$
    - In particular, if $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$

We will develop probability rules more systematically next lecture

[2] A formal, mathematically rigorous definition of probability measure is beyond the scope of STA 35C

Where we are headed

- **Throughout the course**:
  - Learn key ideas of regression, classification, clustering, and more
  - Practice implementing methods and interpreting results
  - Assess when each method is or is not appropriate
- **Next lecture (Fri, Sep 25) and in the first two weeks**:
  - Review probability rules and independence
  - Develop understanding of conditional probability and explore Bayesian ideas
- **Before next lecture**:
  - Complete "Homework 0" at your earliest convenience and seek help if needed
  - Ensure you have access to R and RStudio