Sungkyunkwan University (SKKU) International Summer Semester (ISS) 2025

Statistics in Python

Prof. Kyung Eun Park, Wilkes University

COURSE DESCRIPTION

This course is designed to develop essential statistical thinking skills for learning from data and deriving meaningful insights. It also emphasizes building practical systems to maximize the utility of data. Students will gain hands-on experience with Python, learning programming techniques to perform exploratory data analysis (EDA) and conduct deeper statistical investigations using diverse datasets.
The course introduces the basic methodologies of machine learning, including techniques for performing sampling principles, regression analysis, statistical modeling, and inference using various Python libraries, as well as creating machine learning models and training them for predictive analytics.
By the end of this course, students will have developed a strong foundation in Python, a highly flexible and seamless programming language supported by a vibrant open community. These capabilities empower students to design and implement robust, data-driven solutions.

This course welcomes students from diverse backgrounds, including those with prior experience in statistics, programming, both, or none. While no specific prerequisites are required, students with limited or no background in mathematics, coding, or statistical concepts may need to dedicate additional time and effort to keep up with the course material. Support will be provided to help all students achieve their best outcomes, but success will require consistent effort and engagement.

READING MATERIALS

Textbook 1: Intro Stats, 5th Edition, Richard De Veaux, Paul Velleman, and David Bock, 2018 Pearson, ISBN 13: 978-0-13-421022-3
Textbook 2: Python for probability, statistics, and machine learning, 2nd Edition, Springer, ISBN: 978-3-030-18545-9
Reference:

- Think Stats, 2nd Edition, https://greenteapress.com/thinkstats2/html/index.html
- Python Programming, An Introduction to Computer Science, 3rd Ed., John Zelle, Franklin, Beedle, ISBN-13: 978-1590282755
- Introduction to Computation and Programming Using Python, 2nd Ed., The MIT Press, ISBN-13: 978-0262529624

Additional reading materials will be provided.
Class materials will be provided.
Download and install Python in your laptop: https://www.anaconda.com/distribution/

COURSE REQUIREMENTS AND GRADING

Attendance (20%): It is important that you attend every lecture. SKKU regulations require students to attend at least 80% of all classes. Bring your laptop for in-class exercises. Please do not use your cellphone in class.
Assignments (30%): Homework will be assigned regularly. No late submission will be accepted without a valid excuse in advance. You are welcome to consult each other on assignments, but you should submit your own work. To get full credit on homework and exam problems, you must show everything necessary to establish the validity of your answer and the approach that you use to get it. If you use Python to answer, you must include the relevant code and outputs.

Exams (50%): There will be two exams, and each will count for 25%. No make-up exams will be given unless you have a proper reason not to be present in the exam (should notify the instructor in advance).

You will earn a pass if you receive a grade of 60% or above.
Academic dishonesty will not be tolerated. A failure will be given if academic honesty is violated.

COURSE SCHEDULE

– WEEK I –

Monday (30 June) – Day 1

Introduction to the course
Python basics: Variables, data types, control structures, and functions
Hands-On: Writing basic Python programs

Tuesday (1 July) – Day 2

Introduction to Data: Types of data and basic statistics (mean, median, mode, variance)
Exploratory Data Analysis (EDA): Data cleaning, handling missing data
Hands-On: Using Pandas for basic data exploration

Wednesday (2 July) – Day 3

Data Visualization: Importance and best practices
Visualization Libraries: Matplotlib, Seaborn basics
Hands-On: Creating basic plots (line, bar, histogram, scatter)

Thursday (3 July) – Day 4

Probability Basics: Rules, distributions (normal, uniform, etc.)
Sampling: Distribution models and confidence intervals
Hands-On: Simulating distributions and sampling with Python libraries

– WEEK II –

Monday (7 July) – Day 5

Hypothesis Testing: Concepts, p-values, t-tests, significance levels
Common Tests: hypothesis testing and probability distributions
Hands-On: Performing statistical tests using scientific python tools

Tuesday (8 July) – Day 6

Exam I

Wednesday (9 July) – Day 7

Correlation and Causation: Correlation coefficients, limitations
Linear Regression Introduction: Concepts and assumptions
Hands-On: Implementing linear regression with Scikit-learn

Thursday (10 July) – Day 8

Multiple Linear Regression: Handling multiple features
Model Evaluation: R-squared, adjusted R-squared, residual analysis
Hands-On: Building and evaluating models with real-world datasets

– WEEK III –

Monday (14 July) – Day 9

Logistic Regression: Binary classification basics
Performance Metrics: Accuracy, precision, recall, F1-score
Hands-On: Implementing and evaluating logistic regression models

Tuesday (15 July) – Day 10

Data Preprocessing: Scaling, encoding categorical data, feature selection
Hands-On: Preprocessing datasets for machine learning tasks

Wednesday (16 July) – Day 11

Clustering Basics: K-Means and hierarchical clustering
Use Cases: Customer segmentation, anomaly detection
Hands-On: Clustering with Scikit-learn

Thursday (17 July) – Day 12

Decision Trees: Concepts, Gini index, entropy
Random Forests: Ensemble methods for robust modeling
Hands-On: Implementing trees and forests with Scikit-learn

– WEEK IV–

Monday (21 July) – Day 13

Exam II

Tuesday (22 July) – Day 14

Advanced programming in Python: User-defined classes and GUI programming

Wednesday (23 July) – Day 15

Project presentation