GAVILAN COLLEGE
5055 Santa Teresa Blvd
Gilroy, CA 95023

Course Outline

COURSE: MATH 21 DIVISION: 20 ALSO LISTED AS:

TERM EFFECTIVE: Summer 2024

SHORT TITLE: INTRO DATA SCIENCE

LONG TITLE: Introduction to Data Science

Units	Number of Weeks	Type	Contact Hours/Week	Total Contact Hours
4	18	Lecture:	3	54
		Lab:	3	54
		Other:	0	0
		Total:	6	108

Out of Class Hrs:	108.00
Total Learning Hrs:	216.00

COURSE DESCRIPTION:

An introduction to the foundations of data science that combines three perspectives: inferential thinking, computational thinking, and real-world relevance. The course teaches critical concepts and skills in computer programming and statistical inference, in conjunction with hands-on analysis of real-world datasets. PREREQUISITE: MATH 240 or Intermediate Algebra or equivalent skills or appropriate placement with a grade of 'C' or better.

PREREQUISITES:

Completion of MATH 240, as UG, with a grade of C or better.
OR
Completion of MATH 233, as UG, with a grade of C or better.
OR
Completion of MATH 242, as UG, with a grade of C or better.
OR
Completion of MATH 233A, as UG, with a grade of C or better.
OR
Completion of MATH 233B, as UG, with a grade of C or better.
OR
Completion of MATH 235, as UG, with a grade of C or better.
OR
Completion of MATH 8A, as UG, with a grade of C or better.
OR
Completion of MATH 8B, as UG, with a grade of C or better.
OR
Completion of MATH 11, as UG, with a grade of C or better.
OR
Completion of MATH 1A, as UG, with a grade of C or better.
OR
Completion of MATH 2, as UG, with a grade of C or better.
OR
Completion of MATH 2C, as UG, with a grade of C or better.

COREQUISITES:

CREDIT STATUS: D - Credit - Degree Applicable

GRADING MODES

L - Standard Letter Grade

REPEATABILITY: N - Course may not be repeated

SCHEDULE TYPES:

02 - Lecture and/or discussion
03 - Lecture/Laboratory
04 - Laboratory/Studio/Activity
047 - Laboratory - LEH 0.7
05 - Hybrid
71 - Dist. Ed Internet Simultaneous
72 - Dist. Ed Internet Delayed
73 - Dist. Ed Internet Delayed LAB
737 - Dist. Ed Internet LAB-LEH 0.7

STUDENT LEARNING OUTCOMES:

By the end of this course, a student should:

1. Identify questions that can be answered about real-world datasets while accounting for factors such as bias and non-causality.

2. Write computer code to organize, numerically summarize, and produce visualizations of datasets.

3. Explain the process and results of a statistical analysis.

4. Conduct statistical inference about a population by examining random samples generated using computer iteration.

5. Apply machine-learning techniques such as linear regression and classification to data in order to make predictions.

COURSE OBJECTIVES:

By the end of this course, a student should:

1. Explain the difference between correlation and causation.

2. Write computer code to organize and numerically summarize datasets.

3. Describe relationships among variables by creating and analyzing histograms, bar charts, and scatter plots.

4. Calculate probabilities of events, and perform probability simulations with computer coding.

5. Assess the validity of a statistical model about a population by generating and examining random samples with computer iteration.

6. Estimate a population parameter using a confidence interval produced through bootstrapping, and interpret the confidence interval in context.

7. Describe the effect of sample size on the shape and variability of a distribution of sample means.

8. Make predictions by applying machine-learning techniques, such as linear regression and classification, to data.

COURSE CONTENT:

COURSE APPROVAL DATE: 03/12/2024

LECTURE CONTENT:

1. Introduction to Data Science (3 hours)

What is Data Science?

- Exploration
- Inference
- Prediction

Cause and Effect

- Observational vs. experimental studies
- Correlation vs. causation
- Confounding factors
- Importance of randomization

2. Introduction to Python (or other programming language) Tools (6 hours)
Tables
- Roles of rows vs. columns
- Selecting, dropping, sorting, filtering
Data Types
- Ints, floats, strings, built-in functions, tables
- Conversions
- Type function
Building Tables
- Arrays
- Column arrays
- Ranges
- Creating tables
Census
- Table methods
- Manipulating rows
- Bias in data collection
3. Data Visualization (3 hours)
Numerical vs. categorical variables
Distributions
Categorical Distributions
- Bar charts
Numerical Distributions
- Histograms
- Line plots and graphs
- Scatterplots
Determining which type of visual is appropriate
Visualization fundamentals and best practices
4. Data Analysis Techniques (6 hours)
Functions
- Defining functions
- Applying functions on columns
Groups
- Grouping data by one categorical variable
- Cross-classifying data by multiple categorical variables
- Pivots
- Prediction accuracy with grouping
Joins: integrating data from multiple datasets
Real-life examples of tables and tabling techniques
5. Randomness (6 hours)
Iteration
- Random choice tool
- Appending arrays
- `For? loops and `if? statements
- Simulation
Chance
- Definition of probability
- Multiplication, addition, and complement rules

- Problem-solving strategies

Sampling

- Deterministic sampling vs. random sampling
- Convenience sampling
- Probability distributions vs. empirical distributions
- Law of Large Numbers
- Parameters vs. statistics
- Statistical inference

Statistical Models

- Assessing population models with simulations and data
- Real-world examples of statistical models

6. Hypothesis Testing (8 hours)

Comparing Distributions

- Fundamentals of hypothesis testing
- Model vs. alternative viewpoint
- Steps to compare the model to the alternative

Decisions and Uncertainty

- Null and alternative hypotheses
- Test statistics
- Prediction under the null hypothesis
- Conclusion of the test
- Statistical significance and P-value
- Type 1 and 2 errors

Hypothesis Test Types and Examples

- One sample, one category
- One sample, multiple categories
- One sample, numerical
- Two samples, underlying numerical values (A/B testing)
- Randomized controlled experiment

7. Estimation (3 hours)

Confidence Intervals

- Variability of the estimate
- Bootstrapping

Interpreting Confidence Intervals

- Confidence vs. probability

8. The Central Limit Theorem (7 hours)
Center and Spread
- Mean vs. median
- Standard deviation
- Chebyshev's bounds
- Standard units
The Normal Distribution
- Standard normal curve
- Bounds and normal approximations
Sample Means
- Probability distribution of the sample average
- Central Limit Theorem
- Variability of the sample average
- Confidence intervals for sample means
Designing Experiments
- Choosing sample size
- Margin of error
9. Making Predictions (10 hours)
Correlation
- Linear regression
- Correlation coefficient
- Standard units
- Predicting an output from an input
Least Squares
- Error in estimation
- Numerical optimization
Residuals
- Definition and properties
Regression Inference
Classification
- Machine-learning algorithms
- Training and testing
- Nearest neighbor classifiers
Decisions with Bayes? Rule
- Probability trees
- Subjective probabilities
10. Final Exam (2 hours)

LAB CONTENT:

Lab 1: Expressions (3 hours): Navigating Jupyter notebooks; writing and evaluating basic expressions in Python (or other programming language); naming expressions; calling functions to use code other people have written; and breaking down Python (or other programming language) code into smaller parts to understand it.

Lab 2: Table Operations (3 hours): Importing pre-existing code; performing table operations such as sorting and filtering; and using the operations to do basic analysis on datasets.

Lab 3: Data Types and Arrays (3 hours): Representing and manipulating text and arrays of numbers; creating tables from arrays.

Lab 4: Functions and Visualizations (3 hours): Defining functions and implementing them on data; creating visualizations of data.

Project 1: World Progress (8 hours): Analyzing data about life expectancy, fertility rates, and child mortality across the world over time and determining whether these rates are associated with population growth. Examining trends in global poverty.

Lab 5: Simulations (3 hours): Writing conditional statements; conducting simulations involving randomness.

Lab 6: Examining the Therapeutic Touch (3 hours): Simulating an experiment, using a coin-flipping model, to assess whether some people have the ability to detect "Human Energy Fields".

Lab 7: Great British Bake-Off (A/B Test) (3 hours): Conducting a hypothesis test to determine whether season winners and non-winners of the Great British Bake-Off show received weekly Star Awards in equal number.

Lab 8: Normal Distribution and Variability of Sample Means (3 hours): Examining the characteristics of normal distributions and the shapes and variabilities of distributions of sample means taken from any population distribution, whether normal or not. Exploring the effect of sample size on the latter.

Project 2: Climate Change-- Temperatures and Precipitation (8 hours): Investigating temperature and precipitation trends in various cities and conducting an A/B hypothesis test to determine whether the claim of climate change is plausible.

Lab 9: Regression (3 hours): Constructing a linear regression to help predict wait times of the Old Faithful geyser from the durations of the previous eruptions.

Lab 10: Conditional Probability (3 hours): Calculating conditional probabilities; inspecting the false positive paradox.

Project 3: Movie Classification (8 hours): Predicting genres of movies from the texts of their screenplays.

METHODS OF INSTRUCTION:

Instruction will be by lecture/discussion with periodic cooperative problem solving sessions.

OUT OF CLASS ASSIGNMENTS:

Required Outside Hours 108

Assignment Description

Homework assignments on various topics:

Homework 1: Causality and Expressions

Homework 2: Arrays and Tables

Homework 3: Table Manipulation and Visualization

Homework 4: Functions, Tables, and Groups

Homework 5: Applying Functions and Iteration

Homework 6: Probability, Simulation, Estimation, and Assessing Models

Homework 7: Testing Hypotheses

Homework 8: Confidence Intervals

Homework 9: Sample Sizes and Confidence Intervals

Homework 10: Linear Regression

METHODS OF EVALUATION:

Evaluation Percent 60

Evaluation Description

In-class examinations that combine objective questions (eg. multiple-choice), coding, and free-response questions.

Evaluation Percent 10

Evaluation Description

Labs that involve coding and interpretation of results.

Evaluation Percent 10

Evaluation Description

Weekly homework assignments.

Evaluation Percent 10

Evaluation Description

Projects that require students to apply course knowledge to various questions about real-world datasets.

Evaluation Percent 10

Evaluation Description

Quizzes, either take-home or in-class.

REPRESENTATIVE TEXTBOOKS:

Computational and Inferential Thinking: The Foundations of Data Science, Ani Adhikari, John DeNero, David Wagner, self publishing, 2021 or a comparable textbook/material.

ISBN: N/A

10 Grade Verified by: Erik Medina

Lab materials found at https://github.com/data-8/materials-sp20

ARTICULATION and CERTIFICATE INFORMATION

Associate Degree:
CSU GE:
CSU B4, effective 202450
IGETC:
IGETC 2A, effective 202450
CSU TRANSFER:
Transferable CSU, Effective 202450
UC TRANSFER:
Transferable UC, Effective 202450

SUPPLEMENTAL DATA:

Basic Skills: N
Classification: Y
Noncredit Category: Y
Cooperative Education: N
Program Status: 1 Program Applicable
Special Class Status: N
CAN:
CAN Sequence:
CSU Crosswalk Course Department:
CSU Crosswalk Course Number:
Prior to College Level:
Non Credit Enhanced Funding: N
Funding Agency Code: Y
In-Service: N
Occupational Course: E
Maximum Hours:
Minimum Hours:
Course Control Number: CCC000643714
Sports/Physical Education Course: N
Taxonomy of Program: 170100