MATH 0011 - DATA SCIENCE FOR ALL

Catalog Description

Prerequisite: Completion of Intermediate Algebra or equivalent with grade of "C" or better, or appropriate placement
Hours: 108 (54 lecture, 54 laboratory)
Description: Designed for students from any major, provides high-level understanding of how data, statistics, and inference are inter-related. Introduces the core concepts of data science, including statistical inference and computational thinking. Teaches critical concepts and skills in computer programming and statistical inference while working with real data, such as economic data, geographic data, and social networks. Prepares students to make more data-driven decisions, gaining experience with machine learning and with the practical application of statistical concepts like hypothesis testing, confidence intervals via bootstrapping, regression, inference for regression, and predictive modeling while considering the social issues surrounding data privacy and data ownership. (C-ID MATH 110) (CSU, UC)

Course Student Learning Outcomes

- CSLO #1: Apply numerical methods of descriptive statistics and probabilistic concepts using sampling method to extract relevant information.
- CSLO #2: Interpret data by analysis, draw meaningful conclusions while considering ethics and privacy, and present their findings effectively using various types of graphs.
- CSLO #3: Create computer scripts for data manipulation, visualization, classification and numerical representation.
- CSLO #4: Conduct various statistical techniques, such as hypothesis testing, estimating a parameter using confidence intervals, linear regression and correlation, chi-squared tests, and one-way analysis of variance (ANOVA).

Effective Term

Fall 2025

Course Type

Credit - Degree-applicable

Contact Hours

108

Outside of Class Hours

108

Total Student Learning Hours

216

Course Objectives

Upon Successful completion of the course, students will be able to:

1. Explore various methods for data acquisition and assess their respective strengths and weaknesses;
2. Interpret and describe data displayed in tables;
3. Analyze and interpret data trends, both visually through graphs and numerically;
4. Address potential biases and unintended outcomes inherent in machine learning processes when working with datasets;
5. Compute and interpret key statistical measures such as central tendency and dispersion for given datasets;
6. Explain the central role of variability in statistical analysis and its implications;
7. Apply fundamental probability principles, including sample space concepts and probability rules;
8. Compute the mean and variance for discrete distributions to understand their distributional characteristics;
9. Calculate and interpret probabilities using binomial, normal, and t-distributions in various contexts;
10. Differentiate between sample and population distributions and analyze the significance of the Central Limit Theorem in statistical inference;
11. Choose the most suitable inferential statistics method for resolving a given statistical problem and guiding decision-making;
12. Develop and explain confidence intervals for estimating means and proportions in one or two populations;
13. Execute hypothesis testing procedures for means and proportions in one or two populations and interpret the outcomes;
14. Determine and interpret the significance levels, including interpreting p-values;
15. Identify and describe Type I and II errors in statistical inference processes;
16. Conduct an analysis of variance (ANOVA) and interpret results;
17. Perform and interpret the result of the chi-squared goodness-of-fit test;
18. Conduct a linear regression analysis and apply the findings to make accurate predictions;
19. Compose code to perform various statistical analysis numerically and visually; and
20. Apply appropriate statistical methods to analyze, ethically interpret, and communicate findings based on real-world data from diverse disciplines, such as business, economics, education, psychology, and the social and life sciences.

General Education Information

- Approved College Associate Degree GE Applicability
  - AA/AS - Mathematical Concepts and Quantitative Reasoning
- CSU GE Applicability (Recommended-requires CSU approval)
- Cal-GETC Applicability (Recommended - Requires External Approval)
  - Cal-GETC 2 - Mathematical Concepts
- IGETC Applicability (Recommended-requires CSU/UC approval)

Articulation Information

- CSU Transferable
- UC Transferable

Methods of Evaluation

- Objective Examinations
  - Example: Economic study: Minimum Wage and Unemployment Rates 1. In an economic study investigating the impact of

minimum wage increases on unemployment rates, researchers collect data from different regions over several years. They aim to determine if there is a significant relationship between minimum wage hikes and changes in unemployment levels. After conducting hypothesis testing, the researchers find a p-value of 0.03, which is below the significance level of 0.05. What can be concluded from this result regarding the significance level and interpretation of the p-value in the context of the study? A) The significance level represents the probability of making a Type I error, while the p-value indicates the probability of making a Type II error. B) The significance level is the threshold for rejecting the null hypothesis, typically set at 0.05 or 0.01, while the p-value measures the strength of evidence against the null hypothesis. C) The significance level is calculated as 1 minus the p-value, representing the confidence level of the hypothesis test. D) The significance level is determined by the sample size, while the p-value is determined by the test statistic.

- Problem Solving Examinations
  - Example: Comparing Chances: Online Grocery Shopping In the United States, 28% of adults use Instacart for online grocery shopping. Suppose you sample US adults randomly so that each sampled adult has a chance 0.28 of being a Instacart user independently of all the others. (a) For which sample size below is there a higher chance that the percent of Instacart users in the sample will be at least 25%? # 200, 400 (b) For which sample size below is there a higher chance that the percentage of Instacart users in the sample will be at least 50%? # 200, 400 (c) For which sample size below is there a higher chance that the percentage of Instacart users in the sample will be at least 25% but less than 50%? # 200, 400 (d) (Briefly explain your choices in Parts (a)-(c). Grading: This problem is graded for correctness and meaningful justification for their choices.
- Projects
  - Example: Project Title: Analysis of Housing Prices in a Metropolitan Area Problem Statement: The housing market in a metropolitan area has experienced fluctuations in prices over the past few years. Homebuyers, sellers, and real estate investors are keen to understand the factors influencing these price fluctuations and predict future trends. The objective of this project is to analyze historical housing data, identify key factors affecting housing prices, and develop a predictive model to forecast future prices. Project Objectives: • Explore and clean the housing dataset to ensure data quality and integrity. • Conduct exploratory data analysis to uncover patterns, trends, and relationships among housing variables. • Identify relevant features such as location, property type, size, amenities, and economic indicators that impact housing prices. • Build regression models to predict housing prices based on selected features. • Evaluate model performance using appropriate metrics and techniques. • Interpret model results and identify significant predictors of housing prices. • Develop actionable insights and recommendations for homebuyers, sellers, and real estate professionals based on the analysis. Project Tasks: • Data Collection: Gather historical housing data from reliable sources such as real estate databases or government agencies. • Data Preprocessing: Clean and preprocess the dataset, handle missing values, and encode categorical variables. • Exploratory Data Analysis: Explore the dataset using descriptive statistics, visualizations, and correlation analysis to understand the relationships between variables. • Select Features: Select relevant features that are likely to influence housing prices. • Model Development: Build linear regression model to predict housing prices. • Model Evaluation: Assess model performance using metrics such as correlation coefficient and mean squared error. • Interpretation and Visualization: Interpret model coefficients, feature importance, and visualize key findings using graphs and charts. • Recommendations: Provide actionable insights and recommendations for homebuyers, sellers, and real estate professionals based on the analysis. This project will provide students with hands-on experience in data preprocessing, exploratory data analysis, regression modeling, and interpretation of results in the context of real estate market analysis. Students have the option to collaborate with a partner on this project. They can discuss the project with classmates or seek guidance during office hours from the instructor. However, sharing the code with anyone other than their partner is prohibited.

Repeatable

No

Methods of Instruction

- Laboratory
- Lecture/Discussion
- Distance Learning

Lab:

1. Data science entails analyzing real-world datasets, necessitating a series of labs and projects utilizing authentic data as integral components of the course. For each lab or project, students have the option to collaborate with one partner. The instructor will be accessible virtually during designated student hours to address queries and provide assistance. Students will have the opportunity to message the instructor for guidance if they are unable to attend the designated student hour. Projects are submitted online, and feedback will be provided by the instructor using a rubric. Healthy Living: We will investigate one of the major causes of death in the world: cardiovascular disease. It has two parts. • In Part 1, we'll investigate the major causes of death in the world during the past century (from 1900 to 2015). In order to get a better idea of how we can most effectively prevent deaths, we need to first figure out what the major causes of death are. Run the following cell to read in and view the causes_of_death table, which documents the death rate for major causes of deaths over the last century (1900 until 2015). • In Part 2, we'll look at data from the Framingham Heart Study, an observational study into cardiovascular health. Students will examine one of the main findings of the Framingham study: an association between serum cholesterol (i.e., how much cholesterol is in someone's blood) and whether or not that person develops heart disease using the hypothesis test. We will use the following null and alternative hypotheses: • Null Hypothesis: In the population, the distribution of cholesterol levels among those who develop heart disease is the same as the distribution of cholesterol levels among those who do not. • Alternative Hypothesis: The cholesterol levels of people in the population who develop heart disease are higher, on average, than the cholesterol level of people who do not.

Lecture:

1. Lesson plan: Example-1: Instructor will design a lesson plan that enables students to successfully complete complex lab assignments centered around real-world application problems. Begin with

straightforward examples to stimulate discussion and foster critical thinking, thereby preparing students to tackle more advanced problems. This assignment will be delivered through interactive computing environment. Let's start with a simple example to help you to complete more complex examples about disease in your next lab. Imagine you are a marble. You don't know what you look like (since you obviously have no eyes), but you know that Samy drew you uniformly at random from a bag that contained the following marbles: 4 large shiny marbles, 1 large dull marble, 6 small shiny marbles, and 2 small dull marbles. Question-

2. Knowing only what we've told you so far, what's the probability that you are a large shiny marble? (click on each cell below to see the table, value or graph) probability_large_shiny = ... Here's a table with those marbles: marbles = Table.read_table("marbles.csv") marbles.show() Here are the counts of each type of marble in a pivot table. marbles.pivot('surface', 'size') Here are all the different combinations of surface and size, with the count for each surface-size combination. Each type of marble appears in its own row. marbles.group(['surface', 'size']) Question-
3. What's the probability that you're a shiny marble? Calculate this by hand (using programming tool for arithmetic) by looking at your icon array. Conditional probability: Suppose you overhear Samy say that you are a large marble. Does this somehow change the chance that you are shiny? Let's find out. Question-
4. What's the probability you are a shiny marble, knowing that you are a large marble? Question-
5. Suppose instead Samy had said you are a shiny marble (hooray!). What's the probability that you're large? Run the code cell below to display the icon array, then assign probability_large_given_shiny to the appropriate value. Question-
6. Can you answer the previous two questions just by looking at the full icon array? Hopefully the icon arrays from the above example helped you build intuition for why conditional probabilities can be helpful. Now, let's look at a real-life application. Example-2: The instructor will develop a lesson plan centered around textbook readings to encourage self-directed learning through discussion. Discussion will be conducted through LMS while completing the problem through interactive computing environment. Instructor will provide guidance whenever needed. This discussion is based on textbook reading and an interesting example inspired by a mathematical theorem called "Infinite monkey theorem" (https://en.wikipedia.org/wiki/Infinite_monkey_theorem), which postulates that if you put a monkey in the situation described above for an infinite time, they will eventually type out all of Shakespeare's works. Read the three topics from chapter "Probability, Simulation, Estimation, and Assessing Models".
7. Randomness
8. Sampling and Empirical Distributions
9. Testing Hypotheses Monkeys Typing Shakespeare: A monkey is banging repeatedly on the keys of a keyboard. Each time, the monkey is equally likely to hit any of the 26 lowercase letters of the English alphabet, 26 uppercase letters of the English alphabet, and any number between 0-9 (inclusive), regardless of what it has hit before. There are no other keys on the keyboard. Question
10. Suppose the monkey hits the keyboard 5 times. Compute the chance that the monkey types the sequence Math1
11. (Call this data_chance.) Use algebra and type in an arithmetic equation that your program can evaluate. Question
12. Write a function called simulate_key_strike. It should take no arguments, and it should return a random one-character string that is equally likely to be any of the 26 lower-case English letters, 26 upper-case English letters, or any number between 0-9 (inclusive). Question
13. Write a function called simulate_several_key_strikes. It should take one argument: an integer specifying the number of key strikes to simulate. It should return a string containing that many characters, each one obtained from simulating a key strike by the monkey. Hint: If you make a list or array of the simulated key strikes called key_strikes_array, you can convert that to a string by calling "".join(key_strikes_array) Question
14. Call simulate_several_key_strikes 5000 times, each time simulating the monkey striking 5 keys. Compute the proportion of times the monkey types "Math11", calling that proportion data_proportion. Question
15. Check the value your simulation computed for data_proportion. Is your simulation a good way to estimate the chance that the monkey types "Math11" in 5 strikes (the answer to question 1)? Why or why not? Question
16. Compute the chance that the monkey types the letter "t" at least once in the 5 strikes. Call it t_chance. Use algebra and type in an arithmetic equation that program can evaluate. Question
17. Do you think that a computer simulation is more or less effective to estimate t_chance compared to when we tried to estimate data_chance this way? Why or why not?

Distance Learning

1. Discussion Forums: Students are encouraged to engage in discussions with their peers on each question, refraining from sharing their code. They can discuss mathematical concepts like "absolute value" and "proportion" to bridge gaps in their understanding. Discussions may be anonymized to facilitate open questioning without apprehension. Instructors will introduce specific topics to promote active learning. Instructor will initiate certain topics to enhance the active learning. Example: Discussion based on a virtual homework problem. Birth rate and death rate: The following table gives census-based population estimates for each state on both July 1, 2015 and July 1, 201
2. The last four columns describe the components of the estimated change in population during this time interval. For all questions below, assume that the word "states" refers to all 52 rows including Puerto Rico & the District of Columbia. The data was taken from censos.gov Run the cell below to clean the table and make it easier to work with. (A set of code will be provided to the students to generate the data.) Question
3. Assign us_birth_rate to the total US annual birth rate during this time interval. The annual birth rate for a year-long period is the total number of births in that period as a proportion of the population size at the start of the time period. Hint: Which year corresponds to the start of the time period? us_birth_rate = … us_birth_rate Question
4. Assign movers to the number of states for which the absolute value of the annual rate of migration was higher than 1%. The annual rate of migration for a year-long period is the net number of migrations (in and out) as a proportion of the population size at the start of the period. The MIGRATION column contains estimated annual net migration counts by state. Question
5. Assign west_births to the total number of births that occurred in region 4 (the Western US). Hint: Make sure you double check the type of the values in the region column, and appropriately filter (i.e. the types must match!). Question

6. Assign less_than_west_births to the number of states that had a total population in 2016 that was smaller than the total number of births in region 4 (the Western US) during this time interval. Question 5 : Create a visualization to understand the relationship between birth and death rates. The annual death rate for a year-long period is the total number of deaths in that period as a proportion of the population size at the start of the time period. What visualization is most appropriate to see if there is an association between birth and death rates during a given time interval?
7. Scatter Plot
8. Line Graph
9. Bar Chart Assign visualization below to the number corresponding to the correct visualization. Question
10. In the code cell below, create a visualization that will help us determine if there is an association between birth rate and death rate during this time interval. It may be helpful to create an intermediate table here. Question
11. True or False: There is an association between birth rate and death rate during this time interval. Assign assoc to True or False in the cell below. Instructor will provide the submission instructions.

Typical Out of Class Assignments

Reading Assignments

Weekly reading will be assigned based on the topics covered each week. The Textbook readings will include topics like various types of experimental study, four ways of classifying reality and the result of the test, how to construct confidence interval, and testing hypothesis, analyzing graphical and numerical data summaries, and identifying data characteristics appropriate for machine leaning techniques such as classification. Also, the textbook guide students through the sample programming exercises. Example 1. Read section "Why mean matters?" from the textbook and be prepare to discuss about the following questions. What exactly does the mean measure? How close is the mean to the most of the data? How does the sample size affect the variability of the sample mean? Why does empirical distributions of random sample means exhibit a bell-shaped curve?

Writing, Problem Solving or Performance

Writing assignments will be part of homework, lab and project which include summarizing and analyzing real-world data. Example: 1. Homework: Scary Arithmetic An ad for ADT Security Systems says, "When you go on vacation, burglars go to work [...] According to FBI statistics, over 25% of home burglaries occur between Memorial Day and Labor Day." Does the data in the ad support the claim that burglars are more likely to go to work during the time between Memorial Day and Labor Day? Please explain your answer. Note: You can assume that "over 25%" means only slightly over. Had it been much over, say closer to 30%, then the marketers would have said so. Write your answer here, replacing this text. 2. Programming: The cell below loads an array called president_birth_years. Calling .column(...) on a table returns an array of the column specified, in this case the Birth Year column of the president_births table. The last element in that array is the most recent birth year of any deceased president. Assign that year to most_recent_birth_year. Complete the code given below: president_birth_years = Table.read_table("president_births.csv").column('Birth Year') most_recent_birth_year = ... most_recent_birth_year Finally, assign sum_of_birth_years to the sum of the first, tenth, and last birth year in president_birth_years. sum_of_birth_years = ...

Other (Term projects, research papers, portfolios, etc.)

Students' complete projects using suitable tools to analyze real datasets and present their findings. Each project includes a given data set and a project notebook with questions that align with both in-class and out-of-class tasks.

Required Materials

- Computational and Inferential Thinking: The Foundations of Data Science
  - Author: Ani Adhikari, John DeNero, David Wagner
  - Publisher: UC Berkeley
  - Publication Date: 2022
  - Text Edition: 2
  - Classic Textbook?: No
  - OER Link:
  - OER: https://inferentialthinking.com/chapters/intro.html

Other materials and-or supplies required of students that contribute to the cost of the course.