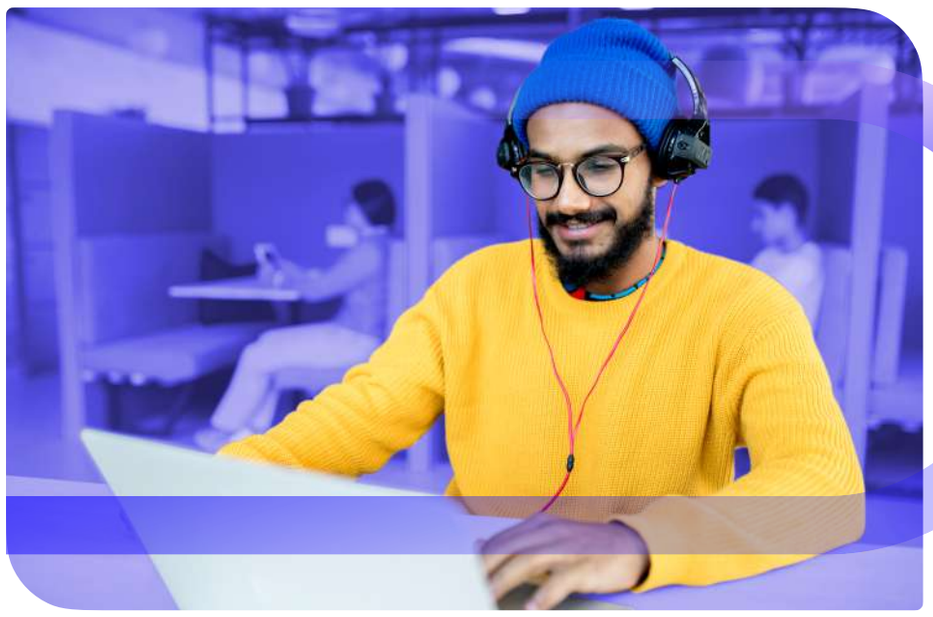
UDACITY

Statistics for Data Analysis

Syllabus

UDACITY.COM



BEFORE YOU START

Overview:

Students will learn essential skills including describing data, understanding probability theory, designing experiments, interpreting statistical results and applying statistical models with Python. After successfully completing this Nanodegree, graduates will be armed with a robust foundation in statistical analysis that can be applied to Data Analyst, Business Analyst, and Data Scientist roles.

Prerequisites

A WELL-PREPARED LEARNER HAS EXPERIENCE WITH:

- ✓ Basic Arithmetic
- ✓ Basic Algebra
- ✓ Basic Python

Educational Objectives

A GRADUATE OF THIS PROGRAM WILL BE ABLE TO:

- Describe data in terms of data types, measures of center, measures of spread, shape, and outliers
- Interpret notation for common mathematical expressions
- Calculate probabilities of independent events
- Apply Bayes Rule and calculate conditional probabilities
- Connect theoretical distributions such as the binomial distribution and normal distribution to real-world data
- Use Python bootstrapping to simulate data distributions
- Calculate and interpret confidence intervals in Python
- Perform and interpret the results of hypothesis tests using Python
- Apply power analysis and hypothesis tests to an A/B testing context
- Analyze relationships between independent variables and a numeric dependent variable using linear regression in Python
- Analyze relationships between independent variables and a categorical dependent variable using logistic regression in Python

LENGTH OF PROGRAM*:
3 months

SKILL LEVEL:
Beginner

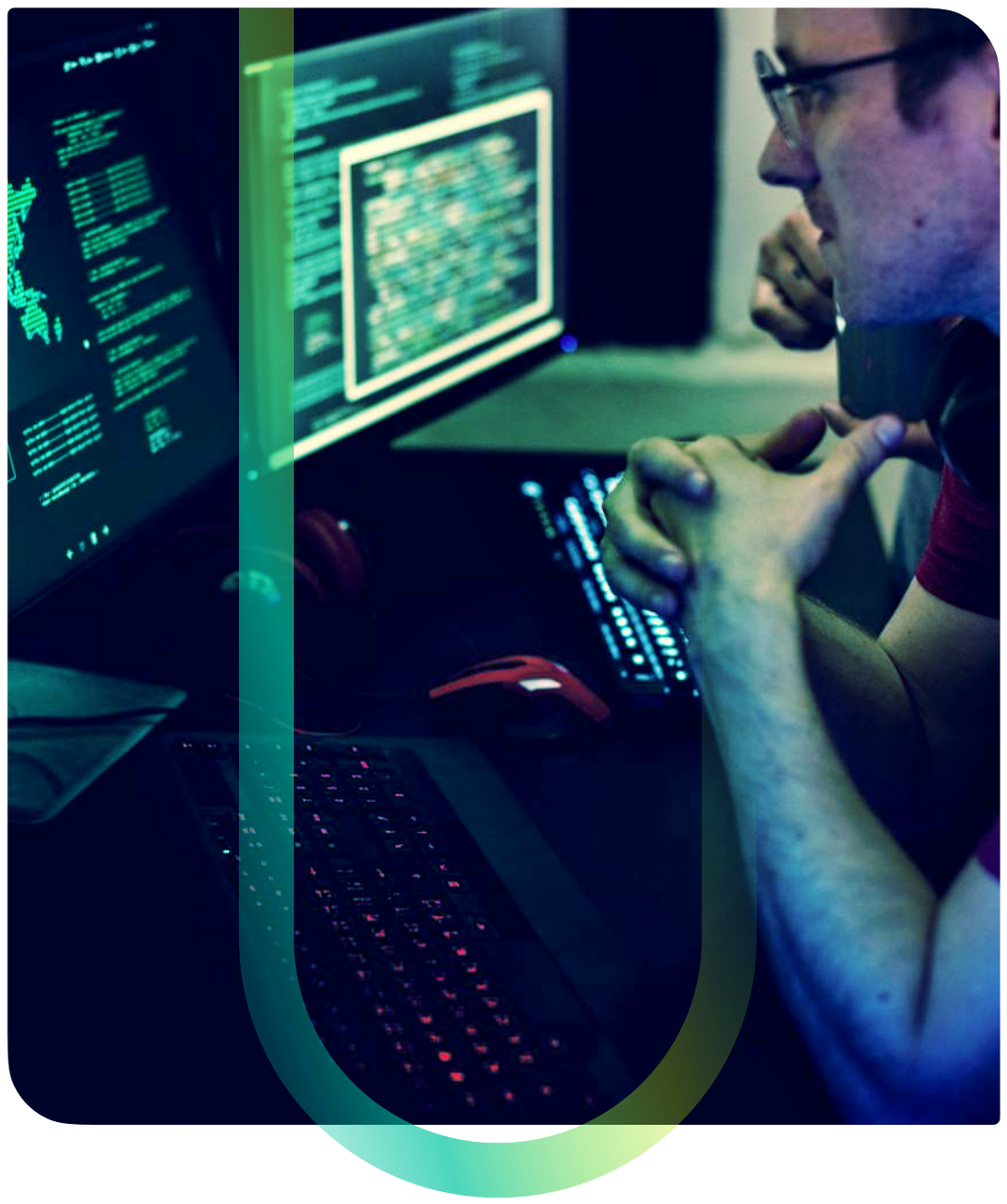
SCHOOL:
Data Science

SOFTWARE/HARDWARE AND VERSION REQUIREMENTS:

- **For this Nanodegree program, you will need access to the Internet.**
- **Additional software such as Python and its common data analysis libraries (e.g., NumPy and pandas) will be required, but the program includes Udacity Workspaces with all of the relevant packages installed, so students will not need to download any additional software.**

*The length of this program is an estimation of total hours the average student may take to complete all required coursework, including lecture and project time. If you spend about 5-10 hours per week working through the program, you should finish within the time provided. Actual hours may vary.

Course #1: Descriptive Statistics



IN THIS COURSE YOU WILL LEARN HOW TO:

- Identify and describe ordinal, nominal, continuous, and discrete data types
- Recognize and interpret the mathematical notation for common measures of center and spread
- Calculate and interpret measures of center
- Calculate and interpret measures of spread
- Describe the shape of a dataset using descriptive statistics and visual methods
- Identify outliers in a dataset using visual methods

HIGH-LEVEL SUMMARY:

Learn how to describe data in terms of data types, measures of center, measures of spread, shape, and outliers. These essential skills in descriptive statistics provide the foundation for more-advanced statistical techniques that are used for data science, data analysis, and machine learning.

Supporting Lesson Content

DATA TYPES

- Distinguish between ordinal and nominal data types
- Distinguish between continuous and discrete data types

MEASURES OF CENTER

- Calculate measures of center (mean, median, and mode) for a given dataset
- Interpret measures of center and the differences between them for a given dataset

MEASURES OF SPREAD

- Calculate measures of spread (range, interquartile range, standard deviation, variance) for a given dataset
- Interpret measures of spread and the differences between them for a given dataset

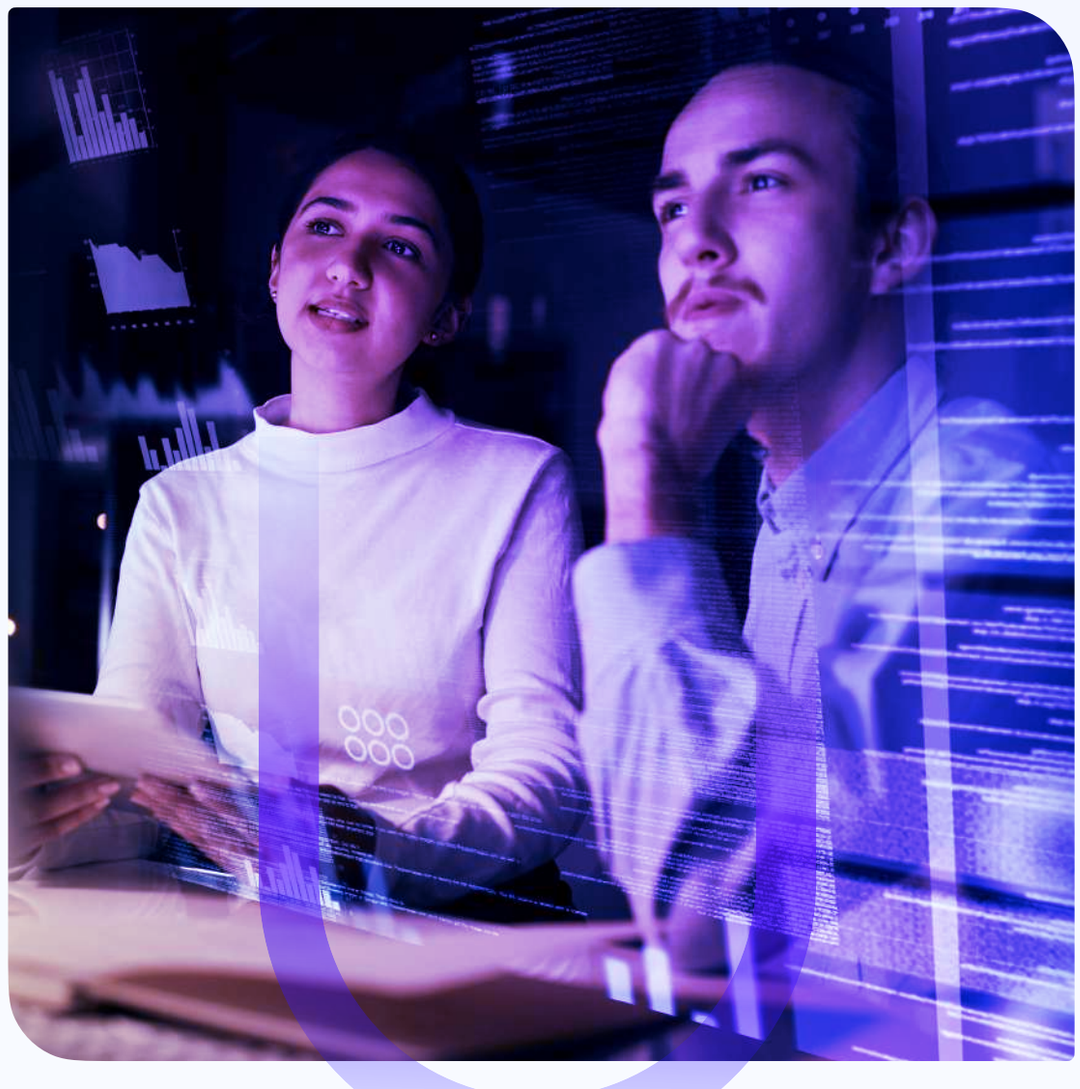
NOTATION

- Interpret notation for common mathematical expressions

SHAPE AND OUTLIERS

- Describe the shape of a distribution using visual methods like histograms and boxplots
- Identify outliers of a dataset using visual methods like histograms and box plots
- Explain the relationship between the shape of a distribution and different measures of center
- Explain the relationship between the shape of a distribution and different measures of spread

Course #2: Probability



IN THIS COURSE YOU WILL LEARN HOW TO:

- Calculate probabilities of independent events
- Describe the binomial distribution and use it to calculate probabilities
- Calculate conditional probabilities
- Apply Bayes' Rule in a real-world scenario

HIGH-LEVEL SUMMARY:

This course is a comprehensive dive into the fundamental concepts and principles of probability. You'll begin with basic probability theory, then progress to more complex topics such as binomial distributions, conditional probability, and Bayes' Rule. These skills will enhance your ability to reason about uncertainty and make claims using data.

Supporting Lesson Content

PROBABILITY

- Differentiate between probability and statistics
- Describe and calculate the probability of one event given another
- Use the law of total probability to calculate the probability of an event
- Use the complementary rule of probability to calculate the probability of an event
- Use truth tables to calculate the probability of independent and conditional events

CONDITIONAL PROBABILITY

- Describe and calculate conditional probabilities
- Calculate the probabilities of multiple dependent events

BINOMIAL DISTRIBUTION

- Identify scenarios where a binomial distribution can be used to calculate the probability of events
- Use the binomial distribution to calculate the probability of independent events

BAYES RULE

- Use conditional probability to calculate posterior probabilities from prior probabilities
- Apply Bayes' Rule to calculate the probability of events

Course #3: Experimentation

IN THIS COURSE YOU WILL LEARN HOW TO:

- Apply the normal distribution function to real-world data
- Use Python to create sampling distributions and apply the central limit theorem to a given dataset
- Calculate and interpret confidence intervals using Python
- Perform and interpret the results of hypothesis tests using Python
- Perform a power analysis to design an A/B test

HIGH-LEVEL SUMMARY:

Experimentation is one of the most important topics in all of statistics because it tells us whether our conclusions are statistically significant. In this course, you will learn about the fundamental role statistics plays in experimentation as well as how to implement statistical concepts in Python.

Supporting Lesson Content

NORMAL DISTRIBUTION THEORY

- Derive the maximum likelihood value from the normal distribution
- Derive the minimum likelihood value from the normal distribution
- Derive the normal distribution function step by step

SAMPLING DISTRIBUTIONS AND THE CENTRAL LIMIT THEOREM

- Differentiate between descriptive and inferential statistics
- Identify the population, parameter, sample, and statistic in a real world situation
- Explain the law of large numbers and central limit theorem
- Use bootstrapping to simulate the law of large numbers and central limit theorem

A/B TESTS

- Analyze results from A/B testing
- Test a hypothesis using Python
- Perform a power analysis
- Design an A/B test

CONFIDENCE INTERVALS

- Explain the relationship of sampling distributions and confidence intervals
- Use bootstrapping to create a confidence interval for a mean
- Use bootstrapping to create a confidence interval for a difference in means
- Differentiate between statistical and practical significance
- Interpret confidence intervals for a mean and difference in means

HYPOTHESIS TESTING

- Explain the importance and use of hypothesis testing
- Set up successful hypothesis tests
- Differentiate between the different types of errors in a hypothesis test
- Choose the correct hypothesis test for a given scenario
- Differentiate between an alternative and null hypothesis
- Calculate and interpret p-values
- Work with large sample sizes and multiple tests

Course #4: Algorithms

IN THIS COURSE YOU WILL LEARN HOW TO:

- Apply simple linear regression to solve problems using Python
- Apply multiple linear regression to solve problems using Python
- Apply logistic regression to solve problems using Python
- Interpret results and metrics from linear models

HIGH-LEVEL SUMMARY:

The algorithms course offers a detailed introduction to fundamental statistical and machine learning algorithms, particularly focusing on regression techniques. The course begins with simple linear regression and progresses to multiple linear regression, equipping students with the ability to analyze relationships between multiple variables. Finally, it covers logistic regression, a powerful tool for classification problems.

Supporting Lesson Content

LINEAR REGRESSION

- Identify regression applications in the real world
- Use correlation to describe relationships between variables
- Use mathematical notation to specify the properties of a regression line
- Set up regression problems using Python
- Interpret the coefficients from a linear regression model

LOGISTIC REGRESSION

- Differentiate between linear and logistic regression
- Set up logistic regression problems using Python
- Evaluate logistic regression models with classification metrics (confusion matrices, precision, and recall)

MULTIPLE LINEAR REGRESSION

- Differentiate between simple linear regression and multiple linear regression
- Set up multiple linear regression problems using Python
- Interpret coefficients in a multiple linear regression model
- Build dummy variables for regression modeling
- Interpret the coefficients of regression models associated with dummy variables
- Identify and address common issues with multiple linear regression
- Incorporate higher order terms and interactions for more complex linear models

Course #5: Capstone Project “Analyze A/B Test Results”

HIGH-LEVEL SUMMARY:

This capstone project gives you hands-on experience with A/B testing, a key practical application of statistical analysis. Students will produce a detailed technical analysis in a Jupyter Notebook, along with a non-technical slide deck presentation designed to effectively communicate their findings and recommendations to business stakeholders.

Nanodegree Program Instructors

Josh Bernhard

STAFF DATA SCIENTIST

Josh has been sharing his passion for data for over a decade. He's used data science for work ranging from cancer research to process automation. He recently has found a passion for solving data science problems within marketplace companies.

Sebastian Thrun

FOUNDER & PRESIDENT

As the founder and president of Udacity, Sebastian’s mission is to democratize education. He is also the founder of Google X, where he led projects including the Self-Driving Car, Google Glass, and more.

Learn More at

WWW.UDACITY.COM