Statistical Modeling and Inference with Python

Chester Ismay, August 2024

Welcome and Introduction

Dr. Chester Ismay

- PhD in Statistics
- Worked in academia, online education, corporate training, tech bootcamps, and independent consulting
- Currently,
  - Vice President of Data and Automation, MATE Seminars
  - Freelance data scientist
- Fun Fact: Slept a night or eaten a meal in all 50 US states

Course repo: https://github.com/ismayc/oreilly-statistical-modeling-and-inference-with-python

Learning Objectives

By the end of this course, you will be able to:

- Construct and interpret linear and non-linear regression models to understand relationships between variables and predict outcomes.
- Implement Analysis of Variance (ANOVA) to investigate differences across multiple group means and employ non-parametric tests for data that doesn't fit normal distribution assumptions.
- Learn bootstrapping methods to assess the reliability of sample statistics and construct confidence intervals to estimate population parameters with a quantifiable level of certainty.
- Apply statistical methods to real-world problems, enhancing decision-making processes in business, science, engineering, and other fields.
- Evaluate the robustness and validity of statistical models and results, ensuring accurate conclusions and recommendations from data analysis projects.

Week 1

Regression Analysis, Correlation Methods, and Analysis of Variance (ANOVA)

Agenda

- Week 1 Module 1: Linear Regression Fundamentals
- Week 1 Module 2: Correlation Analysis
- Week 1 Module 3: Multiple Regression Analysis
- Week 1 Module 4: Logistic Regression and Categorical Data Analysis
- Week 1 Module 5: Introduction to ANOVA

Discussion/Poll Question #1.A (For On24)
What are you most looking forward to in the course?

1. **Fundamental Understanding**: Gain a basic understanding of statistical modeling and its application with Python libraries.
2. **Hands-on Practice**: Apply theoretical knowledge through hands-on exercises and case studies.
3. **Regression Modeling**: Become proficient in using Python libraries for different stages of statistical modeling.
4. **Non-parametric Statistics**: Explore analyzing data that doesn't fit distribution assumptions using non-parametric tests and bootstrap techniques.
5. **Other**

Course repo: https://github.com/ismayc/oreilly-statistical-modeling-and-inference-with-python

Week 1 Module 1

Linear Regression Fundamentals

Introduction to Linear Regression

- **Definition**: A method for modeling the relationship between a dependent variable (y) and a single independent variable (x).
- **Importance**: Fundamental technique for predictive modeling and data analysis.
- **Basic Concept**: Fitting a line to data points to minimize the difference between observed and predicted values.

Assumptions of Linear Regression

- Linearity: The relationship between the dependent and independent variables is linear.
- Independence: Observations are independent of each other.
- Normality: Residuals of the model are normally distributed.
- Error homoscedasticity: Constant variance of residuals/errors.

Train-Test Splitting

- **Purpose**: Evaluate the model's performance on unseen data.
- **Process**: Splitting the dataset into training and testing sets.
- **Proportion**: Common split ratios (e.g., 80/20, 70/30).
- **Avoid Overfitting**: Ensures that the model generalizes well to new data.

Walkthrough and Exercise #1.1
Getting Started

By completing this exercise, you will be able to

1. Set up the Python environment.
2. Select a single feature and target variable.
3. Split the data into training and testing sets.
4. Create and train the model.
5. Make predictions.
6. Evaluate the model.
7. Check if assumptions of linear regression met with visual tools.

Course repo: https://github.com/ismayc/oreilly-statistical-modeling-and-inference-with-python

Questions and Answers

Anything I can clear up regarding the *Week 1 Module 1* content?

Review of Week 1 Module 1

Week 1 Module 2

Correlation Analysis

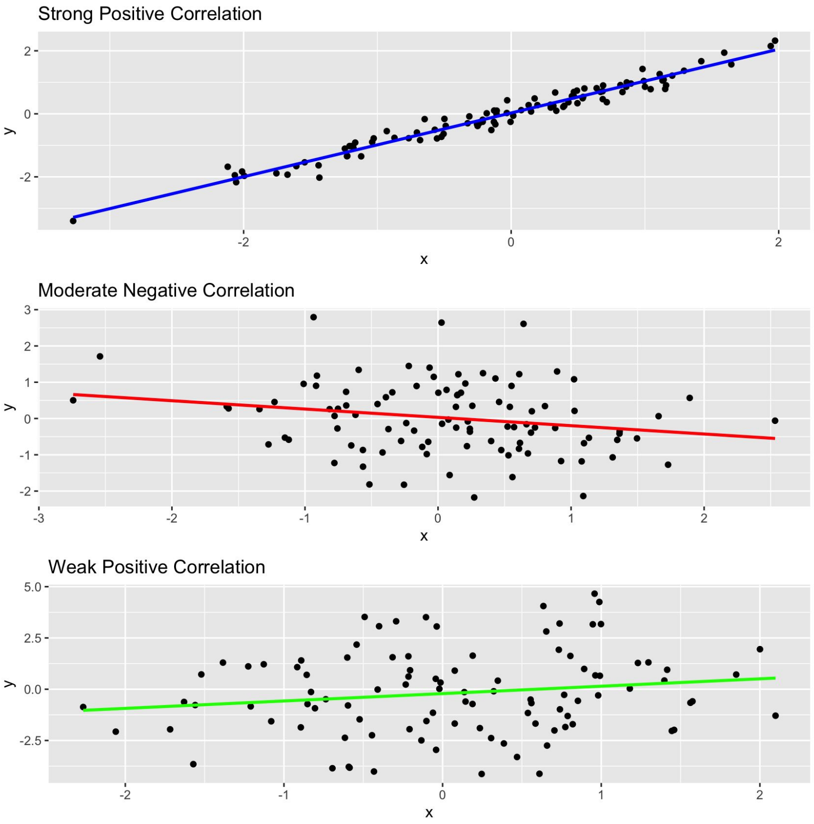
Discussion/Poll Question #1.B (For On24)
Which of the following do you think are key objectives of Correlation Analysis? (Select all that apply)

1. Understanding the strength of relationships between variables
2. Determining cause-and-effect relationships between variables
3. Visualizing the relationship between pairs of variables
4. Summarizing the distribution of individual variables
5. Predicting future values based on past data

Introduction to Correlation

- **Definition**: A statistical measure that describes the degree to which two variables move in relation to each other.
- **Range**: Values range from -1 to 1.
- **Types**: Positive, negative, and zero correlation.

Interpreting Correlation Coefficients



- **Strength**: $|r| > 0.7$ (Strong), $0.3 < |r| < 0.7$ (Moderate), $|r| < 0.3$ (Weak).
- **Direction**: Positive ($r > 0$), Negative ($r < 0$).
- **Limitations**: Only captures linear relationships, sensitive to outliers.

Course repo: https://github.com/ismayc/oreilly-statistical-modeling-and-inference-with-python

Walkthrough and Exercise #1.2
Correlation

By completing this exercise, you will be able to use `pandas` and `seaborn` to

1. Generate a correlation matrix for numeric columns in a DataFrame.
2. Visualize the correlation matrix using a heatmap.
3. Visualize relationships with a pairplot.

Course repo: https://github.com/ismayc/oreilly-statistical-modeling-and-inference-with-python

Questions and Answers

Anything I can clear up regarding the *Week 1 Module 2* content?

Review of Week 1 Module 2

Week 1 Module 3

Multiple Regression Analysis

Discussion/Poll Question #1.C (For On24)
Which of the following statements best describes the purpose of Multiple Regression Analysis? (Select one)

1. It identifies the strongest predictor variable in a dataset.
2. It assesses the combined impact of multiple independent variables on a single dependent variable.
3. It visualizes the relationship between two variables using a scatter plot.
4. It reduces the dimensionality of data by eliminating less important variables.
5. It categorizes data points into distinct groups based on their characteristics.

Course repo: https://github.com/ismayc/oreilly-statistical-modeling-and-inference-with-python

Introduction to Multiple Regression

- Extending linear regression to multiple predictors.
- Importance of multiple regression in statistical modeling.
- Applications in various fields such as finance, healthcare, and marketing.

Definition and Formula of Multiple Regression

- Definition: Models the relationship between a dependent variable and two or more independent variables.
- Equation: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$
- Terms:
  - Y is the dependent variable.
  - β_0 is the y-intercept.
  - $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients for each independent variable $X_1, X_2, \ldots, X_n$.
  - ϵ is the error term.

Assumptions of Multiple Regression

- **L**inearity: The relationship between the independent and dependent variables is linear.
- **I**ndependence: The residuals (errors) are independent.
- **N**ormality: The residuals of the model are normally distributed.
- **E**rror Homoscedasticity: The residuals have constant variance.
- **N**on-multicollinearity: The independent variables are not highly correlated with each other.

Course repo: https://github.com/ismayc/oreilly-statistical-modeling-and-inference-with-python

Difference between Simple and Multiple Regression

- **Simple Regression**: One dependent variable and one independent variable: $Y = \beta_0 + \beta_1 X + \epsilon$.
- **Multiple Regression**: One dependent variable and multiple independent variables: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$.
- **Complexity:** Multiple regression can model more complex relationships and control for more variables.
- **Interpretation**: Multiple regression allows for the assessment of the effect of each independent variable while holding other variables constant.

Walkthrough and Exercise #1.3
Multiple Regression

By completing this exercise, you will be able to use `pandas` and `statsmodels` to

1. Implement a multiple regression model using Python.
2. Interpret the output of multiple regression analysis.

Course repo: https://github.com/ismayc/oreilly-statistical-modeling-and-inference-with-python

Questions and Answers

Anything I can clear up regarding the *Week 1 Module 3* content?

Review of Week 1 Module 3

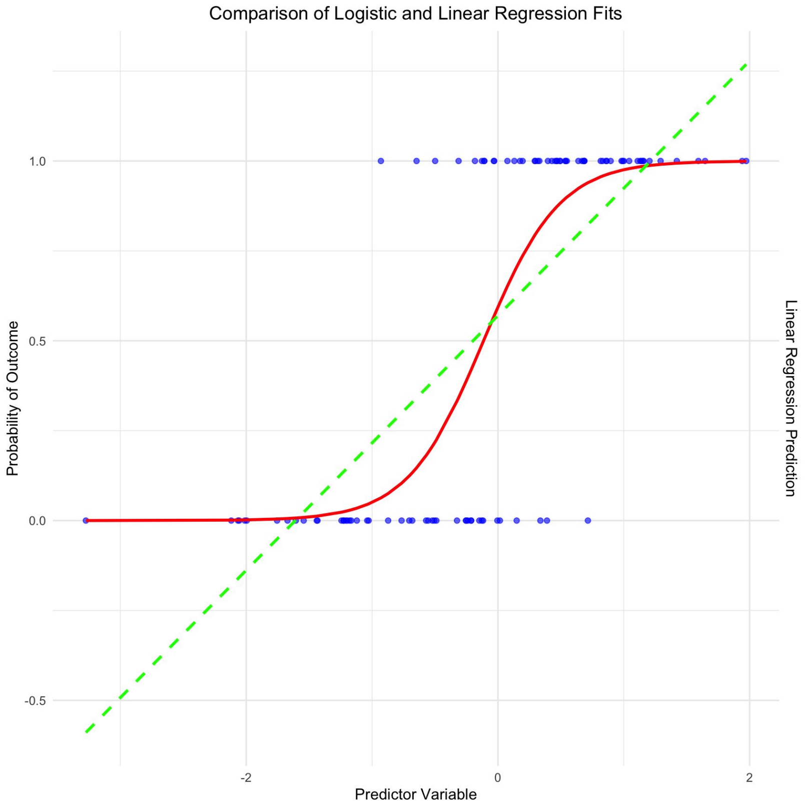
Week 1 Module 4

Logistic Regression and
Categorical Data Analysis

Discussion/Poll Question #1.D (For On24)

1. Predicting binary outcomes based on one or more predictor variables.
2. Visualizing data using scatter plots to observe linear relationships.
3. Estimating the probabilities of different outcomes.
4. Performing dimensionality reduction to simplify datasets.
5. Analyzing the strength and direction of linear relationships between variables.

Logistic Regression for Binary Outcomes



- Used for binary classification problems
- Predicts the probability of an outcome that can be only one of two values
- Different than linear regression
- Applications
  - Medical diagnosis
  - Credit scoring
  - Marketing

Course repo: https://github.com/ismayc/oreilly-statistical-modeling-and-inference-with-python

Logistic Regression Assumptions

- Dependent/target variable is binary.
- Observations are independent.
- Little or no multicollinearity among predictors.
- Linearity of independent variables and likelihood of target variable “success”

Walkthrough and Exercise #1.4
Logistic Regression

By completing this exercise, you will be able to use `statsmodels` and `seaborn` to

1. Fit a logistic regression model using statsmodels in Python.
2. Evaluate the performance of the logistic regression model.
3. Interpret the coefficients and performance metrics.

Course repo: https://github.com/ismayc/oreilly-statistical-modeling-and-inference-with-python

Questions and Answers

Anything I can clear up regarding the *Week 1 Module 4* content?

Review of Week 1 Module 4

Week 1 Module 5

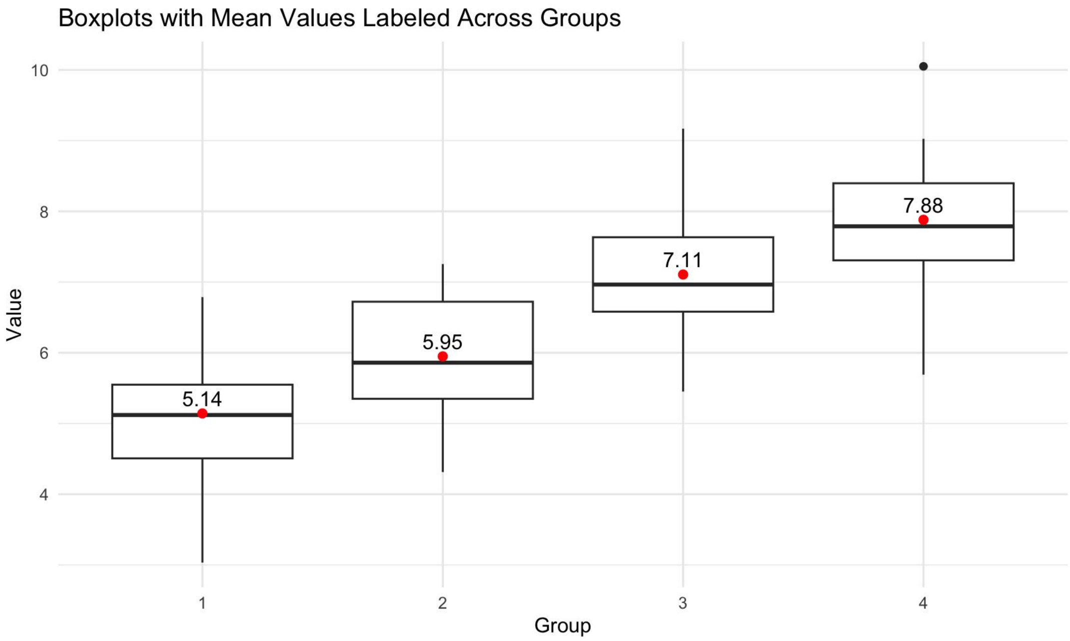
Introduction to ANOVA

Discussion/Poll Question #1.E (For On24)
Which of the following do you believe are primary goals of performing ANOVA?

1. Comparing means across multiple groups.
2. Determining the strength of linear relationships between variables.
3. Testing for significant differences between group means.
4. Reducing the dimensionality of datasets.
5. Identifying interactions between variables.

ANOVA Details

- Analysis of Variance
- Compares means of three or more groups
- Helps determine if observed differences are statistically significant



Course repo: https://github.com/ismayc/oreilly-statistical-modeling-and-inference-with-python

Principles of ANOVA

- ANOVA decomposes total variance into between-group and within-group variance.
- F-statistic: Ratio of between-group variance to within-group variance.
- P-value: Probability that observed differences are due to chance.

Assumptions of ANOVA

- Independence of observations.
- Homogeneity of variances (equal variances among groups).
- Normally distributed residuals.

Different common types of ANOVA

- One-way ANOVA
  - Compares means of three or more groups.
  - Example: Examining differences in test scores among different teaching methods.
- Two-way ANOVA
  - Examines the influence of two different categorical variables on one continuous dependent variable.
  - Example: Studying the effect of diet and exercise on weight loss.

Limitations of ANOVA

- Sensitive to violations of assumptions.
- Does not identify which groups are different.
- Post-hoc tests are necessary for detailed group comparisons.

Walkthrough and Exercise #1.5
Simulating Distributions

By completing this exercise, you will be able to use `statsmodels` and `seaborn` to

1. Perform a one-way ANOVA.
2. Perform a two-way ANOVA.
3. Produce boxplots to compare distributions across groups.

Course repo: https://github.com/ismayc/oreilly-statistical-modeling-and-inference-with-python

Questions and Answers

Anything I can clear up regarding the *Week 1 Module 5* content?

Review of Week 1 Module 5

O'REILLY

Statistical Modeling and Inference with Python

Chester Ismay, August 2024

Week 2

Non-parametric Tests, Bootstrapping Methods, and Confidence Intervals

Agenda

- Week 2 Module 1: Kruskal-Wallis Test and Mann-Whitney U Test
- Week 2 Module 2: Advanced Non-parametric Methods
- Week 2 Module 3: Introduction to Bootstrapping
- Week 2 Module 4: Constructing Confidence Intervals
- Week 2 Module 5: Applications of Bootstrapping in Real-World Scenarios

Week 2 Module 1

Kruskal-Wallis Test and Mann-Whitney U Test

Discussion/Poll Question #2.A (For On24)
Which of the following do you think are key reasons to use non-parametric tests like the Kruskal-Wallis and Mann-Whitney U tests? (Select all that apply)

1. To compare medians across multiple groups.
2. To assess the relationship between two categorical variables.
3. To test for differences in distributions without assuming normality.
4. To evaluate the strength of linear relationships.
5. To analyze data with unequal variances or outliers.

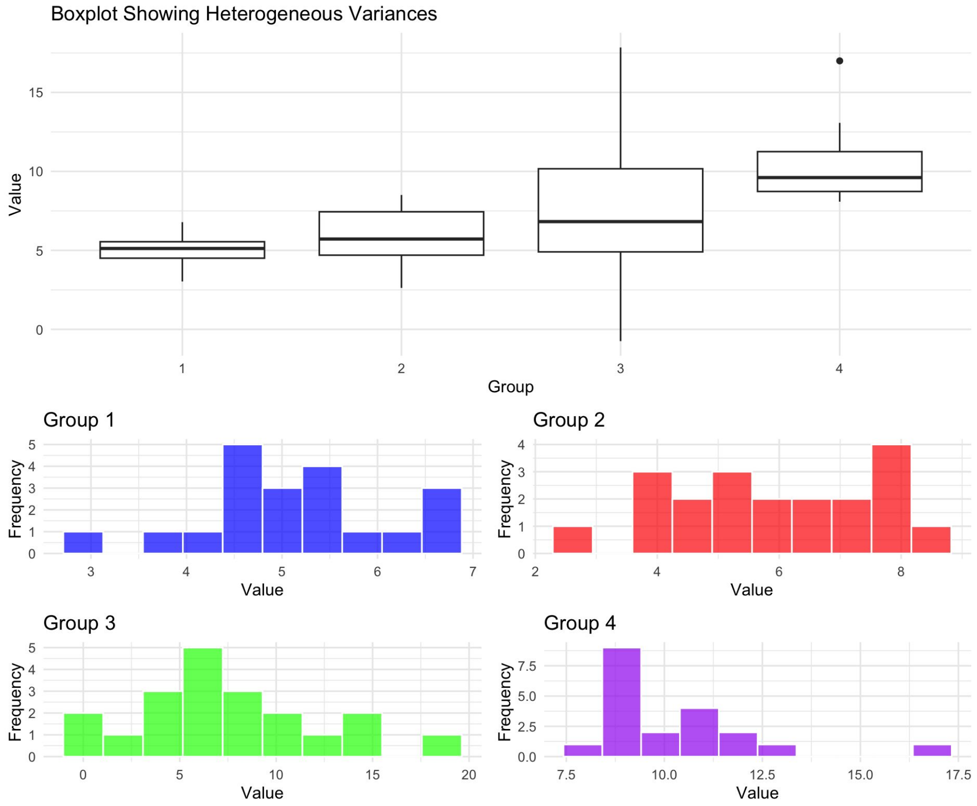
Non-parametric tests

- Introduction to non-parametric tests
- Kruskal-Wallis Test: Overview and when to use it
- Mann-Whitney U Test: Overview and when to use it
- Benefits of non-parametric tests

Course repo: https://github.com/ismayc/oreilly-statistical-modeling-and-inference-with-python

When to Use Non-Parametric Alternatives to ANOVA

- Situations where ANOVA assumptions are violated
- Kruskal-Wallis: more than two groups
- Mann-Whitney U: two independent groups



Course repo: https://github.com/ismayc/oreilly-statistical-modeling-and-inference-with-python

Walkthrough and Exercise #2.1

Kruskal-Wallis and Mann-Whitney U Tests

By completing this exercise, you will be able to use `scipy` to

1. Implement Kruskal-Wallis and Mann-Whitney U tests using Python.
2. Interpret the results of these non-parametric tests.

Course repo: https://github.com/ismayc/oreilly-statistical-modeling-and-inference-with-python

Questions and Answers

Anything I can clear up regarding the *Week 2 Module 1* content?

Review of Week 2 Module 1

Week 2 Module 2

Advanced Non-parametric Methods

Discussion/Poll Question #2.B (For On24)
Which of the following do you think are important reasons to use advanced non-parametric methods like Spearman, Kendall, and Theil-Sen? (Select all that apply)

1. To handle non-linear relationships in data.
2. To assess the strength and direction of monotonic relationships.
3. To perform regression analysis on binary outcomes.
4. To minimize the influence of outliers in correlation analysis.
5. To predict categorical outcomes using logistic regression.

Spearman's Rank Correlation

- **Definition**: Measures the strength and direction of the monotonic relationship between two variables.
- **Use Cases**: Useful when data do not meet the assumptions of parametric tests or when analyzing ordinal data.

Kendall's Tau Correlation

- **Definition:** Measures the strength and direction of association between two variables based on the ranks of the data.
- **Use Cases:** Preferred when the dataset has many tied ranks, providing a more accurate measure of correlation.

Theil–Sen Estimator

- **Definition:** A robust linear regression method that calculates the median of all possible slopes between pairs of points.
- **Use Cases:** Effective for datasets with outliers or non-normal error distributions.

Course repo: https://github.com/ismayc/oreilly-statistical-modeling-and-inference-with-python

Walkthrough and Exercise #2.2
Correlation and Regression Non-parametrics

By completing this exercise, you will be able to use `seaborn`, `scipy`, and `scikit-learn` to

1. Calculate Spearman's rank and Kendall's tau correlation coefficients.
2. Visualize non-parametric correlation matrices using a heatmap.
3. Implement the Theil-Sen estimator for robust regression.

Course repo: https://github.com/ismayc/oreilly-statistical-modeling-and-inference-with-python

Questions and Answers

Anything I can clear up regarding the *Week 2 Module 2* content?

Review of Week 2 Module 2

Week 2 Module 3

Introduction to Bootstrapping

Discussion/Poll Question #2.C (For On24)
Which of the following is the correct interpretation of a 95% confidence interval for a population mean?

1. There is a 95% probability that the population mean falls within the confidence interval.
2. If we were to take many samples and build a confidence interval from each sample, 95% of those intervals would contain the population mean.
3. 95% of the sample data falls within the confidence interval.
4. The population mean is guaranteed to be within the confidence interval.
5. 95% of the time, the sample mean will equal the population mean.

Bootstrapping Fundamentals

- **Concept**: A resampling method used to estimate the distribution of a statistic by sampling with replacement from the original data.
- **Methodology**: Involves repeatedly drawing samples from a dataset and calculating the statistic of interest for each sample.
- **Purpose**: Provides a way to estimate the sampling distribution of a statistic without making strong assumptions about the form of the population distribution.

Bootstrapping Methodology

- **Resampling:** Generate many resamples (typically 1000 or more) from the original dataset.
- **Statistic Calculation:** Compute the statistic of interest (mean, median, standard deviation, etc.) for each resample.
- **Distribution:** Use the distribution of these resampled statistics to make inferences about the population parameter.

Walkthrough and Exercise #2.3
Bootstrapping

By completing this exercise, you will be able to use a custom function and `pandas`, `matplotlib`, and `numpy` to

1. Perform bootstrapping.
2. Visualize the distribution of bootstrap estimates.
3. Estimate the mean and standard deviation visually for the mean of a population.

Course repo: https://github.com/ismayc/oreilly-statistical-modeling-and-inference-with-python

Questions and Answers

Anything I can clear up regarding the *Week 2 Module 3* content?

Review of
Week 2 Module 3

Week 2 Module 4

Constructing Confidence Intervals

Discussion/Poll Question #2.D (For On24)
Which of the following is the correct interpretation of a 95% confidence interval for a population mean?

1. There is a 95% probability that the population mean falls within the confidence interval.
2. If we were to take many samples and build a confidence interval from each sample, 95% of those intervals would contain the population mean.
3. 95% of the sample data falls within the confidence interval.
4. The population mean is guaranteed to be within the confidence interval.
5. 95% of the time, the sample mean will equal the population mean.

Course repo: https://github.com/ismayc/oreilly-statistical-modeling-and-inference-with-python

Discussion/Poll Question #2.C (For On24)
Which of the following do you think are key advantages of using bootstrapping in statistical analysis? (Select all that apply)

1. Estimating the confidence intervals of sample statistics.
2. Automating data cleaning and preprocessing.
3. Reducing the dependency on assumptions about the underlying population distribution.
4. Enhancing the visual appeal of data visualizations.
5. Evaluating the stability and reliability of population parameters.

Course repo: https://github.com/ismayc/oreilly-statistical-modeling-and-inference-with-python

Constructing Confidence Intervals

- **Definition:** Provide a range of values within which the true population parameter is likely to fall.
- **Importance:** They offer a measure of the precision of an estimate and help in understanding the uncertainty associated with sample statistics.
- **Interpretation:** A 95% confidence interval means that if we repeated the sampling process 100 times, approximately 95 of the intervals would contain the true population parameter.

Theory Behind Confidence Intervals

- **Central Limit Theorem:** The distribution of the sample mean approximates a normal distribution as sample size increases.
- **Margin of Error:** The range above and below the sample statistic within which the population parameter is expected to lie.
- **Confidence Level:** The probability that the confidence interval contains the true parameter (commonly 90%, 95%, or 99%).

Walkthrough and Exercise #2.4
Confidence Intervals

By completing this exercise, you will be able to use a custom function, `pandas`, and `matplotlib` to

1. Perform bootstrapping and calculate percentiles.
2. Calculate confidence intervals for sample means.
3. Visualize confidence intervals.

Course repo: https://github.com/ismayc/oreilly-statistical-modeling-and-inference-with-python

Questions and Answers

Anything I can clear up regarding the *Week 2 Module 4* content?

Review of Week 2 Module 4

Week 2 Module 5

Applications of Bootstrapping in Real-World Scenarios

Discussion/Poll Question #2.E (For On24)
Which of the following topics do you feel most confident about as this course concludes? (Select all that apply)

1. Understanding and interpreting correlation coefficients.
2. Conducting multiple regression analysis.
3. Implementing and evaluating logistic regression models.
4. Performing ANOVA and understanding its applications.
5. Applying non-parametric tests like Kruskal-Wallis and Mann-Whitney U.
6. Using bootstrapping to estimate statistics and construct confidence intervals.

Course repo: https://github.com/ismayc/oreilly-statistical-modeling-and-inference-with-python

Applications of Bootstrapping in Real-World Scenarios

- **Case Studies**: Examples of bootstrapping in finance, healthcare, and marketing.
- **Industry Applications**: How different industries use bootstrapping to estimate statistics and make decisions.
- **Bootstrapping Benefits**: Advantages such as making no assumptions about the distribution and flexibility.

Case Study 1: Bootstrapping in Finance

- Risk Assessment: Estimating Value at Risk (VaR) for investment portfolios.
- Stock Prices: Predicting stock price trends and their volatility.
- Portfolio Optimization: Using bootstrapping to optimize asset allocation.

Case Study 2: Bootstrapping in Healthcare

- Clinical Trials: Estimating confidence intervals for treatment effects.
- Survival Analysis: Bootstrapping to analyze patient survival rates.
- Diagnostic Tests: Assessing the accuracy of medical tests.

Case Study 3: Bootstrapping in Marketing

- Customer Segmentation: Bootstrapping to understand customer demographics.
- Sales Forecasting: Predicting future sales and revenue.
- Campaign Analysis: Measuring the effectiveness of marketing campaigns.

Walkthrough and Exercise #2.5
Real-World Scenarios for Bootstrapping

By completing this exercise, you will be able to use custom function, `pandas`, and `matplotlib` to

1. Implement bootstrapping.
2. Calculate and interpret bootstrapped statistics.
3. Apply bootstrapping to a real-world dataset.

Course repo: https://github.com/ismayc/oreilly-statistical-modeling-and-inference-with-python

Questions and Answers

Anything I can clear up regarding the *Week 2 Module 5* content?

Review of Week 2 Module 5

Learning Objectives

By the end of this course, you will be able to:

- Construct and interpret linear and non-linear regression models to understand relationships between variables and predict outcomes.
- Implement Analysis of Variance (ANOVA) to investigate differences across multiple group means and employ non-parametric tests for data that doesn't fit normal distribution assumptions.
- Learn bootstrapping methods to assess the reliability of sample statistics and construct confidence intervals to estimate population parameters with a quantifiable level of certainty.
- Apply statistical methods to real-world problems, enhancing decision-making processes in business, science, engineering, and other fields.
- Evaluate the robustness and validity of statistical models and results, ensuring accurate conclusions and recommendations from data analysis projects.

Conclusion

Additional resources:

- numpy
- pandas
- matplotlib
- seaborn
- scipy
- statsmodels
- scikit-learn

LinkedIn: https://www.linkedin.com/in/chesterismay/

Personal website: https://chester.rbind.io/

Images generated with DALL-E, in Python/R, or via Creative Commons Google Image search

O'REILLY®