O'REILLY®

Fundamentals of Statistics with Python

Chester Ismay

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-fundamentals-of-statistics-with-python

Learning Objectives

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

- Use Python for complex statistical analyses, leveraging libraries like NumPy, Pandas, SciPy, and Matplotlib for data manipulation and visualization.
- Uncover underlying patterns, trends, and anomalies in datasets using exploratory data analysis.
- Model various types of data with probability distributions and utilize hypothesis testing to validate data-driven inferences.

Week 1

Foundations of Statistical Analysis

Agenda

- Week 1 Module 1: Statistical Concepts and Python's Statistical Libraries
- Week 1 Module 2: Descriptive Statistics and Data Summarization
- Week 1 Module 3: Data Visualization Fundamentals
- Week 1 Module 4: Inferential Statistics Introduction

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 analysis and its application with Python libraries.
2. **Hands-on Practice**: Apply theoretical knowledge through hands-on exercises and case studies.
3. **Tool Proficiency**: Become proficient in using Python libraries like NumPy, Pandas, and SciPy for different stages of statistical analysis.
4. **Inferential Statistics**: Explore hypothesis testing to understand how to make predictions and inferences from data samples using Python.
5. **Other**

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Week 1 Module 1

Statistical Concepts and Python's Statistical Libraries

Overview of Statistical Analysis

- What is Statistical Analysis?
- Why is it important?
- Examples of fields where it is used commonly

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Types of Statistical Analysis

- Descriptive Statistics
  - Summarizes data from a sample
- Inferential Statistics
  - Make predictions about a population based on a sample

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Python's Ecosystem for Statistics

- Key Libraries
  - NumPy
  - Pandas
  - SciPy
- Other Useful Libraries
  - Matplotlib
  - Seaborn
  - Statsmodels

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Walkthrough and Exercise #1.1
Getting Started

By completing this exercise, you will be able to

1. Set up the Python environment.
2. Explore a dataset.
3. Perform basic statistical functions using NumPy, Pandas, and SciPy.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-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

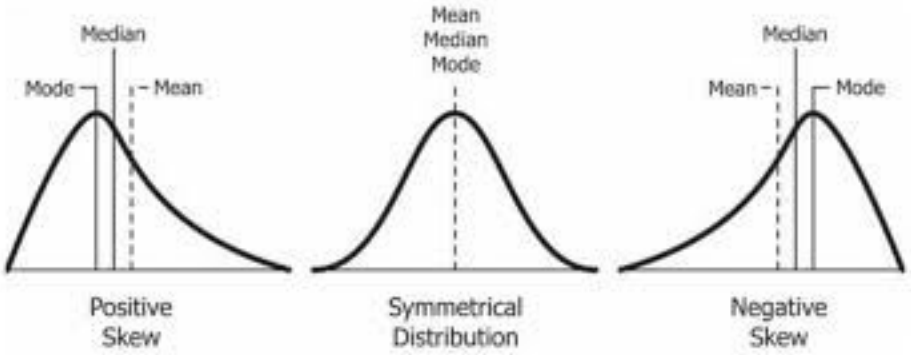
Descriptive Statistics and Data Summarization

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

1. Calculating the average value of a dataset
2. Predicting future trends based on historical data
3. Summarizing and describing the main features of a dataset
4. Identifying patterns and relationships within the data
5. Measuring the spread and variability of the data

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

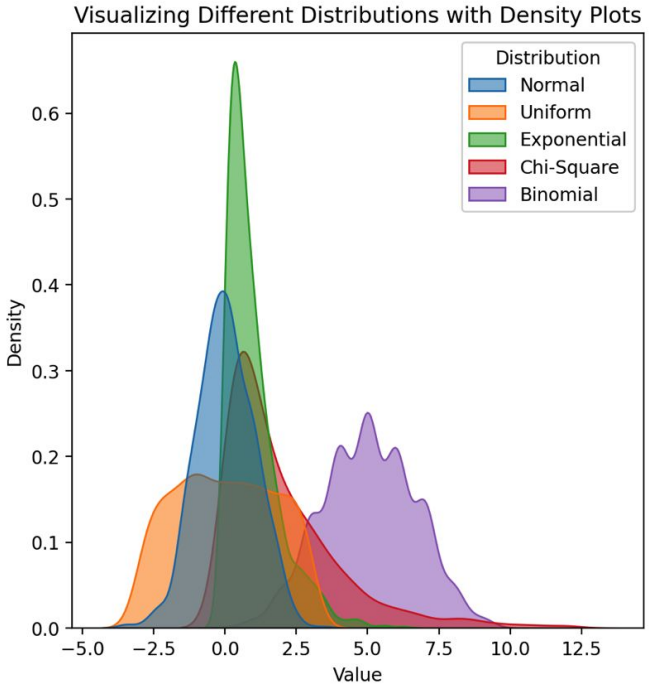
Understanding Measures of Central Tendency



- Mean: The average value of a dataset.
- Median: The middle value when the data is sorted.
- Mode: The most frequently occurring value in the dataset.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Understanding Measures of Variability



- Range
- Variance
- Standard Deviation
- Percentiles

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Walkthrough and Exercise #1.2
Data Summarizing

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

1. Compute and interpret measures of central tendency for different columns in a DataFrame.
2. Compute and interpret measures of variation for different columns in a DataFrame.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-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

Data Visualization Fundamentals

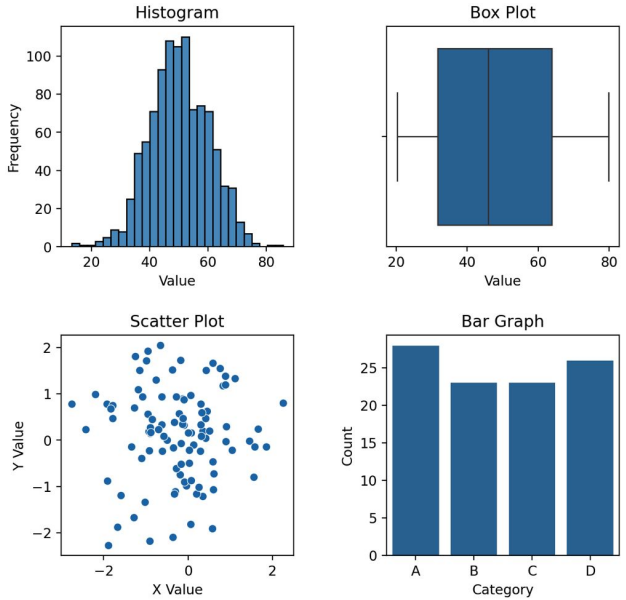
Discussion/Poll Question #1.C (For On24)
Which of the following statements about data visualization tools and techniques is true? (Select all that apply)

1. Matplotlib is a Python library primarily used for creating static visualizations.
2. Seaborn is a data visualization library based on Matplotlib that provides a high-level interface for drawing attractive statistical graphics.
3. Histograms are best used to visualize the relationship between two variables.
4. Box plots are useful for identifying the central tendency and spread of data, as well as detecting outliers.
5. Scatter plots are effective for displaying the distribution of a single variable.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Role of Visualization in Statistics

- Importance of data visualization
- Types of visualizations



Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Matplotlib Foundations

- `matplotlib`: versatile library for static plots
- Common chart types
- Customization options

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Seaborn Basics

- `seaborn` vs. `matplotlib`
- Color palettes and themes

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Walkthrough and Exercise #1.3
Data Visualization with matplotlib and seaborn

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

1. Create and interpret a histogram with title and labels.
2. Create and interpret a boxplot with title and labels.
3. Create an interpret a scatter plot with title and labels.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-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

Inferential Statistics Introduction

Discussion/Poll Question #1.D (For On24)
Which of the following statements about inferential statistics is true? (Select all that apply)

1. Inferential statistics involves making predictions about a population based on a sample.
2. Descriptive statistics and inferential statistics serve the same purpose in data analysis.
3. The Central Limit Theorem applies in only a few contexts.
4. A sampling distribution is the distribution of a sample statistic over repeated sampling from the same population.
5. Inferential statistics does not rely on probability theory.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Descriptive vs. Inferential Statistics

- Descriptive
  - Mean
  - Median
  - Standard deviation
- Inferential
  - Hypothesis tests
  - Confidence intervals

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Introduction to sampling distributions

- Definition: Distribution of a statistic from multiple samples.
  - Not to be confused with sample distribution
- Importance
  - Inferences about population, precision estimation.
- Central Limit Theorem
  - Sample means distribution approaches normality with larger sample sizes.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Walkthrough and Exercise #1.4
Sampling Distribution Generation

By completing this exercise, you will be able to use `matplotlib` and a `for` loop to

1. Generate a sampling distribution for a given sample size.
2. Create a histogram of the sampling distribution.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Questions and Answers

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

Review of Week 1 Module 4


O'REILLY®
Fundamentals of Statistics with Python
Chester Ismay

Week 2

Exploratory Data Analysis and Visualization Techniques

Agenda

- Week 2 Module 1: Advanced Data Visualization Techniques
- Week 2 Module 2: Exploratory Data Analysis (EDA)
- Week 2 Module 3: Data Preprocessing for Statistical Analysis
- Week 2 Module 4: Correlation and Causation

Week 2 Module 1

Advanced Data Visualization Techniques

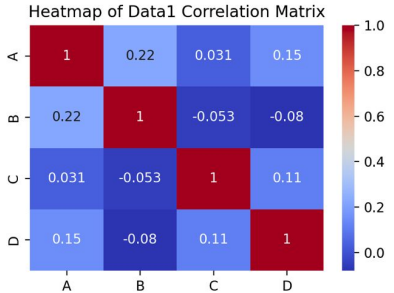
Discussion/Poll Question #2.A (For On24)
Which of the following statements about advanced data visualization techniques is true? (Select all that apply)

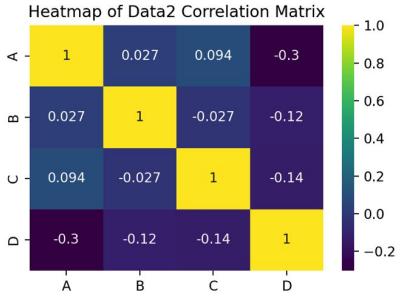
1. Heatmaps are useful for visualizing correlation matrices.
2. Pair plots are ideal for examining the relationship between two variables over time.
3. Time series visualizations are best for displaying changes in data points over time.
4. Heatmaps can only be used for visualizing geographical data.
5. Pair plots can help in identifying patterns among multiple variables simultaneously.

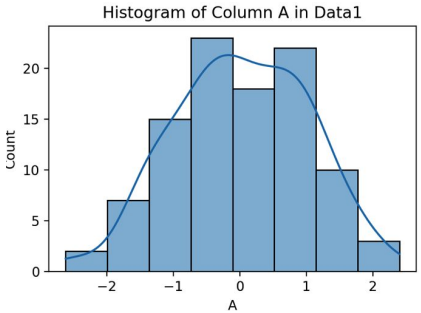
Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

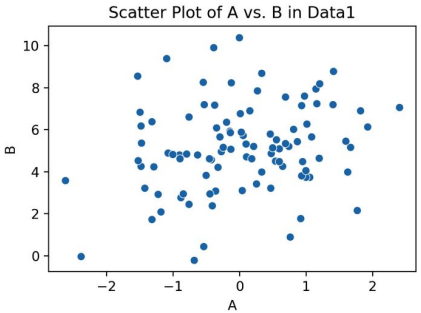
Exploring Complex Visualizations

- Heatmaps
- Pair Plots
- Time Series Plots









Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Walkthrough and Exercise #2.1
Advanced Plots

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

1. Create a heatmap for a correlation matrix.
2. Generate a pair plot for exploring relationships between multiple variables.
3. Plot a time series graph to visualize trends over time.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-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

Exploratory Data Analysis (EDA)

Discussion/Poll Question #2.B (For On24)
Which of the following statements about Exploratory Data Analysis (EDA) is true? (Select all that apply)

1. EDA primarily focuses on confirming hypotheses with statistical tests.
2. One of the main goals of EDA is to understand the data structure.
3. EDA involves using visual methods to summarize the main characteristics of the data.
4. EDA is mainly used for automating data analysis processes.
5. EDA aims to prepare data for further analysis by cleaning and transforming it.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

The EDA Process and Its Objectives

- Understand the data structure
- Detect outliers and anomalies
- Test hypotheses
- Establish relationships
- Techniques: Data visualization, summary statistics, and data transformation

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Steps in EDA

- **Data Collection**: Gather the data
- **Data Cleaning**: Handle missing values, remove duplicates
- **Data Transformation**: Normalize, scale, and encode data
- **Data Visualization**: Use plots and charts to visualize data
- **Summary Statistics**: Calculate mean, median, mode, etc.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Walkthrough and Exercise #2.2
EDA

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

1. Clean the data (handle missing values, remove duplicates).
2. Calculate summary statistics for key variables.
3. Create visualizations (histograms, box plots, scatter plots).

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-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

Data Preprocessing for Statistical Analysis

Discussion/Poll Question #2.C (For On24)
Which of the following data preprocessing steps do you find most challenging? (Select one)

1. Identifying and handling missing values
2. Detecting and managing outliers
3. Normalizing and standardizing data
4. Encoding categorical variables
5. (Unsure/Doesn't apply to me yet)

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Importance of Data Preprocessing

- Ensures data quality and consistency
- Prepares data for accurate statistical analysis
- Involves handling missing values, outliers, and data transformation

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Handling Missing Values

- Identification
- Techniques:
  - Removal: Drop rows/columns with missing values
  - Imputation: Fill missing values with mean, median, mode, or other techniques

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Handling Outliers

- Identification: Use box plots, z-scores, or IQR method
- Techniques:
  - Removal: Remove data points beyond a certain threshold
  - Transformation: Apply log or square root transformation

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Data Transformation

- Normalization: Scale data to a standard range
- Standardization: Scale data to have mean 0 and standard deviation 1
- Encoding Categorical Variables: Convert categorical data to numerical using one-hot encoding or label encoding

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Walkthrough and Exercise #2.3
Data Preprocessing

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

1. Handle missing values by imputation.
2. Identify and handle outliers.
3. Transform the data using normalization and encoding techniques.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-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

Correlation and Causation

Discussion/Poll Question #2.D (For On24)
Which of the following best describes the relationship between correlation and causation? (Select all that apply)

1. If two variables are correlated, one variable must be causing the other.
2. Correlation means that two variables move together, but it doesn't imply causation.
3. Causation always results in a high correlation between two variables.
4. Correlation can only be determined through experimentation.
5. There is no difference between correlation and causation.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Understanding Correlation

- **Definition**: Correlation measures the strength and direction of a linear relationship between two variables.
- **Types**:
  - **Positive Correlation**: Both variables increase together.
  - **Negative Correlation**: One variable increases while the other decreases.
  - **Zero Correlation**: No linear relationship between variables.
- **Measurement**: Correlation coefficient (*r*), ranges from -1 to 1.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Understanding Causation

- **Definition**: Causation implies that changes in one variable directly cause changes in another.
- Difference from Correlation: Correlation does not imply causation.
- Examples:
  - Correlation without Causation: Ice cream sales and drowning incidents.
  - Causation Example: Smoking and lung cancer.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Walkthrough and Exercise #2.4
Correlations

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

1. Calculate the correlation matrix.
2. Visualize the correlation matrix using a heatmap.
3. Create scatter plots for pairs of variables with meaningful correlations.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Questions and Answers

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

Review of Week 2 Module 4


O'REILLY®
Fundamentals of Statistics with Python
Chester Ismay

Week 3

Probability Distributions and Hypothesis Testing

Agenda

- Week 1 Module 1: Statistical Concepts and Python's Statistical Libraries
- Week 1 Module 2: Descriptive Statistics and Data Summarization
- Week 1 Module 3: Data Visualization Fundamentals
- Week 1 Module 4: Inferential Statistics Introduction

Agenda

- Week 2 Module 1: Advanced Data Visualization Techniques
- Week 2 Module 2: Exploratory Data Analysis (EDA)
- Week 2 Module 3: Data Preprocessing for Statistical Analysis
- Week 2 Module 4: Correlation and Causation

Agenda

- Week 3 Module 1: Probability Distributions
- Week 3 Module 2: Fundamentals of Hypothesis Testing
- Week 3 Module 3: Comparing Two or More Groups
- Week 3 Module 4: Introduction to Non-Parametric Tests

Week 3 Module 1

Probability Distributions

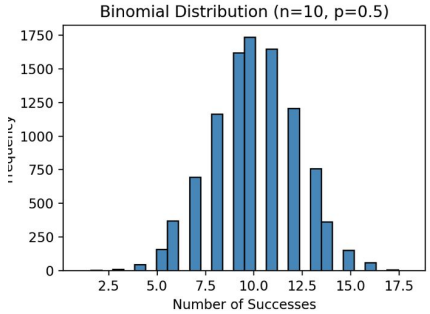
Discussion/Poll Question #3.A (For On24)
Which of the following best describes your current understanding of probability distributions? (Select one)

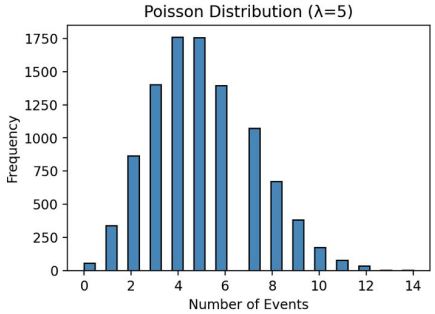
1. I am familiar with basic probability concepts but have limited knowledge of different probability distributions.
2. I have a good understanding of common probability distributions like normal and binomial but need more practice with others.
3. I can calculate and interpret probability distributions but struggle with visualizing them.
4. I am confident in my ability to simulate and analyze various probability distributions using Python.
5. I have no prior experience with probability distributions.

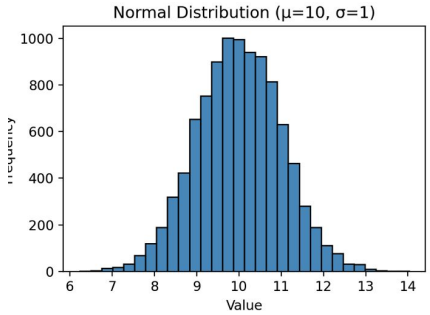
Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

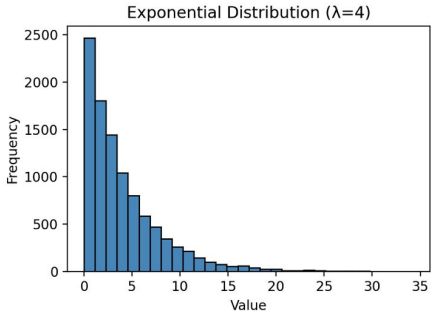
Common Probability Distributions

- Discrete
  - Binomial
  - Poisson
- Continuous
  - Normal
  - Exponential





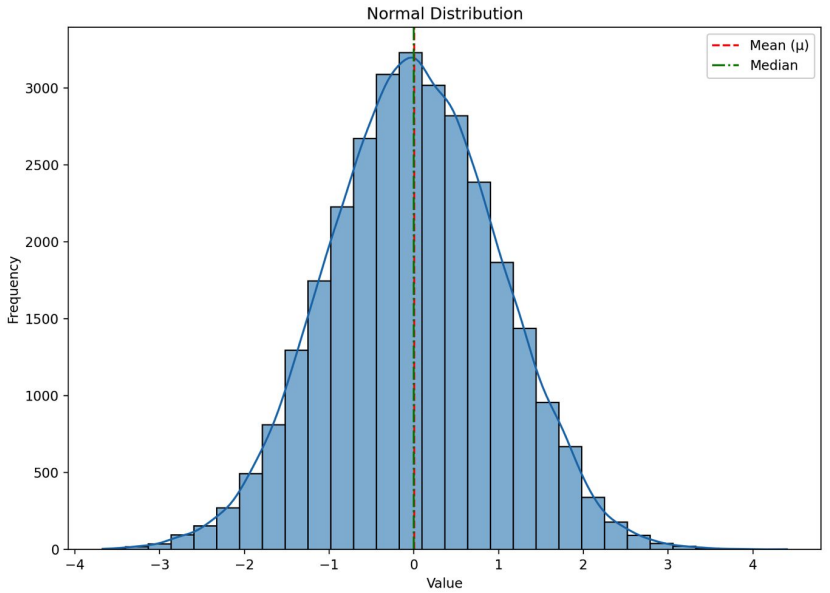




Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Properties of the Normal Distribution

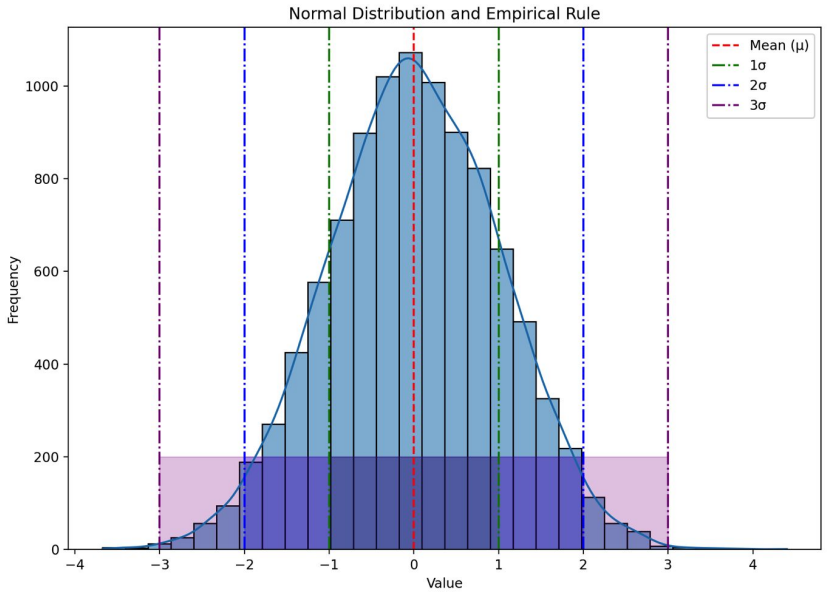
- Characteristics
  - Symmetrical & bell-shaped
  - Mean and median are equal
- Parameters
  - Mean (μ): Location
  - Standard Deviation (σ): Scale



Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Empirical Rule for the Normal Distribution

- 68% of data within 1 standard deviation
- 95% within 2 standard deviations
- 99.7% within 3 standard deviations



Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Walkthrough and Exercise #3.1
Simulating Distributions

By completing this exercise, you will be able to use `numpy` and `matplotlib/seaborn` to

1. Generate random samples from binomial, Poisson, normal, and exponential distributions.
2. Visualize these distributions using histograms and density plots.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Questions and Answers

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

Review of Week 3 Module 1

Week 3 Module 2

Fundamentals of Hypothesis Testing

Discussion/Poll Question #3.B (For On24)
Which of the following statements about hypothesis testing is true? (Select all that apply)

1. The null hypothesis is typically a statement of no effect or no difference.
2. A Type I error occurs when we fail to reject a false null hypothesis.
3. The p-value helps determine whether to reject the null hypothesis.
4. A t-test is used to compare the means of two groups when the population variance is known.
5. The alternative hypothesis represents the hypothesis that the test aims to support.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Hypothesis Testing Introduction

- Definition: A statistical method used to make inferences about population parameters based on sample data.
- Purpose: To determine whether there is enough evidence to reject a null hypothesis in favor of an alternative hypothesis.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Formulating Hypotheses

- Null Hypothesis (H_0): The statement being tested, typically represents no effect or no difference.
- Alternative Hypothesis (H_a): The statement we want to test for, represents an effect or a difference.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Types of Errors

	Truly not guilty	Truly guilty
Verdict		
Not guilty verdict	Correct	Type II error
Guilty verdict	Type I error	Correct

- Type I Error (α): Rejecting the null hypothesis when it is true (false positive).
- Type II Error (β): Failing to reject the null hypothesis when it is false (false negative).
- Significance Level (α): The probability of making a Type I error, commonly set at 0.05.

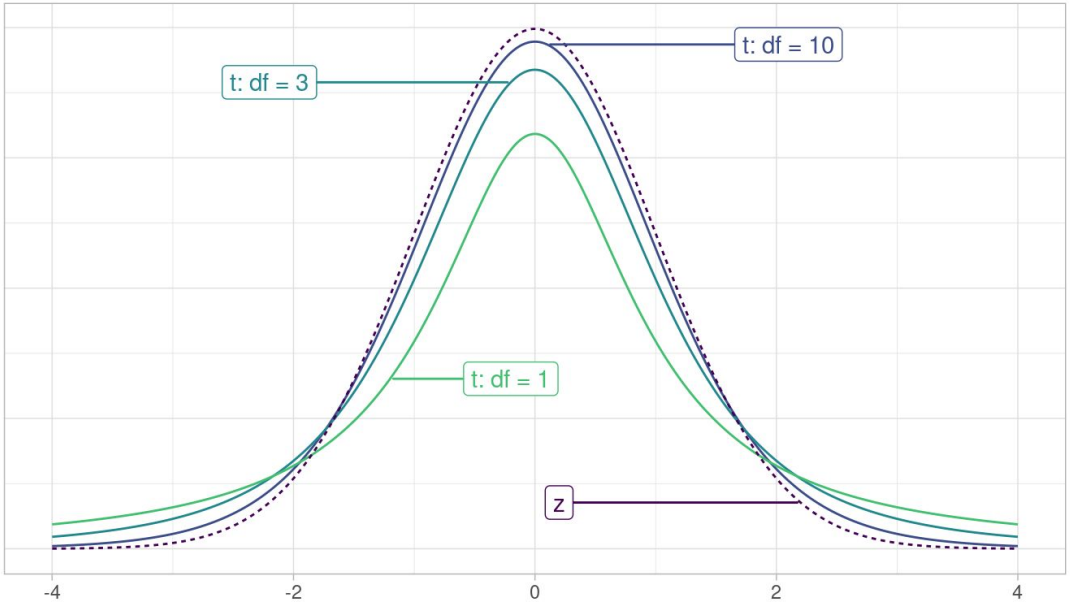
Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Common Hypothesis Tests

- Z-Test: Used when the population variance is assumed to be known and the sample size is large ($n > 30$).
- T-Test: Used when the population variance is unknown and the sample size is small ($n \leq 30$).
- Two-Sample T-Test: Used to compare means from two different groups.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

The *t* Curve



- Heavier tails than normal distribution
- Used when sample size is small
- As sample size increases, *t* approaches normal

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Walkthrough and Exercise #3.2
t-tests

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

1. Perform and interpret the results of a one-sample t-test.
2. Conduct a two-sample t-test comparing a numeric result between two groups and interpret its results.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Questions and Answers

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

Review of Week 3 Module 2

Week 3 Module 3

Comparing Two or More Groups

Discussion/Poll Question #3.C (For On24)
Which of the following statements about comparative tests is true? (Select all that apply)

1. ANOVA is only used to compare the means of two groups.
2. The Chi-Square test is used to examine the relationship between two categorical variables.
3. ANOVA can be used to compare the means of three or more groups.
4. The Chi-Square test can be used to compare the means of continuous variables.
5. ANOVA tests produce an F-statistic to determine if there are significant differences between group means.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

More Comparative Tests

- Common Tests:
  - ANOVA (Analysis of Variance): Compares means across multiple groups.
  - Chi-Square Test: Tests the association between categorical variables.
- When to Use:
  - ANOVA: When comparing three or more group means.
  - Chi-Square: When examining the relationship between two categorical variables.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

ANOVA

- Hypotheses:
  - Null Hypothesis (H_0): All group means are equal.
  - Alternative Hypothesis (H_a): At least one group mean is different.
- F-Statistic: Ratio of the variance between groups to the variance within groups.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Chi-Square Test

- Types of Chi-Square Tests:
  - Chi-Square Test for Independence: Tests if two categorical variables are independent.
  - Chi-Square Goodness of Fit Test: Tests if a sample distribution fits a population distribution.
- Hypotheses:
  - H_0: The variables are independent.
  - H_a: The variables are not independent.

- Chi-Square Statistic: Sum of the squared difference between observed and expected frequencies divided by the expected frequency.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Walkthrough and Exercise #3.3
Comparative Tests

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

1. Conduct a one-way ANOVA on a numeric column across three or more different groups and interpret its results.
2. Perform a chi-square test of independence between two columns and interpret its results.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Questions and Answers

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

Review of Week 3 Module 3

Week 3 Module 4

Introduction to Non-Parametric Tests

Discussion/Poll Question #3.D (For On24)
Which of the following statements about non-parametric tests is true? (Select all that apply)

1. Non-parametric tests require the data to follow a normal distribution.
2. Non-parametric tests are more flexible and can be used with ordinal data.
3. The Mann-Whitney U Test is used to compare the means of two related groups.
4. The Chi-Square test is a non-parametric test used to examine the association between categorical variables.
5. Non-parametric tests are only used when sample sizes are large.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

When and Why to Use Non-Parametric Tests

- When to Use:
  - Data does not meet the assumptions of parametric tests (e.g., normality).
  - Sample sizes are small.
  - Data is ordinal or categorical.
- Advantages:
  - More flexible than parametric tests.
  - Can be used with non-normal distributions.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Common Non-Parametric Tests

- Mann-Whitney *U* Test: Compares differences between two independent groups.
- Kruskal-Wallis *H* Test: Compares differences between three or more independent groups.
- Chi-Square Tests are also non-parametric tests.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Mann-Whitney *U* Test

- Purpose: Tests whether there is a difference between two independent groups.
- Hypotheses:
  - Null Hypothesis (H_0): The distributions of the two groups are equal.
  - Alternative Hypothesis (H_a): The distributions of the two groups are not equal.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Kruskal–Wallis *H* Test

- Purpose: Tests whether there are differences between three or more independent groups.
- Hypotheses:
  - Null Hypothesis (H_0): The distributions of the groups are equal.
  - Alternative Hypothesis (H_a): The distributions of at least one group are different.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Walkthrough and Exercise #3.4
Non-Parametric Tests

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

1. Perform a Mann-Whitney *U* Test between two groups and interpret the results.
2. Apply a Kruskal-Wallis *H* Test across multiple groups.

Course repo: https://github.com/ismayc/oreilly-fundamentals-of-statistics-with-python

Questions and Answers

Anything I can clear up regarding the *Week 3 Module 4* content?

Review of Week 3 Module 4

Learning Objectives

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

- Use Python for complex statistical analyses, leveraging libraries like NumPy, Pandas, SciPy, and Matplotlib for data manipulation and visualization.
- Uncover underlying patterns, trends, and anomalies in datasets using exploratory data analysis.
- Model various types of data with probability distributions and utilize hypothesis testing to validate data-driven inferences.

Conclusion

Additional resources:

- numpy
- pandas
- matplotlib
- seaborn
- scipy

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

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

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


O'REILLY®