Implementing Statistics with Python

Optimize decision-making with statistical inference and Python

Wei-Meng Lee

www.bpbonline.com

First Edition 2024

ISBN: 978-93-55517-104

LIMITS OF LIABILITY AND DISCLAIMER OF WARRANTY

The information contained in this book is true to correct and the best of author's and publisher's knowledge. The author has made every effort to ensure the accuracy of these publications, but publisher cannot be held responsible for any loss or damage arising from any information in this book.

All trademarks referred to in the book are acknowledged as properties of their respective owners but BPB Publications cannot guarantee the accuracy of this information.

Dedicated to

My families, my wife, and my daughter

About the Author

Wei-Meng Lee is a seasoned technologist, author, and educator known for his expertise in a wide range of topics, including software development, data science, and emerging technologies. With a strong background in computer science and a passion for innovation, Wei-Meng has made significant contributions to the tech industry through his writing, teaching, and hands-on experience.

As an author, Wei-Meng has written numerous books and articles that have helped professionals and enthusiasts alike deepen their understanding of programming languages, frameworks, and cutting-edge tools. His works often blend practical insights with theoretical concepts, making complex topics accessible and actionable for readers.

Wei-Meng's teaching experience spans across various platforms, where he imparts his knowledge to students eager to learn about programming, data analysis, and technology trends. His engaging teaching style and ability to simplify complex ideas have earned him praise from learners worldwide. In addition to his writing and teaching, Wei-Meng is actively involved in the tech community, participating in conferences, workshops, and forums where he shares his insights and learns from fellow experts.

About the Reviewer

Kartoue Mady Demdah, Ph.D., is a Data Scientist at Olameter Inc. with a robust background in data science, machine learning, and statistics. He earned his Ph.D. in Mathematics from Université de Rennes I and the University of Pisa. He has expertise in computer vision, Time Series and NLP. Kartoue holds a professional certificate in Digital Transformation from MIT, covering AI, IoT, Cloud Computing, Blockchain, and Cybersecurity. Proficient in Python, SQL, TensorFlow, PyTorch, AWS, and Azure, he is also dedicated to mentoring and community engagement through AI hackathons and data philanthropy.

Acknowledgement

First and foremost, I want to extend my heartfelt appreciation to my families (especially my wife and my daughter) for their love and encouragement throughout this journey. Their love and encouragement have been a constant source of motivation to keep me going.

I am also immensely grateful to BPB Publications for their guidance and expertise in bringing this book to fruition. They have been very supportive and understanding in adjusting the timeline of this project according to my schedule.

I would also like to acknowledge the reviewers, technical experts, and editors who provided valuable feedback and contributed to the refinement of this manuscript. Their insights and suggestions have significantly enhanced the quality of the book. I would like to extend my special thanks to the Technical Reviewer, Kartoue.

Last but not least, I want to express my gratitude to the readers who have shown interest in my book. Your support and encouragement are deeply appreciated.

Preface

In the dynamic and data-driven world we live in today, the ability to analyze and interpret data has become a vital skill across numerous fields. From business and finance to healthcare and engineering, the insights derived from data guide critical decisions and strategies.

This book, "Implementing Statistics with Python," is designed to equip you with the foundational knowledge and practical skills needed to navigate the vast landscape of statistics and data analysis using Python.

We first start with an introduction to the fundamental concepts of statistics in Chapter 1. We will explore the structure of statistical data, the objectives of statistical analysis, and the role of statistics in making informed decisions. This chapter sets the stage for the various techniques and tools that will be covered throughout the book.

In Chapter 2, we will dive into the basics of Python programming, a language renowned for its simplicity and versatility in data analysis. You will learn essential data structures and advanced topics like lambda functions, iterators, and generators, providing you with a solid foundation to implement statistical methods in Python.

In Chapter 3, you will be introduced to two essential libraries for data manipulation: NumPy and Pandas. These powerful tools will enable you to handle data efficiently, laying a solid foundation for more advanced statistical analysis. You will learn how to create and manipulate arrays using NumPy, and how to work with data structures such as Series and DataFrames in Pandas.

Building on the data manipulation skills from the previous chapter, Chapter 4 focuses on data visualization using Matplotlib and Seaborn. These libraries will help you create a wide variety of plots and charts to effectively communicate your data insights. You will learn how to generate basic plots with Matplotlib and enhance them with Seaborn's advanced features, preparing you for comprehensive data analysis and presentation.

Chapter 5 focuses on descriptive statistics, offering techniques for summarizing and presenting data in a meaningful way. You will learn how to calculate measures of central tendency, dispersion, and distribution, enabling you to effectively interpret and communicate your data.

In Chapter 6, you will delve into the realm of probability theory. This chapter covers fundamental concepts and probability distributions that are crucial for understanding data variability and uncertainty. You'll gain a solid foundation in probability, preparing you for more advanced statistical analysis.

Chapter 7 on statistical inference explores methods for drawing meaningful conclusions from data, with a focus on chi-square tests and **Analysis of Variance** (**ANOVA**). These techniques are essential for unraveling patterns and relationships within data.

Regression analysis, covered in Chapter 8, examines the relationships between variables, enabling you to make predictions and draw significant insights from data.

Chapter 9 expands on this by introducing multivariate analysis, a technique critical for understanding complex relationships in datasets and essential for machine learning.

Time series analysis, the subject of Chapter 10, addresses the unique challenges of analyzing data ordered by time, with applications in finance, economics, and beyond. You will learn to perform time series forecasting and visualization using real-world examples.

Chapter 11 bridges the gap between traditional statistical methods and modern machine learning techniques. You'll explore how statistical foundations support machine learning, delve into data preparation and model evaluation, and learn the various machine learning algorithms.

In Chapter 12, you will continue your journey into machine learning by focusing on algorithms and deployment. This chapter covers various machine learning algorithms, advanced data preparation techniques, and model evaluation methods. You will also learn about model deployment, using real-world case studies like the Titanic dataset to illustrate these concepts in action.

This book is designed to be both comprehensive and practical, offering a blend of theoretical knowledge and hands-on experience. Whether you are a student, a professional, or an enthusiast looking to deepen your understanding of statistics and data analysis, this book aims to provide the tools and insights needed to excel in the data-driven landscape of today.

We hope this book serves as a valuable resource in your journey to mastering the art and science of data analysis.

Chapter 1: Introduction to Statistics– This chapter provides an introduction to statistics and the various techniques used for data analysis and inference. We will begin by exploring the fundamental concepts of statistics, including the structure of statistical data, the objectives of statistical analysis, and the role of statistics in making informed decisions based on data. Additionally, we will delve into key topics such as populations and samples, variables (both categorical and quantitative), probability theory, data preprocessing, and data visualization techniques. Through this foundational knowledge, this chapter will provide an introduction to the essential principles and tools required for statistical analytics, which will be covered in the rest of the chapters in this book.

Chapter 2: Python Basics for Statistics– This chapter introduces you to the Python programming language. You will learn the basics of the language, together with some of the most important data structures that you need to be proficient in when implementing statistics in Python. Besides the basics, you will also learn some advanced topics in Python, such as lambda functions, iterators, and generators.

Chapter 3: Introduction to NumPy and Pandas for Data Manipulation– This chapter will lay the foundation for two very important libraries in Python that enable data manipulation – NumPy and Pandas. You will learn how to create NumPy arrays that store collections of data of the same type and explore the use of the Pandas library to represent data in two specific data structures – Series and DataFrame.

Chapter 4: Data Visualization with Matplotlib and Seaborn– Visualizing data is essential for gaining insights, uncovering patterns, and effectively communicating results. In this chapter, you will explore the power of data visualization using two popular Python libraries: Matplotlib and Seaborn. Matplotlib provides a flexible and comprehensive toolkit for creating a wide range of plots and charts, while Seaborn offers a higher-level interface for producing visually appealing and informative statistical graphics.

Chapter 5: Descriptive Statistics– This chapter will discuss a branch of statistics known as descriptive statistics. The field of descriptive statistics involves summarizing, organizing, and presenting data in a meaningful way. Its main use is to provide users with a concise overview of the data, allowing them to understand the key characteristics and patterns of the data.

Chapter 6: Probability Theory– This chapter explores fundamental concepts in probability, starting with an introduction to probability. We will discuss classical, empirical, and subjective probability, followed by key rules like addition and multiplication for event probabilities. Conditional probability and Bayes' theorem are discussed, alongside random variables. We will also explore probability distributions, including discrete (like Binomial and Poisson) and continuous (such as uniform and normal), focusing on statistical transformation techniques like standardization and normalization for data preprocessing.

Chapter 7: Statisical Inference– Statistical inference is the art and science of drawing meaningful conclusions from data in the face of uncertainty. It involves making educated guesses or predictions about a population based on information obtained from a sample. Statistical inference allows researchers, analysts, and decision-makers to extend their insights beyond the observed data, providing a framework for making informed decisions in various fields. In this chapter, we will focus on two powerful and specialized techniques in Statistical Inference: the chi-square test and **Analysis of Variance** (**ANOVA**). By examining these statistical methodologies, we aim to provide a key understanding of how they can unravel patterns and relationships within categorical and continuous data, respectively.

Chapter 8: Regression Analysis– Regression analysis is a statistical method used to examine the relationship between one or more independent variables and a dependent variable. An independent variable is a variable that is manipulated or controlled in a study or experiment. A dependent variable, on the other hand, is a variable that is observed and measured in response to changes in the independent variable.

Understanding the relationship between independent variables and dependent variables is very useful because it allows us to draw important conclusions from various research or statistical studies.

Chapter 9: Multivariate Analysis– Multivariate analysis refers to statistical techniques used to analyze data with multiple variables. It helps you analyze the complex relationships and patterns that may exist among variables. Multivariate analysis is an exercise important for machine learning as it allows you to understand your data and filter the irrelevant features from your dataset to create a more accurate model for your dataset. In this chapter, you will learn the various techniques to study the relationships between the features in your dataset, and how to identify features that are relevant to creating accurate models using your dataset.

Chapter 10: Time Series Analysis– A time series is a set of data points ordered by time. It represents the values of a variable (or multiple variables) over a specified time interval. It has common applications in fields such as finance, economics, engineering, etc. In this chapter, you will learn how to perform time series analysis as well as time series forecasting. You will learn concepts such as autocorrelation and partial autocorrelation. In addition, this chapter will also show you how to perform visualization and evaluation of time series using a real-world example.

Chapter 11: Machine Learning for Statistics– This chapter introduces you to the world of machine learning. You will learn the two main types of machine learning algorithms and have the chance to use the various algorithms implemented in the sklearn library. Note that for brevity, this book does not cover reinforcement learning, a type of machine learning paradigm where an agent learns to make decisions by interacting with an environment.

Chapter 12: Practical Statistical Analysis in Machine Learning– This chapter aims to provide practical insights and guidance on applying statistical methods and machine learning techniques to real-world data. This includes exploring data preparation steps such as importing, cleaning, and encoding features using the Titanic dataset as a case study. Additionally, the chapter aims to cover the evaluation of various machine learning algorithms, including logistic regression, hyper-parameter tuning using GridSearchCV, cross-validation techniques, and model training and deployment.

Code Bundle and Coloured Images

Please follow the link to download the
Code Bundle and the ***Coloured Images*** of the book:

https://rebrand.ly/48d787

The code bundle for the book is also hosted on GitHub at
https://github.com/bpbpublications/Implementing-Statistics-with-Python.
In case there's an update to the code, it will be updated on the existing GitHub repository.

We have code bundles from our rich catalogue of books and videos available at **https://github.com/bpbpublications**. Check them out!

Errata

We take immense pride in our work at BPB Publications and follow best practices to ensure the accuracy of our content to provide with an indulging reading experience to our subscribers. Our readers are our mirrors, and we use their inputs to reflect and improve upon human errors, if any, that may have occurred during the publishing processes involved. To let us maintain the quality and help us reach out to any readers who might be having difficulties due to any unforeseen errors, please write to us at :

errata@bpbonline.com

Your support, suggestions and feedbacks are highly appreciated by the BPB Publications' Family.

Piracy

If you come across any illegal copies of our works in any form on the internet, we would be grateful if you would provide us with the location address or website name. Please contact us at **business@bpbonline.com** with a link to the material.

If you are interested in becoming an author

If there is a topic that you have expertise in, and you are interested in either writing or contributing to a book, please visit **www.bpbonline.com**. We have worked with thousands of developers and tech professionals, just like you, to help them share their insights with the global tech community. You can make a general application, apply for a specific hot topic that we are recruiting an author for, or submit your own idea.

Reviews

Please leave a review. Once you have read and used this book, why not leave a review on the site that you purchased it from? Potential readers can then see and use your unbiased opinion to make purchase decisions. We at BPB can understand what you think about our products, and our authors can see your feedback on their book. Thank you!

For more information about BPB, please visit **www.bpbonline.com**.

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

Table of Contents

Chapter 1
Introduction to Statistics

Introduction

This chapter provides an introduction to statistics and the various techniques used for data analysis and inference. We will begin by exploring the fundamental concepts of statistics, including the structure of statistical data, the objectives of statistical analysis, and the role of statistics in making informed decisions based on data. Additionally, we will delve into key topics such as populations and samples, variables (both categorical and quantitative), probability theory, data preprocessing, and data visualization techniques. Through this foundational knowledge, this chapter will provide an introduction to the essential principles and tools required for statistical analytics, which will be covered in the rest of the chapters in this book.

Structure

The chapter covers the following topics:

- Statistics
- Overview of Python for statistical analytics

Objectives

The objectives of this chapter are to provide a foundational understanding of statistics and their essential components. First, it aims to introduce readers to the basic structure

of statistics and its importance in data analysis and decision-making processes. Second, the chapter aims to clarify the concept of statistics itself, outlining its role in collecting, organizing, analyzing, and interpreting data. Furthermore, it aims to familiarize readers with key concepts such as populations and samples, variables (both categorical and quantitative), and probability theory, laying the groundwork for more advanced statistical analyses. The chapter also intends to highlight the significance of data preprocessing and visualization techniques in preparing and exploring data sets effectively. By achieving these objectives, readers will be well-prepared to delve into subsequent topics such as regression analysis, statistical models, and the use of Python for statistical analytics.

Statistics

Statistics is a branch of mathematics that deals with the collection, analysis, interpretation, presentation, and organization of data. It provides techniques and methods for making inferences and decisions in the presence of uncertainty.

Statistics is especially important in the field of *data analytics*. Data analytics is the process of examining, cleaning, transforming, and modeling data to extract meaningful insights, draw conclusions, and support decision-making. Statistics provides the foundational framework and tools that enable data analysts to make sense of complex datasets, extract meaningful insights, and support data-driven decision-making. Here are some of the areas that are made possible by statistics methods:

- **Descriptive statistics**: This involves summarizing, organizing, and presenting data in a meaningful way. Its main use is to provide users with a concise overview of the data, allowing them to understand the key characteristics and patterns of the data. You will learn more about descriptive statistics in *Chapter 5, Descriptive Statistics*.
- **Inferential statistics**: It is the art and science of drawing meaningful conclusions from data in the face of uncertainty. It involves making educated guesses or predictions about a population based on information obtained from a sample. Statistical inference allows researchers, analysts, and decision-makers to extend their insights beyond the observed data, providing a framework for making informed decisions in various fields. You will learn more about inferential statistics in *Chapter 7, Statistical Inference*.
- **Hypothesis testing**: This helps us assess the validity of assumptions or claims about a population. You will also learn more about hypothesis testing in *Chapter 7, Statistical Inference*.
- **Predictive analytics, machine learning, and data science:** These utilize a broad range of techniques and methods, including statistical approaches, to analyze and extract insights from data and make predictions about future outcomes based on historical data. You will learn more about machine learning in *Chapter 11, Machine Learning for Statistics.*

Some key concepts and techniques in statistics include:

- Population and sample
- Variables
- Probability
- Data visualization
- Regression analysis
- Statistical models

Population and sample

In statistics, there are two terms that are fundamental to making inferences and drawing conclusions about a larger group based on a subset of data. These two terms are:

- Population
- Sample

Population refers to the entire group that is the subject of the study or analysis. It includes all possible observations that share a common characteristic. For example, if you are studying the average **Body Mass Index** (**BMI**) of all the people in a particular country, then the population of the country would be the *population*.

However, it is not feasible (or practical) to measure the BMI of every single person in the country. So, statisticians collect this information from a *subset* of the population. This subset is known as a *sample*. Hence, a sample is a subset of a population that is selected for analysis or study, chosen in a way that represents the characteristics of the population under study.

Figure 1.1 shows the relationship between the population and the sample of a dataset:

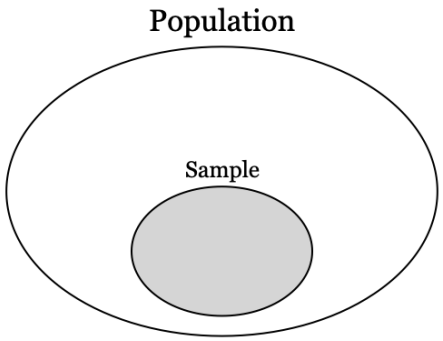

Figure 1.1: *Relationship between sample and population*

Note: A sample should be chosen in a specific way so that it is a representation of the population.

Variables

Variables are characteristics or attributes that can be measured or observed. They can be classified as *categorical* (qualitative) or *numerical* (quantitative):

- **Categorical**: Categorical data represents distinct categories or groups that cannot be measured in terms of numerical values. Categorical data can be further divided into:
  - **Nominal data**: Categories where there is no inherent order, such as colors or genders.
  - **Ordinal data**: Categories where there is an order for the data, such as educational levels or grades. Ordinal data can be further divided into:
    - **Ordered category**: Ordered categories refer to a set of distinct groups or classes where there is a meaningful order or ranking among the categories.
    - **Ranks**: Ranks involve assigning a numerical position or order to a set of values based on some criterion, such as sorting items from lowest to highest or vice versa.
- **Quantitative:** Quantitative data consists of numerical values that can be measured and subjected to mathematical operations. It can be further divided into:
  - **Discrete**: Countable and often represents whole numbers, such as the number of wheels, or the number of students in a class.
  - **Continuous**: Can take any value within a given range and can be measured with great precision, such as height, weight, or temperature.

Tip: The term ordinal is derived from the Latin word "ordo," meaning order or arrangement.

Figure 1.2 shows the breakdown of the different types of data:

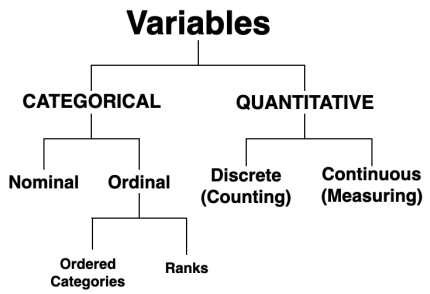

Figure 1.2: *Different types of variables*

To illustrate the different data types, we will use the example of bicycles. Assuming you have a dataset containing a bunch of used bicycles. *Figure 1.3* shows a sample snapshot of the dataset:

Brand	Model	Types	Condition	Weight	Number of Gears	Color	Price
Trek	FX 2	Hybrid	New	25.6	21	Black	699.99
Specialized	Rockhopper	Mountain	Used	30.0	18	Red	849.95
Giant	Escape 3	Hybrid	Like New	27.9	24	Blue	549.00

***Figure 1.3**: A snapshot of the bicycle dataset*

The following sections will outline methods for recognizing different data types within your dataset. This is a crucial step in selecting the right statistical analyses and visualization techniques for effective data exploration and analysis.

Categorical

Categorical values represent categories or groups and can take on a limited, fixed number of distinct and unordered values. From the above sample dataset, you can see that the following fields are categorical variables:

- **Brand**: Trek, Specialized, Giant, and so on
- **Mode**l: FX 2, Rockhopper, Escape 3, and so on
- **Types**: Hybrid, Mountain, and so on
- **Condition**: New, Like New, Used, and so on
- **Color**: Black, Red, Blue, and so on

Here are some characteristics to help you identify categorical variables in your dataset:

- Check the data type. Categorical values are often stored as string types.
- Check for unique values. Categorical values are usually limited and distinct. However, there are exceptions to this rule. For example, if the **Model** fields contain all unique models, it is still considered a categorical field.
- Contextual understanding—often, this requires a basic understanding of your dataset. For example, fields like sex, color, or brand are often categorical.

The next task would be to identify which of the above categorical variables are *nominal* and which are *ordinal*. Of the list of categorical variables identified in the previous section, you can see that the following fields are *nominal* types of data:

- Brand
- Model
- Types
- Color