The Mini Test Book (in development)

G. Alexi Rodríguez-Arelis Kate Manskaia Payman Nickchi

2025-07-24

This mini-book presents fundamental hypothesis tests in statistical inference using different mind maps while incorporating a common test workflow via a frequentist approach. We utilize `Python` and `R` in parallel to demonstrate the execution of these tests.

Table of contents

Preface

Have you ever felt overwhelmed by the numerous fundamental hypothesis tests you need to learn in statistical inference courses?

We have experienced this sense of overwhelm throughout our academic journeys as well. However, we also understand that statistical inference is a **powerful tool** for gaining insights into complex populations across various fields of study. Whether analyzing electoral preferences in political science or assessing the effectiveness of innovative medical treatments in randomized clinical trials, the applications are extensive. Hence, in response to these challenges, we have created this mini-book as a handy resource to help structure and simplify the learning of different fundamental hypothesis tests. Our goal is to present these concepts in a **reader-friendly manner** while clearly explaining the necessary statistical jargon, making these inferential methods accessible to a broader audience.

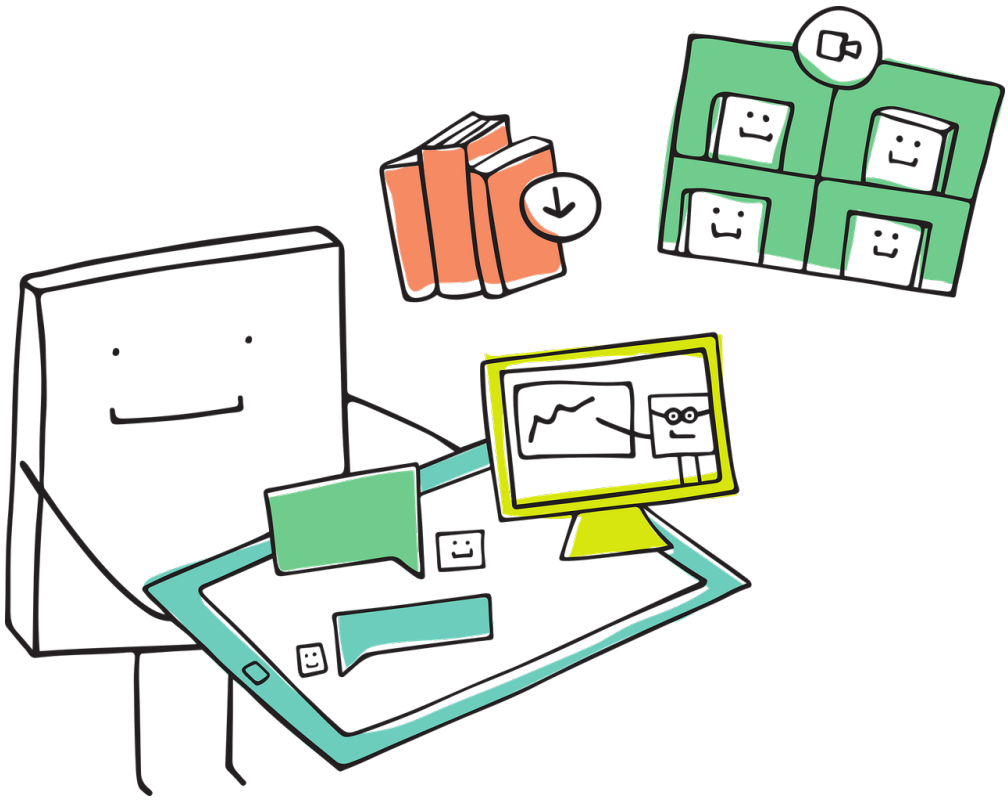


Figure 1: Image by *manfredsteger* via *Pixabay*.

Note that, after conducting extensive research into the available educational literature, we discovered that there is no comprehensive and **frequentist** resource that explains various inferential methods simultaneously using **two essential programming languages** in the field of data science: `R` (R Core Team 2024) and `Python`(Van Rossum and Drake 2009). Furthermore, we could not find reproducible and transparent tools that would enable learners to implement and adapt these methods in their own computational environments. Based on our teaching experience, these shortcomings hinder effective learning in the practice of statistical inference, especially given the numerous tests required to achieve mastery.

To address this gap, we have developed a bilingual and frequentist resource in both `R` and `Python`, which features a **common test workflow** consisting of eight distinct stages applicable to each hypothesis test: *study design*, *data collection and wrangling*, *exploratory data analysis*, *testing settings*, *hypothesis definitions*, *test flavour and components*, *inferential conclusions*, and *storytelling*. Additionally, all the tests we discuss are organized through different mind maps to help readers visualize their learning process. Finally, by offering this mini-book as an Open Educational Resource (OER) in Quarto via a GitHub repository, we aim to inspire and empower academic communities worldwide to share and adapt this knowledge to suit their specific needs.

The Authors

G. Alexi Rodríguez-Arelis

I'm an Assistant Professor of Teaching in the Department of Statistics and Master of Data Science at the University of British Columbia (UBC). Throughout my academic and professional journey, I've been involved in diverse fields, such as credit risk management, statistical consulting, and data science teaching. My doctoral research in statistics is primarily focused on computer experiments that emulate scientific and engineering systems via Gaussian stochastic processes (i.e., kriging regression). I'm incredibly passionate about teaching regression topics while combining statistical and machine learning contexts.

Kate Manskaia

Payman Nickchi

I am a Postdoctoral Research and Teaching Fellow in the Department of Statistics and the Master of Data Science (MDS) program at the University of British Columbia (UBC). I completed my PhD in Statistics at Simon Fraser University (SFU), where my research focused on biostatistics and goodness-of-fit tests using empirical distribution functions. I am currently teaching statistical courses in the MDS program at UBC. My passion for statistics, teaching, and data science led me to this role. Outside of work, I enjoy swimming and capturing the night sky through astrophotography.

License

Website Privacy Policy

Last updated

May 17th, 2025.

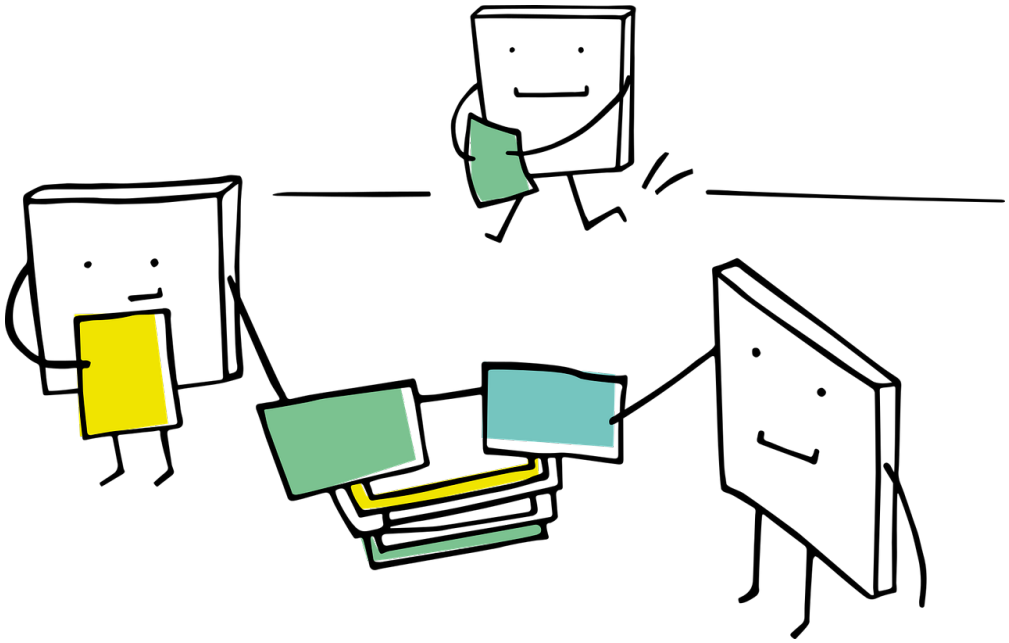


Figure 2: Image by *Manfred Stege* via *Pixabay*.

Your privacy is important to us. This policy outlines how this online textbook created for courses at the University of British Columbia (UBC) ("we," "us," or "our") collects, uses, and protects your information.

Information Collection and Use

We use Google Analytics, a web analytics service provided by Google, LLC. ("Google"). Google Analytics uses cookies to help analyze how students interact with the textbook, including tracking which sections are accessed most frequently. Information generated by cookies about

your use of our website (including IP address) will be transmitted to and stored by Google on servers in the United States.

Google will use this information solely for evaluating textbook usage, compiling usage reports to enhance the educational effectiveness of the textbook, and providing related services.

You may refuse the use of cookies by selecting the appropriate settings in your browser; however, please note this may affect your textbook browsing experience.

Personal Information

We do not collect personally identifiable information through Google Analytics. Any personally identifiable information, such as your name and email address, would only be collected if voluntarily submitted for specific educational purposes (e.g., feedback or course-related inquiries). We will never sell or distribute your personal information to third parties.

For any questions or concerns, please contact us at alexrod@stat.ubc.ca.

1 Introduction

> **The most important maxim for data analysis to heed, and one which many statisticians seem to have shunned, is this: "Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise."**

John W. Tukey (1962, 13)

Data collection worldwide has proven to be a valuable tool for uncovering significant insights across various **populations of interest**. Whether it involves capturing political preferences in a specific demographic ahead of an upcoming election or assessing the effectiveness of an innovative medical treatment through a randomized clinical trial compared to a standard treatment, data plays a crucial role in enhancing our understanding. At times, this understanding can become quite complex, especially when attempting to untangle the relationships between different variables within a given population or even across two or more populations.

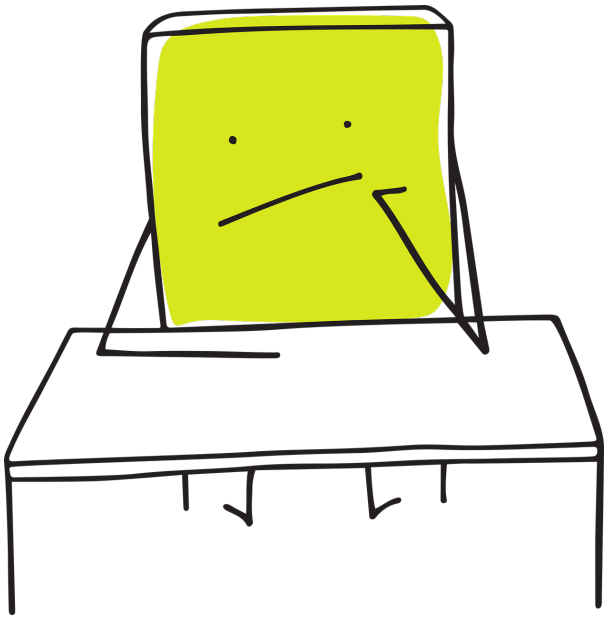


Figure 1.1: Image by *Manfred Stege* via *Pixabay*.

In a vast and diverse field like data science, it is crucial to craft effective and transparent solutions that facilitate proper data analysis. However, conducting a full census to collect data from entire populations can often be impractical due to resource limitations such as budget constraints, workforce shortages, or insufficient technical infrastructure. Despite these challenges, our primary objective remains to gain insights about any population of interest via some class of analysis, even when data availability is limited. In this regard, **statistical inference** is a powerful tool that allows us to draw insights even with limited data. That said, it is important to emphasize that the process of statistical inference begins with asking the **right questions**, even before data collection occurs.

In light of this context, we need to establish the appropriate stages of the statistical inference process, along with a useful tool to help select the right hypothesis test based on our specific context, research questions, variable types, and parameters of interest. This is why this mini-book focuses on two key components:

- **A test workflow:** This workflow will primarily guide us in formulating the right questions about our population(s) of interest, which will involve specific parameters. This process will generally proceed with data collection using a specific sampling method, followed by a thorough analysis that includes exploratory data analysis and the **most suitable** hypothesis testing based on our primary question(s). We will conclude the process by presenting a compelling storytelling to our stakeholders. Section 1.1 will elaborate further on this workflow.
- **A series of test mind maps:** Since the test workflow ultimately involves selecting the most suitable hypothesis test, we require a form of guidance to choose these tests according to the inferential question(s) we want to address. Therefore, Section 1.2 will introduce our core test mind map from Figure 1.14, which will direct us to more detailed mind maps each time we introduce a new chapter.

This mini-book on hypothesis testing is intended to serve as a **practical manual** rather than a traditional statistical textbook. Furthermore, it focuses on providing applied examples in each chapter without any additional exercises for the reader. We aim to explain the necessary mathematical formulas in straightforward language, avoiding formal proofs for these expressions. Additionally, we will establish conventions using admonitions to offer **key insights** and links to **supplementary and more in-depth material**.

Heads-up!

A key insight (or insights) related to a specific hypothesis test or a stage in the test workflow. The reader is advised to keep this heads-up in mind throughout the showcase of the corresponding example in any given chapter.

Tip

An idea or ideas that extend beyond the immediate discussion and can offer valuable context and insightful background. Whenever relevant, we will provide references for further reading to deepen understanding and enhance knowledge.

1.1 The Test Workflow

There is a single test workflow for many different flavours!

The statement above summarizes the essence of our testing workflow, which requires a detailed examination in this section. Primarily, it is crucial to understand that mastering all hypothesis tests involves more than just knowing their mathematical formulas or coding functions; it requires a disciplined and structured process. Whether we are evaluating evidence against a **null hypothesis**—the status quo of our population parameter(s) of interest—or reporting the uncertainty of an **estimated effect**, the workflow outlined by Figure 1.3 is intended to align your main inferential inquiries with the **most suitable test flavour**. Regardless of the flavour chosen, this workflow is designed to ensure that our conclusions are not only statistically valid but also based on clear and purposeful reasoning.

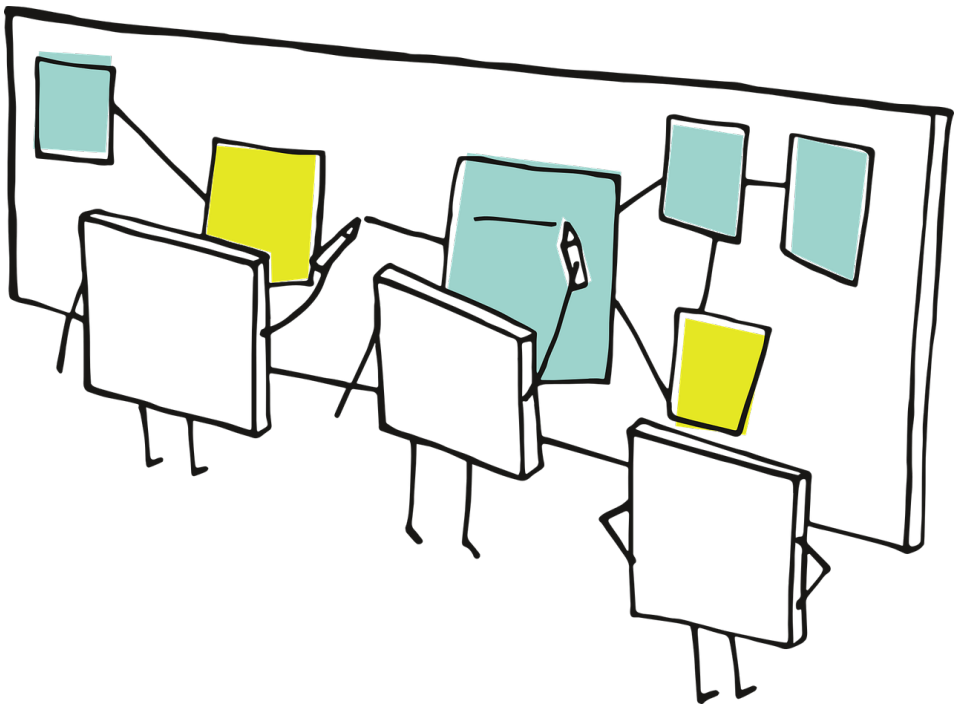


Figure 1.2: Image by *Manfred Stege* via *Pixabay*.

The workflow for hypothesis testing consists of eight stages, which will be discussed in detail in the following sections:

1. **Study design:** This initial stage, referred to as the **main inferential inquiries**, outlines the primary questions we aim to answer through our analysis.

2. **Data collection and wrangling:** The inquiries established in the first stage will guide the design of our data collection, utilizing a specific sampling scheme. Once the data is collected, it must be wrangled and split into two sets: **training** and **test**.
3. **Exploratory data analysis:** In this stage, we classify variables to provide preliminary insights using descriptive statistics and visualizations via the training set.
4. **Testing settings:** We must revisit the **significance level** used in our power analysis (i.e., the procedure used to obtain the minimum sample size n of data points to be collected). Additionally, we need to list all modelling parameters that will be tested.
5. **Hypothesis definitions:** With the modelling parameters to test, we need to define our hypotheses: the **null** hypothesis versus the **alternative** hypothesis. These should be framed in relation to the main inferential inquiries.
6. **Test flavour and components:** At this stage, we choose the most appropriate test flavour and indicate the respective **assumptions**. Depending on whether the test is classical or simulation-based, we will then identify the necessary components to compute the **critical** values or p-values (via the test set) for the next stage.
7. **Inferential conclusions:** The goal of this stage is to determine whether we should reject the null hypothesis based on the critical values or p-values obtained. This stage also includes running the **model diagnostics** to check our corresponding assumptions.
8. **Storytelling:** Finally, communicate the findings through a clear and engaging narrative that is accessible to your stakeholders.

1.1.1 Study Design

This is the initial stage of the hypothesis testing workflow, which involves what we refer to as **main inferential inquiries**. These inquiries are typically posed by stakeholders who wish to conduct a study to better understand a specific population of interest and its associated parameters. In practice, these parameters (such as a population mean or variance) are unknown but considered fixed. This approach, where population parameters are treated as fixed yet unknown, corresponds to the **frequentist paradigm**.

Heads-up on the frequentist paradigm!

In the frequentist paradigm, statistical inference relies on the concept that probabilities represent **long-run** relative frequencies of events observed through repeated experimentation or observation. In this approach, we estimate population parameters by examining the distribution of outcomes derived from multiple independent realizations of a random process.

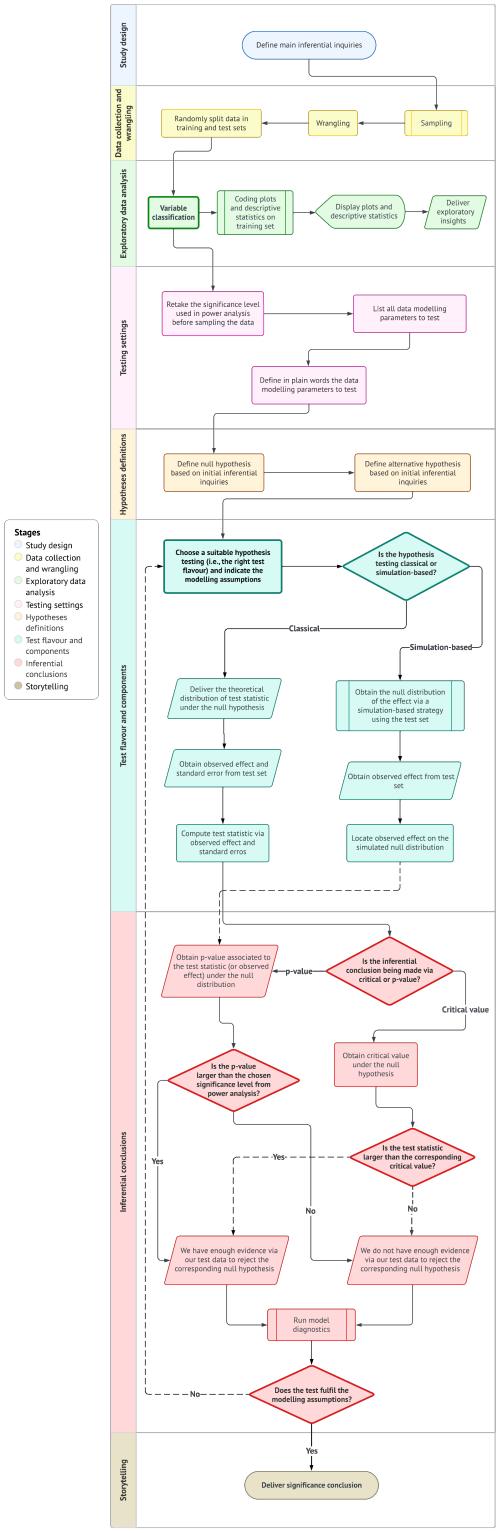


Figure 1.3: A hypothesis testing workflow structured in eight stages: *study design*, *data collection and wrangling*, *exploratory data analysis*, *testing settings*, *hypothesis definitions*, *test flavour and components*, *inferential conclusions*, and *storytelling*.

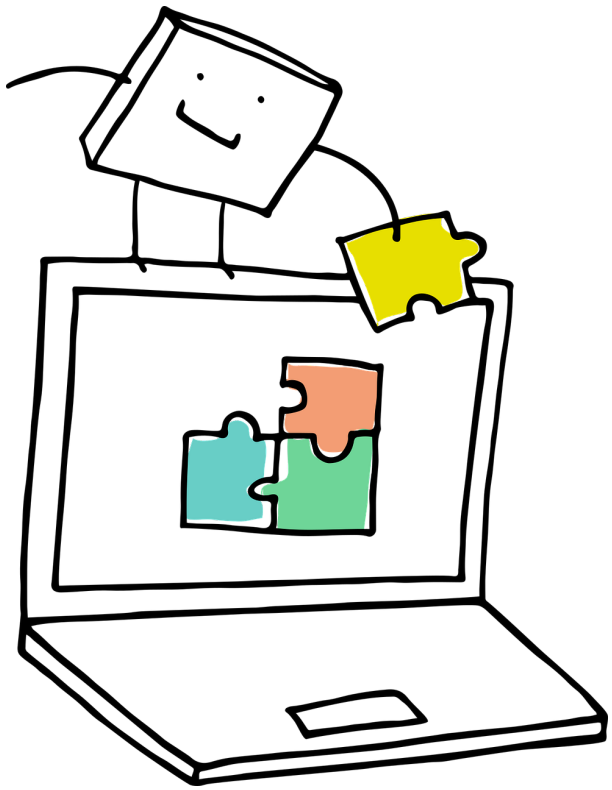


Figure 1.4: Image by *Manfred Stege* via *Pixabay*.

Additionally, this paradigm assumes that the parameters governing a population are **fixed but unknown**. As a result, all randomness is attributed to the data-generating process, not to the parameters themselves.

It is essential to clearly define the main inferential inquiries based on the following principles:

- We need to consult stakeholders about what the study aims to understand regarding their population of interest before entering the second workflow stage, which involves data collection through sampling.
- The main inferential inquiries should align with the stakeholders' research questions. These inquiries should be established at the beginning of this workflow stage and must be meaningful and comprehensive enough to guide the entire inferential investigation.

1.1.2 Data Collection and Wrangling

Once the inferential questions are defined, the next step is to collect and prepare the data for analysis. This stage encompasses sampling strategies to ensure the data is representative of the population and data wrangling to clean and structure the dataset appropriately.

Tip on sampling techniques!

It is important to emphasize the need to choose the **most suitable sampling technique** based on the population's structure and the research questions guiding our main inferential inquiries. Making the right choice of sampling technique is essential for ensuring that our inferential results are accurate, precise, and generalizable to the population of interest. Here are some fundamental (though not exhaustive) sampling techniques:

- **Simple random sampling:** Every individual in the population has an equal probability of being sampled. This is the most basic sampling technique and is probabilistically straightforward, but it may be too simplistic for complex populations in practice.
- **Systematic sampling:** If we have a complete list of individuals in our population, we can sample at regular intervals after selecting a random starting point.
- **Stratified sampling:** The population is divided into distinct groups called strata. These strata are defined in function of the characteristics of the individuals (e.g., age, income, education, etc.). Data is then sampled proportionally from each stratum or through optimal allocation.
- **Cluster sampling:** The population is divided into groups known as clusters, such as households or geographic areas. A random sample is then collected from these clusters.

During the stage of data collection and wrangling of our hypothesis testing workflow, it is crucial to dedicate adequate resources to plan and execute data collection using the most suitable sampling technique. Since the scope of this mini-book does not cover sampling in depth, we recommend reviewing the work by Lohr (2021) for more detailed information on various sampling techniques. This resource includes handy practical examples on this vast field.

When it comes to the wrangling aspect of this stage, once we have sampled our data, it is necessary to structure it in a suitable format (e.g., a proper data frame) using our chosen language, such as the `R` {tidyverse} (Wickham et al. 2019) or `Python` {pandas} (The Pandas Development Team 2024).

Heads-up on coding tabs!

This mini-book is designed to be "*bilingual*," meaning that all hands-on coding can be done in either `R` or `Python`. For each specific example presented in any chapter, you will find two tabs: one for `R` and one for `Python`. We will first display the input code, followed by the corresponding output.

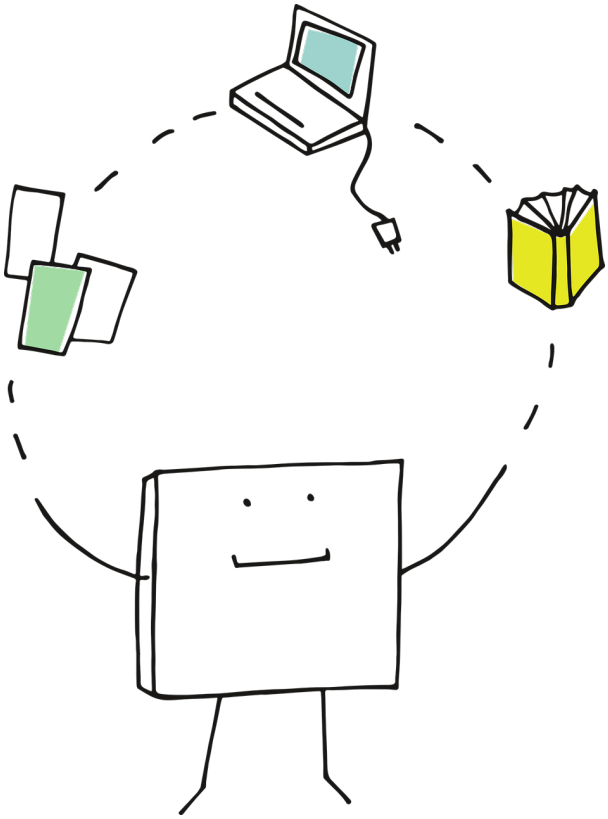


Figure 1.5: Image by *Manfred Stege* via *Pixabay*.

With this format, you can tailor your coding journey based on your language preferences and interests as you advance through the mini-book.

After we have wrangled our data, we need to split it into two sets:

- **Training set.** This set is used solely for exploratory data analysis (EDA) and allows us to gain graphical and descriptive insights into how the sample of individuals behaves concerning our main inferential inquiries.
- **Test set.** This set is reserved for input in our chosen hypothesis testing to be used.

The data splitting is analogous to the standard practice in machine learning to split our data for model training and testing to prevent data leakage in **predictive inquiries**.

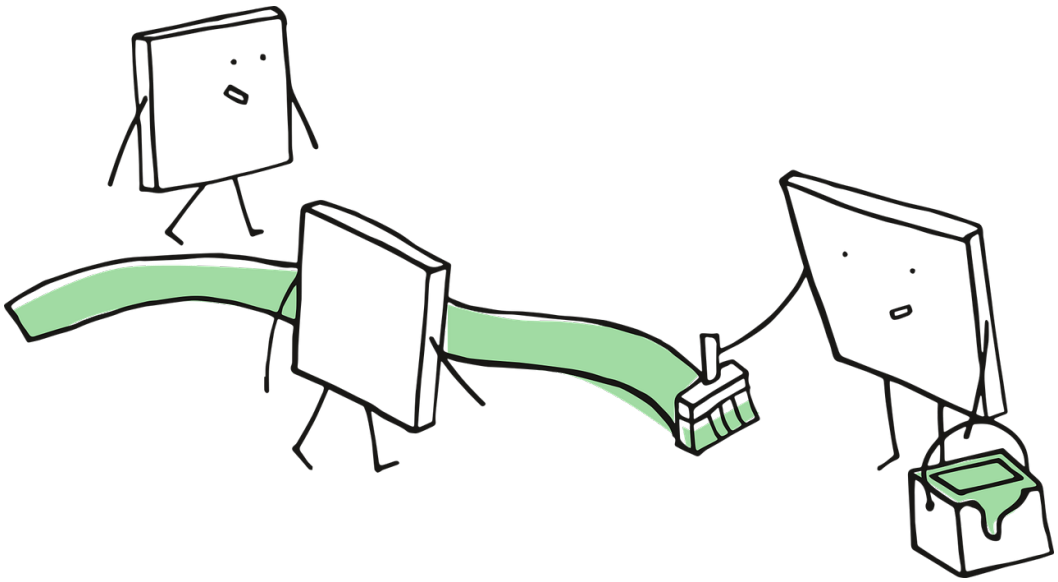


Figure 1.6: Image by *Manfred Steger* via *Pixabay*.

Nevertheless, you might wonder:

Why are we also doing this for an inferential inquiry?

Statistically speaking, the practice of data splitting helps avoid what is known as **double dipping**. Double dipping occurs when the same data is used both for EDA to generate hypotheses and then again for formal statistical testing. Supported by numerical simulations, it can be demonstrated that double-dipping **increases the probability of committing a Type I error**, which occurs when we incorrectly reject the null hypothesis H_0 while it is actually true for the population of interest.

For example, consider a one-sample t-test in a double-dipping context. We might be tempted to formulate our null and alternative hypotheses based on our observed sample mean. For instance, we could state our hypotheses as follows: H_0: $\mu \geq 10$ (null hypothesis) versus H_1: $\mu < 10$ (alternative hypothesis), based on a sample mean of $\bar{x} = 9.5$. If we were to proceed with the statistical test using this same data, we would be falling into the double-dipping trap!

Tip on a further double-dipping resource!

Data splitting is generally not a common practice in statistical inference, despite its frequent use in machine learning. Hence, for more information on double-dipping in statistical inference, Chapter 6 from Reinhart (2015) offers in-depth insights and practical examples.

1.1.3 Exploratory Data Analysis

Once the data is cleaned and structured, it is essential to develop a descriptive understanding of our variables of interest through EDA using the **training set**. The first step is to classify the variables (e.g., numerical, binary, categorical, ordinal, etc.), which will guide us in selecting the most appropriate descriptive statistics and visualizations to examine the relationships between these variables. For instance, we can explore the distribution of these variables and identify any outliers present in the training set. Note EDA is intended to uncover preliminary trends before conducting formal inferential analysis, and these findings should be communicated to our stakeholders during the final storytelling.

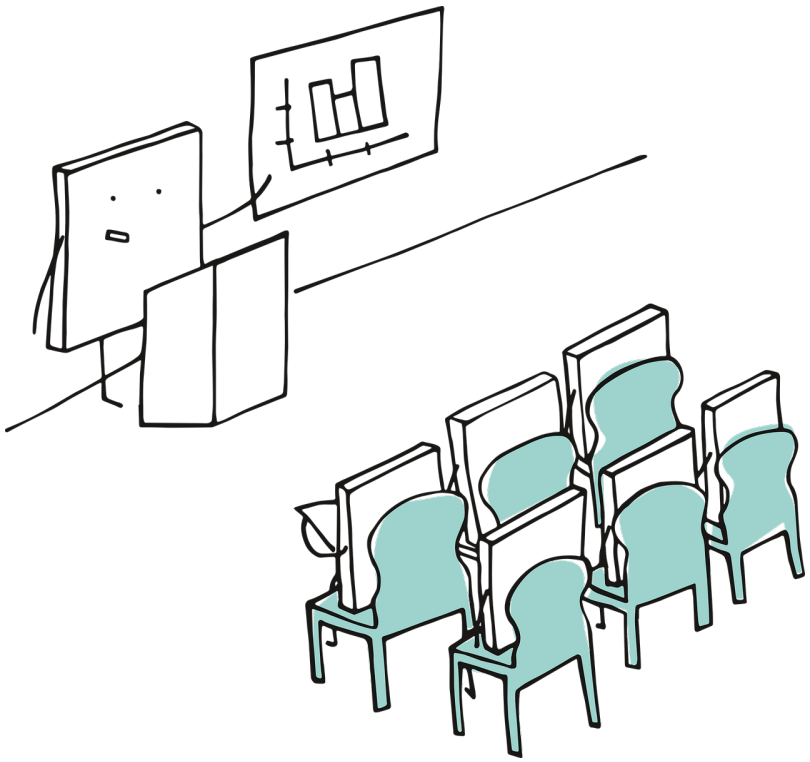


Figure 1.7: Image by *Manfred Stege* via *Pixabay*.

Additionally, the classification of variables during EDA will provide valuable insights for formulating and setting up our hypotheses, while also helping us choose the **most suitable test flavour**. The insights gained from EDA, along with the identified preliminary trends, will shape our expectations for the entire workflow and facilitate a more nuanced statistical interpretation of the main inferential inquiries. Furthermore, EDA aids in justifying our modelling assumptions later in the process. Finally, we must clarify that any insights gained from EDA cannot be generalized to the entire population; they pertain only to the sampled data within the training set.

1.1.4 Testing Settings

This stage allows us to define all our population parameters based on the main inferential inquiries, along with the standards that will guide us through the subsequent stages of the workflow:

1. We need to use the same significance level, denoted as α, that was employed in the **power analysis** when planning data collection for our sampling technique. The significance level denotes the probability of committing Type I error as in Table 1.1 (i.e., the probability of encountering a **false positive**).
2. Regarding the population parameters of interest, we should begin with a formal statistical definition using Greek letters (for further information, see Appendix A).

Table 1.1: Types of inferential conclusions in a frequentist hypothesis testing.

	H_0 is true	H_0 is false
Reject H_0	Type I error (*False positive*)	Correct (*True positive*)
Fail to reject H_0	Correct (*True negative*)	Type II error (*False negative*)

Heads-up on power analysis!

In hypothesis testing, power analysis is a crucial preliminary step used to determine the minimum sample size n necessary to detect a signal that allows us to reject the null hypothesis H_0 in favour of the alternative hypothesis H_1. This analysis ensures that our inferential process can effectively distinguish true population effects from random noise. Note that power analysis requires three key components as inputs:

1. The significance level α.
2. The desired power $1-\beta$ (which relates to correctly rejecting H_0 in favour of H_1, resulting in a **true positive** as in Table 1.1).
3. The effect size—a measure of the magnitude of the association (or causation) that the test is designed to detect.

These three components allow power analysis to provide the minimum sample size n needed to avoid an underpowered study (which occurs when there is a high probability of committing a Type II error as in Table 1.1, denoted as β) or an overly large n that could waste resources.

Figure 1.8: Image by *Manfred Steger* via *Pixabay*.

To effectively communicate our insights to stakeholders via our final storytelling, it is necessary to translate all modelling parameters and hypotheses into clear, plain language for those who may not have a technical background. We should remember that H_0 must be stated in a way that indicates a **status quo** in any given parameter(s), meaning there is nothing noteworthy in the context of our inferential study. On the other hand, H_1 must imply a **departure from this status quo**, indicating that there is indeed something of interest to consider in our inferential analysis.

1.1.5 Hypothesis Definitions

With the settings clarified, the next step is to explicitly define the null and alternative hypotheses. The null hypothesis, H_0, typically asserts that there is no effect or difference—this serves as the default assumption to be tested. As mentioned, H_0 represents the status quo in this context. In contrast, the alternative hypothesis, H_1, indicates the presence of an effect or difference that the data scientist aims to detect. These definitions are derived directly from the main inferential inquiries and insights gained from the EDA **using the training set**.

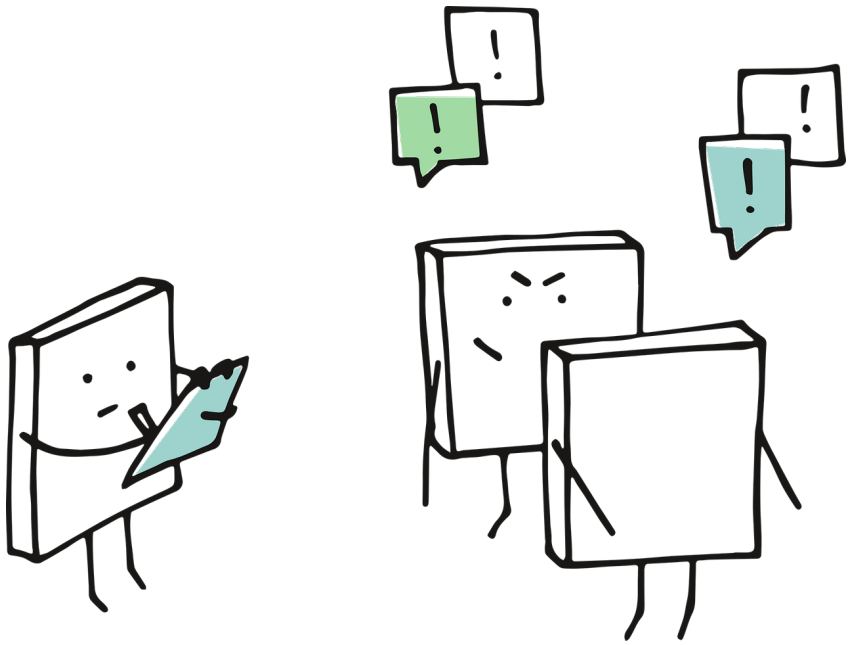


Figure 1.9: Image by *Manfred Steger* via *Pixabay*.

It is crucial to emphasize that clarity in hypothesis definitions is vital for selecting the appropriate type of test and accurately interpreting its results. The formulation of H_0 and H_1 establishes the logical framework for the subsequent analysis. For instance, a null hypothesis might state that the mean test scores for two groups are equal, while the alternative hypothesis claims that they differ. These hypotheses must be mutually exclusive to support a valid statistical decision.

1.1.6 Test Flavour and Components

After specifying the hypotheses, the next step is to choose a suitable statistical test and compute its components. The choice of test—referred to here as the **flavour**—depends on the data structure and the underlying modelling assumptions. We must decide whether to use a classical test, which relies on **theoretical distributions** (e.g., t-test, z-test, chi-squared test, etc.), or a simulation-based method, which employs resampling techniques to **empirically** estimate the null distribution.

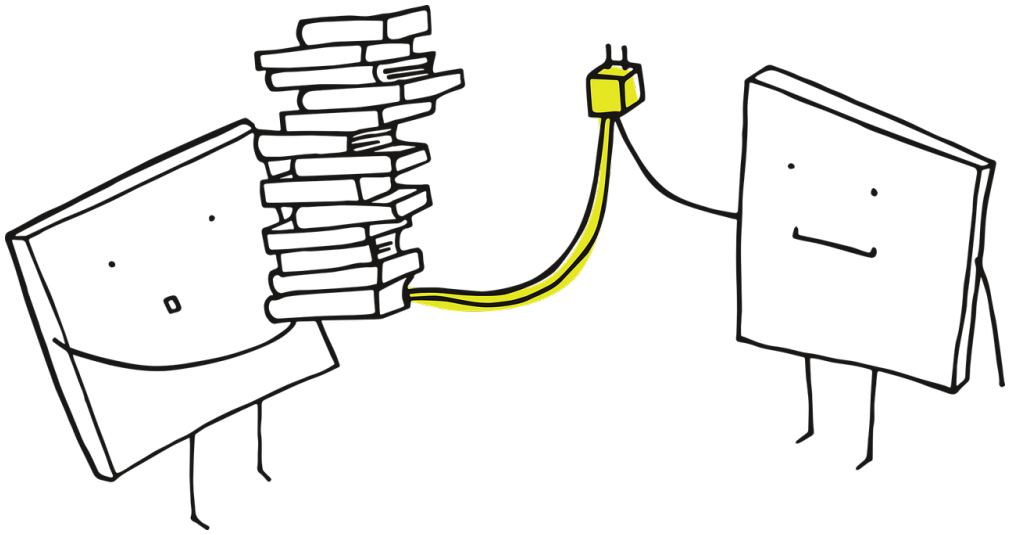


Figure 1.10: Image by *Manfred Steger* via *Pixabay*.

Next, we calculate the observed effect derived from the previously untouched **test set**. For classical tests, the test statistic (calculated using the observed effect and its measure of uncertainty, the standard error) is compared against a theoretical null distribution (i.e., the distribution under the null hypothesis, H_0). In the case of simulation-based tests, the observed effect is situated within a null distribution that is generated from repeated permutations or resamplings. This stage provides the statistical framework necessary to evaluate the strength of evidence against the null hypothesis.

1.1.7 Inferential Conclusions

This stage involves converting the test components obtained from the previous stage into numerical outputs that will allow us to draw meaningful conclusions about our main inferential inquiries. Depending on the type of hypothesis testing, we can obtain the following numerical outputs to support these conclusions:

- **Critical value:** In a classical hypothesis test, the critical value serves as a threshold **under the null theoretical distribution** and is obtained via our our previously set up significance level α. This value helps determine whether we can reject the null hypothesis H_0 in favour of the alternative hypothesis H_1. We directly compare our observed test statistic against this threshold. If the test statistic exceeds the critical value, we reject H_0 in favour of H_1; if it does not, we fail to reject H_0. This approach emphasizes the magnitude of the observed effect relative to the null distribution.
- p-value: This approach can be employed in either a classical test or a simulation-based test. In the case of a classical test, the p-value is derived from the observed test statistic

under the theoretical null distribution. Conversely, when using a simulation-based test, the p-value is associated with the observed effect **under the empirical null distribution**. Regardless of the type of test used, we compare the p-value to our predetermined significance level α. If the p-value is smaller than α, we reject H_0 in favour of H_1; otherwise, we fail to reject H_0. This process provides a probabilistic measure of how surprising the data are under the null distribution (i.e., the status quo).

Heads-up on the practical use of critical and p-values!

Note that the interpretation of critical and p-values must go beyond binary decision-making (statistically speaking); it should also involve the direction, magnitude, and practical implications of the estimated effects, especially in applied contexts. Therefore, we must pay attention not only to statistical significance but also to whether the observed differences or associations are substantively significant in an applied context. This refers to **practical significance**.

Since we are working within a frequentist framework, it is important to remember that critical and p-values are sensitive to sample sizes. In large samples, very small effects can appear statistically significant, even if they have little practical significance. On the other hand, important effects may be overlooked in studies that lack sufficient statistical power. Understanding these nuances helps us avoid misinterpretation and overgeneralization, especially when communicating results to non-statistical stakeholders. Therefore, careful interpretation is essential to bridge the gap between mathematical output and real-world insights.

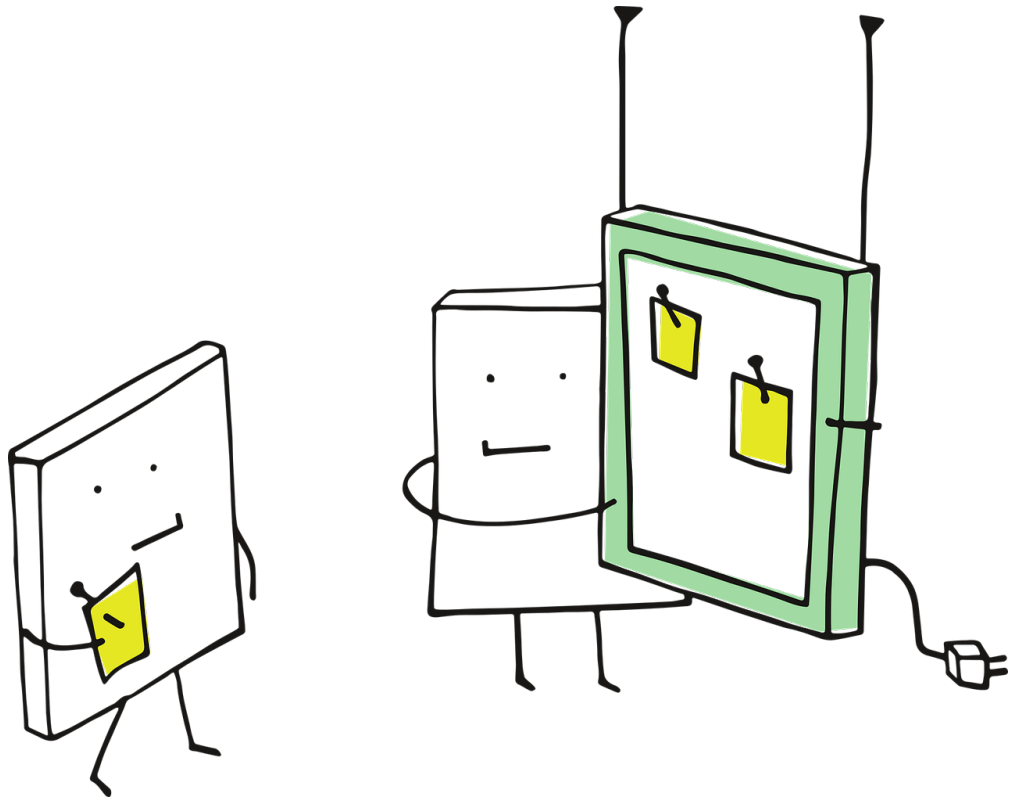


Figure 1.11: Image by *Manfred Steger* via [*Pixabay*](https://pixabay.com/vectors/pixel-cells-pixel-digital-3704070/.

The final step in this stage involves **model diagnostics**. These diagnostics are conducted to verify that the assumptions of data modelling are met (such as normality, independence, homoscedasticity, etc.). If these assumptions are violated, we must reconsider the validity of our inferential conclusions. This process helps prevent drawing incorrect conclusions due to mis-specified models or flawed data structures. If any assumptions are not satisfied, we should return to the previous stage, select a different type of test, and proceed accordingly.

1.1.8 Storytelling

This final stage involves translating our statistical results into language and formats that are accessible primarily to our stakeholders who posed the main inferential questions. The statistical literacy of these stakeholders can vary; they may include research fellows, corporate leadership, policymakers, or even the general public. The ultimate goal of this stage is to craft a data story that either supports or questions a hypothesis based on our study findings. Clear communication should incorporate **key EDA and inferential results**.

Effective storytelling must begin with a **succinct statement that encompasses the main inferential inquiries**, outlines our null and alternative hypotheses in plain language, specifies the chosen significance level (with a tailored explanation for our specific stakeholders), explains why we selected our particular test flavour (i.e., the rationale behind our data modelling assumptions), and reports **uncertainty quantification**. This transparent listing of all our inferential elements allows stakeholders to assess the reliability of our analysis while gaining a fair understanding of any study limitations.

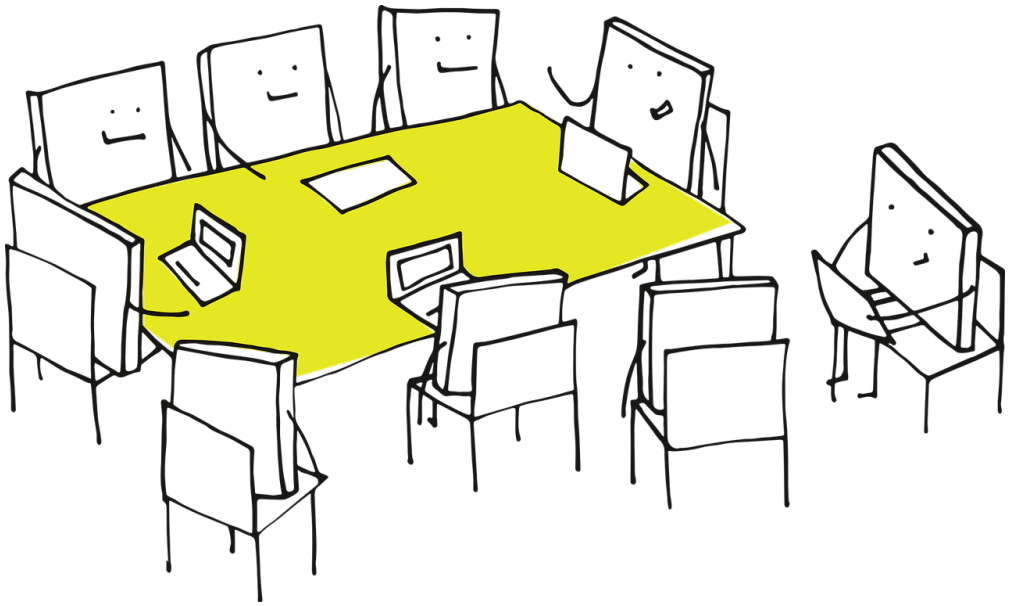


Figure 1.12: Image by *Manfred Stege* via *Pixabay*.

It is important to clarify the difference between statistical significance and practical significance in our storytelling when presenting significant results. Conversely, when discussing non-significant results, we should frame them in a way that indicates the sampled data and study design did not provide enough evidence against the status quo represented by the null hypothesis. This lack of evidence, in light of a non-significant result, is an excellent opportunity to discuss potential issues such as an underpowered study, limitations in sample size, or bias in data collection.

Heads-up on uncertainty quantification!

When presenting our findings to stakeholders, it is essential to include uncertainty quantification as a key component of our storytelling. This process incorporates the uncertainty associated with our point estimates of observed effects, as these estimates are derived from random sampling of our population of interest. We express this uncertainty through **confidence intervals**, which show a range of plausible values where our model parameters may lie (or may not!). The information conveyed by these intervals includes:

- A **narrow confidence interval** around a parameter estimate indicates larger precision in the estimation. For stakeholders, this signifies a more reliable estimate, likely resulting from a sufficiently large sample size, low variability in the sample data or well-specified data modelling.
- A **wide confidence interval** around a parameter estimate indicates lower precision in the estimation. For stakeholders, this represents a less reliable estimate, which may stem from a small sample size, high variability in the sampled data or mis-specification of the data modelling.

1.2 The Test Mind Map

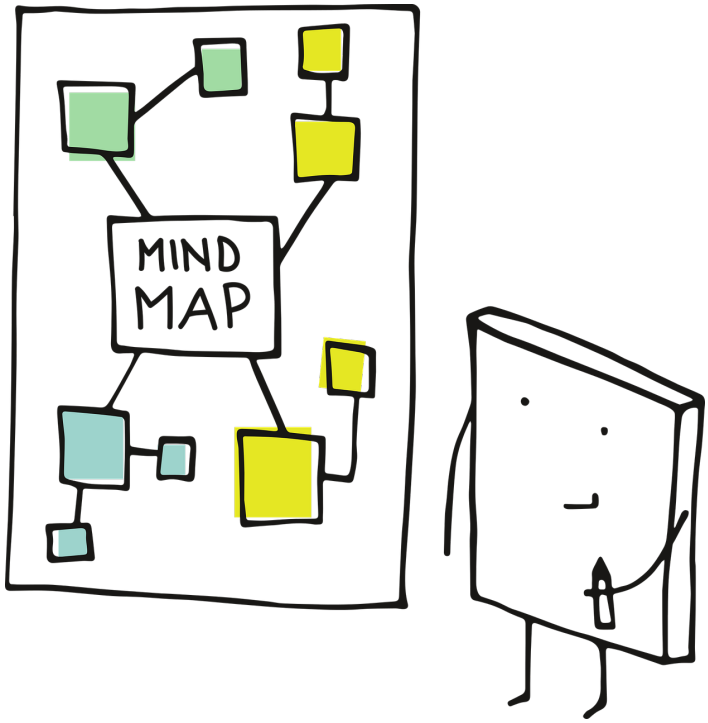


Figure 1.13: Image by *Manfred Steger* via *Pixabay*.

Figure 1.14 outlines the conceptual and organizational structure of this mini-book through its corresponding chapters. This mind map for frequentist hypothesis testings is divided into two main branches: **classical** and **simulation-based** tests. The classical tests are further categorized based on the number of groups being compared (one, two, or k groups), the nature of the variable of interest (unbounded continuous data or proportions derived from binary outcomes), and whether the measurements are independent or related:

- **Chapter 2** focuses on hypothesis tests applied to a single population mean across two different types of responses. For unbounded responses, traditional tests such as the **one-sample t-test** are introduced. For binary responses transformed into proportions (e.g., the fraction of success in a Bernoulli trial), the chapter covers tests for one population proportion, including the **z-test for proportions**.
- **Chapter 3** extends the single-group approach to comparisons between two groups, concentrating on both independent and related populations. For two independent populations with unbounded responses, it discusses the **two-sample t-test**. For binary outcomes between two groups, the chapter explains inference on two proportions using methods like the **z-test for two proportions**. It also addresses related populations (e.g., pre/post measurements or matched pairs) by introducing the **paired-sample t-test**, highlighting how dependency affects the testing framework.
- **Chapter 4** generalizes the two-group comparison to k groups using **analysis of variance (ANOVA)** techniques. This chapter focuses on continuous, unbounded response variables and explains how ANOVA partitions total variation to detect mean differences across groups.

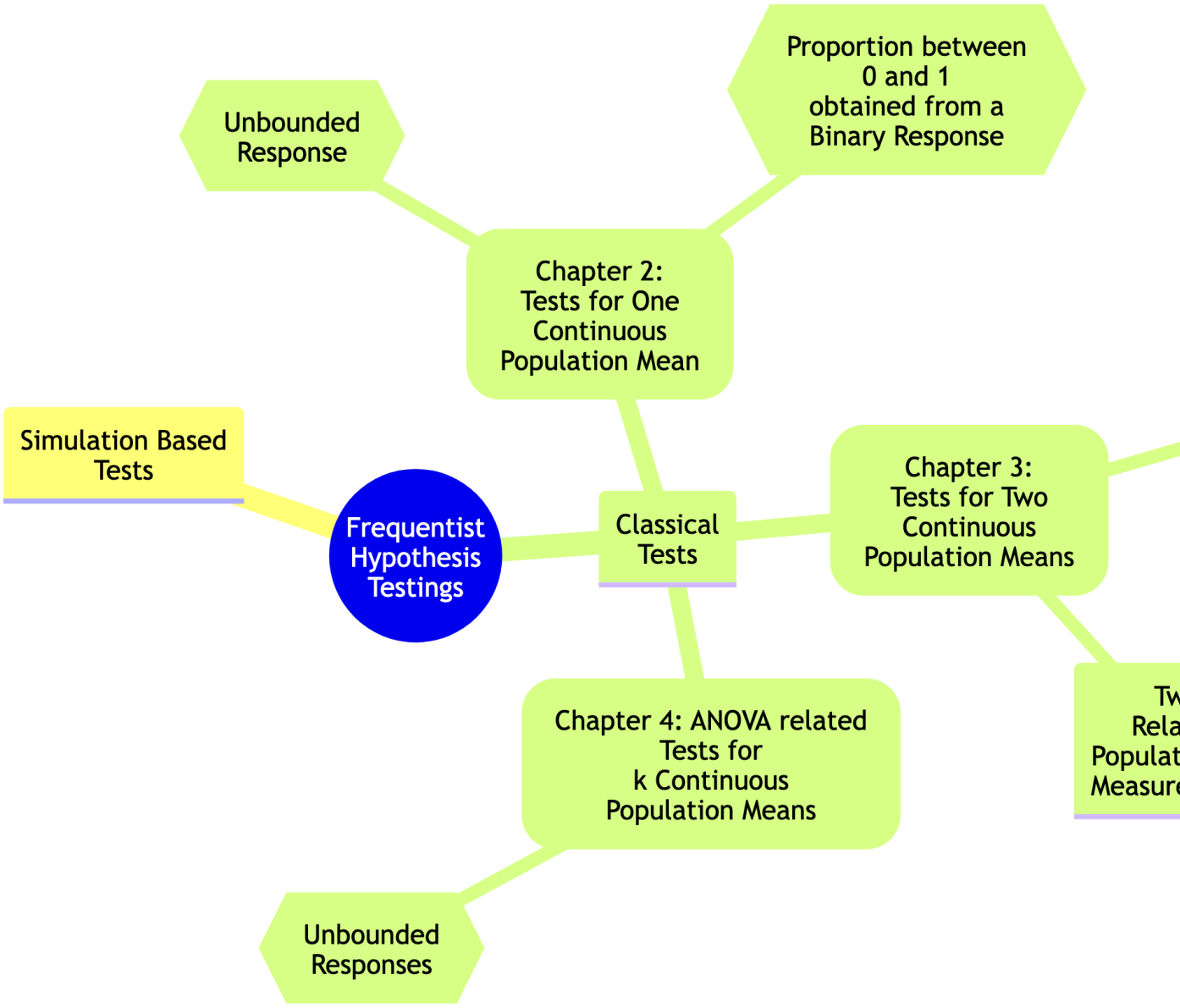


Figure 1.14: A general hypothesis testing mind map outlining all techniques explored in this book. Depending on the overall approach to be used, these techniques are divided into two broad categories: classical and simulation-based tests.

1.3 Chapter Summary

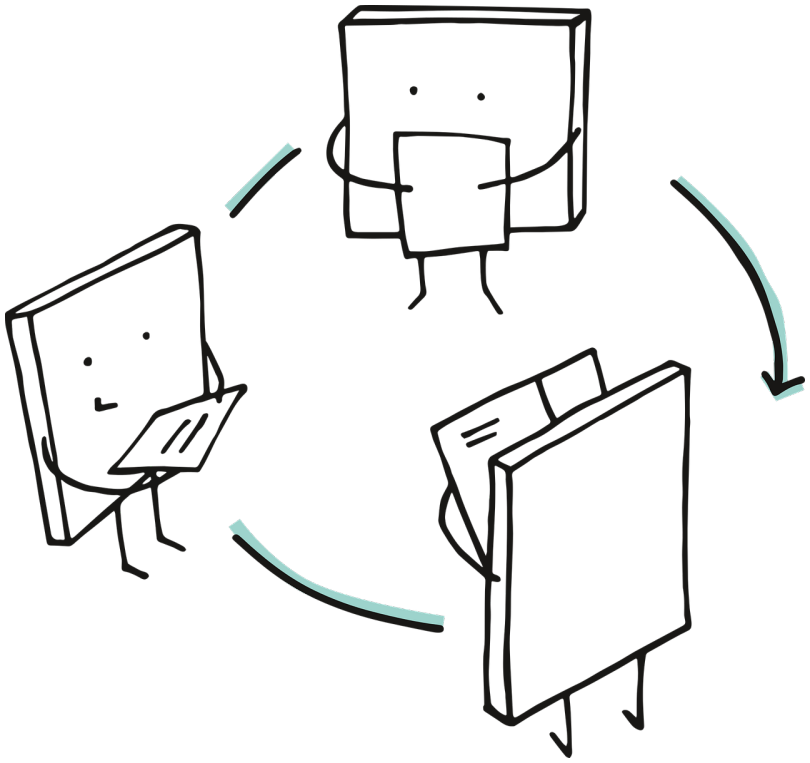


Figure 1.15: Image by *Manfred Steger* via *Pixabay*.

This opening chapter of the mini-book on hypothesis testing introduces the foundational motivations, principles, and practical frameworks that underlie the frequentist approach to statistical inference. We begin by emphasizing hypothesis testing as a central inferential tool in data science and research, which enables practitioners to draw population-level conclusions based on finite sample evidence. Additionally, we revisit the frequentist paradigm, which highlights the logic of repeated sampling and the fixed nature of population parameters.

We also outline the practical components that structure formal hypothesis testing through a comprehensive workflow. This eight-stage framework encompasses everything from study design and data collection to communicating inferential conclusions. Each stage is briefly introduced as a foundation for the upcoming chapters, illustrating how test results are closely linked to modelling assumptions, data preparation strategies, and communication objectives.

2 Tests for One Continuous Population Mean

This chapter introduces statistical tests designed to analyze a single sample, which is a fundamental task in data analysis across many disciplines. Whether you're evaluating whether the average recovery time from a treatment differs from a known standard, assessing whether student test scores exceed a benchmark, or testing if the proportion of success in a group differs from an expected rate, these methods help determine whether the observed values are statistically significant or simply due to chance.

There are several statistical tests used to evaluate hypotheses about a single sample. The appropriate test depends on the type of variable (mean or proportion), sample size, and whether population parameters like variance are known.

We test whether a population mean equals a specific value. The right test depends on:

- Type of response
- Whether the population variance is known
- Sample size

In this chapter, we focus on statistical tests used to evaluate hypotheses about a **single population mean** or **proportion**, based on sample data. These tests help determine whether a sample provides sufficient evidence to conclude that the population mean (or proportion) differs from a specified value.

We cover two cases for the mean — depending on whether the population variance is known or unknown — and one test for binary outcomes where we're testing a population proportion.

Key tests include:

2.1 One-sample z-test for the mean

Use this test when: - The population variance ^{2} **is known**, and - The sample comes from a **normally distributed population**, or the **sample size is large** (typically (n 30)).

The test statistic is:

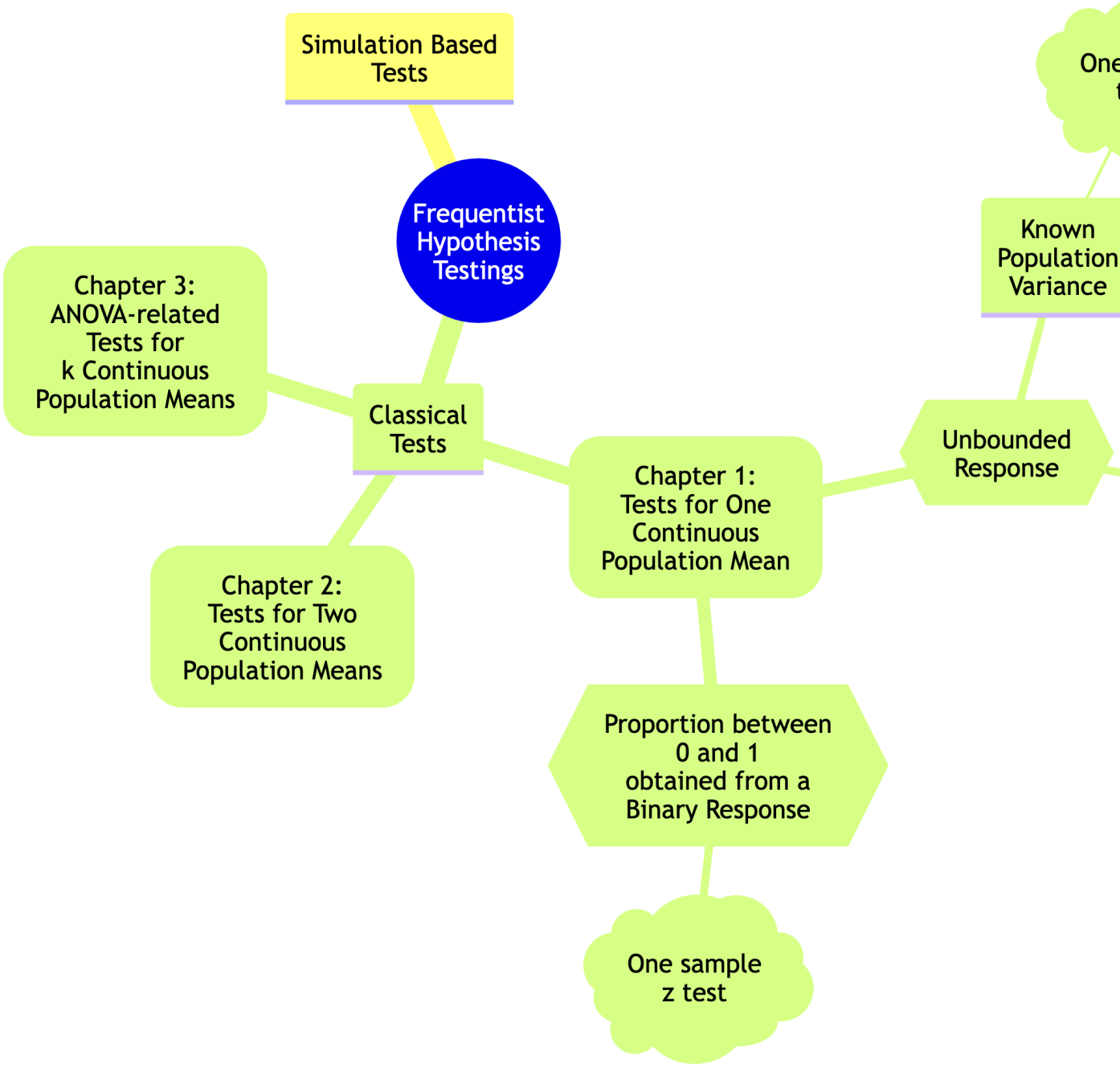


Figure 2.1: A specific hypothesis testing mind map outlining the techniques explored in this chapter, which are classical tests for one continuous population mean.

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Where: - ($\bar{x}$) is the sample mean
- (μ_0) is the hypothesized population mean
- (σ) is the known population standard deviation
- (n) is the sample size

We compare the calculated (z)-value to a standard normal distribution to compute a p-value or make a decision based on a critical value.

2.2 One-sample t-test for the mean

Use this test when: - The population variance **is unknown**, and - The sample is either **normally distributed** or **large enough** to rely on the central limit theorem.

Imagine you want to assess whether a new method of teaching introductory physics improves student performance compared to the traditional method previously used. To explore this, you test the new method at the University of British Columbia (UBC) and compare the results to historical data from students who were taught using the traditional approach. This historical data serves as your **reference value**.

Suppose the population has an unknown average physics score, denoted as:

$$\mu \quad \text{(mean physics score at UBC)}$$

Since we do not have access to the grades of all students, we take a **random sample** from the population. Let this sample consist of n students, with observed scores:

$$X_1, X_2, \ldots, X_n$$

The central question becomes:

> **Is the mean physics score in our sample statistically different from a given reference value?**

If, for example, the historical average physics score is known to be **75**, then our question becomes more specific:

> **Is the mean physics score in the sample statistically different from 75?**

2.2.1 Hypotheses

We can formally express this with the following hypotheses:

- **Null hypothesis** H_0: $\mu = 75$
- **Alternative hypothesis** H_1: $\mu \neq 75$

Under the null hypothesis, we assume that the average score under the new method is equal to the historical average of 75. If the null is rejected, we conclude that there is a **statistically significant difference**, suggesting that the new method may lead to either **higher or lower** average performance.

The test statistic is:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Where: - (s) is the sample standard deviation (used instead of (σ))

This statistic follows a **t-distribution** with (n - 1) degrees of freedom.

2.2.2 Study Design

In this example we use the **Palmer Station Penguins** dataset collected by the LTER in Antarctica (2007 – 2009).
The dataset spans three penguin species and includes continuous variables such as *flipper length*, *bill size*, and *body mass*.

> **Research question:**
> *Is the average flipper length of penguins significantly different from 200 mm?*

2.2.3 Data Collection & Wrangling

We obtain the dataset **Palmer Station Penguins** dataset collected by the 'LTER'

```python
import seaborn as sns
import pandas as pd
from sklearn.model_selection import train_test_split

# Load dataset
penguins = sns.load_dataset("penguins")

# Drop rows with missing values
penguins_clean = penguins.dropna()

# 80/20 train-test split
train_set, test_set = train_test_split(
    penguins_clean, test_size=0.2, random_state=42
)
```

2.2.4 Exploratory Data Analysis (EDA)

Before conducting the statistical test, we begin with an exploratory analysis to understand the distribution and characteristics of the flipper_length_mm variable.

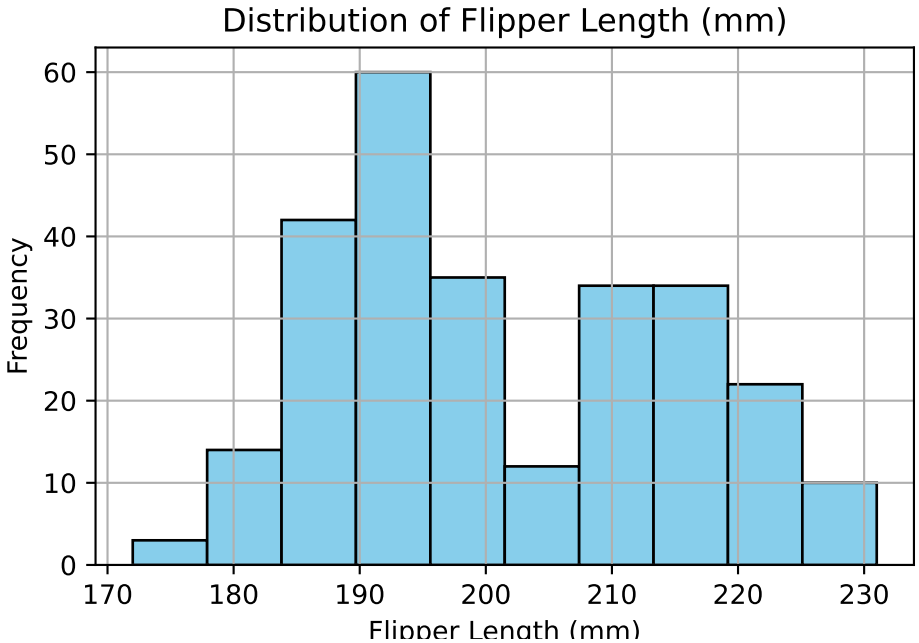
First, we examine summary statistics such as the mean, standard deviation, and quartiles. This helps us get a sense of the central tendency and spread of the data:

```python
print(train_set["flipper_length_mm"].describe())
```

```
count    266.00000
mean     201.00000
std       13.91592
min      172.00000
25%      190.00000
50%      197.00000
75%      213.00000
max      231.00000
Name: flipper_length_mm, dtype: float64
```

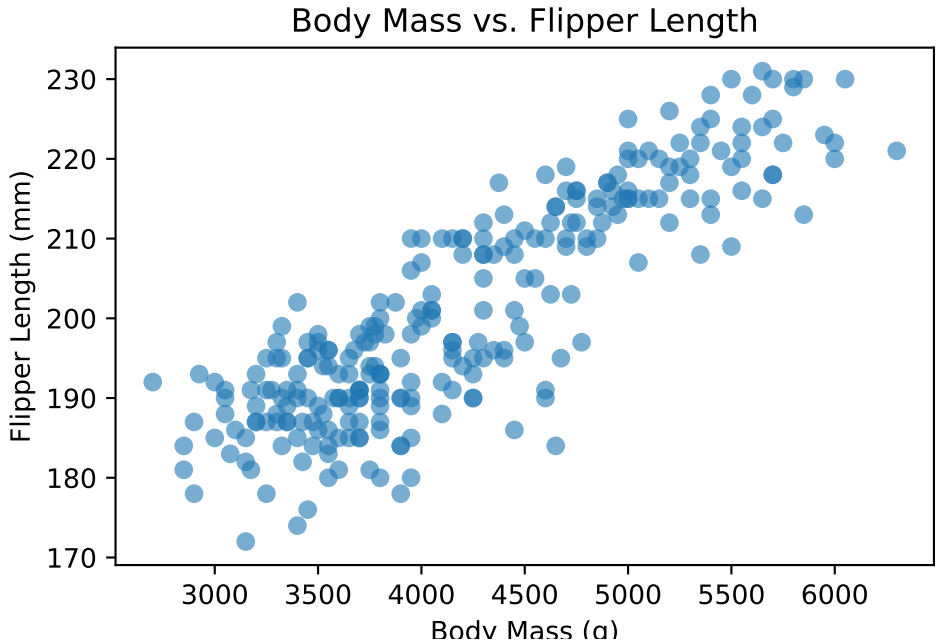
Next, we visualize the distribution of flipper lengths using a histogram. This allows us to assess whether the data are approximately symmetric and whether any outliers are present:

```
import matplotlib.pyplot as plt

train_set["flipper_length_mm"].hist(edgecolor="black", color="skyblue")
plt.title("Distribution of Flipper Length (mm)")
plt.xlabel("Flipper Length (mm)")
plt.ylabel("Frequency")
plt.show()
```



To explore the relationship between flipper length and another continuous variable, we create a scatter plot of flipper length versus body mass. This helps us visually assess whether larger penguins tend to have longer flippers, and whether this relationship is linear or varies across ranges:

```
plt.scatter(
    train_set["body_mass_g"],
    train_set["flipper_length_mm"],
    alpha=0.6
)
plt.title("Body Mass vs. Flipper Length")
plt.xlabel("Body Mass (g)")
plt.ylabel("Flipper Length (mm)")
plt.show()
```



Now, we can perform one-Sample t-Test

```
import scipy.stats as stats

t_stat, p_value = stats.ttest_1samp(
    train_set["flipper_length_mm"], popmean=200
)
print(f"t = {t_stat:.3f},  p = {p_value:.4f}")
```

```
t = 1.172,  p = 0.2422
```

A one-sample t-test was conducted to determine whether the average flipper length of penguins is significantly different from 200 mm. Based on a training sample, the test produced a t-statistic of t and a p-value of p.

Given a significance level of 0.05, if the p-value is less than 0.05, we reject the null hypothesis and conclude that the average flipper length is significantly different from 200 mm. If not, we do not have sufficient evidence to say it differs.

2.3 One-sample z-test for proportions

Use this test when: - The variable is **binary** (success/failure, yes/no, etc.), and - You want to test a **population proportion** (p), using a large enough sample.

The test statistic is:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

Where: - ($\hat{p}$) is the sample proportion
- (p_0) is the hypothesized population proportion
- (n) is the sample size

This test assumes ($np_0 \geq 5$) and ($n(1-p_0) \geq 5$) to justify the normal approximation to the binomial distribution.

---

3 Tests for Two Continuous Population Mean

This chapter introduces statistical tests designed to compare two samples which is a fundamental task in data analysis across many disciplines. Whether you're comparing average recovery times between two medical treatments, student test scores under different teaching methods, comparing the proportion among two samples, or reaction times under varying stress conditions, these methods help determine whether observed differences are statistically significant or simply due to chance.

In this chapter, we review tests for comparing two continuous population means under two conditions: when the populations are independent and when they are dependent. Throughout the sections below, we provide details about these tests and required formula for each case. Broadly speaking, there are two main types of tests to compare the means between two continuous populations:

- Independent samples, where the observations in one group are unrelated to those in the other, and
- Paired (or dependent) samples, where observations are naturally matched in some way, such as before-and-after measurements.

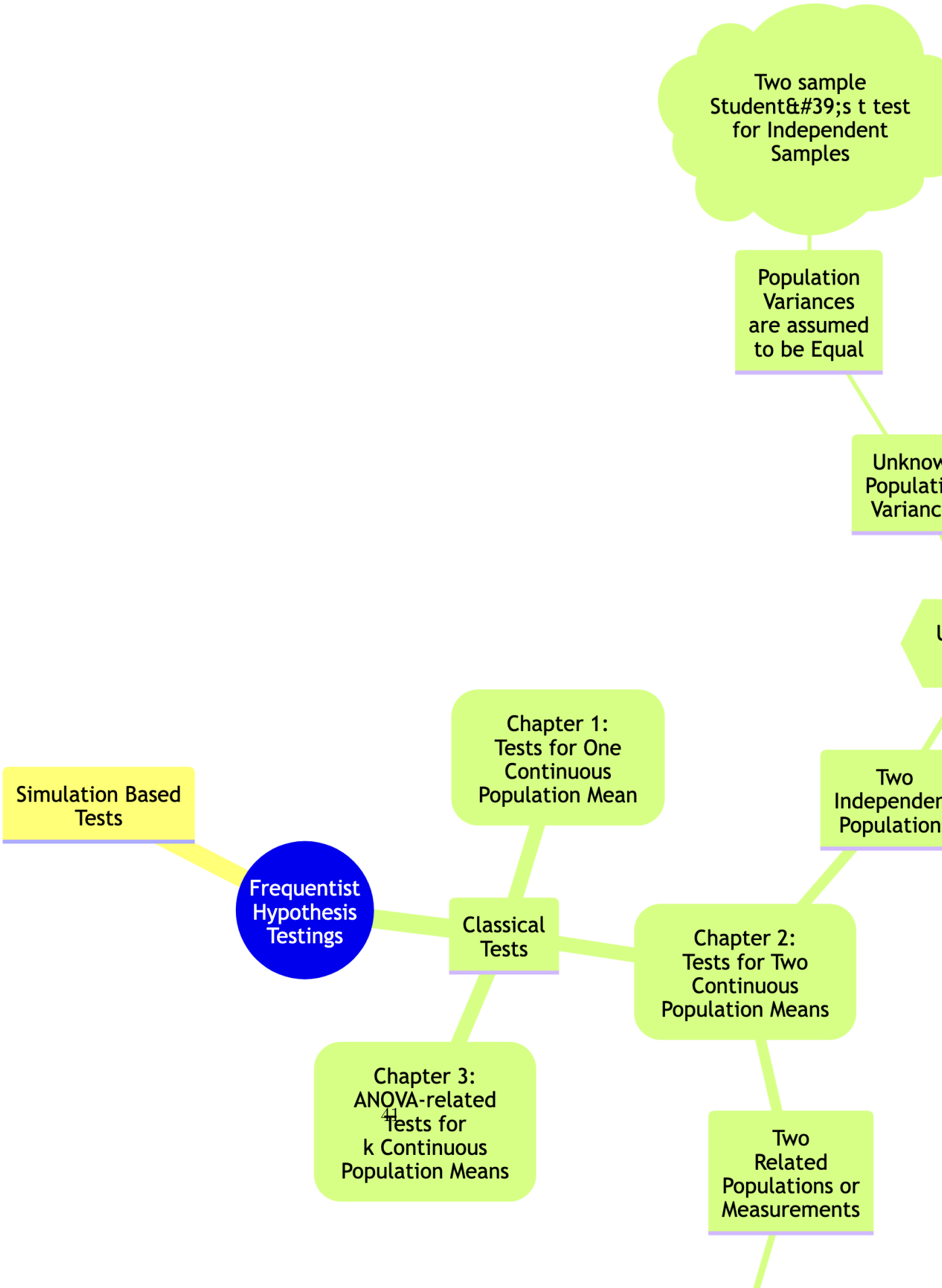
The choice of test depends on the structure of your data. This chapter introduces both types of comparisons, beginning with independent samples. Each section includes definitions, theoretical background, and R/Python code examples using real or simulated datasets to help ground the concepts in practice. We also review the theoretical background and exmaple codes to test two proportions.

3.1 Two sample Student's t-test for Independent Samples

3.1.1 Review

In this section we talk about two sample student's t-test for independent samples. Independent samples arise when the observations in one group do not influence or relate to the observations in the other. In statistical terms we call this two independent samples. A classic example from educational research is described below:

Suppose you're interested in whether a new method of teaching introductory physics improves student performance and learning experience. To investigate this, you decide to test the


Two sample Student's t test for Independent Samples
Population Variances are assumed to be Equal
Unknow Populati Varianc
Chapter 1: Tests for One Continuous Population Mean
Simulation Based Tests
Two Independer Population
Frequentist Hypothesis Testings
Classical Tests
Chapter 2: Tests for Two Continuous Population Means
Chapter 3: ANOVA-related Tests for k Continuous Population Means
Two Related Populations or Measurements

method at two universities: the University of British Columbia (UBC) and Simon Fraser University (SFU). You apply the new teaching method at SFU and compare the results to students taught with the traditional method at UBC.

In this scenario, students at UBC and SFU form two distinct, unrelated groups. Since the students are not paired or matched across schools, and each individual belongs to only one group, the samples are independent. Note that the samples are drawn from two independent population: students at UBC and SFU, respectively.

Let us assume that each population has an unknown average or mean physics score denoted by:

$$\mu_1 \quad \text{(mean for UBC)}, \quad \mu_2 \quad \text{(mean for SFU)}.$$

Since we do not have access to all students' grades, we take a random sample from each school. Suppose:

- From UBC (Population 1), we obtain a sample of size n, denoted as:
$$X_1, X_2, \ldots, X_n$$
- From SFU (Population 2), we obtain a sample of size m, denoted as:
$$Y_1, Y_2, \ldots, Y_m$$

Note that the sample sizes n and m do not necessarily have to be equal. Now, the central question becomes:

Is there a statistically significant difference between the mean physics scores among two groups?

In formal terms, we test the hypotheses:

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_A : \mu_1 \neq \mu_2$$

Now that we reviewed the test concept, let's try to understand it in a read dataset. The steps below follows closely with the roadmap that we introduced in [LINK HERE].

3.1.2 Study design

For this example, we will be using `Auto` dataset from `ISLR` package. This dataset contains gas mileage, horsepower, and other information for 392 vehicles. Some of variables of interest are: 1) `cylinders` an integer (numerical) value between 4 and 8 which indicates the number of cylinders of car, and 2) `horsepower` which shows engine horsepower. You may wondering **if the mean of horsepower in cars with 8 cylinders is statistically different than the means in cars with 4 cylinders?**

3.1.3 Data Collection and Wrangling

To answer this question, we obtain the dataset which is available in `ISLR` package. Note that we consider this data a random sample from population of cars. First we creat a new copy of this dataset to avoid touching the actual data (this is optional). Also we filter rows to those cars with 4 or 8 cylinders only.

```
# Get a copy of dataset
auto_data <- Auto

# Filter rows to cars with 4 or 8 cylinders
auto_data <- auto_data %>% filter(cylinders %in% c(4,8) )
```

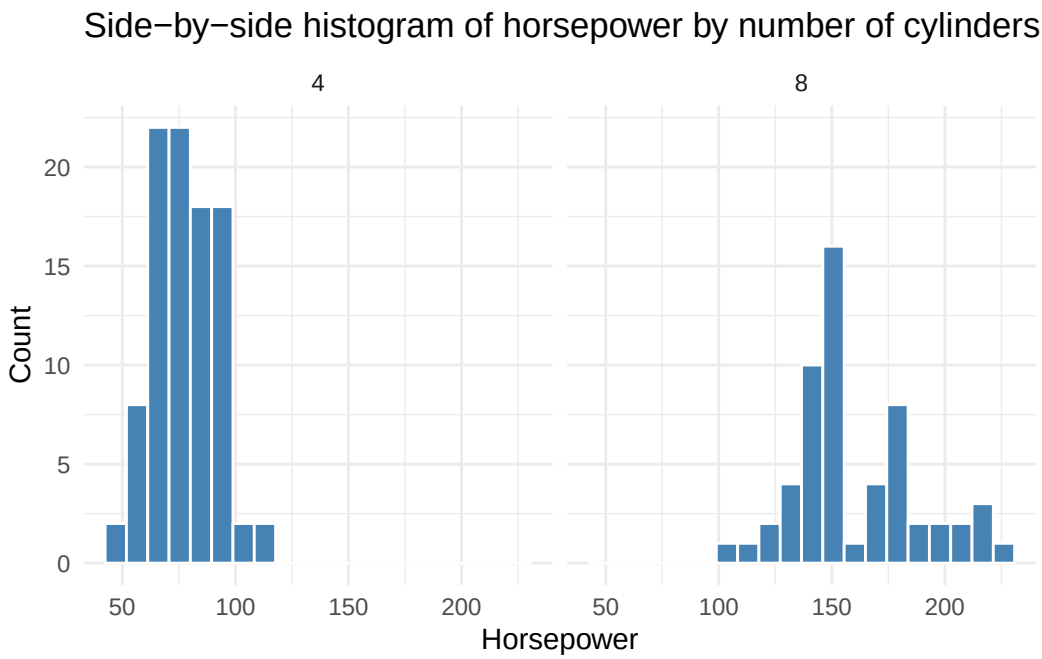
Finally, we randomly create test and train set from this dataset. We use a proportion of 50-50 between train and test.

```
# Set seed for reproducibility
set.seed(123)

# Splitting the dataset into train and test sets
train_indices <- sample(seq_len(nrow(auto_data)), size = 0.50 * nrow(auto_data))
train_auto <- auto_data[train_indices, ]
test_auto <- auto_data[-train_indices, ]
```

3.1.4 Explanatory Data Analysis

Once we have the data and it is split into training and test sets, the next step is to begin exploratory data analysis (EDA) on train set. This step is crucial, as it helps us gain a better understanding of the distribution of variables in our dataset. The `horsepower` variable in dataset is a numerical variable. The `cylinders` variable is an integer variable that helps to divide observations into two groups.

In particular, we are interested in the distribution of `horsepower` in two different groups (cars with 4 cylinders vs cars with 8 cylinders). Using a histogram for this variable is a good choice as we have a variable with numerical values.



We also look at some descriptive statistics of horsepower in both groups for better understanding of data. The descriptive statistics in cars with 4 cylinders:

```
summary(train_auto %>% filter(cylinders == 4) %>% select(horsepower))
```

```
   horsepower
 Min.   : 46.00
 1st Qu.: 68.00
 Median : 78.50
 Mean   : 78.33
 3rd Qu.: 88.00
 Max.   :113.00
```

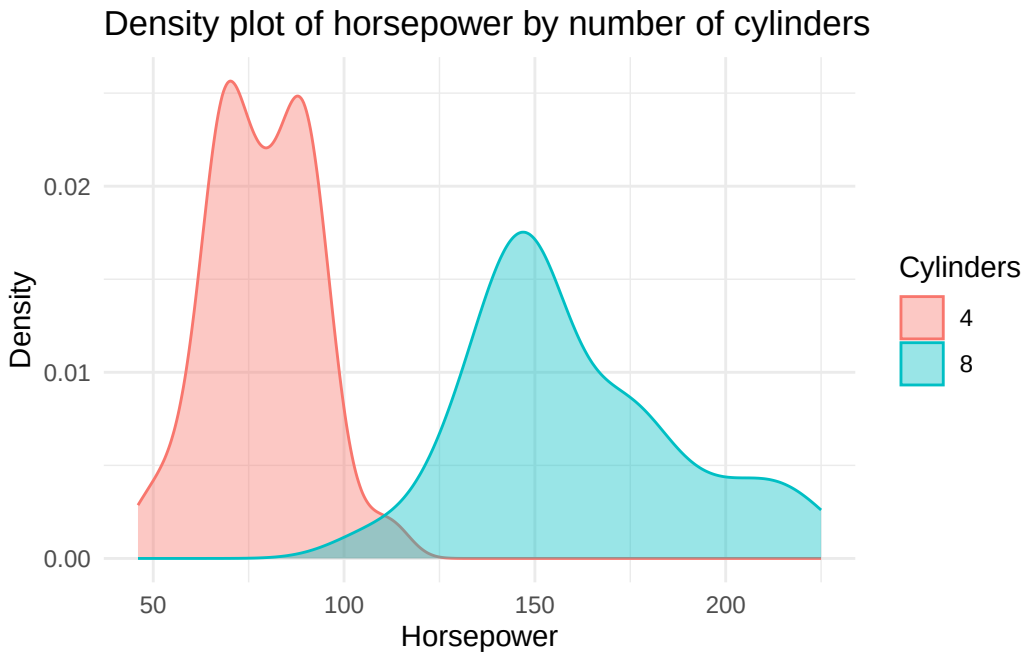
and with 8 cylinders:

```
summary(train_auto %>% filter(cylinders == 8) %>% select(horsepower))
```

```
   horsepower
 Min.   :105
 1st Qu.:140
 Median :150
 Mean   :160
```

```
3rd Qu.:175
Max.   :225
```

Looking at summary statistics, there is a bit of overlap between distribution of `horsepower` among two groups but it does not seem to be much. In fact they seem to be quite separated. Also there is a clear different in their mean and the following plot also confirms this:



3.1.5 Testing Settings

We use a significant level of $\alpha = 0.05$ to run the test. Considering the data we have is a sample from a population of cars we have the following:

- μ_1 is the mean of horsepower for cars with 4 cylinders in the population.
- μ_2 is the mean of horsepower for cars with 8 cylinders in the population.

3.1.6 Hypothesis Definitions

We now define the null and alternative hypothesis. Recall the main inquiry we had:

You may wondering if the average of horsepower in cars with 4 cylinders is statistically different than the means in cars with 8 cylinders?

This translates into the following null and alternative hypotheses:

$$H_0 : \mu_1 = \mu_2 \quad vs \quad H_a : \mu_1 \neq \mu_2$$

Note that the alternative hypothesis is two-sided, as our question does not favor either group and only asks whether the means are different (i.e., group one could be less than or greater than group two). Also the hypothesis tests the unknown parameters in the population which are μ_1 and μ_2.

3.1.7 Test Flavour and Components

To test this hypothesis, we use the **two-sample student's t-test for independent samples**, which compares the sample means and incorporates variability within and between the samples. Note that in this case the samples are independent as clearly cars with 4 cylinders are independent from cars with 8 cylinders.

Now we need to compute a test statistic from the sample. Assuming equal population variances, the test statistic is:

$$t = \frac{(\bar{X} - \bar{Y})}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

where:

- $\bar{X}$ is the mean of horsepower for cars with 4 cylinders in the sample
- $\bar{Y}$ is the mean of horsepower for cars with 8 cylinders in the sample
- S_p is the **pooled standard deviation**, computed as:
- $S_p = \sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}$
- S_X^2 and S_Y^2 are the sample variances of the two groups.

Heads-up!

Note that all elements in this formula (statistic) are computed based on sample.

Tip:

The assumption in this test is that variances among two groups are equal meaning that if we look at the random variable of horsepower in both populations, the variance of this random variable is roughly equal in two groups (cars with 4 cylinders and cars with 8 cylinders).

Note that we do not have access to population and this is rather an assumption that we make with consultation with experts or justifying it based on previous studies. We will introduce the test without equal variance assumption in the next section.

There are some statistical methods designed to test if the variances of different groups are the same or not. Similar to any hypothesis testing, these tests work on a random sample from the population to run the test. Some of the tests are F-test for Equality of Variances, Levene's Test, and Bartlett's Test.

3.1.8 Inferential Conclusions

As you can see, the test statistic computes the difference between $\bar{X}$ and $\bar{Y}$ and scale it based on the variance of this difference. Now the question is whether this difference is significant or not? In order to answer this question we need to know the behavior of statistic that we defined (t) and have a better understanding of what are typical values of this statistic. Knowing the distribution of this statistic helps us to compute *p-value* of the test as follows:

$$p\text{-}value = 2 \times Pr(T_{n+m-2} \geq |t|)$$

Looking at the formula, we can see that we are essentially calculating how much is it likely to see an observation as big as t or as extreme as t (which we computed from our sample).

Heads-up!

Note that t itself is a random variable as it would change from sample to sample.

Tip:

We skipped the theory behind it but under the assumption that null hypothesis is correct (i.e. $\mu_1 = \mu_2$) then the test statistic defined above (t) follows a t-distribution with $n + m - 2$ degrees of freedom (which we denote it by T_{n+m-2}).

Note: The probability is multiplied by two since we have a two sided hypothesis (alternative is $\mu_1 \neq \mu_2$). For a one sided test (when alternative hypothesis is $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$) we do not need to multiply by two.

Now we compare the *p-value* to our significance level. If the *p-value* is less than the significance level, then we have evidence against the null hypothesis. The reasoning is as follows: we performed the calculation under the assumption that the null hypothesis is true. If the null hypothesis is true, then the test statistic we computed should follow a t-distribution with $n + m - 2$ degrees of freedom. If the p-value is smaller than our chosen significance level, this means it is unlikely that our observed result comes from a t-distribution with $n+m-2$ degrees of freedom. In other words, it is unlikely that the null hypothesis is correct.

Note that our observation from the sample might still lead us to an incorrect conclusion (since there is variability among samples). Our tolerance for this type of error is determined by the significance level. If *p-value* is not less than significant level then we do not have any evidence to reject the null hypothesis. Now let us see how to run the two-sample test in R and Python. Note that for the purpose of hypothesis testing we now use test data to avoid double dipping.

3.1.9 How to run the test in R and Python?

The following lines of code in tabset show you how to run the test in R or Python. Note that there are two ways of running this test in R as shown below. They both give the same result and you are welcome to use either of them. Here is a quick explanation from a coding perspective:

- In **Option 1**, we first select the cars with 4 or 8 cylinders and save them in a vector (`cylinders_4` and `cylinders_8`). We then use `t.test` function to run the test.
- In **Option 2**, we use a formula to tell R what is the variable that records the outcome of interest (in this example `horsepower` variable) and what is the grouping variable (in this example `cylinders`). This approach is more concise and easier to read, especially when working directly with a data frame. Note that we need to let R know where it can find `horsepower` and `cylinders` which we do by setting `data = test_auto`.

3.1.10 R Code - Option 1

```r
# Create a vector to hold horsepower values for cars with 4 cylinders
cylinders_4 <- test_auto %>% filter(cylinders == 4) %>% select(horsepower)

# Create a vector to hold horsepower values for cars with 8 cylinders
cylinders_8 <- test_auto %>% filter(cylinders == 8) %>% select(horsepower)

# Run the test
t.test(x = cylinders_4, y = cylinders_8, var.equal = TRUE)
```

```
    Two Sample t-test

data:  cylinders_4 and cylinders_8
t = -21.344, df = 149, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -85.12730 -70.70086
sample estimates:
mean of x mean of y
  78.2381  156.1522
```

3.1.11 R Code - Option 2

```r
# Use the formula horsepower ~ cylinders to run the test
t.test(horsepower ~ cylinders, data = test_auto, var.equal=TRUE)
```

```
    Two Sample t-test

data:  horsepower by cylinders
t = -21.344, df = 149, p-value < 2.2e-16
alternative hypothesis: true difference in means between group 4 and group 8 is not equal to
95 percent confidence interval:
 -85.12730 -70.70086
sample estimates:
mean in group 4 mean in group 8
        78.2381        156.1522
```

3.1.12 Python Code

```python
from scipy import stats
import pandas as pd

# Read test_auto dataframe in Python as df dataframe
df = pd.read_csv('data/test_auto.csv')

# Select cars with 4 and 8 cylinders
cylinders_4 = df[df["cylinders"] == 4]["horsepower"]
cylinders_8 = df[df["cylinders"] == 8]["horsepower"]

# Run the test
t_stat, p_val = stats.ttest_ind(cylinders_4, cylinders_8, equal_var = True)

# Print t statistic value
print(f"T-statistic: {t_stat}")
```

```
T-statistic: -21.34403814660459
```

```python
# Print p-value of the test
print(f"P-value: {p_val}")
```

```
P-value: 3.6294706302411423e-47
```

In order to run this test, similar to what we learned in (LINK to chapter 1) we can use `t.test` function in R. The function can be used to perform one or two sample t-tests. The relevant arguments of the function are as follows:

- `x` is (non-empty) numeric vector of data values.
- `y` is also (non-empty) numeric vector of data values (can be `NULL` if you run a one sample test).
- `var.equal` is a binary value (TRUE/FALSE) to indicate if R needs to assume equal variance or not.

In both outputs, we can see the following:

- `t` is the test statistic.
- `df` is the degrees of freedom for the test.

`p-value` is the p-value of the test. Note that, by default, this is for a two-sided test. If you need to conduct a one-sided test, you can either divide the p-value by two or use the alternative argument in the t.test function.

- `95 percent confidence interval` provides the 95% confidence interval for the parameter of $\mu_1 - \mu_2$.
- `sample estimates` gives the sample means for each group.

Note: By default the value of `var.equal` is `FALSE`. We manully set it to `TRUE` to implement equal variance assumption in our test.

3.1.13 Storytelling

Finally, based on the sample we have and the analysis we conducted, we can draw a conclusion about our initial question: **Is the mean horsepower of cars with 8 cylinders statistically different from that of cars with 4 cylinders?** We observed that the $p-value$ of the test was extremely small compared to the significance level $\alpha = 0.05$. This provides evidence against the null hypothesis. In simple terms, this means: *There appears to be a noticeable difference in the average horsepower between cars with 4 cylinders and those with 8 cylinders.*

3.2 Two sample Welch's t-test for independent samples

3.2.1 Review

In this section we talk about two sample Welch's t-test for independent samples. This test is very similar to two sample Student's t-test for independent samples that we described with a caveat. The two samples are still independent but the only difference is the equal variance assumption. We use this test if we do not have any reason or evidence to believe that the variance of variable of interest is the same among two groups in the population.

3.2.2 Study Design

We will be using `Auto` dataset from `ISLR` package in this section too. Now the main statistical question of interest remains the same as before: **You may wondering if the mean of horsepower in cars with 8 cylinders is statistically different than the means in cars with 4 cylinders?** but we do **not** make an equal variance assumption anymore. Now we are applying a two sample Welch's t-test for independent samples.

3.2.3 Data Collection and Wrangling

To answer this question, we obtain the dataset which is available in `ISLR` package. The following codes are exactly the same as before and are shown here as a review.

```
# Get a copy of dataset
auto_data <- Auto

# Filter rows to cars with 4 or 8 cylinders
auto_data <- auto_data %>% filter(cylinders %in% c(4,8) )
```

Finally, we randomly create test and train set from this dataset. We use a proportion of 50-50 between train and test.

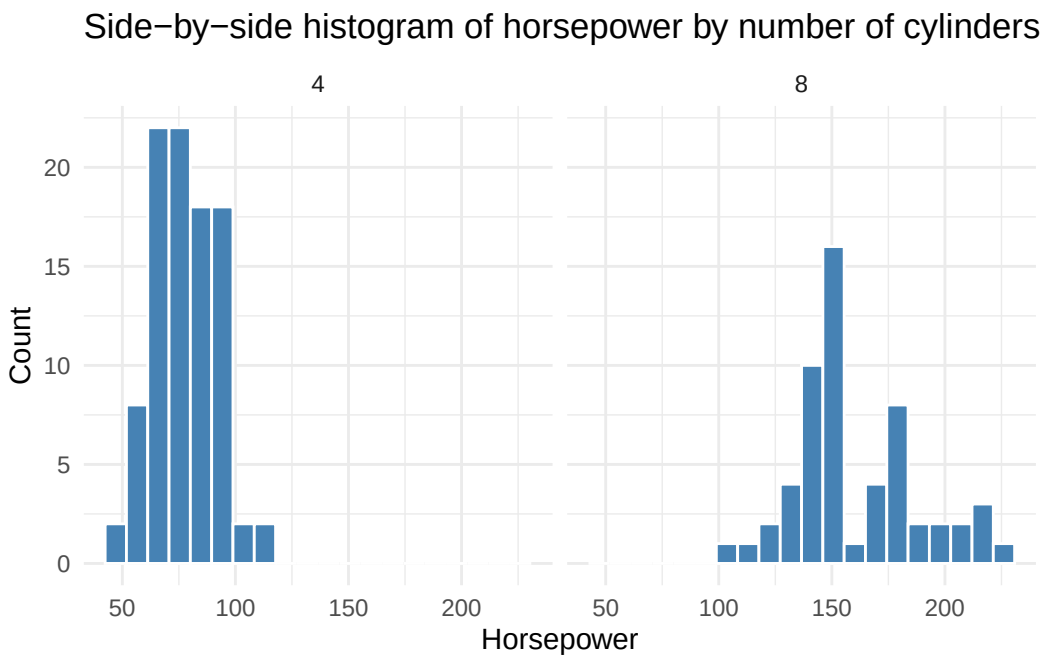
```
# Set seed for reproducibility
set.seed(123)

# Splitting the dataset into train and test sets
train_indices <- sample(seq_len(nrow(auto_data)), size = 0.50 * nrow(auto_data))
train_auto <- auto_data[train_indices, ]
test_auto <- auto_data[-train_indices, ]
```

3.2.4 Explanatory Data Analysis

Once we have the data and it is split into training and test sets, the next step is to begin exploratory data analysis (EDA) on train set. Recall that the `cylinders` variable is an integer variable that helps to divide observations into two groups.

We are still interested in the distribution of `horsepower` in two different groups (cars with 4 cylinders vs cars with 8 cylinders). Using a histogram for this variable is a good choice as we have a variable with numerical values.

The following lines of code are the same as previous section as we are working on the same data. This is shown as a reminder.



We also look at some descriptive statistics of horsepower in both groups for better understanding of data. The descriptive statistics in cars with 4 cylinders:

```
summary(train_auto %>% filter(cylinders == 4) %>% select(horsepower))
```

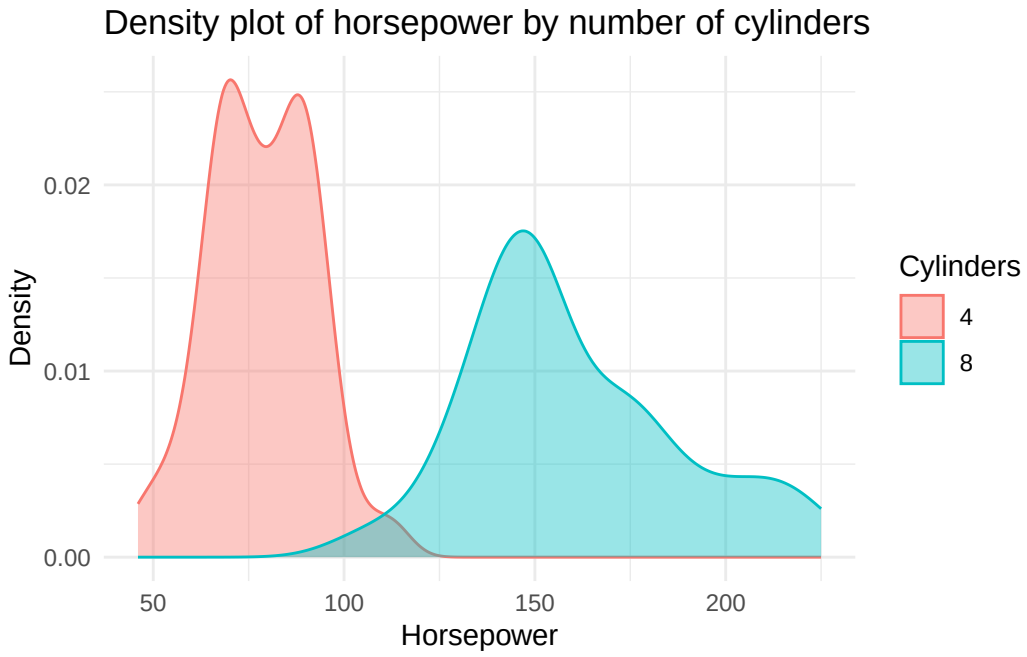
```
  horsepower
 Min.   : 46.00
 1st Qu.: 68.00
 Median : 78.50
 Mean   : 78.33
 3rd Qu.: 88.00
 Max.   :113.00
```

and with 8 cylinders:

```
summary(train_auto %>% filter(cylinders == 8) %>% select(horsepower))
```

```
   horsepower
 Min.   :105
 1st Qu.:140
 Median :150
 Mean   :160
 3rd Qu.:175
 Max.   :225
```

Our conclusion remains the same. Looking at summary statistics, there is a bit of overlap between distribution of `horsepower` among two groups but it does not seem to be much. In fact they seem to be quite separated. Also there is a clear different in their mean and the following plot also confirms this:



3.2.5 Testing Settings

We use a significant level of $\alpha = 0.05$ to run the test. Considering the data we have is a sample from a population of cars we have the following:

- μ_1 is the mean of horsepower for cars with 4 cylinders in the population.
- μ_2 is the mean of horsepower for cars with 8 cylinders in the population.

3.2.6 Hypothesis Definitions

We now define the null and alternative hypothesis. Recall the main inquiry we had:

You may wondering if the average of horsepower in cars with 4 cylinders is statistically different than the means in cars with 8 cylinders?

This translates into the following null and alternative hypotheses:

$$H_0 : \mu_1 = \mu_2 \quad vs \quad H_a : \mu_1 \neq \mu_2$$

Note that the alternative hypothesis is two-sided, as our question does not favor either group and only asks whether the means are different (i.e., group one could be less than or greater than group two). Also the hypothesis tests the unknown parameters in the population which are μ_1 and μ_2.

3.2.7 Test Flavour and Components

As noted before we use Welch's t-test if the assumption of equal variances is questionable. This test adjusts the standard error and degrees of freedom (`df`) of the test accordingly. As a result the test statistic and `df` of the test are different. The Welch's test statistic is computed as:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$$

where:

- $\bar{X}$ is the mean of horsepower for cars with 4 cylinders in the sample
- $\bar{Y}$ is the mean of horsepower for cars with 8 cylinders in the sample
- S_X^2 and S_Y^2 are the sample variances of the two groups.
- n and m are the sample sizes in two groups (not necessarily the same).

Quick review!

Note that, as before, all elements in this formula (statistic) are computed based on the sample. Additionally, the assumption of unequal variance between two populations can be tested using a variety of statistical tests. We do not discuss these tests in this book as the focus of this book lies elsewhere.

3.2.8 Inferential Conclusions

As you can see, the test statistic computes the difference between averages of two samples and adjusts it based on the variance of their differences. The only change from Student's t-test is the variance that is being used in the denominator. Again the question is whether this difference is significant or not? In order to answer this question we need to know the behavior of statistic that we defined and have a better understanding of what are typical values of this statistic. Knowing the distribution of this statistic helps us to compute the $p-value$ of the test as follows:

$$p\text{-}value = 2 \times Pr(T_\nu \geq |t|)$$

Tip on degrees of freedom!

The Greek sign ν is used here to show the degree of freedom of the t-distribution and is computed as

$$\nu = \frac{\left(\frac{s_1^2}{n} + \frac{s_2^2}{m}\right)^2}{\frac{\left(\frac{s_1^2}{n}\right)^2}{n-1} + \frac{\left(\frac{s_2^2}{m}\right)^2}{m-1}}$$

We skipped the theory behind it but under the assumption that null hypothesis is correct, the test statistic defined above still follows a t-distribution but with a different degrees of freedom. Note that this degree of freedom is not necessarily an integer number (could be a real number).

3.2.9 How to run the test in `R` and `Python`?

The following lines of code in tabset show you how to run the Welch's test in `R` or `Python`.

A quick reminder!

- In **Option 1**, we first select the cars with 4 or 8 cylinders and save them in a vector (`cylinders_4` and `cylinders_8`). We then use `t.test` function to run the test.
- In **Option 2**, we use a formula to tell `R` what is the variable that records the outcome of interest (in this example `horsepower` variable) and what is the grouping variable (in this example `cylinders`). This approach is more concise and easier to read, especially when working directly with a data frame. Note that we need to let `R` know where it can find `horsepower` and `cylinders` which we do by setting `data = test_auto`.

3.2.10 R Code - Option 1

```r
# Create a vector to hold horsepower values for cars with 4 cylinders
cylinders_4 <- test_auto %>% filter(cylinders == 4) %>% select(horsepower)

# Create a vector to hold horsepower values for cars with 8 cylinders
cylinders_8 <- test_auto %>% filter(cylinders == 8) %>% select(horsepower)

# Run the test
t.test(x = cylinders_4, y = cylinders_8, var.equal = FALSE)
```

```
    Welch Two Sample t-test

data:  cylinders_4 and cylinders_8
t = -16.92, df = 55.789, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -87.13950 -68.68866
sample estimates:
mean of x mean of y
  78.2381  156.1522
```

3.2.11 R Code - Option 2

```r
# Use the formula horsepower ~ cylinders to run the test
t.test(horsepower ~ cylinders, data = test_auto, var.equal = FALSE)
```

```
    Welch Two Sample t-test

data:  horsepower by cylinders
t = -16.92, df = 55.789, p-value < 2.2e-16
alternative hypothesis: true difference in means between group 4 and group 8 is not equal to
95 percent confidence interval:
 -87.13950 -68.68866
sample estimates:
mean in group 4 mean in group 8
        78.2381        156.1522
```

3.2.12 Python Code

```
from scipy import stats
import pandas as pd

# Read test_auto dataframe in Python as df dataframe
df = pd.read_csv('data/test_auto.csv')

# Select cars with 4 and 8 cylinders
cylinders_4 = df[df["cylinders"] == 4]["horsepower"]
cylinders_8 = df[df["cylinders"] == 8]["horsepower"]

# Run the test
t_stat, p_val = stats.ttest_ind(cylinders_4, cylinders_8, equal_var = False)

# Print t statistic value
print(f"T-statistic: {t_stat}")
```

```
T-statistic: -16.919952924079897
```

```
# Print p-value of the test
print(f"P-value: {p_val}")
```

```
P-value: 1.2443553344442986e-23
```

In order to run this test, similar to what we learned in (LINK to chapter 1) we can use `t.test` function in R. The function can be used to perform one or two sample t-tests. The relevant arguments of the function are as follows:

- `x` is (non-empty) numeric vector of data values.
- `y` is also (non-empty) numeric vector of data values (can be `NULL` if you run a one sample test).
- `var.equal` is a binary value (TRUE/FALSE) to indicate if R needs to assume equal variance or not.

In both outputs, we can see the following:

- `t` is the test statistic.
- `df` is the degrees of freedom for the test.

`p-value` is the p-value of the test. Note that, by default, this is for a two-sided test. If you need to conduct a one-sided test, you can either divide the p-value by two or use the alternative argument in the t.test function.

- `95 percent confidence interval` provides the 95% confidence interval for the parameter of $\mu_1 - \mu_2$.
- `sample estimates` gives the sample means for each group.

Note: By default the value of `var.equal` is `FALSE`. We manully set it to `FALSE` to implement the test without equal variance assumption.

3.2.13 Storytelling

Finally, based on the sample we have and the analysis we conducted, we can draw a conclusion about our initial question: **Is the mean horsepower of cars with 8 cylinders statistically different from that of cars with 4 cylinders?** We observed that the $p-value$ of the test was extremely small compared to the significance level $\alpha = 0.05$. This provides evidence against the null hypothesis. In simple terms, this means: *There appears to be a noticeable difference in the average horsepower between cars with 4 cylinders and those with 8 cylinders.*

3.3 Paired Samples

Paired samples arise when each observation in one group is matched or linked to an observation in the other group. This structure is typical in before-and-after studies, matched-subject designs, or repeated measures on the same individuals. A classic example comes from health sciences.

Suppose you're investigating whether a new diet plan reduces blood pressure. You recruit a group of participants and record their blood pressure **before** starting the diet. **After** following the diet for two months, you measure their blood pressure again. In this scenario, each participant contributes two measurements: one before **the intervention** and one after. These measurements are not independent as they come from the same person. Therefore we treat them as paired.

To formulate the problem and hypothesis, let us assume that each individual has two measurements:

- Before the diet: $X_1, X_2, \dots, X_n$
- After the diet: $Y_1, Y_2, \dots, Y_n$

Note that in this case the sample size is the same (in both before and after diet sample we have n observations). We call this a paired sample. Since the samples are paired, we define the difference for each individual as follows:

$$D_i = Y_i - X_i \quad for \quad i = 1, 2, \dots, n$$

Each D_i is the difference of blood pressure after and before using new diet. The main statistical question now is:

Is there a statistically significant difference in the mean blood pressure before and after the diet?

In other words, we test the following hypothesis:

$$H_0 : \mu_D = 0 \quad \text{versus} \quad H_A : \mu_D \neq 0$$

Here the notation of μ_D is the population mean of the differences of D_i which is an unknown parameter in the population. To test this hypothesis, we use the paired t-test, which is essentially a one-sample t-test on the differences $D_1, D_2, \dots, D_n$. We test $\mu_D = 0$ because if there is an actual effect of diet on blood pressure, we expect the null hypothesis to be rejected.

The test statistic for this hypothesis testing is:

$$t = \frac{\bar{D}}{s_D/\sqrt{n}}$$

where:

- $\bar{D}$ is the sample mean of the differences,
- s_D is the sample standard deviation of the differences,
- n is the number of pairs.

The standard deviation of the differences is calculated as:

$$s_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2}$$

Under the null hypothesis, the test statistic follows a t-distribution with $n-1$ degrees of freedom. For this test, we can compute the *p-value* as:

$$p\text{-}value = 2 \times \Pr(T_{n-1} \geq |t|)$$

When we run t-test, we operate under the assumption that: 1) either the sample size is large enough (we are thinking about $n = 30$ at least) so that central limit theorem assumptions work well, or 2) the distribution of our sample in each group is normal or symmetric enough.

If the normality assumption is also not satisfied (e.g., due to skewed distributions or outliers) or we have a very small sample size, we may turn to a non-parametric alternative, such as the Mann–Whitney–Wilcoxon test, which compares the ranks of the observations across groups rather than the raw values but this book will not cover it. You can read more about it LINK.

4 ANOVA-related Tests for k Continuous Population Means

Figure 4.1: A specific hypothesis testing mind map outlining the techniques explored in this chapter, which include ANOVA-related tests for k population means.

References

Lohr, S. L. 2021. *Sampling: Design and Analysis*. Chapman; Hall/CRC. https://doi.org/https://doi.org/10.1201/9780429298899.

R Core Team. 2024. "R: A Language and Environment for Statistical Computing." Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reinhart, Alex. 2015. *Statistics Done Wrong: The Woefully Complete Guide*. 1st ed. San Francisco, CA: No Starch Press. https://www.statisticsdonewrong.com/index.html.

The Pandas Development Team. 2024. "Pandas-Dev/Pandas: Pandas." Zenodo. https://doi.org/10.5281/zenodo.3509134.

Tukey, John W. 1962. "The Future of Data Analysis." *The Annals of Mathematical Statistics* 33 (1): 1–67. https://doi.org/10.1214/aoms/1177704711.

Van Rossum, Guido, and Fred L. Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

A Greek Alphabet

In the context of hypothesis testing in statistics, mathematical notation serves an important purpose: it distinguishes between unknown **population parameters** and **sample-based estimates**. A key convention in this framework is the use of Greek letters to represent population parameters. For example, μ represents the population mean, σ denotes the population standard deviation, and π signifies the population proportion. In a frequentist framework, these letters indicate unknown and fixed parameters that characterize the entire population of interest. Since hypothesis testing primarily focuses on making inferential conclusions about these parameters, we will consistently use this notation throughout all chapters of this mini-book.

Heads-up on the use of π!

In this textbook, unless otherwise stated, the letter π will represent a population parameter and not the mathematical constant 3.141592...

Figure A.1: Image by *meineresterampe* via *Pixabay*.

It is important to remember that each hypothesis test involves formulating **null** and **alternative hypotheses** regarding the population parameter(s) of interest. For instance, when inferring a population mean μ, the null hypothesis in a two-sided one-sample t-test might indicate that this mean is equal to 100 (i.e., H_0: $\mu = 100$), while the alternative hypothesis will indicate that the mean is not equal to 100 (i.e., H_1: $\mu \neq 100$). Using Greek letters to define our hypotheses helps frame the entire test clearly and precisely. If at any point throughout the chapters this notation feels unfamiliar, we recommend consulting Table A.1 as a reference resource. Regular exposure to this notation will enhance your conceptual clarity when performing statistical inference.

Table A.1: Greek alphabet composed of 24 letters, from *left* to *right* you can find the *name* of letter along with its corresponding *uppercase* and *lowercase* forms.

Name	Uppercase	Lowercase
Alpha	A	α
Beta	B	β
Gamma	Γ	γ
Delta	Δ	δ
Epsilon	E	ϵ
Zeta	Z	ζ
Eta	H	η
Theta	Θ	θ
Iota	I	ι
Kappa	K	κ
Lambda	Λ	λ
Mu	M	μ
Nu	N	ν
Xi	Ξ	ξ
O	O	o
Pi	Π	π
Rho	P	ρ
Sigma	Σ	σ
Tau	T	τ
Upsilon	Υ	υ
Phi	Φ	ϕ
Chi	X	χ
Psi	Ψ	ψ
Omega	Ω	ω