

# **The Mini Test Book**

**(in development)**

G. Alexi Rodríguez-Arelis      Kate Manskaia      Payman Nickchi

2025-06-16

TODO

# Table of contents

<b>Preface</b>	<b>5</b>
<b>The Authors</b>	<b>6</b>
G. Alexi Rodríguez-Arelis . . . . .	6
Kate Manskaia . . . . .	6
Payman Nickchi . . . . .	6
<b>License</b>	<b>7</b>
<b>Website Privacy Policy</b>	<b>8</b>
Information Collection and Use . . . . .	8
Personal Information . . . . .	9
<b>Introduction</b>	<b>10</b>
The Test Mind Map . . . . .	10
The Test Workflow . . . . .	10
<b>Chapter 1: Tests for One Continuous Population Mean</b>	<b>13</b>
One-sample z-test for the mean . . . . .	15
One-sample t-test for the mean . . . . .	15
One-sample z-test for proportions . . . . .	16
<b>Chapter 2: Tests for Two Continuous Population Mean</b>	<b>17</b>
Two sample Student's t-test for Independent Samples . . . . .	17
Data Collection and Wrangling . . . . .	20
Explanatory Data Analysis . . . . .	20
Testing Settings . . . . .	20
Hypothesis Definitions . . . . .	20
Test Flavour and Components . . . . .	20
Inferential Conclusions . . . . .	20
Storytelling . . . . .	20
How to run this test in R? . . . . .	22
Two sample Welch's t-test for independent samples . . . . .	23
Paired Samples . . . . .	24
Example dataset in R . . . . .	25
How to run the test in R? . . . . .	25

**Chapter 3: ANOVA-related Tests for  $k$  Continuous Population Means** 26

**References** 27

# Preface

This is a mini book for hypothesis testing in statistics. This book covers the tests from DSCI 552 in MDS program at UBC.

- Todolist: add more context here

# 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. 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**

# **License**

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

# Website Privacy Policy

Last updated

May 17th, 2025.

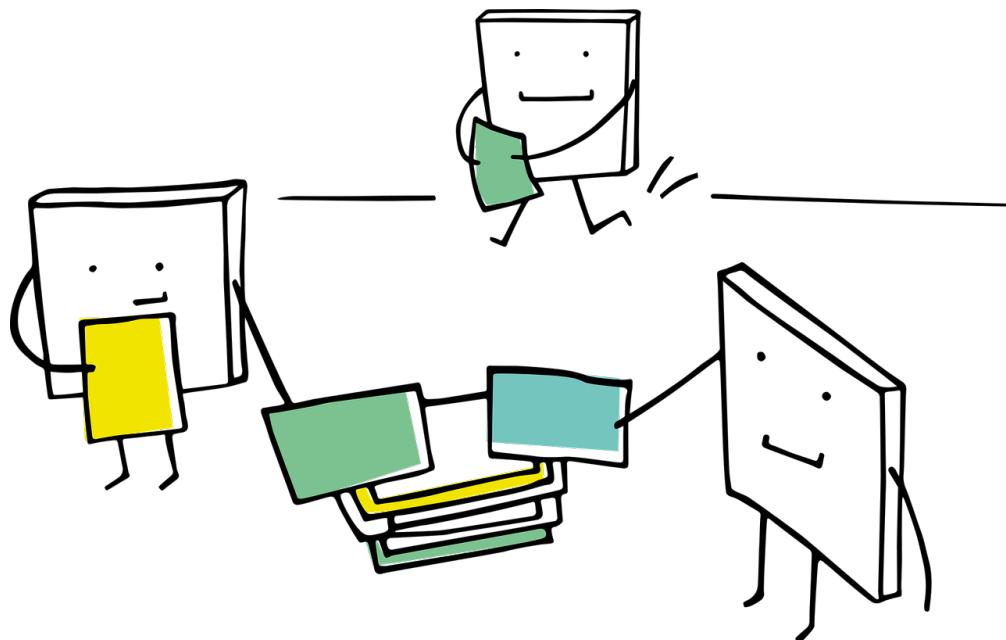


Figure 1: Image by [Manfred Stege](#) via [Pixabay](#).

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](mailto:alexrod@stat.ubc.ca).

# **Introduction**

This is a mini-book on hypothesis testing in statistics. It covers the tests taught in DSCI 552 in the MDS program at UBC.

- Todolist: add more context and information here

## **The Test Mind Map**

## **The Test Workflow**

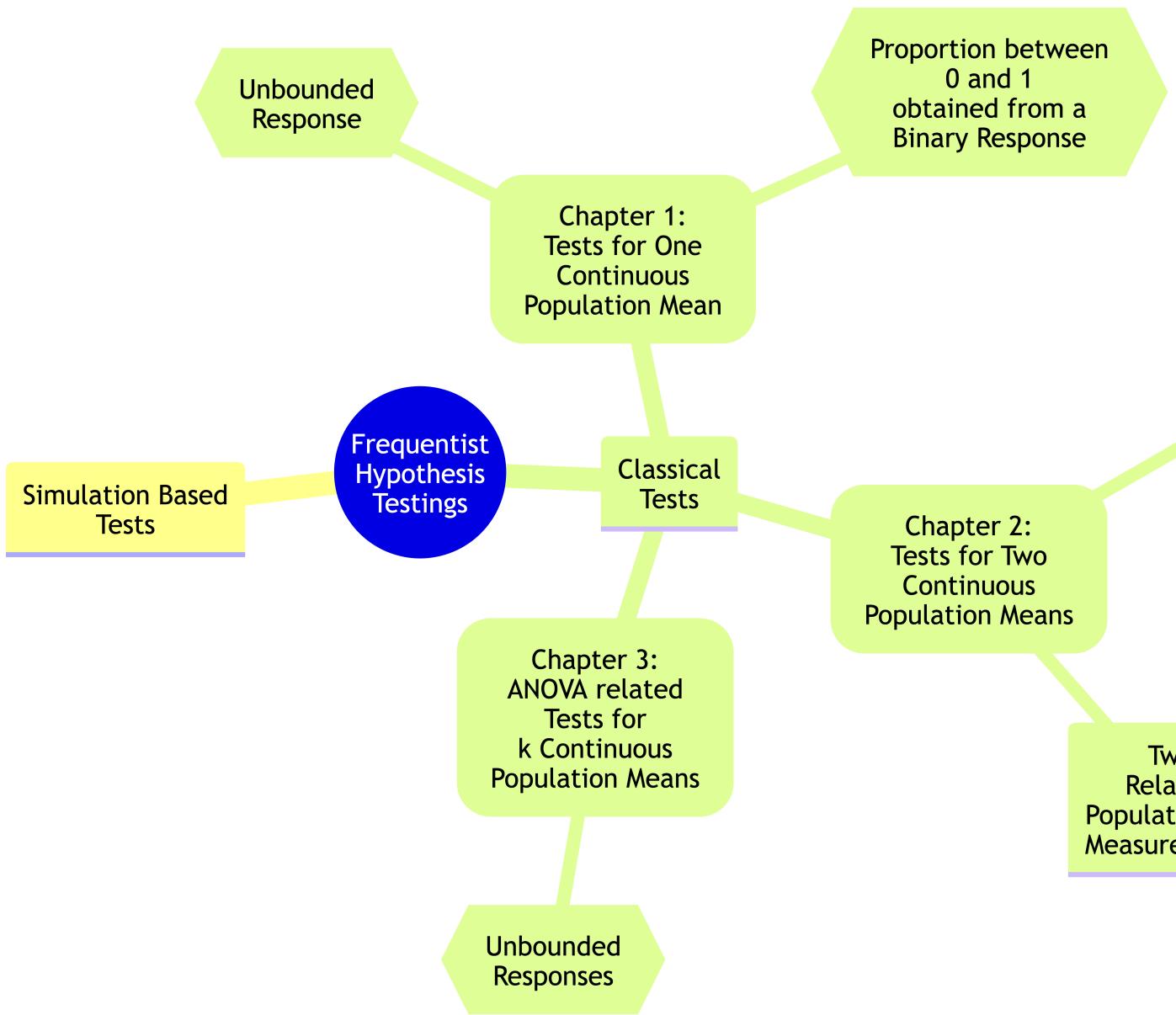


Figure 2: A general hypothesis testing mind map outlining all techniques explored in this book. Depending on the overall approach to be used, these techniques are divided into two broad categories: classical and simulation-based tests.

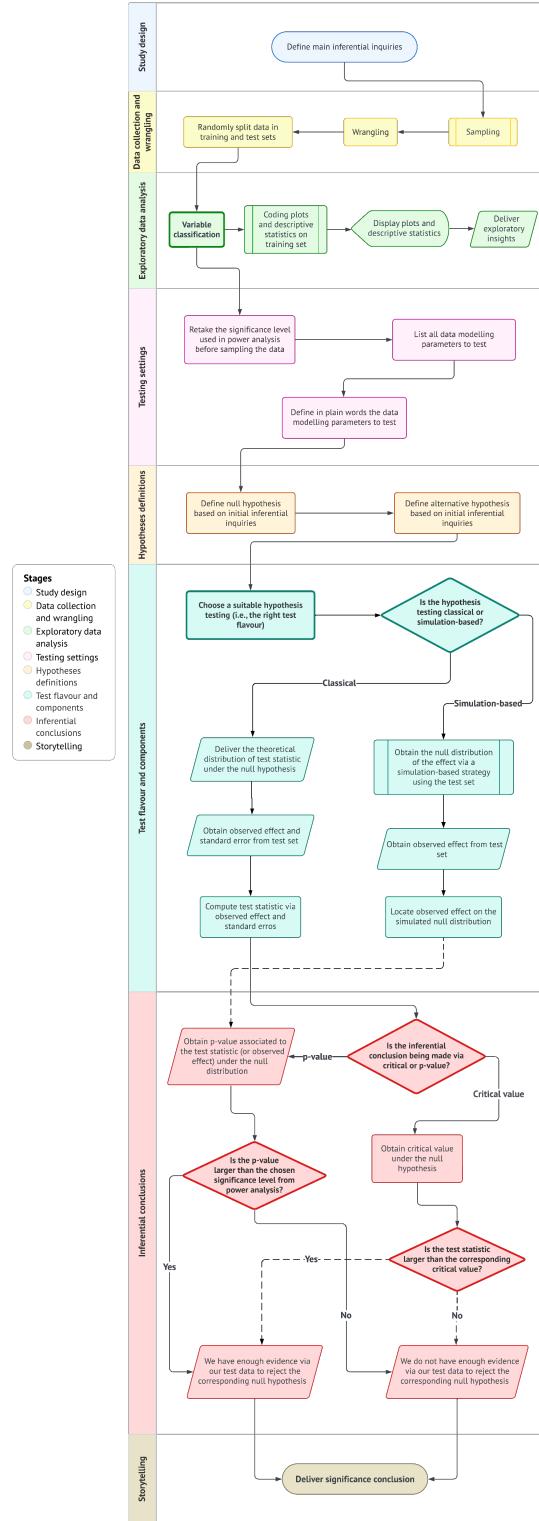


Figure 3: A classical-based hypothesis testing workflow structured in four substages: *general settings*, *hypotheses definitions*, *test flavour and components*, and *inferential conclusions*.

# Chapter 1: 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:

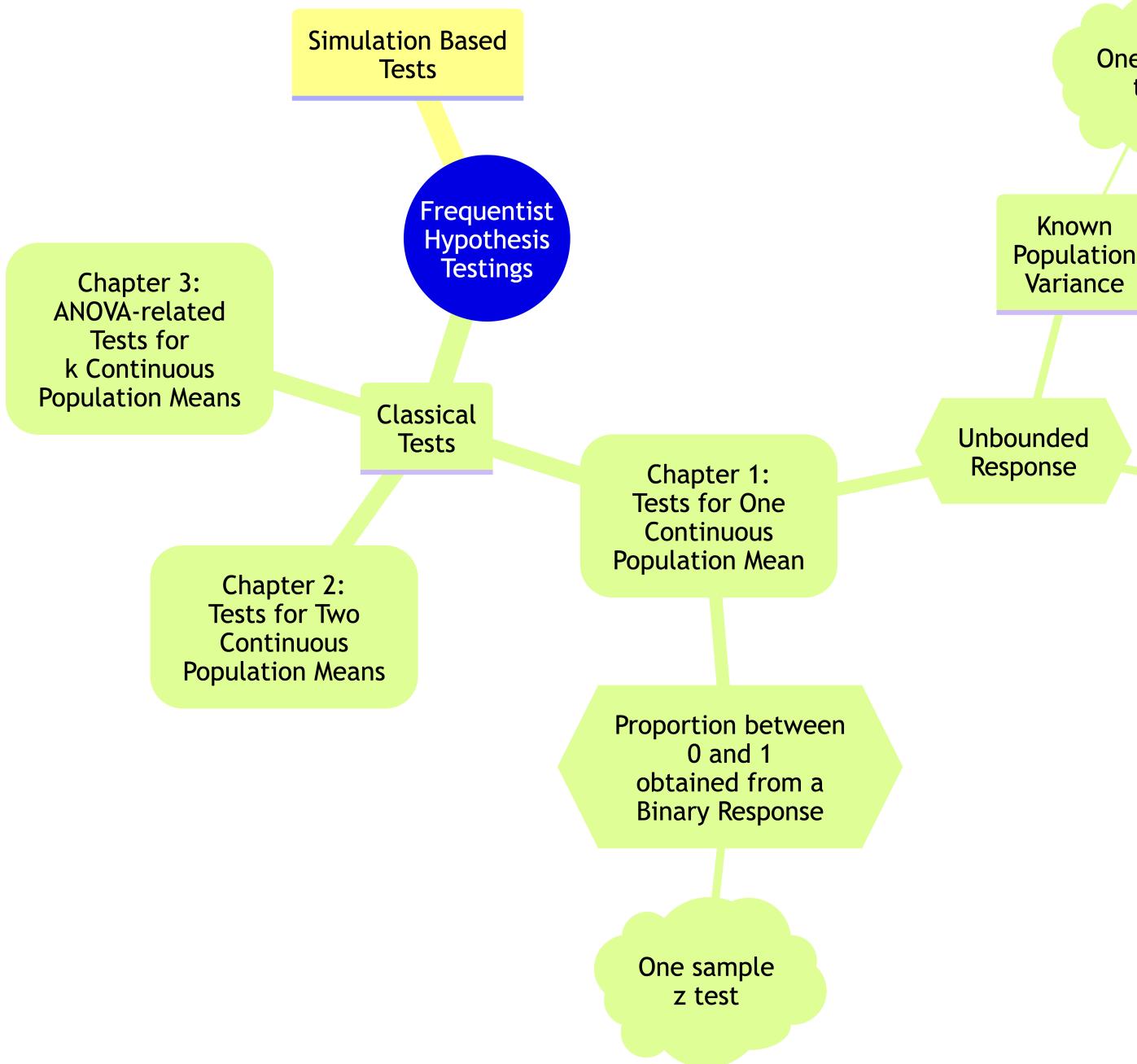


Figure 4: A specific hypothesis testing mind map outlining the techniques explored in this chapter, which are classical tests for one continuous population mean.

## One-sample z-test for the mean

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

The test statistic is:

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

Where: - (  $\bar{x}$  ) is the sample mean

- (  $\mu_0$  ) is the hypothesized population mean

- (  $\sigma$  ) 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.

---

## 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.

The test statistic is:

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

Where: - ( s ) is the sample standard deviation (used instead of (  $\sigma$  ))

This statistic follows a **t-distribution** with ( n - 1 ) degrees of freedom. It is more appropriate for real-world scenarios where we rarely know the true population variance.

---

## 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.

---

# **Chapter 2: 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:

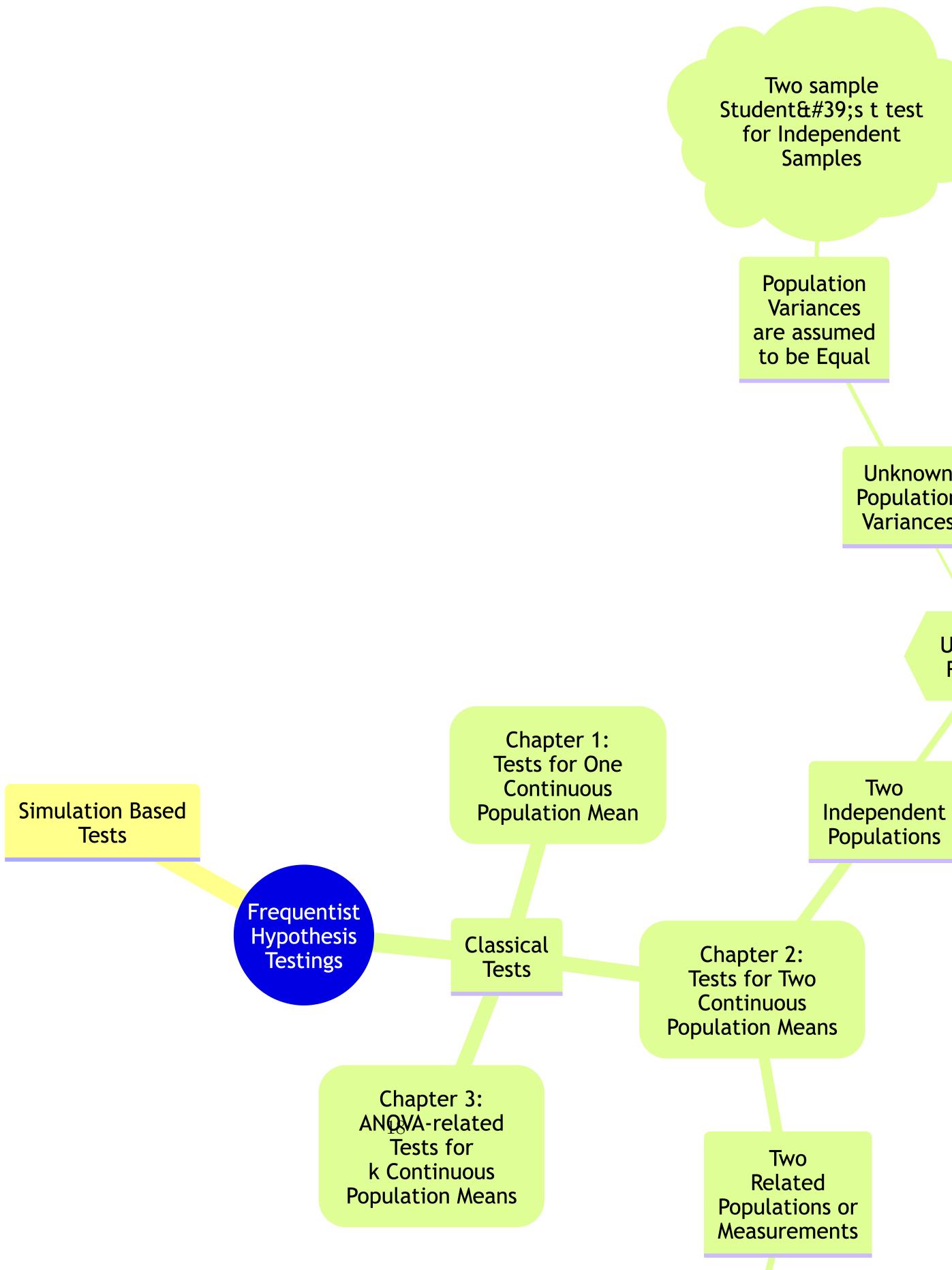
- Independent samples, where the observations in one group are unrelated to those in the other, and
- Paired (or dependent) samples, where observations are naturally matched in some way, such as before-and-after measurements.

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 code examples using real or simulated datasets to help ground the concepts in practice.

## **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. To investigate this, you decide to test the 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.

Let us assume that each population has an unknown average 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:

$$X_1, X_2, \dots, X_n$$

- From SFU (Population 2), we obtain a sample of size  $m$ , denoted as:

$$Y_1, Y_2, \dots, Y_m$$

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

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

In formal terms, we test the hypotheses:

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

To test this, we use the **two-sample t-test**, which compares the sample means and incorporates variability within and between the samples. If we assume equal population variances, the test statistic is:

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

where:

- $\bar{X}$  and  $\bar{Y}$  are the sample means for UBC and SFU, respectively,
- $s_p$  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.

Under the assumption that null hypothesis is correct (i.e.  $\mu_1 = \mu_2$ ) then the test statistic defined above follows a t-distribution with  $n + m - 2$  degrees of freedom (which we denote it by  $T_{n+m-2}$ ). Knowing the distribution of this statistic helps us to compute *p-value* of the test as follows:

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

**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 let us see how to run the two-sample test on some example datasets in R. We will first describe the dataset we want to use, and then show how to run the test in R.

## Data Collection and Wrangling

### Explanatory Data Analysis

### Testing Settings

### Hypothesis Definitions

### Test Flavour and Components

### Inferential Conclusions

### Storytelling

Ignore the following content for now. Still editing

In this example, we start with a dataset that records the time MDS students spend on course website of DSCI 554. You will encounter this dataset again in DSCI 554 when we discuss A/B/n testing. For now, here's what the dataset looks like:

```
ABn_data <- read_csv('data/ABn_data.csv', show_col_types = FALSE)
ABn_data <- ABn_data %>% mutate(Colour = as.factor(Colour), Font = as.factor(Font))
ABn_data
```

```

# A tibble: 72 x 3
  Duration Colour Font
     <dbl> <fct> <fct>
1       90 Red   Small
2       95 Red   Large
3      107 Red   Medium
4       92 Red   Small
5       89 Red   Large
6       92 Red   Medium
7       81 Red   Small
8       92 Red   Large
9       93 Red   Medium
10      80 Blue  Small
# i 62 more rows

```

The columns in the dataset are as follows:

- Font: A factor variable with three levels — small, medium, and large.
- Button Colour: A factor variable with two levels — Red and Blue.
- Duration: A continuous variable representing the duration of each visit, recorded in minutes.

The main statistical question we are asking here is:

**Is there a statistically significant difference in the mean visit duration between websites with red buttons and those with blue buttons?**

This means we are interested in the duration time users spend on website in two different populations: red button design and blue button design. First we do some data selection to create two vector of numbers: one for the visit duration visits in website with red buttons and one for the duration of visits in website with blue buttons. The following code can take care of this. For Red group:

```

duration_red <- ABn_data %>% filter(Colour == 'Red') %>% pull(Duration)
duration_red

```

[1]	90	95	107	92	89	92	81	92	93	83	80	95	98	98	106	74	81	74	85
[20]	88	88	112	104	91	82	78	94	86	78	89	79	86	87	85	89	83		

and for Blue group:

```
duration_blue <- ABn_data %>% filter(Colour == 'Blue') %>% pull(Duration)
```

```
[1] 80 87 100 121 110 119 78 98 122 102 109 105 99 94 123 136 133 132 60  
[20] 104 114 90 118 113 119 122 136 73 114 114 109 131 126 116 136 133
```

## How to run this test in R?

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.

Therefore we can run the test for two different cases.

- Case 1: Under the assumption that variances between two populations are equal:

```
t.test(x = duration_red, y = duration_blue, var.equal = TRUE)
```

### Two Sample t-test

```
data: duration_red and duration_blue  
t = -6.1183, df = 70, p-value = 4.838e-08  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
-28.43489 -14.45400  
sample estimates:  
mean of x mean of y  
89.0000 110.4444
```

- Case 2: Under the assumption that variances between two populations are **NOT** equal:

```
t.test(x = duration_red, y = duration_blue, var.equal = FALSE)
```

### Welch Two Sample t-test

```
data: duration_red and duration_blue
t = -6.1183, df = 50.043, p-value = 1.429e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-28.48424 -14.40465
sample estimates:
mean of x mean of y
89.0000 110.4444
```

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  $\mu_1 - \mu_2$ .
- `sample estimates` gives the sample means for each group.

**Note:** By default the value of `var.equal` is `FALSE`.

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

If the assumption of equal variances is questionable, we instead use Welch's t-test, which adjusts the standard error and degrees of freedom accordingly. The Welch's test statistic is computed as:

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

Under the assumption that null hypothesis is correct, the test statistics defined above still follows a t-distribution but with a different degrees of freedom. The degree of freedom when we do not make equal variance assumption is:

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

Note that this degree of freedom is not necessarily an integer number (could be a real number). 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](#).

## 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:

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

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  $\mu_D$  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:

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

### Example dataset in R

- TBD

### How to run the test in R?

# Chapter 3: ANOVA-related Tests for $k$ Continuous Population Means

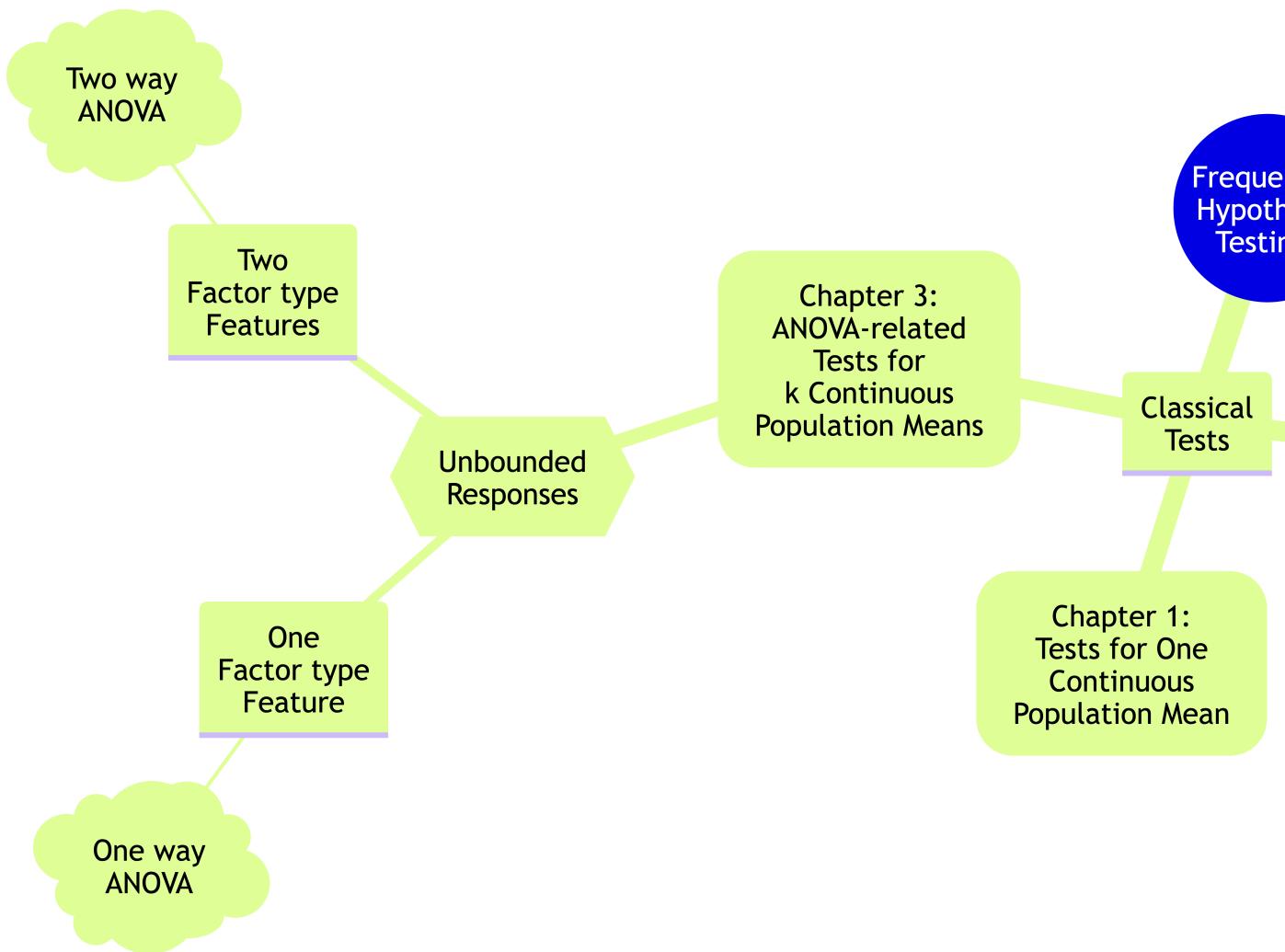


Figure 6: A specific hypothesis testing mind map outlining the techniques explored in this chapter, which include ANOVA-related tests for  $k$  population means.

## **References**