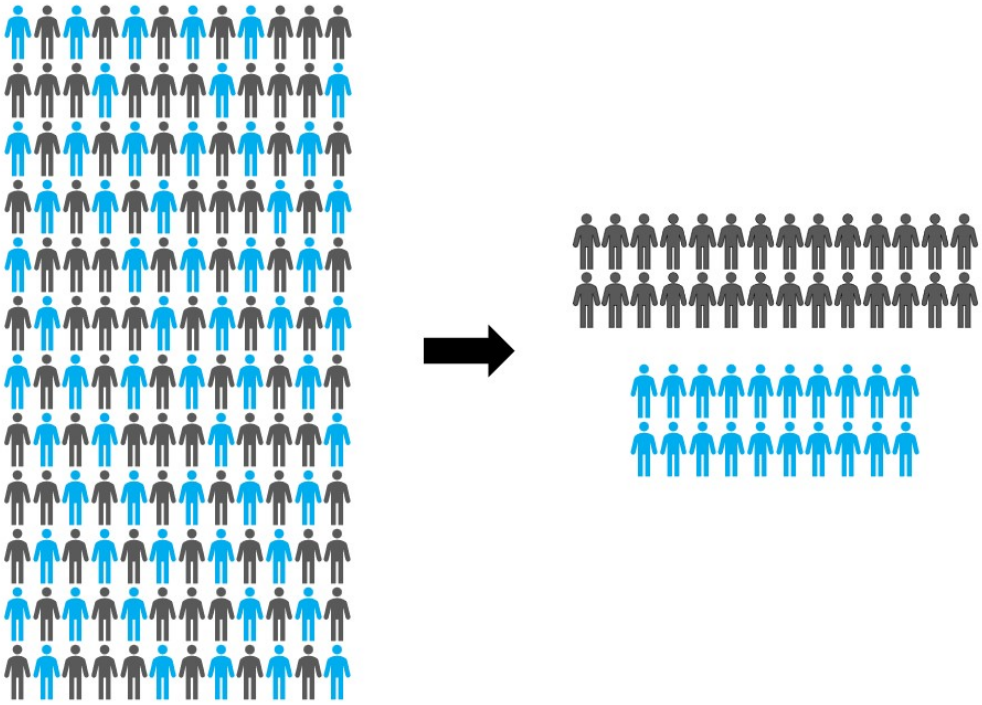
Estimators for the population mean and variance for stratified sampling

The search for unbiased estimators in a suboptimal sample



L. J. Verbeeke

June 26, 2024

TU Delft

Estimators for the population mean and variance for stratified sampling

The search for unbiased estimators in a suboptimal sample

by

L. J. Verbeeke

to obtain the degree of Bachelor of Science
at the Delft University of Technology,
to be defended publicly on July 3, 2024.

Student number:	5650534	
Project duration:	April, 2024 – July, 2024	
Thesis committee:	Dr. A. F. F. Derumigny,	TU Delft, supervisor
	Dr. ir. G. F. Nane,	TU Delft

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

Preface

In this thesis, the goal was to develop estimators such that studies on populations can become more reliable and representative in the case of a bad sample. It marks the end of my Bachelor of Applied Mathematics at the TU Delft. I have enjoyed the feeling of contributing to the interesting field of statistics. Although this project had a more theoretical approach than I thought beforehand, it was great getting more experienced in this kind of work.

First of all, I would like to thank my supervisor, Alexis Derumigny, for all his time and explanations during our meetings. This really helped me in getting a clear view of the project and of what had to be done. He also gave me interesting insights in the world of conducting research. I would also like to thank my family, my girlfriend and my friends for their support throughout the project. Whether it was them showing interest in my project and supporting me on the weekends or studying together at the second floor of EWI, it really helped me in staying motivated and concentrated.

L. J. Verbeeke
Delft, June 26, 2024

Layman's Summary

Researchers can be interested in certain characteristics of a group of people: a population. This population can be divided into different subgroups based on, for instance, gender, age, religion, height or weight. Usually, a smaller sample of the population is created to examine all people. All subgroups are represented in the sample and ideally, the relative size of every subgroup in the sample is equal to the share in which every subgroup appears in the population. If, for instance, one has 70% men and 30% women in the population, these percentages should be more or less the same in the sample to obtain a representative result for all people together. Sometimes, this is not the case, but it can still be necessary to use the acquired data, because it is costly to redo a survey. In this thesis, different ways to use the flawed data are explored, developed and compared in the case of two subgroups. This leads to two suitable methods that take the share of the subgroups into account. However, if one does not know how many people in the population belong to a certain subgroup, these methods become less valuable. Hence, information about a population is very important for the statistical analysis of this population.

Abstract

Dividing a population into subgroups and conducting research on this population including the subgroups comes with a challenge. This stratified sampling relies on information about the share of the subgroups in the population. Sometimes the proportions in the sample are not taken equal to the true proportions of the population. This can be corrected through the use of particular estimators taking these proportions into account. In this thesis, different estimators for the true population mean and variance are defined and examined in terms of bias and variance in the case of two subgroups. Weighing the measurements according to the true proportions creates unbiased estimators for both the mean and the variance. These unbiased estimators are compared with other, biased, estimators, including naive ones in which the influence of different subgroups is not taken into account. The naive estimators are not only biased, they also have a variance of the same order as the unbiased ones. When the true proportions are not available, one can only take a guess. A guess lying close to the true proportions leads to a smaller bias and therefore a better estimator. This underlines the importance of obtaining sufficient knowledge about the population.

Contents

List of Symbols

Symbol	Definition	First appearance
n_1	Number of measurements in group 1	Section 2.1
n	Total number of measurements	Section 2.1
p_1^*	True proportion group 1	Section 2.1
p_1	Guessed proportion group 1	Section 2.1
$\widehat{p}_1$	Fixed proportion group 1: $\frac{n_1}{n}$	Section 2.1
μ^*	True population mean	Section 2.1
σ^{*2}	True population variance	Equation (2.1)
μ_1^*	True mean group 1	Section 2.1
σ_1^{*2}	True variance group 1	Section 2.1
$m_{3,1}^*$	True third moment group 1	Section 2.1
$m_{4,1}^*$	True fourth moment group 1	Section 2.1
Δp	Short notation for $p_1^* - \widehat{p}_1$	Definition 2.1
$\widetilde{\Delta} p$	Short notation for $p_1^* - p_1$	Definition 2.1
$\Delta\mu$	Short notation for $\mu_2^* - \mu_1^*$	Definition 2.1
$\Delta\mu^2$	Short notation for $\mu_2^{*2} - \mu_1^{*2}$	Definition 2.1
$\Delta\sigma^2$	Short notation for $\sigma_2^{*2} - \sigma_1^{*2}$	Definition 2.1
Δn	Short notation for $n_2 - n_1$	Definition 2.1
$\widehat{\mu}$	Estimated population mean	Definition 3.6
$\widehat{\mu}_1$	Estimated mean group 1	Equation (3.3)
$\widehat{\sigma}_1^2$	Estimated variance group 1	Equation (3.4)
$\widetilde{\mu}$	Short notation for $\widehat{p}_1\mu_1^* + \widehat{p}_2\mu_2^*$	Theorem 3.10
$\widetilde{\sigma}^2$	Short notation for $\widehat{p}_1\sigma_1^{*2} + \widehat{p}_2\sigma_2^{*2}$	Theorem 3.10
$\mathcal{O}(u_n)$	Landau's O-symbol	Section 3.2
$\overline{X}$	Naive estimator for μ^*	Equation (2.5)
$\widehat{\theta}_{p_1,p_2}$	Weighted estimator for μ^* with guessed proportions	Equation (2.6)
$\widehat{\theta}_{\text{oracle}}$	Weighted estimator for μ^* with true proportions	Equation (2.7)
$\overline{X}_{\text{random}}$	Estimator $\overline{X}$ with random group sizes	Section 2.3
$\widehat{\text{Var}}[X]$	Naive estimator for σ^{*2}	Equation (3.1)
$\widehat{\gamma}_{p_1,p_2}$	Weighted estimator for σ^{*2} with guessed proportions	Equation (3.2)
$\widehat{\gamma}_{\text{oracle}}$	Weighted estimator for σ^{*2} with true proportions	Equation (3.6)
$\widehat{\psi}_{p_1,p_2}$	Debiased estimator for σ^{*2} with guessed proportions	Equation (3.9)
$\widehat{\psi}_{\text{oracle}}$	Debiased estimator for σ^{*2} with true proportions	Equation (3.8)
$\widehat{\delta}_{p_1,p_2}$	Decomposed estimator for σ^{*2} with guessed proportions	Equation (3.11)
$\widehat{\delta}_{p_1,p_2,\widehat{\theta}_{p_1,p_2}}$	Decomposed estimator for σ^{*2} with $\widehat{\theta}_{p_1,p_2}$ as plug-in estimator	Equation (3.12)
$\widehat{\delta}_{p_1,p_2,\overline{X}}$	Decomposed estimator for σ^{*2} with $\overline{X}$ as plug-in estimator	Equation (3.13)
$\widehat{\delta}_{\text{oracle}}$	Decomposed estimator for σ^{*2} with true proportions	Equation (3.14)

1 Introduction

In social studies, many surveys are held to examine certain attributes of a population. For acquiring data and information about people, a researcher may want to obtain an understanding of different characteristics of the population. With these insights, it is possible to divide the population into different subgroups. Subgroups could be based on, for instance, age, gender, religion, height or weight. This process of selecting a sample is called stratified sampling and the subgroups are called strata. Examining the strata individually can result in new knowledge about their own characteristics, but stratified sampling can also be used to study the complete population. Through the use of stratified sampling, one can make sure every subgroup is correctly represented in the sample, which ensures that the results are representing the complete population.

Acknowledging the existence of the subgroups is important if one wants to assemble a diverse and representative sample of the complete population. A researcher may want to take the influence of these subgroups into account when creating a sample in order to obtain results about the whole population. If, for instance, the population consists of 70% men and 30% women, it is important to know this beforehand. Collecting data from a sample consisting of 50% men and 50% women is then not the best approach to study the population, despite it being a natural choice. This would result in having different group proportions in the sample than in the population and might not give representative results for the population.

Similar to the example given above, there are cases in which the proportions in the sample are not equal or even close to the true proportions of the population. This could be a result of not being able to obtain the true proportions or simply because they were not taken into account. Sometimes the data has already been acquired using the wrong proportions and it is costly to redo the survey. Can the data still be used to obtain reliable results about the population?

When one still wants to draw conclusions from the data, it is important that this is done cautiously. To start with, it must be realised that not necessarily the true proportions are used. This does not mean that the true proportions are known, but the researcher should at least be aware of the fact that their proportions can be inaccurate. A lack of awareness on this could lead to a naive approach to estimating, with major consequences. Wrong conclusions could be drawn, because certain groups can be overrepresented while others are underrepresented or even absent in the sample. Neglecting the inaccurate proportions will lead to a conclusion in which the result of the overrepresented group has too much influence on the main result in comparison with other subgroups. If the outcome of the conducted study determines new legislation, it can even have the discrimination of minorities as a result. The goal of this thesis is to prevent such consequences by developing estimators which can be used even if wrong proportions occur in the sample.

To further illustrate the importance of selecting a proper method to analyse the acquired data, an example is given by Thangavelu and Brunner [4]. They show that a wrong way of sampling and not adjusting for this can cause problems in medical research. In their studies, a new treatment is tested in four different medical centers with all participants equally distributed among the centers. Various standard techniques used in multiple clinical trials show an issue regarding their representativeness for the whole population. This could ultimately lead to a wrong choice of which treatment is best. An advanced analysis of these estimators was done and modifications were made. The techniques used are very different from the ones used in this thesis and are therefore not very applicable here.

Research has already been done on how to select a sample properly in order to have great estimators. For instance, Horvitz and Thompson assumed they could choose a probability for each individual of the population to be part of the sample [2]. They have provided different sampling schemes in relation to the optimal selection probabilities in order for their estimators to work as best as possible. Also Chaudhuri and Stenger [1] and Pfeffermann and Rao [3] studied different sampling schemes. However, in this thesis, we are not able to choose sampling schemes and selection probabilities, but are restricted to using the sample as described before. Only the strata and their shares in the population can be of influence on the way a sample is created.

In this thesis, multiple estimators for the mean and variance of a population with two subgroups are developed. They are defined, modified and compared in terms of bias and variance. For both parameters, an unbiased estimator has been found.

The thesis is structured as follows: Chapter 2 contains the estimation of the population mean for two strata. Various estimators are defined and compared in terms of bias and variance. This is done for both fixed group sizes in the sample and for random group sizes.

Chapter 3 is about estimating the true variance of a population with two subgroups. Different estimators are defined and their biases are derived. The bias of one estimator inspires the creation of a new one, ultimately resulting in an unbiased estimator. Comparing the estimators on their variance leads to a final verdict on the best estimator choice. Because the variance involves quadratic terms, this chapter contains much longer proofs and equations than Chapter 2. Therefore, the chapter is structured such that the definitions and results are given in the beginning and all proofs of the stated theorems are at the end of the chapter. The thesis ends with a conclusion in Chapter 4, followed by some recommendations for future work.

2

Estimating the mean

This chapter starts with setting the framework for this thesis in Section 2.1. In Section 2.2, multiple estimators for the mean are defined and compared for fixed subgroup sizes. Section 2.3 contains a comparison of the previously defined estimators with the case of random group sizes.

2.1. Preliminaries

Let us consider a certain population. This population can be divided into different subgroups based on different characteristics. This is called stratified sampling and the subgroups are called strata. Examples of these characteristics could be gender, age or hair colour. Define an experiment done with this population divided into two strata: 1 and 2. Originally, they have the true proportions p_1^* and p_2^*. Note that $p_1^* + p_2^* = 1$, because the groups form a partition of the population. They are disjoint and their union is the complete set. The true proportions might be available, but this is not necessarily the case.

Fix in advance the numbers of measurements per group: n_1 and n_2. Although the group sizes are fixed, individuals from both strata are still randomly selected as a measurement in the experiment. So n_1 measurements are taken at random in group 1 of the population and n_2 individuals are randomly selected in group 2. Let $n = n_1 + n_2$ be the total number of measurements.

For unknown true group proportions, taking a guess of the proportions comes as a solution when these weights are needed. By p_1 and p_2 the guessed proportions are denoted. These guesses could be based on any available information on the group sizes in the population. Two particular guesses are $\widehat{p}_1 := \frac{n_1}{n}$ and $\widehat{p}_2 := \frac{n_2}{n}$. These are the proportions of measurements in each group in the sample. Because n_1 and n_2 are fixed, these proportions $\widehat{p}_1$ and $\widehat{p}_2$ are also fixed. All guesses are taken such that $p_1 + p_2 = \widehat{p}_1 + \widehat{p}_2 = 1$, because both groups together cover the complete sample.

The measurements of group 1 are denoted by $X_{1,1}, ..., X_{n_1,1}$ and those of group 2 by $X_{1,2}, ..., X_{n_2,2}$. Let $X_{1,1}, ..., X_{n_1,1}$ be independent and identically distributed from an unknown distribution with mean μ_1^* and variance σ_1^{*2} and $X_{1,2}, ..., X_{n_2,2}$ from an unknown distribution with mean μ_2^* and variance σ_2^{*2}. The true mean μ^* of the complete population is: $\mu^* = p_1^*\mu_1^* + p_2^*\mu_2^*$. The true population mean can be of the interest of a researcher or statistician. Multiple ways of estimation of this quantity will therefore be examined in this chapter.

Besides the mean, one could also desire to know the true variance σ^{*2} of the population. Firstly, an expression for this is derived below. Different ways of estimating are presented and compared in Chapter 3.

Let Z denote the group (1 or 2) of a measurement X out of the complete dataset.

Theorem 2.1. *The true variance of the population is:*

$$\sigma^{*2} = p_1^*\sigma_1^{*2} + p_2^*\sigma_2^{*2} + p_1^*\mu_1^{*2} + p_2^*\mu_2^{*2} - \left(p_1^*\mu_1^* + p_2^*\mu_2^*\right)^2. \tag{2.1}$$

Proof of Theorem 2.1. By definition,

$$\begin{aligned}\sigma^{*2} &= \mathbb{E}\left[(X-\mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[\left(X-\mu^*\right)^2\right] \\ &= \mathbb{E}\left[X^2\right]-\mu^{*2} \\ &= p_1^*\mathbb{E}\left[X^2|Z=1\right]+p_2^*\mathbb{E}\left[X^2|Z=2\right]-\left(p_1^*\mu_1^*+p_2^*\mu_2^*\right)^2.\end{aligned}$$

Now, by definition,

$$\mathbb{E}\left[X^2|Z=1\right]=\sigma_1^{*2}+\mathbb{E}[X|Z=1]^2=\sigma_1^{*2}+\mu_1^{*2}, \tag{2.2}$$

and similarly for $Z=2$. □

The first two terms of Equation (2.1) represent the intra-group variance: the group variances averaged according to the true proportions. The other terms represent the inter-group variance: the variance between the different group means. Equation (2.1) could then also be written as:

$$\sigma^{*2}=\mathbb{E}[\mathrm{Var}[X|Z]]+\mathrm{Var}[\mathbb{E}[X|Z]], \tag{2.3}$$

where

$$\begin{aligned}\mathbb{E}[\mathrm{Var}[X|Z]] &= p_1^*\sigma_1^{*2}+p_2^*\sigma_2^{*2}, \\ \mathrm{Var}[\mathbb{E}[X|Z]] &= p_1^*\mu_1^{*2}+p_2^*\mu_2^{*2}-\mu^{*2}.\end{aligned} \tag{2.4}$$

New notation is introduced in the definition below, because multiple proofs and computations contain similar expressions.

Definition 2.1. *Define the following differences:*

$$\begin{aligned}\Delta p &:= p_1^*-\widehat{p}_1=\widehat{p}_2-p_2^* \\ \widetilde{\Delta} p &:= p_1^*-p_1=p_2-p_2^* \\ \Delta\mu &:= \mu_2^*-\mu_1^* \\ \Delta\mu^2 &:= \mu_2^{*2}-\mu_1^{*2} \\ \Delta\sigma^2 &:= \sigma_2^{*2}-\sigma_1^{*2} \\ \Delta n &:= n_2-n_1.\end{aligned}$$

Remember that $p_1^*+p_2^*=p_1+p_2=\widehat{p}_1+\widehat{p}_2=1$. *Note the difference between* $\left(\Delta\mu\right)^2=\left(\mu_2^*-\mu_1^*\right)^2$ *and* $\Delta\mu^2=\mu_2^{*2}-\mu_1^{*2}$.

The true third and fourth moments of the first stratum are denoted by $m_{3,1}^*$ and $m_{4,1}^*$, respectively. These quantities are not estimated in this thesis, but they appear in the variances of some of the developed estimators and are therefore defined here. The true third and fourth moments of the second subgroup are likewise denoted by $m_{3,2}^*$ and $m_{4,2}^*$, respectively.

2.2. Estimating the mean for fixed group sizes

In this section and the next one, all expectations and variances are with respect to μ^* and therefore this is omitted.

Definition 2.2. *Define the following three estimators for* μ^**:* $\overline{X}, \widehat{\theta}_{p_1,p_2}$ *and* $\widehat{\theta}_{\text{oracle}}$*:*

$$\overline{X} := \frac{1}{n}\left(\sum_{i=1}^{n_1}X_{i,1}+\sum_{i=1}^{n_2}X_{i,2}\right), \tag{2.5}$$

$$\widehat{\theta}_{p_1,p_2} := \frac{p_1}{n_1}\sum_{i=1}^{n_1}X_{i,1}+\frac{p_2}{n_2}\sum_{i=1}^{n_2}X_{i,2}, \tag{2.6}$$

$$\widehat{\theta}_{\text{oracle}} := \frac{p_1^*}{n_1}\sum_{i=1}^{n_1}X_{i,1}+\frac{p_2^*}{n_2}\sum_{i=1}^{n_2}X_{i,2}. \tag{2.7}$$

The estimator $\overline{X}$ is a naive one where all measurements are added together and then divided by the total number of measurements. In a situation without stratified sampling, this would be a logical choice. As stated in Theorem 2.2, $\overline{X}$ is biased in general. Therefore, one might consider adjusting the estimator such that the group proportions are taken into account. This gives rise to the estimator $\widehat{\theta}_{p_1,p_2}$. This estimator weighs the measurements with the guessed proportions, when the true proportions are unavailable. However, if these are known, a third estimator $\widehat{\theta}_{\text{oracle}}$ can be used to estimate μ^*. This is a special case of $\widehat{\theta}_{p_1,p_2}$.

Theorem 2.2. *The bias of the estimator $\overline{X}$ is*

$$\mathbb{E}\left[\overline{X}\right] - \mu^* = \Delta p \Delta \mu.$$

The Mean Squared Error (MSE) is

$$MSE\left(\mu^*; \overline{X}\right) = \frac{\widehat{p}_1}{n}\sigma_1^{*2} + \frac{\widehat{p}_2}{n}\sigma_2^{*2} + \left(\Delta p\right)^2\left(\Delta\mu\right)^2.$$

Proof of Theorem 2.2. The bias of $\overline{X}$ is the following:

$$\begin{aligned}
\mathbb{E}\left[\overline{X}\right] - \mu^* &= \mathbb{E}\left[\frac{1}{n}\left(\sum_{i=1}^{n_1} X_{i,1} + \sum_{i=1}^{n_2} X_{i,2}\right)\right] - \mu^* \\
&= \frac{1}{n}\left(\sum_{i=1}^{n_1}\mathbb{E}\left[X_{i,1}\right] + \sum_{i=1}^{n_2}\mathbb{E}\left[X_{i,2}\right]\right) - \mu^* \\
&= \frac{1}{n}\left(n_1\mu_1^* + n_2\mu_2^*\right) - \left(p_1^*\mu_1^* + p_2^*\mu_2^*\right) \\
&= -\Delta p\mu_1^* + \Delta p\mu_2^* \\
&= \Delta p \Delta \mu. \qquad (2.8)
\end{aligned}$$

By definition, the MSE is

$$MSE\left(\mu^*; \overline{X}\right) = \text{Var}\left[\overline{X}\right] + \left(\mathbb{E}\left[\overline{X}\right] - \mu^*\right)^2. \qquad (2.9)$$

Computing the variance gives:

$$\begin{aligned}
\text{Var}\left[\overline{X}\right] &= \text{Var}\left[\frac{1}{n}\left(\sum_{i=1}^{n_1} X_{i,1} + \sum_{i=1}^{n_2} X_{i,2}\right)\right] \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n_1}\text{Var}\left[X_{i,1}\right] + \sum_{i=1}^{n_2}\text{Var}\left[X_{i,2}\right]\right) \\
&= \frac{1}{n^2}\left(n_1\sigma_1^{*2} + n_2\sigma_2^{*2}\right) \\
&= \frac{\widehat{p}_1}{n}\sigma_1^{*2} + \frac{\widehat{p}_2}{n}\sigma_2^{*2}. \qquad (2.10)
\end{aligned}$$

Note that the independence of measurements of different groups is used in the second step. Combining Equation (2.8), Equation (2.9) and Equation (2.10) gives:

$$MSE\left(\mu^*; \overline{X}\right) = \frac{\widehat{p}_1}{n}\sigma_1^{*2} + \frac{\widehat{p}_2}{n}\sigma_2^{*2} + \left(\Delta p\right)^2\left(\Delta\mu\right)^2.$$

□

It follows that $\overline{X}$ is unbiased if $\Delta p = 0$ or $\Delta\mu = 0$, so if $\widehat{p}_1 = p_1^*$ or $\mu_1^* = \mu_2^*$. When $\widehat{p}_1 = p_1^*$, the fixed proportions are equal to the true proportions. This means that the group sizes in the sample exactly match the true proportions. Therefore, all measurements can just be added together and then divided by the total number of measurements, without taking the proportions into account. This is exactly what is done with $\overline{X}$. When $\mu_1^* = \mu_2^*$, the true means of both groups are equal. Although the true group variances can still be very different, the groups are now quite similar and at least have the same expectation. Hence, it is a logical result that $\overline{X}$ is unbiased, because the proportions do not have to be taken into account if the groups are already

equally distributed in terms of mean. In all other cases, $\overline{X}$ is biased, because then $\Delta p \Delta \mu \neq 0$. This raises the question whether it would be better to weigh the measurements according to their proportions. This is done with the $\widehat{\theta}_{p_1,p_2}$ estimator. Guesses of the true proportions are used to weigh the measurements of both groups. The bias and MSE of this estimator are stated below.

Theorem 2.3. *The bias of the estimator $\widehat{\theta}_{p_1,p_2}$ is*

$$\mathbb{E}\left[\widehat{\theta}_{p_1,p_2}\right] - \mu^* = \widetilde{\Delta} p \Delta \mu.$$

The MSE of $\widehat{\theta}_{p_1,p_2}$ is the following:

$$MSE\left(\mu^*; \widehat{\theta}_{p_1,p_2}\right) = \frac{p_1^2}{n_1}\sigma_1^{*2} + \frac{p_2^2}{n_2}\sigma_2^{*2} + \left(\widetilde{\Delta} p\right)^2 \left(\Delta \mu\right)^2. \tag{2.11}$$

Proof of Theorem 2.3. The bias is the following:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\theta}_{p_1,p_2}\right] - \mu^* &= \mathbb{E}\left[\frac{p_1}{n_1}\sum_{i=1}^{n_1} X_{i,1} + \frac{p_2}{n_2}\sum_{i=1}^{n_2} X_{i,2}\right] - \mu^* \\
&= \frac{p_1}{n_1}\sum_{i=1}^{n_1}\mathbb{E}\left[X_{i,1}\right] + \frac{p_2}{n_2}\sum_{i=1}^{n_2}\mathbb{E}\left[X_{i,2}\right] - \left(p_1^*\mu_1^* + p_2^*\mu_2^*\right) \\
&= \left(p_1 - p_1^*\right)\mu_1^* + \left(p_2 - p_2^*\right)\mu_2^*.
\end{aligned}$$

With the differences defined in Definition 2.1, we can rewrite the bias as:

$$\mathbb{E}\left[\widehat{\theta}_{p_1,p_2}\right] - \mu^* = -\widetilde{\Delta} p \mu_1^* + \widetilde{\Delta} p \mu_2^* = \widetilde{\Delta} p \Delta \mu.$$

The MSE again consists of the variance and the square of the bias. The variance is computed below.

$$\begin{aligned}
\mathrm{Var}\left[\widehat{\theta}_{p_1,p_2}\right] &= \mathrm{Var}\left[\frac{p_1}{n_1}\sum_{i=1}^{n_1} X_{i,1} + \frac{p_2}{n_2}\sum_{i=1}^{n_2} X_{i,2}\right] \\
&= \frac{p_1^2}{n_1^2}\sum_{i=1}^{n_1}\mathrm{Var}\left[X_{i,1}\right] + \frac{p_2^2}{n_2^2}\sum_{i=1}^{n_2}\mathrm{Var}\left[X_{i,2}\right] \\
&= \frac{p_1^2}{n_1}\sigma_1^{*2} + \frac{p_2^2}{n_2}\sigma_2^{*2}.
\end{aligned}$$

Adding the variance and the square of the bias together leads to the MSE:

$$MSE\left(\mu^*; \widehat{\theta}_{p_1,p_2}\right) = \frac{p_1^2}{n_1}\sigma_1^{*2} + \frac{p_2^2}{n_2}\sigma_2^{*2} + \left(\widetilde{\Delta} p\right)^2 \left(\Delta \mu\right)^2.$$

□

From Theorem 2.3, it can be concluded that $\widehat{\theta}_{p_1,p_2}$ is biased in general. The bias, denoted by $\widetilde{\Delta} p \Delta \mu$, is of similar form as the bias of $\overline{X}$. Only the guessed proportions are different. The consequence is that $\widehat{\theta}_{p_1,p_2}$ is unbiased in similar cases as $\overline{X}$: when the guessed proportions match the true ones ($p_1 = p_1^*$) or when both groups have equal means ($\mu_1^* = \mu_2^*$).

When the guesses p_1 and p_2 are taken equal to the fixed proportions $\widehat{p}_1$ and $\widehat{p}_2$, the coefficients of $\widehat{\theta}_{p_1,p_2}$ become equal to the ones of $\overline{X}$: $\frac{\widehat{p}_1}{n_1} = \frac{n_1}{n n_1} = \frac{1}{n}$, and similarly for $\frac{\widehat{p}_2}{n_2}$. This means $\widehat{\theta}_{\widehat{p}_1,\widehat{p}_2} = \overline{X}$. So there is no difference in $\overline{X}$ and $\widehat{\theta}_{p_1,p_2}$ as an estimator for μ^* when the guesses p_1, p_2 are taken as the fixed proportions $\widehat{p}_1, \widehat{p}_2$. It can be concluded that $\widehat{\theta}_{p_1,p_2}$ is only useful when one has a reason to take a different guess than the fixed one for the proportions. Otherwise, the estimator $\overline{X}$ suffices, as it is more natural to interpret.

It is possible that the true proportions are known to the researcher. Then it is natural to use them instead of guesses. This is the estimator $\widehat{\theta}_{\text{oracle}}$, given in Equation (2.7). The calculation of the MSE of $\widehat{\theta}_{\text{oracle}}$ is very similar to the one of $\widehat{\theta}_{p_1,p_2}$, only now all p_1's and p_2's are replaced by their true values p_1^* and p_2^*. This creates an unbiased estimator, which makes the second term of Equation (2.11) vanish, leading to the corollary below.

Corollary 2.4. *The estimator $\widehat{\theta}_{\text{oracle}}$ is unbiased and its MSE is*

$$MSE\left(\mu^*; \widehat{\theta}_{\text{oracle}}\right) = \frac{p_1^{*2}}{n_1}\sigma_1^{*2} + \frac{p_2^{*2}}{n_2}\sigma_2^{*2}. \tag{2.12}$$

Because $\widehat{\theta}_{\text{oracle}}$ is always unbiased, it is preferable to use this estimator over $\overline{X}$ and $\widehat{\theta}_{p_1,p_2}$.

2.3. Estimating the mean for random group sizes

In this section, the group sizes n_1 and n_2 are random variables. Still, $n = n_1 + n_2$ is of fixed size, so $n_2 = n - n_1$. Variable n_1 behaves like a binomial distribution with n datapoints and a probability of p_1^*. Let $\overline{X}_{\text{random}}$ be the same estimator as $\overline{X}$, defined in Equation (2.5), but now for random group sizes n_1, n_2.

Theorem 2.5. *The estimator $\overline{X}_{\text{random}}$ is unbiased for μ^*. The MSE of $\overline{X}_{\text{random}}$ is*

$$MSE\left(\mu^*; \overline{X}_{\text{random}}\right) = \frac{p_1^* \sigma_1^{*2} + p_2^* \sigma_2^{*2} + p_1^* \mu_1^{*2} + p_2^* \mu_2^{*2} - \mu^{*2}}{n}. \tag{2.13}$$

Proof of Theorem 2.5. The expected value of this estimator $\overline{X}_{\text{random}}$ is: $\mathbb{E}\left[\overline{X}_{\text{random}}\right] = \mathbb{E}[X] = \mu^*$, with X some measurement of the dataset. Therefore, $\overline{X}_{\text{random}}$ is an unbiased estimator for μ^*. The MSE of $\overline{X}_{\text{random}}$ is then equal to its variance, which is just the true population variance σ^{*2} divided by n. □

The case of fixed n_1, n_2 with known true proportions p_1^*, p_2^* and estimator $\widehat{\theta}_{\text{oracle}}$ can be compared with the case of random n_1, n_2 with the estimator $\overline{X}_{\text{random}}$ to decide which estimator to use.

Theorem 2.6. *The difference between the MSEs of $\overline{X}_{\text{random}}$ and $\widehat{\theta}_{\text{oracle}}$ is*

$$n\left(MSE\left(\mu^*; \overline{X}_{\text{random}}\right) - MSE\left(\mu^*; \widehat{\theta}_{\text{oracle}}\right)\right) = \text{Var}\left[\mathbb{E}[X|Z]\right] - \frac{\Delta p}{\widehat{p}_1 \widehat{p}_2}\left(p_1^* \widehat{p}_2 \Delta\sigma^2 + \Delta p \sigma_2^{*2}\right). \tag{2.14}$$

Proof of Theorem 2.6. The two first terms of Equation (2.12) and Equation (2.13), which look similar, are subtracted and compared below:

$$\begin{aligned}
\frac{p_1^{*2}}{\widehat{p}_1}\sigma_1^{*2} + \frac{p_2^{*2}}{\widehat{p}_2}\sigma_2^{*2} - (p_1^*\sigma_1^{*2} + p_2^*\sigma_2^{*2}) &= \left(\frac{p_1^*}{\widehat{p}_1} - 1\right) p_1^*\sigma_1^{*2} + \left(\frac{p_2^*}{\widehat{p}_2} - 1\right) p_2^*\sigma_2^{*2} \\
&= \left(p_1^* - \widehat{p}_1\right)\frac{p_1^*}{\widehat{p}_1}\sigma_1^{*2} + \left(p_2^* - \widehat{p}_2\right)\frac{p_2^*}{\widehat{p}_2}\sigma_2^{*2} \\
&= \Delta p\left(\frac{p_1^*}{\widehat{p}_1}\sigma_1^{*2} - \frac{p_2^*}{\widehat{p}_2}\sigma_2^{*2}\right) \\
&= \frac{\Delta p}{\widehat{p}_1\widehat{p}_2}\left(p_1^*\widehat{p}_2\sigma_1^{*2} - p_2^*\widehat{p}_1\sigma_2^{*2}\right).
\end{aligned}$$

Rewriting proportions of group 2 into the ones of group 1 with the use of $p_1^* + p_2^* = 1$, leads to:

$$\begin{aligned}
\frac{p_1^{*2}}{\widehat{p}_1}\sigma_1^{*2} + \frac{p_2^{*2}}{\widehat{p}_2}\sigma_2^{*2} - (p_1^*\sigma_1^{*2} + p_2^*\sigma_2^{*2}) &= \frac{\Delta p}{\widehat{p}_1\widehat{p}_2}\left(p_1^*(1-\widehat{p}_1)\sigma_1^{*2} - (1-p_1^*)\widehat{p}_1\sigma_2^{*2}\right) \\
&= \frac{\Delta p}{\widehat{p}_1\widehat{p}_2}\left(-p_1^*\widehat{p}_1\Delta\sigma^2 + p_1^*\sigma_1^{*2} - \widehat{p}_1\sigma_2^{*2}\right) \\
&= \frac{\Delta p}{\widehat{p}_1\widehat{p}_2}\left(-p_1^*\widehat{p}_1\Delta\sigma^2 + p_1^*\sigma_1^{*2} - (p_1^* - \Delta p)\sigma_2^{*2}\right) \\
&= \frac{\Delta p}{\widehat{p}_1\widehat{p}_2}\left(-p_1^*\widehat{p}_1\Delta\sigma^2 + p_1^*\Delta\sigma^2 + \Delta p\sigma_2^{*2}\right) \\
&= \frac{\Delta p}{\widehat{p}_1\widehat{p}_2}\left(p_1^*\widehat{p}_2\Delta\sigma^2 + \Delta p\sigma_2^{*2}\right).
\end{aligned}$$

Note that $p_1^*\mu_1^{*2} + p_2^*\mu_2^{*2} - \mu^{*2} = \text{Var}[\mathbb{E}[X|Z]]$, as in Equation (2.4). The MSEs of both estimators can now be compared:

$$\begin{aligned}
n\left(MSE\left(\mu^*; \overline{X}_{\text{random}}\right) - MSE\left(\mu^*; \widehat{\theta}_{\text{oracle}}\right)\right) &= n\left(\text{Var}\left[\overline{X}_{\text{random}}\right] - \text{Var}\left[\widehat{\theta}_{\text{oracle}}\right]\right) \\
&= \text{Var}\left[\mathbb{E}[X|Z]\right] - \frac{\Delta p}{\widehat{p}_1\widehat{p}_2}\left(p_1^*\widehat{p}_2\Delta\sigma^2 + \Delta p\sigma_2^{*2}\right).
\end{aligned}$$

□

When Δp is equal to zero, this difference will be non-negative, because $\text{Var}[\mathbb{E}[X|Z]]$ is non-negative by definition. It is therefore better to choose $\widehat{\theta}_{\text{oracle}}$ as an estimator. Also when the group means are very different (so $\text{Var}[\mathbb{E}[X|Z]]$ is large), the MSE of $\overline{X}_{\text{random}}$ will be greater than the MSE of $\widehat{\theta}_{\text{oracle}}$. Only when Δp is relatively large, $\overline{X}_{\text{random}}$ is the better choice to estimate μ^*. This means that it could be better to let go the stratified sampling and instead sample completely at random when one can not take proper guesses of the proportions.

3

Estimating the variance

In this chapter, the true variance of the experiment discussed in Section 2.1 is estimated in different ways. Firstly, various estimators are defined and the main results are stated in Section 3.1 and Section 3.2. In the remainder of the chapter, the stated theorems about the biases and variances of these estimators are proved. SVarious lemmas are derived in Section 3.3 to be used in the proofs.

Note that the group sizes n_1, n_2 are fixed in this chapter. All expectations and variances in this chapter are with respect to the true variance σ^{*2}.

3.1. Estimators and their bias

In order to estimate σ^{*2} well with two subgroups, a natural first step might be estimating the variance as if all measurements were coming from just one group. Then all squared differences of the measurements and the mean are added together and afterwards divided by $n-1$. This approach is done initially, resulting in the somewhat naive estimator defined below.

Definition 3.1. *A naive estimator for σ^{*2} is the following one:*

$$\widehat{\mathrm{Var}}[X] := \frac{1}{n-1}\left(\sum_{i=1}^{n_1}\left(X_{i,1}-\overline{X}\right)^2 + \sum_{i=1}^{n_2}\left(X_{i,2}-\overline{X}\right)^2\right). \tag{3.1}$$

As stated in the theorem below, the estimator $\widehat{\mathrm{Var}}[X]$ is in general a biased estimator. The proof of this theorem is given in Section 3.4.

Theorem 3.1. *The bias of $\widehat{\mathrm{Var}}[X]$ is*

$$\mathbb{E}\left[\widehat{\mathrm{Var}}[X]\right] - \sigma^{*2} = \Delta p\left(\Delta\sigma^2 + \Delta\mu^2\right) + \epsilon_n,$$

with

$$\epsilon_n := \frac{1}{n-1}\left(\widehat{p}_1\mu_1^{*2} + \widehat{p}_2\mu_2^{*2} - \left(\widehat{p}_1\mu_1^* + \widehat{p}_2\mu_2^*\right)^2 + 2\widehat{p}_1\widehat{p}_2\Delta\sigma^2\right).$$

Remark that the ϵ_n-term is equal to zero in the asymptotic regime $n \to +\infty$. Therefore, multiple cases exist for which the naive estimator becomes unbiased in under this asymptotic regime. This is similar to the estimation of the mean studied in the previous chapter: we also studied several similar cases in which the naive estimator $\overline{X}$ is unbiased. Firstly, when the fixed proportions $\widehat{p}_1, \widehat{p}_2$ are equal to the true proportions p_1^*, p_2^*, the difference Δp becomes zero. Only the ϵ_n term remains, letting the bias tend to zero. This result does not come as a surprise, because in this case both groups are perfectly represented in the sample. Secondly, when the groups have equal means and variances, the bias tends to zero. Indeed, $\Delta\sigma^2 + \Delta\mu^2$ becomes zero, leaving ϵ_n as the only term of the bias. This is also a natural result, because having similar distributions in both groups means weighing the groups with their proportions becomes unnecessary. Therefore, the naive estimator performs well enough. Lastly, if $\sigma_2^{*2} = \mu_1^{*2}$ and $\sigma_1^{*2} = \mu_2^{*2}$, the bias also tends to zero. In this rare case, $\Delta\sigma^2 + \Delta\mu^2$ becomes zero. Thus, using the naive estimator is acceptable. However, it can be hard to predict in advance whether these equations hold as it is an odd case.

3.1.1. Developing an unbiased estimator

One can be interested in a generally unbiased estimator. Because the naive estimator is generally biased, other approaches must be tried. In search of such an improvement, an estimator looking like the expression for the true variance can be examined. Consider an estimator where all unknown, true parameters in the expression of the true variance (given in Equation (2.1)) are replaced by guessed or estimated parameters. It is defined below.

Definition 3.2. *The estimator $\widehat{\gamma}_{p_1,p_2}$ for σ^{*2} is defined as:*

$$\widehat{\gamma}_{p_1,p_2} := p_1\widehat{\sigma}_1^2 + p_2\widehat{\sigma}_2^2 + p_1\widehat{\mu}_1^2 + p_2\widehat{\mu}_2^2 - \widehat{\theta}_{p_1,p_2}^2, \tag{3.2}$$

with plug-in estimators

$$\widehat{\mu}_1 := \frac{1}{n_1}\sum_{i=1}^{n_1} X_{i,1}, \tag{3.3}$$

$$\widehat{\sigma}_1^2 := \frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(X_{i,1} - \frac{1}{n_1}\sum_{j=1}^{n_1} X_{j,1}\right)^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(X_{i,1} - \widehat{\mu}_1\right)^2, \tag{3.4}$$

and $\widehat{\mu}_2, \widehat{\sigma}_2$ similarly. The estimator $\widehat{\theta}_{p_1,p_2}$ is as defined in Equation (2.6).

The plug-in estimators $\widehat{\mu}_1$ and $\widehat{\sigma}_1^2$ are unbiased for μ_1^* and σ_1^{*2}, respectively. This is proved in Lemma 3.15. Studied in the previous chapter, $\widehat{\theta}_{p_1,p_2}$ is biased in general (Theorem 2.3). Only when the true proportions p_1^* and p_2^* are known (or both groups have equal means), μ^* is estimated without a bias. Then $\widehat{\theta}_{p_1,p_2}$ becomes the estimator $\widehat{\theta}_{\text{oracle}}$, defined in Equation (2.7). Theorem 3.2 is the result of adding the biases of these plug-in estimators, multiplied by their coefficients stated in Equation (3.2). It is proved in Section 3.5.

Theorem 3.2. *The bias of the estimator $\widehat{\gamma}_{p_1,p_2}$ for the true variance σ^{*2} is*

$$\mathbb{E}\left[\widehat{\gamma}_{p_1,p_2}\right] - \sigma^{*2} = \widetilde{\Delta}p\left(\Delta\sigma^2 + \Delta\mu^2\right) + \mu^{*2} - \left(p_1\mu_1^* + p_2\mu_2^*\right)^2 + \frac{p_1p_2}{n_1}\sigma_1^{*2} + \frac{p_1p_2}{n_2}\sigma_2^{*2}. \tag{3.5}$$

Remark that the last two terms tend to zero for $n \to +\infty$. The other terms of Equation (3.5) are only zero when the guessed proportions are exactly equal to the true proportions, except for some rare cases. Thus, in general, $\widehat{\gamma}_{p_1,p_2}$ is also not an unbiased estimator. Replacing the guessed proportions p_1, p_2 by the true proportions p_1^*, p_2^* could potentially solve this problem. Remember that this is not always possible, because the true proportions could be unavailable. When it is possible, it results in the estimator $\widehat{\gamma}_{\text{oracle}}$ defined below.

Definition 3.3. *The estimator $\widehat{\gamma}_{\text{oracle}}$ for the true variance is defined as:*

$$\widehat{\gamma}_{\text{oracle}} := p_1^*\widehat{\sigma}_1^2 + p_2^*\widehat{\sigma}_2^2 + p_1^*\widehat{\mu}_1^2 + p_2^*\widehat{\mu}_2^2 - \widehat{\theta}_{\text{oracle}}^2. \tag{3.6}$$

The corollary below is a result of Theorem 3.2.

Corollary 3.3. *The estimator $\widehat{\gamma}_{\text{oracle}}$ has the following bias for σ^{*2}:*

$$\mathbb{E}\left[\widehat{\gamma}_{\text{oracle}}\right] - \sigma^{*2} = p_1^*p_2^*\left(\frac{\sigma_1^{*2}}{n_1} + \frac{\sigma_2^{*2}}{n_2}\right). \tag{3.7}$$

Proof of Corollary 3.3. When the true proportions are used, $\widetilde{\Delta}p = 0$. Furthermore, remember that $\mu^* = p_1^*\mu_1^* + p_2^*\mu_2^*$. Therefore, Equation (3.21) becomes:

$$\mathbb{E}\left[\widehat{\gamma}_{\text{oracle}}\right] - \sigma^{*2} = \frac{p_1^*p_2^*}{n_1}\sigma_1^{*2} + \frac{p_1^*p_2^*}{n_2}\sigma_2^{*2} = p_1^*p_2^*\left(\frac{\sigma_1^{*2}}{n_1} + \frac{\sigma_2^{*2}}{n_2}\right).$$

□

Having the opportunity to use the true proportions results in a bias that is usually smaller in absolute value than before, except for some rare cases. The estimator $\widehat{\gamma}_{\text{oracle}}$ is asymptotically unbiased for $n \to +\infty$. This is an improvement in comparison with the estimator $\widehat{\gamma}_{p_1,p_2}$. Although $\widehat{\gamma}_{\text{oracle}}$ is asymptotically unbiased,

it is still biased for finite number of measurements. Remark that this estimator is the expression of the true variance where the true parameters have been replaced with naive plug-estimators for the group means and variances. Note that some of these plug-in estimators are biased for what they estimate in the current expression. Although the simple estimator $\widehat{\mu}_1$ is unbiased for μ_1^*, the estimator $\widehat{\mu}_1^2$ is biased for μ_1^{*2}. Likewise, $\widehat{\theta}^2_{\text{oracle}}$ is a biased estimator for μ^{*2}, while $\widehat{\theta}_{\text{oracle}}$ is unbiased for μ^*. These statements are results of Jensen's inequality.

A natural follow-up questions is how to construct unbiased estimators for these quantities. In the following corollary, we present a debiased estimator for μ_1^{*2}, as a consequence of Lemma 3.15.

Corollary 3.4. *The estimator $\widehat{\mu}_1^2 - \frac{\widehat{\sigma}_1^2}{n_1}$ is unbiased for μ_1^{*2}.*

Proof of Corollary 3.4. Using Lemma 3.15, the expectation of the estimator $\widehat{\mu}_1^2 - \frac{\widehat{\sigma}_1^2}{n_1}$ is:

$$\mathbb{E}\left[\widehat{\mu}_1^2 - \frac{\widehat{\sigma}_1^2}{n_1}\right] = \mathbb{E}\left[\widehat{\mu}_1^2\right] - \frac{1}{n_1}\mathbb{E}\left[\widehat{\sigma}_1^2\right] = \frac{\sigma_1^{*2}}{n_1} + \mu_1^{*2} - \frac{\sigma_1^{*2}}{n_1} = \mu_1^{*2}.$$

Therefore, it is an unbiased estimator for μ_1^{*2}. □

Using similar techniques, we now present a debiased estimator for μ^{*2}, as a consequence of Lemma 3.22.

Corollary 3.5. *The estimator $\widehat{\theta}^2_{\text{oracle}} - \frac{p_1^{*2}}{n_1}\widehat{\sigma}_1^2 - \frac{p_2^{*2}}{n_2}\widehat{\sigma}_2^2$ is unbiased for μ^{*2}.*

Proof of Corollary 3.5. With the use of Lemma 3.15 and Lemma 3.22, the expectation of this new estimator is the following:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\theta}^2_{\text{oracle}} - \frac{p_1^{*2}}{n_1}\widehat{\sigma}_1^2 - \frac{p_2^{*2}}{n_2}\widehat{\sigma}_2^2\right] &= \mathbb{E}\left[\widehat{\theta}^2_{\text{oracle}}\right] - \frac{p_1^{*2}}{n_1}\mathbb{E}\left[\widehat{\sigma}_1^2\right] - \frac{p_2^{*2}}{n_2}\mathbb{E}\left[\widehat{\sigma}_2^2\right] \\
&= \frac{p_1^{*2}}{n_1}\sigma_1^{*2} + \frac{p_2^{*2}}{n_2}\sigma_2^{*2} + \left(p_1^*\mu_1^* + p_2^*\mu_2^*\right)^2 - \frac{p_1^{*2}}{n_1}\sigma_1^{*2} - \frac{p_2^{*2}}{n_2}\sigma_2^{*2} \\
&= \mu^{*2}.
\end{aligned}$$

Therefore, $\widehat{\theta}^2_{\text{oracle}} - \frac{p_1^{*2}}{n_1}\widehat{\sigma}_1^2 - \frac{p_2^{*2}}{n_2}\widehat{\sigma}_2^2$ is an unbiased estimator for μ^{*2}. □

With these new plug-in estimators as results from Corollary 3.4 and Corollary 3.5, the estimator $\widehat{\gamma}_{\text{oracle}}$ can be modified such that it is also unbiased for a finite number of measurements. This leads to the estimator $\widehat{\psi}_{\text{oracle}}$ as defined below.

Definition 3.4. *The modified version of estimator $\widehat{\gamma}_{\text{oracle}}$ for σ^{*2} is defined as:*

$$\widehat{\psi}_{\text{oracle}} := p_1^*\widehat{\sigma}_1^2 + p_2^*\widehat{\sigma}_2^2 + p_1^*\widehat{\mu}_1^2 + p_2^*\widehat{\mu}_2^2 - \widehat{\theta}^2_{\text{oracle}} - p_1^*p_2^*\left(\frac{\widehat{\sigma}_1^2}{n_1} + \frac{\widehat{\sigma}_2^2}{n_2}\right). \tag{3.8}$$

Before examining the bias of $\widehat{\psi}_{\text{oracle}}$, a more general case is considered. The true proportions p_1^* and p_2^* are used in the estimator $\widehat{\psi}_{\text{oracle}}$. However, these are not always known to a researcher. To be able to still estimate the variance well, the guessed proportions p_1, p_2 are used. These can be based on the available information one has at their disposal. It leads to $\widehat{\psi}_{p_1,p_2}$ as defined in the definition below, which can be seen as a modification of the estimator $\widehat{\gamma}_{p_1,p_2}$.

Definition 3.5. *A new estimator for σ^{*2} is introduced:*

$$\widehat{\psi}_{p_1,p_2} := p_1\widehat{\sigma}_1^2 + p_2\widehat{\sigma}_2^2 + p_1\widehat{\mu}_1^2 + p_2\widehat{\mu}_2^2 - \widehat{\theta}^2_{p_1,p_2} - p_1p_2\left(\frac{\widehat{\sigma}_1^2}{n_1} + \frac{\widehat{\sigma}_2^2}{n_2}\right). \tag{3.9}$$

Remark that $\widehat{\psi}_{p_1,p_2}$ is equal to $\widehat{\psi}_{\text{oracle}}$, but with the guessed proportions instead of the true ones.

This estimator is biased in general, as stated in the theorem below. It is proved in Section 3.9.

Theorem 3.6. *The bias of $\widehat{\psi}_{p_1,p_2}$ is*

$$\mathbb{E}\left[\widehat{\psi}_{p_1,p_2}\right] - \sigma^{*2} = \widetilde{\Delta}p\left(\Delta\sigma^2 + \Delta\mu^2\right) + \mu^{*2} - \left(p_1\mu_1^* + p_2\mu_2^*\right)^2. \tag{3.10}$$

It is clear that the terms depending on the number of measurements from Equation (3.5) vanished, which was exactly the goal of the new estimator $\widehat{\psi}_{p_1,p_2}$. However, it is still a biased estimator. Remark that these two new estimators are asymptotically the same as the uncorrected ones. This is therefore only a finite-sample correction.

Despite $\widehat{\psi}_{p_1,p_2}$ still being biased, in many cases it does have a smaller squared bias than the previously defined estimator $\widehat{\gamma}_{p_1,p_2}$, especially when the true group variances are big. Having enough information available such that the true proportions are known, is an advantage and is used in the estimator $\widehat{\psi}_{\text{oracle}}$. The guessed proportions of Equation (3.10) are replaced by the true ones, resulting in the corollary below.

Corollary 3.7. *The estimator $\widehat{\psi}_{\text{oracle}}$ is unbiased for σ^{*2}.*

Proof of Corollary 3.7. Substituting the true proportions for the guessed ones in Equation (3.10), results in $\widetilde{\Delta}p$ being 0. Also, remark that $\mu^* = p_1^*\mu_1^* + p_2^*\mu_2^*$. Hence, $\mathbb{E}\left[\widehat{\psi}_{\text{oracle}}\right] - \sigma^{*2} = 0$. □

The goal of finding an unbiased estimator for σ^{*2} is reached. Note that $\widehat{\psi}_{\text{oracle}}$ can only be used if the true proportions are known to the researcher. Otherwise, the estimator $\widehat{\psi}_{p_1,p_2}$ is a possible option for estimation, although it is not unbiased in general. The importance of acquiring data and information about the population is emphasised by this result, because a better approximation of the proportions leads to a smaller bias.

An unbiased estimator is not always the best one to use. Similar as before in Chapter 2, the variances of the estimators are also computed. This is done in Section 3.2, where the estimator variances are also compared. The next subsection is about a different approach to estimating the variance.

3.1.2. A different approach to estimating

Remember the decomposition of the variance that was given in Equation (2.3). In this formula, the total variance of the random variable X is divided into two parts, an intra-group component and an inter-group component. The intra-group component averages the variances according to the proportions of each group, whereas the inter-group component accommodates for the difference in group means. A similar way of thinking can inspire a corresponding estimator for the variance. Such an estimator is presented in the definition below.

Definition 3.6. *The estimator $\widehat{\delta}_{p_1,p_2}$ for σ^{*2} is defined as:*

$$\widehat{\delta}_{p_1,p_2} := p_1\widehat{\sigma}_1^2 + p_2\widehat{\sigma}_2^2 + p_1(\widehat{\mu}_1 - \widehat{\mu})^2 + p_2(\widehat{\mu}_2 - \widehat{\mu})^2, \tag{3.11}$$

where $\widehat{\mu}$ is a plug-in estimator from the previous chapter for the mean μ^* of the complete population. The other plug-in estimators $\widehat{\mu}_1$ and $\widehat{\sigma}_1^2$ are defined in Equations (3.3) *and* (3.4)*, respectively. The used plug-in estimator for μ^* can be denoted in the following way:*

$$\widehat{\delta}_{p_1,p_2,\widehat{\theta}_{p_1,p_2}} := p_1\widehat{\sigma}_1^2 + p_2\widehat{\sigma}_2^2 + p_1(\widehat{\mu}_1 - \widehat{\theta}_{p_1,p_2})^2 + p_2(\widehat{\mu}_2 - \widehat{\theta}_{p_1,p_2})^2, \tag{3.12}$$

$$\widehat{\delta}_{p_1,p_2,\overline{X}} := p_1\widehat{\sigma}_1^2 + p_2\widehat{\sigma}_2^2 + p_1(\widehat{\mu}_1 - \overline{X})^2 + p_2(\widehat{\mu}_2 - \overline{X})^2. \tag{3.13}$$

The first two terms of Equation (3.11) are representing the intra-group variance and the last two terms the inter-group variance. If $\widehat{\mu}$ is chosen to be $\widehat{\theta}_{p_1,p_2}$, the estimator $\widehat{\delta}_{p_1,p_2}$ becomes equal to the previously defined $\widehat{\gamma}_{p_1,p_2}$. However, if a different estimator for μ^* is taken, for example $\overline{X}$, the estimator $\widehat{\delta}_{p_1,p_2}$ is different from $\widehat{\gamma}_{p_1,p_2}$. The theorem below shows the bias of $\widehat{\delta}_{p_1,p_2,\widehat{\theta}_{p_1,p_2}}$. Note that it is indeed equal to the bias of $\widehat{\gamma}_{p_1,p_2}$, given in Equation (3.5). The proof is given in Section 3.6.

Theorem 3.8. *The bias of $\widehat{\delta}_{p_1,p_2,\widehat{\theta}_{p_1,p_2}}$ is*

$$\mathbb{E}\left[\widehat{\delta}_{p_1,p_2,\widehat{\theta}_{p_1,p_2}}\right] - \sigma^{*2} = \widetilde{\Delta}p\left(\Delta\sigma^2 + \Delta\mu^2\right) + \mu^{*2} - \left(p_1\mu_1^* + p_2\mu_2^*\right)^2 + \frac{p_1p_2}{n_1}\sigma_1^{*2} + \frac{p_1p_2}{n_2}\sigma_2^{*2}.$$

Estimating ${\sigma^*}^2$ with $\widehat{\delta}_{p_1,p_2,\widehat{\theta}_{p_1,p_2}}$ is not a new concept, because it is the same estimator as $\widehat{\gamma}_{p_1,p_2}$. Therefore, $\widehat{\mu}$ of Equation (3.11) is chosen to be $\overline{X}$, resulting in the estimator $\widehat{\delta}_{p_1,p_2,\overline{X}}$. The bias is stated in the theorem below, which is proved in Section 3.6.

Theorem 3.9. *The bias of $\widehat{\delta}_{p_1,p_2,\overline{X}}$ is*

$$\mathbb{E}\left[\widehat{\delta}_{p_1,p_2,\overline{X}}\right] - {\sigma^*}^2 = \widetilde{\Delta} p\left(\Delta\sigma^2 + \Delta\mu^2\right) + {\mu^*}^2 + \left(\left(\widehat{p}_1 - 2p_1\right)\mu_1^* + \left(\widehat{p}_2 - 2p_2\right)\mu_2^*\right)\left(\widehat{p}_1\mu_1^* + \widehat{p}_2\mu_2^*\right) + \widehat{p}_2{\sigma_1^*}^2 + \widehat{p}_1{\sigma_2^*}^2.$$

This bias is of similar form as the bias of $\widehat{\delta}_{p_1,p_2,\widehat{\theta}_{p_1,p_2}}$. However, the last two terms do not tend to zero anymore. Moreover, these are always postive, except when the group variances are zero. That is a rare, unusual case, as it would mean that all measurements in a group are equal. Hence, it is unreasonable to estimate the variance with $\widehat{\delta}_{p_1,p_2,\overline{X}}$, as it will be biased in most cases.

Both $\widehat{\delta}_{p_1,p_2,\widehat{\theta}_{p_1,p_2}}$ and $\widehat{\delta}_{p_1,p_2,\overline{X}}$ are depending on the guessed proportions p_1 and p_2. When the true proportions p_1^* and p_2^* are available, it is natural to use them in the estimators. This leads to the estimator $\widehat{\delta}_{\text{oracle}}$, in which $\widehat{\theta}_{\text{oracle}}$ is taken as the plug-in estimator for μ^*.

Definition 3.7. *The estimator $\widehat{\delta}_{\text{oracle}}$ is defined as:*

$$\widehat{\delta}_{\text{oracle}} := p_1^*\widehat{\sigma}_1^2 + p_2^*\widehat{\sigma}_2^2 + p_1^*\left(\widehat{\mu}_1 - \widehat{\theta}_{\text{oracle}}\right)^2 + p_2^*\left(\widehat{\mu}_2 - \widehat{\theta}_{\text{oracle}}\right)^2. \tag{3.14}$$

The estimator defined above is the same as $\widehat{\gamma}_{\text{oracle}}$ and thus has the same bias. Because the estimator $\widehat{\psi}_{\text{oracle}}$ has a smaller bias than $\widehat{\gamma}_{\text{oracle}}$, using $\widehat{\delta}_{\text{oracle}}$ to estimate the variance is not preferable.

3.2. Variance & Comparison of the estimators

In order to compare all the previously defined estimators, their performances can be measured with the Mean Squared Error (MSE). The estimator with the smallest MSE generally has the best overall performance. The MSE consists of the variance of an estimator plus the square of its bias. The biases have already been discussed in the previous section. In this section, the variances are stated in various theorems and compared.

The variance of $\widehat{\text{Var}}[X]$ is stated in the theorem below. Section 3.7 contains the proof.

Theorem 3.10. *The variance of the estimator $\widehat{\text{Var}}[X]$ is*

$$\text{Var}[\widehat{\text{Var}}[X]] = \frac{1}{n}\left(\widehat{p}_1(m_{4,1}^* - {\sigma_1^*}^2) + \widehat{p}_2(m_{4,2}^* - {\sigma_2^*}^2) + \widehat{p}_1^2K_{1,1} + 2\widehat{p}_1\widehat{p}_2K_{1,2} + \widehat{p}_2^2K_{2,2}\right) + \mathcal{O}(1/n^2), \tag{3.15}$$

with

$$\begin{aligned} K_{r,s} := {} & 2\mu_s^*\left((\mu_r^{*2} + {\sigma_r^*}^2)\mu_r^* + 2\widetilde{\mu}{\sigma_r^*}^2 - m_{3,r}^*\right) + 2\mu_r^*\left((\mu_s^{*2} + {\sigma_s^*}^2)\mu_s^* + 2\widetilde{\mu}{\sigma_s^*}^2 - m_{3,s}^*\right) \\ & + 4\widetilde{\sigma}^2\left(\mu_r^* - \widetilde{\mu}\right)\left(\mu_s^* - \widetilde{\mu}\right) + \mathcal{O}(1/n), \end{aligned}$$

where $\widetilde{\mu} := \widehat{p}_1\mu_1^* + \widehat{p}_2\mu_2^*$ and $\widetilde{\sigma}^2 := \widehat{p}_1{\sigma_1^*}^2 + \widehat{p}_2{\sigma_2^*}^2$.

Remark that Landau's $\mathcal{O}(u_n)$-symbol denotes all terms of at least the order of some arbitrary sequence u_n. Consequently, the variance of $\widehat{\text{Var}}[X]$ is of the order $\mathcal{O}(1/n)$. Theorem 3.11 and Theorem 3.12 show that also both other examined estimators have a variance where all terms are depending on the number of measurements. Therefore the variances of all variance estimators are asymptotically equal to zero for $n \to +\infty$.

The variance of the estimator $\widehat{\gamma}_{p_1,p_2}$ is stated in the theorem below. Its proof can be found in Section 3.8.

Theorem 3.11. *The variance of $\widehat{\gamma}_{p_1,p_2}$ is*

$$\text{Var}\left[\widehat{\gamma}_{p_1,p_2}\right] = p_1^2\frac{G_1}{n_1} + p_2^2\frac{G_2}{n_2} + \mathcal{O}\left(\frac{1}{n^2}\right), \tag{3.16}$$

with

$$\begin{aligned} G_1 := {} & m_{4,1}^* - 4m_{3,1}^*\left(p_1\mu_1^* + p_2\mu_2^*\right) + 4{\sigma_1^*}^2\left(p_1^2{\mu_1^*}^2 - p_2\left(2p_2 - 3\right)\mu_1^*\mu_2^* + p_2^2{\mu_2^*}^2\right) \\ & + {\mu_1^*}^3\left(3\mu_1^* + 4p_2\Delta\mu\right) - {\sigma_1^*}^4, \end{aligned}$$

and G_2 equal to G_1 with all indices changed from 1 to 2 and vice versa.

The following theorem is proved in Section 3.8.

Theorem 3.12. *The variance of the estimator $\widehat{\psi}_{p_1,p_2}$ is:*

$$\mathrm{Var}\left[\widehat{\psi}_{p_1,p_2}\right] = p_1^2\frac{Q_1}{n_1} + p_2^2\frac{Q_2}{n_2} + \mathcal{O}\left(\frac{1}{n^2}\right), \tag{3.17}$$

where

$$\begin{aligned} Q_1 := {} & m_{4,1}^* - 4m_{3,1}^*\left(p_1\mu_1^* + p_2\mu_2^*\right) + \sigma_1^{*2}\left(\left(p_1^2 - 2p_2^2\right)\mu_1^{*2} - 4p_2\left(2p_2 - 3\right)\mu_1^*\mu_2^* + 4p_2^2\mu_2^{*2}\right) \\ & + \mu_1^{*3}\left(3\mu_1^* + 4p_2\Delta\mu\right) - \sigma_1^{*4}, \end{aligned}$$

and Q_2 is equal but with the indices changed from 1 to 2 and vice versa.

Remark that the estimators $\widehat{\gamma}_{p_1,p_2}$ and $\widehat{\psi}_{p_1,p_2}$ and $\widehat{\mathrm{Var}}[X]$ all have a variance of the order $\mathcal{O}(1/n)$. Comparing the variance of the naive estimator and the modified estimators precisely is left for further research. However, one can expect that there are cases in which the naive estimator has a smaller variance and cases in which the modified estimators have a smaller variance. Possibly, this depends on the quality of the guesses made for the proportions.

The variances of $\widehat{\gamma}_{p_1,p_2}$ and $\widehat{\psi}_{p_1,p_2}$ are similar and compared in the corollary below. Subtracting Equation 3.17 from Equation 3.16 results directly in this corollary.

Corollary 3.13. *The difference between the variances of the estimators $\widehat{\gamma}_{p_1,p_2}$ and $\widehat{\psi}_{p_1,p_2}$ is*

$$\mathrm{Var}\left[\widehat{\gamma}_{p_1,p_2}\right] - \mathrm{Var}\left[\widehat{\psi}_{p_1,p_2}\right] = \frac{1}{n_1}p_1^2\left(3p_1^2 + 2p_2^2\right)\sigma_1^{*2}\mu_1^{*2} + \frac{1}{n_2}p_2^2\left(2p_1^2 + 3p_2^2\right)\sigma_2^{*2}\mu_2^{*2} + \mathcal{O}\left(\frac{1}{n^2}\right) > 0,$$

for a sufficiently large number of measurements n.

The corollary above shows that the variance of $\widehat{\gamma}_{p_1,p_2}$ is larger than the variance of $\widehat{\psi}_{p_1,p_2}$ for the number of measurements sufficiently large. Because $\widehat{\psi}_{p_1,p_2}$ also has a smaller bias than $\widehat{\gamma}_{p_1,p_2}$, it has the smallest MSE. Therefore, $\widehat{\psi}_{p_1,p_2}$ is the best estimator to use when the true proportions are unavailable.

However, when these true proportions are known, the oracle estimators $\widehat{\gamma}_{\text{oracle}}$ and $\widehat{\psi}_{\text{oracle}}$ are available to use. Recall from the previous section that the estimator $\widehat{\psi}_{\text{oracle}}$ is unbiased. The true proportions p_1^* and p_2^* do not change the variances of the estimators, except changing the guessed proportions into true ones. This leads to the corollary below.

Corollary 3.14. *The difference between the variances of the estimators $\widehat{\gamma}_{\text{oracle}}$ and $\widehat{\psi}_{\text{oracle}}$ is*

$$\mathrm{Var}\left[\widehat{\gamma}_{\text{oracle}}\right] - \mathrm{Var}\left[\widehat{\psi}_{\text{oracle}}\right] = \frac{1}{n_1}p_1^{*2}\left(3p_1^{*2} + 2p_2^{*2}\right)\sigma_1^{*2}\mu_1^{*2} + \frac{1}{n_2}p_2^{*2}\left(2p_1^{*2} + 3p_2^{*2}\right)\sigma_2^{*2}\mu_2^{*2} + \mathcal{O}\left(\frac{1}{n^2}\right) > 0,$$

for a sufficiently large number of measurements n.

Both the variance and the square of the bias of $\widehat{\psi}_{\text{oracle}}$ are smaller than those of $\widehat{\gamma}_{\text{oracle}}$. This results in the MSE of $\widehat{\psi}_{\text{oracle}}$ being smaller than the MSE of $\widehat{\gamma}_{\text{oracle}}$. Thus, the estimator $\widehat{\psi}_{\text{oracle}}$ is the best estimator to use for σ^{*2}.

Because the estimator $\widehat{\delta}_{p_1,p_2}$ was not an improvement over the other estimators in terms of bias, the variance of it is not computed. However, it can be computed with similar techniques as used for the variance of the other estimators. The remainder of this chapter is dedicated to the proofs of all theorems stated in the first two sections.

3.3. Lemmas

In this section, several lemmas are developed in order to compute estimator biases and variances. In the calculations below, X_1 is a measurement X of group 1. Because the measurements are independent and identically distributed, it does not matter which measurement one takes of a certain group.

Lemma 3.15. *Consider the simple estimators $\widehat{\mu}_1$ and $\widehat{\sigma}_1^2$, defined in Equation* (3.3) *and Equation* (3.4)*, respectively. Their expected values, together with the expected value of $\widehat{\mu}_1^2$, are the following:*

$$\mathbb{E}\left[\widehat{\mu}_1\right] = \mu_1^*, \tag{3.18}$$

$$\mathbb{E}\left[\widehat{\mu}_1^2\right] = \frac{\sigma_1^{*2}}{n_1} + \mu_1^{*2}, \tag{3.19}$$

$$\mathbb{E}\left[\widehat{\sigma}_1^2\right] = \sigma_1^{*2}. \tag{3.20}$$

Proof of Lemma 3.15. The proof of Equation (3.18) is very short:

$$\mathbb{E}\left[\widehat{\mu}_1\right] = \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbb{E}\left[X_{i,1}\right] = \frac{n_1}{n_1}\mu_1^* = \mu_1^*.$$

With the use of Equation (2.2), the second result is derived:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\mu}_1^2\right] &= \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\mathbb{E}\left[X_{i,1}X_{j,1}\right] \\
&= \frac{1}{n_1^2}\left(n_1\mathbb{E}\left[X_1^2\right] + n_1\left(n_1-1\right)\mathbb{E}\left[X_1\right]^2\right) \\
&= \frac{1}{n_1^2}\left(n_1\left(\sigma_1^{*2}+\mu_1^{*2}\right) + n_1\left(n_1-1\right)\mu_1^{*2}\right) \\
&= \frac{\sigma_1^{*2}}{n_1} + \mu_1^{*2}.
\end{aligned}$$

Equation (3.20) is established using the result above as a known quantity. Firstly, the square is expanded:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\sigma}_1^2\right] &= \frac{1}{(n_1-1)}\sum_{i=1}^{n_1}\mathbb{E}\left[X_{i,1}^2 - 2X_{i,1}\widehat{\mu}_1 + \widehat{\mu}_1^2\right] \\
&= \frac{1}{(n_1-1)}\sum_{i=1}^{n_1}\left(\mathbb{E}\left[X_{i,1}^2\right] - \frac{2}{n_1}\mathbb{E}\left[\sum_{j=1}^{n_1}X_{i,1}X_{j,1}\right] + \mathbb{E}\left[\widehat{\mu}_1^2\right]\right) \\
&= \frac{n_1}{(n_1-1)}\left(\mathbb{E}\left[X_1^2\right] - \frac{2}{n_1}\mathbb{E}\left[X_1^2\right] - \frac{2\left(n_1-1\right)}{n_1}\mathbb{E}\left[X_1\right]^2 + \mathbb{E}\left[\widehat{\mu}_1^2\right]\right).
\end{aligned}$$

Secondly, the expectations of the equation above are filled in and simplified:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\sigma}_1^2\right] &= \frac{n_1}{(n_1-1)}\left(\sigma_1^{*2}+\mu_1^{*2} - \frac{2}{n_1}\left(\sigma_1^{*2}+\mu_1^{*2}\right) - \left(2-\frac{2}{n_1}\right)\mu_1^{*2} + \frac{\sigma_1^{*2}}{n_1} + \mu_1^{*2}\right) \\
&= \frac{n_1}{(n_1-1)}\left(1-\frac{1}{n_1}\right)\sigma_1^{*2} \\
&= \sigma_1^{*2}.
\end{aligned}$$

□

Lemma 3.16. *The expectation of the third power of the simple estimator $\widehat{\mu}_1$ of Equation* (3.3) *is:*

$$\mathbb{E}\left[\widehat{\mu}_1^3\right] = \mu_1^{*3} + \frac{3}{n_1}\sigma_1^{*2}\mu_1^* + \frac{1}{n_1^2}\left(m_{3,1}^* - 3\sigma_1^{*2}\mu_1^* - \mu_1^{*3}\right).$$

Proof of Lemma 3.16. Starting with expanding the third power and the triple sum gives:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\mu}_1^3\right] &= \frac{1}{n_1^3}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\sum_{k=1}^{n_1}\mathbb{E}\left[X_{i,1}X_{j,1}X_{k,1}\right] \\
&= \frac{1}{n_1^3}\left(n_1\mathbb{E}\left[X_1^3\right] + 3n_1\left(n_1-1\right)\mathbb{E}\left[X_1^2\right]\mathbb{E}\left[X_1\right] + n_1\left(n_1-1\right)\left(n_1-2\right)\mathbb{E}\left[X_1\right]^3\right).
\end{aligned}$$

All expected values of the expression above are known and can be filled in. Rewriting the equation leads to:

$$\begin{aligned}\mathbb{E}\left[\widehat{\mu}_1^3\right] &= \frac{1}{n_1^3}\left(n_1 m_{3,1}^* + 3n_1(n_1-1)\left(\sigma_1^{*2}+\mu_1^{*2}\right)\mu_1^* + n_1(n_1-1)(n_1-2)\mu_1^{*3}\right)\\ &= \frac{1}{n_1^2}m_{3,1}^* + \frac{n_1-1}{n_1^2}\left(3\sigma_1^{*2}\mu_1^* + (n_1+1)\mu_1^{*3}\right)\\ &= \mu_1^{*3} + \frac{3}{n_1}\sigma_1^{*2}\mu_1^* + \frac{1}{n_1^2}\left(m_{3,1}^* - 3\sigma_1^{*2}\mu_1^* - \mu_1^{*3}\right).\end{aligned}$$

□

Lemma 3.17. *Taking the expectation of the fourth power of the simple estimator $\widehat{\mu}_1$, given in Equation* (3.3)*, results in:*

$$\begin{aligned}\mathbb{E}\left[\widehat{\mu}_1^4\right] = \mu_1^{*4} + \frac{6}{n_1}\mu_1^{*2}\sigma_1^{*2} + \frac{1}{n_1^2}\left(4\mu_1^* m_{3,1}^* - 12\mu_1^{*2}\sigma_1^{*2} + 3\sigma_1^{*4} - 4\mu_1^{*4}\right)\\ + \frac{1}{n_1^3}\left(m_{4,1}^* - 4\mu_1^* m_{3,1}^* + 6\mu_1^{*2}\sigma_1^{*2} - 3\sigma_1^{*4} + 3\mu_1^{*4}\right).\end{aligned}$$

Proof of Lemma 3.17. The fourth power results in a quadruple sum, which is expanded below:

$$\begin{aligned}\mathbb{E}\left[\widehat{\mu}_1^4\right] &= \frac{1}{n_1^4}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\sum_{k=1}^{n_1}\sum_{\ell=1}^{n_1}\mathbb{E}\left[X_{i,1}X_{j,1}X_{k,1}X_{\ell,1}\right]\\ &= \frac{1}{n_1^4}\left(n_1\mathbb{E}\left[X_1^4\right] + 4n_1(n_1-1)\mathbb{E}\left[X_1^3\right]\mathbb{E}\left[X_1\right] + 6n_1(n_1-1)(n_1-2)\mathbb{E}\left[X_1^2\right]\mathbb{E}\left[X_1\right]^2\right)\\ &+ \frac{1}{n_1^4}\left(3n_1(n_1-1)\mathbb{E}\left[X_1^2\right]^2 + n_1(n_1-1)(n_1-2)(n_1-3)\mathbb{E}\left[X_1\right]^4\right).\end{aligned}$$

Simplifying this expression leads to:

$$\begin{aligned}\mathbb{E}\left[\widehat{\mu}_1^4\right] &= \frac{1}{n_1^4}\left(n_1 m_{4,1}^* + 4n_1(n_1-1)m_{3,1}^*\mu_1^* + 6n_1(n_1-1)(n_1-2)\left(\sigma_1^{*2}+\mu_1^{*2}\right)\mu_1^{*2}\right)\\ &+ \frac{1}{n_1^4}\left(3n_1(n_1-1)\left(\sigma_1^{*2}+\mu_1^{*2}\right)^2 + n_1(n_1-1)(n_1-2)(n_1-3)\mu_1^{*4}\right)\\ &= \frac{1}{n_1^3}m_{4,1}^* + \frac{4(n_1-1)}{n_1^3}\mu_1^* m_{3,1}^* + \frac{(n_1-1)\left(6(n_1-2)\mu_1^{*2} + 3\left(\sigma_1^{*2}+\mu_1^{*2}\right)\right)}{n_1^3}\left(\sigma_1^{*2}+\mu_1^{*2}\right)\\ &+ \frac{(n_1-1)(n_1-2)(n_1-3)}{n_1^3}\mu_1^{*4}\\ &= \frac{1}{n_1^3}m_{4,1}^* + \frac{n_1-1}{n_1^3}\left(4\mu_1^* m_{3,1}^* + \left(6(n_1-2)\mu_1^{*2} + 3\left(\sigma_1^{*2}+\mu_1^{*2}\right)\right)\left(\sigma_1^{*2}+\mu_1^{*2}\right) + (n_1-2)(n_1-3)\mu_1^{*4}\right).\end{aligned}$$

To get a clear overview of the order of all terms, it is rewritten as:

$$\begin{aligned}\mathbb{E}\left[\widehat{\mu}_1^4\right] &= \frac{1}{n_1^3}m_{4,1}^* + \frac{n_1-1}{n_1^3}\left(4\mu_1^* m_{3,1}^* + 6(n_1-1)\mu_1^{*2}\sigma_1^{*2} + 3\sigma_1^{*4} + \left(n_1^2+n_1-3\right)\mu_1^{*4}\right)\\ &= \mu_1^{*4} + \frac{6}{n_1}\mu_1^{*2}\sigma_1^{*2} + \frac{1}{n_1^2}\left(4\mu_1^* m_{3,1}^* - 12\mu_1^{*2}\sigma_1^{*2} + 3\sigma_1^{*4} - 4\mu_1^{*4}\right)\\ &+ \frac{1}{n_1^3}\left(m_{4,1}^* - 4\mu_1^* m_{3,1}^* + 6\mu_1^{*2}\sigma_1^{*2} - 3\sigma_1^{*4} + 3\mu_1^{*4}\right).\end{aligned}$$

□

Lemma 3.18. *Remark the two simple estimators $\widehat{\sigma}_1^2$ and $\widehat{\mu}_1$, given in Equation* (3.4) *and Equation* (3.3)*, respectively. The expected value of the product of these estimators is:*

$$\mathbb{E}\left[\widehat{\sigma}_1^2\widehat{\mu}_1\right] = \sigma_1^{*2}\mu_1^* + \frac{1}{n_1}\left(m_{3,1}^* - 3\sigma_1^{*2}\mu_1^* - \mu_1^{*3}\right).$$

Proof of Lemma 3.18. Starting with the expansion of all terms and rewriting this to already known quantities, leads to:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\sigma}_1^2\widehat{\mu}_1\right] &= \frac{1}{n_1-1}\mathbb{E}\left[\left(\sum_{i=1}^{n_1}\left(X_{i,1}-\widehat{\mu}_1\right)^2\right)\widehat{\mu}_1\right] \\
&= \frac{1}{n_1-1}\sum_{i=1}^{n_1}\mathbb{E}\left[X_{i,1}^2\widehat{\mu}_1 - 2X_{i,1}\widehat{\mu}_1^2 + \widehat{\mu}_1^3\right] \\
&= \frac{1}{n_1-1}\left(\frac{1}{n_1}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\mathbb{E}\left[X_{i,1}^2X_{j,1}\right] - \frac{2}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\sum_{k=1}^{n_1}\mathbb{E}\left[X_{i,1}X_{j,1}X_{k,1}\right] + \sum_{i=1}^{n_1}\mathbb{E}\left[\widehat{\mu}_1^3\right]\right) \\
&= \frac{1}{n_1\left(n_1-1\right)}\left(n_1\mathbb{E}\left[X_1^3\right] + n_1\left(n_1-1\right)\mathbb{E}\left[X_1^2\right]\mathbb{E}\left[X_1\right]\right) - \frac{2n_1}{n_1-1}\mathbb{E}\left[\widehat{\mu}_1^3\right] + \frac{n_1}{n_1-1}\mathbb{E}\left[\widehat{\mu}_1^3\right] \\
&= \frac{1}{n_1\left(n_1-1\right)}\left(n_1\mathbb{E}\left[X_1^3\right] + n_1\left(n_1-1\right)\mathbb{E}\left[X_1^2\right]\mathbb{E}\left[X_1\right]\right) - \frac{n_1}{n_1-1}\mathbb{E}\left[\widehat{\mu}_1^3\right].
\end{aligned}$$

Filling in all terms (using Lemma 3.16) and simplifying the equation, results in the desired expression:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\sigma}_1^2\widehat{\mu}_1\right] &= \frac{1}{n_1-1}\left(m_{3,1}^* + \left(n_1-1\right)\left(\sigma_1^{*2}+\mu_1^{*2}\right)\mu_1^*\right) - \frac{n_1}{n_1-1}\left(\frac{1}{n_1^2}m_{3,1}^* + \frac{n_1-1}{n_1^2}\left(3\sigma_1^{*2}\mu_1^* + \left(n_1+1\right)\mu_1^{*3}\right)\right) \\
&= \frac{1}{n_1-1}m_{3,1}^* + \left(\sigma_1^{*2}+\mu_1^{*2}\right)\mu_1^* - \frac{1}{n_1\left(n_1-1\right)}m_{3,1}^* - \frac{1}{n_1}\left(3\sigma_1^{*2}\mu_1^* + \left(n_1+1\right)\mu_1^{*3}\right) \\
&= \frac{n_1-1}{n_1\left(n_1-1\right)}m_{3,1}^* + \frac{n_1-3}{n_1}\sigma_1^{*2}\mu_1^* + \frac{n_1-\left(n_1+1\right)}{n_1}\mu_1^{*3} \\
&= \frac{1}{n_1}\left(m_{3,1}^* + \left(n_1-3\right)\sigma_1^{*2}\mu_1^* - \mu_1^{*3}\right) \\
&= \sigma_1^{*2}\mu_1^* + \frac{1}{n_1}\left(m_{3,1}^* - 3\sigma_1^{*2}\mu_1^* - \mu_1^{*3}\right).
\end{aligned}$$

□

Lemma 3.19. *The expectation of the product of the estimators $\widehat{\sigma}_1^2$ and $\widehat{\mu}_1^2$ is:*

$$\mathbb{E}\left[\widehat{\sigma}_1^2\widehat{\mu}_1^2\right] = \sigma_1^{*2}\mu_1^{*2} + \frac{1}{n_1}\left(2m_{3,1}^*\mu_1^* - 6\sigma_1^{*2}\mu_1^{*2} - 2\mu_1^{*4} + \sigma_1^{*4}\right) + \frac{1}{n_1^2}\left(m_{4,1}^* - 4m_{3,1}^*\mu_1^* + 6\sigma_1^{*2}\mu_1^{*2} + 3\mu_1^{*4} - 3\sigma_1^{*4}\right).$$

Proof of Lemma 3.19. The squares of the estimators are expanded, resulting in multiple sums:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\sigma}_1^2\widehat{\mu}_1^2\right] &= \frac{1}{n_1-1}\mathbb{E}\left[\left(\sum_{i=1}^{n_1}\left(X_{i,1}-\widehat{\mu}_1\right)^2\right)\widehat{\mu}_1^2\right] \\
&= \frac{1}{n_1-1}\sum_{i=1}^{n_1}\mathbb{E}\left[X_{i,1}^2\widehat{\mu}_1^2 - 2X_{i,1}\widehat{\mu}_1^3 + \widehat{\mu}_1^4\right] \\
&= \frac{1}{n_1-1}\left(\frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\sum_{k=1}^{n_1}\mathbb{E}\left[X_{i,1}^2X_{j,1}X_{k,1}\right] - \frac{2}{n_1^3}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\sum_{k=1}^{n_1}\sum_{\ell=1}^{n_1}\mathbb{E}\left[X_{i,1}X_{j,1}X_{k,1}X_{\ell,1}\right] + \sum_{i=1}^{n_1}\mathbb{E}\left[\widehat{\mu}_1^4\right]\right).
\end{aligned}$$

This can be rewritten as

$$\begin{aligned}
&\frac{1}{n_1-1}\left(\frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\sum_{k=1}^{n_1}\mathbb{E}\left[X_{i,1}^2X_{j,1}X_{k,1}\right] - 2n_1\mathbb{E}\left[\widehat{\mu}_1^4\right] + n_1\mathbb{E}\left[\widehat{\mu}_1^4\right]\right) \\
&= \frac{1}{n_1-1}\left(\frac{1}{n_1^2}\left(n_1\mathbb{E}\left[X_1^4\right] + 2n_1\left(n_1-1\right)\mathbb{E}\left[X_1^3\right]\mathbb{E}\left[X_1\right] + n_1\left(n_1-1\right)\mathbb{E}\left[X_1^2\right]^2\right)\right. \\
&\qquad \left. + \frac{1}{n_1^2}n_1\left(n_1-1\right)\left(n_1-2\right)\mathbb{E}\left[X_1^2\right]\mathbb{E}\left[X_1\right]^2 - n_1\mathbb{E}\left[\widehat{\mu}_1^4\right]\right).
\end{aligned}$$

Replacing the expectations with their known quantities, using Lemma 3.17, leads to:

$$\begin{aligned}
&\frac{1}{n_1(n_1-1)}\left(m_{4,1}^* + 2(n_1-1)m_{3,1}^*\mu_1^* + (n_1-1)\left(\sigma_1^{*2}+\mu_1^{*2}\right)^2 + (n_1-1)(n_1-2)\left(\sigma_1^{*2}+\mu_1^{*2}\right)\mu_1^{*2}\right)\\
&-\frac{1}{n_1-1}\left(\frac{1}{n_1^2}m_{4,1}^* + \frac{(n_1-1)}{n_1^2}\left(4\mu_1^* m_{3,1}^* + \left(6(n_1-2)\mu_1^{*2} + 3\left(\sigma_1^{*2}+\mu_1^{*2}\right)\right)\left(\sigma_1^{*2}+\mu_1^{*2}\right) + (n_1-2)(n_1-3)\mu_1^{*4}\right)\right)\\
&= \frac{n_1-1}{n_1^2(n_1-1)}m_{4,1}^* + \frac{1}{n_1}\left(2m_{3,1}^*\mu_1^* + \left((n_1-2)\mu_1^{*2} + \left(\sigma_1^{*2}+\mu_1^{*2}\right)\right)\left(\sigma_1^{*2}+\mu_1^{*2}\right)\right)\\
&-\frac{1}{n_1^2}\left(4m_{3,1}^*\mu_1^* + \left(6(n_1-2)\mu_1^{*2} + 3\left(\sigma_1^{*2}+\mu_1^{*2}\right)\right)\left(\sigma_1^{*2}+\mu_1^{*2}\right) + (n_1-2)(n_1-3)\mu_1^{*4}\right).
\end{aligned}$$

Simplifying and collecting all terms with equal order results in the desired expression for $\mathbb{E}\left[\widehat{\sigma}_1^2\widehat{\mu}_1^2\right]$:

$$\begin{aligned}
&\frac{1}{n_1^2}\left(m_{4,1}^* + \left((n_1-2)(n_1-6)\mu_1^{*2} + (n_1-3)\left(\sigma_1^{*2}+\mu_1^{*2}\right)\right)\left(\sigma_1^{*2}+\mu_1^{*2}\right) + 2(n_1-2)m_{3,1}^*\mu_1^* - (n_1-2)(n_1-3)\mu_1^{*4}\right)\\
&= \frac{1}{n_1^2}\left(m_{4,1}^* + \left(n_1^2-8n_1+12\right)\sigma_1^{*2}\mu_1^{*2} + \left(n_1^2-8n_1+12\right)\mu_1^{*4} + (n_1-3)\left(\sigma_1^{*4} + 2\sigma_1^{*2}\mu_1^{*2} + \mu_1^{*4}\right)\right.\\
&\qquad \left. + (2n_1-4)m_{3,1}^*\mu_1^* - \left(n_1^2-5n_1+6\right)\mu_1^{*4}\right)\\
&= \frac{1}{n_1^2}\left(m_{4,1}^* + 2(n_1-2)m_{3,1}^*\mu_1^* + \left(n_1^2-6n_1+6\right)\sigma_1^{*2}\mu_1^{*2} + (-2n_1+3)\mu_1^{*4} + (n_1-3)\sigma_1^{*4}\right)\\
&= \sigma_1^{*2}\mu_1^{*2} + \frac{1}{n_1}\left(2m_{3,1}^*\mu_1^* - 6\sigma_1^{*2}\mu_1^{*2} - 2\mu_1^{*4} + \sigma_1^{*4}\right) + \frac{1}{n_1^2}\left(m_{4,1}^* - 4m_{3,1}^*\mu_1^* + 6\sigma_1^{*2}\mu_1^{*2} + 3\mu_1^{*4} - 3\sigma_1^{*4}\right).
\end{aligned}$$

□

Lemma 3.20. *Remember the estimator $\widehat{\sigma}_1^2$ from Equation* (3.4)*. The expected value of its square is*

$$\mathbb{E}\left[\widehat{\sigma}_1^4\right] = \sigma_1^{*4} + \frac{1}{n_1}\left(m_{4,1}^* - 4m_{3,1}^*\mu_1^* + 6\sigma_1^{*2}\mu_1^{*2} + 3\mu_1^{*4} - \sigma_1^{*4}\right) + \frac{2}{n_1(n_1-1)}\sigma_1^{*4}.$$

Proof of Lemma 3.20. We remark that $\mathbb{E}\left[\widehat{\sigma}_1^4\right]$ can be rewritten as

$$\begin{aligned}
&\frac{1}{(n_1-1)^2}\mathbb{E}\left[\left(\sum_{i=1}^{n_1}\left(X_{i,1}-\widehat{\mu}_1\right)^2\right)^2\right]\\
&= \frac{1}{(n_1-1)^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\mathbb{E}\left[\left(X_{i,1}-\widehat{\mu}_1\right)^2\left(X_{j,1}-\widehat{\mu}_1\right)^2\right]\\
&= \frac{1}{(n_1-1)^2}\left(\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\mathbb{E}\left[X_{i,1}^2X_{j,1}^2\right] - 2\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\mathbb{E}\left[X_{i,1}^2X_{j,1}\widehat{\mu}_1\right] + \sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\mathbb{E}\left[X_{i,1}^2\widehat{\mu}_1^2\right]\right.\\
&\qquad -2\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\mathbb{E}\left[X_{i,1}X_{j,1}^2\widehat{\mu}_1\right] + 4\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\mathbb{E}\left[X_{i,1}X_{j,1}\widehat{\mu}_1^2\right] - 2\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\mathbb{E}\left[X_{i,1}\widehat{\mu}_1^3\right]\\
&\qquad \left. + \sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\mathbb{E}\left[X_{j,1}^2\widehat{\mu}_1^2\right] - 2\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\mathbb{E}\left[X_{j,1}\widehat{\mu}_1^3\right] + \sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\mathbb{E}\left[\widehat{\mu}_1^4\right]\right).
\end{aligned}$$

Rewriting the terms above into known expectations leads to:

$$\begin{aligned}
&\frac{1}{(n_1-1)^2}\left(\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\mathbb{E}\left[X_{i,1}^2X_{j,1}^2\right]-2n_1\sum_{i=1}^{n_1}\mathbb{E}\left[X_{i,1}^2\widehat{\mu}_1^2\right]+n_1\sum_{i=1}^{n_1}\mathbb{E}\left[X_{i,1}^2\widehat{\mu}_1^2\right]-2n_1\sum_{i=1}^{n_1}\mathbb{E}\left[X_{i,1}^2\widehat{\mu}_1^2\right]\right.\\
&\qquad\left.+4n_1^2\mathbb{E}\left[\widehat{\mu}_1^4\right]-2n_1^2\mathbb{E}\left[\widehat{\mu}_1^4\right]+n_1\sum_{i=1}^{n_1}\mathbb{E}\left[X_{i,1}^2\widehat{\mu}_1^2\right]-2n_1^2\mathbb{E}\left[\widehat{\mu}_1^4\right]+n_1^2\mathbb{E}\left[\widehat{\mu}_1^4\right]\right)\\
&=\frac{1}{(n_1-1)^2}\left(\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\mathbb{E}\left[X_{i,1}^2X_{j,1}^2\right]-2n_1\sum_{i=1}^{n_1}\mathbb{E}\left[X_{i,1}^2\widehat{\mu}_1^2\right]+n_1^2\mathbb{E}\left[\widehat{\mu}_1^4\right]\right)\\
&=\frac{1}{(n_1-1)^2}\left(n_1\mathbb{E}\left[X_1^4\right]+n_1(n_1-1)\mathbb{E}\left[X_1^2\right]^2-\frac{2}{n_1}\left(n_1\mathbb{E}\left[X_1^4\right]+2n_1(n_1-1)\mathbb{E}\left[X_1^3\right]\mathbb{E}[X_1]\right)\right.\\
&\qquad\left.-\frac{2}{n_1}\left(n_1(n_1-1)\mathbb{E}\left[X_1^2\right]^2+n_1(n_1-1)(n_1-2)\mathbb{E}\left[X_1^2\right]\mathbb{E}[X_1]^2\right)+n_1^2\mathbb{E}\left[\widehat{\mu}_1^4\right]\right)\\
&=\frac{1}{(n_1-1)^2}\left((n_1-2)\mathbb{E}\left[X_1^4\right]+(n_1-1)(n_1-2)\mathbb{E}\left[X_1^2\right]^2\right.\\
&\qquad\left.-2(n_1-1)\left(2\mathbb{E}\left[X_1^3\right]\mathbb{E}[X_1]+(n_1-2)\mathbb{E}\left[X_1^2\right]\mathbb{E}[X_1]^2\right)+n_1^2\mathbb{E}\left[\widehat{\mu}_1^4\right]\right).
\end{aligned}$$

These expectations are known (using Lemma 3.17). Writing $\mathbb{E}\left[\widehat{\sigma}_1^4\right]$ in terms of $m_{4,1}^*, m_{3,1}^*, \sigma_1^*$ and μ_1^* leads to:

$$\begin{aligned}
&\frac{1}{(n_1-1)^2}\left((n_1-2)m_{4,1}^*+(n_1-1)(n_1-2)\left(\sigma_1^{*2}+\mu_1^{*2}\right)^2-2(n_1-1)\left(2m_{3,1}^*\mu_1^*+(n_1-2)\left(\sigma_1^{*2}+\mu_1^{*2}\right)\mu_1^{*2}\right)\right.\\
&\qquad\left.+\frac{1}{n_1}m_{4,1}^*+\frac{n_1-1}{n_1}\left(4\mu_1^*m_{3,1}^*+\left(6(n_1-2)\mu_1^{*2}+3\left(\sigma_1^{*2}+\mu_1^{*2}\right)\right)\left(\sigma_1^{*2}+\mu_1^{*2}\right)+(n_1-2)(n_1-3)\mu_1^{*4}\right)\right)\\
&=\frac{1}{(n_1-1)^2}\left(\frac{n_1^2-2n_1+1}{n_1}m_{4,1}^*+(n_1-1)\left(-\frac{4(n_1-1)}{n_1}m_{3,1}^*\mu_1^*+(n_1-2)\left(\sigma_1^{*2}+\mu_1^{*2}\right)\left(\sigma_1^{*2}-\mu_1^{*2}\right)\right)\right.\\
&\qquad\left.+\frac{n_1-1}{n_1}\left(\left(6(n_1-2)\mu_1^{*2}+3\left(\sigma_1^{*2}+\mu_1^{*2}\right)\right)\left(\sigma_1^{*2}+\mu_1^{*2}\right)+(n_1-2)(n_1-3)\mu_1^{*4}\right)\right)\\
&=\frac{1}{(n_1-1)^2}\left(\frac{(n_1-1)^2}{n_1}\left(m_{4,1}^*-4m_{3,1}^*\mu_1^*\right)+(n_1-1)(n_1-2)\left(\sigma_1^{*4}-\mu_1^{*4}\right)\right.\\
&\qquad\left.+\frac{n_1-1}{n_1}\left(3\sigma_1^{*4}+6(n_1-1)\sigma_1^{*2}\mu_1^{*2}+\left(n_1^2+n_1-3\right)\mu_1^{*4}\right)\right),
\end{aligned}$$

which is rewritten into:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\sigma}_1^4\right]&=\frac{1}{n_1}\left(m_{4,1}^*-4m_{3,1}^*\mu_1^*\right)+\frac{1}{n_1(n_1-1)}\left(\left(n_1^2-2n_1+3\right)\sigma_1^{*4}+6(n_1-1)\sigma_1^{*2}\mu_1^{*2}+3(n_1-1)\mu_1^{*4}\right)\\
&=\frac{1}{n_1}\left(m_{4,1}^*-4m_{3,1}^*\mu_1^*+6\sigma_1^{*2}\mu_1^{*2}+3\mu_1^{*4}\right)+\frac{1}{n_1(n_1-1)}\left((n_1-1)^2+2\right)\sigma_1^{*4}\\
&=\sigma_1^{*4}+\frac{1}{n_1}\left(m_{4,1}^*-4m_{3,1}^*\mu_1^*+6\sigma_1^{*2}\mu_1^{*2}+3\mu_1^{*4}-\sigma_1^{*4}\right)+\frac{2}{n_1(n_1-1)}\sigma_1^{*4}.
\end{aligned}$$

□

The expected values of the squares of both main estimators of Chapter 2 are needed for other proofs about estimators for the true variance and are therefore computed in the Lemmas below.

Lemma 3.21. *The expectation of the square of the estimator $\overline{X}$, which is defined in Equation* (2.5)*, is:*

$$\mathbb{E}\left[\overline{X}^2\right]=\frac{\widehat{p}_1}{n}\sigma_1^{*2}+\frac{\widehat{p}_2}{n}\sigma_2^{*2}+\left(\widehat{p}_1\mu_1^*+\widehat{p}_2\mu_2^*\right)^2.$$

Proof of Lemma 3.21. Starting with expanding the square and the double sums that appear,

$$\begin{aligned}
\mathbb{E}\left[\overline{X}^2\right]&=\frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^{n_1}X_{i,1}\right)^2+2\left(\sum_{i=1}^{n_1}X_{i,1}\right)\left(\sum_{i=1}^{n_2}X_{i,2}\right)+\left(\sum_{i=1}^{n_2}X_{i,2}\right)^2\right]\\
&=\frac{1}{n^2}\left(\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\mathbb{E}\left[X_{i,1}X_{j,1}\right]+2\left(\sum_{i=1}^{n_1}\mathbb{E}\left[X_{i,1}\right]\right)\left(\sum_{i=1}^{n_2}\mathbb{E}\left[X_{i,2}\right]\right)+\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}\mathbb{E}\left[X_{i,2}X_{j,2}\right]\right)\\
&=\frac{1}{n^2}\left(\left(n_1\mathbb{E}\left[X_1^2\right]+n_1(n_1-1)\mathbb{E}[X_1]^2\right)+2n_1n_2\mathbb{E}\left[X_{i,1}\right]\mathbb{E}\left[X_{i,2}\right]+\left(n_2\mathbb{E}\left[X_2^2\right]+n_2(n_2-1)\mathbb{E}[X_2]^2\right)\right).
\end{aligned}$$

The expression above gets simplified to:

$$\begin{aligned}\mathbb{E}\left[\overline{X}^2\right] &= \frac{1}{n^2}\left(n_1\sigma_1^{*2} + n_2\sigma_2^{*2} + n_1^2\mu_1^{*2} + n_2^2\mu_2^{*2} + 2n_1n_2\mu_1^*\mu_2^*\right)\\ &= \frac{\widehat{p}_1}{n}\sigma_1^{*2} + \frac{\widehat{p}_2}{n}\sigma_2^{*2} + \left(\widehat{p}_1\mu_1^* + \widehat{p}_2\mu_2^*\right)^2.\end{aligned}$$

□

Lemma 3.22. *The expectation of the square of the estimator $\widehat{\theta}_{p_1,p_2}$, which is defined in Equation* (2.6)*, is:*

$$\mathbb{E}\left[\widehat{\theta}^2_{p_1,p_2}\right] = \frac{p_1^2}{n_1}\sigma_1^{*2} + \frac{p_2^2}{n_2}\sigma_2^{*2} + \left(p_1\mu_1^* + p_2\mu_2^*\right)^2.$$

Proof of Lemma 3.22. Expanding the square leads to the following equation:

$$\begin{aligned}\mathbb{E}\left[\widehat{\theta}^2_{p_1,p_2}\right] &= \mathbb{E}\left[\left(\frac{p_1}{n_1}\sum_{i=1}^{n_1}X_{i,1} + \frac{p_2}{n_2}\sum_{i=1}^{n_2}X_{i,2}\right)^2\right]\\ &= \frac{p_1^2}{n_1^2}\mathbb{E}\left[\left(\sum_{i=1}^{n_1}X_{i,1}\right)^2\right] + \frac{2p_1p_2}{n_1n_2}\mathbb{E}\left[\left(\sum_{i=1}^{n_1}X_{i,1}\right)\left(\sum_{i=1}^{n_2}X_{i,2}\right)\right] + \frac{p_2^2}{n_2^2}\mathbb{E}\left[\left(\sum_{i=1}^{n_2}X_{i,2}\right)^2\right]\\ &= \frac{p_1^2}{n_1^2}\left(n_1\mathbb{E}\left[X_1^2\right] + n_1\left(n_1-1\right)\mathbb{E}\left[X_1\right]^2\right) + \frac{2p_1p_2}{n_1n_2}\sum_{i=1}^{n_1}\mathbb{E}\left[X_{i,1}\right]\sum_{i=1}^{n_2}\mathbb{E}\left[X_{i,2}\right]\\ &+ \frac{p_2^2}{n_2^2}\left(n_2\mathbb{E}\left[X_2^2\right] + n_2\left(n_2-1\right)\mathbb{E}\left[X_2\right]^2\right).\end{aligned}$$

Simplifying this expression results in:

$$\begin{aligned}\mathbb{E}\left[\widehat{\theta}^2_{p_1,p_2}\right] &= \frac{p_1^2}{n_1^2}\left(n_1\left(\sigma_1^{*2}+\mu_1^{*2}\right) + n_1\left(n_1-1\right)\mu_1^{*2}\right) + 2p_1p_2\mu_1^*\mu_2^* + \frac{p_2^2}{n_2^2}\left(n_2\left(\sigma_2^{*2}+\mu_2^{*2}\right) + n_2\left(n_2-1\right)\mu_2^{*2}\right)\\ &= \frac{p_1^2}{n_1}\left(\sigma_1^{*2}+n_1\mu_1^{*2}\right) + 2p_1p_2\mu_1^*\mu_2^* + \frac{p_2^2}{n_2}\left(\sigma_2^{*2}+n_2\mu_2^{*2}\right)\\ &= \frac{p_1^2}{n_1}\sigma_1^{*2} + \frac{p_2^2}{n_2}\sigma_2^{*2} + \left(p_1\mu_1^* + p_2\mu_2^*\right)^2.\end{aligned}$$

□

Having stated the Lemmas above, the theorems stated in Section 3.1 and Section 3.2 can be proved. These proofs are shown in the remainder of this chapter.

3.4. Bias of $\widehat{\mathrm{Var}}[X]$ – Proof

In this section, Theorem 3.1 is proved. This theorem states the bias of the estimator $\widehat{\mathrm{Var}}[X]$ from Equation (3.1).

Proof of Theorem 3.1. The bias is by definition equal to $\mathbb{E}\left[\widehat{\mathrm{Var}}[X]\right] - \sigma^{*2}$. To start with, the expectation is computed in Subsection 3.4.1, followed by subtracting the true variance in Subsection 3.4.2.

3.4.1. Computing the expected value

The expectation of the estimator can be split up in the following way:

$$\mathbb{E}\left[\widehat{\mathrm{Var}}[X]\right] = \mathbb{E}\left[\frac{1}{n-1}\left(\sum_{i=1}^{n_1}\left(X_{i,1}-\overline{X}\right)^2 + \sum_{i=1}^{n_2}\left(X_{i,2}-\overline{X}\right)^2\right)\right] = \frac{1}{n-1}A_1 + \frac{1}{n-1}A_2,$$

with

$$\begin{aligned}
A_1 &:= \mathbb{E}\left[\sum_{i=1}^{n_1}\left(X_{i,1}-\overline{X}\right)^2\right] \\
&= \sum_{i=1}^{n_1}\left(\mathbb{E}\left[X_{i,1}^2\right]-2\mathbb{E}\left[X_{i,1}\overline{X}\right]+\mathbb{E}\left[\overline{X}^2\right]\right) \\
&= \sum_{i=1}^{n_1}\left(\mathbb{E}\left[X_{i,1}^2\right]-\frac{2}{n}\left(\sum_{j=1}^{n_1}\mathbb{E}\left[X_{i,1}X_{j,1}\right]+\sum_{j=1}^{n_2}\mathbb{E}\left[X_{i,1}X_{j,2}\right]\right)+\mathbb{E}\left[\overline{X}^2\right]\right).
\end{aligned}$$

This can be further simplified to

$$\begin{aligned}
A_1 &= n_1\left(\mathbb{E}\left[X_1^2\right]-\frac{2}{n}\left(\mathbb{E}\left[X_1^2\right]+(n_1-1)\,\mathbb{E}[X_1]^2+n_2\mathbb{E}[X_1]\,\mathbb{E}[X_2]\right)+\mathbb{E}\left[\overline{X}^2\right]\right) \\
&= \frac{n_1}{n}\left((n-2)\,\mathbb{E}\left[X_1^2\right]-2(n_1-1)\,\mathbb{E}[X_1]^2-2n_2\mathbb{E}[X_1]\,\mathbb{E}[X_2]\right)+n_1\mathbb{E}\left[\overline{X}^2\right] \\
&= \widehat{p}_1\left((n-2)\left(\sigma_1^{*2}+\mu_1^{*2}\right)-2(n_1-1)\,\mu_1^{*2}-2n_2\mu_1^*\mu_2^*\right)+n_1\left(\frac{\widehat{p}_1}{n}\sigma_1^{*2}+\frac{\widehat{p}_2}{n}\sigma_2^{*2}+\left(\widehat{p}_1\mu_1^*+\widehat{p}_2\mu_2^*\right)^2\right) \\
&= \widehat{p}_1\left((n-2)\,\sigma_1^{*2}+(n_2-n_1)\,\mu_1^{*2}-2n_2\mu_1^*\mu_2^*\right)+\widehat{p}_1^2\sigma_1^{*2}+\widehat{p}_1\widehat{p}_2\sigma_2^{*2}+n_1\left(\widehat{p}_1\mu_1^*+\widehat{p}_2\mu_2^*\right)^2.
\end{aligned}$$

The second term, A_2, is equal to A_1 with all indices switched from 1 to 2 and vice versa. Note that Lemma (3.21) is used in the third line above. Dividing by $n-1$ gives:

$$\begin{aligned}
\frac{A_1}{n-1} &= \widehat{p}_1\left(\frac{n-2}{n-1}\sigma_1^{*2}+\frac{n_2-n_1}{n-1}\mu_1^{*2}-\frac{2n_2}{n-1}\mu_1^*\mu_2^*\right)+\frac{\widehat{p}_1^2\sigma_1^{*2}+\widehat{p}_1\widehat{p}_2\sigma_2^{*2}}{n-1}+\frac{n_1}{n-1}\left(\widehat{p}_1\mu_1^*+\widehat{p}_2\mu_2^*\right)^2 \\
&= \widehat{p}_1\left(\left(\frac{n-1}{n-1}+\frac{-1}{n-1}\right)\sigma_1^{*2}+\left(\frac{n_2}{n}-\frac{n_1}{n}+\frac{n_2-n_1}{n(n-1)}\right)\mu_1^{*2}-2\left(\frac{n_2}{n}+\frac{n_2}{n(n-1)}\right)\mu_1^*\mu_2^*\right) \\
&\quad+\frac{\widehat{p}_1^2\sigma_1^{*2}+\widehat{p}_1\widehat{p}_2\sigma_2^{*2}}{n-1}+\left(\frac{n_1}{n}+\frac{n_1}{n(n-1)}\right)\left(\widehat{p}_1\mu_1^*+\widehat{p}_2\mu_2^*\right)^2.
\end{aligned}$$

All terms of the order $\mathcal{O}(1/n)$ are taken separately and are put into the term $\epsilon_{n,1}$ below:

$$\begin{aligned}
\frac{A_1}{n-1} &= \widehat{p}_1\left(\sigma_1^{*2}+\left(\widehat{p}_2-\widehat{p}_1\right)\mu_1^{*2}-2\widehat{p}_2\mu_1^*\mu_2^*\right)+\widehat{p}_1\left(\widehat{p}_1\mu_1^*+\widehat{p}_2\mu_2^*\right)^2+\epsilon_{n,1} \\
&= \widehat{p}_1\left(\sigma_1^{*2}+\left(\Delta p+p_2^*-\widehat{p}_1\right)\mu_1^{*2}-2\widehat{p}_2\mu_1^*\mu_2^*\right)+\widehat{p}_1\left(\widehat{p}_1\mu_1^*+\widehat{p}_2\mu_2^*\right)^2+\epsilon_{n,1} \\
&= \widehat{p}_1\left(\sigma_1^{*2}+\left(\Delta p+1-p_1^*-\widehat{p}_1\right)\mu_1^{*2}-2\widehat{p}_2\mu_1^*\mu_2^*\right)+\widehat{p}_1\left(\widehat{p}_1\mu_1^*+\widehat{p}_2\mu_2^*\right)^2+\epsilon_{n,1} \\
&= \widehat{p}_1\left(\sigma_1^{*2}+\left(1-2\widehat{p}_1\right)\mu_1^{*2}-2\widehat{p}_2\mu_1^*\mu_2^*\right)+\widehat{p}_1\left(\widehat{p}_1\mu_1^*+\widehat{p}_2\mu_2^*\right)^2+\epsilon_{n,1},
\end{aligned}$$

with

$$\begin{aligned}
\epsilon_{n,1} &:= \widehat{p}_1\left(-\frac{1}{n-1}\sigma_1^{*2}+\frac{n_2-n_1}{n(n-1)}\mu_1^{*2}-\frac{2n_2}{n(n-1)}\mu_1^*\mu_2^*\right)+\frac{\widehat{p}_1^2\sigma_1^{*2}+\widehat{p}_1\widehat{p}_2\sigma_2^{*2}}{n-1}+\frac{n_1}{n(n-1)}\left(\widehat{p}_1\mu_1^*+\widehat{p}_2\mu_2^*\right)^2 \\
&= \frac{\widehat{p}_1}{n-1}\left(-\sigma_1^{*2}+\left(\widehat{p}_2-\widehat{p}_1\right)\mu_1^{*2}-2\widehat{p}_2\mu_1^*\mu_2^*+\widehat{p}_1\sigma_1^{*2}+\widehat{p}_2\sigma_2^{*2}+\left(\widehat{p}_1\mu_1^*+\widehat{p}_2\mu_2^*\right)^2\right) \\
&= \frac{\widehat{p}_1}{n-1}\left(\left(\widehat{p}_2-\widehat{p}_1\right)\mu_1^{*2}-2\widehat{p}_2\mu_1^*\mu_2^*-\widehat{p}_2\sigma_1^{*2}+\widehat{p}_2\sigma_2^{*2}+\left(\widehat{p}_1\mu_1^*+\widehat{p}_2\mu_2^*\right)^2\right) \\
&= \frac{\widehat{p}_1}{n-1}\left(\left(\widehat{p}_2-\widehat{p}_1\right)\mu_1^{*2}-2\widehat{p}_2\mu_1^*\mu_2^*+\widehat{p}_2\Delta\sigma^2+\left(\widehat{p}_1\mu_1^*+\widehat{p}_2\mu_2^*\right)^2\right).
\end{aligned}$$

Similarly, $\frac{A_2}{n-1}$ contains a remainder term denoted by $\epsilon_{n,2}$. This one is equal to $\epsilon_{n,1}$, but with the indices changed from 1 to 2 and vice versa. When the number of measurements is large enough, $\epsilon_{n,1}$ and $\epsilon_{n,2}$ become negligible.

3.4.2. Subtracting the true variance

Adding both $\frac{A_1}{n-1}$ and $\frac{A_2}{n-1}$ together and subtracting the true variance leads to the desired expression for the bias:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\mathrm{Var}}[X]\right] - \sigma^{*2} &= \frac{A_1}{n-1} + \frac{A_2}{n-1} - \left(p_1^*\sigma_1^{*2} + p_2^*\sigma_2^{*2} + p_1^*\mu_1^{*2} + p_2^*\mu_2^{*2} - \left(p_1^*\mu_1^* + p_2^*\mu_2^*\right)^2\right) \\
&= \left(\widehat{p}_1 - p_1^*\right)\sigma_1^{*2} + \left(\widehat{p}_2 - p_2^*\right)\sigma_2^{*2} + \left(\widehat{p}_1\left(1-2\widehat{p}_1\right) - p_1^*\right)\mu_1^{*2} + \left(\widehat{p}_2\left(1-2\widehat{p}_2\right) - p_2^*\right)\mu_2^{*2} \\
&\quad - 4\widehat{p}_1\widehat{p}_2\mu_1^*\mu_2^* + \left(\widehat{p}_1 + \widehat{p}_2 + 1\right)\left(\widehat{p}_1\mu_1^* + \widehat{p}_2\mu_2^*\right)^2 + \epsilon_n \\
&= -\Delta p\sigma_1^{*2} + \Delta p\sigma_2^{*2} + \left(-\Delta p - 2\widehat{p}_1^2\right)\mu_1^{*2} + \left(\Delta p - 2\widehat{p}_2^2\right)\mu_2^{*2} \\
&\quad - 4\widehat{p}_1\widehat{p}_2\mu_1^*\mu_2^* + 2\left(\widehat{p}_1^2\mu_1^{*2} + \widehat{p}_2^2\mu_2^{*2} + 2\widehat{p}_1\widehat{p}_2\mu_1^*\mu_2^*\right) + \epsilon_n \\
&= \Delta p\left(\Delta\sigma^2 + \Delta\mu^2\right) + \epsilon_n.
\end{aligned}$$

Note the previously defined differences of Definition 2.1. The remainder term ϵ_n is defined as the sum of $\epsilon_{n,1}$ and $\epsilon_{n,2}$ and can be rewritten as:

$$\begin{aligned}
\epsilon_n &:= \epsilon_{n,1} + \epsilon_{n,2} \\
&= \frac{\widehat{p}_1}{n-1}\left(\left(\widehat{p}_2 - \widehat{p}_1\right)\mu_1^{*2} - 2\widehat{p}_2\mu_1^*\mu_2^* + \widehat{p}_2\Delta\sigma^2 + \left(\widehat{p}_1\mu_1^* + \widehat{p}_2\mu_2^*\right)^2\right) \\
&+ \frac{\widehat{p}_2}{n-1}\left(\left(\widehat{p}_1 - \widehat{p}_2\right)\mu_2^{*2} - 2\widehat{p}_1\mu_1^*\mu_2^* + \widehat{p}_1\Delta\sigma^2 + \left(\widehat{p}_1\mu_1^* + \widehat{p}_2\mu_2^*\right)^2\right) \\
&= \frac{1}{n-1}\left(\widehat{p}_1\left(\left(1-\widehat{p}_1\right) - \widehat{p}_1\right)\mu_1^{*2} - 4\widehat{p}_1\widehat{p}_2\mu_1^*\mu_2^* + \widehat{p}_2\left(\left(1-\widehat{p}_2\right) - \widehat{p}_2\right)\mu_2^{*2}\right. \\
&\qquad \left. + 2\widehat{p}_1\widehat{p}_2\Delta\sigma^2 + \left(\widehat{p}_1 + \widehat{p}_2\right)\left(\widehat{p}_1\mu_1^* + \widehat{p}_2\mu_2^*\right)^2\right) \\
&= \frac{1}{n-1}\left(-2\widehat{p}_1^2\mu_1^{*2} - 4\widehat{p}_1\widehat{p}_2\mu_1^*\mu_2^* - 2\widehat{p}_2^2\mu_2^{*2} + \widehat{p}_1\mu_1^{*2} + \widehat{p}_2\mu_2^{*2} + 2\widehat{p}_1\widehat{p}_2\Delta\sigma^2 + \left(\widehat{p}_1\mu_1^* + \widehat{p}_2\mu_2^*\right)^2\right) \\
&= \frac{1}{n-1}\left(-2\left(\widehat{p}_1\mu_1^* + \widehat{p}_2\mu_2^*\right)^2 + \widehat{p}_1\mu_1^{*2} + \widehat{p}_2\mu_2^{*2} + 2\widehat{p}_1\widehat{p}_2\Delta\sigma^2 + \left(\widehat{p}_1\mu_1^* + \widehat{p}_2\mu_2^*\right)^2\right) \\
&= \frac{1}{n-1}\left(\widehat{p}_1\mu_1^{*2} + \widehat{p}_2\mu_2^{*2} - \left(\widehat{p}_1\mu_1^* + \widehat{p}_2\mu_2^*\right)^2 + 2\widehat{p}_1\widehat{p}_2\Delta\sigma^2\right).
\end{aligned}$$

Remark that ϵ_n goes to zero when the number of measurements becomes large enough. □

3.5. Bias of $\widehat{\gamma}_{p_1,p_2}$ and $\widehat{\psi}_{p_1,p_2}$ – Proof

In this section, Theorem 3.2 and Theorem 3.6 are proved. These theorems are about the biases of the estimators $\widehat{\gamma}_{p_1,p_2}$ and $\widehat{\psi}_{p_1,p_2}$, respectively. The estimators are defined in Equation (3.2) and Equation (3.9).

3.5.1. Proving Theorem 3.2

Let us recall Theorem 3.2. The proof of this theorem is given in the current subsection. Corollary 3.3, which follows from this theorem, has already been proved in Section 3.1.

Proof of Theorem 3.2. By definition, the bias of the estimator is equal to its expectation minus the true variance. Using Lemmas 3.15 and 3.22, the expectation of $\widehat{\gamma}_{p_1,p_2}$ is:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\gamma}_{p_1,p_2}\right] &= p_1\mathbb{E}\left[\widehat{\sigma}_1^2\right] + p_2\mathbb{E}\left[\widehat{\sigma}_2^2\right] + p_1\mathbb{E}\left[\widehat{\mu}_1^2\right] + p_2\mathbb{E}\left[\widehat{\mu}_2^2\right] - \mathbb{E}\left[\widehat{\theta}_{p_1,p_2}^2\right] \\
&= p_1\sigma_1^{*2} + p_2\sigma_2^{*2} + p_1\left(\frac{\sigma_1^{*2}}{n_1} + \mu_1^{*2}\right) + p_2\left(\frac{\sigma_2^{*2}}{n_2} + \mu_2^{*2}\right) - \frac{p_1^2}{n_1}\sigma_1^{*2} - \frac{p_2^2}{n_2}\sigma_2^{*2} - \left(p_1\mu_1^* + p_2\mu_2^*\right)^2,
\end{aligned}$$

by the linearity of the expectation. This can be further simplified to:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\gamma}_{p_1,p_2}\right] &= \frac{n_1 + 1 - p_1}{n_1}p_1\sigma_1^{*2} + \frac{n_2 + 1 - p_2}{n_2}p_2\sigma_2^{*2} + p_1\mu_1^{*2} + p_2\mu_2^{*2} - \left(p_1\mu_1^* + p_2\mu_2^*\right)^2 \\
&= \frac{n_1 + p_2}{n_1}p_1\sigma_1^{*2} + \frac{n_2 + p_1}{n_2}p_2\sigma_2^{*2} + p_1\mu_1^{*2} + p_2\mu_2^{*2} - \left(p_1\mu_1^* + p_2\mu_2^*\right)^2.
\end{aligned}$$

Subtracting the true variance σ^{*2} gives:

$$\begin{aligned}\mathbb{E}\left[\widehat{\gamma}_{p_1,p_2}\right]-\sigma^{*2} &= \frac{n_1+p_2}{n_1}p_1\sigma_1^{*2}+\frac{n_2+p_1}{n_2}p_2\sigma_2^{*2}+p_1\mu_1^{*2}+p_2\mu_2^{*2}-\left(p_1\mu_1^*+p_2\mu_2^*\right)^2\\ &\quad-\left(\sigma^{*2}=p_1^*\sigma_1^{*2}+p_2^*\sigma_2^{*2}+p_1^*\mu_1^{*2}+p_2^*\mu_2^{*2}-\left(p_1^*\mu_1^*+p_2^*\mu_2^*\right)^2\right)\\ &= \left(p_1\frac{n_1+p_2}{n_1}-p_1^*\right)\sigma_1^{*2}+\left(p_2\frac{n_2+p_1}{n_2}-p_2^*\right)\sigma_2^{*2}+\left(p_1-p_1^*\right)\mu_1^{*2}+\left(p_2-p_2^*\right)\mu_2^{*2}\\ &\quad-\left(p_1\mu_1^*+p_2\mu_2^*\right)^2+\left(p_1^*\mu_1^*+p_2^*\mu_2^*\right)^2.\end{aligned}$$

This expression can be simplified with the differences of Defintion 2.1, resulting in:

$$\begin{aligned}\mathbb{E}\left[\widehat{\gamma}_{p_1,p_2}\right]-\sigma^{*2} &= \left(\frac{p_1p_2}{n_1}-\widetilde{\Delta}p\right)\sigma_1^{*2}+\left(\frac{p_1p_2}{n_2}+\widetilde{\Delta}p\right)\sigma_2^{*2}-\widetilde{\Delta}p\mu_1^{*2}+\widetilde{\Delta}p\mu_2^{*2}+\mu^{*2}-\left(p_1\mu_1^*+p_2\mu_2^*\right)^2\\ &= \widetilde{\Delta}p\left(\Delta\sigma^2+\Delta\mu^2\right)+\mu^{*2}-\left(p_1\mu_1^*+p_2\mu_2^*\right)^2+\frac{p_1p_2}{n_1}\sigma_1^{*2}+\frac{p_1p_2}{n_2}\sigma_2^{*2},\end{aligned}\tag{3.21}$$

giving the desired expression for the bias of the estimator $\widehat{\gamma}_{p_1,p_2}$. □

3.5.2. Proving Theorem 3.6

In order to obtain an unbiased estimator for σ^{*2}, Corollary 3.4 and Corollary 3.5 were used. These have already been proved in Section 3.1. The current subsection contains the proof of Theorem 3.6 about the bias of the estimator $\widehat{\psi}_{p_1,p_2}$, defined in Equation (3.9).

Proof of Theorem 3.6. We can rewrite $\widehat{\psi}_{p_1,p_2}$ as

$$\begin{aligned}\widehat{\psi}_{p_1,p_2} &= p_1\widehat{\sigma}_1^2+p_2\widehat{\sigma}_2^2+p_1\widehat{\mu}_1^2+p_2\widehat{\mu}_2^2-\widehat{\theta}_{p_1,p_2}^2-\frac{p_1\left(1-p_1\right)}{n_1}\widehat{\sigma}_1^2-\frac{p_2\left(1-p_2\right)}{n_2}\widehat{\sigma}_2^2\\ &= p_1\widehat{\sigma}_1^2+p_2\widehat{\sigma}_2^2+p_1\left(\widehat{\mu}_1^2-\frac{\widehat{\sigma}_1^2}{n_1}\right)+p_2\left(\widehat{\mu}_2^2-\frac{\widehat{\sigma}_2^2}{n_2}\right)-\widehat{\theta}_{p_1,p_2}^2+\frac{p_1^2}{n_1}\widehat{\sigma}_1^2+\frac{p_2^2}{n_2}\widehat{\sigma}_2^2.\end{aligned}$$

Using Corollary 3.4 and Corollary 3.5, the expectation of $\widehat{\psi}_{p_1,p_2}$ is:

$$\begin{aligned}\mathbb{E}\left[\widehat{\psi}_{p_1,p_2}\right] &= p_1\mathbb{E}\left[\widehat{\sigma}_1^2\right]+p_2\mathbb{E}\left[\widehat{\sigma}_2^2\right]+p_1\mathbb{E}\left[\widehat{\mu}_1^2-\frac{\widehat{\sigma}_1^2}{n_1}\right]+p_2\mathbb{E}\left[\widehat{\mu}_2^2-\frac{\widehat{\sigma}_2^2}{n_2}\right]-\mathbb{E}\left[\widehat{\theta}_{p_1,p_2}^2\right]+\frac{p_1^2}{n_1}\mathbb{E}\left[\widehat{\sigma}_1^2\right]+\frac{p_2^2}{n_2}\mathbb{E}\left[\widehat{\sigma}_2^2\right]\\ &= p_1\sigma_1^{*2}+p_2\sigma_2^{*2}+p_1\mu_1^{*2}+p_2\mu_2^{*2}-\frac{p_1^2}{n_1}\sigma_1^{*2}-\frac{p_2^2}{n_2}\sigma_2^{*2}-\left(p_1\mu_1^*+p_2\mu_2^*\right)^2+\frac{p_1^2}{n_1}\sigma_1^{*2}+\frac{p_2^2}{n_2}\sigma_2^{*2}\\ &= p_1\sigma_1^{*2}+p_2\sigma_2^{*2}+p_1\mu_1^{*2}+p_2\mu_2^{*2}-\left(p_1\mu_1^*+p_2\mu_2^*\right)^2.\end{aligned}$$

Subtracting the true variance leads to a bias of:

$$\begin{aligned}\mathbb{E}\left[\widehat{\psi}_{p_1,p_2}\right]-\sigma^{*2} &= p_1\sigma_1^{*2}+p_2\sigma_2^{*2}+p_1\mu_1^{*2}+p_2\mu_2^{*2}-\left(p_1\mu_1^*+p_2\mu_2^*\right)^2\\ &\quad-\left(p_1^*\sigma_1^{*2}+p_2^*\sigma_2^{*2}+p_1^*\mu_1^{*2}+p_2^*\mu_2^{*2}-\left(p_1^*\mu_1^*+p_2^*\mu_2^*\right)^2\right).\end{aligned}$$

Applying the differences from Definition 2.1 results in:

$$\begin{aligned}\mathbb{E}\left[\widehat{\psi}_{p_1,p_2}\right]-\sigma^{*2} &= -\widetilde{\Delta}p\sigma_1^{*2}+\widetilde{\Delta}p\sigma_2^{*2}-\widetilde{\Delta}p\mu_1^{*2}+\widetilde{\Delta}p\mu_2^{*2}+\left(p_1^*\mu_1^*+p_2^*\mu_2^*\right)^2-\left(p_1\mu_1^*+p_2\mu_2^*\right)^2\\ &= \widetilde{\Delta}p\left(\Delta\sigma^2+\Delta\mu^2\right)+\mu^{*2}-\left(p_1\mu_1^*+p_2\mu_2^*\right)^2,\end{aligned}$$

which is the desired expression for the bias. □

3.6. Bias of $\widehat{\delta}_{p_1,p_2}$ – Proof

In this section, the bias of $\widehat{\delta}_{p_1,p_2}$ is computed for two different plug-in estimators, proving Theorem 3.8 and Theorem 3.9. Firstly, the estimator $\widehat{\theta}_{p_1,p_2}$ is used as a plug-in estimator, resulting in the estimator $\widehat{\delta}_{p_1,p_2,\widehat{\theta}_{p_1,p_2}}$ as defined in Equation (3.12). Secondly, the estimator $\overline{X}$ is applied as plug-in estimator, such that the estimator for σ^{*2} is denoted by $\widehat{\delta}_{p_1,p_2,\overline{X}}$ as in Equation (3.13).

3.6.1. Proving Theorem 3.8

In this subsection, the bias of $\widehat{\delta}_{p_1,p_2,\widehat{\theta}_{p_1,p_2}}$ is derived, thereby proving Theorem 3.8.

Proof of Theorem 3.8. The bias is by definition equal to the expected value of the estimator minus the true variance σ^{*2}. The expectation of $\widehat{\delta}_{p_1,p_2}$ can be split up in the following way:

$$\mathbb{E}\left[\widehat{\delta}_{p_1,p_2,\widehat{\theta}_{p_1,p_2}}\right] = p_1\sigma_1^{*2} + p_2\sigma_2^{*2} + p_1B_1 + p_2B_2,$$

where B_1 is the following:

$$\begin{aligned}
B_1 &:= \mathbb{E}\left[\left(\widehat{\mu}_1 - \widehat{\theta}_{p_1,p_2}\right)^2\right] \\
&= \mathbb{E}\left[\widehat{\mu}_1^2\right] - 2\mathbb{E}\left[\left(\frac{1}{n_1}\sum_{i=1}^{n_1} X_{i,1}\right)\left(\frac{p_1}{n_1}\sum_{i=1}^{n_1} X_{i,1} + \frac{p_2}{n_2}\sum_{i=1}^{n_2} X_{i,2}\right)\right] + \mathbb{E}\left[\widehat{\theta}_{p_1,p_2}^2\right] \\
&= \mathbb{E}\left[\widehat{\mu}_1^2\right] - \frac{2p_1}{n_1^2}\mathbb{E}\left[\left(\sum_{i=1}^{n_1} X_{i,1}\right)^2\right] - \frac{2p_2}{n_1n_2}\mathbb{E}\left[\left(\sum_{i=1}^{n_1} X_{i,1}\right)\left(\sum_{i=1}^{n_2} X_{i,2}\right)\right] + \mathbb{E}\left[\widehat{\theta}_{p_1,p_2}^2\right] \\
&= \mathbb{E}\left[\widehat{\mu}_1^2\right] - \frac{2p_1}{n_1^2}\left(n_1\mathbb{E}\left[X_1^2\right] + n_1\left(n_1-1\right)\mathbb{E}\left[X_1\right]^2\right) - \frac{2p_2}{n_1n_2}\sum_{i=1}^{n_1}\mathbb{E}\left[X_{i,1}\right]\sum_{i=1}^{n_2}\mathbb{E}\left[X_{i,2}\right] + \mathbb{E}\left[\widehat{\theta}_{p_1,p_2}^2\right].
\end{aligned}$$

Using Equation (3.19) and Lemma 3.22, the previous equation can be written as:

$$\begin{aligned}
B_1 &= \frac{\sigma_1^{*2}}{n_1} + \mu_1^{*2} - \frac{2p_1}{n_1}\left(\sigma_1^{*2} + n_1\mu_1^{*2}\right) - 2p_2\mu_1^*\mu_2^* + \frac{p_1^2}{n_1}\left(\sigma_1^{*2} + n_1\mu_1^{*2}\right) + 2p_1p_2\mu_1^*\mu_2^* + \frac{p_2^2}{n_2}\left(\sigma_2^{*2} + n_2\mu_2^{*2}\right) \\
&= \frac{p_1^2 - 2p_1 + 1}{n_1}\left(\sigma_1^{*2} + n_1\mu_1^{*2}\right) + 2\left(p_1-1\right)p_2\mu_1^*\mu_2^* + \frac{p_2^2}{n_2}\left(\sigma_2^{*2} + n_2\mu_2^{*2}\right).
\end{aligned}$$

This expression can be factorised and simplified to:

$$\begin{aligned}
B_1 &= \frac{\left(p_1-1\right)^2}{n_1}\left(\sigma_1^{*2} + n_1\mu_1^{*2}\right) - 2p_2^2\mu_1^*\mu_2^* + \frac{p_2^2}{n_2}\left(\sigma_2^{*2} + n_2\mu_2^{*2}\right) \\
&= \frac{p_2^2}{n_1}\left(\sigma_1^{*2} + n_1\mu_1^{*2}\right) - 2p_2^2\mu_1^*\mu_2^* + \frac{p_2^2}{n_2}\left(\sigma_2^{*2} + n_2\mu_2^{*2}\right) \\
&= p_2^2\left(\frac{\sigma_1^{*2}}{n_1} + \frac{\sigma_2^{*2}}{n_2} + \left(\mu_1^* - \mu_2^*\right)^2\right) \\
&= p_2^2\left(\frac{\sigma_1^{*2}}{n_1} + \frac{\sigma_2^{*2}}{n_2} + \left(\Delta\mu\right)^2\right).
\end{aligned}$$

By symmetry, B_2 is equal, but with all indices 1 and 2 switched.

Subtracting the true variance leads to the bias of $\widehat{\delta}_{p_1,p_2,\widehat{\theta}_{p_1,p_2}}$:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\delta}_{p_1,p_2,\widehat{\theta}_{p_1,p_2}}\right] - \sigma^{*2} &= \left(p_1 - p_1^*\right)\sigma_1^{*2} + \left(p_2 - p_2^*\right)\sigma_2^{*2} + p_1B_1 + p_2B_2 - p_1^*\mu_1^{*2} - p_2^*\mu_2^{*2} + \left(p_1^*\mu_1^* + p_2^*\mu_2^*\right)^2 \\
&= \left(p_1 - p_1^*\right)\sigma_1^{*2} + \left(p_2 - p_2^*\right)\sigma_2^{*2} + p_1\frac{p_1^2 - 2p_1 + 1}{n_1}\left(\sigma_1^{*2} + n_1\mu_1^{*2}\right) \\
&\quad + 2p_1\left(p_1-1\right)p_2\mu_1^*\mu_2^* + \frac{p_1p_2^2}{n_2}\left(\sigma_2^{*2} + n_2\mu_2^{*2}\right) + p_2\frac{p_2^2 - 2p_2 + 1}{n_2}\left(\sigma_2^{*2} + n_2\mu_2^{*2}\right) \\
&\quad + 2p_2\left(p_2-1\right)p_1\mu_1^*\mu_2^* + \frac{p_1^2p_2}{n_1}\left(\sigma_1^{*2} + n_1\mu_1^{*2}\right) - p_1^*\mu_1^{*2} - p_2^*\mu_2^{*2} + \left(p_1^*\mu_1^* + p_2^*\mu_2^*\right)^2.
\end{aligned}$$

We remark that the bias can be rewritten as:

$$\begin{aligned}
&\left(p_1\frac{p_1^2 + p_1\left(p_2-2\right) + 1}{n_1} + p_1 - p_1^*\right)\sigma_1^{*2} + \left(p_2\frac{p_2^2 + p_2\left(p_1-2\right) + 1}{n_2} + p_2 - p_2^*\right)\sigma_2^{*2} \\
&+ \left(p_1^3 + p_1^2\left(p_2-2\right) + p_1 - p_1^*\right)\mu_1^{*2} + \left(p_2^3 + p_2^2\left(p_1-2\right) + p_2 - p_2^*\right)\mu_2^{*2} \\
&+ 2p_1p_2\left(p_1 + p_2 - 2\right)\mu_1^*\mu_2^* + \left(p_1^*\mu_1^* + p_2^*\mu_2^*\right)^2.
\end{aligned}$$

Remember the differences defined in Definition 2.1. With these, the bias can be simplified to:

$$\begin{aligned}
&\left(p_1\frac{p_1^2+p_1(-p_1-1)+1}{n_1}-\widetilde{\Delta}p\right)\sigma_1^{*2}+\left(p_2\frac{p_2^2+p_2(-p_2-1)+1}{n_2}+\widetilde{\Delta}p\right)\sigma_2^{*2}\\
&+\left(p_1^3+p_1^2(-p_1-1)+p_1-p_1^*\right)\mu_1^{*2}+\left(p_2^3+p_2^2(-p_2-1)+p_2-p_2^*\right)\mu_2^{*2}\\
&-2p_1p_2\mu_1^*\mu_2^*+\left(p_1^*\mu_1^*+p_2^*\mu_2^*\right)^2\\
&=\left(p_1\frac{-p_1+1}{n_1}-\widetilde{\Delta}p\right)\sigma_1^{*2}+\left(p_2\frac{-p_2+1}{n_2}+\widetilde{\Delta}p\right)\sigma_2^{*2}+\left(-p_1^2-\widetilde{\Delta}p\right)\mu_1^{*2}\\
&+\left(-p_2^2+\widetilde{\Delta}p\right)\mu_2^{*2}-2p_1p_2\mu_1^*\mu_2^*+\left(p_1^*\mu_1^*+p_2^*\mu_2^*\right)^2.
\end{aligned}$$

Further simplifying this equation leads to the desired expression for the bias of the estimator $\widehat{\delta}_{p_1,p_2,\widehat{\theta}_{p_1,p_2}}$:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\delta}_{p_1,p_2,\widehat{\theta}_{p_1,p_2}}\right]-\sigma^{*2}&=\left(\frac{p_1p_2}{n_1}-\widetilde{\Delta}p\right)\sigma_1^{*2}+\left(\frac{p_1p_2}{n_2}+\widetilde{\Delta}p\right)\sigma_2^{*2}-\widetilde{\Delta}p\mu_1^{*2}+\widetilde{\Delta}p\mu_2^{*2}\\
&\quad-\left(p_1\mu_1^*+p_2\mu_2^*\right)^2+\left(p_1^*\mu_1^*+p_2^*\mu_2^*\right)^2\\
&=\widetilde{\Delta}p\left(\Delta\sigma^2+\Delta\mu^2\right)+\mu^{*2}-\left(p_1\mu_1^*+p_2\mu_2^*\right)^2+\frac{p_1p_2}{n_1}\sigma_1^{*2}+\frac{p_1p_2}{n_2}\sigma_2^{*2}.
\end{aligned}$$

□

3.6.2. Proving Theorem 3.9

Let us recall Theorem 3.9. It states the bias of $\widehat{\delta}_{p_1,p_2,\overline{X}}$. This estimator is defined in Equation (3.13) and uses $\overline{X}$ as a plug-in estimator. The proof of this theorem is given below.

Proof of Theorem 3.9. The bias of an estimator is defined by its expectation minus the true parameter σ^{*2}. We remark that the expectation can be split up in the following way:

$$\mathbb{E}\left[\widehat{\delta}_{p_1,p_2,\overline{X}}\right]=p_1\sigma_1^{*2}+p_2\sigma_2^{*2}+p_1C_1+p_2C_2,$$

with C_1 given below:

$$\begin{aligned}
C_1&:=\mathbb{E}\left[\left(\widehat{\mu}_1-\overline{X}\right)^2\right]\\
&=\mathbb{E}\left[\widehat{\mu}_1^2\right]-2\mathbb{E}\left[\left(\frac{1}{n_1}\sum_{i=1}^{n_1}X_{i,1}\right)\frac{1}{n}\left(\sum_{i=1}^{n_1}X_{i,1}+\sum_{i=1}^{n_2}X_{i,2}\right)\right]+\mathbb{E}\left[\overline{X}^2\right]\\
&=\mathbb{E}\left[\widehat{\mu}_1^2\right]-\frac{2}{n_1n}\mathbb{E}\left[\left(\sum_{i=1}^{n_1}X_{i,1}\right)^2+\left(\sum_{i=1}^{n_1}X_{i,1}\right)\left(\sum_{i=1}^{n_2}X_{i,2}\right)\right]+\mathbb{E}\left[\overline{X}^2\right]\\
&=\mathbb{E}\left[\widehat{\mu}_1^2\right]-\frac{2}{n_1n}\left(n_1\mathbb{E}\left[X_{i,1}^2\right]+n_1(n_1-1)\mathbb{E}\left[X_{i,1}\right]^2+\sum_{i=1}^{n_1}\mathbb{E}\left[X_{i,1}\right]\sum_{i=1}^{n_2}\mathbb{E}\left[X_{i,2}\right]\right)+\mathbb{E}\left[\overline{X}^2\right].
\end{aligned}$$

Recall Equation 3.19 and Lemma 3.21. Hence, we can rewrite the equation above as:

$$\begin{aligned}
C_1&=\frac{\sigma_1^{*2}}{n_1}+\mu_1^{*2}-\frac{2}{n}\left(\sigma_1^{*2}+n_1\mu_1^{*2}\right)-\frac{2n_1n_2}{nn_1}\mu_1^*\mu_2^*+\frac{\widehat{p}_1}{n}\sigma_1^{*2}+\frac{\widehat{p}_2}{n}\sigma_2^{*2}+\left(\widehat{p}_1\mu_1^*+\widehat{p}_2\mu_2^*\right)^2\\
&=\frac{n_2-n_1}{n}\left(\frac{\sigma_1^{*2}}{n_1}+\mu_1^{*2}\right)-2\widehat{p}_2\mu_1^*\mu_2^*+\frac{\widehat{p}_1}{n}\sigma_1^{*2}+\frac{\widehat{p}_2}{n}\sigma_2^{*2}+\left(\widehat{p}_1\mu_1^*+\widehat{p}_2\mu_2^*\right)^2\\
&=\left(\widehat{p}_2-\widehat{p}_1\right)\left(\frac{\sigma_1^{*2}}{n_1}+\mu_1^{*2}\right)-2\widehat{p}_2\mu_1^*\mu_2^*+\frac{\widehat{p}_1}{n}\sigma_1^{*2}+\frac{\widehat{p}_2}{n}\sigma_2^{*2}+\left(\widehat{p}_1\mu_1^*+\widehat{p}_2\mu_2^*\right)^2\\
&=\left(\widehat{p}_2-\widehat{p}_1\right)\mu_1^{*2}-2\widehat{p}_2\mu_1^*\mu_2^*+\left(\widehat{p}_1\mu_1^*+\widehat{p}_2\mu_2^*\right)^2+\left(\frac{\widehat{p}_2-\widehat{p}_1}{\widehat{p}_1}+\widehat{p}_1\right)\frac{\sigma_1^{*2}}{n}+\widehat{p}_2\frac{\sigma_2^{*2}}{n}.
\end{aligned}$$

By symmetry, C_2 is equal to C_1, but with the indices 1 and 2 swapped.

Subtracting the true value of the variance leads to the bias of $\widehat{\delta}_{p_1,p_2,\overline{X}}$:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\delta}_{p_1,p_2,\overline{X}}\right] - \sigma^{*2} &= (p_1 - p_1^*)\sigma_1^{*2} + (p_2 - p_2^*)\sigma_2^{*2} + p_1C_1 + p_2C_2 - p_1^*\mu_1^{*2} - p_2^*\mu_2^{*2} + (p_1^*\mu_1^* + p_2^*\mu_2^*)^2 \\
&= (p_1 - p_1^*)\sigma_1^{*2} + (p_2 - p_2^*)\sigma_2^{*2} + p_1(\widehat{p}_2 - \widehat{p}_1)\mu_1^{*2} - 2p_1\widehat{p}_2\mu_1^*\mu_2^* + p_1(\widehat{p}_1\mu_1^* + \widehat{p}_2\mu_2^*)^2 \\
&+ p_1\left(\frac{\widehat{p}_2 - \widehat{p}_1}{\widehat{p}_1} + \widehat{p}_1\right)\frac{\sigma_1^{*2}}{n} + p_1\widehat{p}_2\frac{\sigma_2^{*2}}{n} + p_2(\widehat{p}_1 - \widehat{p}_2)\mu_2^{*2} - 2\widehat{p}_1p_2\mu_1^*\mu_2^* + p_2(\widehat{p}_1\mu_1^* + \widehat{p}_2\mu_2^*)^2 \\
&+ p_2\left(\frac{\widehat{p}_1 - \widehat{p}_2}{\widehat{p}_2} + \widehat{p}_2\right)\frac{\sigma_2^{*2}}{n} + p_2\widehat{p}_1\frac{\sigma_1^{*2}}{n} - p_1^*\mu_1^{*2} - p_2^*\mu_2^{*2} + (p_1^*\mu_1^* + p_2^*\mu_2^*)^2 \\
&= (p_1 - p_1^*)\sigma_1^{*2} + (p_2 - p_2^*)\sigma_2^{*2} + \left(\widehat{p}_2 - (1-\widehat{p}_2) + (1-\widehat{p}_2)^2\right)\sigma_1^{*2} + \widehat{p}_1\widehat{p}_2\sigma_1^{*2} \\
&+ \left(\widehat{p}_1 - (1-\widehat{p}_1) + (1-\widehat{p}_1)^2\right)\sigma_2^{*2} + \widehat{p}_1\widehat{p}_2\sigma_2^{*2} + p_1(\widehat{p}_2 - \widehat{p}_1)\mu_1^{*2} - 2(p_1\widehat{p}_2 + \widehat{p}_1p_2)\mu_1^*\mu_2^* \\
&+ p_2(\widehat{p}_1 - \widehat{p}_2)\mu_2^{*2} + (\widehat{p}_1\mu_1^* + \widehat{p}_2\mu_2^*)^2 - p_1^*\mu_1^{*2} - p_2^*\mu_2^{*2} + (p_1^*\mu_1^* + p_2^*\mu_2^*)^2.
\end{aligned}$$

Remember that $\widehat{p}_1 + \widehat{p}_2 = 1$. This is used to rewrite the bias as the desired expression:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\delta}_{p_1,p_2,\overline{X}}\right] - \sigma^{*2} &= -\widetilde{\Delta}p\sigma_1^{*2} + \widetilde{\Delta}p\sigma_2^{*2} + \widehat{p}_2^2\sigma_1^{*2} + \widehat{p}_1\widehat{p}_2\sigma_1^{*2} + \widehat{p}_1^2\sigma_2^{*2} + \widehat{p}_1\widehat{p}_2\sigma_2^{*2} \\
&+ (p_1(\widehat{p}_2 - \widehat{p}_1) - p_1^*)\mu_1^{*2} - 2(p_1\widehat{p}_2 + \widehat{p}_1p_2)\mu_1^*\mu_2^* + (p_2(\widehat{p}_1 - \widehat{p}_2) - p_2^*)\mu_2^{*2} \\
&+ (\widehat{p}_1\mu_1^* + \widehat{p}_2\mu_2^*)^2 + (p_1^*\mu_1^* + p_2^*\mu_2^*)^2 \\
&= \widetilde{\Delta}p\Delta\sigma^2 + \widehat{p}_2(\widehat{p}_1 + \widehat{p}_2)\sigma_1^{*2} + \widehat{p}_1(\widehat{p}_1 + \widehat{p}_2)\sigma_2^{*2} + (p_1(1-2\widehat{p}_1) - p_1^*)\mu_1^{*2} \\
&- 2(p_1\widehat{p}_2 + \widehat{p}_1p_2)\mu_1^*\mu_2^* + (p_2(1-2\widehat{p}_2) - p_2^*)\mu_2^{*2} + (\widehat{p}_1\mu_1^* + \widehat{p}_2\mu_2^*)^2 + (p_1^*\mu_1^* + p_2^*\mu_2^*)^2 \\
&= (\widehat{p}_2 - \widetilde{\Delta}p)\sigma_1^{*2} + (\widehat{p}_1 + \widetilde{\Delta}p)\sigma_2^{*2} - (2p_1\widehat{p}_1 + \widetilde{\Delta}p)\mu_1^{*2} - 2(p_1\widehat{p}_2 + \widehat{p}_1p_2)\mu_1^*\mu_2^* \\
&- (2p_2\widehat{p}_2 - \widetilde{\Delta}p)\mu_2^{*2} + \mu^{*2} + (\widehat{p}_1\mu_1^* + \widehat{p}_2\mu_2^*)^2 \\
&= \widetilde{\Delta}p\Delta\sigma^2 + \widetilde{\Delta}p\Delta\mu^2 + \widehat{p}_2\sigma_1^{*2} + \widehat{p}_1\sigma_2^{*2} - 2p_1\widehat{p}_1\mu_1^{*2} - 2(p_1\widehat{p}_2 + \widehat{p}_1p_2)\mu_1^*\mu_2^* \\
&- 2p_2\widehat{p}_2\mu_2^{*2} + \mu^{*2} + (\widehat{p}_1\mu_1^* + \widehat{p}_2\mu_2^*)^2 \\
&= \widetilde{\Delta}p(\Delta\sigma^2 + \Delta\mu^2) + \widehat{p}_2\sigma_1^{*2} + \widehat{p}_1\sigma_2^{*2} + \mu^{*2} - 2(p_1\mu_1^* + p_2\mu_2^*)(\widehat{p}_1\mu_1^* + \widehat{p}_2\mu_2^*) + (\widehat{p}_1\mu_1^* + \widehat{p}_2\mu_2^*)^2 \\
&= \widetilde{\Delta}p(\Delta\sigma^2 + \Delta\mu^2) + \mu^{*2} + ((\widehat{p}_1 - 2p_1)\mu_1^* + (\widehat{p}_2 - 2p_2)\mu_2^*)(\widehat{p}_1\mu_1^* + \widehat{p}_2\mu_2^*) + \widehat{p}_2\sigma_1^{*2} + \widehat{p}_1\sigma_2^{*2},
\end{aligned}$$

which concludes the proof. □

3.7. Variance of $\widehat{\mathrm{Var}}[X]$ – Proof

Let us recall Theorem 3.10 about the variance of the naive estimator $\widehat{\mathrm{Var}}[X]$. It is proved in this section.

3.7.1. Main idea of the proof

Proof of Theorem 3.10. Let $\widetilde{\mu} := \widehat{p}_1\mu_1^* + \widehat{p}_2\mu_2^*$. Let $\widetilde{\sigma}^2 := \widehat{p}_1\sigma_1^{*2} + \widehat{p}_2\sigma_2^{*2}$. Note that

$$\mathrm{Var}[\widehat{\mathrm{Var}}[X]] = \frac{1}{(n-1)^2}\mathrm{Var}\left[\sum_{i=1}^n \left(X_i - \overline{X}\right)^2\right] = \frac{1}{(n-1)^2}\sum_{i,j=1}^n C_{i,j},$$

where

$$C_{i,j} := \mathrm{Cov}\left(\left(X_i - \overline{X}\right)^2, \left(X_j - \overline{X}\right)^2\right).$$

Note that

$$\mathrm{Var}[\overline{X}^2] = \mathrm{Var}\left[\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n X_iX_j\right] = \frac{1}{n^4}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n\sum_{\ell=1}^n \mathrm{Cov}\left(X_iX_j, X_kX_\ell\right) = \mathcal{O}(1/n),$$

since covariances are zero by independence whenever all i, j, k, ℓ are distinct. If $i = j$, then

$$\begin{aligned}
C_{i,i} &= \mathrm{Var}\left(\left(X_i - \overline{X}\right)^2\right) \\
&= \mathrm{Var}\left(X_i^2\right) - 2\mathrm{Cov}(X_i, \overline{X}) + \mathrm{Var}[\overline{X}^2] \\
&= \mathrm{Var}\left(X_i^2\right) - 2\frac{1}{n}\sum_{k=1}^{n}\mathrm{Cov}(X_i, X_k) + \mathcal{O}(1/n) \\
&= \mathrm{Var}\left(X_i^2\right) - 2\frac{1}{n}\mathrm{Var}(X_i) + \mathcal{O}(1/n) = \mathrm{Var}(X_i^2) + \mathcal{O}(1/n).
\end{aligned}$$

By summing up the previous equation over all $i = 1, \ldots, n$, and dividing by n, we obtain

$$\frac{1}{n}\sum_{i=1}^{n} C_{i,i} = \widehat{p}_1(m_{4,1}^* - \sigma_1^{*2}) + \widehat{p}_2(m_{4,2}^* - \sigma_2^{*2}) + \mathcal{O}(1/n).$$

So

$$\mathrm{Var}[\widehat{\mathrm{Var}}[X]] = \frac{\widehat{p}_1(m_{4,1}^* - \sigma_1^{*2}) + \widehat{p}_2(m_{4,2}^* - \sigma_2^{*2})}{n} + \frac{1}{(n-1)^2}\sum_{i\neq j} C_{i,j} + \mathcal{O}(1/n^2).$$

The behavior of $C_{i,j}$ is given in the following lemma, which is proved afterwards.

Lemma 3.23. *For $i \neq j$, we have*

$$\begin{aligned}
C_{i,j} = \frac{2}{n}\Big(&\mathbb{E}[X_j]\left(\mathbb{E}[X_i^2]\mathbb{E}[X_i] + 2\widetilde{\mu}\mathrm{Var}[X_i] - \mathbb{E}[X_i^3]\right) + \mathbb{E}[X_i]\left(\mathbb{E}[X_j^2]\mathbb{E}[X_j] + 2\widetilde{\mu}\mathrm{Var}[X_j] - \mathbb{E}[X_j^3]\right) \\
&+ 2\widetilde{\sigma}^2\left(\mathbb{E}[X_i] - \widetilde{\mu}\right)\left(\mathbb{E}[X_j] - \widetilde{\mu}\right)\Big) + \mathcal{O}(1/n^2).
\end{aligned}$$

As a consequence, when individual i is in subgroup $r \in \{1,2\}$ and individual j is in subgroup $s \in \{1,2\}$, we have

$$\begin{aligned}
n \times C_{i,j} = K_{r,s} := &2\mu_s^*\left((\mu_r^{*2} + \sigma_r^{*2})\mu_r^* + 2\widetilde{\mu}\sigma_r^{*2} - m_{3,r}^*\right) + 2\mu_r^*\left((\mu_s^{*2} + \sigma_s^{*2})\mu_s^* + 2\widetilde{\mu}\sigma_s^{*2} - m_{3,s}^*\right) \\
&+ 4\widetilde{\sigma}^2\left(\mu_r^* - \widetilde{\mu}\right)\left(\mu_s^* - \widetilde{\mu}\right) + \mathcal{O}(1/n)
\end{aligned}$$

Therefore,

$$\mathrm{Var}[\widehat{\mathrm{Var}}[X]] = \frac{1}{n}\left(\widehat{p}_1(m_{4,1}^* - \sigma_1^{*2}) + \widehat{p}_2(m_{4,2}^* - \sigma_2^{*2}) + \widehat{p}_1^2 K_{1,1} + 2\widehat{p}_1\widehat{p}_2 K_{1,2} + \widehat{p}_2^2 K_{2,2}\right) + \mathcal{O}(1/n^2). \tag{3.22}$$

3.7.2. Lemmas and their proofs

In order for Equation (3.22) to hold, we must prove Lemma 3.23. This leads to two other lemmas, which are also proved in this subsection.

Proof of Lemma 3.23.

$$\begin{aligned}
C_{i,j} &= \mathrm{Cov}\left(\left(X_i - \overline{X}\right)^2, \left(X_{j,1} - \overline{X}\right)^2\right) \\
&= \mathrm{Cov}\left(\left(X_i - \widetilde{\mu} + \widetilde{\mu} - \overline{X}\right)^2, \left(X_{j,1} - \widetilde{\mu} + \widetilde{\mu} - \overline{X}\right)^2\right) \\
&= \mathrm{Cov}\left((X_i - \widetilde{\mu})^2 + 2(X_i - \widetilde{\mu})(\widetilde{\mu} - \overline{X}) + (\widetilde{\mu} - \overline{X})^2, (X_j - \widetilde{\mu})^2 + 2(X_j - \widetilde{\mu})(\widetilde{\mu} - \overline{X}) + (\widetilde{\mu} - \overline{X})^2\right) \\
&= \sum_{k=1}^{3}\sum_{\ell=1}^{3} C_{i,j,k,\ell},
\end{aligned}$$

where

$$
\begin{aligned}
C_{i,j,1,1} &:= \mathrm{Cov}\left((X_i-\widetilde{\mu})^2, (X_j-\widetilde{\mu})^2\right),\\
C_{i,j,2,1} &:= 2\mathrm{Cov}\left((X_i-\widetilde{\mu})^2, (X_j-\widetilde{\mu})(\widetilde{\mu}-\overline{X})\right),\\
C_{i,j,3,1} &:= \mathrm{Cov}\left((X_i-\widetilde{\mu})^2, (\widetilde{\mu}-\overline{X})^2\right),\\
C_{i,j,1,2} &:= 2\mathrm{Cov}\left((X_i-\widetilde{\mu})(\widetilde{\mu}-\overline{X}), (X_j-\widetilde{\mu})^2\right),\\
C_{i,j,2,2} &:= 4\mathrm{Cov}\left((X_i-\widetilde{\mu})(\widetilde{\mu}-\overline{X}), (X_j-\widetilde{\mu})(\widetilde{\mu}-\overline{X})\right),\\
C_{i,j,3,2} &:= 2\mathrm{Cov}\left((X_i-\widetilde{\mu})(\widetilde{\mu}-\overline{X}), (\widetilde{\mu}-\overline{X})^2\right),\\
C_{i,j,1,3} &:= \mathrm{Cov}\left((\widetilde{\mu}-\overline{X})^2, (X_j-\widetilde{\mu})^2\right),\\
C_{i,j,2,3} &:= 2\mathrm{Cov}\left((\widetilde{\mu}-\overline{X})^2, (X_j-\widetilde{\mu})(\widetilde{\mu}-\overline{X})\right),\\
C_{i,j,3,3} &= \mathrm{Cov}\left((\widetilde{\mu}-\overline{X})^2, (\widetilde{\mu}-\overline{X})^2\right).
\end{aligned}
$$

By independence, $C_{i,j,1,1}=0$. We know that $\widetilde{\mu}-\overline{X}=\mathcal{O}(1/n)$. This means that the terms $C_{i,j,3,1}, C_{i,j,3,2}, C_{i,j,1,3}$, $C_{i,j,2,3}$ and $C_{i,j,3,3}$ are of the order $\mathcal{O}(1/n^2)$. Note that $C_{i,j,1,2}=C_{j,i,2,1}$. It remains to study $C_{i,j,2,1}$ and $C_{i,j,2,2}$.

The expression of $C_{i,j,2,1}$ and of $C_{i,j,2,2}$ are respectively given by the following lemmas, which conclude the proof. □

Lemma 3.24. *We have*

$$
C_{i,j,2,1} = \frac{2}{n}\mathbb{E}[X_j]\left(\mathbb{E}[X_i^2]\mathbb{E}[X_i] + 2\widetilde{\mu}\mathrm{Var}[X_i] - \mathbb{E}[X_i^3]\right)
$$

Lemma 3.25. *We have*

$$
C_{i,j,2,2} = \frac{4}{n}\widetilde{\sigma}^2\left(\mathbb{E}[X_i]-\widetilde{\mu}\right)\left(\mathbb{E}[X_j]-\widetilde{\mu}\right) + \mathcal{O}(1/n^2).
$$

Proof of Lemma 3.24.

$$
\begin{aligned}
C_{i,j,2,1} &:= 2\mathrm{Cov}\left((X_i-\widetilde{\mu})^2, (X_j-\widetilde{\mu})(\widetilde{\mu}-\overline{X})\right)\\
&= 2\mathrm{Cov}\left((X_i-\widetilde{\mu})^2, (X_j-\widetilde{\mu})\widetilde{\mu}\right) - 2\mathrm{Cov}\left((X_i-\widetilde{\mu})^2, (X_j-\widetilde{\mu})\overline{X}\right)\\
&= -2\mathrm{Cov}\left((X_i-\widetilde{\mu})^2, (X_j-\widetilde{\mu})\overline{X}\right),
\end{aligned}
$$

by independence between X_i and X_j. Therefore,

$$
\begin{aligned}
-C_{i,j,2,1} &= 2\frac{1}{n}\sum_{k=1}^{n}\mathrm{Cov}\left((X_i-\widetilde{\mu})^2, (X_j-\widetilde{\mu})X_k\right)\\
&= 2\frac{1}{n}\mathrm{Cov}\left((X_i-\widetilde{\mu})^2, (X_j-\widetilde{\mu})X_i\right) + 2\frac{1}{n}\mathrm{Cov}\left((X_i-\widetilde{\mu})^2, (X_j-\widetilde{\mu})X_j\right) + 2\frac{n-2}{n}\mathrm{Cov}\left((X_i-\widetilde{\mu})^2, (X_j-\widetilde{\mu})X_k\right).
\end{aligned}
$$

Note that the second and third terms cancel by independence, so that

$$
\begin{aligned}
-(n/2)\times C_{i,j,2,1} &= \mathrm{Cov}\left((X_i-\widetilde{\mu})^2, (X_j-\widetilde{\mu})X_i\right)\\
&= \mathrm{Cov}\left(X_i^2 - 2X_i\widetilde{\mu}+\widetilde{\mu}^2, X_iX_j - X_j\widetilde{\mu}\right)\\
&= \mathrm{Cov}\left(X_i^2 - 2X_i\widetilde{\mu}, X_iX_j - X_j\widetilde{\mu}\right)\\
&= \mathrm{Cov}\left(X_i^2 - 2X_i\widetilde{\mu}, X_iX_j\right),
\end{aligned}
$$

by independence. Therefore

$$
-(n/2)\times C_{i,j,2,1} = \mathrm{Cov}\left(X_i^2, X_iX_j\right) - 2\widetilde{\mu}\mathrm{Cov}\left(X_i, X_iX_j\right).
$$

Note that

$$\begin{aligned}\mathrm{Cov}\big(X_i^2\,,\,X_iX_j\big) &= \mathbb{E}[X_i^2X_iX_j] - \mathbb{E}[X_i^2]\mathbb{E}[X_iX_j]\\ &= \mathbb{E}[X_i^3]\mathbb{E}[X_j] - \mathbb{E}[X_i^2]\mathbb{E}[X_i]\mathbb{E}[X_j]\\ &= \mathbb{E}[X_j]\big(\mathbb{E}[X_i^3] - \mathbb{E}[X_i^2]\mathbb{E}[X_i]\big),\end{aligned}$$

and

$$\begin{aligned}\mathrm{Cov}\big(X_i\,,\,X_iX_j\big) &= \mathbb{E}[X_iX_iX_j] - \mathbb{E}[X_i]\mathbb{E}[X_iX_j]\\ &= \mathbb{E}[X_i^2]\mathbb{E}[X_j] - \mathbb{E}[X_i]\mathbb{E}[X_i]\mathbb{E}[X_j]\\ &= \mathbb{E}[X_j]\big(\mathbb{E}[X_i^2] - \mathbb{E}[X_i]\mathbb{E}[X_i]\big).\end{aligned}$$

So that

$$-(n/2)\times C_{i,j,2,1} = \mathbb{E}[X_j]\big(\mathbb{E}[X_i^3] - \mathbb{E}[X_i^2]\mathbb{E}[X_i] - 2\widetilde{\mu}\big(\mathbb{E}[X_i^2] - \mathbb{E}[X_i]\mathbb{E}[X_i]\big)\big).$$

□

Proof of Lemma 3.25. We have

$$\begin{aligned}C_{i,j,2,2} &:= 4\mathrm{Cov}\Big((X_i-\widetilde{\mu})(\widetilde{\mu}-\overline{X})\,,\,(X_j-\widetilde{\mu})(\widetilde{\mu}-\overline{X})\Big)\\ &= 4\mathrm{Cov}\Big(X_i\widetilde{\mu} - X_i\overline{X} - \widetilde{\mu}^2 + \overline{X}\widetilde{\mu}\,,\,X_j\widetilde{\mu} - X_j\overline{X} - \widetilde{\mu}^2 + \overline{X}\widetilde{\mu}\Big)\\ &= 4\mathrm{Cov}\Big(X_i\widetilde{\mu} - X_i\overline{X} + \overline{X}\widetilde{\mu}\,,\,X_j\widetilde{\mu} - X_j\overline{X} + \overline{X}\widetilde{\mu}\Big),\end{aligned}$$

so

$$\frac{1}{4}C_{i,j,2,2} = \sum_{k,\ell=1}^{3} A_{k,\ell},$$

where

$$\begin{aligned}A_{1,1} &= \mathrm{Cov}\big(X_i\widetilde{\mu}\,,\,X_j\widetilde{\mu}\big),\\ A_{1,2} &= -\mathrm{Cov}\Big(X_i\widetilde{\mu}\,,\,X_j\overline{X}\Big),\\ A_{1,3} &= \mathrm{Cov}\Big(X_i\widetilde{\mu}\,,\,\overline{X}\widetilde{\mu}\Big),\\ A_{2,1} &= -\mathrm{Cov}\Big(X_i\overline{X}\,,\,X_j\widetilde{\mu}\Big),\\ A_{2,2} &= \mathrm{Cov}\Big(X_i\overline{X}\,,\,X_j\overline{X}\Big),\\ A_{2,3} &= -\mathrm{Cov}\Big(X_i\overline{X}\,,\,\overline{X}\widetilde{\mu}\Big),\\ A_{3,1} &= \mathrm{Cov}\Big(\overline{X}\widetilde{\mu}\,,\,X_j\widetilde{\mu}\Big),\\ A_{3,2} &= -\mathrm{Cov}\Big(\overline{X}\widetilde{\mu}\,,\,X_j\overline{X}\Big),\\ A_{3,3} &= \mathrm{Cov}\Big(\overline{X}\widetilde{\mu}\,,\,\overline{X}\widetilde{\mu}\Big).\end{aligned}$$

Note that $A_{1,1} = 0$ by independence. We have

$$A_{1,2} = -\mathrm{Cov}\Big(X_i\widetilde{\mu}\,,\,X_j\overline{X}\Big) = -\frac{1}{n}\widetilde{\mu}\mathrm{Var}[X_i]\mathbb{E}[X_j],$$

$$A_{1,3} = \mathrm{Cov}\Big(X_i\widetilde{\mu}\,,\,\overline{X}\widetilde{\mu}\Big) = \frac{1}{n}\widetilde{\mu}^2\mathrm{Var}[X_i],$$

$$A_{2,1} = -\mathrm{Cov}\left(X_i\overline{X}\,,\,X_j\widetilde{\mu}\right) = -\frac{1}{n}\widetilde{\mu}\mathbb{E}[X_i]\mathrm{Var}[X_j],$$

$$A_{2,2} = \mathrm{Cov}\left(X_i\overline{X}\,,\,X_j\overline{X}\right) = \frac{1}{n^2}\sum_{k=1}^{n}\sum_{\ell=1}^{n}\mathrm{Cov}\left(X_iX_k\,,\,X_jX_\ell\right).$$

This is 0 unless $\ell = k$ or $k = j$ or $\ell = i$. So

$$A_{2,2} = \mathrm{Cov}\left(X_i\overline{X}\,,\,X_j\overline{X}\right) = \frac{1}{n^2}\left(\sum_{k=1}^{n}\mathrm{Cov}\left(X_iX_k\,,\,X_jX_k\right) + \sum_{\ell=1}^{n}\mathrm{Cov}\left(X_iX_j\,,\,X_jX_\ell\right) + \sum_{k=1}^{n}\mathrm{Cov}\left(X_iX_k\,,\,X_jX_i\right) + \mathcal{O}(1)\right).$$

Note that

$$\begin{aligned}
\frac{1}{n}\sum_{k=1}^{n}\mathrm{Cov}\left(X_iX_k\,,\,X_jX_k\right) &= \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[X_iX_k^2X_j] - \mathbb{E}[X_iX_k]\mathbb{E}[X_kX_j] \\
&= \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[X_i]\mathbb{E}[X_k^2]\mathbb{E}[X_j] - \mathbb{E}[X_i]\mathbb{E}^2[X_k]\mathbb{E}[X_j] + \mathcal{O}(1/n) \\
&= \mathbb{E}[X_i]\mathbb{E}[X_j]\left(\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[X_k^2] - \mathbb{E}^2[X_k]\right) + \mathcal{O}(1/n) \\
&= \mathbb{E}[X_i]\mathbb{E}[X_j]\frac{1}{n}\sum_{k=1}^{n}\mathrm{Var}[X_k] + \mathcal{O}(1/n) \\
&= \mathbb{E}[X_i]\mathbb{E}[X_j](\widehat{p}_1\sigma_1^{*2} + \widehat{p}_2\sigma_2^{*2}) + \mathcal{O}(1/n) \\
&= \mathbb{E}[X_i]\mathbb{E}[X_j]\widetilde{\sigma}^2 + \mathcal{O}(1/n).
\end{aligned}$$

Moreover,

$$\begin{aligned}
\frac{1}{n}\sum_{\ell=1}^{n}\mathrm{Cov}\left(X_iX_j\,,\,X_jX_\ell\right) &= \frac{1}{n}\sum_{\ell=1}^{n}\mathbb{E}[X_iX_j^2X_\ell] - \mathbb{E}[X_iX_j]\mathbb{E}[X_jX_\ell] \\
&= \frac{1}{n}\sum_{\ell=1}^{n}\mathbb{E}[X_\ell]\mathbb{E}[X_i]\mathrm{Var}[X_j] + \mathcal{O}(1/n) \\
&= \widetilde{\mu}\mathbb{E}[X_i]\mathrm{Var}[X_j] + \mathcal{O}(1/n).
\end{aligned}$$

Combining the previous computations, we obtain

$$A_{2,2} = \frac{1}{n}\left(\mathbb{E}[X_i]\mathbb{E}[X_j]\widetilde{\sigma}^2 + \widetilde{\mu}\mathbb{E}[X_i]\mathrm{Var}[X_j] + \widetilde{\mu}\mathbb{E}[X_j]\mathrm{Var}[X_i]\right) + \mathcal{O}(1/n^2).$$

Similarly,

$$\begin{aligned}
A_{2,3} &= -\mathrm{Cov}\left(X_i\overline{X}\,,\,\overline{X}\widetilde{\mu}\right) \\
&= -\widetilde{\mu}\mathrm{Cov}\left(X_i\overline{X}\,,\,\overline{X}\right) \\
&= -\widetilde{\mu}\frac{1}{n^2}\sum_{k=1}^{n}\sum_{\ell=1}^{n}\mathrm{Cov}\left(X_iX_k\,,\,X_\ell\right) \\
&= -\widetilde{\mu}\frac{1}{n^2}\left(\sum_{k=1}^{n}\mathrm{Cov}\left(X_iX_k\,,\,X_i\right) + \sum_{k=1}^{n}\mathrm{Cov}\left(X_iX_k\,,\,X_k\right)\right) + \mathcal{O}(1/n^2) \\
&= -\frac{1}{n}\left(\widetilde{\mu}^2\mathrm{Var}[X_i] + \widetilde{\mu}\mathbb{E}[X_i]\widetilde{\sigma}^2\right) + \mathcal{O}(1/n^2).
\end{aligned}$$

Similarly as for $A_{1,3}$, we find that

$$A_{3,1} = \frac{1}{n}\widetilde{\mu}^2\mathrm{Var}[X_j].$$

Similarly as for $A_{2,3}$, we find that

$$A_{3,2} = -\frac{1}{n}\left(\widetilde{\mu}^2\mathrm{Var}[X_j] + \widetilde{\mu}\mathbb{E}[X_j]\widetilde{\sigma}^2\right) + \mathcal{O}(1/n^2).$$

Lastly, we find:

$$A_{3,3} = \frac{1}{n^2}\widetilde{\mu}^2 \sum_{k=1}^{n}\sum_{\ell=1}^{n} \mathrm{Cov}(X_k\,,\, X_\ell) = \frac{1}{n}\widetilde{\mu}^2\widetilde{\sigma}^2 + \mathcal{O}(1/n^2).$$

Combining all results together, we obtain

$$\begin{aligned}
C_{i,j,2,2} = \frac{4}{n}\Big(&-\widetilde{\mu}\mathrm{Var}[X_i]\mathbb{E}[X_j] + \widetilde{\mu}^2\mathrm{Var}[X_i] - \widetilde{\mu}\mathbb{E}[X_i]\mathrm{Var}[X_j] \\
&+ \mathbb{E}[X_i]\mathbb{E}[X_j]\widetilde{\sigma}^2 + \widetilde{\mu}\mathbb{E}[X_i]\mathrm{Var}[X_j] + \widetilde{\mu}\mathbb{E}[X_j]\mathrm{Var}[X_i] \\
&- \widetilde{\mu}^2\mathrm{Var}[X_i] - \widetilde{\mu}\mathbb{E}[X_i]\widetilde{\sigma}^2 + \widetilde{\mu}^2\mathrm{Var}[X_j] - \widetilde{\mu}^2\mathrm{Var}[X_j] - \widetilde{\mu}\mathbb{E}[X_j]\widetilde{\sigma}^2 + \widetilde{\mu}^2 * \widetilde{\sigma}^2\Big) + \mathcal{O}(1/n^2) \\
= \frac{4}{n}\widetilde{\sigma}^2\big(&\mathbb{E}[X_i]\mathbb{E}[X_j] + \widetilde{\mu}^2 - \widetilde{\mu}\mathbb{E}[X_i] - \widetilde{\mu}\mathbb{E}[X_j]\big) + \mathcal{O}(1/n^2) \\
= \frac{4}{n}\widetilde{\sigma}^2\big(&\mathbb{E}[X_i] - \widetilde{\mu}\big)\big(\mathbb{E}[X_j] - \widetilde{\mu}\big) + \mathcal{O}(1/n^2).
\end{aligned}$$

□

With all lemmas now proved, the theorem is proved. □

3.8. Variance of $\widehat{\gamma}_{p_1,p_2}$– Proof

The variance of the estimator $\widehat{\gamma}_{p_1,p_2}$ is computed in this section, proving Theorem 3.11 from Section 3.2. This theorem states the variance of the estimator $\widehat{\gamma}_{p_1,p_2}$. It is proved below.

Proof of Theorem 3.11. The variance of a sum is in general equal to the sum of the variances plus two times the sum of the covariance. This principle is used two times to break the variance of $\widehat{\gamma}_{p_1,p_2}$ down. Before doing this, all estimators of the parameters of equal groups are collected when possible:

$$\begin{aligned}
\mathrm{Var}\left[\widehat{\gamma}_{p_1,p_2}\right] &= \mathrm{Var}\left[p_1\widehat{\sigma}_1^2 + p_2\widehat{\sigma}_2^2 + p_1\widehat{\mu}_1^2 + p_2\widehat{\mu}_2^2 - \widehat{\theta}_{p_1,p_2}^2\right] \\
&= \mathrm{Var}\left[p_1\widehat{\sigma}_1^2 + p_2\widehat{\sigma}_2^2 + p_1\widehat{\mu}_1^2 + p_2\widehat{\mu}_2^2 - \left(p_1\widehat{\mu}_1 + p_2\widehat{\mu}_2\right)^2\right] \\
&= \mathrm{Var}\left[\left(p_1\widehat{\sigma}_1^2 + \left(p_1 - p_1^2\right)\widehat{\mu}_1^2\right) + \left(p_2\widehat{\sigma}_2^2 + \left(p_2 - p_2^2\right)\widehat{\mu}_2^2\right) - 2p_1p_2\widehat{\mu}_1\widehat{\mu}_2\right] \\
&= \mathrm{Var}\left[\left(p_1\widehat{\sigma}_1^2 + p_1p_2\widehat{\mu}_1^2\right) + \left(p_2\widehat{\sigma}_2^2 + p_1p_2\widehat{\mu}_2^2\right) - 2p_1p_2\widehat{\mu}_1\widehat{\mu}_2\right].
\end{aligned}$$

Now the three terms from the expression above are split up below:

$$\begin{aligned}
\mathrm{Var}\left[\widehat{\gamma}_{p_1,p_2}\right] &= \mathrm{Var}\left[p_1\widehat{\sigma}_1^2 + p_1p_2\widehat{\mu}_1^2\right] + \mathrm{Var}\left[p_2\widehat{\sigma}_2^2 + p_1p_2\widehat{\mu}_2^2\right] + \mathrm{Var}\left[-2p_1p_2\widehat{\mu}_1\widehat{\mu}_2\right] \\
&\quad + 2\mathrm{Cov}\left(p_1\widehat{\sigma}_1^2 + p_1p_2\widehat{\mu}_1^2, p_2\widehat{\sigma}_2^2 + p_1p_2\widehat{\mu}_2^2\right) + 2\mathrm{Cov}\left(p_1\widehat{\sigma}_1^2 + p_1p_2\widehat{\mu}_1^2, -2p_1p_2\widehat{\mu}_1\widehat{\mu}_2\right) \\
&\quad + 2\mathrm{Cov}\left(p_2\widehat{\sigma}_2^2 + p_1p_2\widehat{\mu}_2^2, -2p_1p_2\widehat{\mu}_1\widehat{\mu}_2\right) \\
&= p_1^2\mathrm{Var}\left[\widehat{\sigma}_1^2 + p_2\widehat{\mu}_1^2\right] + p_2^2\mathrm{Var}\left[\widehat{\sigma}_2^2 + p_1\widehat{\mu}_2^2\right] + 4p_1^2p_2^2\mathrm{Var}\left[\widehat{\mu}_1\widehat{\mu}_2\right] \\
&\quad - 4p_1^2p_2\mathrm{Cov}\left(\widehat{\sigma}_1^2 + p_2\widehat{\mu}_1^2, \widehat{\mu}_1\widehat{\mu}_2\right) - 4p_1p_2^2\mathrm{Cov}\left(\widehat{\sigma}_2^2 + p_1\widehat{\mu}_2^2, \widehat{\mu}_1\widehat{\mu}_2\right),
\end{aligned} \tag{3.23}$$

by independence of terms of different groups. All terms of the equation above are examined separately, starting with the first three terms in the next subsection.

3.8.1. Computing the variance terms – $\widehat{\gamma}_{p_1,p_2}$

In this subsection, expressions for the first three terms of Equation (3.23) are derived. By definition, the variance is:

$$\mathrm{Var}\left[\widehat{\sigma}_1^2 + p_2\widehat{\mu}_1^2\right] = \mathbb{E}\left[\left(\widehat{\sigma}_1^2 + p_2\widehat{\mu}_1^2\right)^2\right] - \mathbb{E}\left[\widehat{\sigma}_1^2 + p_2\widehat{\mu}_1^2\right]^2. \tag{3.24}$$

The second term is computed below, using Lemma 3.15:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\sigma}_1^2 + p_2\widehat{\mu}_1^2\right] &= \mathbb{E}\left[\widehat{\sigma}_1^2\right] + p_2\mathbb{E}\left[\widehat{\mu}_1^2\right] \\
&= \sigma_1^{*2} + p_2\left(\frac{\sigma_1^{*2}}{n_1} + \mu_1^{*2}\right) \\
&= \sigma_1^{*2} + p_2\mu_1^{*2} + \frac{p_2}{n_1}\sigma_1^{*2}.
\end{aligned} \tag{3.25}$$

The first term of Equation (3.24) can be written as the sum of multiple expectations:

$$\begin{aligned}\mathbb{E}\left[\left(\widehat{\sigma}_1^2+p_2\widehat{\mu}_1^2\right)^2\right] &= \mathbb{E}\left[\widehat{\sigma}_1^4+p_2^2\widehat{\mu}_1^4+2p_2\widehat{\sigma}_1^2\widehat{\mu}_1^2\right]\\ &= \mathbb{E}\left[\widehat{\sigma}_1^4\right]+p_2^2\mathbb{E}\left[\widehat{\mu}_1^4\right]+2p_2\mathbb{E}\left[\widehat{\sigma}_1^2\widehat{\mu}_1^2\right].\end{aligned}$$

With Lemma 3.17, Lemma 3.19 and Lemma 3.20, the expression for $\mathbb{E}\left[\left(\widehat{\sigma}_1^2+p_2\widehat{\mu}_1^2\right)^2\right]$ is rewritten as:

$$\begin{aligned}&\frac{1}{n_1}\left(m_{4,1}^*-4m_{3,1}^*\mu_1^*+6\sigma_1^{*2}\mu_1^{*2}+3\mu_1^{*4}\right)+\frac{n_1-1}{n_1}\sigma_1^{*4}+\frac{2}{n_1\left(n_1-1\right)}\sigma_1^{*4}\\ &+p_2^2\left(\frac{1}{n_1^3}m_{4,1}^*+\frac{\left(n_1-1\right)}{n_1^3}\left(4\mu_1^*m_{3,1}^*+\left(6\left(n_1-2\right)\mu_1^{*2}+3\left(\sigma_1^{*2}+\mu_1^{*2}\right)\right)\left(\sigma_1^{*2}+\mu_1^{*2}\right)+\left(n_1-2\right)\left(n_1-3\right)\mu_1^{*4}\right)\right)\\ &+2p_2\frac{1}{n_1^2}\left(m_{4,1}^*+2\left(n_1-2\right)m_{3,1}^*\mu_1^*+\left(n_1^2-6n_1+6\right)\sigma_1^{*2}\mu_1^{*2}+\left(-2n_1+3\right)\mu_1^{*4}+\left(n_1-3\right)\sigma_1^{*4}\right).\end{aligned}$$

We can further simplify this to

$$\begin{aligned}&\sigma_1^{*4}+p_2^2\mu_1^{*4}+2p_2\sigma_1^{*2}\mu_1^{*2}+\frac{1}{n_1}\left(m_{4,1}^*-4m_{3,1}^*\mu_1^*+6\sigma_1^{*2}\mu_1^{*2}+3\mu_1^{*4}-\sigma_1^{*4}\right)\\ &+\frac{1}{n_1}\left(p_2^2\left(6\mu_1^{*2}\left(\sigma_1^{*2}+\mu_1^{*2}\right)-6\mu_1^{*4}\right)+2p_2\left(2m_{3,1}^*\mu_1^*-6\sigma_1^{*2}\mu_1^{*2}-2\mu_1^{*4}+\sigma_1^{*4}\right)\right)\\ &=\sigma_1^{*4}+p_2^2\mu_1^{*4}+2p_2\sigma_1^{*2}\mu_1^{*2}\\ &+\frac{1}{n_1}\left(m_{4,1}^*+4\left(p_2-1\right)m_{3,1}^*\mu_1^*+6\left(p_2^2-2p_2+1\right)\sigma_1^{*2}\mu_1^{*2}+\left(3-4p_2\right)\mu_1^{*4}+\left(2p_2-1\right)\sigma_1^{*4}\right)+\mathcal{O}\left(\frac{1}{n^2}\right)\\ &=\sigma_1^{*4}+p_2^2\mu_1^{*4}+2p_2\sigma_1^{*2}\mu_1^{*2}+\frac{1}{n_1}\left(m_{4,1}^*-4p_1m_{3,1}^*\mu_1^*+6p_1^2\sigma_1^{*2}\mu_1^{*2}+\left(3-4p_2\right)\mu_1^{*4}+\left(p_2-p_1\right)\sigma_1^{*4}\right)+\mathcal{O}\left(\frac{1}{n^2}\right).\end{aligned} \tag{3.26}$$

Assembling Equation (3.24), Equation (3.25) and Equation (3.26) leads to:

$$\begin{aligned}\operatorname{Var}\left[\widehat{\sigma}_1^2+p_2\widehat{\mu}_1^2\right] &= \sigma_1^{*4}+p_2^2\mu_1^{*4}+2p_2\sigma_1^{*2}\mu_1^{*2}-\left(\left(\sigma_1^{*2}+p_2\mu_1^{*2}\right)+\frac{p_2}{n_1}\sigma_1^{*2}\right)^2\\ &+\frac{1}{n_1}\left(m_{4,1}^*-4p_1m_{3,1}^*\mu_1^*+6p_1^2\sigma_1^{*2}\mu_1^{*2}+\left(3-4p_2\right)\mu_1^{*4}+\left(p_2-p_1\right)\sigma_1^{*4}\right)+\mathcal{O}\left(\frac{1}{n^2}\right)\\ &=\sigma_1^{*4}+p_2^2\mu_1^{*4}+2p_2\sigma_1^{*2}\mu_1^{*2}-\left(\sigma_1^{*2}+p_2\mu_1^{*2}\right)^2-\frac{2p_2}{n_1}\sigma_1^{*2}\left(\sigma_1^{*2}+p_2\mu_1^{*2}\right)-\frac{p_2^2}{n_1}\sigma_1^{*2}\\ &+\frac{1}{n_1}\left(m_{4,1}^*-4p_1m_{3,1}^*\mu_1^*+6p_1^2\sigma_1^{*2}\mu_1^{*2}+\left(3-4p_2\right)\mu_1^{*4}+\left(p_2-p_1\right)\sigma_1^{*4}\right)+\mathcal{O}\left(\frac{1}{n^2}\right).\end{aligned}$$

All terms of the order of $\mathcal{O}(1)$ cancel, meaning only second order terms or higher are remaining. This results in:

$$\begin{aligned}\operatorname{Var}\left[\widehat{\sigma}_1^2+p_2\widehat{\mu}_1^2\right] &= \sigma_1^{*4}+p_2^2\mu_1^{*4}+2p_2\sigma_1^{*2}\mu_1^{*2}-\left(\sigma_1^{*4}+2p_2\sigma_1^{*2}\mu_1^{*2}+p_2^2\mu_1^{*4}\right)\\ &+\frac{1}{n_1}\left(m_{4,1}^*-4p_1m_{3,1}^*\mu_1^*+4p_1^2\sigma_1^{*2}\mu_1^{*2}+\left(3-4p_2\right)\mu_1^{*4}+\left(-p_2-p_1\right)\sigma_1^{*4}\right)+\mathcal{O}\left(\frac{1}{n^2}\right)\\ &=\frac{1}{n_1}\left(m_{4,1}^*-4p_1m_{3,1}^*\mu_1^*+4p_1^2\sigma_1^{*2}\mu_1^{*2}+\left(3-4p_2\right)\mu_1^{*4}-\sigma_1^{*4}\right)+\mathcal{O}\left(\frac{1}{n^2}\right).\end{aligned} \tag{3.27}$$

By symmetry, $\operatorname{Var}\left[\widehat{\sigma}_2^2+p_1\widehat{\mu}_2^2\right]$ is equal to Equation (3.27) with all indices changed from 1 to 2 and vice versa.

The next term of Equation (3.23) is computed below, using the independence of measurements of different groups:

$$\begin{aligned}\operatorname{Var}\left[\widehat{\mu}_1\widehat{\mu}_2\right] &= \mathbb{E}\left[\widehat{\mu}_1^2\widehat{\mu}_2^2\right]-\mathbb{E}\left[\widehat{\mu}_1\widehat{\mu}_2\right]^2\\ &= \mathbb{E}\left[\widehat{\mu}_1^2\right]\mathbb{E}\left[\widehat{\mu}_2^2\right]-\mathbb{E}\left[\widehat{\mu}_1\right]^2\mathbb{E}\left[\widehat{\mu}_2\right]^2.\end{aligned}$$

Replacing the expected value with true parameters leads to:

$$\begin{aligned}\mathrm{Var}\left[\widehat{\mu}_1\widehat{\mu}_2\right] &= \left(\frac{\sigma_1^{*2}}{n_1}+\mu_1^{*2}\right)\left(\frac{\sigma_2^{*2}}{n_2}+\mu_2^{*2}\right)-\mu_1^{*2}\mu_2^{*2}\\ &= \frac{\sigma_1^{*2}\sigma_2^{*2}}{n_1 n_2}+\frac{\sigma_1^{*2}\mu_2^{*2}}{n_1}+\frac{\sigma_2^{*2}\mu_1^{*2}}{n_2}\\ &= \frac{1}{n_1}\sigma_1^{*2}\mu_2^{*2}+\frac{1}{n_2}\sigma_2^{*2}\mu_1^{*2}+\mathcal{O}\left(\frac{1}{n^2}\right).\end{aligned} \tag{3.28}$$

3.8.2. Computing the covariance terms – $\widehat{\gamma}_{p_1,p_2}$

The first three terms of Equation (3.23) have already been derived in the previous subsection. The other two terms, involving a covariance, are worked out in this subsection. They are similar and therefore, computing one directly gives the other term. We firstly rewrite $\mathrm{Cov}\left(\widehat{\sigma}_2^2+p_1\widehat{\mu}_2^2,\widehat{\mu}_1\widehat{\mu}_2\right)$ using its definition. The independence of measurements of different groups is used thereafter. This leads to

$$\begin{aligned}\mathrm{Cov}\left(\widehat{\sigma}_2^2+p_1\widehat{\mu}_2^2,\widehat{\mu}_1\widehat{\mu}_2\right) &= \mathbb{E}\left[\left(\widehat{\sigma}_2^2+p_1\widehat{\mu}_2^2\right)\widehat{\mu}_1\widehat{\mu}_2\right]-\mathbb{E}\left[\widehat{\sigma}_2^2+p_1\widehat{\mu}_2^2\right]\mathbb{E}\left[\widehat{\mu}_1\widehat{\mu}_2\right]\\ &= \left(\mathbb{E}\left[\widehat{\sigma}_2^2\widehat{\mu}_1\widehat{\mu}_2\right]+p_1\mathbb{E}\left[\widehat{\mu}_1\widehat{\mu}_2^3\right]\right)-\left(\mathbb{E}\left[\widehat{\sigma}_2^2\right]+p_1\mathbb{E}\left[\widehat{\mu}_2^2\right]\right)\mathbb{E}\left[\widehat{\mu}_1\widehat{\mu}_2\right]\\ &= \mathbb{E}\left[\widehat{\sigma}_2^2\widehat{\mu}_2\right]\mathbb{E}\left[\widehat{\mu}_1\right]+p_1\mathbb{E}\left[\widehat{\mu}_1\right]\mathbb{E}\left[\widehat{\mu}_2^3\right]-\left(\mathbb{E}\left[\widehat{\sigma}_2^2\right]+p_1\mathbb{E}\left[\widehat{\mu}_2^2\right]\right)\mathbb{E}\left[\widehat{\mu}_1\right]\mathbb{E}\left[\widehat{\mu}_2\right].\end{aligned}$$

With Lemma 3.15, Lemma 3.16 and Lemma 3.18, we can rewrite the previous equation as

$$\begin{aligned}&\frac{1}{n_2}\left(m_{3,2}^*+(n_2-3)\,\sigma_2^{*2}\mu_2^*-\mu_2^{*3}\right)\mu_1^*+p_1\mu_1^*\left(\frac{1}{n_2^2}m_{3,2}^*+\frac{n_2-1}{n_2^2}\left(3\sigma_2^{*2}\mu_2^*+(n_2+1)\,\mu_2^{*3}\right)\right)\\ &-\left(\sigma_2^{*2}+p_1\left(\frac{\sigma_2^{*2}}{n_2}+\mu_2^{*2}\right)\right)\mu_1^*\mu_2^*\\ &=\mu_1^*\mu_2^*\sigma_2^{*2}+p_1\mu_1^*\mu_2^{*3}-\mu_1^*\mu_2^*\sigma_2^{*2}-p_1\mu_1^*\mu_2^{*3}\\ &+\frac{1}{n_2}\left(m_{3,2}^*\mu_1^*-3\mu_1^*\mu_2^*\sigma_2^{*2}-\mu_1^*\mu_2^{*3}+3p_1\mu_1^*\mu_2^*\sigma_2^{*2}-p_1\mu_1^*\mu_2^*\sigma_2^{*2}\right)+\mathcal{O}\left(\frac{1}{n^2}\right)\\ &=\frac{1}{n_2}\left(m_{3,2}^*\mu_1^*-\mu_1^*\mu_2^{*3}+\left(2p_1-3\right)\mu_1^*\mu_2^*\sigma_2^{*2}\right)+\mathcal{O}\left(\frac{1}{n^2}\right),\end{aligned}$$

from which we can conclude that only terms of the order $\mathcal{O}(1/n)$ or higher are remaining. The other term, $\mathrm{Cov}\left(\widehat{\sigma}_1^2+p_2\widehat{\mu}_1^2,\widehat{\mu}_1\widehat{\mu}_2\right)$, is similar, but with the indices switched from 1 to 2 and vice versa.

All terms of Equation (3.23) are now known. Adding the terms leads to the desired expression for the variance of $\widehat{\gamma}_{p_1,p_2}$:

$$\begin{aligned}\mathrm{Var}\left[\widehat{\gamma}_{p_1,p_2}\right] &= \frac{1}{n_1}\left(p_1^2m_{4,1}^*-4p_1^3m_{3,1}^*\mu_1^*-4p_1^2p_2m_{3,1}^*\mu_2^*+4p_1^4\sigma_1^{*2}\mu_1^{*2}+4p_1^2p_2^2\sigma_1^{*2}\mu_2^{*2}-4p_1^2p_2\left(2p_2-3\right)\mu_1^*\mu_2^*\sigma_1^{*2}\right.\\ &\qquad\left.+p_1^2\left(3-4p_2\right)\mu_1^{*4}+4p_1^2p_2\mu_1^{*3}\mu_2^*-p_1^2\sigma_1^{*4}\right)\\ &+\frac{1}{n_2}\left(p_2^2m_{4,2}^*-4p_2^3m_{3,2}^*\mu_2^*-4p_1p_2^2m_{3,2}^*\mu_1^*+4p_2^4\sigma_2^{*2}\mu_2^{*2}+4p_1^2p_2^2\sigma_2^{*2}\mu_1^{*2}-4p_1p_2^2\left(2p_1-3\right)\mu_1^*\mu_2^*\sigma_2^{*2}\right.\\ &\qquad\left.+p_2^2\left(3-4p_1\right)\mu_2^{*4}+4p_1p_2^2\mu_1^*\mu_2^{*3}-p_2^2\sigma_2^{*4}\right)+\mathcal{O}\left(\frac{1}{n^2}\right)\\ &=\frac{p_1^2}{n_1}\left(m_{4,1}^*-4m_{3,1}^*\left(p_1\mu_1^*+p_2\mu_2^*\right)+4\sigma_1^{*2}\left(p_1^2\mu_1^{*2}-p_2\left(2p_2-3\right)\mu_1^*\mu_2^*+p_2^2\mu_2^{*2}\right)\right.\\ &\qquad\left.+\mu_1^{*3}\left(3\mu_1^*+4p_2\Delta\mu\right)-\sigma_1^{*4}\right)\\ &+\frac{p_2^2}{n_2}\left(m_{4,2}^*-4m_{3,2}^*\left(p_1\mu_1^*+p_2\mu_2^*\right)+4\sigma_2^{*2}\left(p_1^2\mu_1^{*2}-p_1\left(2p_1-3\right)\mu_1^*\mu_2^*+p_2^2\mu_2^{*2}\right)\right.\\ &\qquad\left.+\mu_2^{*3}\left(3\mu_2^*-4p_1\Delta\mu\right)-\sigma_2^{*4}\right)+\mathcal{O}\left(\frac{1}{n^2}\right),\end{aligned}$$

which concludes the proof. □

3.9. Variance of $\widehat{\psi}_{p_1,p_2}$ – Proof

Let us recall Theorem 3.12. The proof of this theorem, stating the variance of the estimator $\widehat{\psi}_{p_1,p_2}$, is given below. The proof has the same structure as the proof of Theorem 3.11 given in the previous section.

Proof of Theorem 3.12. All terms inside the variance of $\widehat{\psi}_{p_1,p_2}$ can be collected as in the expression below:

$$\begin{aligned}
\mathrm{Var}\left[\widehat{\psi}_{p_1,p_2}\right] &= \mathrm{Var}\left[p_1\widehat{\sigma}_1^2 + p_2\widehat{\sigma}_2^2 + p_1\left(\widehat{\mu}_1^2 - \frac{\widehat{\sigma}_1^2}{n_1}\right) + p_2\left(\widehat{\mu}_2^2 - \frac{\widehat{\sigma}_2^2}{n_2}\right) - \left(p_1\widehat{\mu}_1 + p_2\widehat{\mu}_2\right)^2 + \frac{p_1^2}{n_1}\widehat{\sigma}_1^2 + \frac{p_2^2}{n_2}\widehat{\sigma}_2^2\right] \\
&= \mathrm{Var}\left[p_1\left(\widehat{\sigma}_1^2 + (1-p_1)\widehat{\mu}_1^2 + \frac{p_1-1}{n_1}\widehat{\sigma}_1^2\right) + p_2\left(\widehat{\sigma}_2^2 + (1-p_2)\widehat{\mu}_2^2 + \frac{p_2-1}{n_2}\widehat{\sigma}_2^2\right) - 2p_1p_2\widehat{\mu}_1\widehat{\mu}_2\right] \\
&= \mathrm{Var}\left[p_1\left(\widehat{\sigma}_1^2 + p_2\widehat{\mu}_1^2 - \frac{p_2}{n_1}\widehat{\sigma}_1^2\right) + p_2\left(\widehat{\sigma}_2^2 + p_1\widehat{\mu}_2^2 - \frac{p_1}{n_2}\widehat{\sigma}_2^2\right) - 2p_1p_2\widehat{\mu}_1\widehat{\mu}_2\right].
\end{aligned}$$

Breaking the variance down into different parts leads to:

$$\begin{aligned}
\mathrm{Var}\left[\widehat{\psi}_{p_1,p_2}\right] &= \mathrm{Var}\left[p_1\left(\widehat{\sigma}_1^2 + p_2\widehat{\mu}_1^2 - \frac{p_2}{n_1}\widehat{\sigma}_1^2\right)\right] + \mathrm{Var}\left[p_2\left(\widehat{\sigma}_2^2 + p_1\widehat{\mu}_2^2 - \frac{p_1}{n_2}\widehat{\sigma}_2^2\right)\right] + \mathrm{Var}\left[-2p_1p_2\widehat{\mu}_1\widehat{\mu}_2\right] \\
&\quad + 2\mathrm{Cov}\left(p_1\left(\widehat{\sigma}_1^2 + p_2\widehat{\mu}_1^2 - \frac{p_2}{n_1}\widehat{\sigma}_1^2\right), p_2\left(\widehat{\sigma}_2^2 + p_1\widehat{\mu}_2^2 - \frac{p_1}{n_2}\widehat{\sigma}_2^2\right)\right) \\
&\quad + 2\mathrm{Cov}\left(p_1\left(\widehat{\sigma}_1^2 + p_2\widehat{\mu}_1^2 - \frac{p_2}{n_1}\widehat{\sigma}_1^2\right), -2p_1p_2\widehat{\mu}_1\widehat{\mu}_2\right) \\
&\quad + 2\mathrm{Cov}\left(p_2\left(\widehat{\sigma}_2^2 + p_1\widehat{\mu}_2^2 - \frac{p_1}{n_2}\widehat{\sigma}_2^2\right), -2p_1p_2\widehat{\mu}_1\widehat{\mu}_2\right) \\
&= p_1^2\mathrm{Var}\left[\widehat{\sigma}_1^2 + p_2\widehat{\mu}_1^2 - \frac{p_2}{n_1}\widehat{\sigma}_1^2\right] + p_2^2\mathrm{Var}\left[\widehat{\sigma}_2^2 + p_1\widehat{\mu}_2^2 - \frac{p_1}{n_2}\widehat{\sigma}_2^2\right] + 4p_1^2p_2^2\mathrm{Var}\left[\widehat{\mu}_1\widehat{\mu}_2\right] \\
&\quad - 4p_1^2p_2\mathrm{Cov}\left(\widehat{\sigma}_1^2 + p_2\widehat{\mu}_1^2 - \frac{p_2}{n_1}\widehat{\sigma}_1^2, \widehat{\mu}_1\widehat{\mu}_2\right) - 4p_1p_2^2\mathrm{Cov}\left(\widehat{\sigma}_2^2 + p_1\widehat{\mu}_2^2 - \frac{p_1}{n_2}\widehat{\sigma}_2^2, \widehat{\mu}_1\widehat{\mu}_2\right),
\end{aligned} \tag{3.29}$$

because of the independence of different groups.

The terms containing a variance are computed in the first subsection, whereas the terms containing a covariance are computed in the second subsection.

3.9.1. Computing the variance terms – $\widehat{\psi}_{p_1,p_2}$

The first term of Equation 3.29 is by definition equal to:

$$\mathrm{Var}\left[\widehat{\sigma}_1^2 + p_2\widehat{\mu}_1^2 - \frac{p_2}{n_1}\widehat{\sigma}_1^2\right] = \mathbb{E}\left[\left(\widehat{\sigma}_1^2 + p_2\widehat{\mu}_1^2 - \frac{p_2}{n_1}\widehat{\sigma}_1^2\right)^2\right] - \mathbb{E}\left[\widehat{\sigma}_1^2 + p_2\widehat{\mu}_1^2 - \frac{p_2}{n_1}\widehat{\sigma}_1^2\right]^2, \tag{3.30}$$

with

$$\begin{aligned}
\mathbb{E}\left[\widehat{\sigma}_1^2 + p_2\widehat{\mu}_1^2 - \frac{p_2}{n_1}\widehat{\sigma}_1^2\right] &= \mathbb{E}\left[\widehat{\sigma}_1^2\right] + p_2\mathbb{E}\left[\widehat{\mu}_1^2\right] - \frac{p_2}{n_1}\mathbb{E}\left[\widehat{\sigma}_1^2\right] \\
&= \sigma_1^{*2} + p_2\left(\frac{\sigma_1^{*2}}{n_1} + \mu_1^{*2}\right) - \frac{p_2}{n_1}\sigma_1^{*2} \\
&= \sigma_1^{*2} + p_2\mu_1^{*2}.
\end{aligned}$$

The first term of Equation 3.30 is equal to:

$$\begin{aligned}
\mathbb{E}\left[\left(\widehat{\sigma}_1^2 + p_2\widehat{\mu}_1^2 - \frac{p_2}{n_1}\widehat{\sigma}_1^2\right)^2\right] &= \mathbb{E}\left[\widehat{\sigma}_1^4 + p_2\widehat{\mu}_1^2\widehat{\sigma}_1^2 - \frac{p_2}{n_1}\widehat{\sigma}_1^4 + p_2\widehat{\mu}_1^2\widehat{\sigma}_1^2 + p_2^2\widehat{\mu}_1^4 - \frac{p_2^2}{n_1}\widehat{\mu}_1^2\widehat{\sigma}_1^2 - \frac{p_2}{n_1}\widehat{\sigma}_1^4 - \frac{p_2^2}{n_1}\widehat{\mu}_1^2\widehat{\sigma}_1^2 + \frac{p_2^2}{n_1^2}\widehat{\sigma}_1^4\right] \\
&= \mathbb{E}\left[\frac{p_2^2 - 2n_1p_2 + n_1^2}{n_1^2}\widehat{\sigma}_1^4 + 2p_2\frac{n_1 - p_2}{n_1}\widehat{\mu}_1^2\widehat{\sigma}_1^2 + p_2^2\widehat{\mu}_1^4\right] \\
&= \frac{p_2^2 - 2n_1p_2 + n_1^2}{n_1^2}\mathbb{E}\left[\widehat{\sigma}_1^4\right] + 2p_2\frac{n_1 - p_2}{n_1}\mathbb{E}\left[\widehat{\sigma}_1^2\widehat{\mu}_1^2\right] + p_2^2\mathbb{E}\left[\widehat{\mu}_1^4\right].
\end{aligned}$$

Remark that this expression only exists of known terms. Using Lemma 3.17, Lemma 3.19 and Lemma 3.20, we can simplify it as:

$$
\begin{aligned}
&\frac{p_2^2-2n_1p_2+n_1^2}{n_1^2}\left(\frac{1}{n_1}\left(m_{4,1}^*-4m_{3,1}^*\mu_1^*+6\sigma_1^{*2}\mu_1^{*2}+3\mu_1^{*4}\right)+\frac{n_1-1}{n_1}\sigma_1^{*4}+\frac{2}{n_1(n_1-1)}\sigma_1^{*4}\right)\\
&+2p_2\frac{n_1-p_2}{n_1}\left(\frac{1}{n_1^2}\left(m_{4,1}^*+2(n_1-2)m_{3,1}^*\mu_1^*+\left(n_1^2-6n_1+6\right)\sigma_1^{*2}\mu_1^{*2}+(-2n_1+3)\mu_1^{*4}+(n_1-3)\sigma_1^{*4}\right)\right)\\
&+p_2^2\left(\frac{1}{n_1^3}m_{4,1}^*+\frac{(n_1-1)}{n_1^3}\left(4\mu_1^*m_{3,1}^*+\left(6(n_1-2)\mu_1^{*2}+3\left(\sigma_1^{*2}+\mu_1^{*2}\right)\right)\left(\sigma_1^{*2}+\mu_1^{*2}\right)+(n_1-2)(n_1-3)\mu_1^{*4}\right)\right)\\
&=\left(1-\frac{2p_2}{n_1}\right)\left(\sigma_1^{*4}+\frac{1}{n_1}\left(m_{4,1}^*-4m_{3,1}^*\mu_1^*+6\sigma_1^{*2}\mu_1^{*2}+3\mu_1^{*4}-\sigma_1^{*4}\right)\right)\\
&+2p_2\left(1-\frac{p_2}{n_1}\right)\left(\sigma_1^{*2}\mu_1^{*2}+\frac{1}{n_1}\left(2m_{3,1}^*\mu_1^*-6\sigma_1^{*2}\mu_1^{*2}-2\mu_1^{*4}+\sigma_1^{*4}\right)\right)\\
&+p_2^2\left(\frac{(n_1-1)}{n_1^3}\left(6n_1\mu_1^{*2}\left(\sigma_1^{*2}+\mu_1^{*2}\right)+\left(n_1^2-5n_1\right)\mu_1^{*4}\right)\right)+\mathcal{O}\left(\frac{1}{n^2}\right)\\
&=\sigma_1^{*4}+2p_2\sigma_1^{*2}\mu_1^{*2}+p_2^2\mu_1^{*4}+\frac{1}{n_1}\left(m_{4,1}^*-4m_{3,1}^*\mu_1^*+6\sigma_1^{*2}\mu_1^{*2}+3\mu_1^{*4}-\left(2p_2+1\right)\sigma_1^{*4}\right)\\
&+\frac{2p_2}{n_1}\left(2m_{3,1}^*\mu_1^*-\left(p_2+6\right)\sigma_1^{*2}\mu_1^{*2}-2\mu_1^{*4}+\sigma_1^{*4}\right)+\frac{p_2^2}{n_1}\left(6\sigma_1^{*2}\mu_1^{*2}\right)+\mathcal{O}\left(\frac{1}{n^2}\right)\\
&=\sigma_1^{*4}+2p_2\mu_1^{*2}\sigma_1^{*2}+p_2^2\mu_1^{*4}\\
&+\frac{1}{n_1}\left(m_{4,1}^*+4\left(p_2-1\right)m_{3,1}^*\mu_1^*+2\left(2p_2^2-6p_2+3\right)\sigma_1^{*2}\mu_1^{*2}+\left(3-4p_2\right)\mu_1^{*4}-\sigma_1^{*4}\right)+\mathcal{O}\left(\frac{1}{n^2}\right).
\end{aligned}
$$

Now Equation 3.30 becomes:

$$
\begin{aligned}
&\mathrm{Var}\left[\widehat{\sigma}_1^2+p_2\widehat{\mu}_1^2-\frac{p_2}{n_1}\widehat{\sigma}_1^2\right]\\
&=\sigma_1^{*4}+2p_2\mu_1^{*2}\sigma_1^{*2}+p_2^2\mu_1^{*4}-\left(\sigma_1^{*2}+p_2\mu_1^{*2}\right)^2\\
&+\frac{1}{n_1}\left(m_{4,1}^*+4\left(p_2-1\right)m_{3,1}^*\mu_1^*+2\left(2p_2^2-6p_2+3\right)\sigma_1^{*2}\mu_1^{*2}+\left(3-4p_2\right)\mu_1^{*4}-\sigma_1^{*4}\right)+\mathcal{O}\left(\frac{1}{n^2}\right)\\
&=\frac{1}{n_1}\left(m_{4,1}^*-4p_1m_{3,1}^*\mu_1^*+\left(\left(p_2-1\right)^2-2p_2^2\right)\sigma_1^{*2}\mu_1^{*2}+\left(3-4p_2\right)\mu_1^{*4}-\sigma_1^{*4}\right)+\mathcal{O}\left(\frac{1}{n^2}\right)\\
&=\frac{1}{n_1}\left(m_{4,1}^*-4p_1m_{3,1}^*\mu_1^*+\left(p_1^2-2p_2^2\right)\sigma_1^{*2}\mu_1^{*2}+\left(3-4p_2\right)\mu_1^{*4}-\sigma_1^{*4}\right)+\mathcal{O}\left(\frac{1}{n^2}\right).
\end{aligned}
$$

The computation of $\mathrm{Var}\left[\widehat{\sigma}_2^2+p_1\widehat{\mu}_2^2-\frac{p_1}{n_2}\widehat{\sigma}_2^2\right]$ is similar, leading to:

$$
\mathrm{Var}\left[\widehat{\sigma}_2^2+p_1\widehat{\mu}_2^2-\frac{p_1}{n_2}\widehat{\sigma}_2^2\right]=\frac{1}{n_2}\left(m_{4,2}^*-4p_2m_{3,2}^*\mu_2^*+\left(p_2^2-2p_1^2\right)\sigma_2^{*2}\mu_2^{*2}+\left(3-4p_1\right)\mu_2^{*4}-\sigma_2^{*4}\right)+\mathcal{O}\left(\frac{1}{n^2}\right).
$$

By Equation (3.28),

$$
\mathrm{Var}\left[\widehat{\mu}_1\widehat{\mu}_2\right]=\frac{1}{n_1}\sigma_1^{*2}\mu_2^{*2}+\frac{1}{n_2}\sigma_2^{*2}\mu_1^{*2}+\mathcal{O}\left(\frac{1}{n^2}\right).
$$

3.9.2. Computing the covariance terms – $\widehat{\psi}_{p_1,p_2}$

The first three terms of Equation 3.29 are now known. Computing one of the two covariance terms of this equation leads to:

$$\begin{aligned}
&\operatorname{Cov}\left(\widehat{\sigma}_2^2 + p_1\widehat{\mu}_2^2 - \frac{p_1}{n_2}\widehat{\sigma}_2^2, \widehat{\mu}_1\widehat{\mu}_2\right) \\
&= \mathbb{E}\left[\left(\widehat{\sigma}_2^2 + p_1\widehat{\mu}_2^2 - \frac{p_1}{n_2}\widehat{\sigma}_2^2\right)\widehat{\mu}_1\widehat{\mu}_2\right] - \mathbb{E}\left[\widehat{\sigma}_2^2 + p_1\widehat{\mu}_2^2 - \frac{p_1}{n_2}\widehat{\sigma}_2^2\right]\mathbb{E}\left[\widehat{\mu}_1\widehat{\mu}_2\right] \\
&= \mathbb{E}\left[\left(\frac{n_2-p_1}{n_2}\widehat{\sigma}_2^2 + p_1\widehat{\mu}_2^2\right)\widehat{\mu}_1\widehat{\mu}_2\right] - \mathbb{E}\left[\frac{n_2-p_1}{n_2}\widehat{\sigma}_2^2 + p_1\widehat{\mu}_2^2\right]\mathbb{E}\left[\widehat{\mu}_1\widehat{\mu}_2\right] \\
&= \frac{n_2-p_1}{n_2}\mathbb{E}\left[\widehat{\sigma}_2^2\widehat{\mu}_1\widehat{\mu}_2\right] + p_1\mathbb{E}\left[\widehat{\mu}_1\widehat{\mu}_2^3\right] - \left(\frac{n_2-p_1}{n_2}\mathbb{E}\left[\widehat{\sigma}_2^2\right] + p_1\mathbb{E}\left[\widehat{\mu}_2^2\right]\right)\mathbb{E}\left[\widehat{\mu}_1\widehat{\mu}_2\right] \\
&= \frac{n_2-p_1}{n_2}\mathbb{E}\left[\widehat{\sigma}_2^2\widehat{\mu}_2\right]\mathbb{E}\left[\widehat{\mu}_1\right] + p_1\mathbb{E}\left[\widehat{\mu}_1\right]\mathbb{E}\left[\widehat{\mu}_2^3\right] - \left(\frac{n_2-p_1}{n_2}\mathbb{E}\left[\widehat{\sigma}_2^2\right] + p_1\mathbb{E}\left[\widehat{\mu}_2^2\right]\right)\mathbb{E}\left[\widehat{\mu}_1\right]\mathbb{E}\left[\widehat{\mu}_2\right].
\end{aligned}$$

Using all known quantities from Lemma 3.15, Lemma 3.16 and Lemma 3.18, this can be rewritten as:

$$\begin{aligned}
&\frac{n_2-p_1}{n_2^2}\left(m_{3,2}^* + (n_2-3)\,\sigma_2^{*2}\mu_2^* - \mu_2^{*3}\right)\mu_1^* + p_1\mu_1^*\left(\frac{1}{n_2^2}m_{3,2}^* + \frac{n_2-1}{n_2^2}\left(3\sigma_2^{*2}\mu_2^* + (n_2+1)\,\mu_2^{*3}\right)\right) \\
&-\left(\frac{n_2-p_1}{n_2}\sigma_2^{*2} + p_1\left(\frac{\sigma_2^{*2}}{n_2} + \mu_2^{*2}\right)\right)\mu_1^*\mu_2^* \\
&= \left(\frac{1}{n_2} - \frac{p_1}{n_2^2}\right)\left(n_2\sigma_2^{*2}\mu_2^* + m_{3,2}^* - 3\sigma_2^{*2}\mu_2^* - \mu_2^{*3}\right)\mu_1^* + p_1\mu_1^*\left(\frac{1}{n_2} - \frac{1}{n_2^2}\right)\left(n_2\mu_2^{*3} + 3\sigma_2^{*2}\mu_2^* + \mu_2^{*3}\right) \\
&-\left(\sigma_2^{*2} + p_1\mu_2^{*2}\right)\mu_1^*\mu_2^* + \mathcal{O}\left(\frac{1}{n^2}\right) \\
&= \mu_1^*\mu_2^*\sigma_2^{*2} + p_1\mu_1^*\mu_2^{*3} - \mu_1^*\mu_2^*\sigma_2^{*2} - p_1\mu_1^*\mu_2^{*3} + \frac{1}{n_2}\left((2p_1-3)\,\mu_1^*\mu_2^*\sigma_2^{*2} + m_{3,2}^*\mu_1^* - \mu_1^*\mu_2^{*3}\right) + \mathcal{O}\left(\frac{1}{n^2}\right) \\
&= \frac{1}{n_2}\left(m_{3,2}^*\mu_1^* + (2p_1-3)\,\mu_1^*\mu_2^*\sigma_2^{*2} - \mu_1^*\mu_2^{*3}\right) + \mathcal{O}\left(\frac{1}{n^2}\right).
\end{aligned}$$

Similarly,

$$\operatorname{Cov}\left(\widehat{\sigma}_1^2 + p_2\widehat{\mu}_1^2 - \frac{p_2}{n_1}\widehat{\sigma}_1^2, \widehat{\mu}_1\widehat{\mu}_2\right) = \frac{1}{n_1}\left(m_{3,1}^*\mu_2^* + (2p_2-3)\,\mu_1^*\mu_2^*\sigma_1^{*2} - \mu_1^{*3}\mu_2^*\right) + \mathcal{O}\left(\frac{1}{n^2}\right).$$

Putting all terms together leads to the following expression for the variance of the estimator $\widehat{\psi}_{p_1,p_2}$:

$$\begin{aligned}
\operatorname{Var}\left[\widehat{\psi}_{p_1,p_2}\right] = &\frac{1}{n_1}\Big(p_1^2m_{4,1}^* - 4p_1^3m_{3,1}^*\mu_1^* - 4p_1^2p_2m_{3,1}^*\mu_2^* + p_1^2\left(p_1^2-2p_2^2\right)\sigma_1^{*2}\mu_1^{*2} + 4p_1^2p_2^2\sigma_1^{*2}\mu_2^{*2} \\
&\quad -4p_1^2p_2\left(2p_2-3\right)\mu_1^*\mu_2^*\sigma_1^{*2} + p_1^2\left(3-4p_2\right)\mu_1^{*4} + 4p_1^2p_2\mu_1^{*3}\mu_2^* - p_1^2\sigma_1^{*4}\Big) \\
&+\frac{1}{n_2}\Big(p_2^2m_{4,2}^* - 4p_2^3m_{3,2}^*\mu_2^* - 4p_1p_2^2m_{3,2}^*\mu_1^* + p_2^2\left(p_2^2-2p_1^2\right)\sigma_2^{*2}\mu_2^{*2} + 4p_1^2p_2^2\sigma_2^{*2}\mu_1^{*2} \\
&\quad -4p_1p_2^2\left(2p_1-3\right)\mu_1^*\mu_2^*\sigma_2^{*2} + p_2^2\left(3-4p_1\right)\mu_2^{*4} + 4p_1p_2^2\mu_1^*\mu_2^{*3} - p_2^2\sigma_2^{*4}\Big) + \mathcal{O}\left(\frac{1}{n^2}\right) \\
= &\frac{p_1^2}{n_1}\Big(m_{4,1}^* - 4m_{3,1}^*\left(p_1\mu_1^* + p_2\mu_2^*\right) + \sigma_1^{*2}\left(\left(p_1^2-2p_2^2\right)\mu_1^{*2} - 4p_2\left(2p_2-3\right)\mu_1^*\mu_2^* + 4p_2^2\mu_2^{*2}\right) \\
&\quad +\mu_1^{*3}\left(3\mu_1^* + 4p_2\Delta\mu\right) - \sigma_1^{*4}\Big) \\
&+\frac{p_2^2}{n_2}\Big(m_{4,2}^* - 4m_{3,2}^*\left(p_1\mu_1^* + p_2\mu_2^*\right) + \sigma_2^{*2}\left(4p_1^2\mu_1^{*2} - 4p_1\left(2p_1-3\right)\mu_1^*\mu_2^* + \left(p_2^2-2p_1^2\right)\mu_2^{*2}\right) \\
&\quad +\mu_2^{*3}\left(3\mu_2^* - 4p_1\Delta\mu\right) - \sigma_2^{*4}\Big) + \mathcal{O}\left(\frac{1}{n^2}\right),
\end{aligned}$$

which is as desired. □

4

Conclusion

In this thesis, estimators for the mean and variance of a population are developed in the case of stratified sampling with two subgroups. These estimators are especially important when the sample is not representative for the population. Therefore, the estimators are modified such that the results are representing the complete population. These estimators are compared with original, naive ones on their bias and variance. It is preferable to obtain an unbiased estimator with a small variance. For both the population mean and the variance, this is achieved.

Estimating the true mean of the population can be done with the naive estimator $\overline{X}$, which adds all measurements and divides it by the total number of measurements. The group proportions are not taken into account with this estimator and it is generally biased. Weighing the measurements according to their proportions in the population results in two other estimators for μ^*. When these true proportions are unknown, they have to be guessed based on the available information, resulting in the estimator $\widehat{\theta}_{p_1,p_2}$. When the true proportions can be used, the estimator is called $\widehat{\theta}_{\text{oracle}}$. The former of these is also biased, while the latter is unbiased.

The estimation of the true variance of the population is a more complicated task. Originally, it can be done with an estimator not taking the subgroups into account: $\widehat{\text{Var}}[X]$. This is in general a biased estimator. In search of an unbiased estimator with a smaller variance, the estimator $\widehat{\gamma}_{p_1,p_2}$ is developed. It represents the true variance of the population where all true parameters are replaced by estimated ones. It is still biased, because some of the used plug-in estimators are biased. Therefore, also when the true proportions are available and used such that the estimator $\widehat{\gamma}_{\text{oracle}}$ appears, the estimator is biased. In search of an unbiased estimator, $\widehat{\gamma}_{p_1,p_2}$ is improved by using unbiased plug-in estimators for μ_1^{*2} and μ^{*2} instead of the biased ones of before. The new estimator is denoted by $\widehat{\psi}_{p_1,p_2}$. Its bias is smaller, but still not equal to zero. Only when the true proportions are used in the estimator $\widehat{\psi}_{\text{oracle}}$, a truly unbiased estimator for the variance is found. This estimator even has a smaller variance than $\widehat{\gamma}_{\text{oracle}}$, which makes it a better estimator. Likewise, $\widehat{\psi}_{p_1,p_2}$ has a smaller variance than $\widehat{\gamma}_{p_1,p_2}$. Therefore, the estimator $\widehat{\psi}_{p_1,p_2}$ should be used to estimate the true population variance if the true proportions are unavailable.

All estimators perform better when the number of measurements is increased. Having information about the population at one's disposal is also of major importance for estimating the population parameters. This creates the opportunity to take a better guess of the proportions, or even use the true ones, which results in an estimate closer to the real value. Using wrong proportions, however, can even lead to a very bad estimate. In some of those cases, the naive estimators will perform better than the newly developed estimators. One must therefore be really careful in choosing how to estimate and in guessing the proportions.

For both the population mean and variance, an unbiased estimator has been found. Both estimators make use of the true proportions, which are not always available. When there is not enough information, it is still important to perform a good estimation, but this might be biased. An interesting question for future research would be whether it is possible to prove that there does not exist an unbiased estimator if the true proportions are not available.

The techniques used in this thesis can also be extended on the matter of linear regression. This possible future work would extend the knowledge on the mean and variance with the estimation of more parameters: the regression coefficients. Another possible extension is the generalisation of the developed estimators for a greater number of subgroups. All proofs are executed for two strata, but one can imagine the existence of multiple subgroups. The computations and proofs are likely to remain similar for k strata, but the expressions will have to be generalised.

Bibliography

[1] A. Chaudhuri and H. Stenger. *Survey Sampling*. Taylor & Francis Group, 2 edition, 2005. ISBN 9780824757540.

[2] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. ISSN 01621459. URL http://www.jstor.org/stable/2280784.

[3] D. Pfeffermann and C.R. Rao. *Sample Surveys: Design, Methods and Applications*. Elsevier, 2009. ISBN 9780444562531.

[4] K. Thangavelu and E. Brunner. Wilcoxon–Mann–Whitney test for stratified samples and Efron's paradox dice. *Journal of Statistical Planning and Inference*, 137(3):720–737, 2007. ISSN 0378-3758. doi: https://doi.org/10.1016/j.jspi.2006.06.005. URL https://www.sciencedirect.com/science/article/pii/S0378375806001376.