Dissertation

ROBUST OPTIMIZATION FOR SURVEY STATISTICAL PROBLEMS

zur Erlangung des akademischen Grades eines
Dr. rer. pol.

Dem Fachbereich IV – VWL
der Universität Trier vorgelegt

von

M.Sc. Mohammad Asim Nomani

Trier, 2019

Eingereicht am: 01.07.2019
Disputation am: 02.08.2019

Gutachter: Dr. habil. Jan Pablo Burgard
Prof. Dr. Mirjam Dür
Prof. Dr. Ralf Münnich

Acknowledgements

First of all, I would like to thank Prof. Mirjam Dür for her support during my research work and in completing my Ph.D dissertation. I am not only grateful for having an opportunity to complete my Ph.D. under her supervision, but even more for her advice, encouragement, helpful suggestions and inspiring discussions. I would also like to thank Prof. Ralf Münnich for his guidance throughout my PhD program and providing his insights in the practical applications of this work.

Further, I would like to thank Dr. Jan Pablo Burgard for his continuous support and help. His help and suggestion provided a fruitful ground for the completion of this dissertation. I would also like to thank all my colleagues at ALOP especially Laura Somorowsky, Julian Wagner and Patrick Groetzner for creating an enjoyable and working atmosphere.

Finally, my thanks goes to all friends and my family, who have always supported me and encouraged me in completion of this work and in my academic career.

The research was financially supported by the German Research Foundation (DFG) within the research training group 2126 Algorithmic Optimization.

Mohammad Asim Nomani
Trier, June 2019

German Summary (Zusamenfassung)

In dieser Arbeit wird das Problem der Stichprobenallokation in stratifizierten Designs unter Unsicherheit untersucht. Im Allgemeinen sind die schichtspezifischen Varianzen, die zur Ermittlung der optimalen Lösung notwendig sind, nur näherungsweise bekannt. Dabei existieren meist keine genaueren Informationen zur Verteilung des Fehlers der Näherung. Ein weiterer Unsicherheitsfaktor bei der Allokation sind die Kosten für die Befragung einer Person in einer Schicht. Diese sind ebenfalls nur näherungsweise bekannt. Beispielsweise sind manchmal Personen für das Interview beim ersten Termin nicht verfügbar, und müssen in Folgeterminen befragt werden, was den Befragungsaufwand und damit die Kosten erhöht. In dieser Dissertation werden robuste Allokationen vorgeschlagen, um der Unsicherheit sowohl bei schichtspezifischen Varianzen als auch bei den schichtspezifischen Kosten zu begegnen. Diese Allokationen sind auch für ausschließlich unsichere Varianzen oder unsichere Kosten geeignet. Insgesamt werden daher drei verschiedene robuste Formulierungen vorgeschlagen, die diese verschiedenen Fälle darstellen. Zum Zeitpunkt der Einreichung dieser Dissertation ist dem Autor keine andere Forschungsarbeit bekannt, die die robuste Allokation für das Stichprobenallokationsproblem berücksichtigt.

Die erste robuste Formulierung für lineare Probleme wurde von (Soyster, 1973) vorgeschlagen. (Bertsimas and Sim, 2004) schlugen eine weniger konservative, robuste Formulierung für lineare Probleme vor. Wir untersuchen diese Formulierungen und erweitern sie für das Problem der nichtlinearen Stichprobenallokation. Es ist sehr unwahrscheinlich, dass alle schichtspezifischen Varianzen und Kosten unsicher sind. Die robusten Formulierungen sind so aufgebaut, dass wir wählen können, wie viele schichtspezifische Varianzen als unsicher gelten. Dies wird als Grad der Unsicherheit bezeichnet. Es wird bewiesen, dass eine Obergrenze für die Wahrscheinlichkeit einer Verletzung der nichtlinearen Beschränkungen berechnet werden kann, bevor das robuste Optimierungsproblem gelöst wird. Wir berücksichtigen verschiedene Arten von Datensätzen und berechnen robuste Allokationen. Wir führen mehrere Experimente durch, um die Qualität der robusten Allokationen zu überprüfen und sie mit den bestehenden Allokationsmethoden zu vergleichen.

Summary

In this thesis, we aim to study the sampling allocation problem of survey statistics under uncertainty. We know that the stratum specific variances are generally not known precisely and we have no information about the distribution of uncertainty. The cost of interviewing each person in a stratum is also a highly uncertain parameter as sometimes people are unavailable for the interview. We propose robust allocations to deal with the uncertainty in both stratum specific variances and costs. However, in real life situations, we can face such cases when only one of the variances or costs is uncertain. So we propose three different robust formulations representing these different cases. To the best of our knowledge robust allocation in the sampling allocation problem has not been considered so far in any research.

The first robust formulation for linear problems was proposed by (Soyster, 1973). (Bertsimas and Sim, 2004) proposed a less conservative robust formulation for linear problems. We study these formulations and extend them for the nonlinear sampling allocation problem. It is very unlikely to happen that all of the stratum specific variances and costs are uncertain. So the robust formulations are in such a way that we can select how many strata are uncertain which we refer to as the level of uncertainty. We prove that an upper bound on the probability of violation of the nonlinear constraints can be calculated before solving the robust optimization problem. We consider various kinds of datasets and compute robust allocations. We perform multiple experiments to check the quality of the robust allocations and compare them with the existing allocation techniques.

Contents

Chapter 1

Introduction

1.1 Motivation

Survey statistics is a branch of statistics where we investigate the estimation of characteristics of the whole population on the basis of data collected from samples of the population. Sampling is the most important factor to decide the accuracy of any survey study. The accuracy of the final results in terms of information about the whole population is directly affected by the selected samples. These samples are selected on the basis of available information of the population. The information about the population can be collected from the data of previous surveys conducted on the same population. The selection of samples becomes more complex when we have some uncertainty in the data. The objective of sample selection is to minimize the total variance of the estimator of the population and/or to minimize the total cost of conducting the survey. If only one characteristic about the population is studied, then the problem is called univariate sampling allocation problem and when several characteristics of the population are studied, then the problem is called multivariate sampling allocation problem. A multivariate sampling allocation problem is a multicriteria optimization problem where we try to find an allocation that minimizes the total variances of the estimators of several characteristics of the population.

The problems of survey statistics are often formulated as optimization problems which generally are of high dimension. Standard optimal allocation might fail or lead to results that are not feasible in the case when uncertainty takes place. Generally, we do not have information about the type of uncertainty existing in the data, so stochastic approaches are not very helpful. A robust optimization approach guarantees feasibility even in the worst case of uncertainty. In robust optimization sometimes in order to ensure feasibility, we lose much of the quality of the solutions if we take the uncertainty in a very pessimistic way. Therefore, new robust allocation techniques need to be developed or existing approaches need to be improved in order to reduce the loss of quality and to ensure feasibility.

This work shows how mathematical optimization can be used to obtain satisfactory solutions of uncertain survey statistics problems. Starting with describing uncertainty in survey statistics (Chapter 2), we introduce some robust optimization modeling techniques (Chapter 3). Some less pessimistic robust models are proposed. These less pessimistic robust allocation models can also be extended to the multicriteria cases. We integrate these less pessimistic robust models for the univariate and multivariate cases and propose new robust allocation models (Chapter 4). Apart from the development of new robust allocation techniques, we present some real data studies and simulation work also.

1.2 Outline

A brief overview of each chapter is presented here. Chapter 2 and Chapter 3 introduce the fundamentals of survey statistics and robust optimization, respectively. Chapter 4 discusses the more specific problems in robust optimization when multiple objectives exist. Chapter 4 also deals with the robust allocation approaches and related robust formulations. In Chapter 5, Chapter 6 and Chapter 7 we present applications of robust formulations from the previous chapters in simulated, synthetic and real life datasets.

Chapter 2: Fundamentals of Survey Statistics

In this chapter, we discuss some fundamentals of survey statistics that are needed in the overall study. We start with the sampling design in stratified sampling and then explain the mathematical formulations of stratified sampling allocation problems. We also discuss the case of uncertainty existing in the stratum specific variances and the formulation of uncertain sampling allocation problems.

Chapter 3: Fundamentals of Robust Optimization

This chapter starts with a case study of a dairy problem. We will see how uncertainty can make a feasible solution completely infeasible and practically meaningless. In order to deal with the uncertainty of parameters some basics of robust optimization are introduced and existing robust counterparts are presented.

Chapter 4: Robust Allocation in Survey Statistics

In this chapter we formulate the uncertain sampling allocation problem presented in Chapter 2 as robust formulation. It has also been proved that the robust approach such as presented by (Soyster, 1973) and (Bertsimas and Sim, 2004) for linear programming can also be applied to our nonlinear stratified sampling allocation problems. Theoretical results of robust formulations are also proved again for our specific problem.

Chapter 5: Analysis with Simulated Data

In this chapter simulated data is generated for the survey statistical problem. The variables are considered to be distributed diversely. Robust formulations are solved for three different cases when only cost is uncertain, when only variance is uncertain and when both cost and variance parameters are uncertain. In order to check the robust solutions, stability and feasibility tests have been carried out.

Chapter 6: Robust Allocation in the AMELIA Dataset

In this chapter, we consider a synthetically generated dataset with around 3.7 million observations for our sampling allocation problem. The idea is to check the robust approach for more complex datasets. Provinces in the AMELIA dataset have been merged together in order to make the data more complex statistically but computationally easier. Feasibility tests have been carried out for the robust solutions.

Chapter 7: NRW Income and Taxation Data

After working with simulated and synthetic datasets in the previous chapters, this chapter continues with a real life complex dataset. We consider here income and taxation data from 2001 of the German state of NRW. Characteristics of the data have been studied and on the basis of the characteristics, stratification of the population has been carried out. Robust allocations are achieved using the robust formulations introduced in the previous chapters. Feasibility tests have been carried out for the robust solutions.

Chapter 2

Fundamentals of Survey Statistics

In order to make strategic decisions about a whole population, surveys play an important role, see for example (LeRoux and Wright, 2010). Surveys provide information such as the health status of a population, economic activity and educational situation of the population, and this statistical information is the basis for taking decisions. For example, (Hollederer, 2011) studied the German microcensus 2005 and found that there are various types of interactions between health and occupational status of people. This kind of insight helps the policy makers to take major decisions about the population. This is why survey statistics are widely used in these decision making processes in order to obtain accurate decisions, see (Hoffmann et al., 2000) and (LeRoux and Wright, 2010).

Survey statistics also involves financial risks and accuracy risks. If the implementation of a survey is not done using optimal strategies it can result in including more people in the research than required or too few people which will result in increased cost or accuracy problems respectively, for details see (Swanson and Holton, 2005), p. 50. One of the common approaches is to take samples from the population such that they represent the whole of the population. Inference about the whole population is drawn from the selected samples. German microcensus is one of the examples of how survey statistics can be helpful in conducting surveys efficiently. In German microcensus single stage stratified cluster sampling is used since 1972 and federal states are considered as strata. The sampling fraction is taken to be 1% of the total population (about 800,000 people), see (GESIS, 2019).

Definition 2.1 (UNECE (2000))**.** *A survey is an investigation about the characteristics of a given population by means of collecting data from a sample of that population and estimating their characteristics through the systematic use of statistical methodology.*

Sampling can be classified in probability sampling and non-probability sampling. Non-probability sampling is often referred to as purposive sampling, see (Teddlie and Yu, 2007). Probability sampling is used when the probability of selecting a sample

according to the given objective is known. When the probability of selecting a sample is not known, we use the non-probability sampling. In this study we will not use non-probability sampling as the non-probability sampling relies more on researchers decision of selection of the samples and the results of non-probability sampling can sometimes lead to a biased outcome, (Yeager et al., 2011). Some of the differences in probability sampling and non-probability sampling can be listed as follows in Table 2.1.

Alternatively, probability sampling is also known as random sampling because the basis of probability sampling is randomization. In probability sampling some of the following methods are used:

- **Simple Random Sampling (SRS)**

  Simple random sampling is the easiest form of probability sampling and provides a basis for almost all kinds of complex probability sampling. In simple random sampling, the samples can be selected in two ways: with replacement and without replacement. In SRS with replacement, the same unit may be included in the sample more than once whereas in SRS without replacement all units in the sample are distinct. Simple random sampling is discussed in detail in Section 2.1.

- **Stratified Sampling**

  In stratified sampling the whole population is divided into many subgroups that we call strata. Then simple random sampling is applied to take samples from each stratum. The population is divided into heterogeneous strata which are homogeneous within themselves. The strata are formed according to the interest of the investigator. For example, for a study related to income and taxation we can divide the population into subgroups on the basis of their income level or the taxation level. Stratified sampling is discussed in detail in Section 2.2.

- **Cluster Sampling**

  In cluster sampling, the whole population is divided into a collection of population units called clusters. So in cluster sampling each unit of the sample is a collection of population units. Simple random sampling is used for selecting clusters. Suppose we want to survey all the people living in villages of Rhineland-Palatinate. However we do not have a list of all people living in villages. We can consider the villages as clusters and take a simple random sampling of all the villages. This is useful in reducing the total cost of conducting the survey as it can directly reduce the travelling cost of the interviewer. However, people living in the same village might have the same characteristrics, so stratified sampling or simple random sampling can provide better precision than cluster sampling. In several studies where the travelling cost can be very high, for example in forestry, cluster sampling is very efficient. For example (Philippi, 2005) uses adaptive cluster sampling to estimate the abundances of low abundance plants.

Probability sampling	**Non-probability sampling**
The samples are selected randomly so each unit of the population has an equal chance of being selected.	The samples are not selected randomly. Hence, each unit does not have an equal chance of being selected.
The probability of selecting a unit is known and equal for all the units.	The probability of selecting a unit is not equal and it is either unknown or not specified.
Probability sampling is helpful when the research is conclusive, for example in the German microcensus (Schwarz, 2001).	If the research is non exploratory, then non-probability sampling can provide better results, see for example (Schreuder and Alegria, 1995)
(Schreuder and Alegria, 1995) discuss the case when probability sampling is used to estimate population totals if the probabilities of selection were unequal and unknown. They also mention that this can introduce a probabilistic bias, which can be large.	(Särndal et al., 1992) discussed several non-probability sampling methods including balance sampling and quota sampling methods. These two methods select samples in such a way that many units of the population have zero probability of selection. This approach might provide an accurate estimation of population characteristics but an objective measure of precision is not possible.
Probability samples represent the population in a more effective way and they have a broader appeal and support, see (Hansen et al., 1983).	Many authors discussed that the word "representative" can be subject to a wide interpretation. Some other authors say that a sample with $n < N$ elements of a population of size N can never represent the population, whether it is chosen probabilistically or not. See, (Kruskal and Mosteller, 1979) and (Schreuder et al., 2001).

Table 2.1: Probability and non-probability sampling

- **Systematic Sampling**

  In systematic sampling, the first sample unit is selected randomly and then other samples are seleted according to the first sample but with a fixed interval size. This interval size can be calculated by dividing the population with the total

sample size. For example, if we have a population of 100 people and we need to select a sample of 8 people for a survey, then we have an interval size of $\lfloor \frac{100}{8} \rfloor = 12$. The first unit of the sample is selected randomly, let us say 9 is selected. Now according to systematic sampling, the other sample units wil be 21, 33, 45, 57, 69, 81, 93. The selected sample units represent the population as it is very unlikely to happen that each 12^{th} unit of the population has the same characteristic.

There are some other methods also, such as two stage sampling and multi stage sampling. For details, we suggest the readers to see (Lohr, 2010).

2.1 Simple Random Sampling

Simple random sampling is a very basic form of probability sampling. In simple random sampling the sample units are selected randomly from the population and each unit of the population has an equal chance of being selected, see (Olken and Rotem, 1986). There are two ways to draw samples in simple random sampling: with replacement in which the same unit can be chosen again, and without replacement in which each unit can be selected only once.

In simple random sampling with replacement (SRSWR), if we have a sample size of n and a population size of N then the first sample unit is selected randomly from the population with an equal probability of being selected. The selected sample unit is replaced into the population so for selecting the second sample unit each population unit has again equal probability of being selected. The procedure is repeated till n sample units are selected. The probability of selecting a unit in each draw is $1/N$.

In SRSWoR each sample is equally likely and there are $\binom{N}{n}$ possible samples. In the first draw each unit has equal probability of $1/N$ of being selected. For the next draw, unlike SRSWR, the selected unit is not replaced and a unit is chosen from the remaining population of size $N-1$. This process is repeated untill n sample units are selected. In SRSWoR, the probability of selecting any individual sample S can be written as follows:

$$P(S) = \frac{1}{\binom{N}{n}} = \frac{n!\,(N-n)!}{N!}$$

Let i be a unit contained in the population of size N. Let π_i denote the probability for unit i to be included in a sample of size n using SRSWoR. If unit i is in the sample, then the remaining $n-1$ sample units must be chosen from the remaining $N-1$ units of the population. Since there are $\binom{N-1}{n-1}$ possibilities to do so, we obtain

$$\pi_i = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$$

In SRSWR the selected sample can have duplicates from the population and sampling the same individual twice does not increase any information about the population. This is the reason that simple random sampling without replacement(SRSWoR) is preferred.

Definition 2.2 (Lohr (2010))**.** *For any sampling design, we define the sampling weight of unit i to be the reciprocal of the inclusion probability:*

$$w_i = \frac{1}{\pi_i}$$

Sampling weights are used in design based sampling to achieve proportionality. The sampling weight of an individual unit i can be interpreted as the number of units in the population which are represented by unit i. In SRS each unit has equal probability of inclusion and thus equal sampling weight. We can interpret this as follows: each unit in the sample represents itself and all other units of the population that are not in the sample. We refer to (Lohr, 2010) and (Särndal et al., 2003) for a detailed study.

2.2 Stratified Sampling Allocation Problems

One important kind of sampling design is stratified random sampling that is widely used in practice e.g. in the German microcensus, (Schwarz, 2001). Stratified sampling leads to efficient selection of samples when the total population is heterogeneous in nature. In stratified sampling, the total heterogeneous population of size N is divided into H subsets of sizes $N_1, N_2, \ldots, N_H$ as denoted in (Sukhatme, 1954), where

$$N = N_1 + N_2 + \ldots + N_H.$$

The process of dividing the population into smaller subsets is called stratification. These subsets are called strata, they are homogeneous within themselves but heterogeneous among each other. We need some prior information about the population in order to identify homogeneous strata.

Figure 2.1 taken from (Pinterest, 2019) explains how stratified sampling exactly works by identifying homogeneous strata and then selecting samples from these strata. In stratified sampling we take a sample of size n_h from stratum h, and these n_h units are selected using simple random sampling. The population quantities for a variable of interest y can be defined as follows:

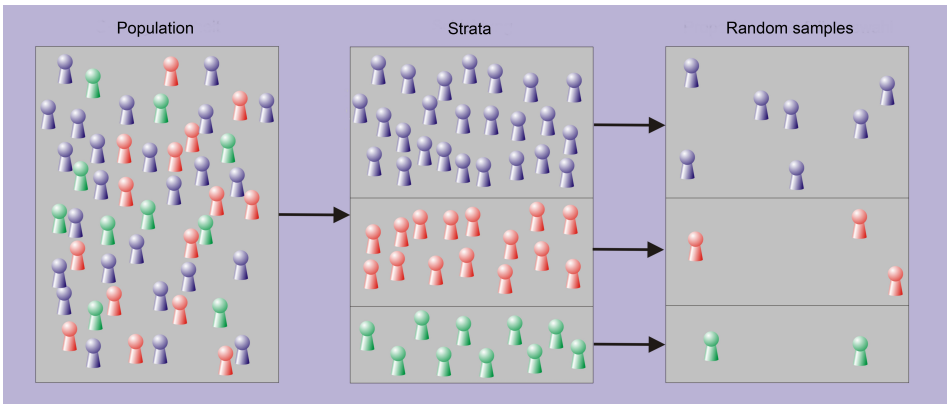


Figure 2.1: Representation of stratification and random samples (Pinterest, 2019)

$$
\begin{aligned}
y_{hj} &= \text{Value of unit } j \text{ in stratum } h \\
Y_h &= \sum_{j=1}^{N_h} y_{hj} = \text{Population total in stratum } h \\
Y &= \sum_{h=1}^{H} Y_h = \text{Population total} \\
\bar{Y}_h &= \frac{\sum_{j=1}^{N_h} y_{hj}}{N_h} = \text{Population mean in stratum } h \\
\bar{Y}_U &= \frac{Y}{N} = \frac{\sum_{h=1}^{H}\sum_{j=1}^{N_h} y_{hj}}{N} = \text{Overall population mean} \\
S_h^2 &= \sum_{j=1}^{N_h} \frac{(y_{hj} - \bar{Y}_h)^2}{(N_h - 1)} = \text{Population variance in stratum } h
\end{aligned}
\tag{2.1}
$$

For convinience reasons we assume WOR sampling if not stated otherwise. Define $\mathbb{S}_h$ to be the set of n_h units in simple random sampling for stratum h. We have $|\mathbb{S}_h| = n_h$ for each $h = 1, \ldots, H$. The notations in (2.1) for the sample within each stratum can be defined as follows:

$$
\begin{aligned}
\bar{y}_h &= \frac{\sum_{j\in\mathbb{S}_h} y_{hj}}{n_h} = \text{ Sample mean} \\
s_h^2 &= \sum_{j\in\mathbb{S}_h} \frac{(y_{hj} - \bar{y}_h)^2}{(n_h - 1)} = \text{Sample variance} \\
\hat{y}_h &= \frac{N_h}{n_h} \sum_{j\in\mathbb{S}_h} y_{hj} = N_h \bar{y}_h = \text{Estimate of the population total in stratum } h
\end{aligned}
$$

In stratum h, we have a population of N_h units and a sample of n_h units selected using SRS. Then we can estimate $\bar{Y}_U$ and Y_h by $\bar{y}_h$ and $\hat{y}_h$ respectively. As given in (2.1) the population total is $Y = \sum_{h=1}^{H} Y_h$ and can be estimated as follows:

$$\hat{Y}_{str} = \sum_{h=1}^{H} \hat{y}_h = \sum_{h=1}^{H} N_h \bar{y}_h,$$

and the overall population mean $\bar{Y}_U$ can be estimated as follows:

$$\bar{Y}_{str} = \frac{\hat{Y}_{str}}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h.$$

Some more details on estimators can be found in Cochran (1977).

In stratified sampling, the variance $V(\hat{Y}_{str})$ of the estimator $\hat{Y}_{str}$ can be calculated as follows:

$$V(\hat{Y}_{str}) = \sum_{h=1}^{H} \frac{N_h^2 S_h^2}{n_h} (1 - \frac{n_h}{N_h}), \tag{2.2}$$

where S_h^2 represents the stratum specific variance of the characteristic under study in stratum h.

In SRS, we saw in Section 2.1 that the inclusion probabilities π_i and the sampling weights $w_i = 1/\pi_i$ are equal. However, in stratified sampling, the inclusion probabilities can vary from one stratum to another. Thus the sampling weights can also be different. The stratified sampling estimator $\hat{Y}_{str}$ can be written as weighted sum of the individual sampling units as follows:

$$\hat{Y}_{str} = \sum_{h=1}^{H} N_h \bar{y}_h = \sum_{h=1}^{H} \sum_{j \in \mathbb{S}_h} \frac{N_h}{n_h} y_{hj}$$

and following this, the estimator of the population total can be written as follows:

$$\hat{Y}_{str} = \sum_{h=1}^{H} \sum_{j \in \mathbb{S}_h} w_{hj} y_{hj},$$

where $w_{hj} = \frac{N_h}{n_h}$ is the sampling weight for unit j in stratum h. Here this sampling weight can be understood as the number of population units represented by the selected unit j in stratum h. The probability π_{hj} of inclusion of unit j in stratum h can be written as follows:

$$\pi_{hj} = \frac{n_h}{N_h}$$

Thus the sampling weight can be calculated as the reciprocal of the inclusion probability, i.e.

$$w_{hj} = \frac{1}{\pi_{hj}} = \frac{N_h}{n_h}.$$

There are various approaches for the selection of samples from these strata such as, for example, equal allocation where equal sample sizes are allocated to each stratum irrespective of its size and its stratum specific variance. If the total sample size is β, then the sample size n_h of the sample in stratum h in equal allocation is calculated as follows:

$$n_h = \frac{\beta}{H}, \quad \forall h = 1, \ldots, H.$$

In some applications, equal allocation might allocate a sample size bigger than the stratum size, i.e. $n_h > N_h$, and that is practically impossible.

This problem can be avoided by using a different well known method of allocating samples in stratified sampling: proportional allocation. In proportional allocation the sample sizes are allocated according to the size of the strata i.e.:

$$n_h \propto N_h, \quad \forall h = 1, \ldots, H.$$

However, the proportional allocation allocates sample sizes irrespective of the stratum specific variances which might lead to an increase in the total variance of the estimator. This shows an important perspective to be considered while allocating samples to the strata, that is, minimizing the total variance of the estimator.

Neyman (1959) and Tschuprow (1923) proposed the following optimum allocation minimizing the variance of the estimator using a standard Lagrangian approach.

$$n_h = \frac{N_h S_h}{\sum_{i=1}^{H} N_i S_i} \beta, \quad \forall h = 1, \ldots, H \tag{2.3}$$

In the sampling allocation problem sometimes the fixed cost of selecting a unit sample is also known. In this case, we can consider the total cost of allocating samples as an additional objective. If C_h denotes the cost of selecting a unit sample in stratum h ($h = 1, \ldots, H$), then Cochran suggested an allocation method which minimizes the product of the variance of the total estimate and the total cost (see Section 5.5, Cochran (1977)). Abbreviating $W_h = N_h/N$, we can calculate the allocation using the following formula:

$$n_h = \beta \frac{W_h S_h / \sqrt{C_h}}{\sum_{h=1}^{H} W_h S_h \sqrt{C_h}} \quad \forall h = 1, \ldots, H.$$

However, there exist problems where there are several characteristics of the total population under study. These characteristics are referred to as variables in survey statistics. A sampling allocation problem with multiple variables is called multivariate sampling allocation problem. In this situation we want to minimize several total variances in the optimum sampling allocation problem. We assume there are $K > 1$ characteristics. For $k = 1, \ldots, K$ and $h = 1, \ldots, H$, we define S_{hk}^2 to represent the stratum specific variance of characteristic k in stratum h. In multivariate sampling allocation, we estimate the various characteristics of the population. Conidering the K charactersitics separately might lead to K different sample allocations in the population. However, considering them together may lead to yet another allocation. Chatterjee (1967) proposed an allocation method by minimizing the sum of the relative increases in the variance of the estimates. Chatterjee's allocation can be calculated as follows:

$$n_h = \frac{C_h \sqrt{\sum_{k=1}^{K} n_{hk}^{*2}}}{\sum_{h=1}^{H} C_h \sqrt{\sum_{k=1}^{K} n_{hk}^{*2}}} \quad \forall h = 1, \ldots, H$$

where

$$n_{hk}^{*} = \frac{C_h W_h S_{hk} / \sqrt{C_h}}{\sum_{h=1}^{H} W_h S_{hk} \sqrt{C_h}} \quad \forall h = 1, \ldots, H; k = 1, \ldots, K$$

All of the above mentioned allocation methods suffer from three disadvantages:

- Sometimes we encounter the problem of over allocation in a stratum, i.e., $n_h > N_h$. This happens for example if both the stratum specific variance S_h^2 and the unit cost C_h are very low.
- If the stratum specific variance S_h^2 and the cost C_h are very high, then that stratum might be assigned an extremely low sample size which might not give a good representation of the stratum population.
- All of these methods give non integer solutions which we need to round off before actual use.

These disadvantages in the sampling allocation inspired the mathematical formulation of box constrained sampling allocation problems by Münnich et al. (2012).

2.3 Mathematical Formulation

In order to deal with the problem of over allocation and low sample size, Münnich et al. (2012) and Gabler et al. (2012) added box constraints to the sampling allocation

problem where lower and upper bounds m_h and M_h on the variables n_h are defined, such that $m_h \geq 2$ (as we need at least two sample units from each stratum in order to calculate the stratum specific variances) and $M_h \leq N_h$ (as n_h should not be bigger than the size N_h of the population in stratum h). The following constraints are added to the optimization problem:

$$m_h \leq \ n_h \leq M_h \quad \forall h = 1, \ldots, H \quad \text{and} \quad \sum_{i=1}^{H} n_i \leq \beta.$$

These bounds make sure that the optimal allocation does not have more sample units allocated to a stratum than the stratum size and the sum of the sample sizes does not exceed the total sample size. The objective function of this mathematical formulation is the total variance of the estimator defined in (2.2), i.e.

$$V(\hat{Y}_{str}) = \sum_{h=1}^{H} \frac{N_h^2 S_h^2}{n_h}(1 - \frac{n_h}{N_h}) = \sum_{h=1}^{H} \frac{N_h^2 S_h^2}{n_h} - \sum_{h=1}^{H} N_h S_h^2.$$

However, the constant part in this variance can be ignored in the mathematical formulation. We get the following mathematical formulation:

$$\begin{aligned} \min & \sum_{h=1}^{H} \frac{N_h^2 S_h^2}{n_h} \\ \text{s.t.} & \sum_{h=1}^{H} n_h \leq \beta \\ & m_h \leq \ n_h \leq M_h \quad \forall h = 1, \ldots, H \end{aligned} \tag{2.4}$$

In this model, it is assumed that $\beta > 0$ and $2 \leq m_h < M_h \leq N_h$ for all h. We define $d_h := N_h^2 S_h^2$ for all $h = 1, \ldots, H$ and the above formulation can be rewritten as follows:

$$\begin{aligned} \min & \sum_{h=1}^{H} \frac{d_h}{n_h} \\ \text{s.t.} & \sum_{h=1}^{H} n_h \leq \beta \\ & m_h \leq \ n_h \leq M_h \quad \forall h = 1, \ldots, H \end{aligned} \tag{2.5}$$

(Friedrich et al., 2015) proposed three algorithms based on Greedy strategies to obtain an integer solution of the univariate sampling allocation problem (2.5). They use the fact that an integer solution can be obtained if the feasible set is a polymatroid.

The univariate problem (2.5) can be extended to the multivariate case. Assuming that we have K variables of interest, let us denote by S_{hk}^2 the stratum specific variance of variable k in statum h. We define

$$d_{hk} := N_h^2 S_{hk}^2 \quad \forall h = 1, \ldots, H \text{ and } \forall k = 1, \ldots, K$$

We get the following multi-criteria optimization problem:

$$\begin{aligned} \min & \sum_{h=1}^{H} \frac{d_{hk}}{n_h} \quad \forall k = 1, \ldots, K \\ \text{s.t.} & \sum_{h=1}^{H} n_h \leq \beta \\ & m_h \leq \ n_h \leq M_h \quad \forall h = 1, \ldots, H \end{aligned} \tag{2.6}$$

We can see here that the above problem (2.6) is a multi-objective optimization problem and it must be scalarized in order to get some Pareto optimal solution (Friedrich et al., 2018). There are several ways of doing this: one option is the so called weighted sum method, where we transform the problem to a single objective problem by considering a weighted sum of the K objective functions. For more details see Chapter 3 in (Ehrgott, 2005). Another option is the so called ε-constraint method, where we minimize only one of the original K objective functions, while the others are moved into the constraints by introducing an upper bound of ε_k for the objectrive function f_k. For more details we refer to Chapter 4.1 in (Ehrgott, 2005). However, it is always difficult to find appropriate weights of the different characteristics in the weighted sum method or to find the proper bounds ε_k in the ε-constraint method.

Several authors used other well established multiobjective optimization techniques to solve the multivariate sampling allocation problem in the form of problem (2.6). (Díaz-García and Cortez, 2008) solve the multivariate problem using the value function method. They also suggest a distance based method to find a compromise solution by minimizing the distance to the ideal point. (Friedrich et al., 2018) proposed several scalarization techniques such as weighted sum and p-norm scalarization method to deal with the multiple objectives and they proposed some standardization techniques where the objectives have been standardized by the unique univariate optimal allocations.

If the cost C_h of selecting a unit sample in stratum h is provided, then we have one more objective function to minimize the total cost

$$\min \sum_{h=1}^{H} C_h n_h + C_0,$$

where C_0 is the overhead cost which is a constant. However, we can ignore the constant term in this objective function. We get the following problem with $K + 1$ objective functions:

$$
\begin{aligned}
&\min \sum_{h=1}^{H} \frac{d_{hk}}{n_h} \quad \forall k = 1, \ldots, K \\
&\min \sum_{h=1}^{H} C_h n_h \\
&\text{s.t.} \sum_{h=1}^{H} n_h \leq \beta \\
&\quad m_h \leq \ n_h \leq M_h \quad \forall h = 1, \ldots, H
\end{aligned}
\tag{2.7}
$$

Whenever we have a given total budget C that our total cost must not exceed, we can transfer the cost function to the constraints by adding the following constraint:

$$
\sum_{h=1}^{H} C_h n_h \leq C
$$

The same can be done with the other objective functions: If an upper bound V_k on the the objective function $\sum_{h=1}^{H} \frac{d_{hk}}{n_h}$ is somehow available, then we can transfer this objective function to the constraints by adding the following constraint:

$$
\sum_{h=1}^{H} \frac{d_{hk}}{n_h} \leq V_k \tag{2.8}
$$

The problem of allocating samples in stratified sampling has been discussed continuously since the 1950s. Dalenius (1953) discussed two geometric approaches for allocating samples in stratified sampling where he solved the problem with two strata. Following that Kokan and Khan (1967) present an analytical solution for a certain multi stage sampling and double sampling problem. They also show existence and uniqueness of the solution in their specific allocation problem. Chaddha et al. (1971) proposed a dynamic programming approach to find an optimum solution for a univariate stratified sampling problem. Omule (1985) solved the multivariate sampling allocation problem using the same dynamic programming approach. (Chatterjee, 1967) solved the sampling allocation problem considering the cost objective function and calculating an upper bound on the variance constraint. (Sukhatme, 1954) formulated the sampling allocation problem as nonlinear optimization problem with cost as objective function. Recently, (Díaz-García and Garay-Tápia, 2007) investigated the same nonlinear optimization problem as (Sukhatme, 1954) by considering a cost objective function and nonlinear variance constraints with known upper bounds on the variances.

2.4 Uncertainty in Sampling Allocation Problems

In the stratified sampling allocation problem, the stratum specific variances are some given values used in the optimization process. Generally, the exact values for the stratum specific variances are not known because they are often based on inexact data or data from previous surveys that may not be valid at present time. So there is a very high chance that there are uncertainties in the stratum specific variances. The same applies to the cost of selecting a unit sample in each stratum. These costs are also subject to change.

Denote the true values by $\tilde{d}_{hk}$ and $\tilde{C}_h$. We assume interval uncertainty which means that we assume

$$\tilde{d}_{hk} \in [d_{hk} - \hat{d}_{hk}, d_{hk} + \hat{d}_{hk}] \quad \text{and} \quad \tilde{C}_h \in [C_h - \hat{C}_h, C_h + \hat{C}_h]$$

respectively. Here, d_{hk} and C_h are fixed numbers and $\hat{d}_{hk} \geq 0$ and $\hat{C}_h \geq 0$ are the possible deviations from d_{hk} and C_h respectively.

The box constrained optimum sampling allocation problem with interval uncertainty can be written as follows:

$$\begin{aligned} &\min \sum_{h=1}^{H} \frac{\tilde{d}_{hk}}{n_h} \qquad \forall k = 1, \ldots, K \\ &\min \sum_{h=1}^{H} \tilde{C}_h n_h \\ &\text{s.t.} \sum_{h=1}^{H} n_h \leq \beta \\ &\qquad m_h \leq \ n_h \leq M_h \quad \forall h = 1, \ldots, H \end{aligned} \tag{2.9}$$

Many authors have considered the uncertainty existing in the stratum specific variance. (Díaz-García and Garay-Tápia, 2007) solved the univariate nonlinear sampling allocation optimization problem using a stochastic optimization approach considering the fact that S_h^2 are generally unknown and the sample variances s_h^2 are random variables. (Diaz-Garcia and Ramos-Quiroga, 2011) proposed a stochastic matrix optimization approach to solve the multivariate sampling allocation problem where they minimize the estimated covariance matrix of the estimated means. In order to deal with the uncertainty, fuzzy programming has also been used for sampling allocation problems. (Gupta et al., 2014) proposed a chance constrained multivariate sampling allocation approach and used a fuzzy goal programming approach to find a compromise solution. (Ullah et al., 2015) proposed a fuzzy geometric programming approach for a two stage multivariate problem considering linear and quadratic cost functions. These methods do

not consider interval uncertainty but more general types of uncertainty. However, these approaches are computationally tractable only for small scale problem.

Another very useful approach to deal with the interval uncertainty is robust optimization. To the best of our knowledge, robust optimization has not been applied yet in the sampling allocation problems. In robust optimization, one generally considers optimization problems where the objective function does not contain uncertainties as it would result in an interval valued optimal value of the objective function which is very difficult to obtain from a computational point of view. Instead, an uncertain objective function is moved to the constraints by introducing an additional variable. We introduce new variables $\phi_k \in \mathbb{R}$ for all $k = 1, \ldots, K$ and $\phi_0 \in \mathbb{R}$ and rewrite (2.9) as follows:

$$
\begin{aligned}
&\min \phi_k \qquad \forall k = 1, \ldots, K \\
&\min \phi_0 \\
&\text{s.t.} \sum_{h=1}^{H} \frac{\tilde{d}_{hk}}{n_h} \leq \phi_k \qquad \forall k = 1, \ldots, K \\
&\qquad \sum_{h=1}^{H} \tilde{C}_h n_h \leq \phi_0 \\
&\qquad \sum_{h=1}^{H} n_h \leq \beta \\
&\qquad m_h \leq \ n_h \leq M_h \quad \forall h = 1, \ldots, H \\
&\qquad \phi_0 \in \mathbb{R}, \phi_k \in \mathbb{R} \quad \forall k = 1, \ldots, K
\end{aligned}
\tag{2.10}
$$

We can deal with the univariate problem ($K = 1$) in just the same way. If the cost function is considered as the second objective of the problem, then the uncertain univariate sampling allocation problem reads as follows:

$$
\begin{aligned}
&\min \ \phi \\
&\min \ \phi_0 \\
&\text{s.t.} \ \sum_{h=1}^{H} \frac{\tilde{d}_h}{n_h} \leq \phi \\
&\quad \sum_{h=1}^{H} \tilde{C}_h n_h \leq \phi_0 \\
&\quad \sum_{h=1}^{H} n_h \leq \beta \\
&\quad m_h \leq \ n_h \leq M_h \quad \forall h = 1, \ldots, H \\
&\quad \phi \in \mathbb{R}, \quad \phi_0 \in \mathbb{R}
\end{aligned} \tag{2.11}
$$

Here we have a bi-objective formulation of the sampling allocation problem. As discussed in Section 2.3, we can easily convert it to a single objective problem by calculating an upper bound on one of the objectives. Another issue with this problem is that we have uncertain parameters both in the linear and in the nonlinear constraints. Presence of uncertain parameters in the nonlinear constraints makes the above problem difficult from the computational point of view.

In this chapter, we presented some basics of survey statistics. We discussed various mathematical formulations of sampling allocation problems in stratified sampling. However, new formulations of sampling allocation problems are needed because of uncertainty existing in the parameters. Before discussing robust formulations of sampling allocation problems, we discuss some basics of robust optimization in the next chapter.

Chapter 3

Fundamentals of Robust Optimization

3.1 Why Robust Optimization?

It has been discussed that we have some data uncertainty in sampling allocation problems. In the presence of uncertainty we do not get the desired optimality using uncertain values of parameters, as the true values of the parameters might differ. In real world situations, this may lead to serious problems. In order to show how badly uncertainty can affect our solutions, we present a case study on a sampling allocation problem in dairy industry.

The sample allocation problem is taken from (Khan et al., 1997) (Example 2 of Section 4) and originally reported in Jessen (1942). The problem considers three characteristics under study with stratum specific variances and costs of selecting a sample unit in different strata given in Table 3.1.

h	1	2	3	4	5
C_h	3	4	5	6	7
N_h	39,552	38,347	43,969	36,942	41,760
S_{h1}^2	4.6	3.4	3.3	2.8	3.7
S_{h2}^2	11.7	9.8	7.0	6.5	9.8
S_{h3}^2	332	357	246	173	278

Table 3.1: Data for the dairy problem

The notations are as defined in Section 2.2. From the data in Table 3.1 we have $K = 3$ and $H = 5$ and we set $\beta = 1082$.

In this case study, we take into consideration proportional allocation, Cochran's allocation and Chatterjee's allocation. We calculated the sampling allocation with R

software using these techniques and we found

$$n^p = (197, 191, 219, 184, 208) \tag{3.1}$$

using proportional allocation,

$$n^C = (330, 244, 195, 123, 189) \tag{3.2}$$

using Cochran's allocation and

$$n^{CH} = (330, 245, 195, 123, 189) \tag{3.3}$$

using Chatterjee's allocation. We can calculate the total variances of these three methods using the formula (2.2).

Since Cochran's allocation and Chatterjee's allocation are optimal allocation methods using a compromise function, they have smaller total variances than the total variance of the proportional allocation method. For this reason we use the total variance V_k $(k = 1, \ldots, K)$ for proportional allocation as an upper bound on the total variance of the estimator as discussed in Formula (2.8). We get the following optimization problem:

$$\begin{aligned} \min \ & \sum_{h=1}^{H} C_h n_h \\ \text{s. t.} \ & \sum_{h=1}^{H} n_h \leq \beta \\ & \sum_{h=1}^{H} \frac{d_{hk}}{n_h} \leq V_k \qquad \forall k = 1, \ldots, K \end{aligned} \tag{3.4}$$

Observe here that the stratum specific variances are hardly known with high accuracy, so it is natural to consider them as uncertain. We study the effect of uncertainty only on the nonlinear constraints. The non uncertain constraint $\sum_{h=1}^{H} n_h \leq \beta$ is fulfilled by proportional, Cochran's and Chatterjee's allocation.

Let us assume that the uncertain values d_{hk} are 10% approximations of the unknown true values $\tilde{d}_{hk}$ and the true value $\tilde{d}_{hk}$ lies in the interval $[d_{hk} \pm 0.1 d_{hk}]$. For each of the three characteristics, we have one nonlinear constraint.

We conduct the following experiment: For each $h = 1, \ldots, H$ and $k = 1, \ldots, K$ we took 100 uniformly distributed random d_{hk} within the interval $[d_{hk} \pm 0.1 d_{hk}]$. With each set of parameters d_{hk}, we tested whether the proportional, Cochran and Chatterjee's allocation from (3.1)-(3.3) fulfills the constraints

$$\sum_{h=1}^{H} \frac{d_{hk}}{n_h} \leq V_k \qquad \forall k = 1, \ldots, K.$$

We investigated the effect of this uncertainty and here is what we found:

• We found that for most of the random parameter sets d_{hk} one or more of the nonlinear constraints $\sum_{h=1}^{H} \frac{d_{hk}}{n_h} \leq V_k$ $(k = 1, \ldots, K)$ were violated. The worst violations of the constraints (instead of $\sum_{h=1}^{H} \frac{\tilde{d}_{hk}}{n_h^*} - V_k \leq 0$) which we observed were:

Proportional allocation

for $k = 1$,

$$\sum_{h=1}^{H} \frac{\tilde{d}_{h1}}{n_h^*} - V_1 \geq 0.00035702$$

for $k = 2$,

$$\sum_{h=1}^{H} \frac{\tilde{d}_{h2}}{n_h^*} - V_2 \geq 0.00124568$$

for $k = 3$,

$$\sum_{h=1}^{H} \frac{\tilde{d}_{h3}}{n_h^*} - V_3 \geq 0.04095227$$

Cochran allocation

for $k = 1$,

$$\sum_{h=1}^{H} \frac{\tilde{d}_{h1}}{n_h^*} - V_1 \geq 0.0002639279$$

for $k = 2$,

$$\sum_{h=1}^{H} \frac{\tilde{d}_{h2}}{n_h^*} - V_2 \geq 0.0008612519$$

for $k = 3$,

$$\sum_{h=1}^{H} \frac{\tilde{d}_{h3}}{n_h^*} - V_3 \geq 0.02671216$$

Chatterjee allocation

for $k = 1$,

$$\sum_{h=1}^{H} \frac{\tilde{d}_{h1}}{n_h^*} - V_1 \geq 0.000261641$$

for $k = 2$,

$$\sum_{h=1}^{H} \frac{\tilde{d}_{h2}}{n_h^*} - V_2 \geq 0.0008546603$$

for $k = 3$,

$$\sum_{h=1}^{H} \frac{\tilde{d}_{h3}}{n_h^*} - V_3 \geq 0.02647204$$

• Considering the above worst case scenario could be a very pessimistic approach. So we consider a more realistic approach to know the violation intensity. We tested feasibility of the allocations from (3.1)-(3.3) with all of the randomly generated d_{hk}. What we found is that in many cases these allocations are heavily infeasible. With this experiment, Figure 3.1 is generated using R software. It explains the pattern of constraint violation. The green region in Figure 3.1 represents the feasibility with respect to the nonlinear constraints and the values outside this green region show infeasibility.

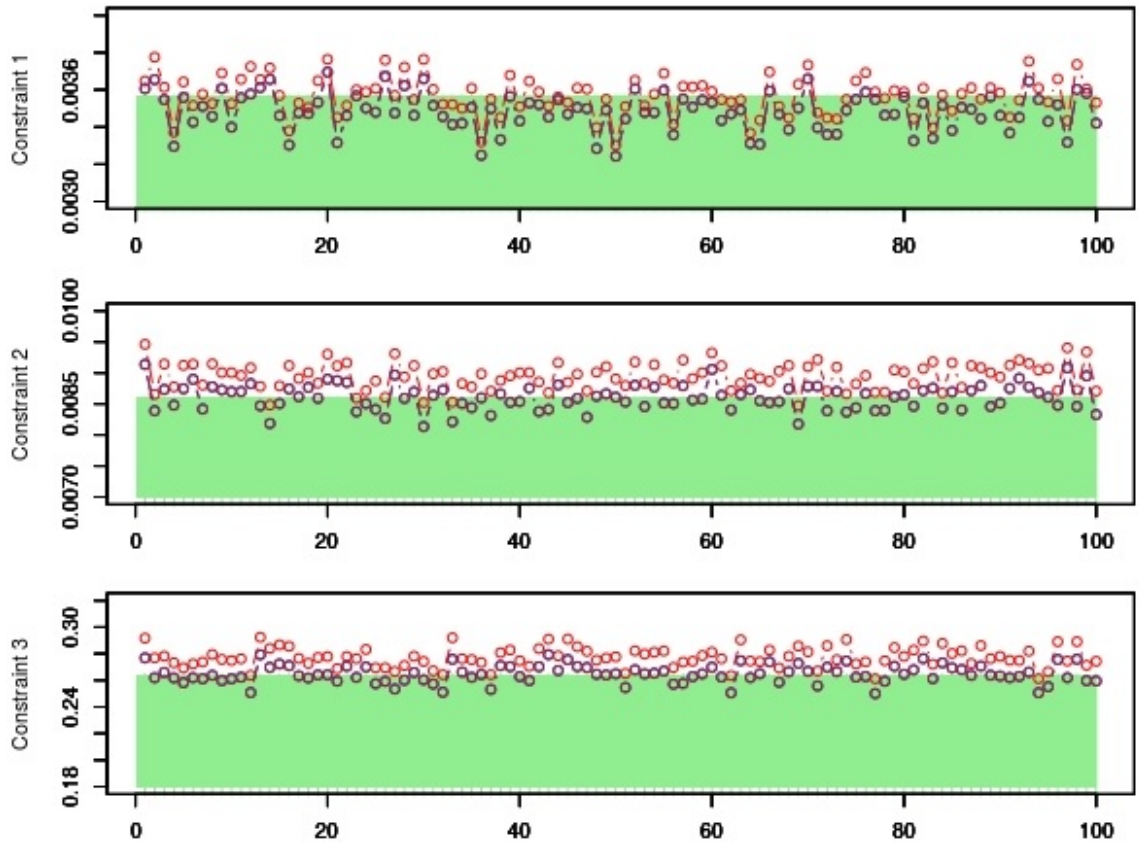


Figure 3.1: Constraint feasibility with 100 random paremeters

• Figure 3.2 represents the density plots for the scaled total variances obtained in

our experiment. For each V_1, V_2, V_3 and each allocation from (3.1)-(3.3), Figure 3.2 shows how often the quantity

$$\frac{1}{V_k}\sum_{h=1}^{H}\frac{\tilde{d}_{hk}}{n_h^*}$$

takes a certain value. Clearly, an allocation is feasible if and only if this quantity is less than or equal to 1. We can see that the first characteristic is least affected by the uncertainty whereas for the third characteristic, the allocations are highly infeasible.

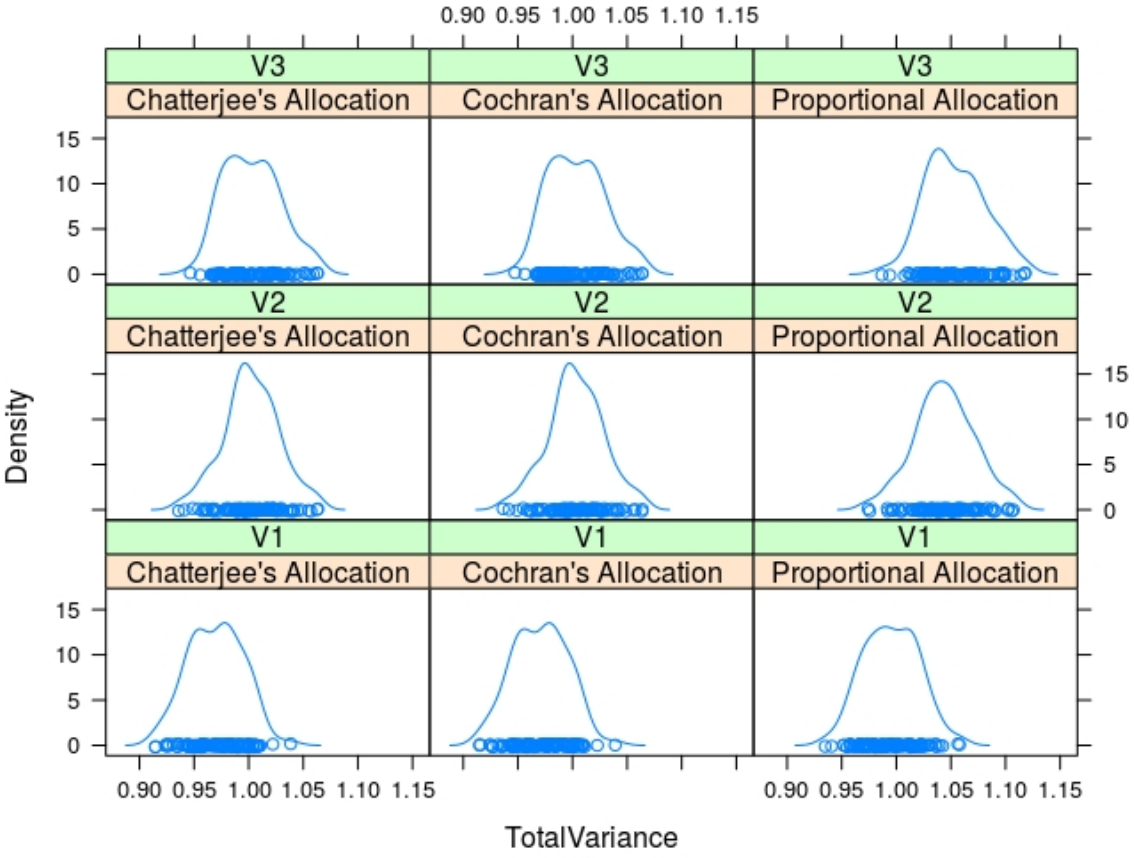


Figure 3.2: Density plot with uncertain parameters

• The same data can also be visualized as boxplots in Figure 3.3. We see that the uncertainty existing in the parameters makes the allocation infeasible in most of the cases.

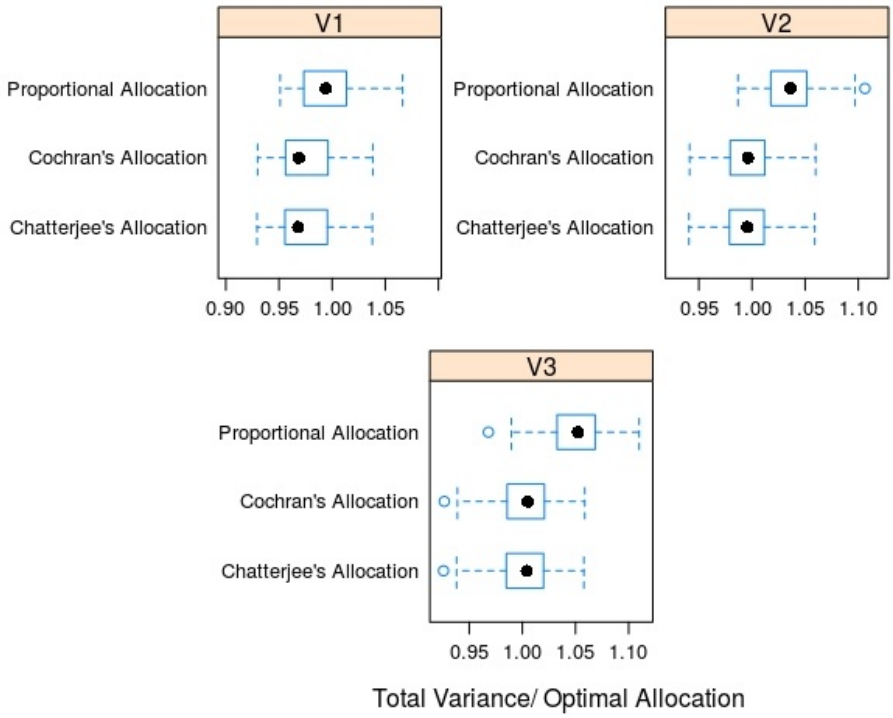


Figure 3.3: Boxplots of total variances for different allocations

Summarizing, we see that just 10% perturbation of the obviously uncertain stratum specific variances can make an allocation heavily infeasible. Such allocations, if uncertainty exists, are practically meaningless. We have seen in this section that it is necessary to consider the uncertainty and that it is important to compute a solution that is robust.

3.2 Principles of Robust Optimization

As we have seen in the previous section, uncertainty in the problem parameters is a serious issue which can severely affect the solutions of the underlying optimization problems. This was also observed by (Ben-Tal and Nemirovsky, 2000) who carried out a case study on linear optimization problems from the Net-Lib library. We quote their words about how important robustness is:

In real-world applications of Linear Programming, one cannot ignore the possibility that a small uncertainty in the data can make the usual optimal solution completely meaningless from a practical viewpoint.

We will now explain the principles of robust optimization and we will illustrate

this by looking at linear optimization problems. Although our sampling allocation problems lead to nonlinear optimization problems, it will be clear how to formulate robust versions for these problems. This will be studied in Chapter 4. So consider a linear optimization problem of the form:

$$\begin{aligned} &\min\ c^\top x \\ &\text{s.t. } Ax \leq b \\ &\qquad x \geq 0 \end{aligned} \tag{3.5}$$

where $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, and $b \in \mathbb{R}^m$. This problem is parametrized by (c, A, b), all of which can be affected by uncertainty. In this case, one considers an uncertainty set $\mathbb{U} \subseteq \mathbb{R}^n \times \mathbb{R}^{m \times n} \times \mathbb{R}^m$ such that

$$(c, A, b) \in \mathbb{U}.$$

Ignoring the uncertainty by simply taking any $(c, A, b) \in \mathbb{U}$ leads to the so called nominal problem. However, we are interested in robust verisons of the problem. Before we outline the possible approaches, note that it is no loss of generality to assume that the objective function is not affected by the uncertainty. This is true because we can always write (3.5) equivalently as

$$\begin{aligned} &\min\ t \\ &\text{s.t. } Ax \leq b \\ &\qquad c^\top x \leq t \\ &\qquad x \geq 0 \\ &\qquad t \in \mathbb{R} \end{aligned}$$

by introducing an additional variable t.

Likewise we can always assume that the right-hand side is not affected by uncertainty. To see this, consider the LP

$$\begin{aligned} &\min\ c^\top x \\ &\text{s.t. } a_i^\top x \leq b_i \quad \forall i = 1, 2, \ldots, m \\ &\qquad x \geq 0 \end{aligned}$$

where the right hand side is uncertain. We can write this LP in the equivalent form

$$\begin{aligned} &\min\ c^\top x \\ &\text{s.t. } a_i^\top x - b_i x_{n+1} \leq 0 \quad \forall i = 1, 2, \ldots, m \\ &\qquad x_{n+1} = 1 \\ &\qquad x_1, \ldots, x_n \geq 0 \end{aligned}$$

with the additional variable $x_{n+1} \in \mathbb{R}$. Therefore, we can assume that neither the objective function nor the right hand side of the constraints are affected by uncertainty.

(Soyster, 1973) was the first to formulate a robust counterpart of an uncertain linear optimization problem. He proposed a linear optimization formulation of an uncertain problem such that the solution of the new formulation is feasible for all uncertain parameters. (Soyster, 1973) considers column-wise uncertainty which means that the columns A_j of the constraint matrix A belong to a convex set K_j. This is the most conservative robust optimization approach because in order to ensure feasibility, the worst case was considered. This results in loosing a lot of optimality of the nominal problem in terms of the objective function value (see (Ben-Tal and Nemirovski, 2000)). (El Ghaoui and Lebret, 1997) presented a less conservative robust model considering ellipsoidal uncertainty which can be solved using second order cone programming. (Ben-Tal and Nemirovski, 1999) also proposed a less conservative robust model considering ellipsoidal uncertainty. Generally for a large scale uncertain optimization problem, robustness becomes expensive in terms of computational complexity. The practical drawback of these less conservative models are that the robust counterparts are nonlinear. (Ben-Tal and Nemirovski, 1999) proved that many robust counterparts of linear programs with ellipsoidal uncertainty are polynomially solvable inspite of the fact that some robust counterparts are not linear programs.(Bertsimas and Sim, 2004) presented a controlled conservative approach where the level of conservatism of the robust model can be controlled or various robust solutions can be achieved on the basis of the preferred level of conservatism. The highest level of conservatism leads to the robust formulation of (Soyster, 1973). A detailed discussion is available in Section 3.3.

3.3 Robust Counterparts

We have already explained that we can always include the uncertain objective function parameters and right hand side uncertain parameters in the constraint matrix. We also consider that we have only interval uncertainty as discussed in Section 2.4. Let $\hat{a}_{ij}$ be the maximum deviation from the nominal value a_{ij}. In the constraint matrix A, the true value of each element a_{ij} is a symmetric and bounded random variable, represented by $\tilde{a}_{ij} \in [a_{ij} - \hat{a}_{ij}, a_{ij} + \hat{a}_{ij}]$. We define the random variable $\eta_{ij} \in [-1, 1]$ which follows an unknown but symmetric distribution such that $\eta_{ij} = (\tilde{a}_{ij} - a_{ij})/\hat{a}_{ij}$.

3.3.1 Robust counterpart of Soyster

(Soyster, 1973) considered the following model:

$$\begin{aligned} &\min\ c^\top x \\ &\text{s.t.}\ \sum_{j=1}^{n} A_j x_j \le b \quad \forall A_j \in K_j, j = 1, 2, \ldots, n \\ &\quad x \ge 0 \end{aligned} \tag{3.6}$$

where $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, and $A_j \in \mathbb{R}^m$ is considered to be the column j of the constraint matix A. The uncertainty model considered by Soyster is column wise uncertainty for the constraint matrix, i.e., $A_j \in K_j$ for some convex sets $K_j\,(j = 1, \ldots, n)$.

Define a matrix $\bar{A}$ whose entries are $\bar{a}_{ij} = \sup_{A_{j\in K_j}}(a_{ij})$. (Soyster, 1973) showed that (3.6) is equivalent to the following problem:

$$\begin{aligned} &\min\ c^\top x \\ &\text{s.t.}\ \sum_{j=1}^{n} \bar{A}_j x_j \le b \\ &\quad x \ge 0. \end{aligned} \tag{3.7}$$

Soyster considered LPs with nonnegativity constaints. Here we study a more general model and we consider the following nominal problem:

$$\begin{aligned} &\min\ c^\top x \\ &\text{s.t.}\ Ax \le b \\ &\quad l \le x \le u \end{aligned} \tag{3.8}$$

where $c, l, u \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Note that in this case the lower bound l is allowed to be strictly negative. Now we consider entrywise uncertainty which means that

$$\tilde{a}_{ij} \in [a_{ij} - \hat{a}_{ij}, a_{ij} + \hat{a}_{ij}] \text{ for all } i = 1, \ldots, m \text{ and } j = 1, \ldots, n,$$

where $\tilde{a}_{ij}$ follows an unknown but symmetric distribution in $[a_{ij} - \hat{a}_{ij}, a_{ij} + \hat{a}_{ij}]$ and $\hat{a}_{ij} \ge 0$ is the maximal deviation from the nominal value a_{ij}. We use the random variables $\eta_{ij} = (\tilde{a}_{ij} - a_{ij})/\hat{a}_{ij} \in [-1, 1]$, such that

$$\tilde{a}_{ij} = a_{ij} + \eta_{ij}\hat{a}_{ij} \text{ for all } i = 1, \ldots, m \text{ and } j = 1, \ldots, n$$

Then the constraint

$$\sum_{j=1}^{n} \tilde{a}_{ij} x_j \le b_i \text{ for all } \tilde{a}_{ij} \in [a_{ij} - \hat{a}_{ij}, a_{ij} + \hat{a}_{ij}]$$

becomes

$$\sum_{j=1}^{n} a_{ij} x_j + \sum_{j=1}^{n} \eta_{ij}\hat{a}_{ij} x_j \le b_i.$$

Since x_j can be positive or negative, we have to introduce auxiliary variables $y_j \geq 0$ and we can rewrite this constraint equivalently as

$$\begin{aligned}
&\sum_{j=1}^{n} a_{ij}x_j + \sum_{j=1}^{n} \hat{a}_{ij}y_j \leq b_i \quad \forall i = 1, 2, \ldots, m \\
&-y_j \leq x_j \leq y_j \quad \forall j = 1, \ldots, n \\
&y \geq 0
\end{aligned}$$

We can therefore formulate a roust version according to Soyster for problem (3.8) as follows:

$$\begin{aligned}
&\min\ c^\top x \\
&\text{s.t.}\ \sum_{j=1}^{n} a_{ij}x_j + \sum_{j=1}^{n} \hat{a}_{ij}y_j \leq b_i \quad \forall i = 1, 2, \ldots, m \\
&\quad -y_j \leq x_j \leq y_j \quad \forall j = 1, \ldots, n \\
&\quad l \leq x \leq u \\
&\quad y \geq 0
\end{aligned} \tag{3.9}$$

Let (x^*, y^*) be an optimal solution of problem (3.9). Note that we can always assume that $y_j^* = |x_j^*|$ for all j: Clearly, the constraint in (3.9) entails $|x_j^*| \leq y_j^*$ for all j. Assume that $|x_k^*| < y_k^*$ for some k. Then we can define

$$y_j^{**} = \begin{cases} |x_j^*| & \text{if} \quad j = k \\ y_j^* & \text{else.} \end{cases}$$

With this definition the point (x^*, y^{**}) is also an optimal solution of problem (3.9) and it fulfills $y_j^{**} = |x_j^*|$ for all j.

Now we can show that for every possible realization of $\tilde{a}_{ij}$ of the uncertain data, the optimal solution x^* is feasible for the original uncertain problem (3.6): indeed, we have

$$\begin{aligned}
\sum_{j=1}^{n} \tilde{a}_{ij}x_j^* &= \sum_{j=1}^{n} a_{ij}x_j^* + \sum_{j=1}^{n} \eta_{ij}\hat{a}_{ij}x_j^* \\
&\leq \sum_{j=1}^{n} a_{ij}x_j^* + \sum_{j=1}^{n} \hat{a}_{ij}|x_j^*| \leq b_i \quad \text{for all } i.
\end{aligned}$$

Note that if the lower bound l in problem (3.9) is 0 or strictly positive, then it is not necessary to introduce the auxiliary variable y_j. In this case problem (3.9) becomes

$$
\begin{aligned}
&\min\ c^\top x \\
&\text{s.t.}\ \sum_{j=1}^{n}(a_{ij}+\hat{a}_{ij})x_j \le b_i \quad \forall i=1,2,\dots,m \\
&\qquad 0 \le l \le x \le u
\end{aligned}
\tag{3.10}
$$

Soyster's robust optimization problem (3.9) is equivalent to solving the nominal problem for the worst case. As we have seen the robust solution according to Soyster is feasible for the original uncertain problem for any realization of the uncertain parameters within the uncertainty interval. This is why this approach is considered to be the most conservative robust formulation. This robustnss comes with a cost: the optimal objective function value is usually worse than the optimal value of the nominal problem. It is also very pessimistic to consider that all of the parameters are uncertain in the worst possible way. Therefore, we study some less conservative robust formulation in the next section.

3.3.2 Robust counterpart of Ben-Tal and Nemirovski

Consider the following nominal problem:

$$
\begin{aligned}
&\min\ c^\top x \\
&\text{s.t.}\ Ax \le b \\
&\qquad l \le x \le u
\end{aligned}
\tag{3.11}
$$

where $c, l, u \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, and $A \in \mathbb{R}^{m\times n}$. In order to address the conservatism of Soyster's model, (Ben-Tal and Nemirovski, 2000) consider a different setting. They assume again that the uncertainty concerns the entries $\tilde{a}_{ij}$ which can take values in the interval $[a_{ij}-\hat{a}_{ij}, a_{ij}+\hat{a}_{ij}]$.

For $i=1,\dots,m$, let $J_i \subseteq \{1,\dots,n\}$ be the set of indices of the parameters that are considered uncertain in row i of the constraint matrix A, i.e., a_{ij} is an uncertain parameter if and only if $j \in J_i$ (and consequently a_{ij} is not considered uncertain if $j \notin J_i$). (Ben-Tal and Nemirovski, 2000) introduce a parameter $\Omega_i > 0$ $(i=1,\dots,m)$ for each uncertain constraint and consider the following robust optimization problem:

$$
\begin{aligned}
&\min\ c^\top x \\
&\text{s.t.}\ \sum_{j=1}^{n} a_{ij}x_j + \sum_{j\in J_i}\hat{a}_{ij}y_{ij} + \Omega_i\sqrt{\sum_{j\in J_i}\hat{a}_{ij}^2 z_{ij}^2} \le b_i \quad \forall i=1,2,\dots,m \\
&\qquad -y_{ij} \le x_j - z_{ij} \le y_{ij} \quad \forall i,j \in J_i \\
&\qquad l \le x \le u \\
&\qquad y \ge 0
\end{aligned}
\tag{3.12}
$$

The authors show the following: if (x, y, z) is an optimal solution of (3.12), then x is feasible for the nominal problem, and for any realization of the uncertain parameters $\tilde{a}_{ij}$, the probability that x violates constraint i is at most $\exp(-\Omega^2/2)$.

So in a sense, the parameter Ω_i guides the probability that constraint i is violated. Since this parameter is chosen by the user, he can decide the level of conservatism.

The drawback of the approach by (Ben-Tal and Nemirovski, 2000) is that (3.12) is not a linear problem. It has a certain quadratic structure which is called second order cone problem (SOCP). Although nonlinear, these problems can be solved efficiently by interior point algorithms, see e.g. (Boyd and Vandenberghe, 2004), and efficient software implementations are available for this.

However, in this thesis we will restrict ourselves to robust models which are linear. In the next section, we describe a robust model introduced by (Bertsimas and Sim, 2004) which combines the benefits of being less conservative than Soyster's aproach, but still resulting in a linear optimization problem.

3.3.3 Robust counterpart of Bertsimas and Sim

In this section, we present the robust formulation of (Bertsimas and Sim, 2004) which is less conservative than that of Soyster.

Consider again the nominal problem as follows:

$$\begin{aligned} \min\ & c^\top x \\ \text{s.t. } & Ax \leq b \\ & l \leq x \leq u \end{aligned} \tag{3.13}$$

Similar to the approach of (Ben-Tal and Nemirovski, 2000), the approach of Bertsimas and Sim is based on the assumption that we know which of the problem parameters are uncertain and which are not. For $i = 1, \ldots, m$, let $J_i \subseteq \{1, \ldots, n\}$ be the set of indices of the parameters that are considered uncertain in row i of the constraint matrix A, i.e., a_{ij} is an uncertain parameter if and only if $j \in J_i$ (and consequently a_{ij} is not considered uncertain if $j \notin J_i$). Note that this setting includes both the case of all parameters being uncertain (if $J_i = \{1, \ldots, n\}$ for all i) and the case of absence of uncertainty (if $J_i = \emptyset$ for all i). We assume again that the uncertain parameters a_{ij} can take values in $[a_{ij} - \hat{a}_{ij}, a_{ij} + \hat{a}_{ij}]$.

Next, introduce parameters $\Gamma_i \in [0, |J_i|]$ (for $i = 1, \ldots, m$) and assume for the moment that Γ_i is integer (we will relax this assumption later). The philosophy behind this is that it seems unlikely that all of the parameters a_{ij} ($j \in J_i$) will change, and Γ_i reflects how many of them do. The robust formulation of (Bertsimas and Sim, 2004) guarantees that the optimal solution is feasible if at most Γ_i out of the $|J_i|$ uncertain parameters are allowed to change, however we do not know which of them do change.

We need to investigate how this model of uncertainty affects problem (3.13), and in particular the constraint

$$\sum_{j=1}^{n} \tilde{a}_{ij} x_j \leq b_i.$$

In Soyster's approach we saw that if all uncertain parameters change we arrive at the constraints

$$\begin{aligned} &\sum_{j=1}^{n} a_{ij} x_j + \sum_{j=1}^{n} \hat{a}_{ij} y_j \leq b_i \quad \forall i = 1, 2, \ldots, m \\ &-y_j \leq x_j \leq y_j \quad \forall j = 1, \ldots, n \\ &y \geq 0 \end{aligned}$$

If we now assume that only $\Gamma_i \leq |J_i|$ of the uncertain parameters change, then clearly the maximal possible left hand side value is

$$\sum_{j=1}^{n} a_{ij} x_j + \max_{S_i \subseteq J_i, |S_i| = \Gamma_i} \sum_{j \in S_i} \hat{a}_{ij} y_j. \tag{3.14}$$

Therefore the robust version of the problem (3.13) according to (Bertsimas and Sim, 2004) is

$$\begin{aligned} &\min c^\top x \\ &\text{s.t.} \ \sum_{j=1}^{n} a_{ij} x_j + \max_{S_i \subseteq J_i, |S_i| = \Gamma_i} \sum_{j \in S_i} \hat{a}_{ij} y_j \leq b_i \quad \forall i = 1, \ldots, m \\ &\quad -y_j \leq x_j \leq y_j \quad \forall j \in 1, \ldots, n \\ &\quad l \leq x \leq u \\ &\quad y \geq 0 \end{aligned} \tag{3.15}$$

Note that if $\Gamma_i = |J_i|$ for all i, then we recover Soyster's model, while $\Gamma_i = 0$ for all i is the case of no data uncertainty and we recover the nominal problem (3.13). In this problem (3.15), it is inconvinient that there appear the terms

$$\beta_i(y, \Gamma_i) := \max_{S_i \subseteq J_i, |S_i| = \Gamma_i} \sum_{j \in S_i} \hat{a}_{ij} y_j \quad \forall i = 1, \ldots, m.$$

Bertsimas and Sim showed that $\beta_i(y, \Gamma_i)$ can be replaced by a system of linear expressions.

Lemma 3.1. *Let $i \in \{1, \ldots, m\}$, let $y \in \mathbb{R}^n$ and let $\Gamma_i \in [0, |J_i|]$. Then $\beta_i(y, \Gamma_i)$ equals the optimal value of the following linear problem:*

$$\begin{aligned} \beta_i(y, \Gamma_i) = \min_{u,v} \; & \Gamma_i u_i + \sum_{j \in J_i} v_{ij} \\ & s.t. \;\; u_i + v_{ij} \geq \hat{a}_{ij} y_j \quad \forall j \in J_i \\ & v_{ij} \geq 0, \quad \forall j \in J_i \\ & u_i \in \mathbb{R}. \end{aligned} \tag{3.16}$$

Proof. First consider the linear problem which is dual to (3.16)

$$\begin{aligned} \max_{z} \; & \sum_{j \in J_i} \hat{a}_{ij} y_j z_{ij} \\ \text{s.t.} \; & \sum_{j \in J_i} z_{ij} = \Gamma_i \\ & 0 \leq z_{ij} \leq 1, \quad \forall j \in J_i. \end{aligned} \tag{3.17}$$

Note that since we consider y to be fixed, this is indeed an LP. Clearly, an optimal solution z_i^* of (3.17) is a binary vector with Γ_i of its components equal to 1 and $|J_i| - \Gamma_i$ components equal to 0. The nonzero components clearly fulfill $z_{ij}^* = 1$ if and only if $j \in S_i^*$, where $S_i^* \subseteq J_i$ is a set of cardinality $|S_i^*| = \Gamma_i$ which maximizes $\sum_{j \in S_i} \hat{a}_{ij} y_j$. In other words, the optimal value of (3.17) equals $\beta_i(y, \Gamma_i)$. Since problem (3.17) is feasible and bounded, strong LP duality yields that the optimal values of (3.17) and (3.16) are equal, which proves the lemma. □

Substituting this result into model (3.15) gives the following equivalent but more tractable formulation of the robust model (3.15) according to Bertsimas and Sim:

$$\begin{aligned} \min \; & c^\top x \\ \text{s.t.} \; & \sum_{j=1}^{n} a_{ij} x_j + \Gamma_i u_i + \sum_{j \in J_i} v_{ij} \leq b_i \quad \forall i = 1, \ldots, m \\ & u_i + v_{ij} \geq \hat{a}_{ij} y_j \quad \forall i = 1, \ldots, m \quad \forall j \in J_i \\ & -y_j \leq x_j \leq y_j \quad \forall j \in 1, \ldots, n \\ & l \leq x \leq u \\ & y \geq 0 \\ & v_{ij} \geq 0 \quad \forall i = 1, \ldots, m \quad \forall j \in J_i \\ & u_i \in \mathbb{R} \quad \forall i = 1, \ldots, m \end{aligned} \tag{3.18}$$

Note that this problem is a linear problem with $2n + m + mn$ variables, $m + \sum_{i=1}^{m} |J_i| + 4n$ constraints and $n + \sum_{i=1}^{m} |J_i|$ nonnegativity constraints.

(Bertsimas and Sim, 2004) also considered the case when $\Gamma_i \in [0, |J_i|]$ is non integer. The interpretation is that this provides robustness against the case that $\lfloor \Gamma_i \rfloor$ of the uncertain parameters change by their worst value $\hat{a}_{ij}$, while one more parameter changes by $(\Gamma_i - \lfloor \Gamma_i \rfloor)\hat{a}_{ij}$. In this setting it is easy to see that in analogy to (3.14) the maximum possible left hand side value in constraint i is:

$$\sum_{j=1}^{n} a_{ij}x_j + \max_{\{S_i \cup \{t_i\} | S_i \subseteq J_i, |S_i| = \lfloor \Gamma_i \rfloor, t_i \in J_i \setminus S_i\}} \left\{ \sum_{j \in J_i} \hat{a}_{ij} y_j + (\Gamma_i - \lfloor \Gamma_i \rfloor)\hat{a}_{it_i} y_{t_i} \right\} \leq b_i \quad (3.19)$$

Similar as in Lemma 3.1 Bertsimas and Sim showed that the maximum in (3.19) can be computed by solving the following linear problem:

$$\begin{aligned} \min_{u,v} \ & \Gamma_i u_i + \sum_{j \in J_i} v_{ij} \\ \text{s.t.} \ & u_i + v_{ij} \geq \hat{a}_{ij} y_j \quad \forall j \in J_i \\ & v_{ij} \geq 0 \quad \forall j \in J_i \\ & u_i \geq 0 \quad \forall i = 1, \ldots, m. \end{aligned} \quad (3.20)$$

Therefore, the robust version of problem (3.13) in this setting reads as follows:

$$\begin{aligned} \min \ & c^\top x \\ \text{s.t.} \ & \sum_{j=1}^{n} a_{ij}x_j + \Gamma_i u_i + \sum_{j \in J_i} v_{ij} \leq b_i \quad \forall i = 1, \ldots, m \\ & u_i + v_{ij} \geq \hat{a}_{ij} y_j \quad \forall i = 1, \ldots, m \quad \forall j \in J_i \\ & -y_j \leq x_j \leq y_j \quad \forall j \in 1, \ldots, n \\ & l \leq x \leq u \\ & y \geq 0 \\ & v_{ij} \geq 0 \quad \forall i = 1, \ldots, m \quad \forall j \in J_i \\ & u_i \in \mathbb{R} \quad \forall i = 1, \ldots, m \end{aligned} \quad (3.21)$$

Again this is a linear problem and hence easy to solve.

Chapter 4

Robust Allocation in Survey Statistics

In the sample allocation problem, it is well known that some of the parameters are uncertain in nature. (Díaz-García and Garay-Tápia, 2007) solved the optimum allocation problem considering the stratum specific variances as random variables. However a stochastic approach might not be suitable as we do not have much information about the distribution of the uncertain parameters. So in this chapter, we propose a robust allocation approach for solving the sample allocation problem.

We can always solve the sampling allocation problem by minimizing the total variance and transferring the cost function into the constraints when an upper bound on the total budget is given. The problem can also be solved by minimizing the total cost by transferring the variance function to the constraints when an upper bound on the total variance is available, for details see (Díaz-García and Garay-Tápia, 2007). If we have a fixed total budget C then we can write the uncertain multivariate sampling allocation problem from (2.10) as follows:

$$
\begin{aligned}
&\min \phi_k \qquad \forall k = 1, \ldots, K \\
&\text{s.t.} \ \sum_{h=1}^{H} \tilde{C}_h n_h \leq C \\
&\quad \sum_{h=1}^{H} \frac{\tilde{d}_{hk}}{n_h} \leq \phi_k \qquad \forall k = 1, \ldots, K \\
&\quad \sum_{h=1}^{H} n_h \leq \beta \\
&\quad m_h \leq \ n_h \leq M_h \quad \forall h = 1, \ldots, H \\
&\quad \phi_k \in \mathbb{R} \quad \forall k = 1, \ldots, K
\end{aligned}
\tag{4.1}
$$

In this section, we present robust formulations of this sampling allocation problem. We follow the approaches of (Soyster, 1973) and (Bertsimas and Sim, 2004) which were already discussed in Section 3.3.

4.1 Robust allocation according to Soyster

The approach of (Soyster, 1973) follows a worst case philosophy, so it is extremely conservative and guarantees feasibility of the optimal allocation under any realization of the uncertain parameters. Soyster mainly considered linear optimization problems, but his approach can easily be extended to our nonlinear problem (4.1).

Using Soyster's approach, we formulate a robust version of the uncertain multivariate sampling allocation problem (4.1) as follows:

First we define for all $k = 1, \ldots, K$ and for all $h = 1, \ldots, H$, the quantities:

$$\bar{d}_{hk} := \max\{\tilde{d}_{hk} \mid \tilde{d}_{hk} \in [d_{hk} - \hat{d}_{hk}, d_{hk} + \hat{d}_{hk}]\} = d_{hk} + \hat{d}_{hk},$$

and for all $h = 1, \ldots, H$, we define:

$$\bar{C}_h := \max\{\tilde{C}_h \mid \tilde{C}_h \in [C_h - \hat{C}_h, C_h + \hat{C}_h]\} = C_h + \hat{C}_h.$$

The robust formulation according to Soyster reads as follows:

$$\begin{aligned}
&\min \quad \phi_k \quad \forall k = 1, \ldots, K \\
&\text{Subject to:} \\
&\sum_{h=1}^{H} \bar{C}_h n_h \leq C \\
&\sum_{h=1}^{H} \frac{\bar{d}_{hk}}{n_h} \leq \phi_k \qquad \forall k = 1, \ldots, K \\
&\sum_{h=1}^{H} n_h \leq \beta \\
&m_h \leq \ n_h \leq M_h \quad \forall h = 1, \ldots, H \\
&\phi_k \in \mathbb{R} \quad \forall k = 1, \ldots, K
\end{aligned} \tag{4.2}$$

The robust formulation of Soyster admits the highest protection and hence is the most conservative in practice. The next statement shows that Soyster's approach guarantees feasibility even in the worst case.

Theorem 4.1. *Let* $(n_1^*, \ldots, n_H^*, \phi_1^*, \ldots, \phi_K^*)$ *be a feasible solution of the problem* (4.2). *Then* $(n_1^*, \ldots, n_H^*, \phi_1^*, \ldots, \phi_K^*)$ *is feasible for problem* (4.1) *under any realization of parameters*

$$\tilde{d}_{hk} \in [d_{hk} - \hat{d}_{hk}, d_{hk} + \hat{d}_{hk}] \quad \textit{and} \quad \tilde{C}_h \in [C_h - \hat{C}_h, C_h + \hat{C}_h].$$

In particular, any Pareto optimal solution of problem (4.2) *is feasible for problem* (4.1) *under any realization of the uncertain parameters.*

Proof. We have to show that $(n_1^*, \ldots, n_H^*, \phi_1^*, \ldots, \phi_K^*)$ is feasible for problem (4.2). Let $\tilde{d}_{hk} \in [d_{hk} - \hat{d}_{hk}, d_{hk} + \hat{d}_{hk}]$ and $\tilde{C}_h \in [C_h - \hat{C}_h, C_h + \hat{C}_h]$. Then we have

$$\sum_{h=1}^{H} \tilde{C}_h n_h^* \leq \sum_{h=1}^{H} \bar{C}_h n_h^* \leq C.$$

So the cost constraint in problem (4.1) is fulfilled. Due to the definition of $\bar{d}_{hk}$ and $n_h^* \geq 0$, we have that

$$\sum_{h=1}^{H} \frac{\tilde{d}_{hk}}{n_h^*} \leq \sum_{h=1}^{H} \frac{\bar{d}_{hk}}{n_h^*} \leq \phi_k^* \quad \forall k = 1, \ldots, K.$$

So the other nonlinear constraints of problem (4.1) are also fulfilled. The remaining constraints

$$\begin{aligned} &\sum_{h=1}^{H} n_h^* \leq \beta \\ &m_h \leq \ n_h^* \leq M_h \quad \forall h = 1, \ldots, H \\ &\phi_k^* \in \mathbb{R} \quad \forall k = 1, \ldots, K \end{aligned}$$

are trivialy fulfilled. Thus, the solution $(n_1^*, \ldots, n_H^*, \phi_1^*, \ldots, \phi_K^*)$ is feasible for (4.1) under any realization of uncertain parameters defined in the interval. □

Soyster's approach guarantees feasibility even in the worst case but in order to ensure feasibility, it loses a lot of optimality in terms of the quality of the objective function value. For this reason, we will not consider Soyster's approach any further in this thesis. Instead, we follow a less conservative aproach by (Bertsimas and Sim, 2004), who proposed a robust formulation which reduces the conservatism and gives a probabilistic bound on the constraint violation.

4.2 Robust allocation according to Bertsimas and Sim

In this section, we formulate robust models for the sampling allocation problem that are less conservative than that of Soyster. We follow closely an approach outlined by Bertsimas and Sim (2004), which we already introduced in Section 3.3.3.

4.2.1 Robust allocation if costs and variances both are uncertain (RobCV)

Let us return to problem (4.1). First introduce the abbreviation $\mathbb{H} := \{1, \ldots, H\}$. Following the philosophy outlined in Section 3.3.3, let $J_0 \subseteq \mathbb{H}$ be the set of uncertain cost parameters C_h in stratum h, and for $k = 1, \ldots, K$, let $J_k \subseteq \mathbb{H}$ be the set of uncertain variance parameters d_{hk} of variable k in stratum h.

We also introduce parameters $\Gamma_i \in [0, |J_i|]$ for $i = 0, \ldots, H$. Note that we do not assume that Γ_i are integers. Our goal is to be protected against all cases in which $\lfloor \Gamma_0 \rfloor$ of $\tilde{C}_h$ are allowed to change in the interval $[C_h + \hat{C}_h, C_h - \hat{C}_h]$ and one coefficient changes by $(\Gamma_0 - \lfloor \Gamma_0 \rfloor)\hat{C}_h$. Similarly for each $k = 1, \ldots, K$, we allow $\lfloor \Gamma_k \rfloor$ of the parameters $\tilde{d}_{hk}$ to change in the interval $[d_{hk} + \hat{d}_{hk}, d_{hk} - \hat{d}_{hk}]$ and one coefficient changes by $(\Gamma_k - \lfloor \Gamma_k \rfloor)\hat{d}_{hk}$.

Denoting the vector $n := (n_1, \ldots, n_H)$, we define

$$\beta_0(n, \Gamma_0) := \max_{\{\{i\} \cup S | S \subseteq \mathbb{H}, S = \lfloor \Gamma_0 \rfloor, i \in \mathbb{H} \setminus S\}} \left\{ \sum_{h \in S} \hat{C}_h n_h + (\Gamma_0 - \lfloor \Gamma_0 \rfloor)\hat{C}_i n_i \right\} \tag{4.3}$$

and for all $k = 1, \ldots, K$, we define

$$\beta_k(n, \Gamma_k) := \max_{\{\{i\} \cup S | S \subseteq \mathbb{H}, S = \lfloor \Gamma_k \rfloor, i \in \mathbb{H} \setminus S\}} \left\{ \sum_{h \in S} \frac{\hat{d}_{hk}}{n_h} + (\Gamma_k - \lfloor \Gamma_k \rfloor)\frac{\hat{d}_{ik}}{n_i} \right\}. \tag{4.4}$$

By the same arguments that we used in Section 3.3.3 to derive problem (3.21), we obtain the following robust version of problem (4.1).

$$
\begin{aligned}
\min \quad & \phi_k \quad \forall k = 1, \ldots, K \\
\text{s.t.} \quad & \sum_{h=1}^{H} C_h n_h + \beta_0(n, \Gamma_0) \leq C \\
& \sum_{h=1}^{H} \frac{d_{hk}}{n_h} + \beta_k(n, \Gamma_k) \leq \phi_k \quad \forall k = 1, \ldots, K \\
& \sum_{h=1}^{H} n_h \leq \beta \\
& m_h \leq \ n_h \leq M_h \quad \forall h \in \mathbb{H} \\
& \phi_k \in \mathbb{R} \quad \forall k = 1, \ldots, K
\end{aligned}
\tag{4.5}
$$

As we will see next, a Pareto optimal solution of problem (4.5) is feasible for the problem (4.1) with a very high probability even if more than Γ_0 of the cost parameters and/or more than Γ_k of the variance parameters are uncertain.

Probability bound on Constraint Violation

By construction, a Pareto optimal solution of problem (4.5) is feasible for the uncertain problem (4.1) if at most Γ_0 of the cost parameters $\tilde{C}_h$ and/or at most Γ_k of the variance parameters $\tilde{d}_{hk}$ are uncertain. Here, we prove that even if more parameters are uncertain, then the robust solution is feasible with a very high probability. We prove the following theorem to support our statement. We use the abbreviation $\phi^* := (\phi_1^*, \ldots, \phi_K^*)$.

Theorem 4.2. *Let (n^*, ϕ^*) be a Pareto optimal solution of model (4.5). For all $k = 0, 1, \ldots, K$, let S_k^* and i_k^* be the set and index, respectively, that achieve the maximum for $\beta_k(n^*, \Gamma_k)$. Then we have:*

(a) The probability that the cost constraint is violated can be bounded as follows:

$$Pr\left(\sum_{h\in\mathbb{H}} \tilde{C}_h n_h^* > C\right) \leq Pr\left(\sum_{h\in\mathbb{H}} \gamma_h \tilde{a}_h > \Gamma_0\right),$$

where for all $h \in \mathbb{H}$, we define

$$\tilde{a}_h = \frac{\tilde{C}_h - C_h}{C_h}, \qquad \gamma_{h0} = \begin{cases} 1, & if \quad h \in S_0^* \\ \frac{\hat{C}_h n_h^*}{\hat{C}_{e_0^*} n_{e_0^*}}, & if \quad h \in \mathbb{H}\backslash S_0^* \end{cases}$$

and

$$e_0^* = \text{argmin}\{\hat{C}_e n_e^* \mid e \in S_0^* \cup \{i_0^*\}\}.$$

(b) The probability that the k-th nonlinear constraint is violated can be bounded as follows

$$Pr\left(\sum_{h\in\mathbb{H}} \frac{\tilde{d}_{hk}}{n_h^*} > \phi_k^*\right) \leq Pr\left(\sum_{h\in\mathbb{H}} \gamma_{hk} \tilde{b}_h > \Gamma_k\right)$$

where for all $h \in \mathbb{H}$ and for all $k = 1, \ldots, K$, we define:

$$\tilde{b}_{hk} = \frac{\tilde{d}_{hk} - d_{hk}}{d_{hk}}, \qquad \gamma_{hk} = \begin{cases} 1, & if \quad h \in S_k^* \\ \frac{\hat{d}_{hk} n_{e_k^*}}{\hat{d}_{e_k^*} n_h^*}, & if \quad h \in \mathbb{H}\backslash S_k^* \end{cases}$$

and

$$e_k^* = \text{argmin}\left\{\frac{\hat{d}_{rk}}{n_e^*} \mid e \in S_k^* \cup \{i_k^*\}\right\}$$

Proof. (a) The probability that (n^*, ϕ^*) violates the cost constraint can be written as follows:

$$
\begin{aligned}
& Pr\left(\sum_{h\in\mathbb{H}} \tilde{C}_h n_h^* > C\right) \\
=& Pr\left(\sum_{h\in\mathbb{H}} C_h n_h^* + \sum_{h\in J_0} \tilde{a}_h \hat{C}_h n_h^* > C\right) \\
\leq& Pr\left(\sum_{h\in\mathbb{H}} C_h n_h^* + \sum_{h\in J_0} \tilde{a}_h \hat{C}_h n_h^* > \sum_{h\in\mathbb{H}} C_h n_h^* + \sum_{h\in S_0^*} \hat{C}_h n_h^* + (\Gamma_0 - \lfloor\Gamma_0\rfloor)\hat{C}_{i_0^*} n_{i_0^*}\right) \\
=& Pr\left(\sum_{h\in J_0} \tilde{a}_h \hat{C}_h n_h^* > \sum_{h\in S_0^*} \hat{C}_h n_h^* + (\Gamma_0 - \lfloor\Gamma_0\rfloor)\hat{C}_{i_0^*} n_{i_0^*}\right) \\
=& Pr\left(\sum_{h\in J_0\setminus S_0^*} \tilde{a}_h \hat{C}_h n_h^* > \sum_{h\in S_0^*} (1-\tilde{a}_h)\hat{C}_h n_h^* + (\Gamma_0 - \lfloor\Gamma_0\rfloor)\hat{C}_{i_0^*} n_{i_0^*}\right) \\
\leq& Pr\left(\sum_{h\in J_0\setminus S_0^*} \tilde{a}_h \hat{C}_h n_h^* > \hat{C}_{e_0^*} n_{e_0^*}^* \left(\sum_{h\in S_0^*} (1-\tilde{a}_h) + (\Gamma_0 - \lfloor\Gamma_0\rfloor)\right)\right) \\
=& Pr\left(\sum_{h\in J_0\setminus S_0^*} \tilde{a}_h \frac{\hat{C}_h n_h^*}{\hat{C}_{e_0^*} n_{e_0^*}^*} > \left(\lfloor\Gamma_0\rfloor - \sum_{h\in S_0^*} \tilde{a}_h + \Gamma_0 - \lfloor\Gamma_0\rfloor\right)\right) \\
=& Pr\left(\sum_{h\in S_0^*} \tilde{a}_h + \sum_{h\in J_0\setminus S_0^*} \tilde{a}_h \frac{\hat{C}_h n_h^*}{\hat{C}_{e_0^*} n_{e_0^*}^*} > \Gamma_0\right) \\
=& Pr\left(\sum_{h\in S_0^*} \gamma_{h0}\tilde{a}_h + \sum_{h\in J_0\setminus S_0^*} \gamma_{h0}\tilde{a}_h > \Gamma_0\right) \\
=& Pr\left(\sum_{h\in J_0} \gamma_{h0}\tilde{a}_h > \Gamma_0\right) \\
\leq& Pr\left(\sum_{h\in J_0} \gamma_{h0}\tilde{a}_h \geq \Gamma_0\right)
\end{aligned}
$$

Thus,

$$
Pr\left(\sum_{h\in\mathbb{H}} \tilde{C}_h n_h^* > C\right) \leq Pr\left(\sum_{h\in J_0} \gamma_{h0}\tilde{a}_h \geq \Gamma_0\right)
$$

This proves part (a).

Similarly, part (b) can also be proved. For each $k = 1, \ldots, K$, we have;

$$
\begin{aligned}
& Pr\left(\sum_{h\in\mathbb{H}} \frac{\tilde{d}_{hk}}{n_h^*} > \phi_k^*\right) \\
=& Pr\left(\sum_{h\in\mathbb{H}} \frac{d_{hk}}{n_h^*} + \sum_{h\in J_k} \tilde{b}_{hk}\frac{\hat{d}_{hk}}{n_h^*} > \phi_k^*\right) \\
\leq& Pr\left(\sum_{h\in\mathbb{H}} \frac{d_{hk}}{n_h^*} + \sum_{h\in J_k} \tilde{b}_{hk}\frac{\hat{d}_{hk}}{n_h^*} > \sum_{h\in\mathbb{H}} \frac{d_{hk}}{n_h^*} + \sum_{h\in S_k^*} \frac{\hat{d}_{hk}}{n_h^*} + (\Gamma_k - \lfloor\Gamma_k\rfloor)\frac{\hat{d}_{i_k^* k}}{n_{i_k}^*}\right) \\
=& Pr\left(\sum_{h\in J_k} \tilde{b}_{hk}\frac{\hat{d}_{hk}}{n_h^*} > \sum_{h\in S_k^*} \frac{\hat{d}_{hk}}{n_h^*} + (\Gamma_k - \lfloor\Gamma_k\rfloor)\frac{\hat{d}_{i_k^* k}}{n_{i_k}^*}\right) \\
=& Pr\left(\sum_{h\in J_k\setminus S_k^*} \tilde{b}_{hk}\frac{\hat{d}_{hk}}{n_h^*} > \sum_{h\in S_k^*} (1-\tilde{b}_{hk})\frac{\hat{d}_{hk}}{n_h^*} + (\Gamma_k - \lfloor\Gamma_k\rfloor)\frac{\hat{d}_{i_k^* k}}{n_{i_k}^*}\right) \\
\leq& Pr\left(\sum_{h\in J_k\setminus S_k^*} \tilde{b}_{hk}\frac{\hat{d}_{hk}}{n_h^*} > \frac{\hat{d}_{e_k^*}}{n_{e_k^*}^*}\left(\sum_{h\in S_k^*} (1-\tilde{b}_{hk}) + (\Gamma_k - \lfloor\Gamma_k\rfloor)\right)\right) \\
=& Pr\left(\sum_{h\in J_k\setminus S_k^*} \tilde{b}_{hk}\frac{\hat{d}_{hk}/n_h^*}{\hat{d}_{e_k^* k}/n_{e_k^*}^*} > \left(\lfloor\Gamma_k\rfloor - \sum_{h\in S_k^*} \tilde{b}_{hk} + \Gamma_k - \lfloor\Gamma_k\rfloor\right)\right) \\
=& Pr\left(\sum_{h\in S_k^*} \tilde{b}_{hk} + \sum_{h\in J_k\setminus S_k^*} \tilde{b}_{hk}\frac{\hat{d}_{hk}/n_h^*}{\hat{d}_{e_k^* k}/n_{e_k^*}^*} > \Gamma_k\right) \\
=& Pr\left(\sum_{h\in S_k^*} \gamma_{hk}\tilde{b}_{hk} + \sum_{h\in J_k\setminus S_k^*} \gamma_{hk}\tilde{b}_{hk} > \Gamma_k\right) \\
=& Pr\left(\sum_{h\in J_k} \gamma_{hk}\tilde{b}_{hk} > \Gamma_k\right) \\
\leq& Pr\left(\sum_{h\in J_k} \gamma_{hk}\tilde{b}_{hk} \geq \Gamma_k\right)
\end{aligned}
$$

Thus, we have that for each $k = 1, \ldots, K$

$$
Pr\left(\sum_{h\in\mathbb{H}} \frac{\tilde{d}_{hk}}{n_h^*} > \phi_k^*\right) \leq Pr\left(\sum_{h\in J_k} \gamma_{hk}\tilde{b}_{hk} \geq \Gamma_k\right).
$$

□

Theorem (4.2) provides upper bounds on the probability that (n^*, ϕ^*) violates the constraints. However, from the computational point of view these upper bounds are difficult to compute. The upper bound depends on the robust solution (n^*, ϕ^*) so we have to solve the problem first to know how good the solution is. However, it would be much better to have an upper bound that is independent of the robust solution. Next we present upper bounds on the probability bounds which are easier to compute and are independent of the robust solution. This result closely follows Theorem 2 of (Bertsimas and Sim, 2004).

Theorem 4.3. *(a) If for all $h \in J_0$, the random variables $\tilde{a}_h = \frac{\tilde{C}_h - C_h}{C_h}$ are independent and symmetrically distributed in $[-1, 1]$, then*

$$Pr\left(\sum_{h \in J_0} \gamma_{h0} \tilde{a}_h > \Gamma_0\right) \leq \exp\left(\frac{-\Gamma_0^2}{2|J_0|}\right).$$

(b) If for all $h \in J_k$ and for all $k \in \{1, \ldots, K\}$, the random variables $\tilde{b}_{hk} = \frac{\tilde{d}_{hk} - d_{hk}}{d_{hk}}$ are independent and symmetrically distributed in $[-1, 1]$, then

$$Pr\left(\sum_{h \in J_k} \gamma_{hk} \tilde{b}_{hk} > \Gamma_k\right) \leq \exp\left(\frac{-\Gamma_k^2}{2|J_k|}\right).$$

Proof. (a) For any $t > 0$, we have using Tschebyshev inequality

$$\begin{aligned}
Pr\left(\sum_{h \in J_0} \gamma_{h0} \tilde{a}_h \geq \Gamma_0\right) &\leq \exp(-t\Gamma_0) \mathbb{E}\left(\exp\left(t \sum_{h \in J_0} \gamma_{h0} \tilde{a}_h\right)\right) \\
&= \exp(-t\Gamma_0) \prod_{h \in J_0} \mathbb{E}\left(\exp\left(t\gamma_{h0} \tilde{a}_h\right)\right) \qquad (4.6) \\
&= \exp(-t\Gamma_0) \prod_{h \in J_0} 2 \int_0^1 \sum_{k=0}^{\infty} \frac{(t\gamma_{h0} \tilde{a}_h)^{2k}}{(2k)!} \mathrm{d}F_{\tilde{a}_h}(a_h) \qquad (4.7) \\
&\leq \exp(-t\Gamma_0) \prod_{h \in J_0} \sum_{k=0}^{\infty} \frac{(t\gamma_{h0})^{2k}}{(2k)!} \\
&\leq \exp(-t\Gamma_0) \prod_{h \in J_0} \exp\left(\frac{t^2 \gamma_{h0}^2}{2}\right) \\
&= \exp(-t\Gamma_0) \exp\left(\frac{t^2}{2} \sum_{h \in J_0} \gamma_{h0}^2\right)
\end{aligned}$$

Transformations (4.6) and (4.7) come from the independence and symmetry of the distribution of $\tilde{a}_h \in [-1, 1]$.

Now choose $t = \Gamma_0 / |J_0|$. Then we get

$$
\begin{aligned}
Pr\left(\sum_{h\in J_0}\gamma_{h0}\tilde{a}_h \geq \Gamma_0\right) &\leq \exp\left(-\frac{\Gamma_0^2}{|J_0|}\right)\exp\left(\frac{\Gamma_0^2\sum_{h\in J_0}\gamma_{h0}^2}{2|J_0|^2}\right)\\
&\leq \exp\left(-\frac{\Gamma_0^2}{|J_0|}\right)\exp\left(\frac{\Gamma_0^2}{2|J_0|}\right)\\
&= \exp\left(\frac{\Gamma_0^2}{2|J_0|}-\frac{\Gamma_0^2}{|J_0|}\right)\\
&= \exp\left(\frac{-\Gamma_0^2}{2|J_0|}\right)
\end{aligned}
$$

This proves part (a). We can similarly prove part (b) also. □

It is very difficult to solve the problem (4.5) in its current form. In order to reformulate it and get rid of (4.3) and (4.4) in the constraints, we proceed as in Section 3.3.3

Lemma 4.4. *Given a vector n^*, then $\beta_0(n^*,\Gamma_0)$ has the same optimal value as*

$$
\begin{aligned}
\beta_0(n^*,\Gamma_0) = \min\ & r\Gamma_0 + \sum_{h\in J_0} q_h \\
s.t.\ & r + q_h \geq \hat{C}_h n_h^* \quad \forall h \in J_0 \\
& r \geq 0 \\
& q_h \geq 0 \qquad \forall h \in J_0
\end{aligned}
\tag{4.8}
$$

Proof. The dual of problem (4.8) can be written as follows:

$$
\begin{aligned}
\max\ & \sum_{h\in J_0}\hat{C}_h n_h^* z_h \\
\text{s.t.}\ & \sum_{h\in J_0} z_h \leq \Gamma_0 \\
& 0 \leq z_h \leq 1 \quad \forall h \in J_0
\end{aligned}
\tag{4.9}
$$

Since problem (4.8) is feasible and bounded, its dual problem (4.9) is also feasible and bounded and hence we have strong duality.

An optimal solution z^* of (4.9) has the property that $\lfloor\Gamma_0\rfloor$ of its coordinates equal 1, one of its coordinates equals $(\Gamma_0 - \lfloor\Gamma_0\rfloor)$ and the rest of the coordinates equal zero. Define a set $S_0^* := \{h \in J_0 | z_h^* = 1\}$. Clearly S_0^* is a solution of the maximization problem (4.3) defining $\beta_0(n^*,\Gamma_0)$. This proves the lemma. □

Lemma 4.5. *Let $k \in \{1,\ldots,K\}$ and n^* be a given vector, then $\beta_k(n^*,\Gamma_k)$ has the same optimal value as*

$$
\begin{aligned}
\beta_k(n^*, \Gamma_k) = \min\ & l_k\Gamma_k + \sum_{h\in J_k} p_{hk} \\
s.t.\ & l_k + p_{hk} \geq \frac{\hat{d}_{hk}}{n_h^*} \quad \forall h \in J_k \\
& l_k \geq 0 \\
& p_{hk} \geq 0 \qquad \forall h \in J_k
\end{aligned} \tag{4.10}
$$

Proof. Note that the problem (4.10) is linear as we have a fixed n^*. Now, using the dual of (4.10), we can prove this lemma similarly as Lemma 4.4. □

Thus, substituting the values of $\beta_0(n^*, \Gamma_0)$ and $\beta_k(n^*, \Gamma_k)$ from (4.8) and (4.10) respectively, in problem (4.5), we get the following:

$$
\begin{aligned}
\min\quad & \phi_k \qquad \forall k \\
\text{s.t.}\ & \sum_{h=1}^{H} C_h n_h + r\Gamma_0 + \sum_{h\in J_0} q_h \leq C \\
& r + q_h \geq \hat{C}_h n_h \quad \forall h \in J_0 \\
& \sum_{h=1}^{H} \frac{d_{hk}}{n_h} + l_k\Gamma_k + \sum_{h\in J_k} p_{hk} \leq \phi_k \qquad \forall k = 1, \ldots, K \\
& l_k + p_{hk} \geq \frac{\hat{d}_{hk}}{n_h} \quad \forall h \in J_k \text{ and } \forall k = 1, \ldots, K \\
& \sum_{h=1}^{H} n_h \leq \beta \\
& m_h \leq\ n_h \leq M_h \quad \forall h = 1, \ldots, H \\
& \phi_k \in \mathbb{R} \quad \forall k = 1, \ldots, K \\
& r, q \geq 0 \\
& l_k, p_{hk} \geq 0 \qquad \forall h \in J_k \text{ and } \forall k = 1, \ldots, K
\end{aligned} \tag{4.11}
$$

Hence, problem (4.11) is equivalent to problem (4.5). Note that we get the robust formulation of a univariate sampling allocation problem by putting $k = 1$ in formulation (4.11).

4.2.2 Robust allocation if only costs are uncertain (RobC)

In this robust formulation of the sampling allocation problem (4.1), we consider uncertainty only in the cost parameters. We refer to this robust model as RobC. The following robust formulation is presented:

$$
\begin{aligned}
\min \quad & \phi_k \quad \forall k = 1, \ldots, K \\
\text{s.t.} \quad & \sum_{h=1}^{H} C_h n_h + \beta_0(n, \Gamma_0) \leq C \\
& \sum_{h=1}^{H} \frac{d_{hk}}{n_h} \leq \phi_k \quad \forall k = 1, \ldots, K \\
& \sum_{h=1}^{H} n_h \leq \beta \\
& m_h \leq \ n_h \leq M_h \quad \forall h = 1, \ldots, H \\
& \phi_k \in \mathbb{R} \quad \forall k = 1, \ldots, K
\end{aligned}
\tag{4.12}
$$

where $\beta_0(n, \Gamma_0)$ is defined in (4.3).

The results of Theorem 4.2, Theorem 4.3 and Lemma 4.4 also hold for problem (4.12). In a similar way, we can get rid of $\beta_0(n, \Gamma_0)$ in the constraints of (4.12). An equivalent version of problem (4.12) can be written as follows:

$$
\begin{aligned}
\min \quad & \phi_k \quad \forall k = 1, \ldots, K \\
\text{s.t.} \quad & \sum_{h=1}^{H} C_h n_h + r\Gamma_0 + \sum_{h \in J_0} q_h \leq C \\
& r + q_h \geq \hat{C}_h n_h \quad \forall h \in J_0 \\
& \sum_{h=1}^{H} \frac{d_{hk}}{n_h} \leq \phi_k \quad \forall k = 1, \ldots, K \\
& \sum_{h=1}^{H} n_h \leq \beta \\
& m_h \leq \ n_h \leq M_h \quad \forall h = 1, 2, \ldots, H \\
& r \geq 0 \\
& q_h \geq 0 \qquad \forall h \in J_0
\end{aligned}
\tag{4.13}
$$

where r and q_h are the optimization variables introduced in the robustification process.

4.2.3 Robust allocation if only variances are uncertain (RobV)

In this section, we formulate the univariate sampling allocation problem (4.1) considering that we have uncertain stratum specific variances but not uncertain costs. This robust formulation is referred to as RobV and can be stated as follows:

$$
\begin{aligned}
\min \quad & \phi_k \quad \forall k = 1, \ldots, K \\
\text{s.t.} \quad & \sum_{h=1}^{H} C_h n_h \leq C \\
& \sum_{h=1}^{H} \frac{d_{hk}}{n_h} + \beta_k(n, \Gamma_k) \leq \phi_k \quad \forall k = 1, \ldots, K \\
& \sum_{h=1}^{H} n_h \leq \beta \\
& m_h \leq \ n_h \leq M_h \quad \forall h \in \mathbb{H} \\
& \phi_k \in \mathbb{R} \quad \forall k = 1, \ldots, K
\end{aligned}
\tag{4.14}
$$

where $\beta_k(n, \Gamma_k)$ is as defined in (4.4).

Here again, the results of Theorem 4.2, Theorem 4.3 and Lemma 4.4 hold for problem (4.14). We can get rid of $\beta_k(n, \Gamma_k)$ in the constraints of (4.14) in a similar way. The problem (4.14) can be written equivalently as follows:

$$
\begin{aligned}
\min \quad & \phi_k \qquad \forall k = 1, \ldots, K \\
\text{s.t.} \quad & \sum_{h=1}^{H} C_h n_h \leq C \\
& \sum_{h=1}^{H} \frac{d_{hk}}{n_h} + l_k \Gamma_k + \sum_{h \in J_k} p_{hk} \leq \phi_k \qquad \forall k = 1, \ldots, K \\
& l_k + p_{hk} \geq \frac{\hat{d}_{hk}}{n_h} \quad \forall h \in J_k \text{ and } \forall k = 1, \ldots, K \\
& \sum_{h=1}^{H} n_h \leq \beta \\
& m_h \leq \ n_h \leq M_h \quad \forall h = 1, \ldots, H \\
& \phi_k \in \mathbb{R} \quad \forall k = 1, \ldots, K \\
& l_k, p_{hk} \geq 0 \qquad \forall h \in J_k \text{ and } \forall k = 1, \ldots, K
\end{aligned}
\tag{4.15}
$$

Here l_k and p_{hk} are the optimization variables introduced during robustification process.

Chapter 5

Analysis with Simulated Data

In this chapter we generate some large scale simulated data of survey statistical problems. Simulation can enable us to work with diversely distributed variables of a population. Sometimes it is difficult to gather exact information about the population such as the distribution of variables in subgroups of the population. In this simulation study we generate such data using R software. We generate a population of fixed size with variables having different distributions within the total population and also within the subgroups of the population. We use this simulated data to calculate robust allocations from our robust formulations. As we have already discussed in Chapter 4, Bertsimas and Sim's approach is less conservative than Soyster's apporach. We formulate three different robust formulations of the sampling allocation problem (SAP) using Bertsimas and Sim's approach and compare the results. We perform various experiments on the robust allocations obtained for the simulated data. These experiments are helpful in explaining the benefits of using robust formulations. In these experiments we check if the uncertain parameters of the optimization problems can make the robust solutions infeasible.

5.1 Simulated Data Generation

We consider both the univariate and the multivariate case of stratified sampling. The nominal problem of the multivariate SAP can be written as follows:

$$\begin{aligned} \min \ & \sum_{h=1}^{H} \frac{d_{hk}}{n_h} \qquad \forall k = 1, \ldots, K \\ \text{s.t.} \ & \sum_{h=1}^{H} C_h n_h \leq C \\ & \sum_{h=1}^{H} n_h \leq \beta \\ & m_h \leq \ n_h \leq M_h \quad \forall h = 1, \ldots, H \end{aligned} \tag{5.1}$$

As before, N_h denotes the size of stratum h and $N = \sum_{h=1}^{H} N_h$ denotes the population size. We abbreviate $d_{hk} := N_h^2 S_{hk}^2 / N^2$. Here, S_{hk}^2 is the variance of variable k in stratum h. The optimization variable defining the sample size in stratum h is denoted by n_h and C_h is the unit cost of selecting a sample in stratum h. The upper bound on the total sample size is denoted by β and C is the upper bound on the total cost. In addition we have lower bounds m_h and upper bounds M_h on the optimization variables n_h.

If we consider only one variable of the population then the problem (5.1) can be reduced to the following:

$$\begin{aligned} \min \quad & \sum_{h=1}^{H} \frac{d_h}{n_h} \\ \text{s. t.} \quad & \sum_{h=1}^{H} n_h \leq \beta \\ & \sum_{h=1}^{H} c_h n_h \leq C \\ & m_h \leq \ n_h \leq M_h \quad \forall h = 1, 2, ..., H \end{aligned} \tag{5.2}$$

In this simulation study, we generate a population of size $N = 10,000$ with $H = 5$ strata of sizes $N_1 = 1000$, $N_2 = 2000$, $N_3 = 1000$, $N_4 = 2000$ and $N_5 = 4000$. The unit cost of selecting a sample in each stratum is taken $(C_1, \ldots, C_5) = (31, 32, 33, 34, 35)$ and the deviation of the cost in case of uncertainty is taken to be $\hat{C}_h = 30$ for all $h = 1, 2 \ldots, 5$. The upper bound on the total cost is considered to be $C = 2000$. We allow a maximal sample size of 1% of the total population, i.e. $\beta = 0.01 \sum_{h=1}^{H} N_h = 100$. We take $m_h = 2$ and $M_h = N_h$ (to avoid over allocation) as lower and upper bounds on the optimization variable n_h. The data can be summarized in following table:

Table 5.1: Assumed data for the sample allocation

	N_h	Cost (C_h)	$\hat{C}_h$
Stratum 1	1000	31	30
Stratum 2	2000	32	30
Stratum 3	1000	33	30
Stratum 4	2000	34	30
Stratum 5	4000	35	30

We generate this population using simulation for three different characteristics. In stratum 1 and 2 the variables are Normally distributed and generated using R software as follows:

```
Sigma.h1 <- matrix(c(1,.9,-.2,.9,1,0,-.2,0,1),3,3)
is.positive.definite(Sigma.h1)
Sigma.h2 <- matrix(c(1,-.9,+.2,-.9,1,0,+.2,0,1),3,3)
is.positive.definite(Sigma.h2)
```

```
Pop.h1 <- rmvnorm(Nh[1],mean=rep(10,3),sigma = Sigma.h1)
Pop.h2 <- rmvnorm(Nh[2],mean=rep(10,3),sigma = Sigma.h2)
```

*Sigma.h*1 and *Sigma.h*2 are the 3×3 covariance matrices of the Normal distributions. The mean of the Normal distribution is considered to be $(10, 10, 10)$.

In stratum 3 and 4 the variables are χ^2 distributed and generated as follows:

```
Pop.h3 <- rmvnorm(Nh[3],mean=rep(10,3),sigma = Sigma.h1)^2
Pop.h4 <- rmvnorm(Nh[4],mean=rep(10,3),sigma = Sigma.h2)^2
```

For a more diverse population we consider that in stratum 5, two variables are Normally distributed and one is χ^2 distributed.

```
Pop.h5 <- rmvnorm(Nh[5],mean=rep(10,3),sigma = Sigma.h1)
Pop.h5[,3] <- Pop.h5[,3]^2
cor(Pop.h5)

POP <- rbind(cbind(Pop.h1,1),cbind(Pop.h2,2),cbind(Pop.h3,3),
cbind(Pop.h4,4),cbind(Pop.h5,5))
```

As discussed in the previous chapter we can always solve the multivariate sampling allocation problem as single objective optimization problem by minimizing the cost function and transferring the variance functions into the constraints. In next sections, many experiments are carried out to investigate the stability and robustness of the allocations. Note that we aim here to present the robust allocations and to test how the robust allocations behave in different situations. We start by computing robust allocations considering the first variable generated in the simulation which is also defined in Table 5.2. The deviation $\hat{d}_h$ in the stratum specific variance has been taken to be 10% of d_h.

Table 5.2: Simulated data for the sample allocation

	S^2_{h1}	d_{h1}	$\hat{d}_{h1}$	Distribution
Stratum 1	0.9761	0.0098	0.0010	Normal
Stratum 2	1.3414	0.0537	0.0054	Normal
Stratum 3	286.7397	2.8674	0.2867	χ^2
Stratum 4	494.4716	19.7789	1.9779	χ^2
Stratum 5	0.8799	0.1408	0.0141	Normal

5.2 Robust Formulations of SAP

5.2.1 Robust Formulation with Uncertain Costs (RobC)

In this setting the robust formulation of the univariate sampling allocation problem with uncertain cost parameters is formulated. We consider the situation where only the cost vector is affected by uncertainty. We refer to the robustification of this problem as RobC.

$$
\begin{aligned}
\min \quad & \sum_{h=1}^{H} \frac{d_h}{n_h} \\
\text{subject to:} \\
& \sum_{h=1}^{H} C_h n_h + r\Gamma_0 + \sum_{h=1}^{H} q_h \leq C \\
& r + q_h \geq \hat{C}_h n_h \quad \forall h = 1, 2, \ldots, H \\
& \sum_{h=1}^{H} n_h \leq \beta \\
& m_h \leq \ n_h \leq M_h \quad \forall h = 1, 2, \ldots, H \\
& r \geq 0 \\
& q_h \geq 0 \qquad \forall h = 1 \ldots, H
\end{aligned}
\tag{5.3}
$$

where r and q_h are the optimization variables introduced in the robustification process. We solve this problem for different values of Γ_0 using the R package nloptr. The results are given in Table 5.3.

Table 5.3: Results of RobC for different values of Γ_0.

Gammas	$\Gamma_0 = 0$	$\Gamma_0 = 1$	$\Gamma_0 = 2$	$\Gamma_0 = 3$	$\Gamma_0 = 4$	$\Gamma_0 = 5$
Total Variance ($\sum_{h=1}^{H} \frac{d_h}{n_h}$)	0.99187	1.24533	1.43268	1.47989	1.52927	1.58702
Total Cost ($\sum_{h=1}^{H} C_h n_h$)	1646.719	1315.226	1155.962	1118.947	1087.109	1053.719
n_1	2	2	2	2	2	2
n_2	2	2	2	2	2	2
n_3	14.25	9.84	7.80	7.57	7.50	7.38
n_4	27.73	22.82	20.32	19.78	18.92	18.04
n_5	3.06	2.52	2.31	2.00	2.00	2.04
$\sum_{h=1}^{5} n_h$	49.05	39.19	34.45	33.36	32.42	31.44

Here we have monotonically increasing total variance with increasing level of uncertainty. This increment in the variance can be considered as the cost of robustness. The more robust we want to be the more optimality we loose. However, the total cost is decreasing as we increase the level of uncertainty. The reason is that the optimal solution of the robust formulation gives smaller stratum specific sample sizes due to the fact that uncertainty reduces the sample sizes in the strata with high uncertainty which results in reduced total costs. If we see the cost constraint in (5.3), we have an upper bound on it of $C = 2000$. This gap between the total cost and the total budget is for the protection of the cost constraint in case of uncertainty.

5.2.2 Robust Formulation with Uncertain Variance (RobV)

Next, we formulate the univariate sampling allocation problem considering that we have uncertain stratum specific variances but no uncertainty in the cost. This robust formulation is referred to as RobV and can be stated as follows:

$$
\begin{aligned}
\min \quad & \sum_{h=1}^{H} \frac{d_h}{n_h} + l\Gamma_1 + \sum_{h=1}^{H} p_h \\
& \text{subject to:} \\
& \sum_{h=1}^{H} C_h n_h \leq C \\
& l + p_h \geq \frac{\hat{d}_h}{n_h} \quad \forall h = 1, 2, \ldots, H \\
& \sum_{h=1}^{H} n_h \leq \beta \\
& m_h \leq \ n_h \leq M_h \quad \forall h = 1, 2, \ldots, H \\
& l \geq 0 \\
& p_h \geq 0 \qquad \forall h = 1, 2, \ldots, H
\end{aligned}
\tag{5.4}
$$

Here l and p_h are the optimization variables introduced during robustification process. We solve this problem using the R package nloptr and the results obtained are as follows:

Table 5.4: Results for RobV with different values of Γ_1.

Gammas	$\Gamma_1 = 0$	$\Gamma_1 = 1$	$\Gamma_1 = 2$	$\Gamma_1 = 3$	$\Gamma_1 = 4$	$\Gamma_1 = 5$
Total Variance	0.7973683	0.8493959	0.8693515	0.8739389	0.8766170	0.8771051
Total Cost	2000	2000	2000	2000	2000	2000
n_1	2	2	2	2	2	2
n_2	2.02	2.00	2.00	2.00	2.00	2.02
n_3	14.57	14.11	14.62	14.58	14.57	14.57
n_4	37.71	38.29	37.83	37.73	37.71	37.71
n_5	3.13	3.03	2.99	3.14	3.13	3.13
$\sum_{h=1}^{5} n_h$	59.45	59.44	59.46	59.45	59.45	59.45

The total variance of the estimator is increasing as we increase the level of uncertainty in the variance. The total variance in RobV is smaller than the total variance in RobC. The reason is that in RobC we have uncertain costs but the total budget is fixed and the total cost can not exceed the total budget and hence the optimal solution of RobC has smaller sample sizes, whereas in RobV we have bigger sample sizes than in RobC and this results in a bigger total variance in RobV. However we can conclude from the results of RobV that if we increase the total budget C in RobC then in the optimal solution of RobC the sample sizes will increase

and consequently the total variance would be reduced. We can also notice in Table 5.4 that the total cost seems to be not affected by the change of Γ_1. The cost depends directly on the sample sizes for each stratum whereas the total variance also depends on the stratum specific variances. The sum of the allocated sample sizes is constant for all uncertainty levels however the sample sizes for each stratum change which is why the total variance is changing. When dealing with the uncertain stratum specific variances we prefer a solution that concentrates more on reducing the total variance while keeping the total budget. The results show that the formulation RobV fufills this aim.

5.2.3 Robust Formulation with Uncertain Cost and Variance (RobCV)

In real life survey statistical problems we might face a situation when there is uncertainty in both the variance and the cost. The solutions of RobC and RobV are not enough for this kind of situation. So we assume that $\tilde{C}_h$ and $\tilde{d}_h$ are allowed to change in the intervals $[C_h + \hat{C}_h, C_h - \hat{C}_h]$ and $[d_h + \hat{d}_h, d_h - \hat{d}_h]$ respectively. In this setting the robust formulation of the univariate sampling allocation problem with uncertain stratum specific variances and uncertain costs is referred to as RobCV and can be formulated as follows:

$$
\begin{aligned}
\min \quad & \sum_{h=1}^{H} \frac{d_h}{n_h} + l\Gamma_1 + \sum_{h=1}^{H} p_h \\
\text{subject to:} \\
& \sum_{h=1}^{H} C_h n_h + r\Gamma_0 + \sum_{h=1}^{H} q_h \leq C \\
& r + q_h \geq \hat{C}_h n_h \quad \forall h = 1, 2, \ldots, H \\
& l + p_h \geq \frac{\hat{d}_h}{n_h} \quad \forall h = 1, 2, \ldots, H \\
& \sum_{h=1}^{H} n_h \leq \beta \\
& m_h \leq \ n_h \leq M_h \quad \forall h = 1, 2, \ldots, H \\
& l, r \geq 0 \\
& p_h, \ \ q_h \geq 0 \qquad \forall h = 1, \ldots, H
\end{aligned}
\tag{5.5}
$$

Here, Γ_0 represents the number of uncertain cost parameters and Γ_1 represents the number of uncertain stratum specific variance parameters. It has already been established that in Bertsimas and Sim's aproach of robust formulation, if both $\Gamma_0 = 0$ and $\Gamma_1 = 0$ then we get back to the original nominal problem (3.4). Therefore, problem (5.3) and problem (5.4) are special cases of problem (5.5) when Γ_1 or Γ_0 are 0, respectively. However in practice this might not be completely true. If we have $\Gamma_0 = 0$ in problem (5.5) then our input is that there is no uncertainty in the costs. However we still have some other optimization variables such

as r, q in problem (5.5) that were added during the robustification process. These variables might affect the total cost in the robust solution for the case when we have $\Gamma_0 = 0$ or $\Gamma_1 = 0$. Hence the robust formulations RobC and RobV give more accurate robust solutions for the cases where $\Gamma_0 = 0$ or $\Gamma_1 = 0$.

We solve the problem (5.5) with the R package nloptr. The following Table 5.5 shows the total variances as we increase the number of uncertain parameters.

Table 5.5: Total variances for RobCV with different values of Γ_0 and Γ_1.

Gammas	$\Gamma_0 = 0$	$\Gamma_0 = 1$	$\Gamma_0 = 2$	$\Gamma_0 = 3$	$\Gamma_0 = 4$	$\Gamma_0 = 5$
$\Gamma_1 = 0$	0.7974	1.2370	1.4327	1.4799	1.5290	1.5817
$\Gamma_1 = 1$	0.8494	1.3267	1.5293	1.5792	1.6319	1.6884
$\Gamma_1 = 2$	0.8694	1.3518	1.5666	1.6177	1.6717	1.7297
$\Gamma_1 = 3$	0.8739	1.3575	1.5728	1.6247	1.6787	1.7367
$\Gamma_1 = 4$	0.8766	1.3602	1.5755	1.6274	1.6814	1.7394
$\Gamma_1 = 5$	0.8771	1.3607	1.5760	1.6279	1.6819	1.7399

The following Table 5.6 shows various total costs generated as we increase the number of uncertain parameters.

Table 5.6: Total costs for different values of Γ_0 and Γ_1.

Gammas	$\Gamma_0 = 0$	$\Gamma_0 = 1$	$\Gamma_0 = 2$	$\Gamma_0 = 3$	$\Gamma_0 = 4$	$\Gamma_0 = 5$
$\Gamma_1 = 0$	2000.00	1342.39	1155.96	1118.94	1087.19	1055.44
$\Gamma_1 = 1$	2000.00	1334.60	1154.83	1119.06	1087.31	1055.55
$\Gamma_1 = 2$	2000.00	1340.96	1154.26	1118.94	1087.19	1055.44
$\Gamma_1 = 3$	2000.00	1342.39	1155.96	1118.94	1087.19	1055.44
$\Gamma_1 = 4$	2000.00	1342.39	1155.96	1118.94	1087.19	1055.44
$\Gamma_1 = 5$	2000.00	1342.39	1155.96	1118.94	1087.19	1055.44

In Theorem 4.2 and Theorem 4.3, we have developed an upper bound on the probability that a constraint is violated. Figure 5.1 illustrates this upper bound for both the cost constraint and the variance constraint.

We find that when we increase Γ_0 or Γ_1 the upper bound on the probability is decreasing exponentially. This gives the decision maker some flexibility to choose how conservative they would like to be. It makes sense to assume that not all of the parameters are uncertain. Decreasing the uncertainty level is completely practical and efficient especially when we have a strong probability that the robust solutions will be feasible even if our assumptions are wrong.

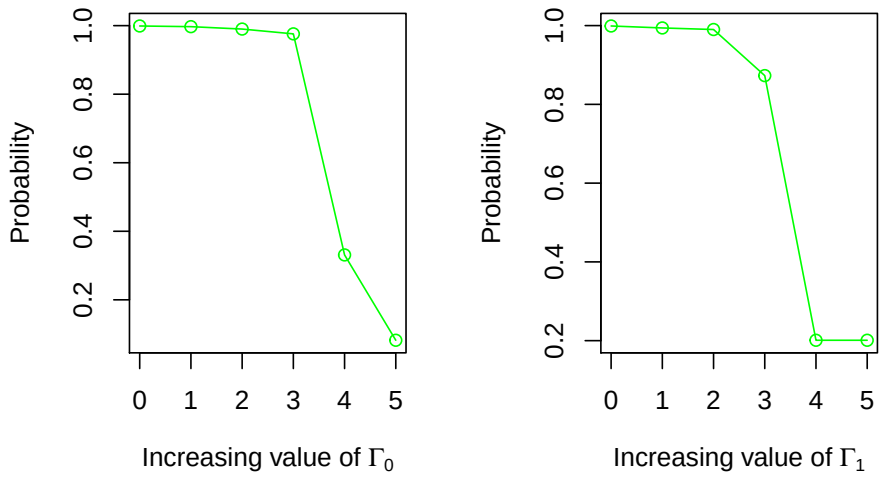


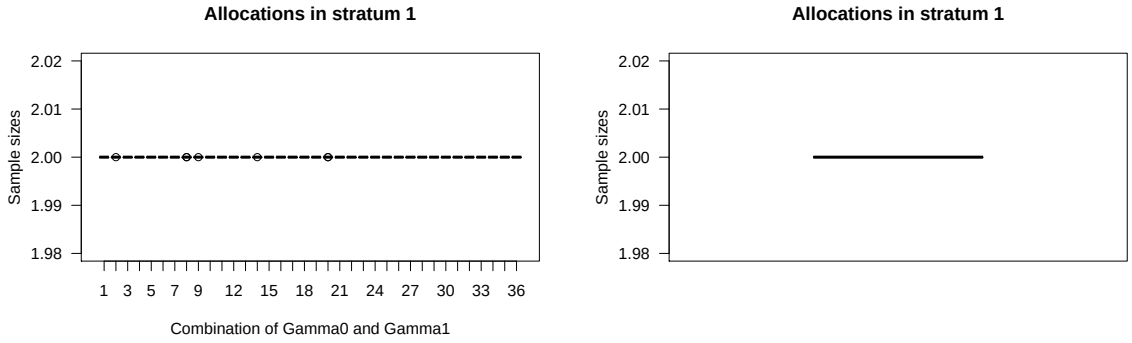
Figure 5.1: Upper bound on probabilities of the constraint violation in RobCV with increasing Γ_0 and Γ

5.3 Experiments

5.3.1 Stability Analysis

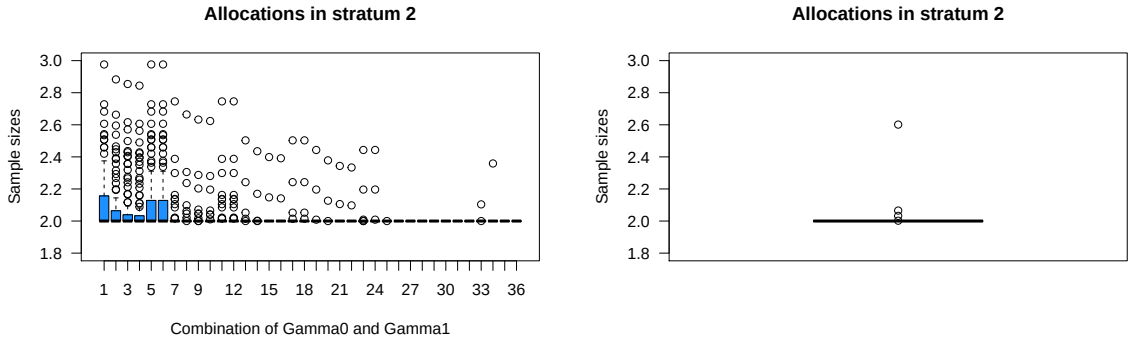
Robustness always comes with a cost, it decreases the quality of optimality *i.e.* we can achieve a better optimal value without robustness. We do some experiments to see how the robustness affects the allocation in our SAP. We want to find out the stability of robust solutions against varying stratum specific variances and we would like to see how sample sizes, total cost and total variance are affected by increasing levels of uncertainty. In this experiment, we generated 100 sets of stratum specific variances using simulation on the same population. For these 100 cases, we always considered the same cost. We calculated both the nominal (non-robust) and robust allocations with RobC, RobV and RobCV for each set of stratum specific variances.

For each of the 100 simulation runs, we compare the robust allocations from RobCV for different values of Γ_0 and Γ_1 with the nominal allocations using boxplots. The results are shown in the following figures. In the figures displaying the robust allocations we see on the x-axis the 36 combinations of $\Gamma_0, \Gamma_1 \in \{0, \ldots, 5\}$ in the following order: $(0,0), (0,1), \ldots, (0,5), (1,0), \ldots, (5,5)$.



(a) RobCV allocations (b) Nominal allocations

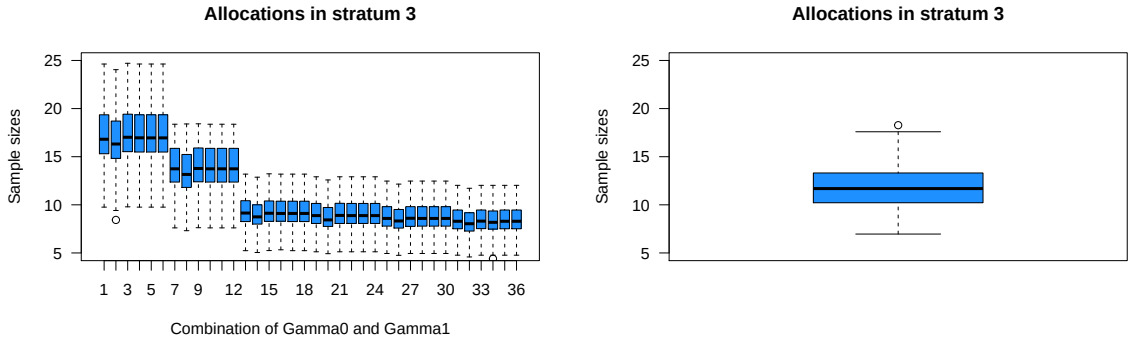
Figure 5.2: Allocations in stratum 1



(a) RobCV allocations (b) Nominal allocations

Figure 5.3: Allocations in stratum 2

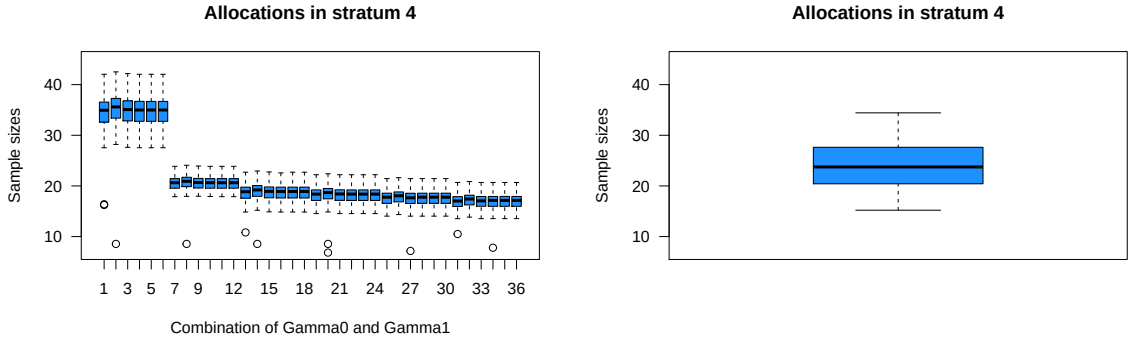
A very low variance within stratum 1 leads to $n_1 = 2 = m_1$ in stratum 1 in all cases. There is not much difference in nominal allocations and robust allocations in stratum 1 as shown in Figure 5.2a and Figure 5.2b. In Figure 5.3a, robust allocations in stratum 2 become more stable as the level of uncertainty is increasing. However with low level of uncertainty, the robust allocations are similar to the nominal allocation. The sample sizes in RobCV decrease when the uncertainty is increasing. This can be understood by the increasing level of uncertainty in the costs. A bigger sample size in a stratum might be very expensive when the cost is uncertain and that might result in an extremely high total cost. The same pattern can be observed in the allocations in stratum 3, stratum 4 and stratum 5. As the uncertainty level is increasing the robust solutions are more stable, i.e. the range of the optimal sample sizes is smaller.



(a) RobCV allocations

(b) Nominal allocations

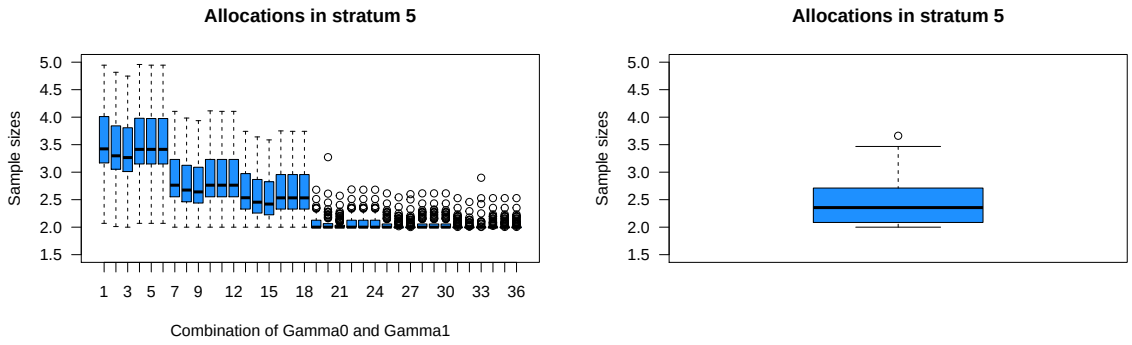
Figure 5.4: Allocations in stratum 3



(a) RobCV allocations

(b) Nominal allocations

Figure 5.5: Allocations in stratum 4



(a) RobCV allocations

(b) Nominal allocations

Figure 5.6: Allocations in stratum 5

In this experiment we also study how the total cost and the total variance of RobC, RobV and RobCV behave as compared to the nominal allocations.

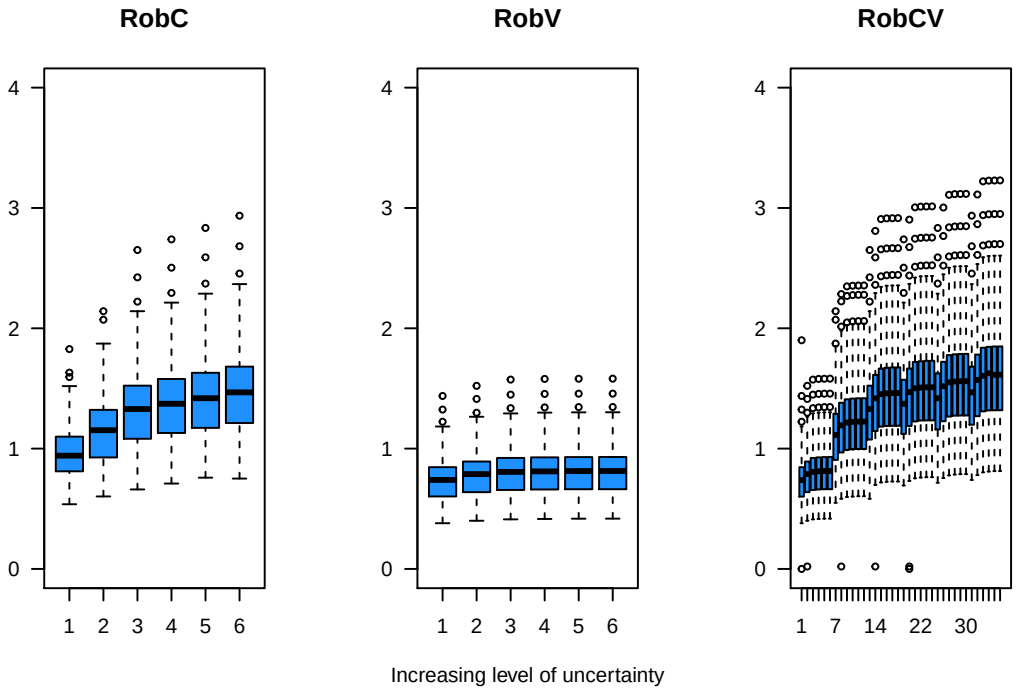


Figure 5.7: Boxplots of total variance with increasing Γ_0 and Γ_1

We see in Figure 5.7 that robust allocation has better variance as compared to the nominal allocation when the uncertainty level is lower. As the level of uncertainty is increasing the total variance is also increasing. In order to ensure feasibility we loose some optimality and that results in an increased total variance.

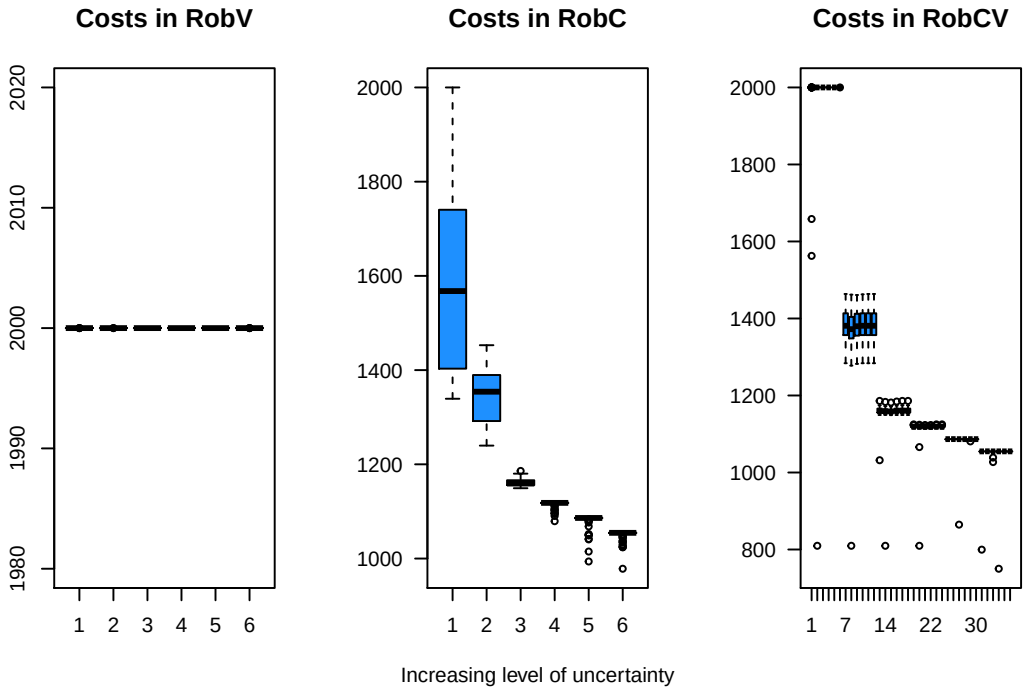


Figure 5.8: Representation of costs with increasing Γ_0 and Γ_1

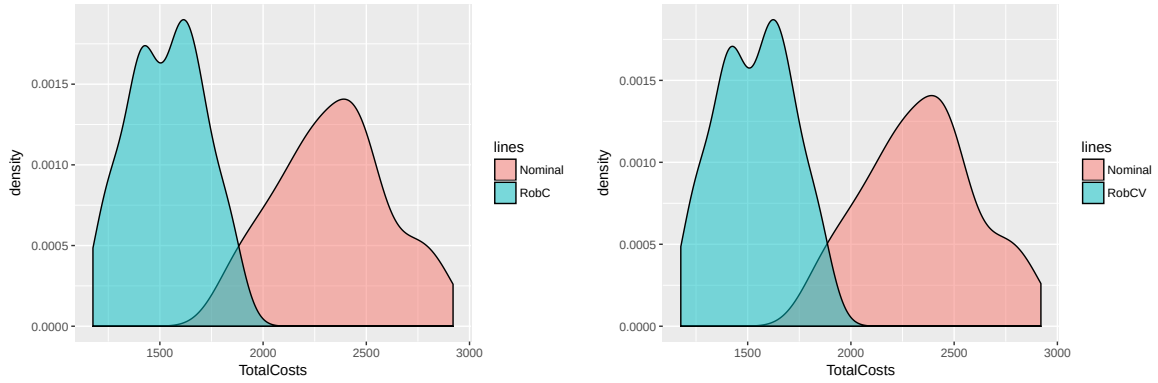
We can see in Fig 5.8 that the costs in RobV are not affected by the uncertainty level because we have uncertainty only in the variance. The costs in RobC and RobCV are monotonically decreasing with increasing level of uncertainty. This is because the robust models reduce the sample sizes when the cost is uncertain in order to avoid higher total cost.

Summarizing, we see from these expirements that in the nominal case the total variance is less than the total variances in RobC, RobV and RobCV. However, the robust allocations still have an advantage over the nominal allocation: in case there are some changes in the stratum specific variances the robust allocation is still feasible whereas the nominal allocation might violate some constraints. In order to illustrate this we do one more experiment:

5.3.2 Feasibility Analysis

We have already seen that the allocations obtained from RobC, RobV and RobCV are more stable as compared to the nominal allocation. Now we want to see the effect of uncertainty

on the nominal allocation and the robust allocations. We check whether the optimal solutions of RobC, RobV and RobCV satisfy the cost constraint by taking 100 random values between C_h and $C_h + \hat{C}_h$ as defined in Table 5.1. The density graph of various costs obtained by using 100 random parameters in the defined interval is as follows:



(a) Nominal vs RobC (b) Nominal Vs RobCV

Figure 5.9: Density plots of costs

Recall that the upper bound on the cost constraint was $C = 2000$ which, in general, is the given fixed total budget. It is clear from Fig 5.9a and Fig 5.9b that RobC and RobCV have no infeasibility whereas the nominal allocation is highly infeasible when uncertainty is considered. The infeasibility is clearer if we look at the following boxplots of nominal, RobC and RobCV allocations.

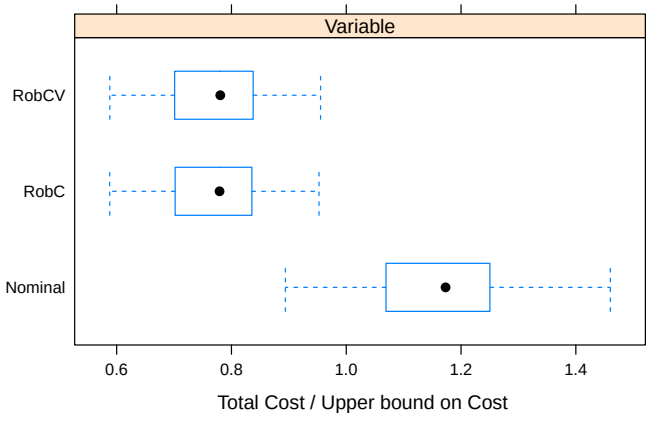
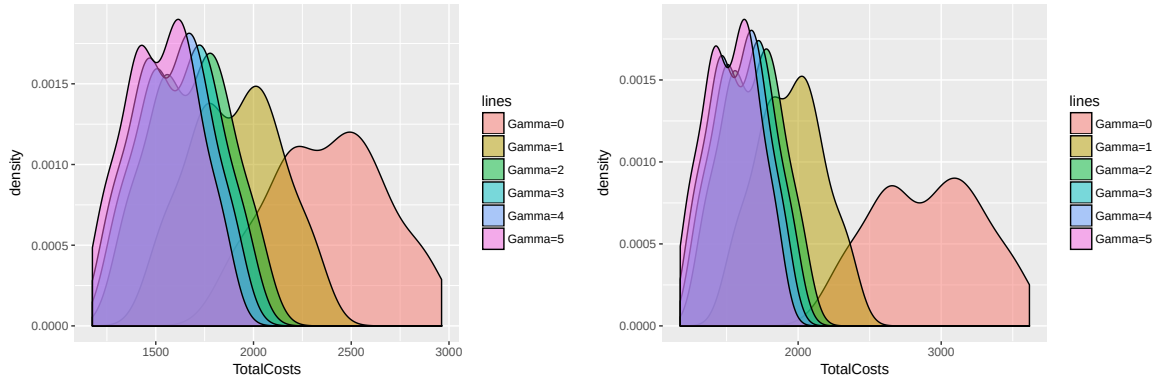


Figure 5.10: Boxplots of costs for the nominal, RobC and RobCV solutions

In Figure 5.10 we divide the total costs obtained from the random parameters by the

upper bound on the cost. Hence all values bigger than 1 represent infeasible cases and values smaller than 1 represent the feasible cases. Also, in these comparisons we considered the value of $\Gamma_0 = 5$ in RobC and RobCV which is why we do not have any infeasible cases. If we decrease the value of Γ_0, we might get some infeasible cases as illustrated in Figure 5.11.



(a) RobC (b) RobCV

Figure 5.11: Density plots of costs in all cases of RobC and RobCV for different values of Γ_0

We can see in Fig 5.11a and Fig 5.11b that as the value Γ_0 for the cost is increasing we have more feasible solutions. We know that the cases $\Gamma_0 = 0$ and $\Gamma_0 = 5$ repersent the two extreme situations where no uncertainty is present or all parameters are uncertain, respectively. We can see that at $\Gamma_0 = 3$ we have feasibility in most of the cases as we found in the probability given in Fig 5.1. It also shows that we can be less conservative on the basis of probability calculations. It can help in improving optimality without having to fear a high amount of infeasibility.

Chapter 6

Robust Allocation in the AMELIA Dataset

In this chapter, we work with the dataset AMELIA, see (Burgard et al., 2017). This is a synthetically generated dataset with approximately 3.7 million observations of 27 variables on household level and approximately 10 million observations of 33 variables on personal level. The AMELIA dataset is an artificially generated data set where the variables follow conditional distributions. Generally access to some household level and personal level data is restricted for real data and that makes the research and exploration of the data very complex. In this data set a synthetic population, generated using simulation is available and can be used for sampling allocation. Both the population and samples drawn using various techniques are provided at the AMELIA platform and can be accessed using the url `http://www.amelia.uni-trier.de/`. The generation process of the AMELIA with detailed description can be found in (Alfons et al., 2011). The data generation approach for the AMELIA dataset is explained in (Münnich and Schürle, 2003). The AMELIA dataset and its samples are provided in the form of csv and RData files.

The AMELIA is a synthetically generated dataset that can be used for sampling allocation problem. Using synthetic data generation, we can produce the required dataset with some specific information which in real life datasets is not available or sometimes it is anonymized. The synthetic data is generated from real data by anonymization, merging and taking subsets of the real data, see (Machanavajjhala et al., 2008). Synthetic data is generated by filtering the confidential informations that is not allowed to use for research purposes by individuals, such as the geographical location, contact number and IP address etc.

In the sampling allocation problem, cost is often considered as an uncertain parameter, see (Díaz-García and Garay-Tápia, 2007). If we do not have enough information about cost then uncertainty existing in the cost can make a survey very expensive. Information about cost can be obtained using the population distribution structure for various variables. In the sampling allocation problems the geographical location plays a very important role in the optimization process. The total cost of conducting a survey directly depends on the cost of interviewing individual units of the population. The cost of interviewing varies with the geographical location of persons. For example, it is less expensive to interview a person living

in a big city as compared to interviewing someone at a remote location due to the better transport connections and lesser distance from one interview location to another.

It is interesting to see how household income and other characteristics are distributed in the region under study, it could also be helpful in the stratification process. The geocoded data of household incomes of the US population for the years 2006–2010 is available at the website of the University of Michigan, (Population studies centre, 2010). This data includes zip codes with mean and median households sizes in the US. (Amunategui, 2014) has allocated this data on a Google map of the US in order to show how household incomes are distributed. The geocoded census data for the German microcensus is so far not available due to political and data security issues. However, a similar study could be conducted about Germany also if geocoded data were available.

6.1 Description of the AMELIA Dataset

The AMELIA dataset has a very large population size and we consider the household level data for our sampling allocation problem. In this dataset, the household size varies from 1 to 16 people. This dataset provides 33 household level variables however, we focus on the income variables and investigate their structure and distribution among different regional levels. On the basis of the structure and the distribution of the variables among regions, we select one of the variables and solve the sampling allocation optimization problems by considering both certainty and uncertainty in the parameters.

This dataset also provides household structure in different degrees of urbanisation. The structure and distribution of variables among the population is in well defined form. The AMELIA dataset provides 4 levels of population distribution: regions, provinces, districts and cities. The population is distributed among 4 regions, 11 provinces, 40 districts and 1592 cities. This information on the regional distribution of the population can be helpful in the stratification process. We draw some insights of this structure as follows:

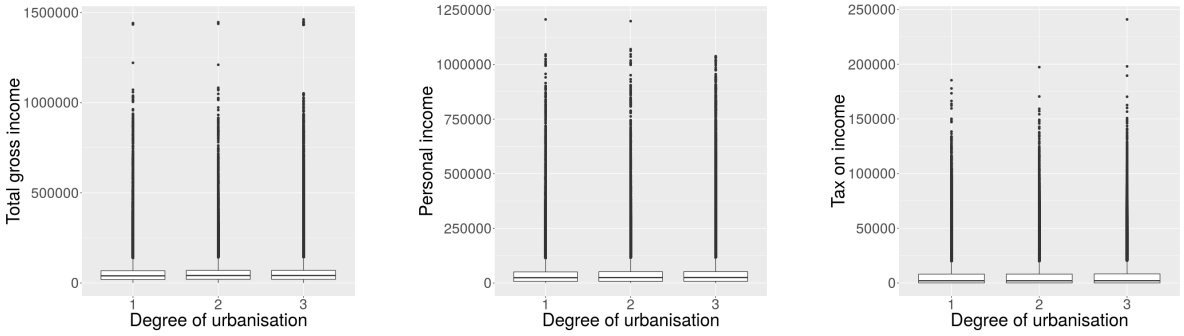


Figure 6.1: Distribution of variables for different degree of urbanisation

We start with looking into the effect of urbanisation on two income variables (total gross income and personal income) and one tax variable (tax on income and social insurance contribution). In the AMELIA dataset the degree of urbanisation is provided in 3 levels. We

show boxplots of the variables for different degrees of urbanisation in Figure 6.1. We can see that the income and tax variables are homogeneously distributed for different degrees of urbanisation.

In the AMELIA dataset the total population is divided in 11 provinces. It is also interesting to see how the variables are distributed in the different provinces. We generated boxplots for the income and tax variables in each province. We can see in Figure 6.2 that province 1 does not have as many outliers as the other provinces. Provinces 4 and 8 have the highest number of outliers for total gross income and personal income whereas provinces 2 and 6 have the highest number of outliers for the tax variable.

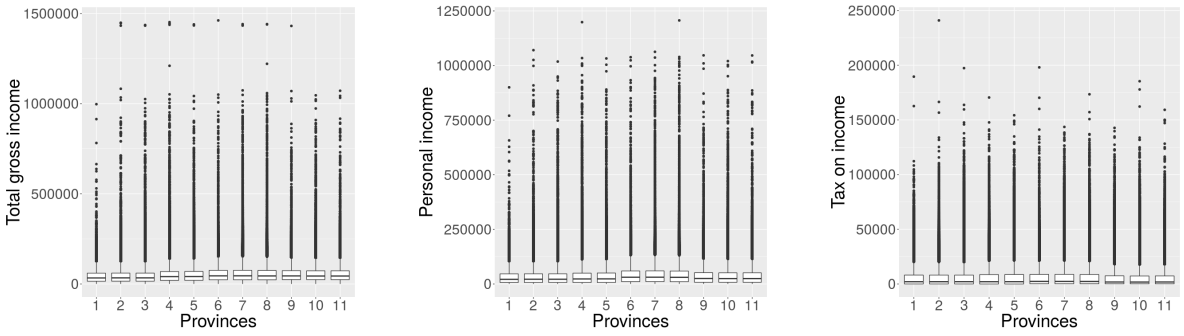


Figure 6.2: Distribution of variables in different provinces

It is also interesting to see how the whole population is distributed in various provinces. We show it using a density plot given in Figure 6.3.

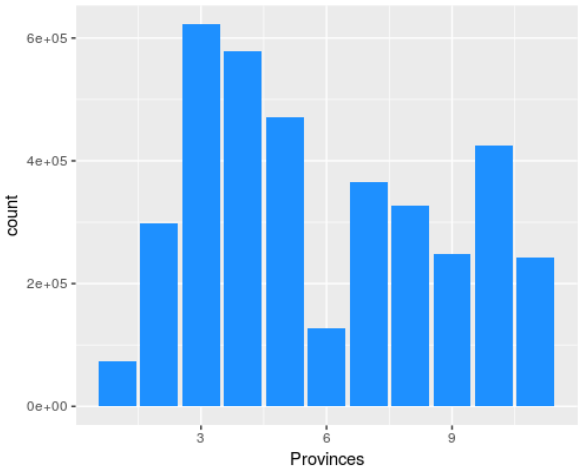


Figure 6.3: Barplot of population distribution in each province

We can see here that province 1 and province 6 have the least population and province 3 has the highest population. We can stratify our population on the basis of provinces. Thus

the cost of interviewing a person can be defined on the basis of the population of the province: the provinces with higher population will have lower cost of interviewing and the provinces with lower population have higher cost of interviewing. The reason behind such costs is that in the regions where population is higher, transportation is usually easily available and the distances in between two interview locations is usually smaller. However, regions with smaller population are generally remote locations with less transportation facilities and bigger distances between two interview locations. This is the reason why we assume the costs in the mentioned way when we do not know exact costs.

In stratified sampling, for such a large population, stratification is a very important step. Stratification can make the survey very complicated and expensive if the strata are not defined considering administrative efficiency. For example if a stratum includes population from different provinces and different cities, it will be a very complex situation from the administrative point of view. Complexity can be avoided by doing stratification using regional structure that is already available in the AMELIA dataset. So in order to avoid complexities we assume the provinces as the strata in our study.

6.2 Sampling Allocation with the Provinces as Strata

In this section, we consider the personal income as our variable of the interest for the sampling allocation problem. The problem is to select samples in each province such that we can minimize the total variance of the variable. We assume that the cost of interviewing a person living in the highest populated province is lowest, and in the least populated province the cost is highest. We also consider uncertainty in the data and assume that 10% of the stratum specific variances are subject to uncertainty and 20% of the cost parameters are subject to uncertainty.

Table 6.1: Data for the variable personal income in the AMELIA Dataset

h	N_h	S_h^2	$d_h = \frac{N_h^2 S_h^2}{N^2}$	$\hat{d}_h$	C_h	$\hat{C}_h$
1	74429	1182924552.29	458312.47	45831.25	150	30
2	298381	1225672233.33	7631974.93	763197.49	110	22
3	621864	1206393366.65	32628719.47	3262871.95	50	10
4	577935	1441844867.51	33681915.80	3368191.58	60	12
5	471154	1424703280.49	22119252.52	2211925.25	70	14
6	127344	1949091889.49	2210596.85	221059.68	140	28
7	366365	1898144764.94	17818762.68	1781876.27	90	18
8	327763	1918627099.08	14415532.31	1441553.23	100	20
9	249249	1433520514.58	6228608.28	622860.83	120	24
10	424330	1437897700.64	18107384.40	1810738.44	80	16
11	242475	1423128816.40	5851920.35	585192.03	130	26

The notations in Table 6.1 are as defined in the previous chapters. Recall that N_h denotes

the population size in stratum (or province) h given in the AMELIA dataset. The stratum specific variance S_h^2 of the variable personal income is calculated from the AMELIA dataset. We calculate $N = \sum_{h=1}^{11} N_h$ which is the total population size and $d_h = \frac{N_h^2 S_h^2}{N^2}$. The deviation $\hat{d}_h$ from d_h in case of uncertainty is calculated as $\hat{d}_h = 0.1 d_h$. The cost of interviewing one person in province h is defined to be C_h as given in Table 6.1. As mentioned earlier, we assume higher cost for strata with a smaller population and lower cost for strata with bigger population. The cost deviation $\hat{C}_h$ from C_h in the case of uncertainty is calculated as $\hat{C}_h = 0.2 C_h$.

We used the above data to solve the nominal sampling allocation problem where we minimize the variance function and we have the cost function in the set of constraints along with other constraints of the sampling allocation problem as discussed in the Section 2.3. We also solve three different cases of robust allocation problems as discussed in Section 4.2.

(i) Robust allocation when cost in uncertain (RobC)

(ii) Robust allocation when variance is uncertain (RobV)

(iii) Robust allocation when both variance and cost are uncertain (RobCV)

First of all, the upper bounds on the probability of violation of the cost constraint, as we derived in Theorem 4.3, have been calculated and are represented in Figure 6.4. We can calculate the upper buonds on the probabaility of violation before solving the optimization problems in RobC and RobCV. We can see that at $\Gamma_0 = 7$ till $\Gamma_0 = 10$ the proability upper bound on the constraint violation is very small.

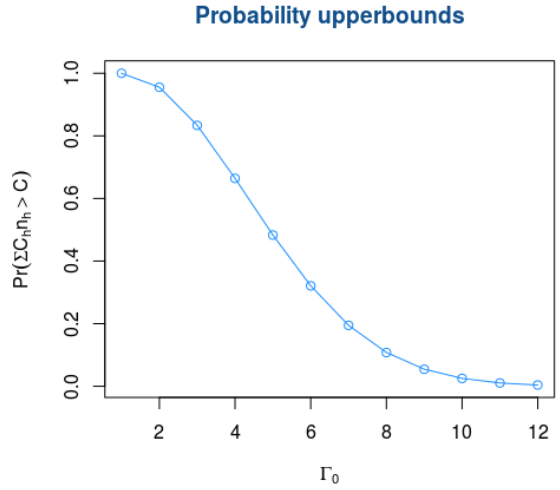


Figure 6.4: Probability bound on the constraint violation with increasing Γ_0

All the problems are solved using the **NLopt** package of R software on a computer with an Intel Core i7 processor running at 3.40GHz using 8 GB of RAM, running on ubuntu

version 14.04. RobC and RobV are solved 11 times for different values of $\Gamma_0 = 1, \ldots, 11$ and $\Gamma_1 = 1, \ldots, 11$ respectively. RobCV is solved 121 times for all combinations of values of $\Gamma_0 = 1, \ldots, 11$ and $\Gamma_1 = 1, \ldots, 11$. The total computation time in solving RobC, RobV and RobCV using R Software is 166.23 minutes, 12.19 minutes and 4577.25 minutes (76.45 hours) respectively.

Allocation

Allocation of samples for the nominal problem is plotted in Figure 6.5. The sample size in province 1 is smallest as the population size in province 1 is smallest and the cost of interviewing a person is highest. Province 5 has the biggest sample size as in this province the cost of interviewing is very small and the population size is big.

In RobC, we have 11 different allocations in each province corresponding to the 11 different levels of uncertainty ($\Gamma_0 = 1, \ldots, 11$). The different allocations in each province are represented by the boxplots in Figure 6.6. We can see that the mean value of the sample sizes is highest in province 1 and smallest in province 3. As we minimize the variance function in RobC and since province 1 has the lowest stratum specific variance, we get a bigger sample size allocated to this province.

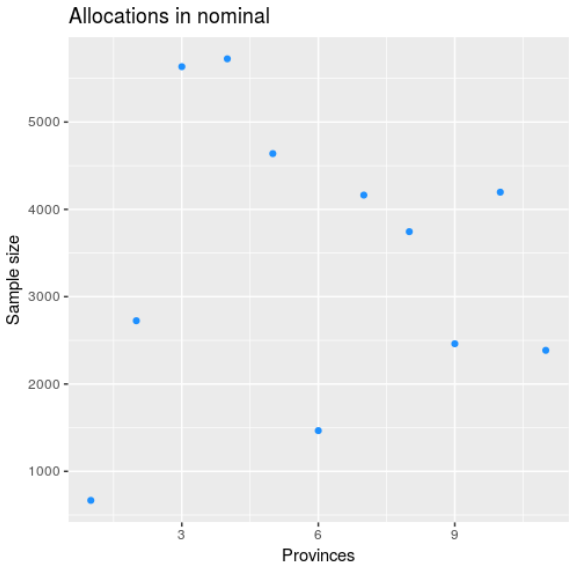


Figure 6.5: Allocations in nominal problem

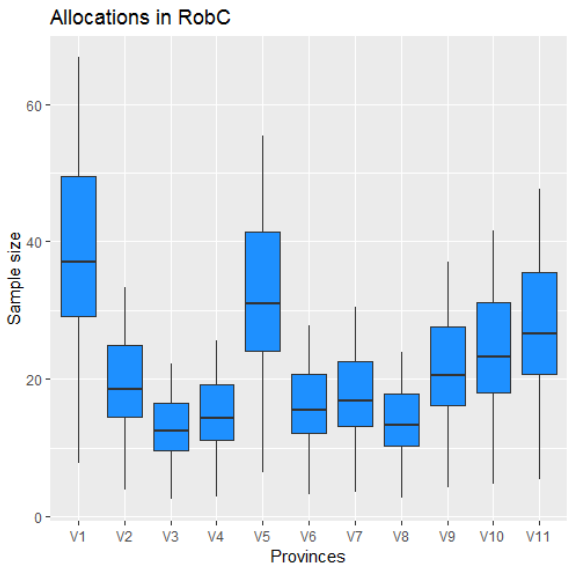


Figure 6.6: Allocations in RobC problem

Allocations in RobV are also shown using boxplots for each province in Figure 6.7. Here we have 11 different allocations for RobV for different levels of uncertainty in each province. We can also notice that allocations are more stable in RobV as compared to RobC and are not changing a lot with the change in the uncertainty level.

Allocations in RobCV are shown in the Figure 6.8. We have and 121 different allocations in RobCV for all the combinations of $\Gamma_0 = 1, \ldots, 11$ and $\Gamma_1 = 1, \ldots, 11$. Allocations in RobCV

are more unstable than RobC and RobV. The possible reason might be that the AMELIA dataset has many outliers and RobCV deals with the highest amount of uncertainty.

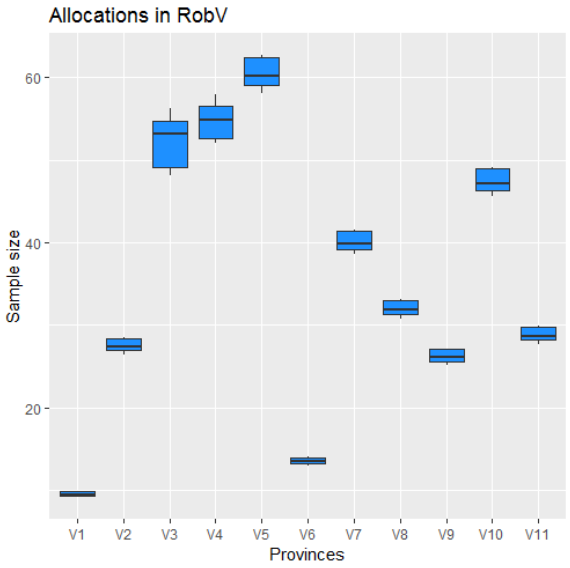


Figure 6.7: Allocations in RobV problem

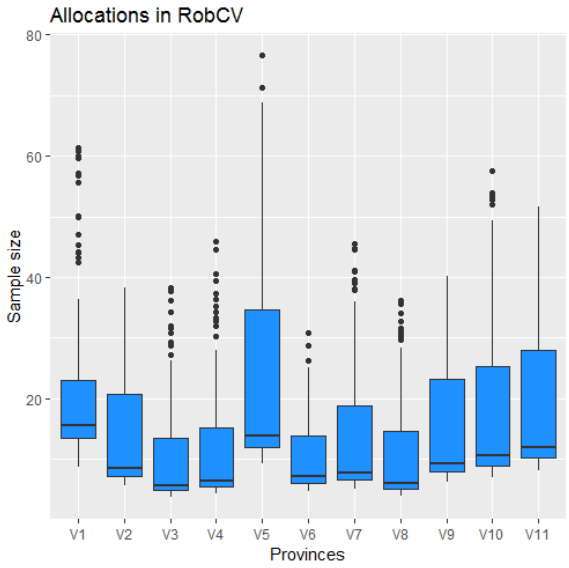


Figure 6.8: Allocations in RobCV

Figure 6.9 and Figure 6.10 show the total sample sizes in RobC and RobV respectively for increasing level of uncertainty. We can see that in both RobC and RobV the sample sizes are not decreasing continuously. The possible reason is the very high stratum specific variance in the datasets and the presence of outliers which affects the total sample size in both RobC and RobV. If the total sample sizes are not decreasing monotonically then the total cost will also be not decreasing monotonically as the total cost directly depends on the sample sizes.

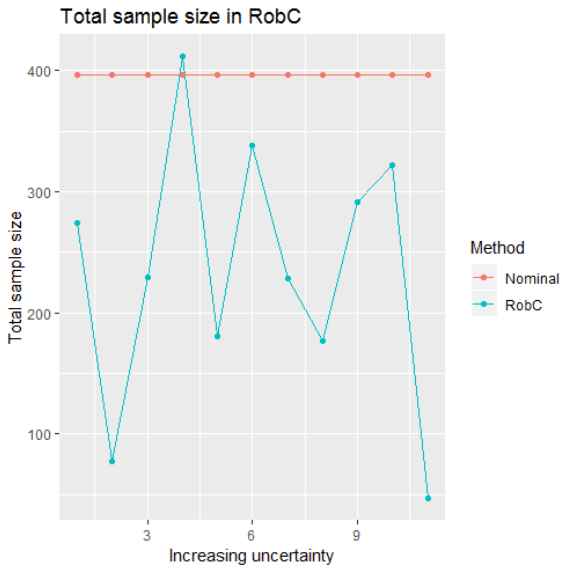


Figure 6.9: Total sample size in RobC

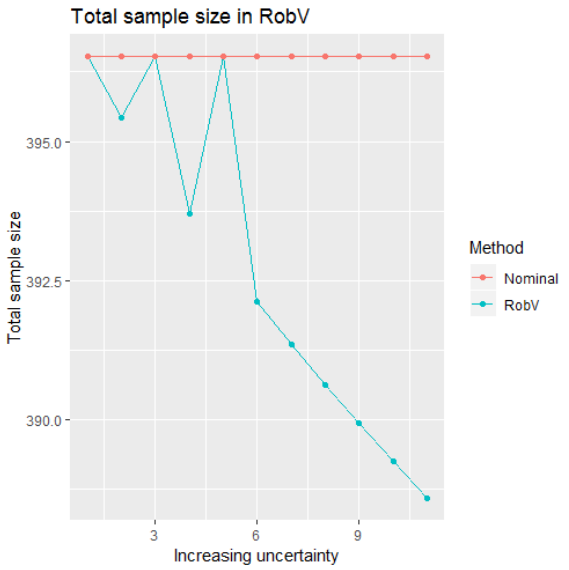


Figure 6.10: Total sample size in RobV

Total Variance

Figures 6.11 and 6.12 show the total variances in RobC and RobV for different uncertainty levels. We can see that the total variances in both RobC and RobV are bigger than the total variance in the nominal problem. This difference in the total variance can be considered as the cost of robustness. We can also see that the total variances of RobC and RobV are not continuously increasing. The reason here is the heterogeneity of the strata and the selection of outliers in the sample.

Figure 6.13 and Figure 6.14 show the total variances in RobCV when the uncertainty level Γ_0 in the cost and Γ_1 in the variance parameters are increasing. In Figure 6.13 for each Γ_0 we have 11 total variances for $\Gamma_1 = 1, \ldots, 11$ represented in the boxplot. In a similar way, in Figure 6.14 for each Γ_1 we have 11 total variances for $\Gamma_0 = 1, \ldots, 11$ represented in the boxplot. We can see that in Figure 6.13 total variances are not changing much with the change in uncertainty level of the cost parameter which is as expected. In Figure 6.14 the total variances are continuously increasing with the increase in uncertainty level in stratum specific variances. However, we can see a decrease in the total variance from $\Gamma_1 = 3$ to $\Gamma_1 = 4$, from $\Gamma_1 = 7$ to $\Gamma_1 = 8$ and from $\Gamma_1 = 9$ to $\Gamma_1 = 10$. This decrease or increase in the total variance is recorded because we have many outliers in the total population. As we select the sample completely randomly from each stratum, we can have sometimes more outliers and sometimes less outliers in the sample. This is an interesting effect and therefore we will study it in more detail in Section 6.3 by investigating more heterogeneous strata and the effect of this heterogeneity on the robust solutions.

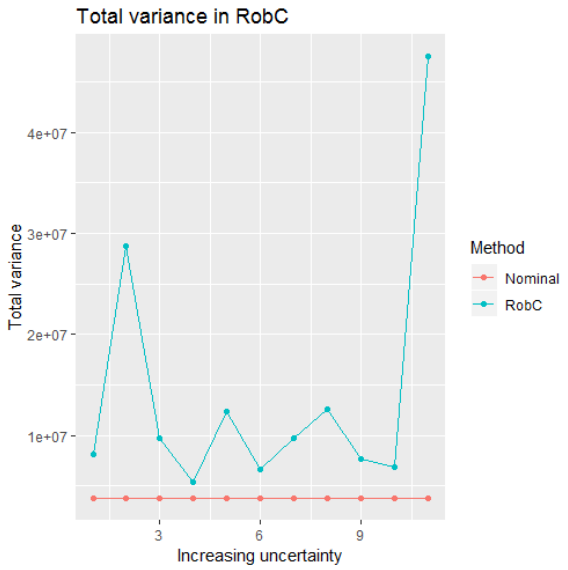


Figure 6.11: Total variance in RobC with increasing Γ_0

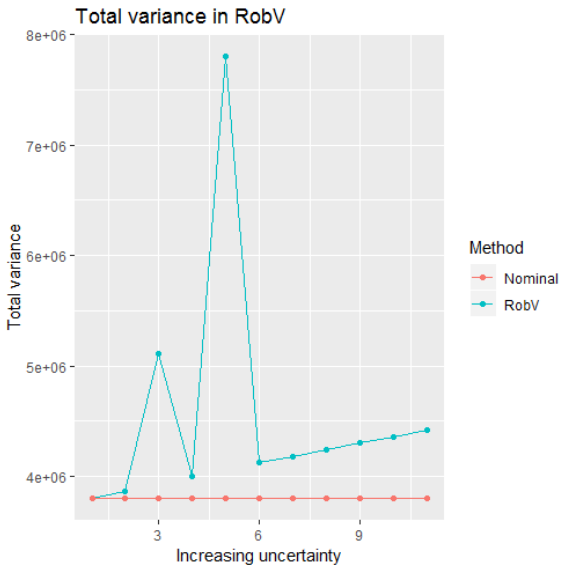


Figure 6.12: Total variance in RobV with increasing Γ_1

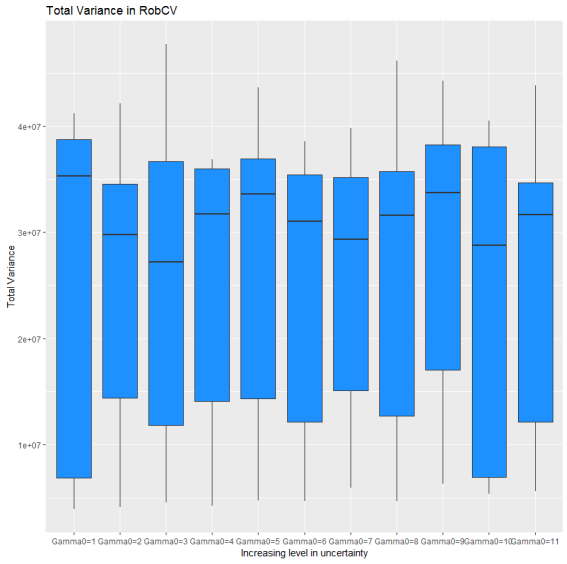


Figure 6.13: Total variance in RobCV with increasing Γ_0

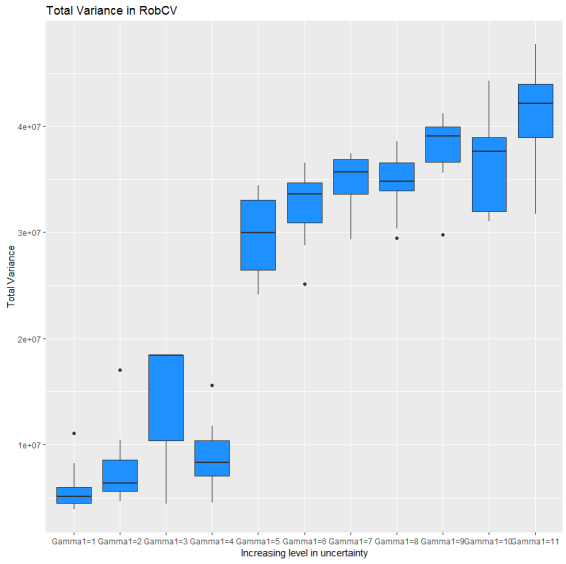


Figure 6.14: Total variance in RobCV with increasing Γ_1

Total Cost

Figure 6.15 and 6.16 show the total costs in RobC and RobV for increasing level of uncertainty in cost parameters and variance parameters respectively. In RobC total costs for all values of

Γ_0 are less than the total cost of the nominal problem. The reason behind this is that RobC has smaller sample sizes as compared to the nominal allocation in order to ensure feasibility when uncertainty exists. The total cost is expected to increase strongly for uncertain cases and RobC tries to compensate that cost in advance. In Figure 6.16 we can see that the total costs for all Γ_1 in RobV is equal to the total cost in the nominal problem because we do not consider any uncertainty in the cost parameters in problem RobV.

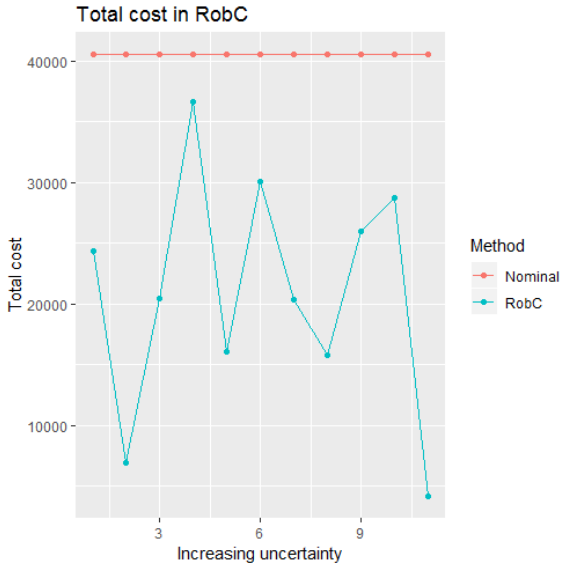


Figure 6.15: Total cost in RobC with increasing $\Gamma_0 = 1, \ldots, 11$

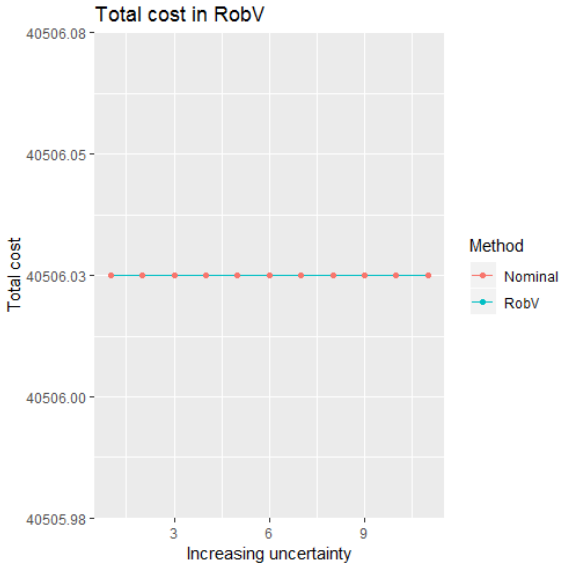


Figure 6.16: Total cost in RobV with increasing $\Gamma_1 = 1, \ldots, 11$

Figure 6.17 shows the total costs in the RobCV with increasing $\Gamma_0 = 1, \ldots, 11$. In this figure for each Γ_0 we have 11 total costs for different values of Γ_1. Figure 6.18 shows the total cost in RobCV with increasing $\Gamma_1 = 1, \ldots, 11$ where for each Γ_1 we have 11 total costs for $\Gamma_0 = 1, \ldots, 11$. The total cost in the RobCV is decreasing when uncertainty in variance is increasing. We also notice in the Figure 6.17 that uncertainty in the cost parameters does not affect the total costs of RobCV and uncertainty in the variance parameters directly affects the total costs of RobCV. It is interesting to see that when we have heterogeneous strata uncertainty of variance parameters affetcs the total cost. The possible reason here is that we have the variance function as the objective of the optimization problem of RobCV. Another reason is also that in each stratum the cost of interviewing a person is equal for all the units. However, uncertainty in the total costs will be also effective if we have different costs for each unit in a stratum.

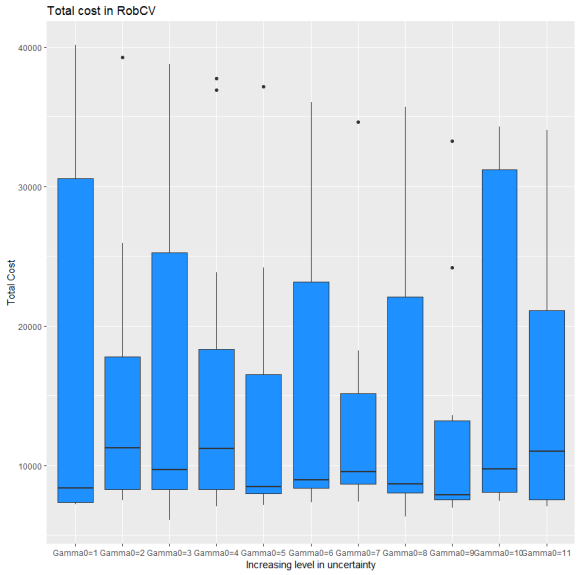


Figure 6.17: Total cost in RobCV with increasing Γ_0

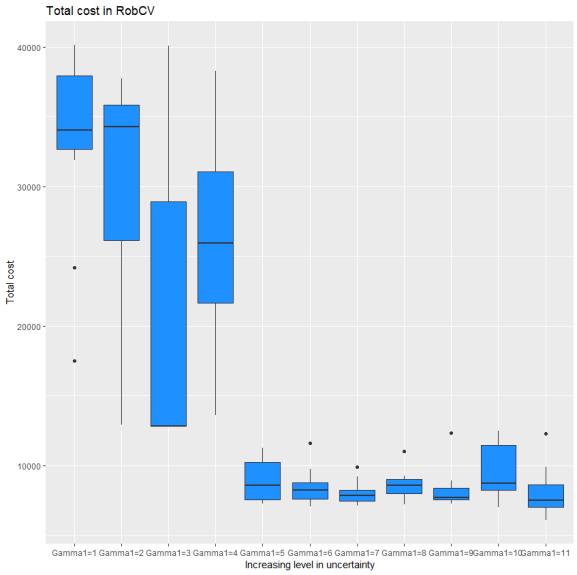


Figure 6.18: Total cost in RobCV with increasing Γ_1

We can see here that heterogeneity in the strata affects the total costs in the robust solutions of RobC and RobCV. We noticed that the total cost in RobC is not decreasing continuously. In RobV we can see that the total cost takes the value of the upper bound. The total variance of RobV is also not increasing with the increase in the uncertainty parameter Γ_1. We can state that RobC and RobV are not very stable when the population is heterogeneous and has many outliers. However performance of RobCV is still good in case of a heterogeneous population.

Feasibility Analysis

We have seen in this section that with a heterogeneous population the total variances and total costs are not completely in control of robust solutions. However, robust solutions still have an advantage that they are feasible for all the values of uncertain parameters in the uncertainty interval. We know that in RobV there is no variance constraint and the costs are considered to be certain parameters, hence the feasibility is guaranteed.

We have uncertain costs in both RobC and RobCV in the following constraint:

$$\sum_{h=1}^{11} \tilde{C}_h n_h \leq C \tag{6.1}$$

where $\tilde{C}_h$ represent the uncertain costs and $C = 40500$ as considered in the RobC, RobCV and the nominal problem. We check whether the robust optimal allocations $(n_{\Gamma_0}^{RobC})$ of RobC and the robust optimal allocations $(n_{\Gamma_0}^{RobCV})$ of RobCV for each value of Γ_0 fulfill the cost

constraint. We know that the cost constraint will not be violated for $\tilde{C}_h \in [C_h - \hat{C}_h, C_h]$ so we ignore this interval in this feasibility analysis.

We take 100 uniformly distributed random cost vectors

$$(\tilde{C}_1, \ldots, \tilde{C}_{11}) \in [C_1, C_1 + \hat{C}_1] \times \ldots \times [C_{11}, C_{11} + \hat{C}_{11}]$$

and for each value of $\Gamma_0 = 1, \ldots, 11$ we calculate the 100 total costs with the optimal allocations $n_{\Gamma_0}^{RobC}$ of RobC and $n_{\Gamma_0}^{RobCV}$ of RobCV. We also calculate the total costs of the nominal optimal allocation n^{Nom} for these 100 random cost vectors. We investigate in each case whether the cost constraint (6.1) is fulfilled.

We see in Figures 6.19 and 6.20 the density plots of the total costs in RobC and RobCV respectively for all 100 random cost vectors. We consider here only the case when we have $\Gamma_0 = 11$ in RobC and RobCV for comparison with the nominal problem. We can see that the cost constraint in the nominal problem is violated in all 100 cases whereas the cost constraint in RobC and RobCV is always fulfilled. However as mentioned, in this comparison we have taken the total costs of RobC and RobCV at the maximum uncertainty, i.e. at $\Gamma_0 = 11$. If we consider other values of Γ_0 then we might get some cases where we have infeasibility as predicted in the probability upper bound, see Figure 6.4.

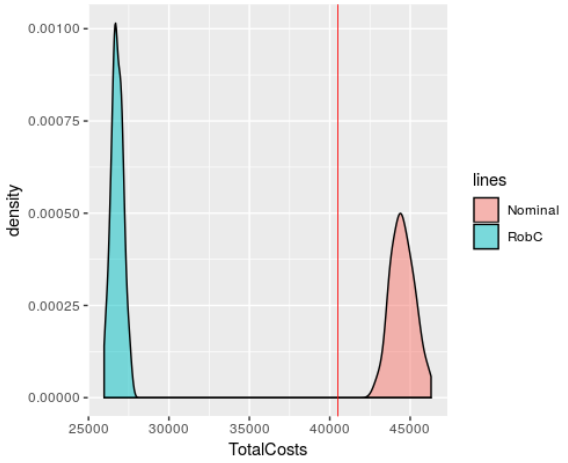


Figure 6.19: Densityplot of total cost in RobC with $\Gamma_0 = 11$ and nominal considering uncertainty. The red line displays the upper bound $C = 40506.03$

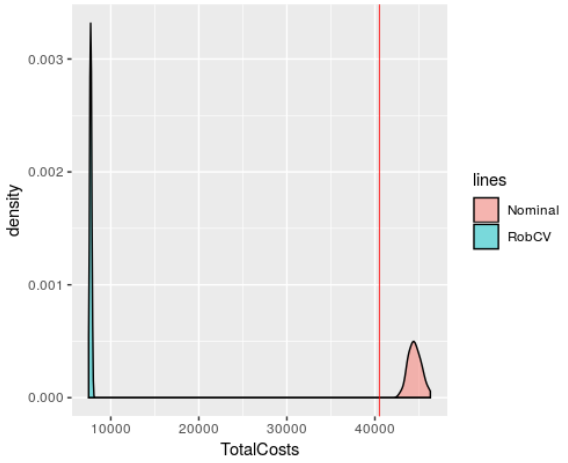


Figure 6.20: Densityplot of total cost in RobCV with $\Gamma_0 = 11$ and nominal considering uncertainty. The red line displays the upper bound $C = 40506.03$

In Figure 6.21 we can see the densityplots of the total costs in RobC for all 100 random cost vectors at different uncertainty levels. It is interesting to see that only for $\Gamma_0 = 3$ some of the robust optimal solutions violate the cost constraint (6.1). For all other values of Γ_0 the cost constraint is always fulfilled.

In Figure 6.22 we see the densityplots of the total costs in RobCV for all random cost vectors for $\Gamma_0 = 1, \ldots, 11$. For each Γ_0, we have a 100 random cost vectors and 11 different

robust optimal allocations corresponding to $\Gamma_1 = 1, \ldots, 11$. This leads to 1100 total cost values for each value of Γ_0. We see that for $\Gamma_0 = 1, 2$ and 3 none of the robust optimal solutions fulfill the cost constraint. For $\Gamma_0 = 4$ and 5 the cost constraint is fulfilled for some of the cases and violated for others. Moreover we see that for $\Gamma_0 = 6, \ldots, 11$ the cost constraint is fulfilled by all of the cases.

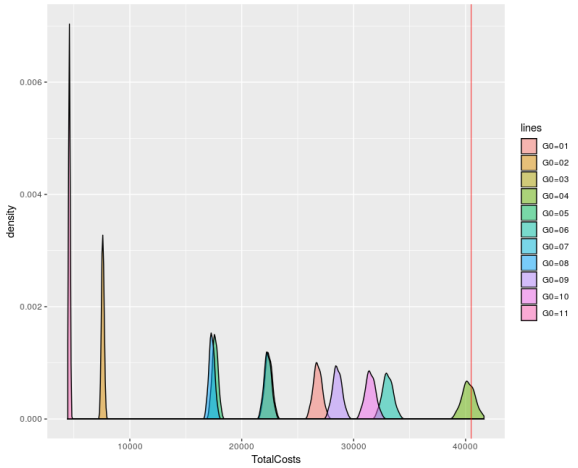


Figure 6.21: Densityplot of total cost in RobC for $\Gamma_0 = 1, \ldots, 11$. The red line displays the upper bound $C = 40506.03$

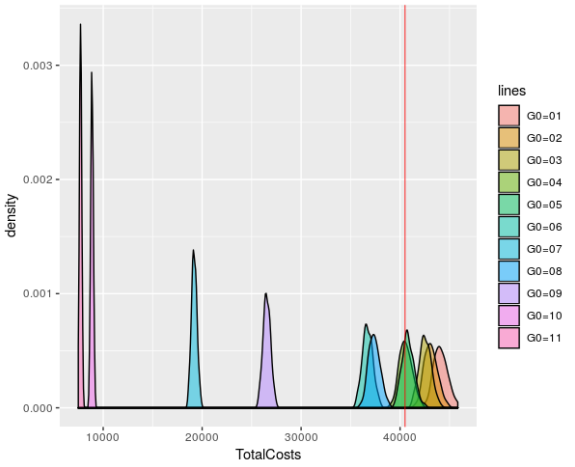
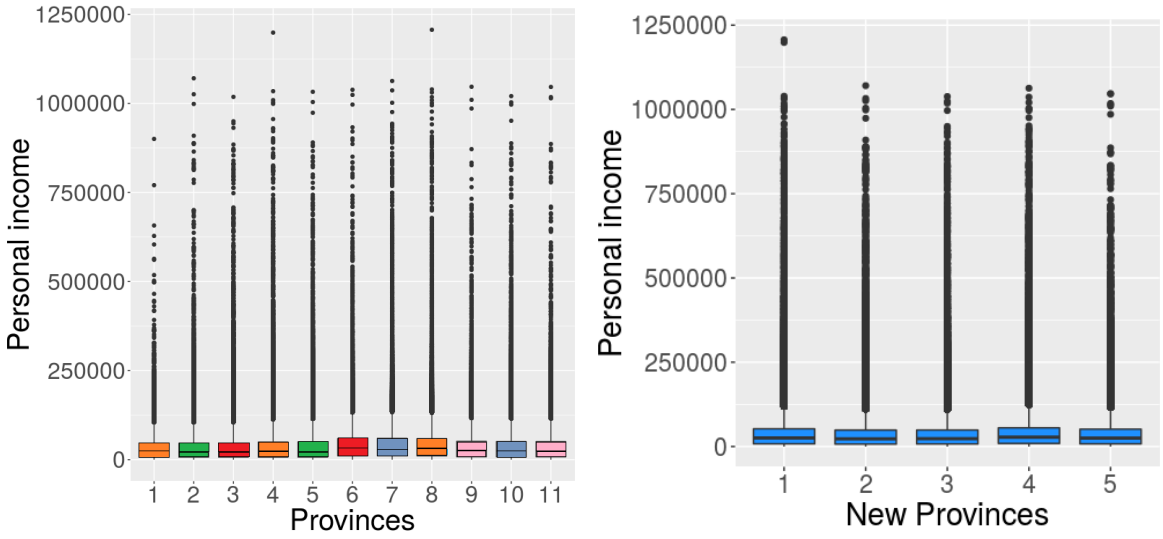


Figure 6.22: Densityplot of total cost in RobCV for $\Gamma_0 = 1, \ldots, 11$. The red line displays the upper bound $C = 40506.03$

However, we have seen in this section that with a heterogeneous data RobC does not perform very well specially when we have a very high amount of stratum specific variance. Performance of RobCV is still very good with this dataset but the computation time of solving RobCV using the AMELIA dataset is very long. It would be interesting from statistical and computational point of view if we decrease the number of strata by merging various provinces together and hence making the strata population heterogeneous and much diversely distributed. It might help in reducing the total computation time but mmakes it more complex for dealing with uncertainty in a much more heterogeneous strata.

6.3 Sampling Allocation with more Heterogeneous Strata

In order to define heterogeneous strata, we merge provinces that are taken originally from different regions and represent heterogeneous populations. In these new strata, stratum 1 is made by merging province 1, 4 and 8. Stratum 2 is made by merging province 2 and 5. Stratum 3 is made by merging province 3 and 6. Stratum 4 is made by merging province 7 and 10 and stratum 5 is made by merging province 9 and 11. The resut is a complex dataset from a statistical point of view. However the optimization problem to compute the sampling allocation has fewer variables and therefore the computation time is reduced.



(a) Boxplots of personal income in different provinces

(b) Boxplots of personal income in the new strata

Figure 6.23: Distribution of peronsal income among population before and after merge of provinces

We can see that in Figure 6.23a that the provinces with same colour boxplots are merged with each other and we get 5 strata instead of 11. In Figure 6.23b, each stratum has all the values of personal incomes from two or more provinces that were merged to make new strata. Clearly, the new strata is much more hetrogeneous than the provinvces provided in the AMELIA dataset.

Table 6.2: Data for sampling allocation with new strata in the AMELIA Dataset

Stratum h	N_h	S_h^2	d	$\hat{d}_h$	C_h	$\hat{C}_h$
1	980127	1595234057	107178920	10717892.00	5	1.0
2	769535	1348854293	55865289	5586528.90	3	0.6
3	749208	1346753826	52870484	5287048.40	2	0.4
4	790695	1662381151	72689026	7268902.60	4	0.8
5	491724	1428433269	24155889	2415588.90	1	0.2

In the Table 6.2, the new costs are chosen according to the population size of the stratum. The stratum with the biggest population size has the highest cost and the stratum with the smallest population has the lowest cost. The sample size of proportional allocation is taken to be 1% of the stratum sizes rounded off to nearest integer, i.e.

$$n^{prop} = (9801, 7695, 7492, 7906, 4917).$$

The total cost of proportional allocation

$$C = C^{prop} = \sum_{h=1}^{5} n_h^{prop} C_h = 123619$$

is considered as the upper bound on the cost constraint in RobCV. The upper bounds M_h and lower bounds m_h on the optimization variable n_h are considered to be $M_h = 10000$ and $m_h = 2$ for all $h = 1, \ldots, 5$.

The sampling allocation problem becomes much more complex when the uncertainty is introduced and solutions are not robust against uncertainty. In our case we can be sure that the probability of the robust solution being infeasible is bounded as discussed in Theorem 4.3. Figure 6.24 shows the probability upper bound that the robust solution will violate the cost constraint of the nominal optimization problem at different uncertainty levels. Again these uncertainty levels represent how many strata have uncertainty. We can see that at $\Gamma_0 = 5$ we have the maximum guarantee that the robust solution is feasible for the nominal problem, however we can also be less conservative and consider $\Gamma_0 = 4$ with a little risk known in the form of

$$Pr(\sum_{h=1}^{5} C_h n_h > C) = 0.2.$$

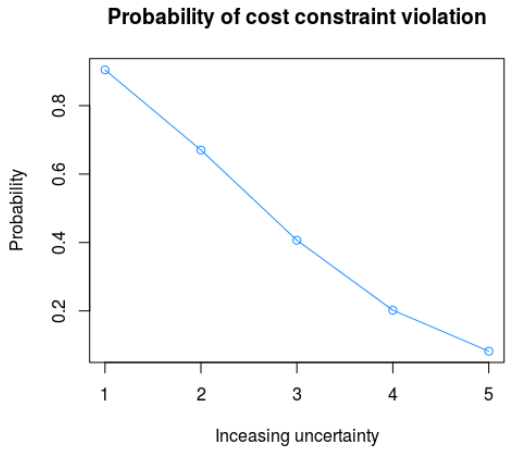
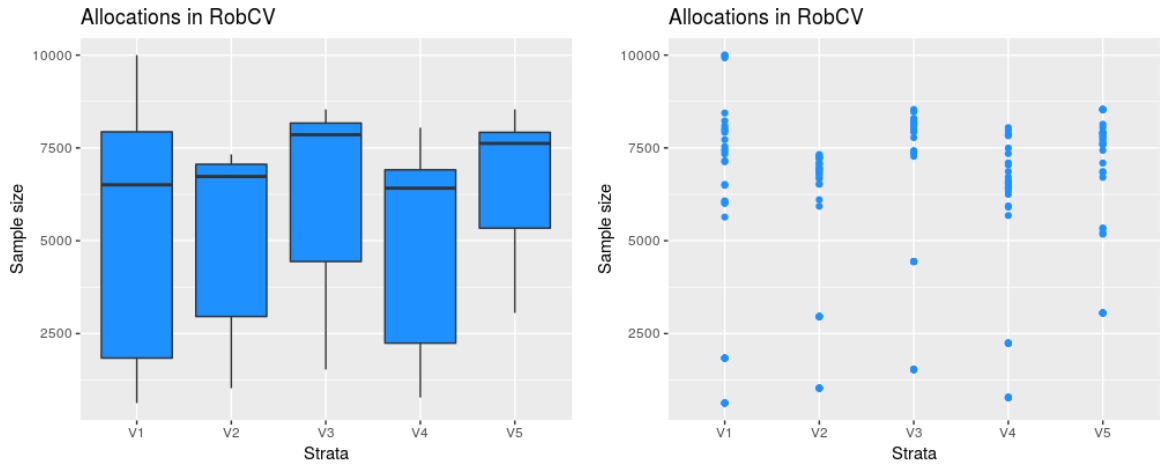


Figure 6.24: Probability upper bound on the cost constraint violation with increasing Γ_0

The problem RobCV was solved 25 times for all comibations of values of $\Gamma_0 = 1, \ldots, 5$ and $\Gamma_1 = 1, \ldots, 5$ using **NLopt** package and **Auglag** package of R software. The total computation time for solving the 25 problems was 3.321 hours.

Figure 6.25a and Figure 6.25b both represent the robust allocations of RobCV in each stratum. For each stratum we have 25 allocations for all combinations of $\Gamma_0 = 1, \ldots, 5$ and

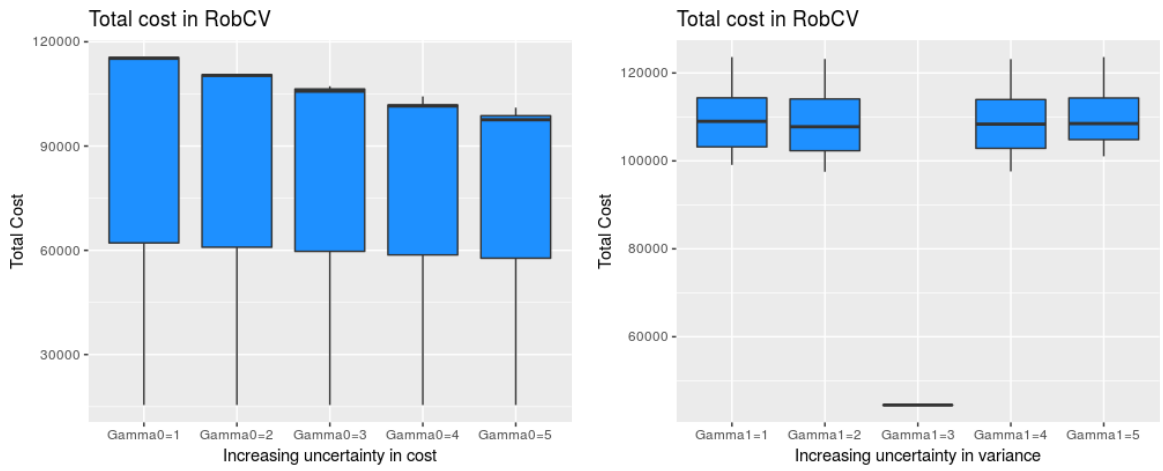


(a) (b)

Figure 6.25: Sample allocations in RobCV with more heterogeneous strata

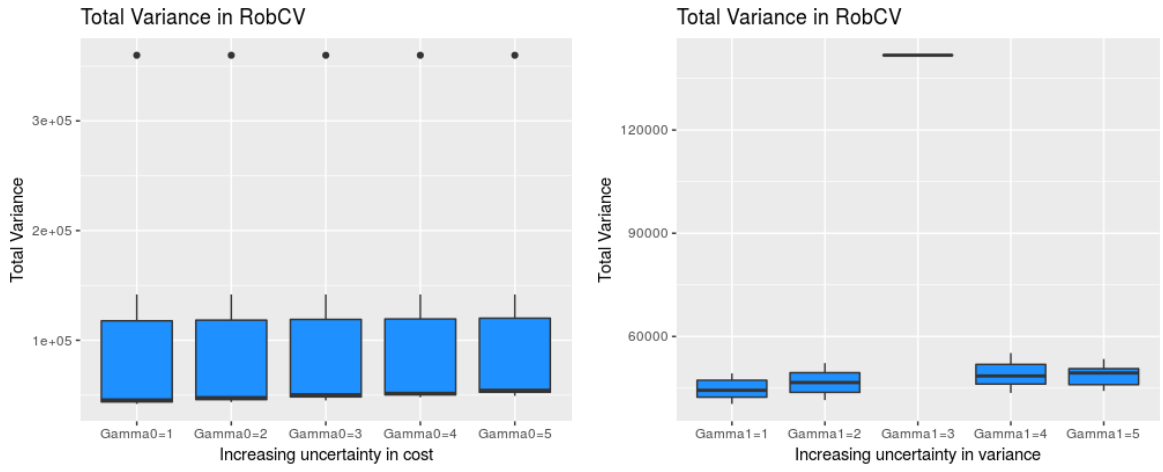
$\Gamma_1 = 1, \ldots, 5$. These different allocations are represented by boxplots in Figure 6.25a and by scatterplots in Figure 6.25b. We can see in the boxplots that the mean of the sample sizes is highest in stratum 3 and stratum 5 as these strata have the smallest cost of selecting a sample unit. We can also see that in stratum 1 the sample size equals the upper bound $M_1 = 10000$ two times even though stratum 1 has the highest cost of selecting a sample unit.

The total costs in RobCV are represented by boxplots in Figure 6.26a when uncertainty is increasing in cost parameters by $\Gamma_0 = 1, \ldots, 5$. For each value of Γ_0 we have 5 values (for $\Gamma_1 = 1, \ldots, 5$) of total costs and these 5 values are represented in the boxplots of Figure 6.26a. In a similar way we generated the boxplots of Figure 6.26b that represent the total costs in RobCV when uncertainty in the stratum specific variance $\Gamma_1 = 1, \ldots, 5$ is increasing. We can see in Figure 6.26a that with increasing uncertainty in Γ_0 the total cost is continuously decreasing. We can see some jumps in the boxplots for example in Figure 6.26b at $\Gamma_1 = 3$. The possible reasons is the heterogeneity of the popoulation as shown in Figure 6.23b.



(a) (b)

Figure 6.26: Total costs in problem RobCV with increasing Γ_0 and Γ_1



(a) (b)

Figure 6.27: Total variance in problem RobCV with increasing Γ_0 and Γ_1

We can see in Figure 6.27 that the total variance in RobCV is very slightly increasing with the increase of $\Gamma_0 = 1, \ldots, 5$ whereas the total variance is not increasing or decreasing continuously with the increase in Γ_1. We note here that in Figure 6.27b the total variance for $\Gamma_1 = 1, 2, 4, 5$ is roughly the same whereas we observe a big jump for $\Gamma_1 = 3$. This is consistent with what we saw for the total costs of RobCV where we have a very small total cost at $\Gamma_1 = 3$, see Figure 6.26b. This shows that the total sample size is very small at $\Gamma_1 = 3$.

These jumps in the total variances and total costs are because of the heterogeneity of the population. When we select only one sample then we might get outliers in our sample and that can increase or decrease our total variance. We try to deal with this problem by taking the samples 100 times for each allocation of RobCV.

100 times samples selection and outliers removal

We study three ways of dealing with the heterogenity of the dataset: First we study the effect of stratum specific variances. These were given in Table 6.2. In order to see effect of these values we generate estimated stratum specific variances following a normal distribution. We show the stimated stratum specific variances in the following table:

Table 6.3: Known and estimated stratum specific variances. The known values are the same as in Table 6.2

Stratum h	S_h^2 (known)	S_h^2 (estimated)
1	1595234057	1595197725
2	1348854293	1348821639
3	1346753826	1346744676
4	1662381151	1662394233
5	1428433269	1428408121

For the estimated stratum specific variances we solve RobCV for all combinations of $\Gamma_0 = 1, \ldots, 5$ and $\Gamma_1 = 1, \ldots, 5$ as we solved RobCV for the known stratum specific variances. Now we have two types of allocations in RobCV: (i) allocations with the estimated stratum specific variances and (ii) allocations with the known stratum specific variances. We also solve the nominal problem with both estimated and known stratum specific variances.

Second, we deal with outliers: We know that there are many outliers in the data and the sample units are selected randomly from the population. In random selection we can have outliers in our sample and that can increase or decrease the total variance and hence the sample sizes and therefore the total cost is also affected. This is one of the problems that we faced in the calculation of the total variance. Due to the amount of heterogeneity we have in our stratum, the data is very complex statistically. We illustrate this in Figure 6.28: The left most boxplot shows the personal income in the AMELIA dataset. The middle boxplot shows the cumulative personal income data in the selected samples for all 25 allocations of RobCV computed with the known stratum specific variances. We can see that there are many outliers in the sample. Since they heavily affect the total variance, we delete these outliers. The right most boxplot shows the personal income when the outliers of the samples are deleted.

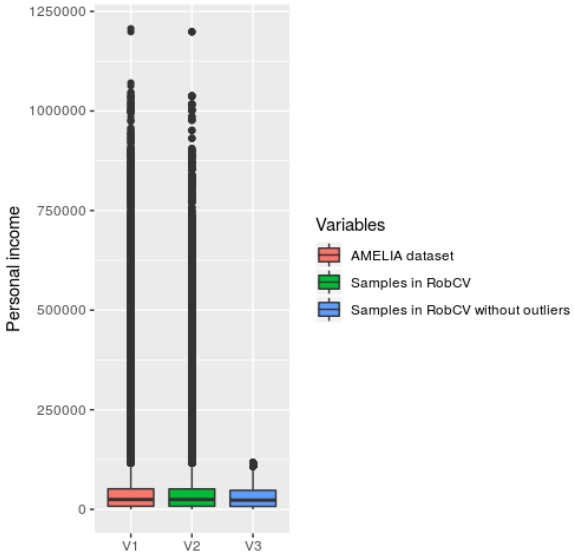


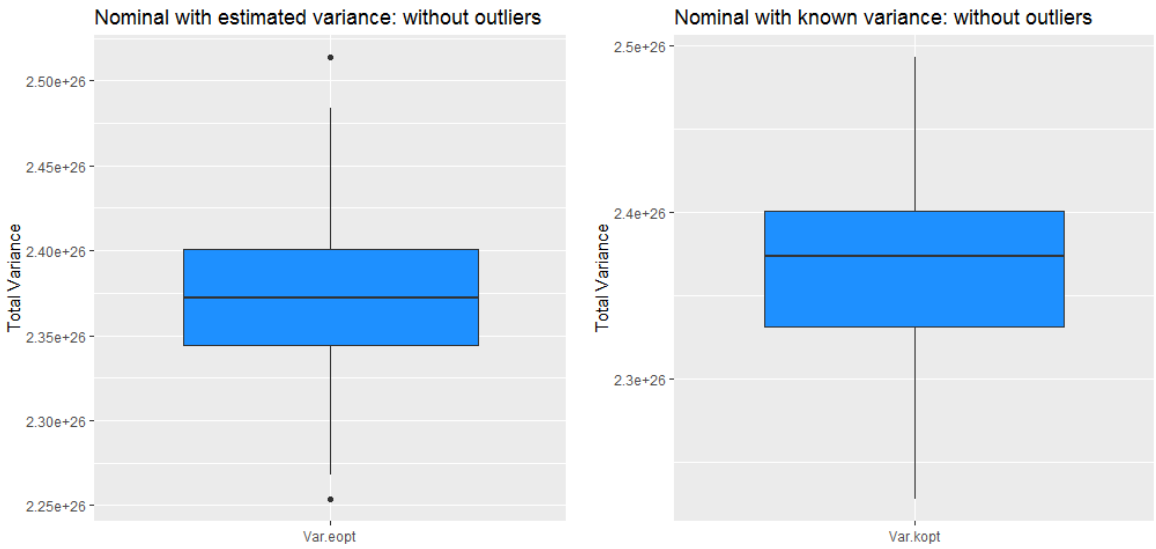
Figure 6.28: Boxplots of personal income in different phases

Third, we repeat this experiment 100 times: We have 25 different allocations in RobCV with the known variance and 25 allocations in RobCV with the estimated variance. We have one allocation for the nominal problem with known variance and one allocation for the nominal problem with estimated variance. Now for each of these allocations we select 100 samples. We delete the outliers from each sample. In RobCV with known variance 907601 sample units are selected each time and on an average 32523 outliers (3.58%) have been removed. In RobCV with estimated variance 1122480 samples are selected each time and on an average 40222 outliers (3.58%) have been removed. In nominal problem with known variance 37814 sample units are selected each time and on an average 1355 outliers (3.58%) have been removed. In nominal problem with estimated variance 37812 sample units are selected each time and on an average 1358 outliers (3.59%) have been removed. The number of population units for the variable of ineterst personal income in the AMELIA dataset is 3781289 with 135838 outliers (3.59%).

For each of the resulting 100 samples without outliers, we calculate the total variances for all the allocations of RobCV with known and estimated variance and for the nominal problem with known and estimated variance. In Figures 6.29 to 6.31, we compare these total variances.

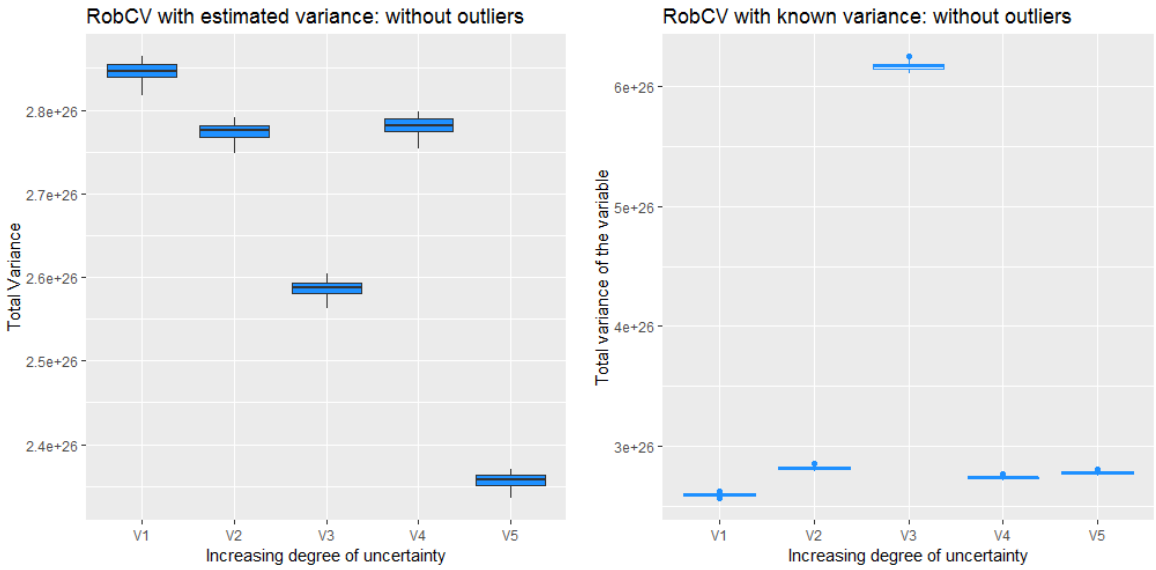
The total variances with estimated and known stratum specific variances for the nominal problem are calculated for each of the 100 samples. These 100 total variances are represented in the boxplots of Figure 6.29a and Figure 6.29b. We can see that the median of the total variances of the nominal problem with known stratum specific variances is a bit bigger than the median of the total variances of the nominal problem with estimated stratum specific variances. It is interesting that the total variance in nominal problem with estimated variance is better than the total variance of nominal problem with known variance. However, in the

overall 100 variances we do not see a big difference.



(a) (b)

Figure 6.29: Nominal problem using estimated and known stratum specific variances



(a) $\Gamma_1 = 1, \ldots, 5$ (b) $\Gamma_1 = 1, \ldots, 5$

Figure 6.30: RobCV using estimated and known stratum specific variances with increasing Γ_1

In Figure 6.30 we see the boxplots of the total variances in RobCV for increasing uncertainty levels in cost and variance represented by $\Gamma_0 = 1, \ldots, 5$ and $\Gamma_1 = 1, \ldots, 5$ and for both known and estimated stratum specific variances. For each value of Γ_1 we have 5 allocations for $\Gamma_0 = 1, \ldots, 5$. So for each Γ_1 in Figure 6.30 we have 500 values of total variance corresponding to $\Gamma_0 = 1, \ldots, 5$ and 100 samples.

It is interesting that we see again in Figure 6.30b a jump in the total variance of RobCV with known variances at the $\Gamma_1 = 3$. The reason is that the allocations (i.e. the sample sizes) of RobCV were calculated considering the known stratum specific variances of personal income without removal of outliers. However, in Figure 6.30a we see a small jump at $\Gamma_1 = 4$ in the total variance of RobCV with estimated stratum specific variances. We can see that for $\Gamma_0 = 1, \ldots, 5$ the difference in the total variances is smaller as comapared to the difference for $\Gamma_1 = 1, \ldots, 5$ and that is why we get the boxplot as a line in Figures 6.30a and 6.30b.

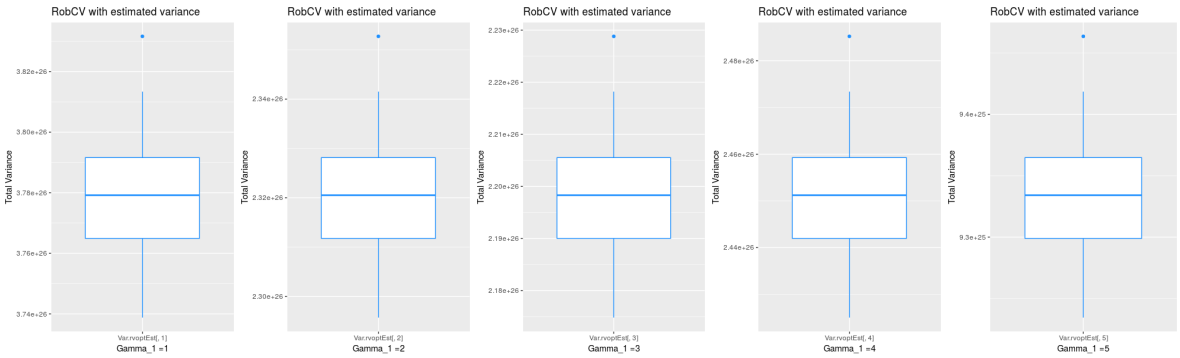


Figure 6.31: Total variance in RobCV for $\Gamma = 1, \ldots, 5$

In Figure 6.31 we can see separate boxplots of the total variances for each Γ_1 to visualise them better. As mentioned earlier in the boxplots for each Γ_1, we have all the cases of $\Gamma_0 = 1, \ldots, 5$. We can see that the interquartile range is rather small so we can conclude that the total variances are not affected much by Γ_0. In a similar way we could generate boxplots of the total variances of RobCV with known stratum specific variances and draw the same conculsion.

Feasibility Analysis

We now return to the setting where we only consider the known stratum specific variances and we do not delete outliers. We have seen how the robust solutions perform when there are many outliers in the dataset. We have noticed that in an extremely heterogeneous dataset the total variances and total costs are hard to control according to the uncertainty level. However, robust solutions always have the advantage that they are feasible and it does not matter what value the uncertain parameters take in the defined uncertainty interval.

We check the feasibility of our robust solutions for the cost constraint as done for the robust allocations in Section 6.2. We know that the total costs will not be violated in the interval $[C_h - \hat{C}_h, C_h]$ so we have ignored it in this analysis. We take 100 random values of $\tilde{C}_h \in [C_h, C_h + \hat{C}_h]$ in the defined uncertainty interval of cost parameters for each $h = 1, \ldots, 5$. For each of these 100 set of cost parameters we computed the total costs and checked whether they violate the cost constraint. We recall that the upper bound C in the cost constraint is $C = 123619$.

For the nominal problem, the total costs for 100 sets of cost parameters from the uncertainty interval are represented in Figure 6.32. In all of the cases the total costs in the nominal problem exceed the cost upper bound.

In Figure 6.33 a comparison is presented between the total costs of the nominal problem and the total costs of RobCV for the 100 sets of cost parameters. The 100 total costs in this figure are represented by the densityplot. We can see that the total costs of RobCV are always smaller than the upper bound and the total costs of the nominal problem are always bigger than the upper bound. In this figure, we have considered the case of $\Gamma_1 = 1$ and $\Gamma_0 = 5$ for RobCV which provides the maximum ensurance that the cost constraint will not be violated.

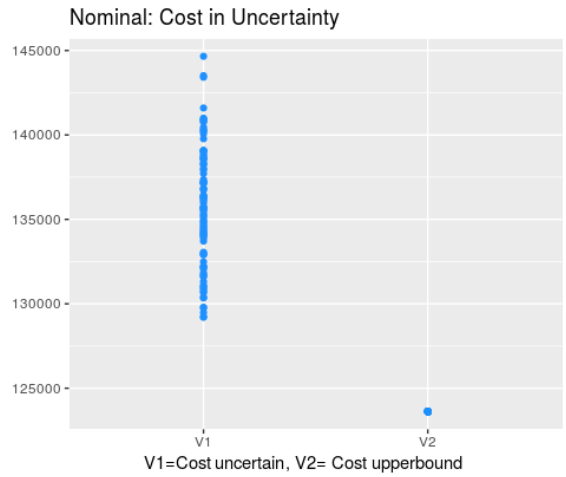


Figure 6.32: Total costs in the nominal problem when uncertainty exists

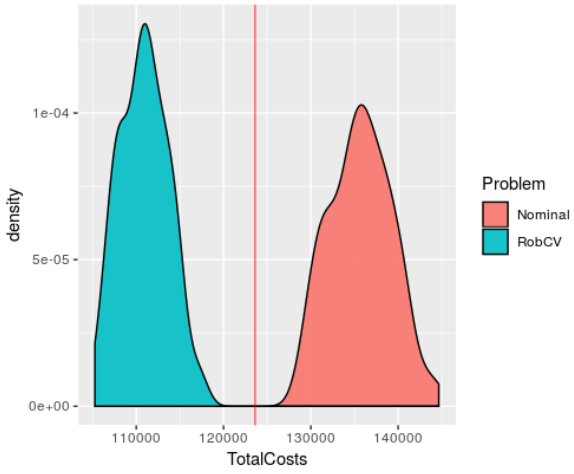


Figure 6.33: Comparison of total costs in RobCV and the nominal problem. The red line displays the upper bound $C = 123619$

However if we look at the different levels of uncertainty in the costs i.e., $\Gamma_0 = 1, \ldots, 5$ then there might be some violations.

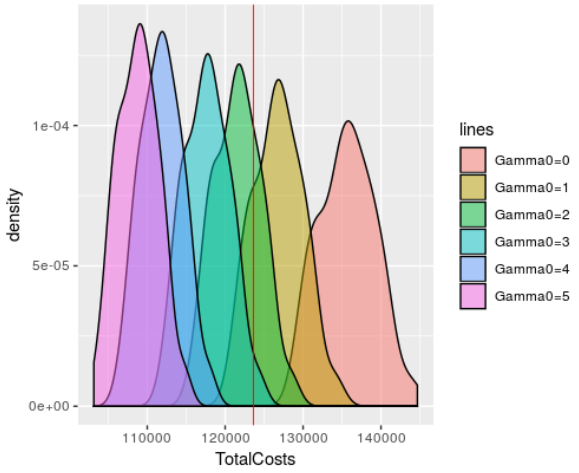


Figure 6.34: Densityplots of total costs in RobCV for $\Gamma_0 = 1, \ldots, 5$ and $\Gamma_1 = 1$. The red line displays the upper bound $C = 123619$

In Figure 6.34 we can see different levels of Γ_0 and the corresponding densityplots of the total costs. It is very interesting to see that at $\Gamma_0 = 0$ the densityplot is very similar to the densityplot of the nominal problem, however, this is not always the case. We see that at $\Gamma_0 = 1$ we have only a few of the total costs which are smaller than the upper bound. For $\Gamma_0 = 2, 3$ we have only few of the total costs bigger than the upper bound and for $\Gamma_0 = 4, 5$ all of the 100 total costs are smaller than the upper bound in the cost constraint.

Chapter 7

NRW Income and Taxation Data

Microcensus in Germany is being carried out since 1957. The German microcensus survey is integrated into the Labour Force Survey of the European Union (EU Labour Force Survey). The microcensus aims to collect official statistical figures about the population. It is helpful to have an inference about the labour market, economic and social activity of the population, education and training situations and on health and housing situations. Stratified sampling techniques are used to select samples and the sample size is taken to be 1% of the people and households in Germany. The microcensus provides very important data not just for administrative purposes but also for research purposes. We have already seen the robust allocation approach for simulated datasets in Chapter 5 and Chapter 6. In this chapter, we focus on the real dataset on income and taxation of North Rhine-Westphalia (NRW), Germany. This data is available for research purposes from (Forschungsdatenzentrum, 2001).

Income and taxes are two important factors that have a direct impact on the financial situation of a country (Alesina and Perotti, 1997). Income and taxation data are collected from different states of a country. This data helps the government in taking financial decisions. For our survey statistical problem we take the income and taxation data of North Rhine-Westphalia. This data includes information about a population of size $274,743$ with their age, sex, income (yearly and monthly), amount of tax paid in a year, taxation level and social structure etc.

We do not have geocoded data available from the statistical office in Germany. This would be desirable because it can be helpful to identify the interview cost per sample unit based on the location. The NRW income and taxation dataset does not contain cost of interviewing a person which is needed if we have a fixed budget of conducting a survey. Self interviews and online interviews can reduce the total costs but they are poor in quality and can sometimes result in non response. This is why the interviews are generally carried out by visiting the persons. Some of the surveys are made mandatory by the German law so that non responses can be eradicated, (Schwarz, 2001).

A sampling allocation problem has to be solved for this data considering the objective that the selected samples should represent the whole NRW population in terms of tax and income. The first step is to clean the data as some values in the data might be typos and do not make any sense, for example NAs and some negative values. We can not have these

values as input during the stratification and optimization process. The dataset contains 32101 such units where the values were either wrong or NAs. We simply delete these values and their corresponding data lines. In this way we can get rid of the clearly visible wrong values, however there still might be other typos which we can not detect. This is why we need a robust allocation process so that we can achieve good compromising solutions even if some entries in our data are wrong. Deletion of rows can also affect the quality of the optimal allocations, however this can be compensated by increasing the uncertainty level in the remaining data.

After cleaning the data, the first step is stratification of the population. We have 5 levels of taxes in Germany and as our objective is related to the income and tax, we stratify the whole population in 5 strata each representing a tax level.

We have some variables available in the NRW data which are not directly related to the income and taxation such as age group and social structure of the population units. The age group variable is defined as follows:

- 0 denotes no entry in the age column
- 1 denotes 0 to 19 years of age
- 2 denotes 20 to 29 years of age
- 3 denotes 30 to 39 years of age
- 4 denotes 40 to 49 years of age
- 5 denotes 50 to 59 years of age
- 6 denotes 60 to 69 years of age and
- 7 denotes 70 years or older.

In a similar way the social structure variable is defined as follows:

- 0 denotes no income
- 1 denotes mostly non-self-employed with reduced provisioning allowance
- 2 denotes mostly non-self-employed with unrestricted provision
- 3 denotes predominantly pension recipients with a reduced pension plan
- 4 denotes predominantly pension recipients with unrestricted provision
- 5 denotes mostly self-employed with gross wages
- 6 denotes mostly self-employed without gross wage.

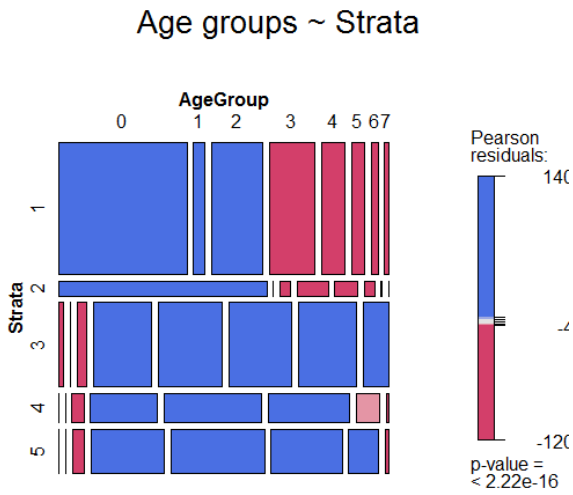


Figure 7.1: Age groups distriution in the strata

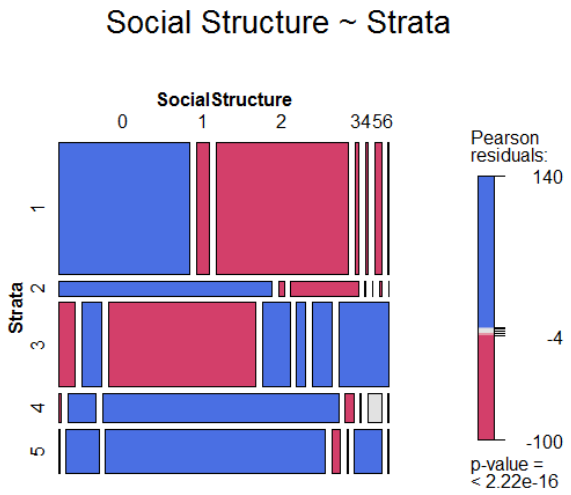


Figure 7.2: Distribution of social structure in the strata

In Figure 7.1 we can see the distribution of the different age groups among the strata. We can see that the people of age 0 to 29 years (age group 1 and 2) pay mostly the level 1 and level 2 tax whereas people of age 40 to 69 years (age group 4, 5 and 6) pay mostly in the tax level 3. People of age 30 to 39 years (age group 3) pay mostly level 4 and level 5 taxes. We can also see that there are is a big group who pay level 1 tax but they have no entry in the age group. Pearson correlation test suggests that there is a correlation between the age group and tax levels.

In Figure 7.2 we can see the distribution of the variable social structure among the strata. We can see that stratum 1 has a majority of those who have no income (social structure 0). We can see that level 1 tax has the least number of self employed people (social structure 5 and 6) whereas level 3 tax has highest number of self employed people. Pearson correlation test suggests that there is a correlation between social structure and tax levels.

Now we look at the income variables provided by this dataset. We have 3 main income variables:

1. Income (Einkommen)
2. Sum of the income (Summe der Einkommen)
3. Total income (Gesamtbetrag der Einkünfte)

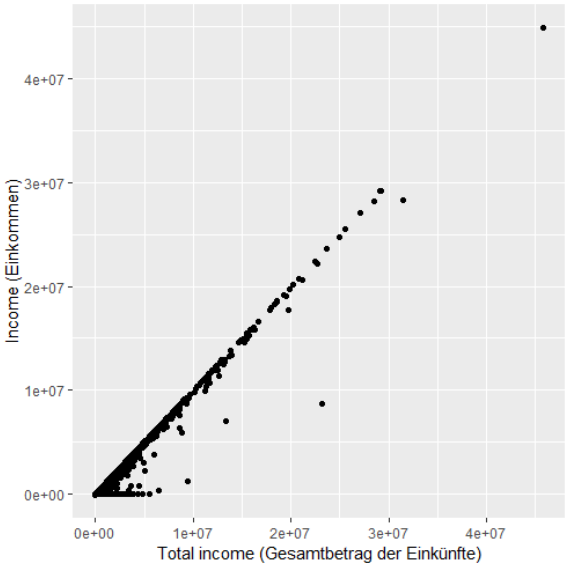


Figure 7.3: Income vs Total income scatter plot

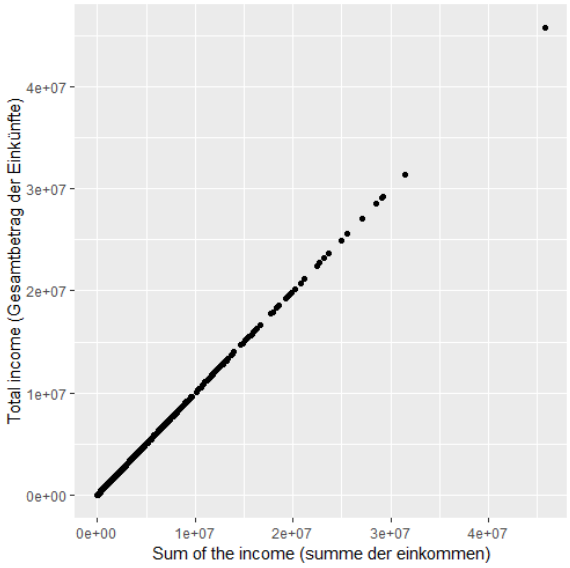


Figure 7.4: Sum of the income vs Total income scatterplot

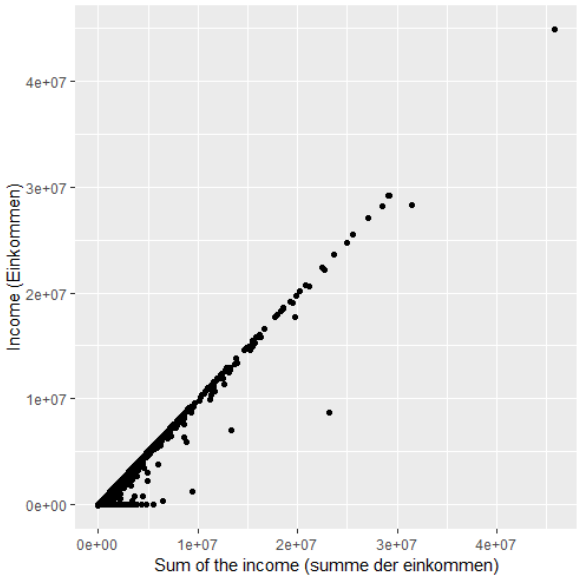


Figure 7.5: Sum of the income vs Income scatterplot

We check the possible correlations among these variables. In Figures 7.3, 7.4 and 7.5 we can see that there is a linear correlation between these variables. For this reason, we will consider "Sum of the income" as our variable of interest for the robust optimization problems.

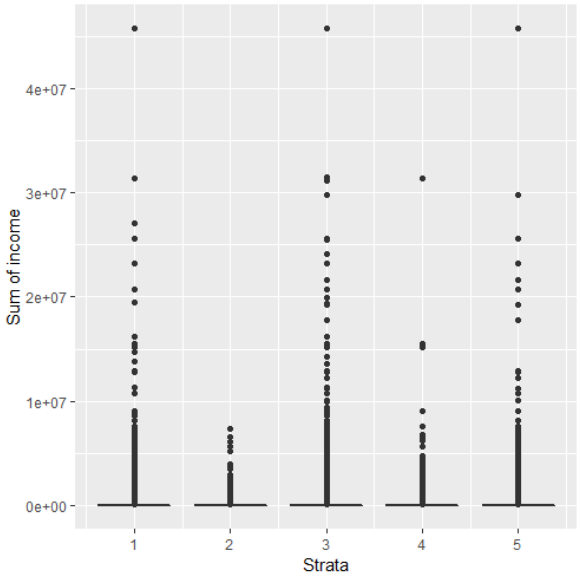


Figure 7.6: Boxplots of sum of the income in each taxation level

It is interesting to see how the variable of interest is distributed among the 5 different tax levels. In Figure 7.6 we show boxplots of the sum of the income in each tax level which we consider as strata. We observe heterogeneity in each of the strata. However, the heterogeneity in stratum 2 is not as bad as in the other strata. The heterogeneity existing in the variable of interest makes the sampling allocation process more complex.

7.1 Allocations and Analysis

In this section we present the results of robust allocations for the income and taxation data of NRW. As mentioned earlier the unavailability of cost parameters in the data leads us to solve the problem without considering cost constraints. We assume that the costs of conducting the survey do not affect the allocation of samples. The sampling allocation problem in this scenario can be formulated as follows:

Table 7.1: Stratum size (N_h) and stratum specific variances (S_h^2) for the variable sum of the income

h	N_h	S_h^2	$d_h = \frac{N_h^2 S_h^2}{N^2}$	$\hat{d}_h = 10\% d_h$
1	105584	$1.336060365 \times 10^{11}$	$2.5006562817 \times 10^{10}$	$2.500656282 \times 10^{9}$
2	12178	$3.8389023459 \times 10^{10}$	9.5585211×10^{7}	9.558521×10^{6}
3	67204	$3.66309611529 \times 10^{11}$	$2.7776035552 \times 10^{10}$	$2.777603555 \times 10^{9}$
4	23701	$1.0979788868 \times 10^{11}$	$1.035521281 \times 10^{9}$	1.03552128×10^{8}
5	35386	$2.57341088083 \times 10^{11}$	$5.410081133 \times 10^{9}$	5.41008113×10^{8}

We calculate the stratum specific variances of the variable sum of the income for each stratum. This stratum specific variance is an uncertain parameter in the optimization process. We assume that there is 10% uncertainty in the stratum specific variance. The final parameters used for the sampling allocation problem are given in Table 7.1.

We take the sample size to be 1% of the total population size, i.e.,

$$\beta = 0.01 \sum_{h=1}^{5} N_h = 2440.53.$$

We use **NLopt** package of R software to solve all the optimization problems. The nominal problem without a cost constraint can be written as problem (7.1).

$$\begin{aligned} &\min \sum_{h=1}^{5} \frac{d_h}{n_h} \\ &\text{s. t.} \\ &\sum_{h=1}^{5} n_h \leq \beta \\ &m_h \leq \ n_h \leq M_h \quad \forall h = 1, 2, ..., 5 \end{aligned} \tag{7.1}$$

where the notations are same as explained in Chapter 4. We consider $m_h = 2$ and $M_h = N_h$ for all $h = 1, \ldots, 5$. The robust formulation RobV of the above problem can be written as follows:

$$
\begin{aligned}
\min \quad & \sum_{h=1}^{5} \frac{d_h}{n_h} + l\Gamma_1 + \sum_{h=1}^{5} p_h \\
\text{s.t. } & l + p_h \geq \frac{\hat{d}_h}{n_h} \quad \forall h = 1, 2, \ldots, 5 \\
& \sum_{h=1}^{5} n_h \leq \beta \\
& m_h \leq \ n_h \leq M_h \quad \forall h = 1, 2, \ldots, 5 \\
& l \geq 0 \\
& p_h \geq 0 \qquad \forall h = 1, 2, \ldots, 5
\end{aligned}
\tag{7.2}
$$

Here $\Gamma_1 = 0, \ldots, 5$ defines the number of uncertain variance parameters.
The solutions of above robust optimization problem are provided in Table 7.2:

Table 7.2: Robust allocations with different values of Γ_1

Gammas	Stratum 1	Stratum 2	Stratum 3	Stratum 4	Stratum 5
$\Gamma_1 = 0$	877.1488	54.16350	924.1846	178.2392	406.7939
$\Gamma_1 = 1$	860.6172	53.20779	951.2789	175.1295	400.2966
$\Gamma_1 = 2$	863.5282	52.29401	959.1640	172.1219	393.4218
$\Gamma_1 = 3$	868.3603	51.43075	964.5310	169.2805	386.9274
$\Gamma_1 = 4$	829.0089	114.83812	920.8217	183.2365	392.6248
$\Gamma_1 = 5$	877.2793	49.83732	974.4379	164.0359	374.9396

Table 7.2 provides the allocation for each stratum and for different uncertainty levels. We can see in this table that the uncertainty level does not affect much the allocations in the strata. For each value of Γ_1, stratum 3 has the largest sample size allocated whereas stratum 2 has the smallest sample size allocated. The reason is that stratum 3 has the largest stratum specific variance and stratum 2 has the smallest. In Figure 7.7 we can see the same results displayed as boxplots for each stratum. In each stratum we have 5 allocations for all values of $\Gamma_1 = 0, \ldots, 5$.

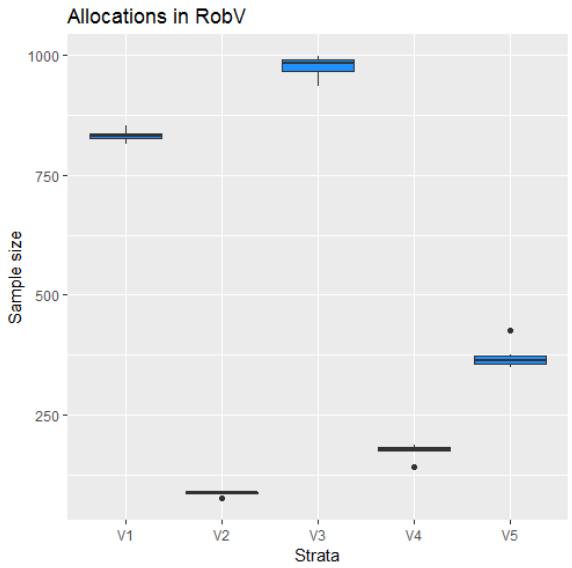


Figure 7.7: Boxplots of allocations in RobV in different strata

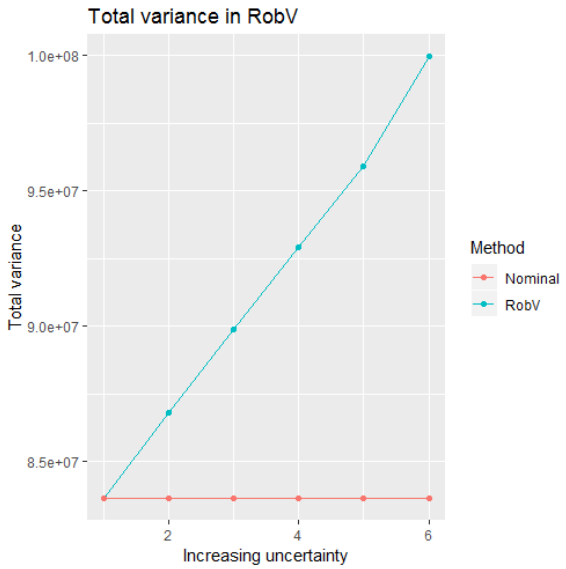


Figure 7.8: Total variances in the nominal problem and RobV for $\Gamma_1 = 0, \ldots, 5$

We wanted to see how the total variances of RobV and of the nominal problem differ. So we solved the nominal problem (7.1) and calculated its total variance. In Figure 7.8 we can see that at $\Gamma_1 = 0$, the total variance in RobV and in the nominal problem are equal. The total variance of RobV increases with the increase in Γ_1 and is highest at $\Gamma_1 = 5$.

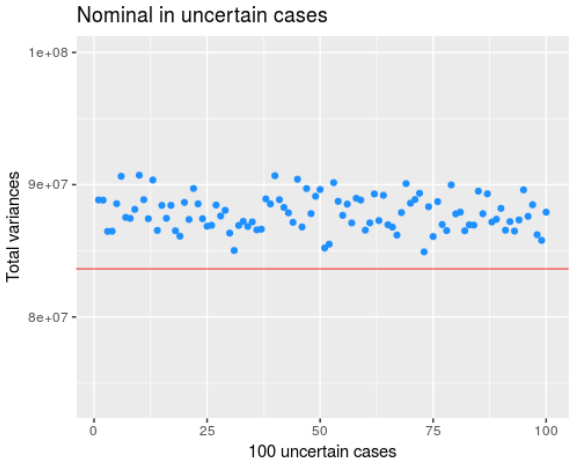


Figure 7.9: Total variance in the nominal problem and effect of uncertainty. The red line displays the total variance of the nominal problem.

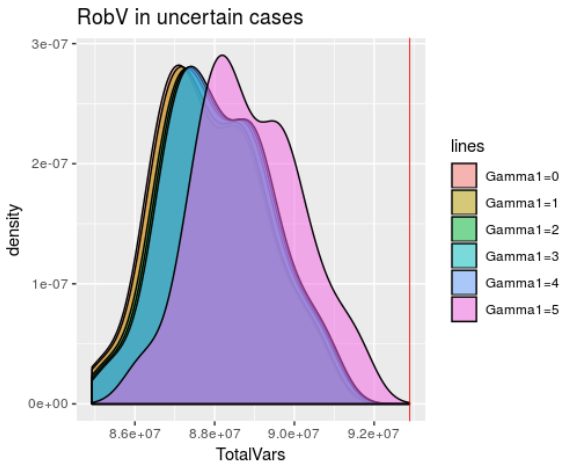


Figure 7.10: Total variance in RobV and effect of uncertainty. The red line displays the total variance of RobV at $\Gamma_1 = 3$

It is interesting to see how, when uncertainty is introduced, total variances of the nominal problem and RobV change. So we performed an experiment here and took 100 sets of random stratum specific variances within the uncertainty level. For each of these random stratum specific variances, we calculated 100 total variances using the nominal allocation and 100 total variances for each $\Gamma_1 = 0, \ldots, 5$ using the allocations of RobV. We can see in Figure 7.9 that the total variances calculated using random stratum specific variances are larger than the total variance of the nominal problem in all the 100 cases. In Figure 7.10 we generated the densityplots of 100 total variances for each Γ_1. We have shown here that the 100 total variances of RobV for each Γ_1 are smaller than the total variance of RobV at $\Gamma_1 = 3$. We can conclude that we do not need to consider always that all of the parameters are uncertain.

7.2 Inclusion of Cost and Robust Allocations

In real life sampling allocation problems the cost of carrying out a survey plays an important role. Sometimes it also happens that the selected person is not available for interview and the interviewer has to revisit the person in order to complete the survey process. This results in an increased cost and shows that the cost is very uncertain in nature.

So here we assume that the cost C_h of selecting a unit sample in stratum h is as given in Table 7.3. We also assume that there is 20% of uncertainty in the costs, i.e., the uncertainty interval for the cost C_h is

$$[C_h - \hat{C}_h, C_h + \hat{C}_h] \text{ with } \hat{C}_h = 0.2C_h.$$

The resulting assumed data is provided in the following Table 7.3:

Table 7.3: Assumed data for the sample allocation

Strata (h)	1	2	3	4	5
C_h	100	120	140	160	180
$\hat{C}_h$	20	24	28	32	36

We now include a cost constraint in the robust sampling allocation problems RobC, RobV and RobCV. The upper bound on the cost constraint

$$\sum_{h=1}^{5} C_h n_h \leq C$$

is assumed to be $C = 157629$.

Now the sampling allocation problem can be divided into the three cases discussed in Section 4.2: First, when only cost is uncertain (RobC). Second, when only variance is uncertain (RobV) and third, when both cost and variance are uncertain (RobCV). Before discussing these cases, we would like to recall that the robust formulations of these cases are similar to (5.3), (5.4) and (5.5) from Section 5.2. We use the data given in Tables 7.1 and 7.3 for solving all the problems in this section.

7.2.1 When only cost is uncertain (RobC)

This scenario is based on the situation when only the cost in the sampling allocation problem is uncertain. We solved the problem RobC given in (5.3) 6 times for different values of $\Gamma_0 = 0, \ldots, 5$ using **NLopt** package of R software. Solving these 6 problems took 22 minutes and 37 seconds. We also computed allocations of the nominal problem given in (5.2) in which we minimize the variance function subject to the cost and other constraints. We compare the total cost of the nominal problem and the total costs of RobC in the following Figure 7.11.

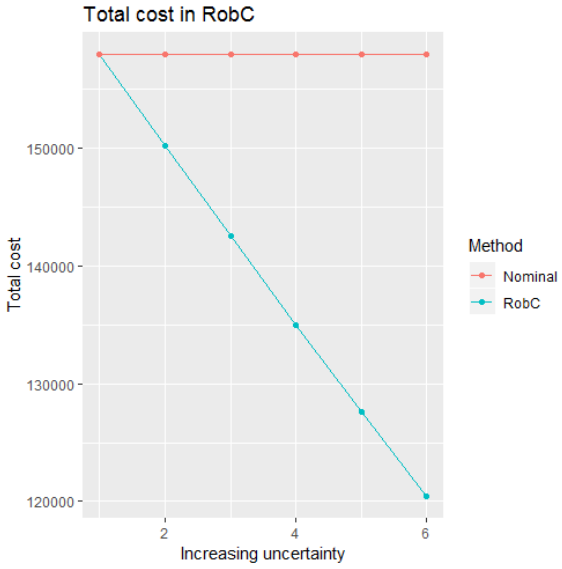


Figure 7.11: Total costs in the nominal problem and RobC for $\Gamma_0 = 0, \ldots, 5$

We can see here that the total cost in RobC is continuously decreasing with the increase in the uncertainty level of cost parameters $\Gamma_0 = 0, \ldots, 5$. We can see that at $\Gamma_0 = 0$ the total cost is equal to the total cost of the nominal problem which is not always the case. The total cost in a sampling allocation problem directly depends on the sample sizes allocated to the strata. Now we look at the allocations of RobC in Table 7.4

Table 7.4: Sample sizes n_h for stratum h in RobC with different values of Γ_0

Gammas	n_1	n_2	n_3	n_4	n_5	$\sum_{h=1}^{5} n_h$
$\Gamma_0 = 0$	389.2061	71.08409	278.0182	203.98859	216.2467	1158.54
$\Gamma_0 = 1$	387.8898	58.47588	277.0779	167.44065	215.5154	1106.39
$\Gamma_0 = 2$	385.9911	46.24693	275.7218	132.17464	214.4606	1054.59
$\Gamma_0 = 3$	382.5667	35.05459	273.2758	100.03876	212.5581	1003.49
$\Gamma_0 = 4$	379.4524	30.72769	271.0513	87.60385	188.8454	957.68
$\Gamma_0 = 5$	374.9896	27.01899	267.8636	77.02346	166.0117	912.90

We can see in the Table 7.4 that the total sample size is decreasing with the increase in the uncertainty level Γ_0 of cost parameters. This decrease in total sample size is reflected in the decrease of the total cost of RobC which we saw above. Boxplots of the allocations in RobC are given in the Figure 7.12. We can see that for all uncertainty levels $\Gamma_0 = 0, \ldots, 5$, stratum 1 has the highest allocation and a possible reason is that stratum 1 has the biggest population size. Stratum 2 has the smallest population size and also the smallest allocation.

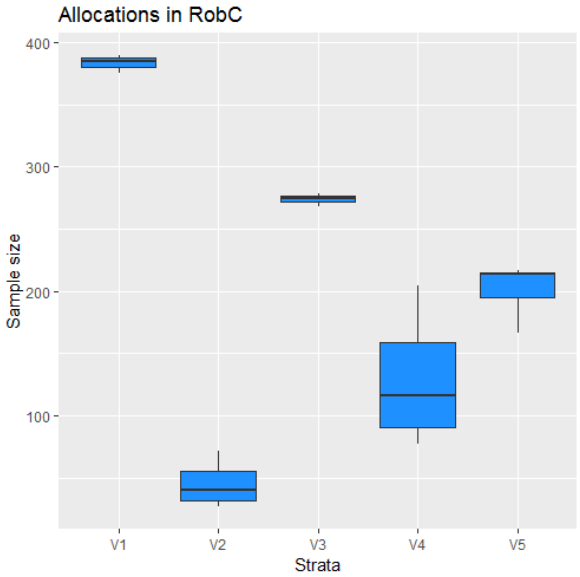


Figure 7.12: Boxplots of allocations in RobC for each stratum

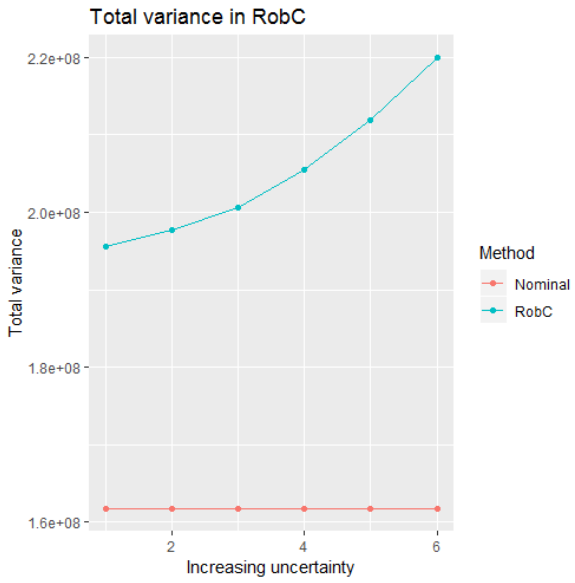


Figure 7.13: Total variance in RobC for $\Gamma_0 = 0, \ldots, 5$

We can see in Figure 7.13 that the total variance for RobC is continuously increasing with the increase in uncertainty in the cost parameters Γ_0. We saw earlier that with the increase in uncertainty we have decreasing total sample sizes. However, a smaller sample size leads to a bigger total variance.

7.2.2 When only variance is uncertain (RobV)

The second scenario is based on the case when only the variance is considered to be uncertain. We solved the prolem RobV 6 times for $\Gamma_1 = 0, \ldots, 5$ using the **NLopt** package of R software. Solving these 6 problems took 1 minute and 32 seconds. The allocations in RobV are given in Table 7.5 and Figure 7.14.

Table 7.5: Sample sizes n_h for stratum h in RobV with different values of Γ_1

Gammas	n_1	n_2	n_3	n_4	n_5	$\sum_{h=1}^{5} n_h$
$\Gamma = 0$	481.1024	47.03223	446.4876	83.12790	157.7009	1215.45
$\Gamma = 1$	473.0996	46.12478	458.7384	81.52297	154.6501	1214.13
$\Gamma = 2$	463.3346	45.32172	471.4763	80.10446	151.9641	1212.20
$\Gamma = 3$	458.2498	44.57075	480.4101	78.78336	149.5154	1211.52
$\Gamma = 4$	449.6980	43.84004	491.8766	77.48480	146.9895	1209.88
$\Gamma = 5$	442.9153	43.17934	501.4704	76.31706	144.7743	1208.65

We can see that there is not much difference in the total sample size for different uncer-

tainty levels $\Gamma_1 = 0, \ldots, 5$. However, the allocations in each stratum are changing with the increase in Γ_1. We can see in Figure 7.14 that stratum 3 has the largest allocation. This is consistent with the fact that stratum 3 has the largest stratum specific variance (see Table 7.1). In RobV we minimize the total variance function and this is inversely proportional to the sample sizes. We can conclude that the bigger the sample size is the smaller the total variance is.

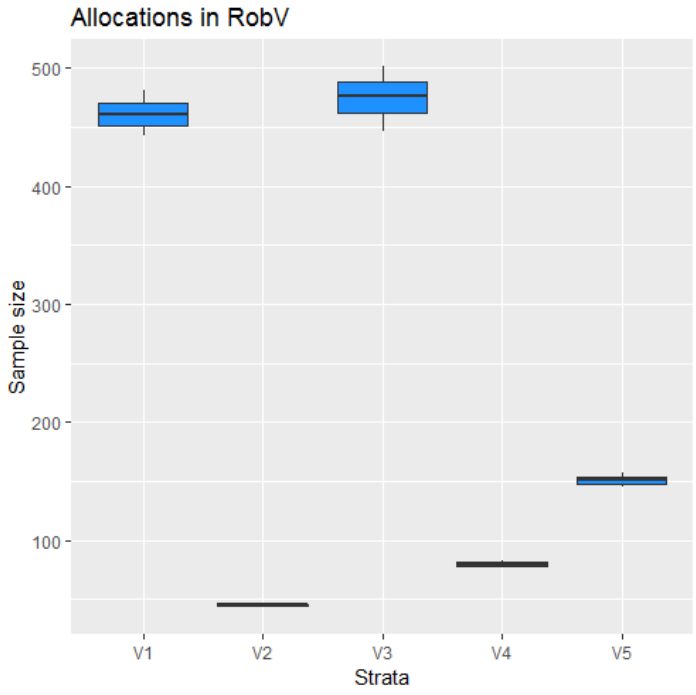


Figure 7.14: Boxplots of allocations in RobV for each stratum

We can see in Figure 7.15 that there is no difference in the total costs of RobV for different uncertainty levels in the variance parameter $\Gamma_1 = 0, \ldots, 5$. The total cost of RobV is equal to the total cost of the nominal problem for all values of Γ_1 which is as expected. From Table 7.5, we can see that the total sample size in RobV for different values of Γ_1 is not much increasing or decreasing and since the total cost depends on the sample size, this total cost is also not increasing or decreasing.

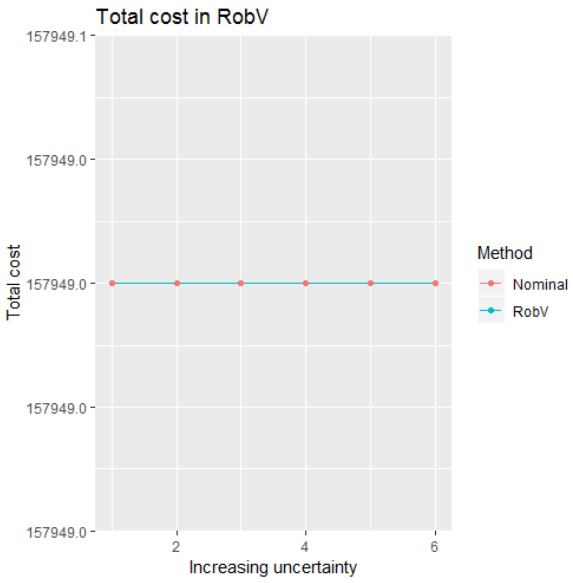


Figure 7.15: Comparison of total costs in the nominal problem and RobV for $\Gamma_1 = 0, \ldots, 5$

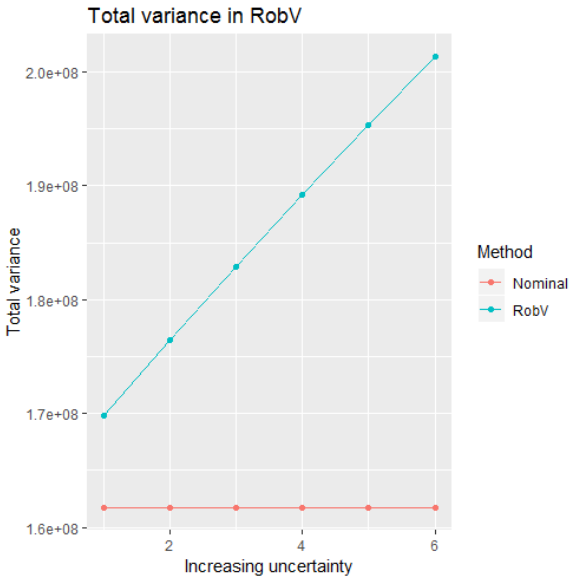


Figure 7.16: Comparison of total variances in the nominal problem and RobV for $\Gamma_1 = 0, \ldots, 5$

Figure 7.16 shows the total variances of the nominal problem and RobV for different uncertainty levels $\Gamma_1 = 0, \ldots, 5$. For $\Gamma_1 = 0$, the total variance of RobV is bigger than the total variance of the nominal problem. An increase in the uncertainty level Γ_1 results in an increased total variance of RobV. This increase in total variance can be understood as the cost of robustness.

7.2.3 When both cost and variance are uncertain (RobCV)

This is the realistic case because usually both costs and variances are uncertain parameters in the sampling allocation problem. Neverthless, RobC and RobV can be very useful if one of the parameters is known for sure. We recall again that due to the artificial variables introduced in the robustification process, RobCV with $\Gamma_0 = 0$ is not the same as RobV, and RobCV with $\Gamma_1 = 0$ is not the same as RobC.

We solve the optimization problem RobCV given in (5.5) 36 times for all combinations of $\Gamma_0 = 0, \ldots, 5$ and $\Gamma_1 = 0, \ldots, 5$ using **NLopt** package of R software. Solving these 36 problems took 4 hours 24 minutes 37 seconds. The allocations in RobCV for each stratum are given in Figure 7.17:

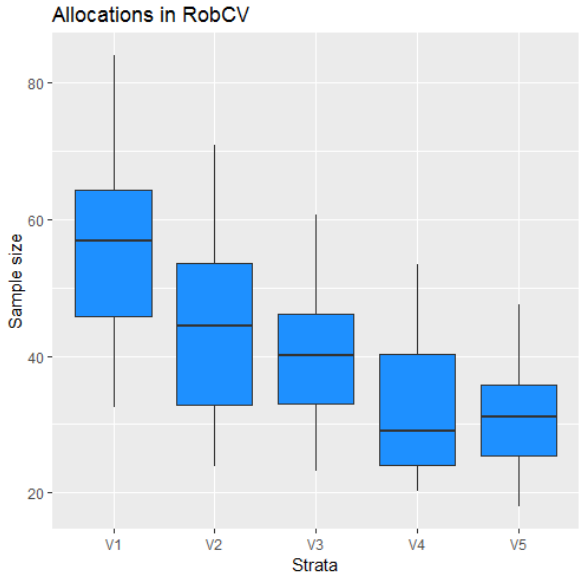


Figure 7.17: Boxplot of allocations in RobCV for all combinations of Γ_0 and Γ_1

We can see in Fig 7.17 that the sample sizes in stratum 1 are bigger than the other sample sizes. The reason is that stratum 1 has a large stratum specific variance and the minimal cost of selecting a unit sample.

We can see in Figure 7.18 that the total costs are not increasing or decreasing with the increase in the uncertainty level Γ_0. We can also see that the total cost is much smaller than

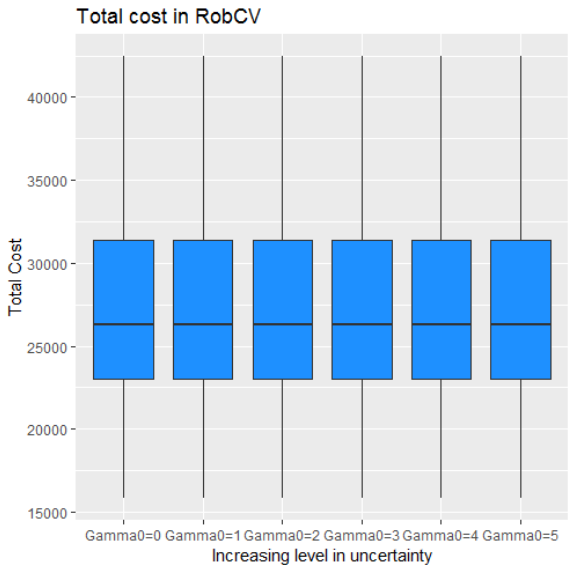


Figure 7.18: Total cost in RobCV for increasing Γ_0

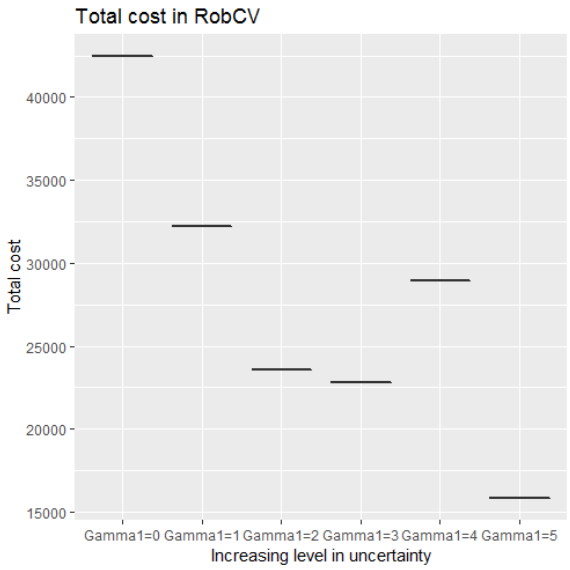


Figure 7.19: Total cost in RobCV for increasing Γ_1

the upper bound $C = 157629$ in the cost constraint. In Figure 7.19, we can see that for $\Gamma_1 = 0, \ldots, 5$ the total costs is changing.

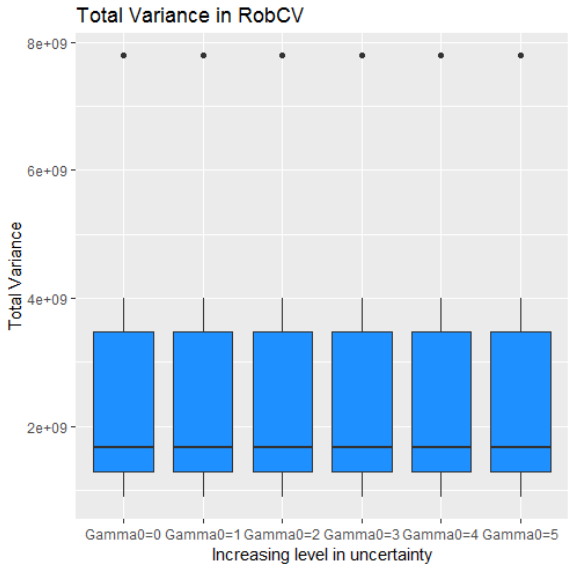


Figure 7.20: Total variance in RobCV for increasing Γ_0

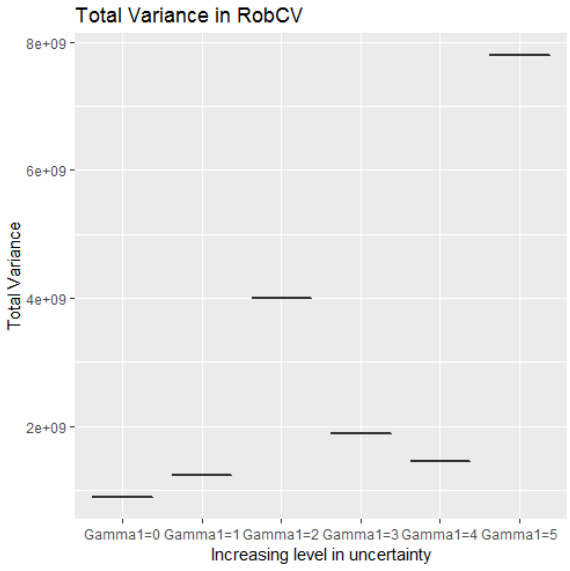


Figure 7.21: Total variance in RobCV for increasing Γ_1

We can see in Figures 7.20 and 7.21 that the total variance is not affected by the increasing

level of Γ_0 and remains constant whereas the total variance is directly affected by the increasing level of Γ_1, which is as expected. The total variance of RobCV is expected to increase with increasing Γ_1 however, we can see small jumps in the total variance at $\Gamma_1 = 3, 4$. The reason for this is probably that the strata are heterogeneous and at some places its effect was expected.

7.2.4 Feasibility Analysis

Now we want to see the effect of uncertainty on the nominal and the robust allocations. We check whether the optimal solutions of the nominal problem, RobC and RobCV satisfy the cost constraint by taking 100 random values between C_h and $C_h + \hat{C}_h$ as defined in Table 7.3. The densityplots of the total costs obtained by using these 100 random parameters are as follows:

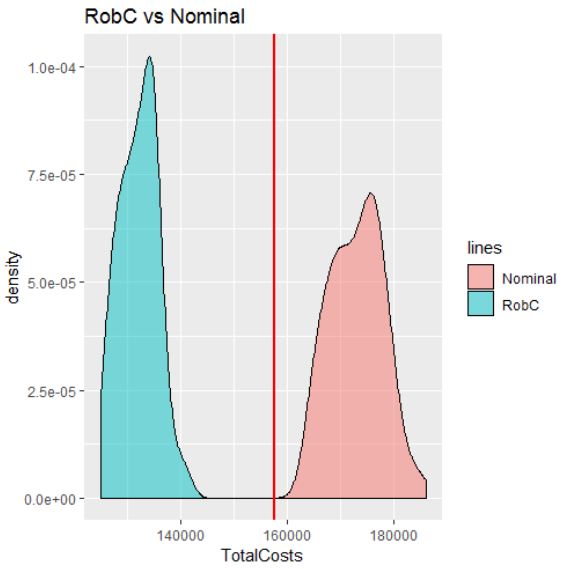


Figure 7.22: Total costs in the nominal problem and RobC with $\Gamma_0 = 5$. The red line displays the upper bound $C = 157629$

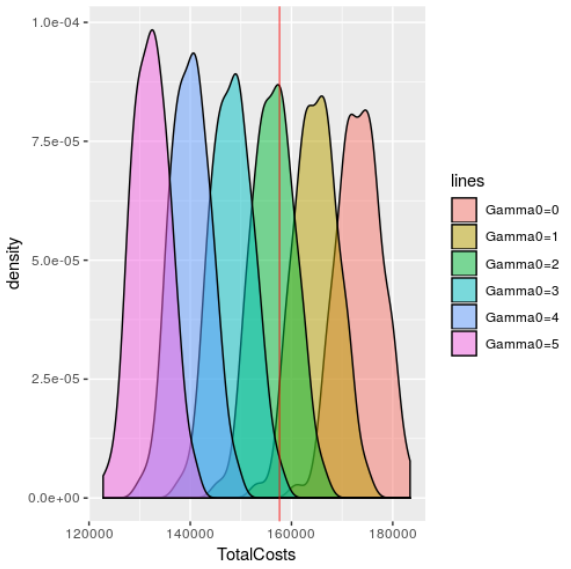


Figure 7.23: Total costs in the RobC for increasing Γ_0. The red line displays the upper bound $C = 157629$

In Figure 7.22 we can see a comparison of the total costs in the nominal problem and in RobC. We see that the total costs of the nominal problem are always bigger than the upper bound in the cost constraint and hence the cost constraint is always violated. In Figure 7.23 we can see different levels of Γ_0 and the corresponding densityplots of the total costs in RobC. It is very interesting to see that at $\Gamma_0 = 0$ the densityplot is very similar to the densityplot of the nominal problem, however, this is not always the case. We see that at $\Gamma_0 = 1$ we have a few of the total costs which are smaller than the upper bound. For $\Gamma_0 = 2$ we have half of the total costs bigger than the upper bound and at $\Gamma_0 = 3$ only few of the total costs are bigger than the upper bound. For $\Gamma_0 = 4, 5$ all of the 100 total costs are smaller than the upper bound in the cost constraint.

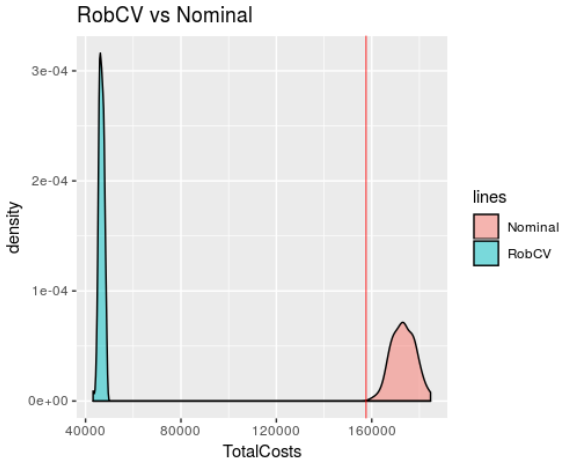


Figure 7.24: Total costs in the nominal problem and in RobCV with $\Gamma_0 = \Gamma_1 = 5$

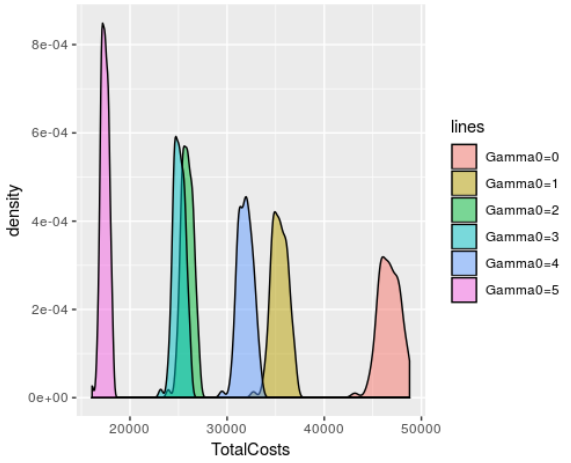


Figure 7.25: Total costs in RobCV with $\Gamma_1 = 5$ and increasing $\Gamma_0 = 0, \ldots, 5$

In Figure 7.24 we see a comparison between the total costs of the nominal problem and the total costs of RobCV. The total costs of the nominal problem are again always bigger than the upper bound in the cost constraint whereas the total costs of RobCV is always smaller than the upper bound. In Figure 7.25 we can see the densityplots of the total costs of RobCV with $\Gamma_1 = 5$ and different values of $\Gamma_0 = 0, \ldots, 5$. Due to the very small sample sizes in RobCV all of the costs are much smaller than the upper bound in the cost constraint.

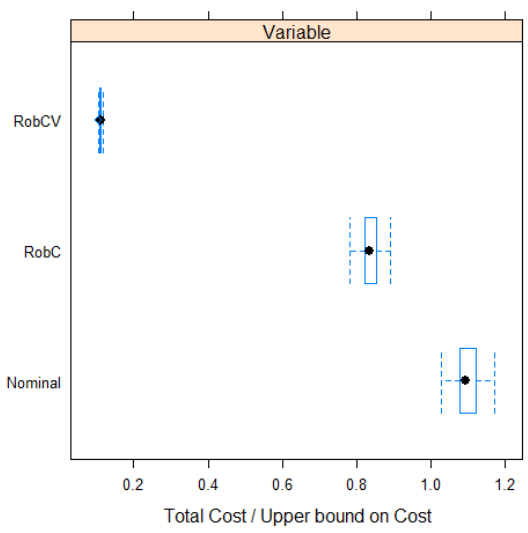


Figure 7.26: Boxplots of scaled total costs in the nominal, RobC and RobCV

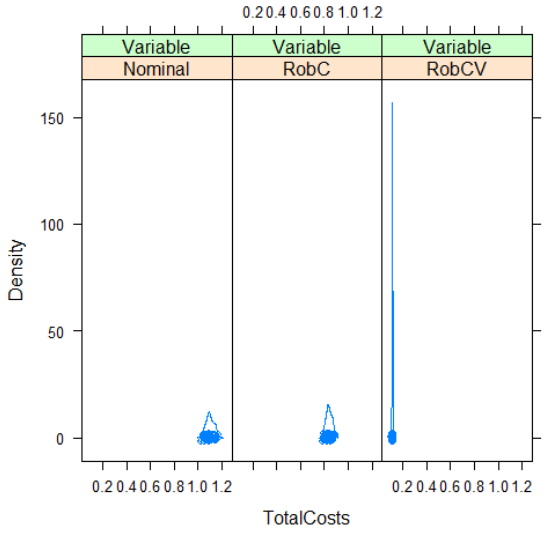


Figure 7.27: Densityplot of scaled total costs in the nominal, RobC and RobCV

In Figures 7.26 and 7.27 we divide the total costs of the nominal problem, RobC and RobCV by the upper bound in the cost constraint. Hence in both the boxplots and densityplots all the values bigger than 1 represent infeasible cases and values smaller than 1 represent the feasible cases. Also, in these comparisons we considered the value of $\Gamma_0 = 5$ in RobC which is why we do not have any infeasible cases. For the total costs of RobCV we consider $\Gamma_0 = \Gamma_1 = 5$.

Chapter 8

Conclusion and Outlook

In this thesis, we dealt with the sampling allocation problem with uncertain stratum specific variances and uncertain costs. We proposed three robust formulations for optimal sampling allocations: if the only cost is uncertain (RobC), if the only stratum specific variances are uncertain (RobV) and if both stratum specific variances and cost are uncertain (RobCV). To the best of our knowledge, this is the first time robust allocations are proposed for sampling allocation problems. We proved that the upper bound on the probability of nonlinear constraint violation can be calculated. It is Interesting that the calculated upper bound does not depend on the robust solutions, so we know in advance that how good the robust solutions are. In order to check the stability and feasibility of the robust allocations, we performed several experiments. We considered three different datasets: First, we considered a simulated data where the strata are diversely distributed. Second, we consider the synthetically generated heterogeneous dataset AMELIA with around 3.7 million observations. Third, we considered a real life and very complex dataset: an income and taxation dataset from the German state of NRW for the year 2001.

We found that the robust allocations for a population generated through simulation are very stable. We noticed that in such cases the total cost decreases monotonically with the increase in uncertainty level. However, for a heterogeneous population, the robust solutions are not very stable and change with the change in uncertainty level. A feasibility test for the robust allocations of simulated data was also carried out and proved that the robust solutions are feasible no matter what values the uncertain parameter takes in the defined uncertainty interval. The computation time can be very large so we merged some of the strata and made more heterogeneous strata. Robust allocations in a smaller number of strata are easy to compute but very complex statistically. We also did an experiment to see how robust solutions are affected if the outliers of heterogeneous populations are deleted. We saw that in a very heterogeneous dataset it is hard to control the total costs and total variance but we have feasibility in all the cases when uncertainty exists.

The main applications of these robust allocations are in the future censuses. However, generally, the censuses have much bigger number of strata, for example in the German census 2011 the total number of strata was around 19,144. For such a large problem, the computation time of robust allocations might be a problem. The integrality constraint should also be

added in the robust formulations. In this case, we will have a mixed integer nonlinear robust formulation of the sampling allocation problem. Hence, an algorithmic development is also needed for the robust formulations of the sampling allocation problem.

Bibliography

Alberto Alesina and Roberto Perotti. Fiscal adjustments in oecd countries: composition and macroeconomic effects. *Staff Papers*, 44(2):210–248, 1997.

Andreas Alfons, Peter Filzmoser, Beat Hulliger, Jan-Philipp Kolb, Stefan Kraft, Ralf Münnich, and Matthias Templ. Synthetic data generation of silc data. *AMELI Research Project Report WP6 - D6.2.*, 2011.

Manuel Amunategui. Data exploration and machine learning, hands-on: Mapping the united states census with ggmap, 2014. URL `http://amunategui.github.io/ggmap-example/`. Accessed on 11-04-2019.

Aharon Ben-Tal and Arkadi Nemirovski. Robust solutions of uncertain linear programs. *Oper. Res. Lett.*, 25(1):1–13, 1999. ISSN 0167-6377. doi: 10.1016/S0167-6377(99)00016-4. URL `https://doi.org/10.1016/S0167-6377(99)00016-4`.

Aharon Ben-Tal and Arkadi Nemirovski. Robust solutions of linear programming problems contaminated with uncertain data. *Math. Program.*, 88(3, Ser. A):411–424, 2000. ISSN 0025-5610. doi: 10.1007/PL00011380. URL `https://doi.org/10.1007/PL00011380`.

Dimitris Bertsimas and Melvyn Sim. The price of robustness. *Oper. Res.*, 52(1):35–53, 2004. ISSN 0030-364X. doi: 10.1287/opre.1030.0065. URL `https://doi.org/10.1287/opre.1030.0065`.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Jan Pablo Burgard, Jan-Philipp Kolb, Hariolf Merkle, and Ralf Münnich. Synthetic data for open and reproducible methodological research in social sciences and official statistics. *AStA Wirtschafts- und Sozialstatistisches Archiv*, 11(3):233–244, Dec 2017. doi: 10.1007/s11943-017-0214-8. URL `https://doi.org/10.1007/s11943-017-0214-8`.

RL Chaddha, WW Hardgrave, DJ Hudson, M Segal, and JW Suurballe. Allocation of total sample size when only the stratum means are of interest. *Technometrics*, 13(4):817–831, 1971.

Samprit Chatterjee. A note on optimum allocation. *Skand. Aktuarietidskr.*, 1967:40–44, 1967. doi: 10.1080/03461238.1967.10406206. URL https://doi.org/10.1080/03461238.1967.10406206.

William G. Cochran. *Sampling techniques*. John Wiley & Sons, New York-London-Sydney, third edition, 1977. Wiley Series in Probability and Mathematical Statistics.

Tore Dalenius. The multi-variate sampling problem. *Skand. Aktuarietidskr.*, 36:92–102, 1953. doi: 10.1080/03461238.1953.10419460. URL https://doi.org/10.1080/03461238.1953.10419460.

José A Díaz-García and Liliana Ulloa Cortez. Multi-objective optimisation for optimum allocation in multivariate stratified sampling. *Survey Methodology*, 34(2):215–222, 2008.

José A. Díaz-García and Ma. Magdalena Garay-Tápia. Optimum allocation in stratified surveys: stochastic programming. *Comput. Statist. Data Anal.*, 51(6):3016–3026, 2007. ISSN 0167-9473. doi: 10.1016/j.csda.2006.01.016. URL https://doi.org/10.1016/j.csda.2006.01.016.

Jose A Diaz-Garcia and Rogelio Ramos-Quiroga. Multivariate stratified sampling by stochastic multiobjective optimisation. *arXiv preprint arXiv:1106.0773*, 2011.

Matthias Ehrgott. *Multicriteria optimization*. Springer-Verlag, Berlin, second edition, 2005. ISBN 3-540-21398-8.

Laurent El Ghaoui and Hervé Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM J. Matrix Anal. Appl.*, 18(4):1035–1064, 1997. ISSN 0895-4798. doi: 10.1137/S0895479896298130. URL https://doi.org/10.1137/S0895479896298130.

Forschungsdatenzentrum. Lohn- und einkommensteuerstatistik 2001, 2001. URL https://www.forschungsdatenzentrum.de. Accessed on 05-10-2018.

Ulf Friedrich, Ralf Münnich, Sven de Vries, and Matthias Wagner. Fast integer-valued algorithms for optimal allocations under constraints in stratified sampling. *Comput. Statist. Data Anal.*, 92:1–12, 2015. ISSN 0167-9473. doi: 10.1016/j.csda.2015.06.003. URL https://doi.org/10.1016/j.csda.2015.06.003.

Ulf Friedrich, Ralf Münnich, and Martin Rupp. Multivariate optimal allocation with box-constraints. *Austrian Journal of Statistics*, 47(2):33–52, 2018.

Siegfried Gabler, Matthias Ganninger, and Ralf Münnich. Optimal allocation of the sample size to strata under box constraints. *Metrika*, 75(2):151–161, 2012. ISSN 0026-1335. doi: 10.1007/s00184-010-0319-3. URL https://doi.org/10.1007/s00184-010-0319-3.

GESIS. Gml:microcensus, 2019. URL https://www.gesis.org/en/gml/microcensus/. Accessed on 18-06-2019.

Neha Gupta, Irfan Ali, and Abdul Bari. An optimal chance constraint multivariate stratified sampling design using auxiliary information. *J. Math. Model. Algorithms Oper. Res.*, 13 (3):341–352, 2014. ISSN 2214-2487. doi: 10.1007/s10852-013-9237-5. URL `https://doi.org/10.1007/s10852-013-9237-5`.

Morris H. Hansen, William G. Madow, and Benjamin J. Tepping. An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78(384):776–793, 1983.

Christiane Hoffmann, J-Matthias Graf von der Schulenburg, et al. The influence of economic evaluation studies on decision making.: A european survey. *Health policy*, 52(3):179–192, 2000.

Alfons Hollederer. Unemployment and health in the german population: results from a 2005 microcensus. *Journal of Public Health*, 19(3):257–268, 2011.

Raymond J. Jessen. Statistical investigation of a sample survey for obtaining farm facts. *Research Bulletin (Iowa Agriculture and Home Economics Experiment Station)*, 26(304):1, 1942.

MGM Khan, Mohammad J. Ahsan, and Nujhat Jahan. Compromise allocation in multivariate stratified sampling: An integer solution. *Naval Research Logistics (NRL)*, 44(1):69–79, 1997.

A. R. Kokan and Sanaullah Khan. Optimum allocation in multivariate surveys: An analytical solution. *J. Roy. Statist. Soc. Ser. B*, 29:115–125, 1967. ISSN 0035-9246. URL `http://links.jstor.org/sici?sici=0035-9246(1967)29:1<115:OAIMSA>2.0.CO;2-L&origin=MSN`.

William Kruskal and Frederick Mosteller. Representative sampling, iii: The current statistical literature. *International Statistical Review/Revue Internationale de Statistique*, pages 245–265, 1979.

Kelly LeRoux and Nathaniel S Wright. Does performance measurement improve strategic decision making? findings from a national survey of nonprofit social service agencies. *Nonprofit and Voluntary Sector Quarterly*, 39(4):571–587, 2010.

Sharon L. Lohr. *Sampling: design and analysis*. Brooks/Cole, Cengage Learning, Boston, MA, second edition, 2010. ISBN 978-0-495-10527-5; 0-495-10527-9.

Ashwin Machanavajjhala, Daniel Kifer, John Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: Theory meets practice on the map. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, pages 277–286. IEEE Computer Society, 2008.

Ralf Münnich and J. Schürle. On the simulation of complex universes in the case of applying the german microcensus. *DACSEIS research paper series, No.4.*, 2003.

Ralf T. Münnich, Ekkehard W. Sachs, and Matthias Wagner. Numerical solution of optimal allocation problems in stratified sampling under box constraints. *AStA Adv. Stat. Anal.*, 96(3):435–450, 2012. ISSN 1863-8171. doi: 10.1007/s10182-011-0176-z. URL https://doi.org/10.1007/s10182-011-0176-z.

Jerzy Neyman. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Estadística*, 17:587–651, 1959. ISSN 0014-1135.

Frank Olken and Doron Rotem. Simple random sampling from relational databases, 1986. URL https://escholarship.org/uc/item/9704f3dr. Accessed on 20-06-2019.

S. A. Y. Omule. Optimum design in multivariate stratified sampling. *Biometrical J.*, 27(8): 907–912, 1985. ISSN 0323-3847. doi: 10.1002/bimj.4710270813. URL https://doi.org/10.1002/bimj.4710270813.

Thomas Philippi. Adaptive cluster sampling for estimation of abundances within local populations of low-abundance plants. *Ecology*, 86(5):1091–1100, 2005.

Pinterest. Discover ideas about statistics, 2019. URL https://www.pinterest.com/pin/357754764120670109/. Accessed on 15-04-2019.

University of Michigan Population studies centre. Zip code characteristics: Mean and median household income, 2010. URL https://www.psc.isr.umich.edu/dis/census/Features/tract2zip/. Accessed on 11-04-2019.

Carl E. Särndal, Bengt Swensson, and Jan Wretman. *Model assisted survey sampling*. Springer-Verlag New York, 1992.

Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model assisted survey sampling*. Springer Science & Business Media, 2003.

Hans T. Schreuder and James Alegria. *Stratification and plot selection rules: Misuses and consequences*, volume 536. USDA Forest Service, Rocky Mountain Forest and Range Experiment Station, 1995.

Hans T. Schreuder, Timothy G. Gregoire, and Johann P. Weyer. For what applications can probability and non-probability sampling be used? *Environmental Monitoring and Assessment*, 66(3):281–291, 2001.

Norbert Schwarz. The german microcensus. *Schmollers Jahrbuch*, 121(649):654, 2001.

Allen L. Soyster. Technical note - convex programming with set-inclusive constraints and applications to inexact linear programming. *Operations Research*, 21(5):1154–1157, 1973. doi: 10.1287/opre.21.5.1154. URL https://doi.org/10.1287/opre.21.5.1154.

Pandurang V. Sukhatme. *Sampling theory of surveys with applications*. The Indian Society of Agricultural Statistics, New Delhi, India; The Iowa State College Press, Ames, Iowa, 1954.

Richard A Swanson and Elwood F Holton. *Research in organizations: Foundations and methods in inquiry*. Berrett-Koehler Publishers, 2005.

Charles Teddlie and Fen Yu. Mixed methods sampling: A typology with examples. *Journal of mixed methods research*, 1(1):77–100, 2007.

Aleksandr A. Tschuprow. On the mathematical expectation of the moments of frequency distributions in the case of correlated observations (chapters 4-6). *Metron*, 2:646–683, 1923.

Shafi Ullah, Irfan Ali, and Abdul Bari. Fuzzy geometric programming approach in multivariate stratified sample surveys under two stage randomized response model. *J. Math. Model. Algorithms Oper. Res.*, 14(4):407–424, 2015. ISSN 2214-2487. doi: 10.1007/s10852-015-9276-1. URL `https://doi.org/10.1007/s10852-015-9276-1`.

UNECE. Terminology on statistical metadata, conference of european statisticians statistical standards and studies, 2000. URL `https://ec.europa.eu/eurostat/ramon/coded_files/UNECE_TERMINOLOGY_STAT_METADATA_2000_EN.pdf`. Accessed on 26-12-2018.

David S Yeager, Jon A Krosnick, LinChiat Chang, Harold S Javitz, Matthew S Levendusky, Alberto Simpser, and Rui Wang. Comparing the accuracy of rdd telephone surveys and internet surveys conducted with probability and non-probability samples. *Public opinion quarterly*, 75(4):709–747, 2011.