Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Bergitte Viste

The Social and Spiritual Situation in Lilleby

A Statistical Simulation Study with a Questionnaire Survey

Master's thesis in Applied Physics and Mathematics
Supervisor: Henning Omre
July 2020

NTNU
Norwegian University of Science and Technology

Bergitte Viste

The Social and Spiritual Situation in Lilleby

A Statistical Simulation Study with a Questionnaire Survey

Master's thesis in Applied Physics and Mathematics
Supervisor: Henning Omre
July 2020

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Soli Deo Gloria

Summary

Oslo Monitor 1.0 was released in January 2018 by The Think Tank Skaperkraft. The report accounts for the spiritual situation, social suffering and cultural challenges in Oslo. The data presented in the report are given a probability distribution with corresponding parameter estimates. The specified marginal distributions only provide insight concerning the individual nature of the variables. The goal is to include the marginal distributions in an interaction model to account for the interplay among the variables as well. The concept of copulas is introduced to derive the interaction model. From the interaction model a sequential simulation algorithm is developed for categorical variables with either a binomial or multinomial distribution. The algorithm generates a realization of the population in a city called Lilleby. Biplots visualize the dependence assumed between the variables included in the interaction model. The realized population of Lilleby reflects both the marginal distributions from Oslo Monitor 1.0 and the dependence assumed to exist.

The population of Lilleby participates in a statistical survey with questionnaires. The questionnaire is distributed to a representative and stratified sample of Lilleby residents. Data collection deals with two major types of correction: Stratification and bias correction. Stratification is enforced when the questionnaires are distributed. But some gender, age groups or districts might be over- or underrepresented in the responses and must be weighted to restore the stratification. The weights are set by solving the prevailing minimization problem by Lagrange multipliers. A likelihood model expresses the psychological aspects of answering a questionnaire, such as potential prejudices. We apply a posterior model to the responses to correct for bias from potential prejudices. The response sample is evaluated by its sensitivity to the stratification and bias correction by the comparison of proportion estimates. Bias correction has major impact on the centering of the proportion estimates. The centering can be further improved by stratification but on the expense of somewhat larger spread. The bias corrected proportion estimate compared to the stratified and bias corrected proportion estimate by their RMSE calls them even. Still, the stratified and bias corrected proportion estimate is centered closest to the true Lilleby proportion compared to the bias corrected proportion estimate. The stratification model and, especially, the bias correction model appear as effective tools to correct for skewness in a response sample and to deal with the bias caused by potential prejudices in a statistical survey including the subjectivity and unpredictable behaviour of humans.

Sammendrag

Oslo Monitor 1.0 ble publisert i januar 2018 av Tankesmien Skaperkraft. Rapporten beskriver den åndelige situasjonen, den sosiale smerten og de kulturelle utfordringene i Oslo. Data presentert i rapporten tildeles en sannsynlighetsfordeling med tilhørende parameterestimater. De definerte marginalfordelingene gir innsikt i den individuelle naturen til variablene. Vi ønsker å inkludere marginalfordelingene i en interaksjonsmodell slik at også samspillet mellom variablene kan beskrives. Interaksjonsmodellen utledes ved hjelp av copula-konseptet. Vi bruker en sekvensiell simuleringsalgoritme laget for kategoriske variabler med en binomisk eller multinomisk fordeling til å simulere en realisert befolkning av Lilleby. Biplot visualiserer den antatte avhengigheten mellom de inkluderte variablene i interaksjonsmodellen. Den realiserte Lilleby reflekterer marginalfordelingene fra Oslo Monitor 1.0 i tillegg til den antatte avhengigheten.

Innbyggerne i Lilleby deltar i en statistisk undersøkelse ved å svare på et spørreskjema. Spørreskjemaet sendes ut til et representativt og stratifisert utvalg av innbyggere. Ved en datainnsamling oppstår behovet for to hovedtyper korreksjon: Stratifisering og korreksjon av usikkerhet i svarene. Noen kjønn, aldersgrupper eller bydeler vil kunne være over- eller underrepresentert i utvalget som responderer på spørreskjemaet. Vi ønsker å gjenopprette et stratifisert respondentutvalg og dette gjøres ved at alle innsamlede spørreskjema vektes. Vektene bestemmes ved å løse det aktuelle minimeringsproblemet ved hjelp av Lagrange-multiplikatorer. En rimelighetsmodell uttrykker det psykologiske aspektet som spiller inn når spørreskjemaer fylles ut. Vi ønsker å korrigere for usikkerheten som oppstår i svarene på grunn av dette. Derfor anvendes en posteriori-modell på de innsamlede spørreskjemaene. Respondentutvalget evalueres ved å se hvor sensitive de innsamlede spørreskjemaene er til korreksjon ved hjelp av stratifisering og korreksjon av potensielle usikkerheter i svarene. Dette gjøres ved å sammenligne fire ulike estimerte andeler. Usikkerhetskorreksjon er avgjørende for riktig sentrering av de estimerte andelene. Sentreringen kan forbedres ytterligere ved stratifisering, men på bekostning av større spredning. Den estimerte usikkerhetskorrigerte andelen kommer like godt ut som den estimerte andelen som både er stratifisert og usikkerhetskorrigert når deres RMSE sammenlignes. Ved å kun sammenligne sentreringen til disse to estimatene, presterer den stratifiserte og usikkerhetskorrigerte best. Korreksjon av respondentutvalget ved hjelp av stratifisering og korreksjon av potensielle usikkerheter i svarene synes å være effektive verktøy. De korrigerer for svarskjevheter og usikkerhet i svarene som oppstår i en statistisk undersøkelse hvor menneskers uforutsigbare atferd er involvert.

Preface

This master's thesis is submitted to complete my degree of Master of Science (M.Sc.) in Applied Physics and Mathematics with Industrial Mathematics as main profile with further specialization in statistics. The degree is accomplished at the Department of Mathematical Sciences (IMF) at the Norwegian University of Science and Technology (NTNU) in Trondheim.

Throughout the years at NTNU it has become clear to me that I want to use my skills and knowledge to something meaningful. To me it is meaningful to use statistics to solve problems that hopefully improve the society in one way or another. I have engaged in a local church in Trondheim, as well as being a student at NTNU. Through our involvement with the most vulnerable people in our city I have come to know many people with a background so far away from my own privileged background.

This master's thesis is a result of my motivation to combine my interest in statistics with my care for people. In many ways it is a daring choice since modelling data in social science leads to quite a few challenges. But in the end I have learned a lot. Especially that every collection of data requires a lot of work to ensure that the data are representative. It was a time-consuming task to trace the origin of the data used in Oslo Monitor 1.0.

I want to thank NTNU for being so attentive and flexible and my family and friends for keep telling me that I can do this. A special thanks goes to Bergljot Matre Gåsland for your extensive work in proof reading the master's thesis and to Oscar Christian Ameln for your great company and indispensable help this last month. Last but not least, I want to give thanks to my supervisor, Henning Omre, for being (almost) as 'unorthodox' as myself. Thank you for being willing to supervise this alternative master's thesis, as well as being such an encouraging life coach!

Bergitte Viste
NTNU, Trondheim
July 1, 2020

Contents

Chapter 1

Introduction

To understand human behaviour and opinions is a complex challenge that occurs in social science. It is hard to define relevant aspects to account for when combining complex personalities with individual experiences. To generalize and conclude on what to be true is even harder. In addition, there is a lot of uncertainty connected to the collection of responses as humans by nature want to portray themselves in a good light. Still, it is worth trying to get insight to the human behaviour and opinions, as it might give useful information. Both qualitative and quantitative approaches may be used depending on the goal. Either approach includes a process starting with some sort of preparation, followed by the collection of data, an analysis of the data and then a presentation of the research as a report. The goal of the preparation is to decide on a problem to look into and why. As this study intends to present the statistical aspect of social science, the quantitative approach is used.

1.1 Problem and Motivation

A typical problem in social science is to monitor the situation of a city or area by different factors. This is done in Oslo Monitor 1.0, which is a report that was released in January 2018 by The Think Tank Skaperkraft in cooperation with church leaders in Oslo. The target groups are church leaders and leaders of Christian organizations. Still, the findings probably are interesting for a wider range of readers.

1.1.1 Introduction to Oslo Monitor 1.0

The report accounts for the spiritual situation, social suffering and cultural challenges in Oslo and provides the base for decisions concerning activities supporting the ultimate goal: Make Oslo an even better city to live in for everybody (Talset, 2018).

Spiritual Situation

The factors considered regarding the Christian spiritual situation are the population's attitudes towards religion, involvement in a Christian community and Bible usage. The different attitudes tells if a person believes in God or not and/or define themself as a personal Christian or not. The level of involvement is measured by the number of people attending activities at church weekly. Bible usage is measured by how often a person reads the Bible. The report also accounts for the distribution of churches in the different districts, their challenges and an overview of new churches that have been planted the last 15-25 years.

Social Suffering

Important factors considered for depicting social suffering are loneliness, child neglect, child poverty, life expectancy, divorce and social differences measured by disability, income and education. The report accounts for the number of people finding themselves lonely, children that are in need of foster care, children raised in a home of low income, the expected lifetime when a child is born and the number of children experiencing their parents getting a divorce. The social differences between the districts in the east and the west of the city are also quantified.

Cultural Challenges

The cultural challenges monitor mindsets and attitudes inspired by the trends in the society and their patterns. The spheres involved are: Media, high school drop-outs, illegal employment and volunteering. The report looks into the number of teenagers not finishing high school within five years as well as the amount of illegal employment in Norway. At last an insight in the volunteering culture is given.

Action Points

The report seeks to present the current situation within the three main areas in Oslo. As a result of the analysis the report proposes some actions to take in the upcoming years within each of the three main areas:

- There is a need for establishing a strategy for church planting and reaching people with the gospel the next 15-20 years. Moreover the use of the Bible among the church members and the population should be stimulated.
- The churches should reach out for people finding themselves lonely.
- Information and training with respect to being a foster care is needed. Parents should be informed and guided to preclude and support the children. Support should be given to those working to prevent teenagers from dropping out of high school.
- An effort to change the attitudes towards illegal employment is needed. Moreover the business should be encouraged to be more purpose driven and to spend resources to finance social and religious volunteering.

Collection of Data

The discussion in the report is based on data from secondary sources like Statistics Norway (SSB), Norwegian Institute of Public Health (FHI), KIFO, IPSOS MMI, NOVA, PISA, NLA Gimlekollen and the municipality of Oslo. In addition, a questionnaire was distributed to church leaders in Oslo in January 2017; whereas 31 out of 164 replied. A collection of the number of church attendees in various churches in Oslo was done directly during the autumn of 2017. The findings in Oslo Monitor 1.0 are descriptive representations of the different factors, or variables, considered in the report. A further analysis of the report is the object of interest in chapter 3.

1.1.2 Introduction to Lilleby Monitor

Oslo Monitor 1.0 is the inspiration for this master's thesis. But since the available data in Oslo Monitor 1.0 originate from different data sources, the interactions between the variables are not accounted for. Hence, the only available insight comes from the individual or univariate variables. A data set should include data on multiple variables collected on the same person, to get insight to the interactions among variables.

The relevant variables in Oslo Monitor 1.0 are given parametric distributions. The idea is to generate a realization of the population of a city, Lilleby. This is done by a statistical model that models the interplay among the variables. Insight into the interactions can be obtained by simulating a true Lilleby and discuss the results from Oslo Monitor 1.0 relative to the simulated Lilleby.

Further we simulate and distribute questionnaires to the residents of the realized Lilleby. Their answers with all their subjectivity make up the primary data. Correction models are applied and the effect is measured and compared by proportion estimates. This is the idea behind Lilleby Monitor.

The ultimate goal is inspired by Oslo Monitor 1.0: To describe the social and spiritual situation for the sake of indicating the primary needs in the different districts of Lilleby.

1.2 Data Collection and Method

To model the interactions of variables, inspired by Oslo Monitor 1.0, a multivariate statistical model has to be used. The multivariate Gaussian distribution is commonly used to handle big data sets where multiple variables are included. Still, it is not always applicable. This is the case when the outcomes of the individual variables and the interactions between them are not continuous. A multinomial distribution could be the answer. But as any other distribution it can only model the nature of the variables if its parameters are known or can be estimated from an existing data set.

Based on the concept of copulas, strategies have been developed to model data when the only available information comes from univariate variables. The strategies work fine for continuous or discrete variables. But for categorical variables it is more complicated. Hence, options are lacking as to model categorical data. Based on the concept of pair-copula constructions, we develop a sequentially computing strategy to simulate categorical

bi- or multivariate data. The strategy accounts for the relevant interactions between the variables.

Lilleby is the simulated town where the residents follow the interaction model. A questionnaire is distributed to the realized population of Lilleby. We apply a likelihood model to the responses. This is done to express the psychological aspects that might affect the responses of a questionnaire. Data collection deals with two major types of correction: Stratification and bias correction. We develop a stratification model to enforce a stratified sample of Lilleby and a posterior model to correct for bias from potential prejudices in the responses.

The goal of the distributed questionnaire is to monitor the true state of the population of Lilleby by the use of stratification and bias correction. Hence, the effect of the unpredictable behaviour of humans is to some extent limited.

1.3 Outline

The following chapter introduce the statistical definitions and models we use throughout the study. Chapter 3 contains a presentation of a statistical analysis of Oslo Monitor 1.0. In chapter 4 some general characteristics are introduced and defined. The interaction model used to account for interplay among the included variables is derived. Additionally, the simulation of Lilleby is carried out and dependence among the variables is visualized. Chapter 5 introduces the extensive process behind every questionnaire. As well as the concepts of stratification and likelihood modelling of the psychological aspects of answering a questionnaire. The response sample of Lilleby is simulated and evaluated by its sensitivity to the stratification and posterior model, by the comparison of proportion estimates. Chapter 6 yields some concluding remarks.

Chapter 2

Statistical Definitions and Models

We introduce some basic statistical terminology and definitions. Relevant statistical models are presented and an introduction to statistical inference is included. Notation and definitions are inspired by Walpole et al. (2012), Geer (2019) and Casella and Berger (2002).

2.1 Random Variable and Sample Space

Data are gathered as samples; being a collection of observations drawn from a population. The sample is represented by a random variable, X. In general, the random variable $X \in \Omega_X$, with outcome x, takes one element in Ω_X; the sample space of X.

Countable Sample Space

A countable sample space, Ω_X, is usually a finite set of outcomes and can be either categorical or discrete.

The categorical sample space may be non-ordered. An example from Oslo Monitor 1.0 is the sample space: $\Omega_X = \{$'I believe in God', 'I do not believe in God'$\}$. This sample space usually takes on a binomial distribution.

The discrete sample space may be ordered. In Oslo Monitor 1.0 an example of such a sample space is the measured number of children experiencing their parents getting a divorce. The sample space is given by: $\Omega_X = \mathbb{N}_{\oplus}$; being positive, natural numbers. This sample space usually takes on a Poisson distribution.

Continuous Sample Space

A continuous sample space, Ω_X, is an infinite set of outcomes and can be open, bounded or an interval.

An example of an open sample space in Oslo Monitor 1.0 is the differences in life expectancy by birth. The open sample space is given by: $\Omega_X = \mathbb{R}$. Such a sample space could, for instance, take on a Gaussian distribution.

A bounded sample space is given by: $\Omega_X = \mathbb{R}_{[0,\infty)} = [0, \infty) \subset \mathbb{R}$ or by $\Omega_X = \mathbb{R}_{\oplus}$. In Oslo Monitor 1.0 the data of the average income provides an example of a bounded sample space and could either take on a log-Gaussian distribution or the Pareto distribution.

The sample space is an interval when $\Omega_X = \mathbb{R}_{[a,b]} = [a, b] \subset \mathbb{R}$, where $a < b$. An example from Oslo Monitor 1.0 is the proportion of people getting in-disability support. Such a sample space takes on the beta distribution.

Multivariate Sample Space

The idea of Lilleby requires a model that can account for the interaction among variables. We denote the vector of multiple random variables by: $\mathbf{X} = (X_1, X_2, ..., X_k)$, where k is the number of variables in the model. This is hence a k-variate model. The sample space of $\mathbf{X}$ is given by $\mathbf{X} \in \Omega_{\mathbf{X}}$, where: $\Omega_{\mathbf{X}} = \Omega_{X_1} \times \Omega_{X_2} \times \ldots \times \Omega_{X_k}$. This is a k-variate, or multivariate, sample space.

2.2 Probability Distributions

Assumptions are made of the random variables on the sample space. This allows us to assign a probability distribution, $p(x)$, to the random variable, X, where $x \in \Omega_X$.

In the categorical case we define the probability mass function (pmf), $p(x)$, to satisfy the following:

1. $\sum_{x \in \Omega_X} p(x) = 1, \quad$ where $x \in \Omega_X$ is countable,
2. $p(x) \geq 0 \quad$ and
3. $P(X = x) = p(x)$.

In the continuous case $p(x)$ is called the probability density function (pdf) and the following holds:

1. $\int_{\Omega_X} p(x)dx = 1, \quad$ where $x \in \Omega_X$ is continuous,
2. $p(x) \geq 0, \quad$ for all $x \in \Omega_X, \quad$ and
3. $P(a < X < b) = \int_a^b p(x)dx$.

For a multivariate random variable, $\mathbf{X}$, we assign a multivariate probability distribution, $p(\mathbf{x})$, where $\mathbf{x} \in \Omega_{\mathbf{X}}$ is a vector. The sample space can be either categorical or continuous. An example of a multivariate probability distribution is the multinomial distribution; to be introduced.

Expected Value and Variance

We have different scalar measures for a distribution. These are only relevant for discrete and ordered, and continuous variables. The two most common are the expected value and the variance.

The expected value is the probability weighted average, denoted by μ_X and defined by:

$$\mu_X = \mathrm{E}[X] = \sum_{x \in \Omega_X} xp(x), \quad \text{when } X \text{ is a discrete random variable}$$

and

$$\mu_X = \mathrm{E}[X] = \int_{\Omega_X} xp(x)dx, \quad \text{when } X \text{ is a continuous random variable.}$$

The variance, denoted by σ_X^2, is the spread of values centered at the expected value and is defined by:

$$\sigma_X^2 = \mathrm{Var}[X] = \mathrm{E}[(X - \mu_X)^2] = \sum_{x \in \Omega_X} (x - \mu_X)^2 p(x),$$

when X is a discrete random variable

and

$$\sigma_X^2 = \mathrm{Var}[X] = \mathrm{E}[(X - \mu_X)^2] = \int_{\Omega_X} (x - \mu_X)^2 p(x)dx,$$

when X is a continuous random variable.

Alternatively the variance of the random variable X can be expressed as:

$$\sigma_X^2 = \mathrm{Var}[X] = \mathrm{E}[X^2] - \mu_X^2.$$

Taking the positive square root of σ_X^2 yields the standard deviation of X, denoted by σ_X or $\mathrm{Sd}[X]$.

Joint Probability Distributions

Random variables with their probability distributions can be considered jointly to evaluate the simultaneous outcome of them. Consider $X_1 \in \Omega_{X_1}$ and $X_2 \in \Omega_{X_2}$. Their joint probability distribution is then denoted by $p(x_1, x_2)$, yielding a bivariate distribution. The joint probability distribution of k random variables defines a k-variate distribution.

If $X_1 \in \Omega_{X_1}$ and $X_2 \in \Omega_{X_2}$ are both categorical the following holds:

1. $p(x_1, x_2) \geq 0, \quad \text{for } (x_1, x_2) \in \Omega_{X_1} \times \Omega_{X_2},$
2. $\sum_{x_1 \in \Omega_{X_1}} \sum_{x_2 \in \Omega_{X_2}} p(x_1, x_2) = 1,$
3. $P(X_1 = x_1, X_2 = x_2) = p(x_1, x_2),$

for any region $A \subset \Omega_{X_1} \times \Omega_{X_2}$, $P[(X_1, X_2) \in A] = \sum \sum_A p(x_1, x_2)$.

If $X_1 \in \Omega_{X_1}$ and $X_2 \in \Omega_{X_2}$ are continuous the following holds:

1. $p(x_1, x_2) \geq 0, \quad$ for $(x_1, x_2) \in \Omega_{X_1} \times \Omega_{X_2}$,
2. $\int_{\Omega_{X_2}} \int_{\Omega_{X_1}} p(x_1, x_2) dx_1 dx_2 = 1$,
3. $P[(X_1, X_2) \in A] = \int \int_A p(x_1, x_2) dx_1 dx_2, \quad$ for any region $A \subset \Omega_{X_1} \times \Omega_{X_2}$.

The marginal distributions of X_1 and X_2 are found by summing or integrating over $X_2 \in \Omega_{X_2}$ and $X_1 \in \Omega_{X_1}$, respectively. They are denoted by:

$$p(x_1) = \sum_{x_2 \in \Omega_{X_2}} p(x_1, x_2) \quad \text{and} \quad p(x_2) = \sum_{x_1 \in \Omega_{X_1}} p(x_1, x_2)$$

for the discrete or categorical case. For the continuous case:

$$p(x_1) = \int_{\Omega_{X_2}} p(x_1, x_2) dx_2 \quad \text{and} \quad p(x_2) = \int_{\Omega_{X_1}} p(x_1, x_2) dx_1.$$

The probability of X_1 given X_2, $p(x_1 \mid x_2)$, is called the conditional pmf for the categorical case and the conditional pdf for the continuous case. By definition it follows that:

$$p(x_1 \mid x_2) = \frac{p(x_1, x_2)}{p(x_2)}, \quad \text{when} \quad p(x_2) > 0,$$

for $X_1 \in \Omega_{X_1}$ and $X_2 \in \Omega_{X_2}$. This is true for $p(x_2 \mid x_1)$ as well:

$$p(x_2 \mid x_1) = \frac{p(x_1, x_2)}{p(x_1)}, \quad \text{when} \quad p(x_1) > 0,$$

for $X_2 \in \Omega_{X_2}$ and $X_1 \in \Omega_{X_1}$.

If X_1 and X_2 are statistically independent one has that:

$$p(x_1, x_2) = p(x_1 \mid x_2) p(x_2) = p(x_2 \mid x_1) p(x_1) = p(x_1) p(x_2),$$

for all $(x_1, x_2) \in \Omega_{X_1} \times \Omega_{X_2}$. Hence, $p(x_1 \mid x_2) = p(x_1)$ and $p(x_2 \mid x_1) = p(x_2)$.

As for the case with k random variables, $\mathbf{X} = (X_1, X_2, ..., X_k) \in \Omega_{\mathbf{X}}$, the joint probability function is denoted by $p(x_1, x_2, \ldots, x_k)$. The marginal distribution of $(X_1, ..., X_i)$ is hence given by:

$$p(x_1, ..., x_i) = \sum_{x_{i+1} \in \Omega_{X_{i+1}}} \cdots \sum_{x_k \in \Omega_{X_k}} p(x_1, x_2, \ldots, x_k),$$

in the discrete case, and

$$p(x_1, ..., x_i) = \int_{\Omega_{X_k}} \cdots \int_{\Omega_{X_{i+1}}} p(x_1, x_2, \ldots, x_k) dx_{i+1} dx_{i+2} \cdots dx_k,$$

in the continuous case.

We now denote each marginal distribution by $p_1(x_1), \ldots, p_k(x_k)$. The conditional probability distribution of X_i given $\mathbf{X}_{-i}$, where $\mathbf{X}_{-i} = (X_1, ..., X_{i-1}, X_{i+1}, ..., X_k)$, is given by:

$$p(x_i \mid \mathbf{x}_{-i}) = \frac{p(\mathbf{x})}{p(\mathbf{x}_{-i})},$$

for $X_i \in \Omega_{X_i}$ and $\mathbf{X}_{-i} \in \Omega_{\mathbf{X}_{-i}}$ as long as $p(\mathbf{x}_{-i}) > 0$.

The random variables $X_1, \ldots, X_k$ are mutually statistically independent if:

$$p(x_1, x_2, \ldots, x_k) = p_1(x_1)p_2(x_2)\ldots p_k(x_k), \quad \text{for all } (x_1, x_2, \ldots, x_k) \in \Omega_{\mathbf{X}}.$$

Hence, $p(x_i \mid \mathbf{x}_{-i}) = p(x_i)$ for all $x_i \in \Omega_{\mathbf{X}}$ if $X_1, ..., X_k$ are mutually statistically independent.

Covariance and Correlation

The nature of the association between two random variables, $X_1 \in \Omega_{X_1}$ and $X_2 \in \Omega_{X_2}$, is measured by the covariance, given by:

$$\sigma_{X_1X_2} = \text{Cov}[X_1, X_2] = \text{E}[X_1X_2] - \mu_{X_1}\mu_{X_2},$$

where μ_{X_1} and μ_{X_2} are the respective means of X_1 and X_2. In other words, the covariance is a measure of the joint variability of two random variables:

$$\sigma_{X_1X_2} = \text{E}[(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})] = \sum_{x_1 \in \Omega_{X_1}} \sum_{x_2 \in \Omega_{X_2}} (x_1 - \mu_{X_1})(x_2 - \mu_{X_2})p(x_1, x_2),$$

if X_1 and X_2 are discrete, and

$$\sigma_{X_1X_2} = \text{E}[(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})] = \int_{\Omega_{X_2}} \int_{\Omega_{X_1}} (x_1 - \mu_{X_1})(x_2 - \mu_{X_2})p(x_1, x_2)dx_1dx_2,$$

if X_1 and X_2 are continuous.

The covariance is normalized to measure the strength of the linear relation, resulting in the correlation coefficient given by:

$$\rho_{X_1X_2} = \frac{\sigma_{X_1X_2}}{\sigma_{X_1}\sigma_{X_2}}, \quad \text{where} -1 \leq \rho_{X_1X_2} \leq 1.$$

2.3 Parametric Probability Distributions

We can assign a parametric probability distribution, $p(x; \boldsymbol{\theta})$, to the random variable, X, where $\boldsymbol{\theta}$ are the model parameters. The categorical probability distributions and their parameters are introduced as well as their expected value and variance. These measures are now defined as functions of the model parameters for the given distribution: $\mu_X = \mu_X(\boldsymbol{\theta})$ and $\sigma_X^2 = \sigma_X^2(\boldsymbol{\theta})$. We denote Θ as the parameter space (Geer, 2019).

2.3.1 The Bernoulli Distribution

The categorical variable $X \in \Omega_X = \mathbb{N}_{[0,1]}$ with two outcomes is termed the Bernoulli distribution. The random variable takes on the value 1 with probability p and the value 0 with probability $1 - p$. The distribution is given by:

$$p(x; p) = p^x(1 - p)^{1-x}, \quad \text{for } x = 0, 1,$$

where $\boldsymbol{\theta} = p$ and $p \in \Theta = \mathbb{R}_{[0,1]}$.

The Bernoulli distribution models a single Bernoulli trial meaning that the outcome of a single trial will be either success or failure. This results in a boolean value, $X \in \mathbb{N}_{[0,1]}$.

The expected value and variance is:

- $\mu_X = \mathrm{E}[X] = p$,
- $\sigma_X^2 = \mathrm{Var}[X] = p(1 - p)$.

2.3.2 The Binomial Distribution

The categorical variable $X \in \Omega_X = \mathbb{N}_{[0,n]}$ with n outcomes is termed the binomial distribution and is given by:

$$p(x; n, p) = \binom{n}{x} p^x(1 - p)^{n-x}; \quad x = 0, 1, ..., n,$$

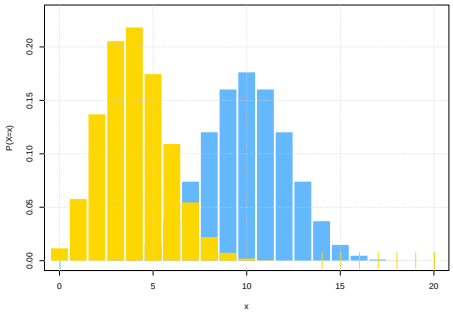
where $\boldsymbol{\theta} = (n, p)$ and $\boldsymbol{\theta} \in \Theta = \mathbb{N}_+ \times \mathbb{R}_{[0,1]}$ are the model parameters. The effect of different values of the parameters p and n is displayed in figure 2.1a and b.

The categorical variable $x \in \mathbb{N}_{[0,n]}$ represents the number of successes in a sequence of n independent, identically distributed Bernoulli trials, Y_i, with probability p for success and $1 - p$ for failure. Hence, the categorical variable can be denoted by: $X = \sum_{i=1}^{n} Y_i$.

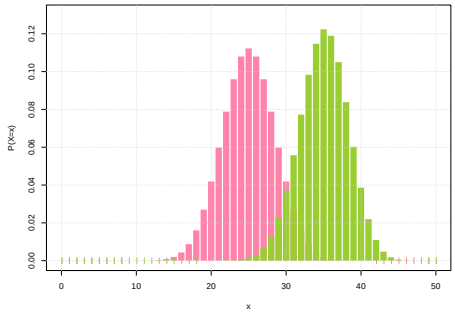
The expected value and variance is:

- $\mu_X = \mathrm{E}[X] = np$,
- $\sigma_X^2 = \mathrm{Var}[X] = np(1 - p)$.

We add some additional remarks regarding the binomial distribution. When the number of trials, n, is sufficiently large and p is sufficiently small, the binomial distribution converges towards the Poisson distribution, with parameter $\lambda = np$; given by $p(x; np)$. The product of n and p must remain constant which it will as p tends to zero. In addition the binomial distribution can be approximated by the Gaussian distribution as long as n is large enough and p is not too close to either 0 or 1. The corresponding Gaussian distribution is then given by $p(x; np, np(1 - p))$.



(a) The binomial distribution for $n = 20$ with $p = 0.5$ (blue) and $p = 0.2$ (yellow).



(b) The binomial distribution for $n = 50$ with $p = 0.5$ (pink) and $p = 0.7$ (green).

Figure 2.1: The binomial distribution for different values of the parameters p and n.

2.3.3 The Multinomial Distribution

In the case of k different outcomes for each $x_i \in \mathbb{N}_{[0,n]}$, where $i \in \mathbb{N}_{[0,k]}$ and $\sum_{i=1}^{k} x_i = n$, the multinomial distribution is used. The vector of multiple random variables is given by $\mathbf{X} = (X_1, X_2, ..., X_k)$. The sample space of $\mathbf{X}$ is given by $\mathbf{X} \in \Omega_{\mathbf{X}}$, where: $\Omega_{\mathbf{X}} = \Omega_{X_1} \times \Omega_{X_2} \times \ldots \times \Omega_{X_k}$.

The multinomial distribution is defined by:

$$p(\mathbf{x}; n, \mathbf{p}) = \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k}, \quad \text{where} \sum_{i=1}^{k} x_i = n.$$

The k possible mutually, exclusive outcomes has a corresponding probability, p_k, where $\sum_{i=1}^{k} p_i = 1$, $\mathbf{p} = (p_1, ..., p_k)$ and each $p_i \in \mathbb{R}_{[0,1]}$. Hence, the model parameters are $\boldsymbol{\theta} = (n, \mathbf{p})$ and $\boldsymbol{\theta} \in \Theta = \mathbb{N}_+ \times \mathbb{R}_{[0,1]}^k$.

Each trial in an experiment has one of k categorical outcomes with probability p_k. The number of independent trials are n. The random variable, $\mathbf{x} = (x_1, \ldots x_k)$, contains the number of outcomes of each category and is multinomial distributed. The multinomial distribution is a generalization of the binomial distribution.

The expected value, variance and covariance of the multinomial distribution is defined by:

- $\mu_{X_i} = \mathrm{E}[X_i] = np_i$,
- $\sigma_{X_i}^2 = \mathrm{Var}[X_i] = np_i(1 - p_i)$,
- $\sigma_{X_i, X_j} = \mathrm{Cov}[X_i, X_j] = -np_i p_j \quad \text{for} \quad i \neq j$.

2.4 Statistical Inference

Each X_i in the observed data, $\mathbf{X}_n = (X_1, \ldots, X_n) \in \Omega_{\mathbf{X}_n}$, is assumed to be independent and identically distributed (iid) from an infinite population with a given distribution, $p(x; \boldsymbol{\theta})$. We want to estimate the function of a given parameter $\tau(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of model parameters for the given distribution. An estimator is a function of the random variable $\mathbf{X}_n$ denoted by $W = W(\mathbf{X}_n)$ (Casella and Berger, 2002).

Uniform Minimum Variance Unbiased Estimation (UMVUE)

Often we require the estimators of the model parameters to be unbiased; meaning that the expected value of the estimator equals the quantity ought to estimate. If there are two unbiased candidates for $\tau(\boldsymbol{\theta})$ we use the estimator with smallest variance; the most efficient estimator of $\tau(\boldsymbol{\theta})$. Hence, W^* is said to be a uniform minimum variance unbiased estimator (UMVUE) of $\tau(\boldsymbol{\theta})$ if:

$$W^* = \arg\min_{W \in \mathcal{W}} \text{Var}[W]$$

where $\mathcal{W} = \{W : \text{E}[W] = \tau(\boldsymbol{\theta})\}$.

Maximum Likelihood Estimation (MLE)

A common method used to find the estimator of a model parameter in a probability distribution is the method of maximizing the likelihood function.

The iid observations, $\mathbf{X}_n$, with outcome $\mathbf{x}_n$ have a discrete or continuous distribution, $p(\mathbf{x}_n; \boldsymbol{\theta})$, with parameters $\boldsymbol{\theta}$. The joint distribution of the random variables is given by:

$$p(\mathbf{x}_n; \boldsymbol{\theta}) = p(x_1; \boldsymbol{\theta}) \ldots p(x_n; \boldsymbol{\theta}).$$

If we insert the outcome, $\mathbf{x}_n$, and consider the expression to be a function of $\boldsymbol{\theta}$, we obtain the likelihood function, $L(\boldsymbol{\theta}; \mathbf{x}_n)$.

We want to maximize the likelihood function with respect to $\boldsymbol{\theta}$. Taking the natural logarithm of a function does not change its maximizer, since the logarithm is a continuous strictly increasing function over the range of the likelihood. The logarithm also has some convenient properties which allows for simplifications when computing the maximizer. The log-likelihood is given by: $l(\boldsymbol{\theta}; x) = \ln L(\boldsymbol{\theta}; x)$. Deriving the log-likelihood function by the parameters, $\boldsymbol{\theta}$, yields the parameter value that produces the largest probability of obtaining the sample, defined as:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}; \mathbf{x}_n).$$

This is called the maximum likelihood estimate (MLE) of the parameter. The expression is an explicit function of the observed data. The MLE converges in probability and is consistent with asymptotic efficiency.

Moment Estimation

Estimation of the first two moments of X yields the estimated expected value of X, $\hat{\mu}_X$, and the estimated variance of X, $\hat{\sigma}_X^2$:

- $\hat{\mu}_X = \widehat{\mathrm{E}}[x] = \frac{1}{n}\sum_{i=1}^{n} X_i$,
- $\hat{\sigma}_X^2 = \widehat{\mathrm{Var}}[x] = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_X)^2$.

Alternatively, the estimates are found by directly substituting $\hat{\boldsymbol{\theta}}$ for the given distribution into the expressions for $\mu_X(\boldsymbol{\theta})$ and $\sigma_X^2(\boldsymbol{\theta})$ for the respective distribution. The estimated covariance between two variables for a multivariate distribution is found in the same way.

The variance for the estimated expected value obtains the lowest possible variance:

$$\sigma_{\hat{\mu}_X}^2 = \mathrm{Var}[\hat{\mu}_X] = \mathrm{Var}[\widehat{\mathrm{E}}[x]] = \frac{1}{n}\mathrm{Var}[x],$$

where n is the total number of observations. The actual value is obtained by inserting $\hat{\sigma}_X^2$.

2.5 Assumptions

So far we have assumed that each X_i is iid from a distribution $p(x;\boldsymbol{\theta})$ for a sample of $\mathbf{X}_n = (X_1, ..., X_n) \in \Omega_{\mathbf{X}_n}$. By iid it is meant that all random variables must have the same probability distribution and they must all be mutually independent. We have also assumed that the population size is infinite.

When Assumptions Fail

Problems may arise when looking into the assumptions made on the sample $\mathbf{X}_n = (X_1, ..., X_n) \in \Omega_{\mathbf{X}_n}$, where each X_i are assumed to be iid from a distribution $p(x;\boldsymbol{\theta})$.

First of all, as a given probability distribution is assigned to $\mathbf{X}_n$, each random variable is assumed to have the same probability distribution. But this might not always be the case. Especially if the data in $\mathbf{X}_n$ are from different sources and if the sample spaces are not clearly defined.

Secondly, each variable in $\mathbf{X}_n$ should be independent of the others. Independence is less likely to be satisfied if data are collected within a group of people with some sort of relationship among them; like a family, school class or neighbourhood. In practice it is impossible to ensure that the sample is perfectly random.

Lastly, an infinite population is assumed when calculating the uncertainty and variance in the data, as this yields attractive limiting properties. In practice, the population of a city is finite and for certain sub-groups it can be fairly small, hence the sample may be more representative than expected.

Finite Population Inference

The variable $X \in \Omega_X = \mathbb{N}_{[0,n]}$ follows a binomial distribution, $p(x; n, p)$, and is used to demonstrate the features of finite population inference. The number of successes, with probability p for success and $1 - p$ for failure, in a sequence of n iid Bernoulli trials, is collected in $x \in \mathbb{N}_{[0,n]}$. To simplify the notation we denote a success by the number 1 and a failure by 0.

In the infinite population case, for a sample of $\mathbf{X}_n = (X_1, ..., X_n) \in \Omega_{\mathbf{X}_n}$, iid Bernoulli trials, the parameter estimator of p is given by: $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} I(X_i = 1)$. By using the expressions for the expected value and variance, introduced for the binomial distribution, the following is true for $\hat{p}$:

- $\mu_{\hat{p}} = \mathrm{E}[\hat{p}] = p$,
- $\sigma^2_{\hat{p}} = \mathrm{Var}[\hat{p}] = \frac{p(1-p)}{n}$.

In the finite population case with population $\mathbf{X}_n$, defined by the sample above, we no longer focus on the parameter p. We focus on a stochastic variable of the population proportion defined as: $p_n = \frac{1}{n}\sum_{i=1}^{n} I(X_i = 1)$. Let the sample of the finite population be of size $m \leq n$ and denote it by $\mathbf{X}^*_m = (X^*_1, ..., X^*_m) \in \Omega_{\mathbf{X}^*_m}$. Each X^*_i, for $i = 1, ..., m$, is uniformly drawn from $\mathbf{X}_n$ without replacement. The estimate of p_n is now given by: $\hat{p}_n = \frac{1}{m}\sum_{j=1}^{m} I(X^*_j = 1)$.

The goal in finite population inference is to assess the population proportion, p_n, based on its estimate, $\hat{p}_n$. The following properties can be used to evaluate the estimator of the population proportion:

- $\mu_{p_n - \hat{p}_n} = \mathrm{E}[p_n - \hat{p}_n] = \mathrm{E}[p_n] - \mathrm{E}[\hat{p}_n] = 0$,
- $\sigma^2_{p_n - \hat{p}_n} = \mathrm{Var}[p_n - \hat{p}_n] = \mathrm{Var}[p_n] + \mathrm{Var}[\hat{p}_n] - 2\mathrm{Cov}[p_n, \hat{p}_n]$
$= \Big(\frac{1}{m} - \frac{1}{n}\Big)p(1-p), \quad \text{for} \quad m \leq n,$

with

$$\mathrm{Cov}\Big[\frac{1}{n}\sum_{i=1}^{n} I(X_i = 1), \frac{1}{m}\sum_{j=1}^{m} I(X^*_j = 1)\Big] = \frac{1}{nm}\sum_{j=1}^{m} \mathrm{Var}\big[I(X^*_j = 1)\big] = \frac{1}{n}p(1-p).$$

Some examples of different sample sizes, m, are plotted in figure 2.2 to illustrate the nature of a finite population with $n = 10$. The variance decreases as m increases and approaches n. When $m = n$ the variance is equal to zero. For fixed m, if $n \to \infty$ then $\sigma^2_{p_n - \hat{p}_n} \to \frac{p(1-p)}{m}$ for every m, resulting in the variance corresponding to an infinte population.

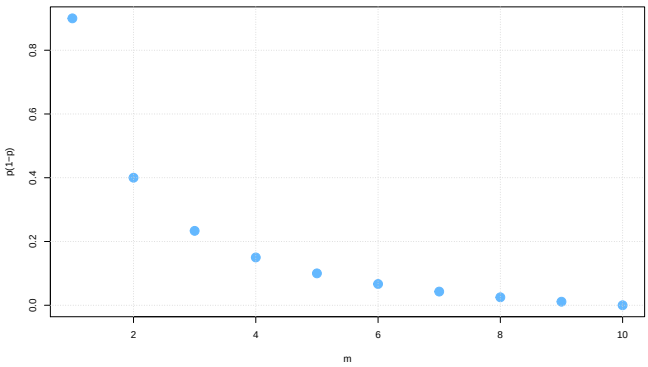


Figure 2.2: Examples of finite sample sizes when $n = 10$.

Population Proportion Estimation

Consider a very large population of size N, i.e. the population of Oslo. Let $X \in \Omega_X = \mathbb{N}_{[0,1]}$ be a binary characteristic of each inhabitant in the population, with probability p for $x = 1$. Collect a random subsample of size $n \ll N$ and let n_0 and n_1 be the number of zeros and ones, respectively. Hence, $n = n_0 + n_1$. Since $n \ll N$, assume that n_1 is binomial, with $p(n_1; n, p)$. Then p is estimated by its MLE:

$$\hat{p} = \frac{n_1}{n},$$

with

- $\mathrm{E}[\hat{p}] = p$,
- $\mathrm{Var}[\hat{p}] = \frac{p(1-p)}{n} \approx \frac{\hat{p}(1-\hat{p})}{n}$.

Define N_0 and N_1 to be the number of zeros and ones in the large population, hence $N = N_0 + N_1$. The predictor for N_1 is then:

$$\hat{N}_1 = N\hat{p},$$

with

- $\mathrm{E}[\hat{N}_1] = N\mathrm{E}[\hat{p}] = Np = N_1$,
- $\mathrm{Var}[\hat{N}_1] = N^2\mathrm{Var}[\hat{p}] = \frac{N^2}{n}p(1-p) \approx \frac{N^2}{n}\hat{p}(1-\hat{p})$.

The corresponding approximated 95% prediction interval is given by:

$$\left\{ \hat{N}_1 \pm 2\left[\frac{N^2}{n}\hat{p}(1-\hat{p}) \right]^{\frac{1}{2}} \right\}. \tag{2.1}$$

A population characteristic may have k possible outcomes, with probability p_k for the corresponding outcome, x_k. If this is the case, we assume that $\mathbf{n} = (n_1, .., n_k)$ is multinomial distributed with $p(\mathbf{n}; n, \mathbf{p})$, where $\mathbf{p} = (p_1, ..., p_k)$. Each p_k is estimated as for the binomial case.

Assumptions for Oslo Monitor 1.0

The MLEs are a direct result of the infinite population and iid assumptions. We also know that the MLE of a model parameter is consistent for a large sample and has the lowest possible variance. In Oslo Monitor 1.0 the overall number of respondents is high and the data originate from well-known sources. Hence, the infinite population and iid assumptions are assumed to be reasonable and valid as the data in Oslo Monitor 1.0 are discussed.

Chapter 3

Revisiting Oslo Monitor 1.0

Oslo Monitor 1.0 was released in January 2018 by The Think Tank Skaperkraft. We investigate the sources of the data behind the spiritual situation, social suffering and cultural challenges in the report. The relevant factors to include as variables in the simulation of Lilleby are assigned a categorical distribution. The corresponding parameter estimates are calculated based on the available data in Oslo Monitor 1.0. Some comments are made on the remaining variables as well. Note that N varies according to the relevant year of data collection for the variable of interest.

3.1 Spiritual Situation

The Christian spiritual situation considers the population's attitude towards religion, involvement in a Christian community and relationship to the Bible. The data originate from different studies, a collection of the number of church attendees done directly and a targeted questionnaire.

Attitudes Towards Religion

The attitudes towards religion tell us if a person believes in the Abrahamic God, define themself as a Christian with a personal relationship with God and how often they attends a church of any kind.

The data originate from the study 'Norsk Monitor 2015/16' by IPSOS (Ingebretsen, Holbæk-Hanssen, and Dalen, 2016). This is a report made for the Ministry of Children and Equality. Data were collected between September 2015 and January 2016. The collection of data involved an interview by phone, followed up by a questionnaire containing 129 pages to fill out. The questionnaire were completed by 3981 respondents in total over the age of 15. Out of these 3981 there were 376 respondents between 15 and 20 years old and 470 respondents in the age between 21 and 26. An audience analysis were used and the two groups were continuously compared to each other. At estimation of the results, a weighting of gender and age were made within each of the 5 regions of Norway (Nord-Norge, Trøndelag, Vestlandet, Østlandet og Sørlandet). This to ensure that the composition

of the sample is more statistically representative and to reduce the effect of any selection bias.

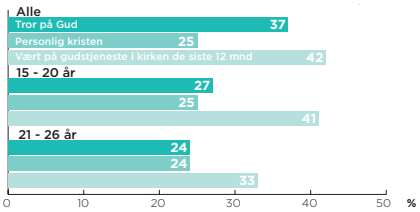
In Hellevik (2015) the implications of non-response in Norsk Monitor is discussed. Hellevik concludes that such surveys are representative despite the low response rate, as long as random selection is used and that the occurrences of non-responses are not systematic.

Oslo Monitor 1.0 refers to the changes in attitudes towards religion among youths in Norsk Monitor 2015/16. Figure 3.1a, b and c present the results used in Oslo Monitor 1.0. The size of the sample of respondents between 15 and 26 years old is $n = 846$. The different attitudes are: 'Belief in God', 'Personal Christian' and 'Attended church last 12 months'. Personal Christians are assumed to also believe in God. Each attitude is assigned a sample space as follow: $\Omega_{X_1} = \{$'Do believe in God', 'Do not believe in God'$\}$, $\Omega_{X_2} = \{$'Personal Christian', 'Not personal Christian'$\}$ and $\Omega_{X_3} = \{$'Have attended church the last 12 months', 'Have not attended church the last 12 months'$\}$. The number of individuals that believes in God, n_1, the number of personal Christians, n_2, and the number of individuals that attended church the last 12 months, n_3; each one follows a binomial distribution, with $p(n_1; n, p_1)$, $p(n_2; p_1 n, p_2)$ and $p(n_3; n, p_3)$, respectively. The total population of Oslo in 2015 were $N = 658390$ (Oslo Kommune, 2019). N_1, N_2 and N_3 denote the number of people believing in God, defining themselves as a personal Christian and have been attending church the last 12 months in all of Oslo.

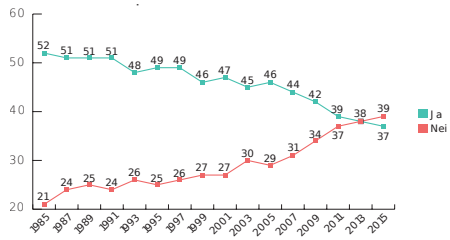
We estimate p_1, p_2 and p_3 and predict N_1, N_2 and N_3 by their corresponding approximated 95% prediction interval, PI_{N_1}, PI_{N_2} and PI_{N_3}, given in equation (2.1):

1. $\hat{p}_1 = 0.37, \quad \text{PI}_{N_1} = [243604 \pm 21857]$,
2. $\hat{p}_2 = 0.68, \quad \text{PI}_{N_2} = [165651 \pm 12846]$,
3. $\hat{p}_3 = 0.42, \quad \text{PI}_{N_3} = [276524 \pm 22344]$.

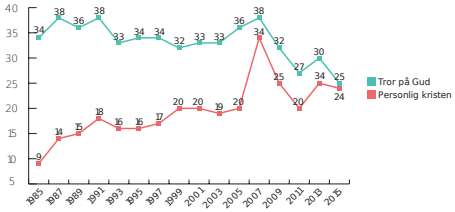
Because of the similarities between the two groups most of the analysis done in Norsk Monitor 2015/16 only assumes one group. This is verified by the proportions presented in figure 3.1a. The proportion that have attended church the least year is higher than the one for people that actually believe in God. This is interesting and might be explained by the fact that people seek the church either for special occasions or in grief, even though they do not believe in God. Figure 3.1b and c present changes over time in people believing in God or not, and in people defining themselves as personal Christian versus people believing in God. In both cases the total percentage is constant from 1985 till 2015. The people that either believe in God or not accounts for 75 percent of the population both in 1985 and 2015 meaning that an overall of 25 percent of the population do not take a stand. People that define themselves as personal Christians have increased, while people believing in God have decreased. An hypothesis to explain this is that the church has experienced a secularization over the last decades. This results in a need for Christians to either define themselves as personal Christian or cultural Christian; like attending church for special occasions.



(a) Amount of people believing in God, considering themselves as personal Christians and have been attending church the last year (Talset, 2018).



(b) Changes over time in people believing or not believing in God (Talset, 2018).



(c) Changes over time in people considering themselves to be personal Christians and people believing in God (Talset, 2018).

Figure 3.1: Data regarding attitudes towards religion.

Active Church Attendees

Active church attendees are defined as persons attending activities at church on a weekly basis.

The data shown in figure 3.2 are from what the Church of Norway (DNK) calls counting weeks. As well as a collection done by Skaperkraft by directly contacting 40 different churches in Oslo by phone during the autumn of 2017; whereas 34 responded. They were asked to account for the number of persons attending activities at church on a weekly basis. DNK arranges their counting weeks twice a year and the numbers in figure 3.2 are from the counting done in week 13 in 2017 for 33 churches. The number of weekly attendees from the other church communities were gathered by asking the church leaders to give an approximate estimate on how many persons they would say to join churchly activities on a weekly basis.

Even though n is unknown, we assign a sample space: $\Omega_X = \{$'Do attend church weekly', 'Do not attend church weekly'$\}$. The number of active church attendees, n_1, follows a binomial distribution, $p(n_1; n, p)$. The total population of Oslo in 2017 used in figure 3.2 is $N = 666757$. Furthermore, N_1 denotes the number of people that attends activities at church on a weekly basis in all of Oslo.

We estimate p and predict N_1 but cannot calculate the corresponding approximated 95% prediction interval of N_1, because n is unknown:

$$\hat{p} = 0.05, \quad \hat{N}_1 = 33338.$$

There is a high level of uncertainty in the numbers from other churches than the DNK churches since the leaders estimated the number without any actual countings from activities. It is primarily a guess. For a more accurate picture of number of persons attending the different churches, an actual counting should be done; like in the DNK churches.

Figure 3.2 shows that the proportions of the total population of Oslo that attends churches are small. The 'free churches' accounts for the highest attendance. An interesting note is the fact that only 1.21 percent of the population of Oslo attends a DNK church on a weekly basis, while 70 percent of the population of Norway are members of DNK (SSB, 2018d). The church appears to play an important part in people's lives as they stay a member even though they do not attend the church regularly.

Oslos befolkning	DKK+Ort	DNK	Frikirker	Immigrantk	Sum
666 757	6 935	8 076	11 237	6 006	32 254
I % av Oslos befolkning	1,04%	1,21%	1,69%	0,90%	4,84%
Relativ andel	22%	25%	35%	19%	100%

Figure 3.2: Amount and proportion of people attending different churches at a weekly basis (Talset, 2018).

Bible Usage

Bible usage is measured in Oslo Monitor 1.0 by how often a person reads the Bible.

The data used originate from a study called 'Nordmenns Bibelbruk' from 2017 by KIFO (Rafoss, 2017) made in cooperation with Bibelselskapet. The study discusses Norwegians use of the Bible and their attitudes towards it. The available data for the study originate from both Norwegian and international surveys regarding the Bible. An overview of the available data is shown in figure 3.3.

The problem regarding these surveys is that the questions are not asked in the same way with identical options each time. Also the number of questions regarding Bible usage is only one or two in each survey. The problem occurs when comparing the changes in Bible usage over time. The surveys made by TNS Gallup on behalf of Bibelskapet are marked by * in figure 3.3. In these surveys the same questions were asked but unfortunately the surveys from 1985, 1992 and 2002 were not possible to access, according to KIFO. When looking at changes the available data are used. For the rest of the study, data are provided by 'Tro- og livssynsundersøkelsen (TLU)' from 2012 made by Norstat on behalf of KIFO. As well as the survey done in 2016 by TNS Gallup on behalf of Bibelselskapet. The TLU survey yields a lot of information about Norwegians religious attitudes and practices, as well as having a high number of respondents.

The data actually used in Oslo Monitor 1.0 to state that 11 percent of the population reads the Bible once a week or more is shown in figure 3.4. The data originate from TNS Gallup's survey from 2009. Total number of respondents in Norway were $n = 1000$. Figure 3.4 shows the different levels of Bible reading which exclude each other. Hence, an appropriate sample space for the level of Bible reading is: $\Omega_X = \{$'Never', 'Not so often', 'Some times a year', 'Once a month', 'Once a week', 'Every day'$\}$. The number of individuals for each outcome, $\mathbf{n} = (n_1, ..., n_6)$, takes on a corresponding

Navn	År	Tilgang til rådata	Antall respondenter
TNS Gallup	1972	Ja	1630
TNS Gallup*	1985	Nei	-
TNS Gallup*	1987	Ja	1001
Opinionen	1989	Ja	611
ISSP	1991	Ja	1506
TNS Gallup*	1992	Nei	-
Opinionen	1992	Ja	1015
ISSP	1998	Ja	1532
TNS Gallup*	2002	Nei	-
Skandinavisk Bibelbarometer	2009	Nei	-
TLU (Norstat)	2012	Ja	4001
Infact	2014	Nei	-
TNS Gallup*	2016	Ja	1070

* Undersøkelser utført av TNS Gallup på vegne av Det norske Bibelselskap

Figure 3.3: An overview of surveys done regarding the use of a Bible (Rafoss, 2017).

multinomial distribution, $p(\mathbf{n}; n, \mathbf{p})$, where $\mathbf{p} = (p_1, ..., p_6)$. The total population of Oslo in 2009 were $N = 575475$ (Statistikkbanken, 2019). N_k denotes the number of people that belongs to the corresponding level of Bible reading in all of Oslo.

We estimate each p_k and predict their corresponding N_k by their approximated 95% prediction interval, PI_{N_k}, given in equation (2.1), for $k = 1, ..., 6$,:

1. $\hat{p}_1 = 0.51, \quad \text{PI}_{N_1} = [293492 \pm 18194]$,
2. $\hat{p}_2 = 0.24, \quad \text{PI}_{N_2} = [138114 \pm 15544]$,
3. $\hat{p}_3 = 0.10, \quad \text{PI}_{N_3} = [57548 \pm 10919]$,
4. $\hat{p}_4 = 0.04, \quad \text{PI}_{N_4} = [23019 \pm 7132]$,
5. $\hat{p}_5 = 0.06, \quad \text{PI}_{N_5} = [34529 \pm 8644]$,
6. $\hat{p}_6 = 0.05, \quad \text{PI}_{N_6} = [28774 \pm 7932]$.

There is reason to question the numbers in figure 3.4 as they are from TNS Gallup's survey from 2009; 10 years ago. Oslo Monitor 1.0 states that the amount of people owning and reading the Bible has decreased. The amount of people reading the Bible once a week or more is probably even less today. This might be explained by the secularization taking place in the Christian communities.

	Norge			Sverige			Danmark		
	Total	Menn	Kvinner	Total	Menn	Kvinner	Total	Menn	Kvinner
Hver dag	5	4	5	1	1	1	2	2	2
En gang i uken	6	5	7	1	1	2	3	2	4
En gang i måneden	4	4	4	4	3	5	3	3	3
Noen ganger i året	10	10	11	11	10	11	10	9	10
Sjeldnere	24	24	23	34	34	34	44	41	46
Aldri	51	53	50	48	49	47	38	42	34
Vet ikke	0	0	0	1	1	1	0	0	0
Totalt	100	100	100	100	99	101	100	99	99
N	1000	508	492	1000	475	525	1001	492	509

Figure 3.4: Frequency by proportions of Bible reading in Scandinavia (Rafoss, 2017).

Challenges in Churches

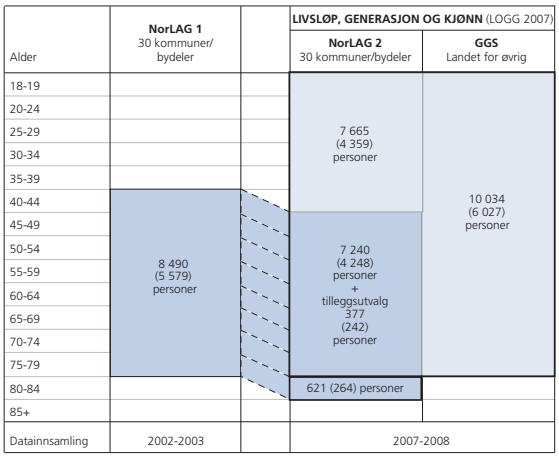
A number of 164 church leaders were asked to respond to a questionnaire in January 2017; whereas 31 replied. As stated in Oslo Monitor the answers are not representative for all churches in Oslo. The results used in Oslo Monitor are only the church's priorities the coming years and what they think of as the most critical social needs in Oslo. The more quantitative responses are few, as well as lacking information.

Conclusion

The data presenting the spiritual situation in Oslo demonstrate that in the last decades Norway has experienced a secularization. There are far less people believing in God today than 30 years ago. Still, the amount of people regarding themselves as personal Christians has increased. Especially among young people. Church attendance is decreasing and per January 2018 only 4.8 percent of Oslo's population engaged in churchly activities on a regular basis. The gap between church attendance and personal Christians is explained by the likely fact that a lot of personal Christians are not involved in a Christian community. The proportion of people in possession of and reading the Bible has decreased.

3.2 Social Suffering

To depict social suffering the factors loneliness, child neglect, child poverty, life expectancy, divorce and social differences measured by disability, income and education are considered. The data originate from different studies, as well as the municipality of Oslo and SSB.



[1] Størrelsen på et *utvalg* er antall personer i en befolkningsgruppe som er trukket ut for å intervjues. *Brutto-utvalget* er de vi sitter igjen med etter at det er tatt hensyn til at enkelte er utvandret eller døde etter at utvalget ble trukket. *Nettoutvalget* (i parentes) er antall gjennomførte telefonintervjuer etter frafall, fordi noen personer ikke ville la seg intervjue eller det ikke ble oppnådd kontakt med dem.

(a) Overview of selections made for LOGG 2007 (Tønder, 2009).

Variabler	Prosent en-somme	Antall personer totalt
Alle	**21,2**	**15 048**
Kjønn		
Menn	18,0	7 414
Kvinner	24,1	7 637
Alder		
18-29 år	22,7	2 552
30-39 år	19,0	3 067
40-49 år	18,5	2 895
50-59 år	19,4	2 675
60-69 år	22,5	2 267
70-79 år	27,0	1 269
80 år og over	31,7	325

(b) Amount of people regarding themselves as lonely at least occasionally (Tønder, 2009).

Figure 3.5: The data used to account for loneliness in Oslo Monitor 1.0.

Loneliness

Loneliness tells us to what degree a person finds themself lonely.

Oslo Monitor 1.0 refers to a journal called 'Samfunnsspeilet', published by SSB in 2009 (Tønder, 2009), to account for loneliness. The journal is based on data from the research called 'Studien av livsløp, generasjon og kjønn (LOGG)' from 2007. LOGG 2007 is a national research done by SSB and NOVA and consists of the international study 'Generations and Gender Survey (GSS)' and the second round of the Norwegian study 'Livsløp, aldring og generasjon (NorLAG)'.

In GGS a representative sample of men and women were used. The same person was interviewed by three years apart each time; called a longitudinal study. NorLAG is also a longitudinal study. First amount of data were collected in 2002-2003 by interviewing 5559 persons between 40 and 79 years old. The same persons did participate when collecting data for LOGG in 2007; also being the second round of NorLAG. The selection is from 30 local communities from Agder, Oslo and Akershus, Nord-Trøndelag and Troms. LOGG collected their data through phone, questionnaires by mail and records. The total base of data is complex since two different studies were merged. In figure 3.5a there is an overview of the selections made for LOGG 2007. They ended up getting responses from 43.2 percent of the gross sample, which is a low response rate. They therefore had to weight the numbers to get representative results for the whole country. Already existing records also contributed with important data.

The data used in Oslo Monitor to state that more than every fifth Norwegian feels lonely are shown in figure 3.5b. The total number of respondents were $n = 15048$. The question asked is whether they find themselves lonely at least occasionally, or not. The

sample space is therefore given by: $\Omega_X = \{$'Do feel lonely some times', 'Do not feel lonely some times'$\}$. The number of individuals that occasionally finds themselves lonely, n_1, takes on a binomial distribution, $p(n_1; n, p)$. The total population of Oslo in 2007 was $N = 548617$ (Statistikkbanken, 2019). N_1 denotes the number of people that occasionally finds themselves lonely in all of Oslo.

We estimate p and predict N_1 by its approximated 95% prediction interval, PI_{N_1}, given in equation (2.1):

$$\hat{p} = 0.212, \quad \text{PI}_{N_1} = [116307 \pm 3656].$$

Every fifth person in Oslo feels lonely some times. This is a high amount. Loneliness is an important factor to take into consideration when mapping the social suffering in Oslo.

Child Neglect

According to NSPCC (2007) child neglect is defined as "the persistent failure to meet a child's basic physical and/or psychological needs resulting in serious impairment of health and/or development". The Child Welfare is involved to help the child whenever such a case is uncovered.

The data used in Oslo Monitor 1.0 are numbers of children that received help from the Child Welfare in 2015. The data are gathered from the municipality of Oslo; from its 'child welfare statistics' in their 'bank of statistics'. The exact data are picked from 'Bydelsstatistikken 2015' (Oslo Kommune, 2015) and tell us that the number of children that received help from the Child Welfare in Oslo municipality in 2015 were 5684. Among these cases some are so serious that the child is in need of a foster care. In 2015 this was the case for 941 children in Oslo.

There were $n = 127639$ children under the age of 18 in Oslo in 2015 (SSB, 2015). A person could either have experienced child neglect or not. The sample space is given by: $\Omega_X = \{$'Child neglected', 'Not child neglected'$\}$. The number of individuals that have experienced child neglect, n_1, follows a binomial distribution: $p(n_1; n, p)$. The total number of people in Oslo in 2015 were $N = 647676$ (Statistikkbanken, 2019). N_1 denotes the number of people that have been neglected as a child in all of Oslo.

We estimate p and predict N_1 by its approximated 95% prediction interval, PI_{N_1}, given in equation (2.1):

$$\hat{p} = 0.04, \quad \text{PI}_{N_1} = [25907 \pm 710].$$

The numbers are alarming. Hence, it is important to look further into how the variable interacts with other variables. Child neglect may have disturbing consequences for the social situation of a person.

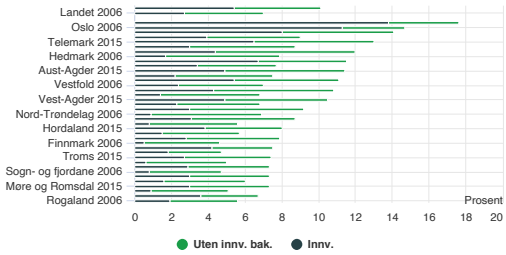
Child Poverty

Child poverty is here meant by children growing up in a household with persistent low income. Oslo Monitor 1.0 states that 17.6 percent of all children in Oslo grew up in households with persistent low income in 2015. The data represented in figure 3.6a are retrieved

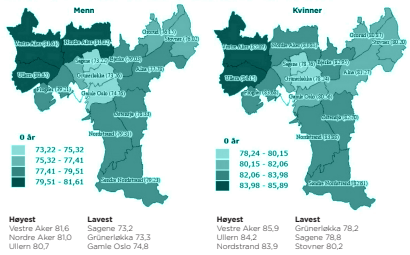
from SSB (Epland and Kirkeberg, 2017) and shows the regions with the highest proportions. Still, the original data set includes all of the regions of Norway. There were 98200 children in households with persistent low income in all of Norway in 2015. They account for 10 percent of the children under 18 years old. In addition the proportion has increased the last years, especially among children from an immigrant family. The proportions of children either from a Norwegian family or an immigrant family is presented in figure 3.6a.

Life Expectancy

Life expectancy reflects the quality of the populations health. It is defined for a given year as the expected lifetime of a child born the given year. An important assumption is that the rate of death will be constant in the future. This is of course not the case in real life (FHI, 2018). In Oslo Monitor 1.0 the life expectancy by birth is presented by the difference from 2006 to 2010 in the different districts of Oslo. This is shown in figure 3.6b. The data come from 'Samfunnsspeilet' published by SSB in 2013 (Nørgaard, 2013). But data collected on change of life expectancy over time are hard to compare because of variation in methods used. Another factor is that the data used are based on both three and five vintages of death. Still, the average life expectancy of a person might tell us a lot about the social situation of that person. Especially, when combined with factors like loneliness, disability benefits, education and income.



(a) Amount of children, with and without immigrant background, in households with persistent low income by county (Epland and Kirkeberg, 2017).



(b) Change from 2006 to 2010 in life expectancy by birth in the different districts of Oslo (Talset, 2018).

Figure 3.6: Data concerning poverty and life expectancy.

Divorce

We consider the number of children, under the age of 18, experiencing their parents getting divorced in 2015. The number were 8743. This number is collected from a SSB report (SSB, 2018b). The data originate from information about parents and their relation to their children from 'Det sentrale folkeregister' (DSF). But the data from DSF do not take into consideration the children with parents living in a cohabitation experiencing their parents leaving each other. Most likely the number of children experiencing their parents splitting up is a lot higher.

Disability

Oslo Monitor 1.0 includes three graphs concerning social differences, presented in figure 3.7. The first graph in figure 3.7a models the disability benefits among people in the different districts of Oslo. By disability benefits it is meant that a person receives financial support because their ability to make an income is permanently reduced because of sickness or an injury (NAV, 2019). Hence, by disability it is meant that a person is permanently reduced because of the reasons just mentioned.

The data are collected from the municipality of Oslo's 'bank of statistics' (Statistikkbanken, 2019) but the data originate from PESYS/NAV and SSB. The data are both from 2010 and 2016.

In 2016 there were 24014 registered persons that received disability benefits in Oslo. Thus, n is unknown. Nevertheless, we assume that: $\Omega_X = \{$'Disabled', 'Not disabled'$\}$. The number of people that are disabled, n_1, follows a binomial distribution, $p(n_1; n, p)$. The total population of Oslo above the age of 18 in 2016 were $N = 546536$. Furthermore, N_1 denotes the number of people in all of Oslo that because of sickness or an injury are permanently not able to make an income.

We estimate p and predict N_1 but cannot calculate the corresponding approximated 95% prediction interval of N_1, because n is unknown:

$$\hat{p} = 0.05, \quad \hat{N_1} = 27327.$$

The number of persons that received disability benefits in each district of Oslo in 2016 might reflect the magnitude of people dealing with challenges caused by their childhood. At the same time the number also includes persons that all of a sudden become ill or injured.

Income

The data for the average income of the different districts of Oslo from 2010 and 2014 are presented in figure 3.7b. The data are collected from the municipality of Oslo's 'bank of statistics' (Statistikkbanken, 2019) but the data originate from PESYS/NAV and SSB. In the 'bank of statistics' it is possible to get the same data for groups of different levels of income instead.

The size of the population in the data set is $n = 283986$. We assign a sample space for the income levels given, in thousands, by: $\Omega_X = \{$'0-199', '200-399', '400-599', '600-799', '800+'$\}$. The number of individuals for each outcome, $\mathbf{n} = (n_1, ..., n_5)$, takes on a corresponding multinomial distribution, $p(\mathbf{n}; n, \mathbf{p})$, where $\mathbf{p} = (p_1, ..., p_5)$. The total population of Oslo in 2014 were $N = 634463$ (Statistikkbanken, 2019). N_k denotes the number of people with the corresponding income level in all of Oslo.

We estimate each p_k and predict their corresponding N_k by their approximated 95% prediction interval, PI_{N_k}, given in equation (2.1), for $k = 1, ..., 5$:

1. $\hat{p}_1 = 0.17, \quad \text{PI}_{N_1} = [107859 \pm 894]$,
2. $\hat{p}_2 = 0.22, \quad \text{PI}_{N_2} = [139582 \pm 986]$,
3. $\hat{p}_3 = 0.30, \quad \text{PI}_{N_3} = [190339 \pm 1091]$,

4. $\hat{p}_4 = 0.15, \quad \text{PI}_{N4} = [95169 \pm 850]$,

5. $\hat{p}_5 = 0.16, \quad \text{PI}_{N_5} = [101514 \pm 873]$.

The lowest income possible is actually the lowest disability benefit available. According to NAV (2018) this is 2.28 times the absolutely lowest amount called 1G. According to Skatteetaten (2019), the value of 1G in 2014 was 88370 NOK, while in 2019 it was 99858 NOK. Hence, the lowest possible income in 2014 was 201484 NOK and in 2019 it was 227676 NOK. This indicates that the amount of people accounting for the proportion of people with an income under 200000 most likely are children, youth and young adults still under education. They neither earn their own money or receive disability benefits.

Education

Education is measured by what kind of school a person has completed, resulting in different levels of education. The two highest levels of education are 'University - Lower level' and 'University - Higher level'. According to Statistikkbanken (2019) the lower level includes 4 years of a completed degree at a university, while the higher level covers completed degrees above 4 years, as well as researchers. The levels are presented in figure 3.7c for the different districts of Oslo.

The data are collected from the municipality of Oslo's 'bank of statistics' but the data originate from PESYS/NAV and SSB (Statistikkbanken, 2019).

The size of the population in the data set is $n = 553365$. From the data in figure 3.7c we decide on a sample space for the levels of education: $\Omega_X = \{$'Not applicable', 'Elementary School', 'High School', 'University - Lower Level', 'University - Higher Level'$\}$. The number of individuals for each outcome, $\mathbf{n} = (n_1, ..., n_5)$, takes on a corresponding multinomial distribution, $p(\mathbf{n}; n, \mathbf{p})$, where $\mathbf{p} = (p_1, ..., p_5)$. The total population of Oslo in 2017 were $N = 666759$ (Statistikkbanken, 2019). N_k denotes the number of people in all of Oslo with the corresponding level of education.

We estimate each p_k and predict their corresponding N_k by their approximated 95% prediction interval, PI_{N_k}, given in equation (2.1), for $k = 1, ..., 5$:

1. $\hat{p}_1 = 0.02, \quad \text{PI}_{N_1} = [13335 \pm 251]$,

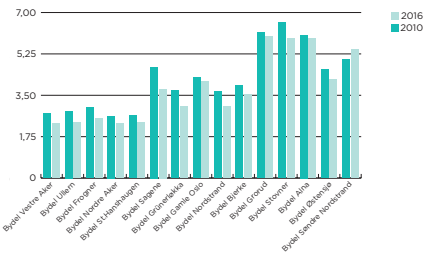
2. $\hat{p}_2 = 0.20, \quad \text{PI}_{N_2} = [133352 \pm 717]$,

3. $\hat{p}_3 = 0.28, \quad \text{PI}_{N_3} = [186693 \pm 805]$,

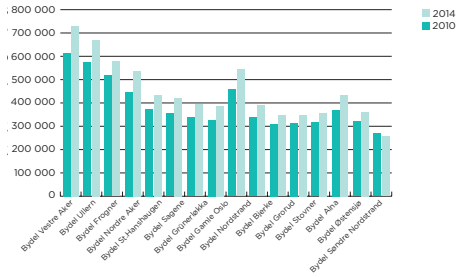
4. $\hat{p}_4 = 0.30, \quad \text{PI}_{N_4} = [200028 \pm 821]$,

5. $\hat{p}_5 = 0.20, \quad \text{PI}_{N_5} = [133352 \pm 717]$.

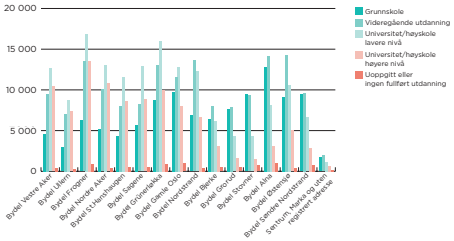
Education is an important variable to include in a model accounting for the social situation of a person. In general your education is the foundation on which your carrier is built in Norway.



(a) Percentages of people getting disability benefits in the different districts of Oslo (Talset, 2018).



(b) Average income in the different districts of Oslo (Talset, 2018).



(c) Levels of education in the different districts of Oslo (Talset, 2018).

Figure 3.7: Data regarding social differences.

Conclusion

Oslo Monitor 1.0 states that at any time 20 percent of the population find themselves lonely. Loneliness is here correlated to a low social support. On the contrary, a social network encouraging mutual commitment antagonizes loneliness. Child neglect is widely common in Oslo. In addition, 17.6 percent of all children in Oslo are raised in a home of low income. The differences are huge among the districts in the east and the west of the city. Not only do the eastern districts have a lower mean income and level of education but also the expected life span is lower. Divorce is also widely common; not only in Oslo but all of Norway. 8743 children in Norway, before turning 18 years old, experience their parents getting a divorce.

3.3 Cultural Challenges

The cultural challenges monitor mindsets and attitudes towards the following spheres of society: Media, high school drop-outs, illegal employment and volunteering. The data are mostly from SSB.

Media

The use of religious words in media has increased and is more frequent than 10 years ago. 'Tro', or 'faith', was written about 174.000 times in 2016 and 159.000 in 2007 (Talset, 2018). As Oslo Monitor 1.0 states, there is no way to tell if the word 'faith' is actually used in a non-religious setting. Still, the trend is the same when looking at words like 'church', 'God', 'Christians' and 'Jesus'.

High School Drop-outs

High school drop-outs are people not completing high school within five years.

SSB (2018e) is used as source to account for the amount of high school drop-outs. Oslo Monitor 1.0 states that 27 percent of those starting high school do not complete within five years. There are more boys (61 percent) than girls (39 percent). Of the high school drop-outs 74 percent followed the 'yrkesfaglig', or technical, study program and 26 percent followed the 'studieforberedende', or academical, study program. The data are from the period 2010-2015 (Talset, 2018).

The number of students that started high school in 2010 and finished it in 2015 or earlier is $n = 63837$. The question considered is whether a high school student finishes or not within five years. An appropriate sample space is given by: $\Omega_X = \{$'Did not finish high school within five years', 'Did finish high school within five years'$\}$. The number of individuals that did not finish high school within five years, n_1, follows a binomial distribution, $p(n_1; n, p)$. The total population of Oslo in 2010 were $N = 586860$ (Statistikkbanken, 2019). Furthermore, N_1 denotes the number of high school drop-outs in all of Oslo.

We estimate p and predict N_1 by its approximated 95% prediction interval, PI_{N_1}, given in equation (2.1):

$$\hat{p} = 0.27, \quad \text{PI}_{N_1} = [158452 \pm 2062]. \tag{2.1}$$

There might be a lot of factors involved to why a youth decides to drop out of high school. Factors like child neglect and parents receiving disability benefits are assumed to be correlated to drop-outs. Also their parents educational level might play a significant role, as well as their relationship to their parents.

Illegal Employment

The distribution of illegal employment accounts for almost 15 percent of the turnover in Norway and is estimated to be 430 billion NOK a year by Skatteetaten (Talset, 2018). But it follows a huge amount of uncertainty to such a number, obviously. Also, the question whether a person take advantage of illegal employment or not will not present the true distribution.

Volunteering

Approximately 5 percent of the economic value in Norway in 2014 were added by volunteering work (Talset, 2018). Oslo Monitor 1.0 continues to account for who finances the volunteering work; where 43 percent comes from private households. The data are from

'Satelittregnskap for ideelle og frivillige organisasjoner' by SSB from 2014 (SSB, 2018g). The method used to estimate the numbers presented by SSB is assumed to cover the population as a whole within each area of activity defined by UN's standard for classification (ICNPO): Culture, education, health, social services, conservation, local communities, political organisations, centers for volunteering, international organisations, religion and labor unions. The estimated value of the number of non-payed volunteering work units is found from assuming expenses per 'årsverk', or 'annual amount of work', in NOK to be the same as for regular work within the different areas of activities. Then by dividing the total expenses of non-payed work in NOK on the relevant 'annual amount of work'. The number of non-payed volunteering work units within each area of activity reflects the volunteering mindset that most Norwegians share.

Conclusion

From their analysis of cultural challenges there are 27 percent of Norwegian teenagers who do not finish high school within five years. This often results in an unstable prospect for their future. This raises the question if there should be offered other alternatives than attending high school. Another big challenge is the distribution of illegal employment accounting for almost 15 percent of the turnover in Norway. The attitude towards this kind of employment is the first thing that needs to change, to fight this. To care about the society beyond an organization's primarily goal, often in terms of maximizing the return, will be more and more important in the future. The volunteering aspect of Norwegian society contributes with 72 billion Norwegian crowns a year.

Chapter 4

Population of Lilleby

Different factors within the spiritual situation, social suffering and cultural challenges are introduced in the previous chapter. The factors make up variables that are interesting to include in an interaction model to describe the population of Lilleby. Still, some modifications are made as to simplify the interaction model.

The theory behind pair-copula constructions is introduced and proposes a method to build an interaction model. A simplified version of the method is derived for categorical variables with either a binomial or multinomial distribution.

The population of Lilleby with all their characteristics is made by simulating a realization of residents using the interaction model. Estimates of the marginal probabilities are compared to their respective marginal probabilities. The degree of dependence in bivariate characteristics of the population is visualized by two-dimensional biplots.

4.1 Description of Lilleby

'Spiritual situation' is modified to only include the following attitudes towards religion: 'Belief in God' and 'Personal Christian'. 'Church attendance last 12 months' is removed as 'Active church attendee', renamed as 'Church activity', is more interesting to include in the model. 'Bible usage' is still included.

'Social suffering' is renamed to 'Social background' and 'Cultural challenges' is removed but 'High school drop-outs' is included in 'Education'. The following variables are removed from 'Social background': 'Child poverty', 'Life expectancy' and 'Divorce'. Hence, 'Social background' includes the following variables: 'Child neglect', 'Disability', 'Education', 'Income' and 'Loneliness'.

In addition some 'General characteristics' are included with their marginal distributions and parameter estimates. They are 'District', 'Cultural origin', 'Age', 'Marital status', 'Gender' and 'Number of children'.

The resulting interaction model accounts for interplay among the introduced variables for the spiritual situation and the social background, as well as variables for the more general characteristics now to be introduced.

4.1.1 General Characteristics of Lilleby

The general characteristics for a person in Lilleby are district, cultural origin, age, marital status, gender and number of children. The assumptions regarding the marginal distributions of the general characteristics are inspired by data retrieved from SSB (2019a) and Statistikkbanken (2019).

District

Lilleby consists of four districts called 'West', 'South', 'North' and 'East'. Some assumptions are made regarding each of the districts. In the north the majority of the population is assumed to be students and young adults while the south is dominated by families with children at home. The western part of the city inhabits rich and elderly people while the majority of the eastern part is poor people and immigrants.

An appropriate sample space for the district, X, of a person is given by: $\Omega_X = \{$'West', 'South', 'North', 'East'$\}$. The size of each district is inspired by the data from the inner part of Oslo in 2018 (Statistikkbanken, 2019). Amount of people living in each district follows a multinomial distribution with corresponding parameters, $\mathbf{p} = (p_1, ..., p_4)$.

The relevant parameter estimates for the proportion of persons living in the different districts of Lilleby are found by their respective MLE:

$$\hat{p}_1 = 0.25, \quad \hat{p}_2 = 0.15, \quad \hat{p}_3 = 0.25, \quad \hat{p}_4 = 0.35.$$

Cultural Origin

A persons cultural origin is hard to define. Both religion and economy is taken into account, in our case. Religion is decided to be the most common religion or belief system in the given part of the world. While economy tells to what extend a person is assumed to manage the Norwegian job market. On behalf of religion and economy the different countries of origin are categorized into the following groups of origin:

1. North of Europe including Norway and the Nordic countries as well as North America and Oceania where the common religion is Lutheran Christianity and the people do have an easy approach to the Norwegian job market.

2. Middle of Europe, South of Europe and South America where the common religion is Catholicism and the people do have a similar approach to the Norwegian job market as the first group.

3. East of Europe and North of Asia including Russia among others where the common religion is Orthodox Christianity and the people do often enter the market for craftsmen.

4. The Far East including China, Korea, Japan, India and Thailand among others where the common religions are Buddhism and Hinduism and the people do not enter the Norwegian job market as easy as the first two groups.

5. Africa and The Near East including The Middle East, Iran, Afghanistan and Turkey among others where the common religion is Islam and the majority are refugees with a challenging time entering the Norwegian job market.

An appropriate sample space for the origin, X, of a person is given by: $\Omega_X = \{\text{'Origin 1'}, \text{'Origin 2'}, \text{'Origin 3'}, \text{'Origin 4'}, \text{'Origin 5'}\}$. The corresponding estimated parameters for the proportion of persons having the different origins in Lilleby are inspired by the data from SSB (2018a). They are found by their respective MLE for the multinomial distribution:

$$\hat{p}_1 = 0.91, \quad \hat{p}_2 = 0.02, \quad \hat{p}_3 = 0.04, \quad \hat{p}_4 = 0.01, \quad \hat{p}_5 = 0.02.$$

Age

The population of Lilleby is divided into five different age groups inspired by the data from SSB (2018c). Children under the age of 15 are not included. The five groups are youths at high school (15-19 years old), young adults with children at home (20-39 years old), adults with youths at home (40-49 years old), adults still working but no kids at home (50-69 years old) and retired, elderly people (70+ years old).

An appropriate sample space for the age group, X, of a person is given by: $\Omega_X = \{\text{'Age 15-19'}, \text{'Age 20-39'}, \text{'Age 40-49'}, \text{'Age 50-69'}, \text{'Age 70+'}\}$. The corresponding estimated parameters for the proportion of persons from the different age groups in Lilleby are found by their respective MLE for the multinomial distribution:

$$\hat{p}_1 = 0.07, \quad \hat{p}_2 = 0.33, \quad \hat{p}_3 = 0.17, \quad \hat{p}_4 = 0.28, \quad \hat{p}_5 = 0.15.$$

Marital Status

According to SSB (2018f) it is reasonable to define a persons marital status as either single, married or living in a cohabitation. An appropriate sample space for the marital status, X, of a person is given by: $\Omega_X = \{\text{'Single'}, \text{'Cohabitation'}, \text{'Married'}\}$. The corresponding estimated parameters for the proportion of persons with different marital status in Lilleby are found by their respective MLE for the multinomial distribution:

$$\hat{p}_1 = 0.40, \quad \hat{p}_2 = 0.20, \quad \hat{p}_3 = 0.40.$$

Gender

Data from SSB (2018c) state that gender, X, follows a binomial distribution. The sample space is therefore given by: $\Omega_X = \{\text{Male}, \text{Female}\}$. The relevant parameter estimates of $\mathbf{p} = (p_1, p_2)$ are found from the MLE for the binomial distribution. The estimated proportion parameters for a person being a male or a female is, respectively:

$$\hat{p}_1 = 0.50, \quad \hat{p}_2 = 0.50.$$

Number of Children

Data from SSB (2019b) show the proportions of persons having different number of children under the age of 18. Either a person has no children, one child, two children or three children or more. An appropriate sample space for the number of children, X, is given by: $\Omega_X = \{$'No children', '1 child', '2 children', '3 children or more'$\}$. The corresponding estimated parameters for the proportion of persons in Lilleby with the different amounts of children are found by their respective MLE for the multinomial distribution:

$$\hat{p}_1 = 0.74, \quad \hat{p}_2 = 0.11, \quad \hat{p}_3 = 0.11, \quad \hat{p}_4 = 0.04.$$

4.1.2 Interaction Model

Each variable, $X_i \in \Omega_{X_i}$, now have a fully specified and fixed marginal distribution; $p(x_i); i = 1, ..., n$, where n is the total number of included variables in the interaction model. Hence, $\mathbf{X} = (X_1, ..., X_n)$ has a multivariate probability mass function (pmf) defined by $p(\mathbf{x})$ where $\mathbf{X} \in \Omega_{\mathbf{X}}$ is given by: $\Omega_{\mathbf{X}} = \Omega_{X_1} \times \Omega_{X_2} \times \ldots \times \Omega_{X_n}$. The multivariate sample space is fully specified as long as no dependence is assumed between the variables. The goal is to create an interaction model that accounts for the interplay among the $\mathbf{X}$'s. Hence, the challenge is to describe a fully specified sample space for $\mathbf{X} \in \Omega_{\mathbf{X}}$ such that $p(\mathbf{x})$ given $p(x_i); i = 1, ..., n$, is fully specified when dependence is assumed between the X_i's. This is necessary to be able to sample from $p(\mathbf{x})$. The interactions are described and defined in the following by the use of conditional independence.

Interactions

The population of Lilleby is a simulated realization of $p(\mathbf{x})$ where each of the n_L residents are assigned a value for each of the following categorical variables:

C: General Characteristics	**U: Social Background**	**S: Spiritual Situation**
C_1: District	U_1: Child neglect	S_1: Belief in God
C_2: Cultural origin	U_2: Disability	S_2: Personal Christian
C_3: Age	U_3: Education	S_3: Church activity
C_4: Marital status	U_4: Income	S_4: Bible usage
C_5: Gender	U_5: Loneliness	
C_6: Number of children		

The simulated realizations are contained in a matrix of dimension $(n_c + n_u + n_s) \times n_L$, where n_c, n_u and n_s are the number of variables in **C**, **U** and **S**, respectively. The realizations for person i; where $i = 1, ..., n_L$, are collected in:

$$\mathbf{X}_i = [\mathbf{C}, \mathbf{U}, \mathbf{S}]_i = [C_1, ..., C_6, U_1, ..., U_5, S_1, ..., S_4]_i.$$

To describe the interactions between the different variables for the full matrix of realizations they are conditioned on each other as follows:

$$\mathbf{X} = [\mathbf{S} \mid \mathbf{U}, \mathbf{C}][\mathbf{U} \mid \mathbf{C}]\mathbf{C}. \tag{4.1}$$

Should this be fully written out for each C_i, U_j and S_k, for $i = 1, ..., n_c, j = 1, ..., n_u$ and $k = 1, ..., n_s$, they will all be conditioned on every preceding variable. But conditional independence is assumed, resulting in the following conditioned variables:

$$[\mathbf{C}] = C_1[C_2 \mid C_1][C_3 \mid C_2, C_1][C_4 \mid C_3][C_5 \mid C_4][C_6 \mid C_4, C_3, C_2],$$

$$\begin{aligned}[\mathbf{U} \mid \mathbf{C}] =& [U_1 \mid C_5, C_4][U_2 \mid U_1, C_3][U_3 \mid U_1, C_5, C_2][U_4 \mid U_3, C_3] \\ & [U_5 \mid U_2, C_5, C_3]\end{aligned}$$

and

$$[\mathbf{S} \mid \mathbf{U}, \mathbf{C}] = [S_1 \mid C_3, C_2][S_2 \mid S_1, C_4, C_2][S_3 \mid S_2, U_5][S_4 \mid S_2].$$

Interactions are defined as in the preceding and all the marginal distributions are specified and fixed for each $X_i \in \Omega_{X_i}$. The multivariate sample space for $\mathbf{X} = [\mathbf{S} \mid \mathbf{U}, \mathbf{C}][\mathbf{U} \mid \mathbf{C}]\mathbf{C} \in \Omega_{\mathbf{X}}$ is to be specified, such that a $p(\mathbf{x}) = p(\mathbf{s} \mid \mathbf{u}, \mathbf{c})p(\mathbf{u} \mid \mathbf{c})p(\mathbf{c})$ can be specified as well and used for simulation. A method based on the theory of pair-copula constructions is now developed.

Theory of Pair-Copula Constructions

Multivariate models are used to describe the interaction between a multiple of variables. The objective is hence to describe more than just the marginal properties of a variable. Pair-copula constructions (PCCs) discussed by Haff (2012) are frequently used. The idea behind PCCs is to build a multivariate copula from bivariate copulas.

A copula is used to model dependence. According to Cont and Tankov (2004), a d-dimensional copula is a function, C, with domain $\mathbb{R}^d_{[0,1]}$ such that:

1. C is grounded and d-increasing,
2. C has margins $C_l, l = 1, 2, ..., d$, which satisfy $C_l(u) = u$ for all $u \in \mathbb{R}_{[0,1]}$.

We focus on $d = 2$ because this is the case for the pair-copula constructions. Let S_1 and S_2 be two possible infinite closed intervals. The bivariate copula, C, with domain $S_1 \times S_2$ is said to be grounded if for every $x \in S_1$, $C(x, \min S_2) = 0$ and for every $y \in S_2$, $C(\min S_1, y) = 0$. Let $x_1, x_2 \in S_1$ where $x_1 \leq x_2$ and $y_1, y_2 \in S_2$ where $y_1 \leq y_2$. Then C is 2-increasing if $C(x_2, y_2) - C(x_2, y_1) - C(x_1, y_2) + C(x_1, y_1) \geq 0$.

Sklar's theorem states that if F is a d-dimensional cumulative distribution function, with uniformly distributed marginal cdfs, $F_1, \ldots, F_d$, then there exists a copula C such that:

$$F(x_1, ..., x_d) = C(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)).$$

The copula, C, is unique for continuous distributions. This is not the case for discrete distributions. No theory could be found for the non-ordered categorical case as we deal with in the Lilleby study. Hence, we develop a copulae-inspired technique to include dependence in multivariate categorical variables, given their marginal distributions.

Derivation of the Interaction Model

The variables associated with each person in Lilleby and their interactions are defined in equation (4.1). Each x_i is assumed to be bi- or multinomially distributed with known, fixed marginal distribution, $p(x_i)$; for $i = 1, ..., n_x$, where $n_x = n_c + n_u + n_s$ are the total number of variables in **x**.

Consider $x \in \Omega_x$ and $y \in \Omega_y$ having k_x and k_y elements, respectively, also being the number of possible outcomes for x and y. For y it may consist of a subset of the sample space, Ω_y. Note that the subset is not explicitly reflected in the notation.

Consider the known and fixed probabilities for x and y from the already defined marginal distributions. For x they are defined as p_i^x for $i = 1, ..., k_x$, while for y the probabilities are p_j^y for $j = 1, ..., k_y$. The interaction between x and y is defined by their bivariate distribution, $p(x, y)$. Since they are both categorical they are increasingly ordered on the axis of the corresponding bivariate matrix such that the pairs $(x_1, y_1), ..., (x_{k_x}, y_{k_y})$ are assumed to have increasing positive dependence. Each element of the bivariate matrix, assuming independence, has joint distribution $p(x_i, y_j) = p_i^x p_j^y$ for the given pair (x_i, y_j).

Dependence is created by adding the set of $\alpha_{ij} \in \mathbb{R}$ for $i = 1, ..., k_x, j = 1, ..., k_y$ to the corresponding element of the bivariate matrix. The (i, j)-th element of the bivariate matrix is hence given by:

$$p(x_i, y_j) = p_i^x p_j^y + \alpha_{ij}; \quad i = 1, ..., k_x, \quad j = 1, ..., k_y.$$

Note that for $p(x, y)$ to be a valid bivariate pmf reproducing the marginals, $p_i^x, i = 1, ..., k_x$, and $p_j^y, j = 1, ..., k_y$, the following must hold:

$$\sum_i \alpha_{ij} = 0, \quad \sum_j \alpha_{ij} = 0$$

and

$$0 \leq p_i^x p_j^y + \alpha_{ij} \leq 1; \quad \forall \quad i = 1, ..., k_x, \quad j = 1, ..., k_y. \tag{4.2}$$

The written-out bivariate matrix is:

$$\begin{array}{c} \\ p_{k_x}^x \\ p_{k_x-1}^x \\ \vdots \\ p_2^x \\ p_1^x \end{array} \begin{array}{c} \begin{array}{ccccc} p_1^y & p_2^y & & \cdots & p_{k_y-1}^y & p_{k_y}^y \end{array} \\ \begin{pmatrix} & \cdots & \cdots & \cdots & \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ \vdots & \cdots & p_i^x p_j^y + \alpha_{ij} & \cdots & \vdots \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ & \cdots & \cdots & \cdots & \end{pmatrix} \end{array}, \tag{4.3}$$

where k_x and k_y are number of outcomes in Ω_x and Ω_y, respectively, and α_{ij} is a measure of deviance from independence.

We make a procedure to adjust the inital bivariate matrix by adding α_{ij} to each corresponding element. Give each α_{ij} a positive or negative integer, b_{ij}, multiplied with the

unit u. Hence, $\alpha_{ij} = b_{ij}u$, where $b_{ij} \in \mathbb{N}$ for $i = 1, ..., k_x, j = 1, ..., k_y$ and $u \geq 0$. Specify all b_{ij} such that $\sum_i b_{ij} = \sum_j b_{ij} = 0$ and such that each b_{ij} reflects the relative deviation from independence for the given matrix element. In order to fulfill equation (4.2), calculate the largest value u may have given that $\alpha_{ij} = b_{ij}u$. Keep in mind that each α_{ij} is assigned subjectively based on our understanding of the conditions of the relevant aspects included in **x** in the Norwegian society.

The bivariate matrix now defines the probabilities for the different outcomes of (x_i, y_i) which can be used to calculate the conditional probabilities we need to sample:

$$p(x_i \mid y_j) = \frac{p(x_i, y_j)}{p(y_j)} = \frac{p_i^x p_j^y + \alpha_{ij}}{p_j^y}, \quad \forall \quad i = 1, ..., k_x, \quad j = 1, ..., k_y. \tag{4.4}$$

The bivariate matrix can be thought of as a categorical bivariate copula used as a building-block to construct a multivariate copula (Haff, 2012). After sequentially repeating this procedure for all the conditioned variables in **x**, the outcome is a fully specified sample space for $\mathbf{x} \in \Omega_{\mathbf{x}}$. Hence, it is possible to sample from the joint pdf, $p(\mathbf{x})$, given the marginal pdfs, $p(x_i); i = 1, ..., n$. See algorithm 1.

Algorithm 1 Sampling from a bivariate joint distribution with categorical outcomes

1: Decide on dependencies to account for as done in (4.1)
2: Let the interaction be between y (known) and x (to be found)
3: **if** $x \mid y$ **then**
4: Decide on a natural ordering of y and x
5: Calculate their bivariate matrix as in (4.3)
6: **else if** $x \mid y_1, y_2, ...$ **then**
7: Decide on reasonable groups for the $k_{y_1} \times k_{y_2} \times ...$ possible combinations of $y_1, y_2, ...$
8: Sum the combinations within each group and let the groups be the new possible outcomes of y
9: Decide on a natural ordering of y and x
10: Calculate their bivariate matrix as in (4.3)
11: **end if**
12: Calculate the measure of dependence, α_{ij}
13: Weight the probability elements of the bivariate matrix to account for the amount of dependence between the different y's and x's
14: **for** $i = 1, 2, \ldots, n_L$ **do**
15: Given the already simulated outcome(s) of y_i, sample x_i from a multinomial distribution with the relevant conditional probability, $p_i^{x|y}$, given by (4.4)
16: **end for**
17: Repeat this procedure for the next categorical variable, x

4.2 The Realized Lilleby

The interaction model is used to simulate a realization of Lilleby residents with all their characteristics. Estimates of the marginal probabilities are compared to their respective marginal probabilities. The degree of dependence in bivariate characteristics of the population is visualized by two-dimensional biplots.

4.2.1 Simulation of Lilleby

The population of Lilleby with all their characteristics is made by simulating a realization of $n_L = 100000$ residents using the method summarized in algorithm 1. Hence, each person, i, of Lilleby posses all the variables within general characteristics, $\mathbf{c}_i$, social background, $\mathbf{u}_i$, and spiritual situation, $\mathbf{s}_i$, where $i = 1, ..., n_L$. The interactions between the variables, $\mathbf{x}_i = [\mathbf{s} \mid \mathbf{u}, \mathbf{c}]_i[\mathbf{u} \mid \mathbf{c}]_i\mathbf{c}_i$, are defined in equation (4.1).

The proportions for each variable outcome is an estimate of the marginal probabilities and they are compared to their respective marginal probabilities. The deviations are of course small since n_L is large.

We inspect Lilleby further by visualizing the degree of dependence in bivariate characteristics of the population by two-dimensional biplots. A contingency table is made to be able to visualize the degree of dependence in the bivariate characteristics of a given pair of categorical variables, x and y, for $x \neq y$. We make sure that the values follow a natural ordering. The contingency table is then used to compare the number of persons for each combination of x and y in the simulated realization to the corresponding probabilities assuming independence. The degree of dependence is hence the deviation from $p^{xy} = p^x p^y$ and is given by a deviation factor from independence:

$$\boldsymbol{\tau}^{xy} = \frac{\tilde{p}^{xy}}{p^{xy}},$$

where $\tilde{p}^{xy}$ is the fraction of persons for each combination of x and y in the simulated realization.

The bivariate matrix is visualized by its respective $\tilde{p}^{xy}$; being related to a bubble of corresponding size. The color of each bubble indicates the degree of dependence, hence deviation from p^{xy}. A yellow bubble is assigned whenever $\boldsymbol{\tau}^{xy} \approx 1$ and indicates weak dependence. The color strength of the bubbles vary with correlation strength, where red indicates negative dependence ($\boldsymbol{\tau}^{xy} < 1$) and green indicates positive dependence ($\boldsymbol{\tau}^{xy} > 1$). The biplots are made to assure that the dependence accounted for while sampling is reflected in the simulated realization.

4.2.2 Evaluation of Lilleby

Biplots are made of the interactions defined in equation (4.1) as well as of variables interacting through one or two steps of variables. Comments are made to what degree the visualized dependencies corresponds to the assumptions made on dependence when simulating Lilleby.

General Characteristics

We select some bivariate characteristics for the general characteristics. The given pairs of categorical variables are visualized as biplots.

In figure 4.1 the degree of dependence is visualized for 'Cultural origin' and 'District'. In the realization of Lilleby a resident of origin 5 (Africa and The Near East) is more likely to live in the east of Lilleby and less likely to live in the west. The assumption is confirmed in the biplot. The largest proportion of residents are of origin 1 (North America, Oceania and North of Europe including Norway) and the degree of dependence given each district is weak, yet reasonable. Figure 4.2 presents the degree of dependence for 'Age' and 'District'. The south is assumed to be dominated by families with children at home which corresponds to the strong degree of dependence in age 15-19 living in the south. The biplot also confirms that inhabitants in the west are more likely to be elderly people over the age of 70. Figure 4.3 is a biplot of 'Marital status' and 'Age'. Adults at age 50-69, where the majority is still working but have no kids living at home, are more likely to be married and less likely to live in cohabitation. At age 15-19 one is more likely to be single and not likely to be married at all. Young adults at age 20-39 are more likely to live in cohabitation. These assumptions are confirmed in the biplot. The biplot also shows that the youths at age 15-19 are more likely to live in cohabitation but the strong degree of dependence is questionable. Figure 4.4 confirms that a male is more likely to be single, while a female is more likely to be married. In figure 4.5 the degree of dependence for 'Number of children' and 'Age' is overall weak. But the biplot confirms that having three children or more is most unlikely for a youth at age 15-19. The degree of dependence between having one child and being at age 15-19 does not correspond to the assumptions made when sampling.

Figure 4.6 includes 'Gender' and 'Cultural origin'. They interact through two steps of variables. Naturally the degree of dependence is weak. In figure 4.7 the variables 'Number of children' and 'District' are interacting through one step of variables. The degree of dependence is weak, yet reasonable.

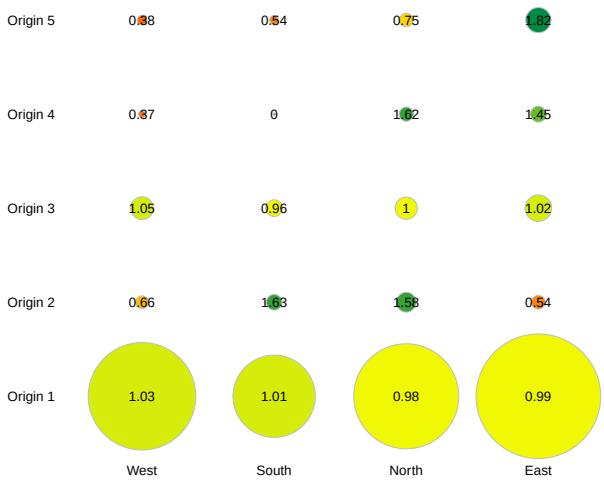


Figure 4.1: Degree of dependence for 'Cultural origin' and 'District', $\tilde{\rho}^{c_2 c_1}$.

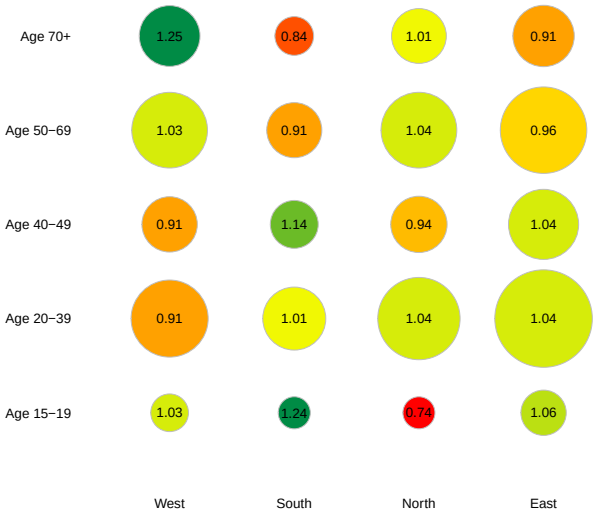


Figure 4.2: Degree of dependence for 'Age' and 'District', $\tilde{\rho}^{c_3 c_1}$.

Figure 4.3: Degree of dependence for 'Marital status' and 'Age', $\tilde{\boldsymbol{\rho}}^{c_4 c_3}$.

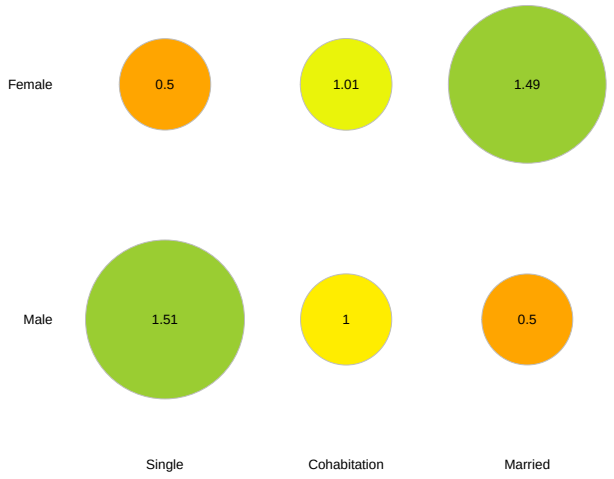


Figure 4.4: Degree of dependence for 'Gender' and 'Marital status', $\tilde{\boldsymbol{\rho}}^{c_5 c_4}$.

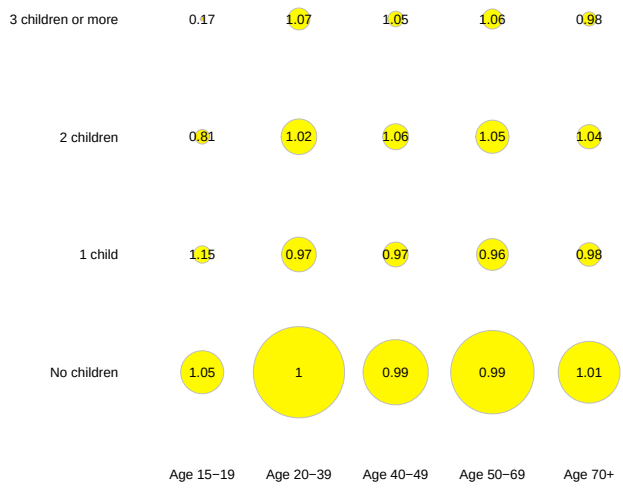


Figure 4.5: Degree of dependence for 'Number of children' and 'Age', $\tilde{\boldsymbol{\rho}}^{c_6 c_3}$.

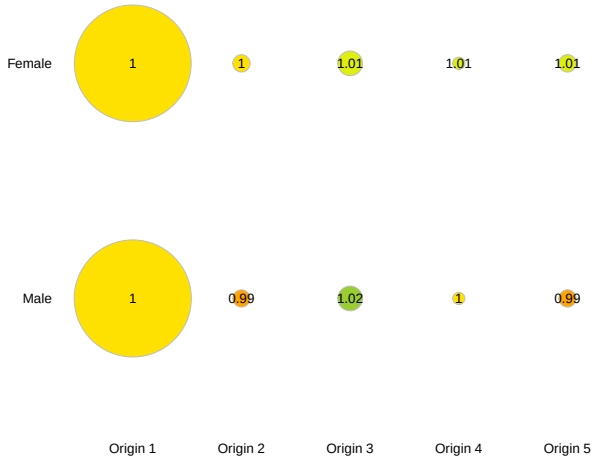


Figure 4.6: Degree of dependence for 'Gender' and 'Cultural origin', $\tilde{\boldsymbol{\rho}}^{c_5 c_2}$.

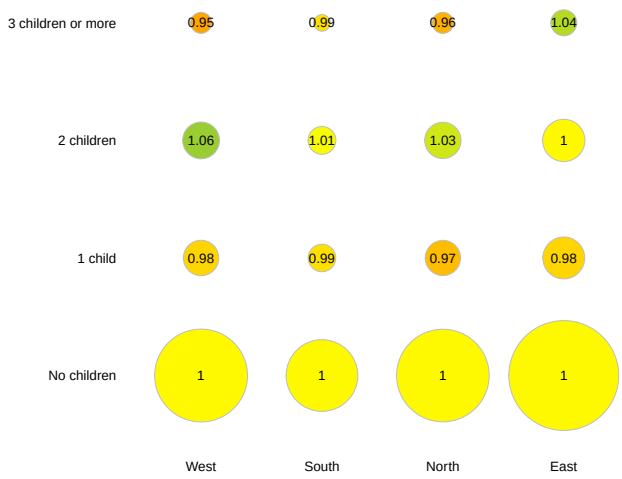


Figure 4.7: Degree of dependence for 'Number of children' and 'District', $\tilde{\rho}^{c_6 c_1}$.

Social Background

The following is accounted for in the simulation: High school is the highest level of education that a youth at age 15-19 can have. Hence, if the realized person is at age 15-19 the probability of being a high school drop-out, defined in chapter 3, is used to account for this. We select some bivariate characteristics for the social background. The given pairs of categorical variables are visualized as biplots.

In figure 4.8 the degree of dependence for 'Child neglect' and 'Marital status' is visualized. In the realization of Lilleby a child neglected resident is more likely to be single and less likely to be married. The assumption is confirmed in the biplot. The degree of dependence for 'Child neglect' and 'Disability' is presented in figure 4.9 and confirms that a disabled resident is most likely to also have been child neglected. Figure 4.10 shows the degree of dependence for 'Education' and 'Gender'. The biplot confirms that males are more likely to have high school as their highest level of education and females more likely to have either lower or higher level of University as their highest level of education. In figure 4.11 the degree of dependence for 'Income' and 'Age' confirms that the income level increases with age. At age 15-19 one is more likely to have income level 1 as assumed. The degree of dependence for 'Loneliness' and 'Disability' is visualized in figure 4.12. A disabled resident is more likely to also find themself lonely corresponding to the assumptions made while sampling.

Figure 4.13 includes 'Child neglect' and 'Age'. They interact through one step of variables. At age 15-19 one is more likely to experience child neglect according to the biplot. In figure 4.14 the variables 'Education' and 'Age' are interacting through two steps of variables. The degree of dependence is overall weak except from the natural strong degree of dependence in age 15-19 having high school as the highest level of education.

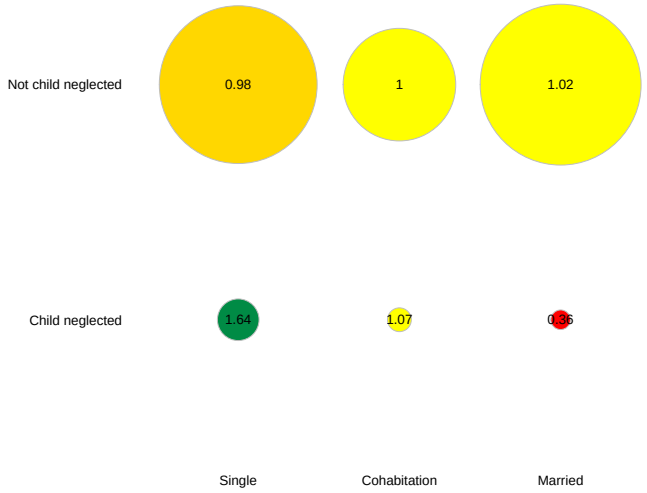


Figure 4.8: Degree of dependence for 'Child neglect' and 'Marital status', $\tilde{\rho}^{u_1 c_4}$.

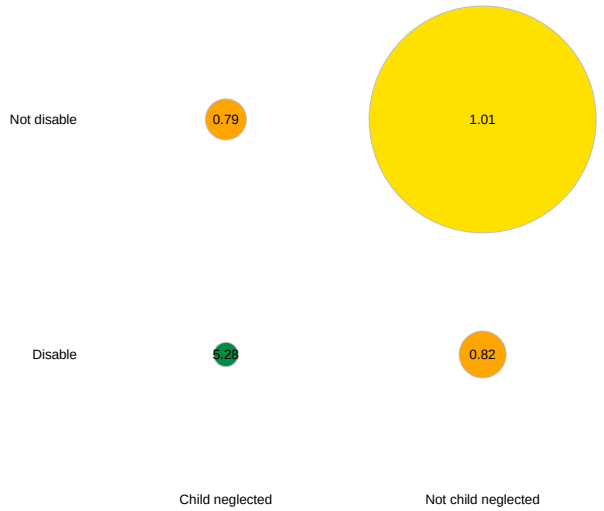


Figure 4.9: Degree of dependence for 'Disability' and 'Child neglect', $\tilde{\rho}^{u_2 u_1}$.

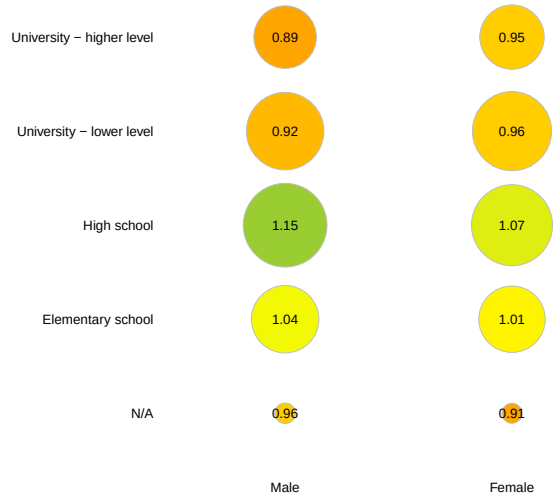


Figure 4.10: Degree of dependence for 'Education' and 'Gender', $\tilde{\boldsymbol{\rho}}^{u_3 c_5}$.

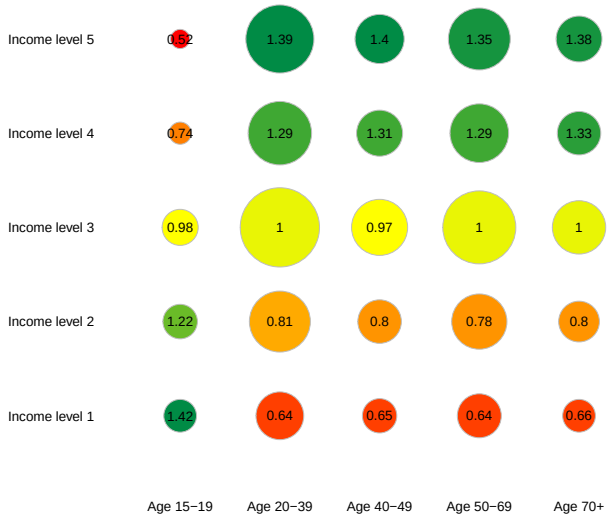


Figure 4.11: Degree of dependence for 'Income' and 'Age', $\tilde{\boldsymbol{\rho}}^{u_4 c_3}$.

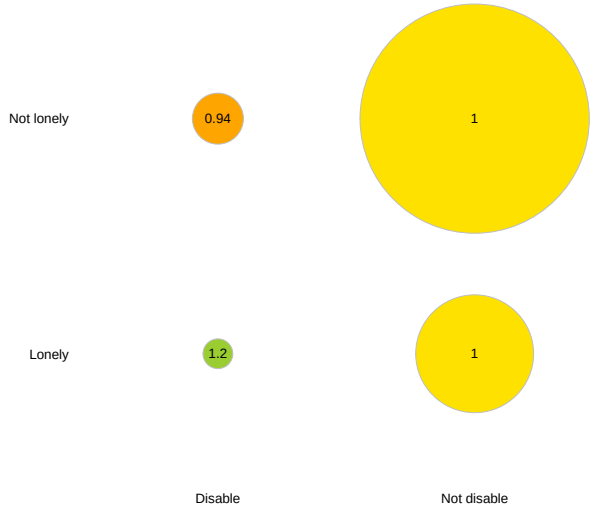


Figure 4.12: Degree of dependence for 'Loneliness' and 'Disability', $\tilde{\boldsymbol{\rho}}^{u_5 u_2}$.

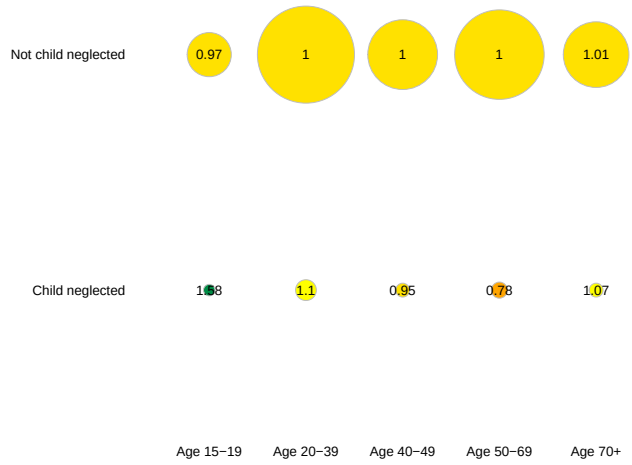


Figure 4.13: Degree of dependence for 'Child neglect' and 'Age', $\tilde{\boldsymbol{\rho}}^{u_1 c_3}$.

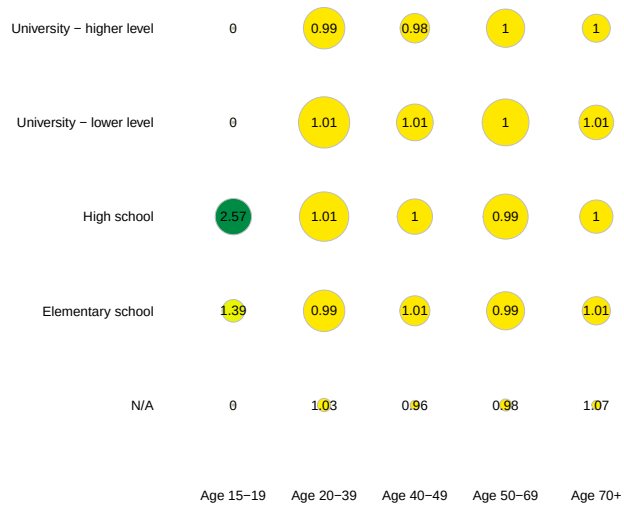


Figure 4.14: Degree of dependence for 'Education' and 'Age', $\tilde{\boldsymbol{\rho}}^{u_3 c_3}$.

Spiritual Situation

The following has been accounted for in the simulation: If a person defines themself as a personal Christian the person is also assumed to believe in God. We select some bivariate characteristics for the spiritual situation. The given pairs of categorical variables are visualized as biplots.

In figure 4.15 the degree of dependence is visualized for 'Belief in God' and 'Cultural origin'. Residents of origin 2 (Middle and South of Europe and South America), origin 3 (East of Europe and North of Asia) and origin 5 (Africa and The Near East) are more likely to believe in the Abrahamic God. A resident of origin 4 (The Far East) is less likely to believe in the Abrahamic God because the common religions in the Far East are Buddhism and Hinduism. These assumptions are confirmed in the biplot. Figure 4.16 presents the degree of dependence for 'Personal Christian' and 'Marital status'. A personal Christian is less likely to live in cohabitation, as assumed. Figure 4.17 shows that the degree of dependence for 'Church activity' and 'Loneliness' is overall weak. Still a resident that do attend church weekly is a little less likely to find themself lonely. The degree of dependence for 'Bible usage' and 'Personal Christian', visualized in figure 4.18, primarily reflects that less people define themselves as personal Christians in Lilleby than the marginal probabilities assume. Residents in general do not read the Bible frequently.

Figure 4.19 includes 'Personal Christian' and 'Age'. They interact through one step of variables. In figure 4.20 the variables 'Bible usage' and 'District' are interacting through two steps of variables. In both biplots the degree of dependence is weak.

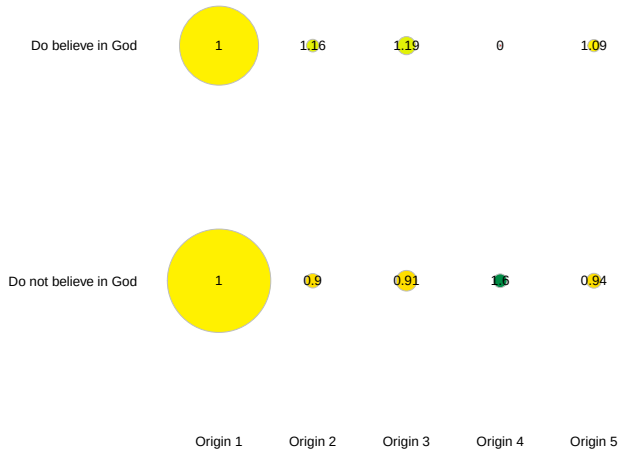


Figure 4.15: Degree of dependence for 'Belief in God' and 'Cultural origin', $\tilde{\boldsymbol{\rho}}^{s_1 c_2}$.

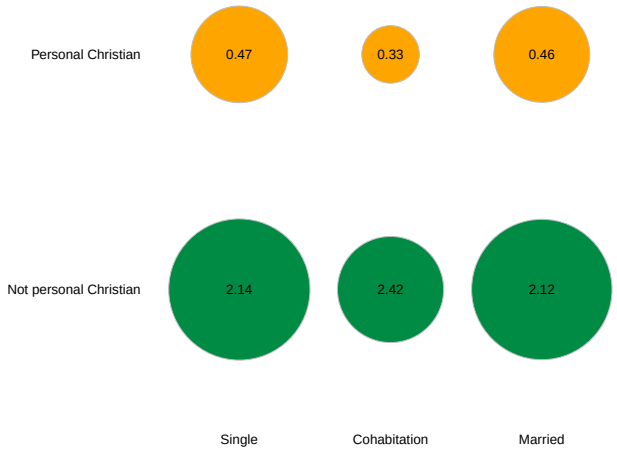


Figure 4.16: Degree of dependence for 'Personal Christian' and 'Marital status', $\tilde{\boldsymbol{\rho}}^{s_2 c_4}$.

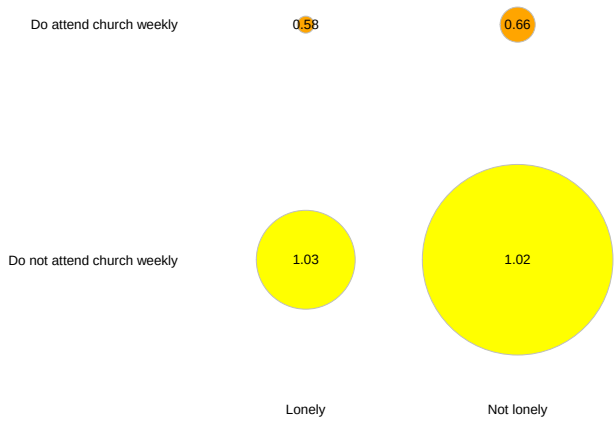


Figure 4.17: Degree of dependence for 'Church activity' and 'Loneliness', $\tilde{\rho}^{s_3 u_5}$.

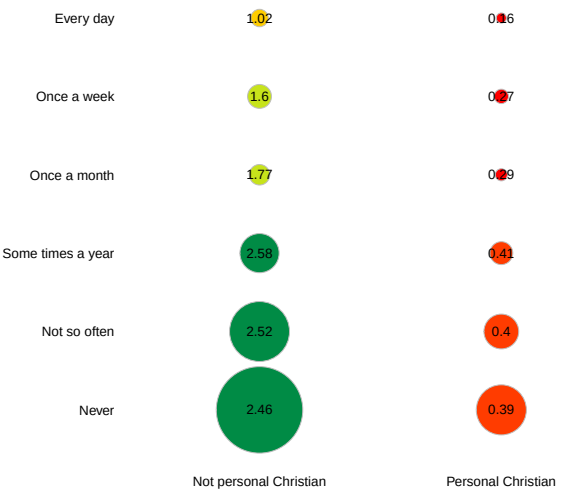


Figure 4.18: Degree of dependence for 'Bible usage' and 'Personal Christian', $\tilde{\rho}^{s_4 s_2}$.

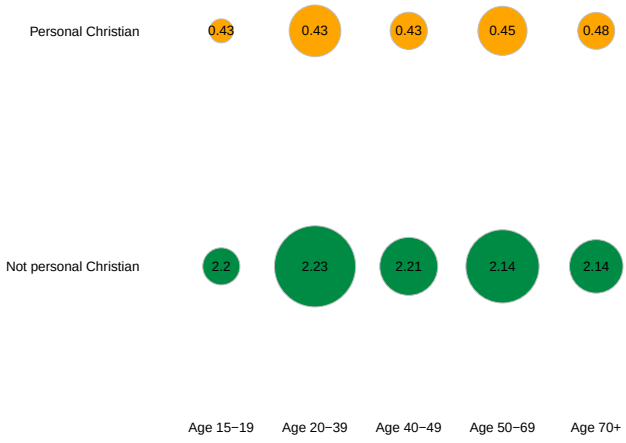


Figure 4.19: Degree of dependence for 'Personal Christian' and 'Age', $\tilde{\boldsymbol{\rho}}^{s_2 c_3}$.

Figure 4.20: Degree of dependence for 'Bible usage' and 'District', $\tilde{\boldsymbol{\rho}}^{s_4 c_1}$.

Summary

The biplots of some of the interactions defined in equation (4.1) confirms that dependence exists in the realization of Lilleby. Hence, the dependence accounted for while sampling is reflected in Lilleby. Naturally the degree of dependence between variables interacting through one or two steps is weak, if at all existing. It is important to keep in mind that the reflected dependence might come from spurious relationships. Still, the overall evaluation of the simulated realization of Lilleby is that the residents follow the assumptions made on dependence in the sampling.

Chapter 5

Lilleby Monitor

The realized population of Lilleby is ready to participate in a statistical survey with questionnaires. The formal regulations concerning the collection of data in real life studies is accounted for and the questionnaire is made. A stratified sample of respondents is desired in order to obtain representative, valid and reliable responses for the response data. We introduce the concept of stratification and derive a likelihood model with a corresponding posterior model to deal with bias correction. Proportion estimators are derived. The questionnaire is distributed to a representative and stratified sample of Lilleby. The response sample is evaluated by its sensitivity to the stratification and bias correction by the comparison of proportion estimates.

5.1 Collection and Correction of Data

The design of the experiment and questionnaire is usually time-consuming and the process is described in the following. Also the formal regulations concerning permission and confidentiality in real life studies is accounted for. The questionnaire is made and the correction models are derived, as well as the proportion estimators. Additionally, a measure of the goodness of fit is introduced.

5.1.1 Collection of Data

The collection of data in real life studies is subject to formal regulations and it is essential to keep the privacy of each respondent. The data collection includes a random sample of respondents, an information note for the potential respondents, a questioning strategy and the questionnaire itself (Johannessen, Tufte, and Christoffersen, 2016).

Norwegian Centre for Research Data (NSD, 2020) contributes to and shares research data by ensuring open access and making opportunities for research by offering their support and information.

Privacy

The formal regulations concerning permission and confidentiality depends on what kind of data the questionnaire asks for. According to NSD (2019a) a project should be notified to NSD by a Notification Form if it includes personal data and/or if the processing of data, either personal or anonymous, is done electronically. A Notification Form is required even if the data is not to be published.

Personal data means that a person is either indirectly or directly identifiable by the information left in the questionnaire. An indirectly identifiable respondent might or might not be traced based on what they answers. On the contrary, a directly identifiable respondent would have been asked to leave some kind of unique ID.

An anonymous respondent cannot be identified because the data contain no information that directly or indirectly identify them. A study including only anonymous respondents do not need to be notified in the first place. Still, if the processing of the anonymous data such as collecting, storing, sharing and publishing is done electronically the project must be notified.

A research institution might have its own agreements with NSD concerning the Notification Form. In addition, it is important to make a data processor agreement in advance if the supplier of the online survey is not affiliated with the research institution (NSD, 2018).

In 2018 EUs General Data Protection Regulation (GDPR) took effect in Norway. In addition to handle the Notification Form, NSD also makes sure that the proposed project fulfills the GDPR by a process called DPIA. Each research institution is also required to make sure that their projects fulfill the GDPR but this is often done in cooperation with NSD.

As long as the respondent has given their consent there are few restrictions on which questions that are allowed to ask. The exceptions are questions opposing the respondents confidentiality, questions involving a third party and questions not following the guidelines defined by NESH (2015).

Sample Design

Segmentation of respondents is important in order to obtain representative, valid and reliable data. The segmentation precludes systematic bias of the collected data. The quality of the sample of respondents depends on the stratification used. To ensure that the stratification is of high quality it is sensible to use an already existing panel when collecting answers. Two recognized panels in Norway are KANTAR (2020) and CINT (2020). The downside is the high expenses that follow because of the amount of work put into ensuring a sample of high quality.

Information Note

The information note requests participation from the potential respondents and presents information about the study such as its goal and terms of privacy. It is important that the study is presented in such a way that the potential respondent is inspired to participate. Then the note should explain the procedure to follow if the respondent wants to participate.

The note must also follow the guidelines for research ethics in the social sciences, the humanities, law and theology defined by The National Committee for Research Ethics in the Social Sciences and the Humanities (NESH, 2015). NSD (2019b) offers and recommends a template for the information note.

Questioning Strategy

The quantitative approach of data collection is the questionnaire. A questionnaire yields standardized answers which makes it possible to generalize the results. The weakness compared to a qualitative approach is that once the responses are collected no more information can be added. Hence, the most important part of the data collection using a questionnaire is developing a questioning strategy to define each question to be included in the questionnaire.

Making a complete questionnaire involves different stages. Start out with creativity combined with inspiration from already existing questionnaires. This is part of developing the questioning strategy. Find a structure and decide on the order of the questions. At last decide on the layout. Let a few people test the questionnaire to see if the questions make sense.

The most important aspects to consider when developing the questioning strategy, according to Johannessen, Tufte, and Christoffersen (2016):

- Every question should be relevant and unequivocal.
- Each question should have an easy and clear formulation and the way to respond should be intuitive.
- If the question requires an answer presented by a scale the different levels of the scale should be mutual exclusive and well-explained.
- Loaded questions, meaning that a question prefers one answer above another, should be avoided. This is to ensure that the respondent gives an answer that is as subjective as possible.
- By nature the respondent will have an unconscious need to present themself in the best social acceptable manner as possible. The questions should avoid the urge for the respondent to do this.
- Where to place the questions regarding the respondents background information must be considered carefully. If included in the beginning they might affect the following answers but at the same time they could function as a warm-up for the respondent.
- The questions should be complementary but not too many. They should all together answer the goal of the study.
- The questionnaire itself should be self-explaining and the layout should be universal such that anyone is able to participate.

5.1.2 The Questionnaire

The questionnaire used to collect data in Lilleby is called 'The Social and Spiritual Situation in Lilleby' and is found in appendix A. The questions collect information on each of the variables included in the realized Lilleby population.

Data collection in real life studies deals with two major types of correction. Firstly, stratification is crucial in order to obtain representative responses. Stratification means that the questionnaire is distributed to a representative sample of Lilleby segmented by the variables 'Gender', 'Age' and 'District', concerning question 1, 2 and 3, respectively. Stratification is enforced when the questionnaires are distributed to a subset of residents of Lilleby. But some gender, age groups or districts might be over- or underrepresented in the sample of respondents. To correct for the potential skewness, the stratification variables are weighted to restore the correct stratification in the responses. Secondly, correction is made due to the effect of the questioning ambiguity and the potential prejudices in the responses caused by the subjective interpretation of the questions.

We consider the questioning strategy developed in the preceding section. Questions regarding the respondents background information used for stratification are assumed to be well-defined and unbiased. Comments on the assumptions are listed in the following:

- **Question 1:** A person's biological gender is either male or female. The gender variable is assumed to be well-defined with two mutual exclusive categories.
- **Question 2:** The age groups are mutual exclusive and each group is defined such that it includes residents with a similar life situation. The question restricts the questionnaire to respondents from age 15 to 70+. Hence, the questionnaire is not distributed to children under the age of 15. The age variable is assumed to be well-defined.
- **Question 3:** A resident of Lilleby is currently living in either one district or another according to their permanent address. The district variable is assumed to be well-defined with four mutual exclusive categories.

The questions regarding the stratification variables are placed in the beginning of the questionnaire and hence function as a warm-up session. The order of the questions is considered carefully to avoid any effect on the current answer from the preceding ones. The question formulations are as clear as possible and loaded questions are avoided. There exist an urge to present oneself in the best social acceptable manner. This causes a potential bias in the answer. Comments are made on the questions where bias might occur and hence correction is needed. The comments are listed in the following:

- **Question 4:** The country where the respondent's mother were raised might link the respondent to a history of humiliation or a unpopular political opinion among other things. Some bias might occur within the variable of origin because of this.
- **Question 5:** A respondent or their partner might have children from previous relationships. The result might be some bias in the variable of number of children caused by a person's subjective definition of whom to be their children. The bias is assumed to skew the answers in both directions.

- **Question 6:** A person's marital status reflects a lifestyle. Some people might value to be in a lifelong relationship and wants to portray themselves as married. People who are engaged might also answer that they are married. Hence, some bias is assumed to occur in the variable of a respondent's current situation when it comes to marital status. A higher proportion is assumed to answer that they live together with their husband/wife.

- **Question 7:** A person's level of education has become a measure of success in Norway. A higher level of education is linked to a higher intellectual and even social status. Hence, we assume some bias in the education level variable because of answers being skewed to the higher education levels.

- **Question 8:** The same urge applies to a person's income level since being wealthy is synonymous to a higher level of success in life. Some bias is assumed in the income level variable as well, with the answers being skewed to the higher income levels. We also assume some bias because of answers being skewed to a lower income level. People tend to avoid the extreme outcomes and gather around the average.

- **Question 9:** If a person receives a disability benefit he or she might not be open about it because of the stigma connected to not being able to work for a living. Bias may occur in the disability variable because more people will answer that they do not receive a disability benefit even though they do.

- **Question 10:** In some cases people might suppress incidents that happened to them in their childhood. Some bias might occur in the child neglect variable as a result of this. A person might not know if he or she ever experienced to be neglected as a child but the expected proportion to answer that they do not know is small.

- **Question 11:** When it comes to loneliness it is hard to put a limit for when you are lonely or not. The question seeks to find the respondents that subjectively experience a feeling of loneliness as a part of their everyday life. Still there is a chance that more people will answer that they feel lonely quite often. At the same time people might want to portray themselves in a better light. A respondent might find it hard to decide whether they feel lonely on a regular basis or not and hence a relative huge proportion is assumed to say that they do not know. Hence, some bias is assumed to occur in the loneliness variable.

- **Question 12:** A person might want to portray themself as dutiful on the one hand or show disapproval on the other hand when it comes to reading the Bible. The variable of Bible usage might contain some bias because of this.

- **Question 13:** Inaccurate answers might occur because of a person's wish to either portray themself as dutiful on the one hand or show disapproval on the other hand. Still most bias is assumed to be a result of people answering that they attend church weekly, while in practice they attend church less frequently.

- **Question 14:** Some people do not know if they believe in a monotheistic God or even if there is something more to the world than what we can actually see. Bias may occur regarding belief in God if people have not made up their minds yet.

- **Question 15:** People might not know if they define themselves as a Christian with a personal relationship with God or not. Hence, some bias is assumed to occur in the personal Christian variable.

5.1.3 Correction Models

The many regards concerning the collection of data in real life studies must be accounted for in the processing and interpretation of the responses to the questionnaire. Stratification is introduced and the psychological aspects of answering a questionnaire such as potential prejudices is expressed in mathematical terms by a likelihood model. The bias caused by prejudices is corrected by a corresponding posterior model.

Stratification Model

The questionnaire is distributed to a representative and stratified sample of Lilleby inhabitants of size n. The sample is stratified by the variables 'Gender', 'Age' and 'District'; called the stratification variables. The sample of size n is denoted by $\mathbf{x} = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \Omega_{\mathbf{X}}$ and the stratification variables take on the following sample space and corresponding marginal distribution:

- $x_G \in \Omega_{X_G} : p(x_G; \mathbf{p}^G); \quad \mathbf{p}^G = (p_1^G, p_2^G)$,
- $x_A \in \Omega_{X_A} : p(x_A; \mathbf{p}^A); \quad \mathbf{p}^A = (p_1^A, ..., p_5^A)$,
- $x_D \in \Omega_{X_D} : p(x_D; \mathbf{p}^D); \quad \mathbf{p}^D = (p_1^D, ..., p_4^D)$.

The collection of actual observations is the response of $n^* \leq n$ inhabitants and the sample is collected in $\mathbf{x}^* = [\mathbf{x}_1^*, ..., \mathbf{x}_{n^*}^*]$. Since we do not control the population response, these n^* answers may not be stratified with respect to the stratification variables. Some gender, age groups or districts might be over- or underrepresented causing skewness in the observations. The number of returned questionnaires within each combination of the stratification variables is denoted by n_{ijk}^*, for $i = 1, 2; j = 1, ..., 5; k = 1, ..., 4$. To correct for the skewness each possible combination of the stratification variables is assigned a weight, w_{ijk}.

To determine the weights we solve the following minimization problem:

$$w_{ijk} = \min_{w_{ijk}} \left\{ \sum_{ijk} (w_{ijk} - 1)^2 \right\}, \tag{5.1}$$

given by the $i + j + k$ equality constraints:

$$\sum_{jk} w_{ijk} n_{ijk}^* = n^* p_i^G; \quad i = 1, 2,$$

$$\sum_{ik} w_{ijk} n_{ijk}^* = n^* p_j^A; \quad j = 1, ..., 5,$$

$$\sum_{ij} w_{ijk} n_{ijk}^* = n^* p_k^D; \quad k = 1, ..., 4,$$

with

$$\sum_{ijk} n^*_{ijk} = n^*.$$

The equality constraints yield a system of eleven linearly dependent equations. Hence, the system is reduced by the two equations for $j = 5$ and $k = 4$, resulting in the following system of nine linearly independent equality constraints:

$$\begin{aligned}
\sum_{jk} w_{ijk} n^*_{ijk} &= n^* p^G_i; \quad i = 1, 2, \\
\sum_{ik} w_{ijk} n^*_{ijk} &= n^* p^A_j; \quad j = 1, ..., 4, \\
\sum_{ij} w_{ijk} n^*_{ijk} &= n^* p^D_k; \quad k = 1, ..., 3.
\end{aligned}$$

The minimization problem can now be solved by using Lagrange multipliers. The notation is inspired by Nocedal and Wright (2006). A Lagrange multiplier is a scalar quantity, λ_m, introduced for each constraint, where $m = 1, ..., 9$. The Lagrangian function is defined by:

$$\begin{aligned}
\mathcal{L}(\mathbf{w}, \boldsymbol{\lambda}) = &\sum_{ijk} (w_{ijk} - 1)^2 - \sum_{m=1}^{2} \lambda_m \big(\sum_{jk} w_{ijk} n^*_{ijk} - n^* p^G_i \big) \\
&- \sum_{m=3}^{6} \lambda_m \big(\sum_{ik} w_{ijk} n^*_{ijk} - n^* p^A_j \big) - \sum_{m=7}^{9} \lambda_m \big(\sum_{ij} w_{ijk} n^*_{ijk} - n^* p^D_k \big),
\end{aligned}$$

where $\mathbf{w} = (w_{111}, w_{211}, w_{121}, w_{221}, ..., w_{254})$ and $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_9)$. The necessary optimality conditions for the Lagrangian function require that $\nabla \mathcal{L}(\mathbf{w}^*, \boldsymbol{\lambda}^*) = 0$, where $\mathbf{w}^* = (w^*_{111}, ..., w^*_{254})$ and the vector $\boldsymbol{\lambda}^* = (\lambda^*_1, ..., \lambda^*_9)$ are the solution candidates to the minimization problem. The sufficient optimality conditions require that $\nabla^2 \mathcal{L}(\mathbf{w}^*, \boldsymbol{\lambda}^*)$ is positive definite.

The set of w^*_{ijk}, for $i = 1, 2; j = 1, ..., 5; k = 1, ..., 4$, are the weights corresponding to each possible combination of the stratification variables. The weights are assigned to their corresponding element in $\mathbf{x}^*$ to correct for the skewness in the observations.

Likelihood Model

The urge to present ourselves in the best social acceptable manner causes a potential bias in the answers to the questionnaire. Comments are made on the potential bias in the preceding section and the effect is now expressed in mathematical terms.

The observations in $\mathbf{x}^* = [\mathbf{x}^*_1, ..., \mathbf{x}^*_{n^*}]$ can be expressed by their likelihood denoted by $p(\mathbf{x}^*_i \mid \mathbf{x}_i)$, where $\mathbf{x}_i$ are the correct states of the person i. The answers of the stratification variables 'Gender', 'Age' and 'District' are assumed to be correct.

The likelihood model assumes that each answer for a particular person, x^*_j, for $j = 1, ..., 15$, does not depend on the other answers. The total likelihood for the person is

given by:

$$p(\mathbf{x}^* \mid \mathbf{x}) = \prod_{j=1}^{15} p(x_j^* \mid x_j).$$

We parametrize the likelihood function by deviations from the correct answer. Potential bias to the answer, x_j^*, is represented by deviance from the correct state of the person, x_j, by a parameter of deviance denoted by $\alpha_t \in \mathbb{R}_{[0,1]}$, for $t = 1, ..., k_{x_j}$. Let k_{x_j} be the number of outcomes for the given variable, x_j. The likelihood, $p(x_j^* \mid x_j)$, is expressed by a matrix of dimension $k_{x_j} \times k_{x_j}$. Respondents are assumed to tend to avoid the extreme outcomes of the given variable and gather around the average.

The likelihood model, if $k_{x_j} = 3$, is expressed in matrix form as:

$$p(x_j^* \mid x_j) = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 1-\alpha_1 & \alpha_1 & 0 \\ 0 & 1 & 0 \\ 0 & \alpha_3 & 1-\alpha_3 \end{pmatrix} \end{array},$$

where the rows correspond to x_j and the columns to x_j^*. Note that $0 < \alpha_t < 1$. For $\alpha_1 = \alpha_3 = 0$ the answer correctly reflects the truth.

The independence assumed between each x_j^* results in the following expression for the posterior distribution of the correct states of a person, $\mathbf{x}$, given the answers, $\mathbf{x}^*$:

$$p(\mathbf{x} \mid \mathbf{x}^*) = \frac{p(\mathbf{x})p(\mathbf{x}^* \mid \mathbf{x})}{p(\mathbf{x}^*)} = \prod_{j=1}^{15} \frac{p(x_j)p(x_j^* \mid x_j)}{p(x_j^*)} = \prod_{j=1}^{15} p(x_j \mid x_j^*), \tag{5.2}$$

where $p(\mathbf{x}) = \prod_{j=1}^{15} p(x_j)$ is the prior distribution for the correct states of a person.

The corresponding prior model for the case when $k_{xj} = 3$ is assumed to be uniformly distributed, hence $p(x_j) = \frac{1}{k_{x_j}}$:

$$p(x_j) = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \\ \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix} \end{array}.$$

The posterior distribution for a given variable, x_j^*, is by Bayes rule:

$$p(x_j \mid x_j^*) = \frac{p(x_j)p(x_j^* \mid x_j)}{\sum_{x_j \in \Omega_{X_j}} p(x_j)p(x_j^* \mid x_j)}, \tag{5.3}$$

where the sum is over all outcomes for x_j. Hence, the posterior model for $k_{xj} = 3$ can be calculated from the fully specified matrix $p(x_j^* \mid x_j)$ and vector $p(x_j)$:

$$p(x_j \mid x_j^*) = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ \frac{\alpha_1}{1+\alpha_1+\alpha_3} & \frac{1}{1+\alpha_1+\alpha_3} & \frac{\alpha_3}{1+\alpha_1+\alpha_3} \\ 0 & 0 & 1 \end{pmatrix} \end{array}.$$

The posterior is found for each variable. The posterior for each person is found by equation (5.2). Hence, the potential ambiguity or bias in the responses caused by prejudices are corrected by the posterior model.

The parameter of deviance, α_t, for $t = 1, ..., k_{x_j}$, represents the potential bias to an answer. The prejudices are caused by the subjective interpretation of the questions, the urge to present oneself in the best social acceptable manner and the fact that people tend to avoid the extreme outcomes and gather around the average. Consequently, it is difficult to assess α_t. Affirmatively, a search for relevant literature on the psychology behind answering a questionnaire yields lacking information. Especially for questionnaires concerning social factors. Further attempts are encouraged to quantify potential bias caused by prejudices such that more realistic values may be found.

5.1.4 Proportion Estimators

The response sample is collected in $\mathbf{x}^* = [\mathbf{x}_1^*, ..., \mathbf{x}_{n^*}^*]$. Each response, $\mathbf{x}_s^*$ for $s = 1, ..., n^*$, contains $\mathbf{x}_{s(ijk)}^*$ for the stratification variables, where $i = 1, 2; j = 1, ..., 5; k = 1, ..., 4$, and $x_{s(l)}^*$ for the additional variables, where $l = 4, ..., 15$. The response sample is evaluated by its sensitivity to stratification and bias correction. The sensitivity is measured by different proportion estimators.

Let L denote the variables l from $4, ..., 15$. Hence, $L = \{l_4, ..., l_{15}\}$. Consider a set of variables $H \subset L$, e.g. $H = \{l_4, l_7\}$, where $\mathbf{x_H} = (x_{l_4}, x_{l_7})$ denotes the pair of variables of interest. We estimate the proportion of the population where the responses $\mathbf{x_H^*} = (x_{s(l_4)}^*, x_{s(l_7)}^*)$ correspond to $\mathbf{x_H}$ for every respondent $s = 1, ..., n^*$. The proportion estimator is denoted by $p_{\mathbf{x_H^*}=\mathbf{x_H}}$ and is derived for four different approaches to the response sample.

Naive Estimator

The naive approach assumes that the respondents answer is the correct state of the person for every respondent s, where $s = 1, ..., n^*$. Hence, neither stratification nor bias correction is applied. The naive estimator is given by:

$$p_{\mathbf{x_H^*}=\mathbf{x_H}}^* = \frac{1}{n^*} \sum_{s=1}^{n^*} I(\mathbf{x}_{s(\mathbf{H})}^* = \mathbf{x_H}) = \frac{1}{n^*} \sum_{s=1}^{n^*} I(x_{s(l_4)}^* = x_{l_4}) I(x_{s(l_7)}^* = x_{l_7}).$$

Stratified Estimator

The stratified approach only takes stratification into account to correct the response sample. Each observation in $\mathbf{x}^*$ is assigned a corresponding weight, $w^*_{s(ijk)}$ for $s = 1, ..., n^*; i = 1, 2; j = 1, ..., 5; k = 1, ..., 4$, to correct for the skewness in the responses caused by the lack of stratification in the response sample. The stratified estimator is given by:

$$\tilde{p}_{\mathbf{x}^*_{\mathbf{H}}=\mathbf{x}_{\mathbf{H}}} = \frac{1}{n^*}\sum_{s=1}^{n^*} w^*_{s(ijk)} I(\mathbf{x}^*_{s(\mathbf{H})} = \mathbf{x}_{\mathbf{H}}) = \frac{1}{n^*}\sum_{s=1}^{n^*} w^*_{s(ijk)} I(x^*_{s(l_4)} = x_{l_4}) I(x^*_{s(l_7)} = x_{l_7}).$$

Bias Corrected Estimator

We apply the posterior model to the response sample to correct for the psychological bias that might occur when answering a questionnaire. The bias corrected approach corrects for the prejudices in the response sample by including the posterior, $p(\mathbf{x}_{s(\mathbf{H})} = \mathbf{x}_{\mathbf{H}} \mid \mathbf{x}^*_{s(\mathbf{H})})$, for every respondent $s = 1, ..., n^*$; where $\mathbf{x}_{s(\mathbf{H})}$ is the correct state of the person. The bias corrected estimator is given by:

$$\begin{aligned}\hat{p}_{\mathbf{x}^*_{\mathbf{H}}=\mathbf{x}_{\mathbf{H}}} &= \frac{1}{n^*}\sum_{s=1}^{n^*} p(\mathbf{x}_{s(\mathbf{H})} = \mathbf{x}_{\mathbf{H}} \mid \mathbf{x}^*_{s(\mathbf{H})}) \\ &= \frac{1}{n^*}\sum_{s=1}^{n^*} p(x_{s(l_4)} = x_{l_4} \mid x^*_{s(l_4)}) p(x_{s(l_7)} = x_{l_7} \mid x^*_{s(l_7)}).\end{aligned}$$

Stratified and Bias Corrected Estimator

The stratified and bias corrected estimator corrects for both stratification and potential prejudices, where $s = 1, ..., n^*; i = 1, 2; j = 1, ..., 5; k = 1, ..., 4$. The estimator is given by:

$$\begin{aligned}\hat{\tilde{p}}_{\mathbf{x}^*_{\mathbf{H}}=\mathbf{x}_{\mathbf{H}}} &= \frac{1}{n^*}\sum_{s=1}^{n^*} w^*_{s(ijk)} p(\mathbf{x}_{s(\mathbf{H})} = \mathbf{x}_{\mathbf{H}} \mid \mathbf{x}^*_{s(\mathbf{H})}) \\ &= \frac{1}{n^*}\sum_{s=1}^{n^*} w^*_{s(ijk)} p(x_{s(l_4)} = x_{l_4} \mid x^*_{s(l_4)}) p(x_{s(l_7)} = x_{l_7} \mid x^*_{s(l_7)}).\end{aligned}$$

5.1.5 Goodness of Fit

To indicate whether a proportion estimate is a good fit or not we introduce a measure for the deviance of multiple estimated proportion estimates from the true value. In our case the true value is the Lilleby proportion. Let $b = 1, ..., m$, where m is the number of simulated response samples.

According to James et al. (2017), the mean squared error (MSE) is the most commonly-used measure. The MSE is a metric to quantify the goodness of fit of a model. To simplify

the notation let p be the true Lilleby proportion and $\hat{p_b}$, for $b = 1, ..., m$, be the proportion estimates for the relevant approach. Then the MSE is found by:

$$\text{MSE} = \frac{1}{m}\sum_{b=1}^{m}(p - \hat{p_b})^2.$$

The MSE is small if the m proportion estimates are close to the true Lilleby proportion and large if some of them are far away.

The root mean squared error (RMSE) allows for easier interpretation because of its applicable scale, which makes it possible to compare the overall goodness of fit between the different types of proportion estimates. The RMSE is found by taking the square root of the MSE:

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{b=1}^{m}(p - \hat{p_b})^2}. \tag{5.4}$$

The RMSE evaluates whether the given proportion estimates are centered close to the true Lilleby proportion or not. Penalty is added when the spread in the proportion estimates increases.

5.2 The Survey of Lilleby

The questionnaire 'The Social and Spiritual Situation in Lilleby' in appendix A is distributed to a representative and stratified sample of residents from the realized Lilleby population. The many regards concerning the collection of data in real life studies are accounted for in the sampling of the responses to the questionnaire. The stratification model is used to correct for over- or under-representation of respondents according to stratification groups. To correct for potential prejudices we apply the posterior model to the responses. The proportion estimates are used to evaluate the response sample by its sensitivity to the stratification and bias correction.

5.2.1 Collection

The questionnaire is distributed to a sample of $n = 1000$ residents of the realized Lilleby population by drawing a random but stratified sample of Lilleby. The sample is stratified by the stratification variables 'Gender', 'Age' and 'District'. Hence, the number of residents, n_{ijk}, is equal in each combination of the stratification variables, for $i = 1, 2; j = 1, ..., 5; k = 1, ..., 4$. Within each group of stratification variables we specify a corresponding probability that a resident in the given group actually respond to the questionnaire, $p^*_{ijk} \in [0.5, 0.9]$. We assume that an elderly person is more likely to answer the questionnaire than a younger person. People between the age of 40-49 is assumed to have a busy lifestyle and hence to be the less frequent group of respondents. Based on the given p^*_{ijk} a random sample of actual respondents within the corresponding group is drawn. The collection of the actual observations is the response of $n^* \leq n$ residents where $n^* = \sum_{ijk} n^*_{ijk}$, for $i = 1, 2; j = 1, ..., 5; k = 1, ..., 4$, and n^*_{ijk} is the actual number of

respondents within each stratification group. We apply the likelihood to account for the bias caused by the psychological aspects of answering a questionnaire. The likelihood, $p(x_l^* \mid x_l)$ for $l = 4, .., 15$, is parametrized by deviations from the correct answer by $\alpha_t \in [0.05, 0.3]$, for $t = 1, ..., k_{x_l}$. Each α_t is assigned a value based on the preceding comments made on the potential bias regarding each question in the questionnaire. The corresponding likelihood is applied to every variable except from the stratification variables. Hence, every variable for each person contains the persons actual answer to the question. The response sample is collected in $\mathbf{x}^* = [\mathbf{x}_1^*, ..., \mathbf{x}_{n^*}^*]$.

5.2.2 Stratification

Some gender, age groups or districts might be over- or under-represented in the response sample, $\mathbf{x}^*$. To correct for the potential skewness, the stratification variables are weighted to restore stratification of the responses. The marginal distributions of the stratification variables, $\mathbf{p}^G, \mathbf{p}^A$ and $\mathbf{p}^D$, and the number of respondents within each stratification group, n_{ijk}^* for $i = 1, 2; j = 1, ..., 5; k = 1, ..., 4$, is used to determine the corresponding weights, w_{ijk}^*. This is done by solving the minimization problem in equation (5.1) by Lagrange multipliers.

5.2.3 Bias Correction

The likelihood model is used in the sampling of the response sample to account for potential prejudices. We calculate the posterior model using the given marginal distribution as prior, $p(x_l)$, and the likelihood, $p(x_l^* \mid x_l)$ for $l = 4, .., 15$. The posterior, $p(x_l \mid x_l^*)$, is now given by equation (5.3) for a given variable, x_l.

5.2.4 Proportion Estimates

We measure the sensitivity of the responses in $\mathbf{x}^*$ to stratification and bias correction by the four different proportion estimators. This is done for the following outcomes of either a single variable or a given set of variables:

Single Outcomes	**Pairs of Outcomes**
Origin 1	Married and age 20-39
1 child	Not child neglected and not disabled
Married	Do believe in God and origin 5
University - higher level	High school and male
Child neglected	Lonely and not disabled
Do not believe in God	Personal Christian and married
Not disabled	
Lonely	

Note that the prior distribution, $p(x)$, used in the calculation of the posterior model is the marginal distribution of x. Thus, the proportion estimates for the single outcomes are less interesting. We focus our discussion on the evaluation of the pairs of outcomes.

Histograms

We simulate $m = 100$ response samples and calculate the four proportion estimates for the relevant single outcomes and pairs of outcomes for each response sample. The m naive, stratified, bias corrected as well as the stratified and bias corrected proportion estimates are plotted as histograms for each single outcome and pair of outcomes. The histograms for the relevant single outcomes are found in figure 5.1 through 5.8. The histograms for the relevant pairs of outcomes are found in figure 5.9 through 5.14. The true Lilleby proportion is also plotted.

The sensitivity of the response sample to the stratification and bias correction is visualized in the histograms. The different proportion estimates are evaluated by whether they are centered away from or close to the true Lilleby proportion. We also evaluate the spread around the center for a given proportion estimate.

We consider one example. Histograms of the m naive, stratified, bias corrected as well as the stratified and bias corrected proportion estimates for the pair of outcomes 'Not child neglected and not disabled' are presented in figure 5.10. The naive proportion estimate is centered farthest away from the true Lilleby proportion while the stratified proportion estimate is centered closer to the true Lilleby proportion. Bias correction critically draws the proportion estimate closer to the true Lilleby proportion. The stratified and bias corrected proportion estimate is centered closest to the true Lilleby proportion. The spread is larger for the stratified proportion estimate than for the naive proportion estimate. Bias correction decreases the spread. Hence, the stratified and bias corrected proportion estimate has a spread that is larger than for the bias corrected proportion estimate.

We look at the overall sensitivity to the correction models in the histograms. The bias correction model effectively makes sure that the proportion estimate is centered closer to the true Lilleby proportion. The stratification model corrects for skewness but causes a larger spread in the proportion estimate.

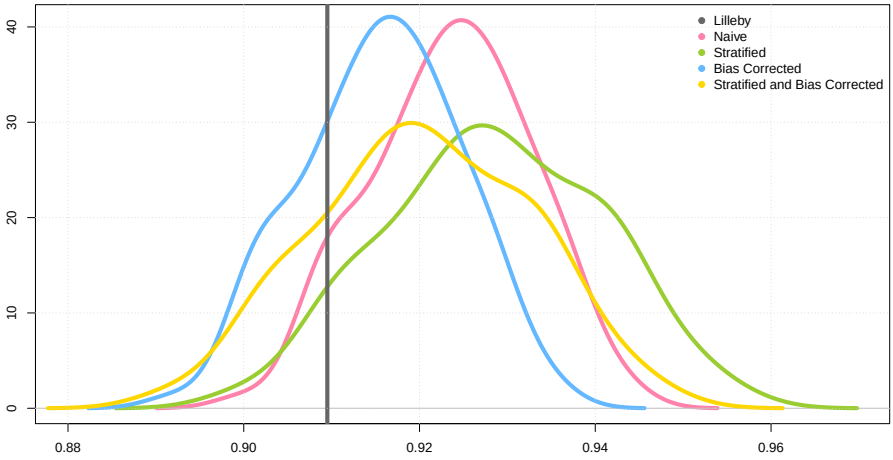


Figure 5.1: Histograms of the $m = 100$ naive (pink), stratified (green), bias corrected (blue) as well as the stratified and bias corrected (yellow) proportion estimates for the single outcome 'Origin 1'. The true Lilleby proportion (grey) is represented by a vertical line.

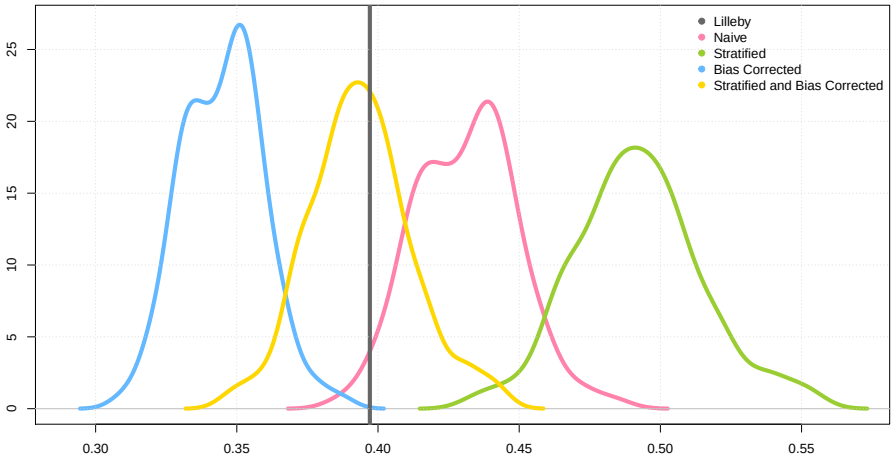


Figure 5.2: Histograms of the $m = 100$ naive (pink), stratified (green), bias corrected (blue) as well as the stratified and bias corrected (yellow) proportion estimates for the single outcome 'Married'. The true Lilleby proportion (grey) is represented by a vertical line.

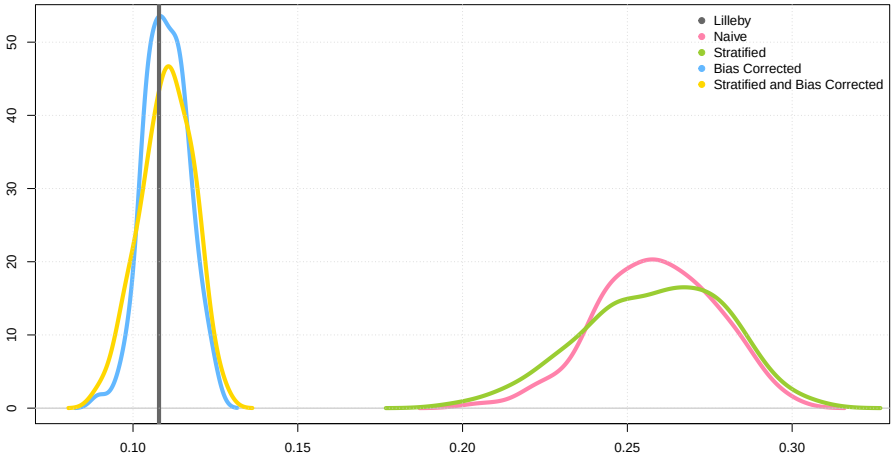


Figure 5.3: Histograms of the $m = 100$ naive (pink), stratified (green), bias corrected (blue) as well as the stratified and bias corrected (yellow) proportion estimates for the single outcome '1 child'. The true Lilleby proportion (grey) is represented by a vertical line.

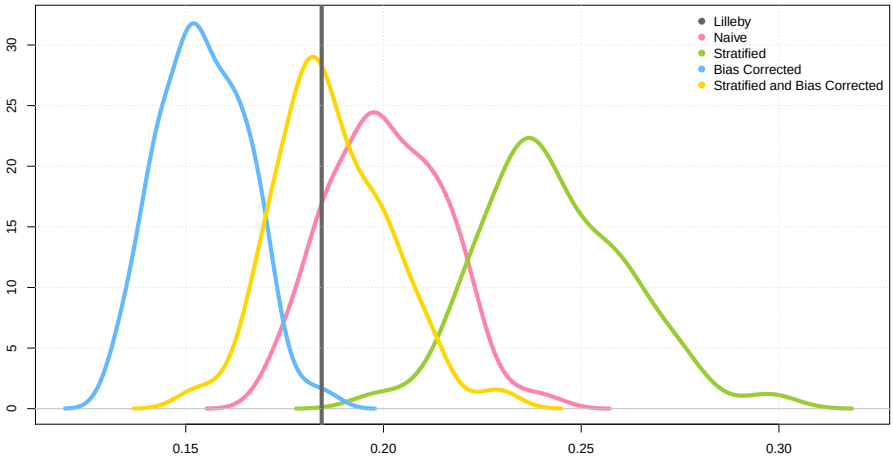


Figure 5.4: Histograms of the $m = 100$ naive (pink), stratified (green), bias corrected (blue) as well as the stratified and bias corrected (yellow) proportion estimates for the single outcome 'University - higher level'. The true Lilleby proportion (grey) is represented by a vertical line.

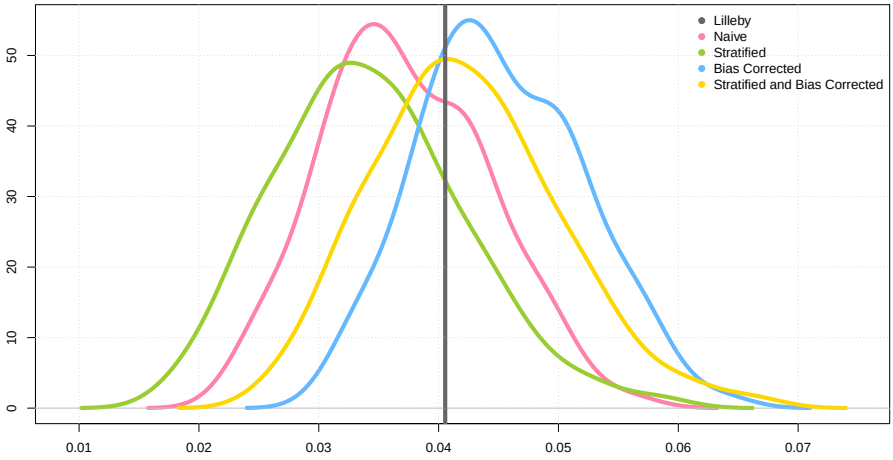


Figure 5.5: Histograms of the $m = 100$ naive (pink), stratified (green), bias corrected (blue) as well as the stratified and bias corrected (yellow) proportion estimates for the single outcome 'Neglect'. The true Lilleby proportion (grey) is represented by a vertical line.

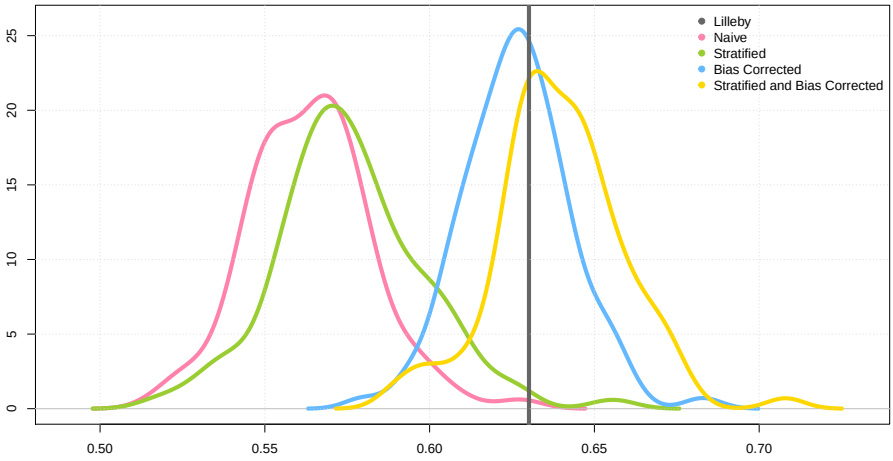


Figure 5.6: Histograms of the $m = 100$ naive (pink), stratified (green), bias corrected (blue) as well as the stratified and bias corrected (yellow) proportion estimates for the single outcome 'Do not believe in God'. The true Lilleby proportion (grey) is represented by a vertical line.

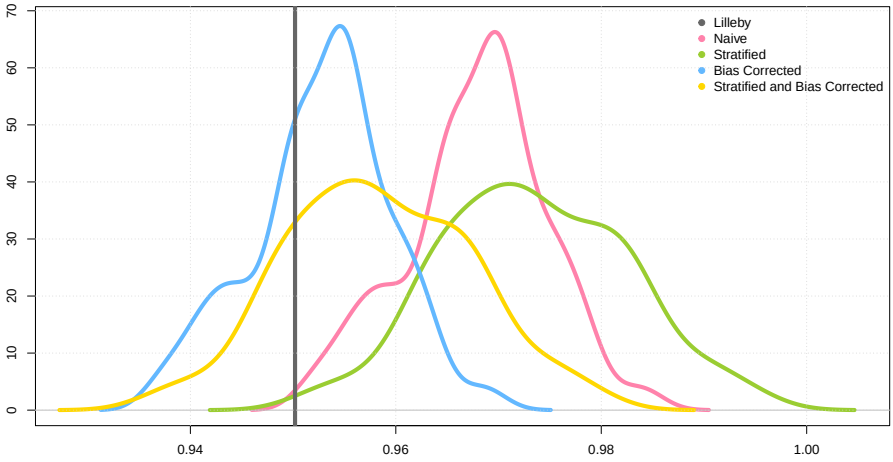


Figure 5.7: Histograms of the $m = 100$ naive (pink), stratified (green), bias corrected (blue) as well as the stratified and bias corrected (yellow) proportion estimates for the single outcome 'Not disabled'. The true Lilleby proportion (grey) is represented by a vertical line.

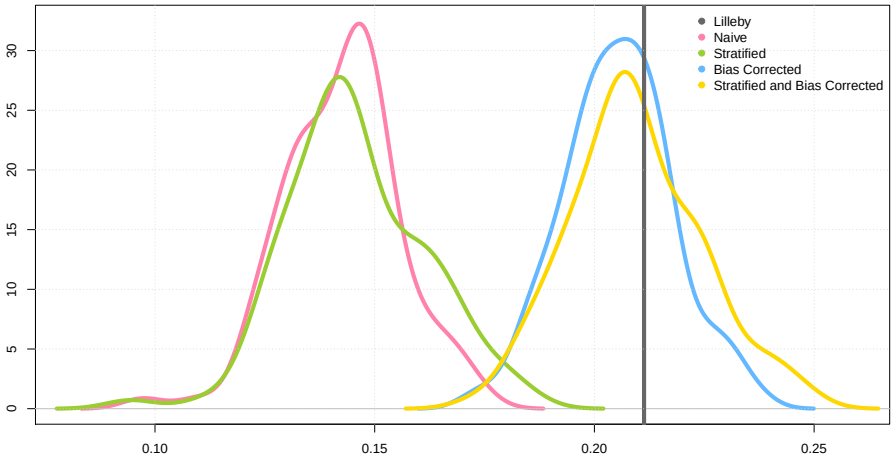


Figure 5.8: Histograms of the $m = 100$ naive (pink), stratified (green), bias corrected (blue) as well as the stratified and bias corrected (yellow) proportion estimates for the single outcome 'Lonely'. The true Lilleby proportion (grey) is represented by a vertical line.

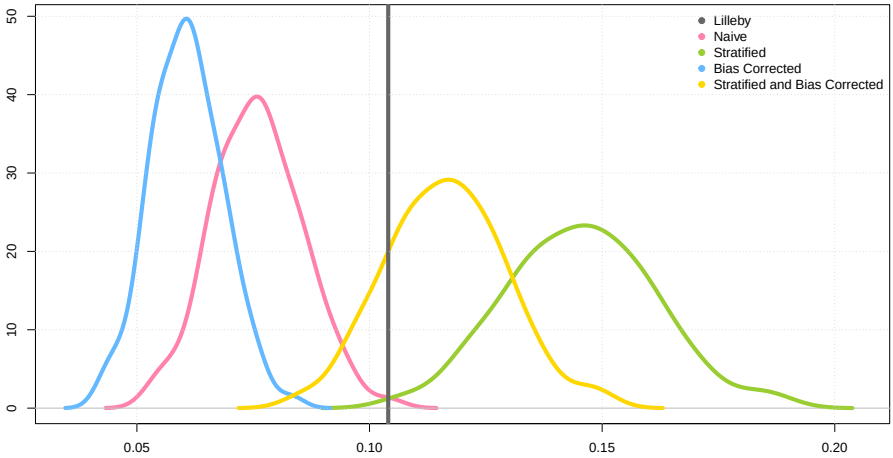


Figure 5.9: Histograms of the $m = 100$ naive (pink), stratified (green), bias corrected (blue) as well as the stratified and bias corrected (yellow) proportion estimates for the pair of outcomes 'Married and age 20-39'. The true Lilleby proportion (grey) is represented by a vertical line.

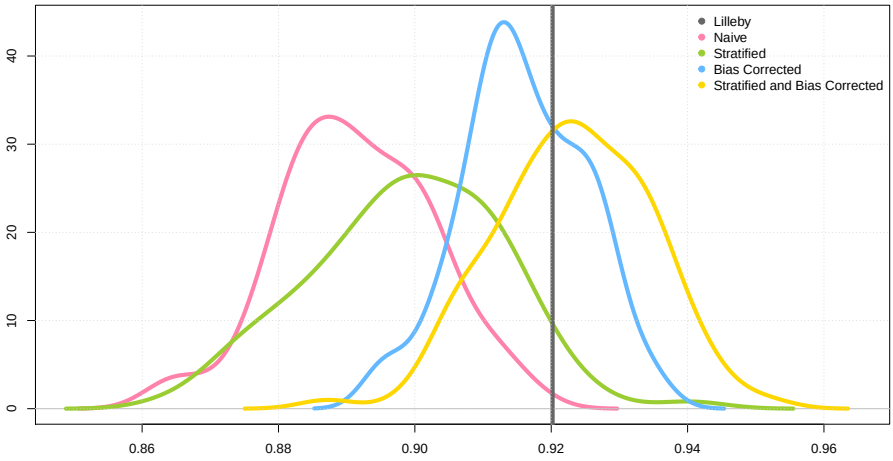


Figure 5.10: Histograms of the $m = 100$ naive (pink), stratified (green), bias corrected (blue) as well as the stratified and bias corrected (yellow) proportion estimates for the pair of outcomes 'Not child neglected and not disabled'. The true Lilleby proportion (grey) is represented by a vertical line.

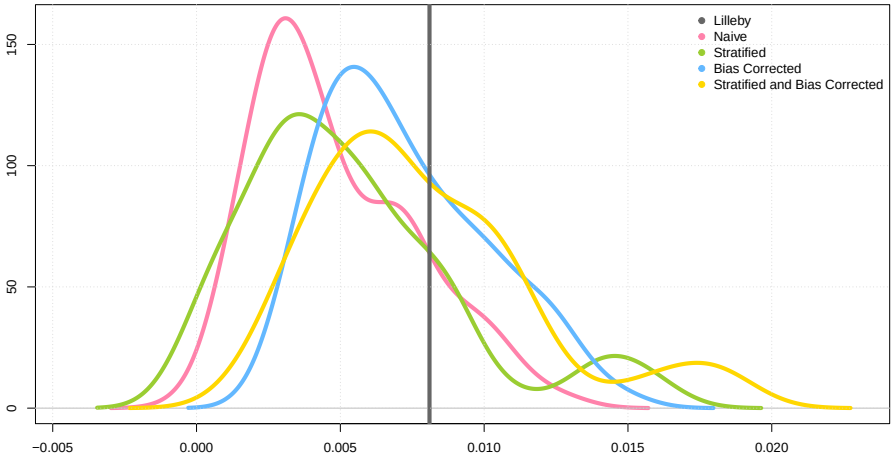


Figure 5.11: Histograms of the $m = 100$ naive (pink), stratified (green), bias corrected (blue) as well as the stratified and bias corrected (yellow) proportion estimates for the pair of outcomes 'Do believe in God and origin 5'. The true Lilleby proportion (grey) is represented by a vertical line.

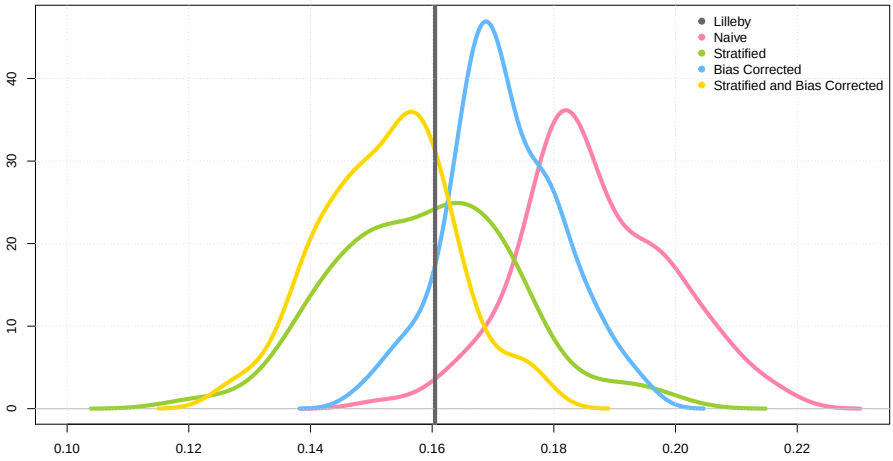


Figure 5.12: Histograms of the $m = 100$ naive (pink), stratified (green), bias corrected (blue) as well as the stratified and bias corrected (yellow) proportion estimates for the pair of outcomes 'High school and male'. The true Lilleby proportion (grey) is represented by a vertical line.

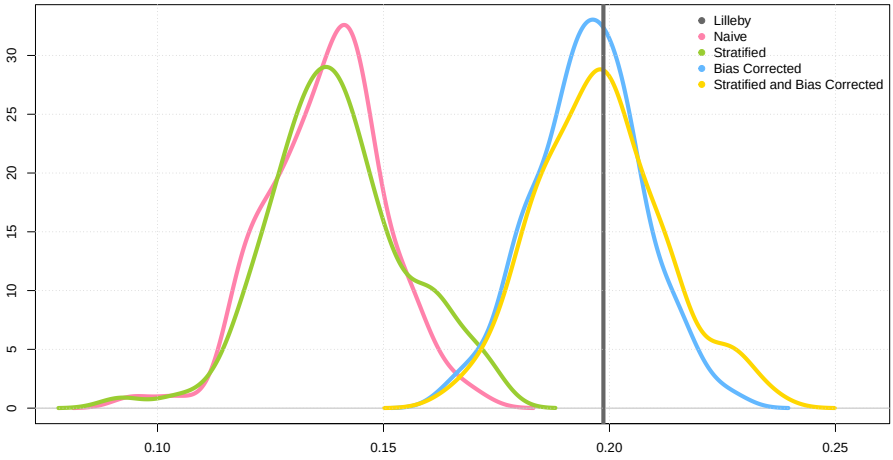


Figure 5.13: Histograms of the $m = 100$ naive (pink), stratified (green), bias corrected (blue) as well as the stratified and bias corrected (yellow) proportion estimates for the pair of outcomes 'Lonely and not disabled'. The true Lilleby proportion (grey) is represented by a vertical line.

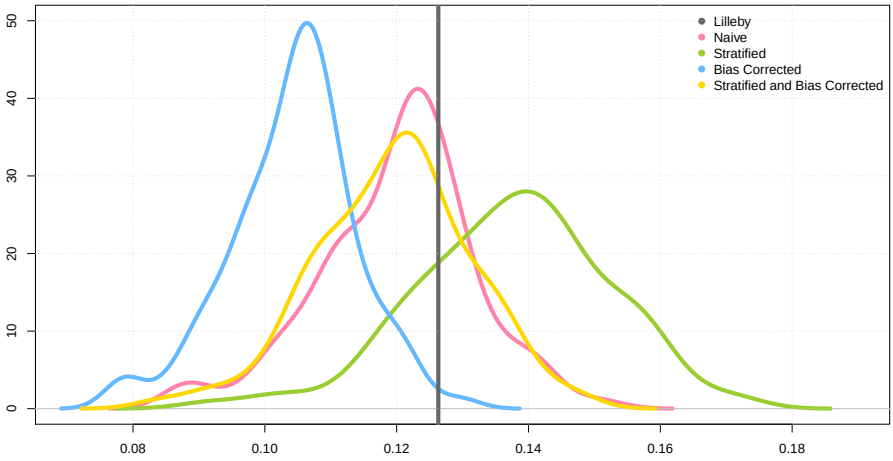


Figure 5.14: Histograms of the $m = 100$ naive (pink), stratified (green), bias corrected (blue) as well as the stratified and bias corrected (yellow) proportion estimates for the pair of outcomes 'Personal Christian and married'. The true Lilleby proportion (grey) is represented by a vertical line.

Centering and Spread

The centering of the m proportion estimates for each of the four types and the corresponding spread is presented in table 5.1 and 5.2 for the relevant single outcomes and pairs of outcomes, respectively. We focus on the latter.

The naive proportion estimates and the stratified proportion estimates are systematically centered away from the true Lilleby proportion. Each stratified and bias corrected proportion estimate is centered closest to the true Lilleby proportion compared to the bias corrected proportion estimate. The spread increases when stratification is applied to the naive proportion estimates. Furthermore, the spread is always larger in the stratified and bias corrected proportion estimates than in the bias corrected proportion estimates.

	Lilleby	Naive	Stratified	Bias Corrected	Stratified and Bias Corrected
Origin 1	0.910	0.924(0.009)	0.928(0.012)	0.916(0.009)	0.920(0.012)
Married	0.397	0.431(0.018)	0.492(0.023)	0.345(0.014)	0.394(0.018)
1 child	0.108	0.259(0.018)	0.258(0.021)	0.110(0.007)	0.110(0.008)
University (Higher Level)	0.184	0.201(0.014)	0.243(0.019)	0.154(0.011)	0.187(0.014)
Child neglected	0.041	0.037(0.007)	0.034(0.008)	0.045(0.007)	0.042(0.008)
Do not believe in God	0.630	0.564(0.019)	0.576(0.023)	0.626(0.016)	0.639(0.019)
Not disabled	0.950	0.968(0.007)	0.973(0.009)	0.953(0.007)	0.958(0.009)
Lonely	0.211	0.142(0.013)	0.146(0.016)	0.205(0.012)	0.209(0.015)

Table 5.1: The centering of proportion estimates with its spread for $m = 100$ response samples for the relevant single outcomes.

	Lilleby	Naive	Stratified	Bias Corrected	Stratified and Bias Corrected
Married / Age 20-39	0.104	0.076(0.010)	0.145(0.016)	0.061(0.008)	0.116(0.013)
Not child neglected / Not disabled	0.920	0.891(0.011)	0.899(0.014)	0.916(0.009)	0.923(0.011)
Do believe in God / Origin 5	0.008	0.005(0.003)	0.005(0.004)	0.007(0.003)	0.008(0.004)
High School / Male	0.160	0.187(0.012)	0.159(0.015)	0.172(0.009)	0.152(0.011)
Lonely / Not disabled	0.199	0.137(0.013)	0.140(0.015)	0.195(0.012)	0.198(0.014)
Personal Christian / Married	0.126	0.120(0.012)	0.138(0.015)	0.104(0.010)	0.119(0.012)

Table 5.2: The centering of proportion estimates with its spread for $m = 100$ response samples for the relevant pairs of outcomes.

Root Mean Squared Error

Table 5.3 and 5.4 present the RMSE found by equation (5.4) for the relevant single outcomes and pairs of outcomes, respectively. We focus on the latter.

The naive proportion estimates and the stratified proportion estimates have a RMSE that is systematically worse than for the bias corrected proportion estimates and the stratified and bias corrected proportion estimates. We compare the bias corrected proportion estimates to the stratified and bias corrected proportion estimates by their RMSE. The number of proportion estimates with the smallest RMSE for these two types are even. Thus, it is not possible to conclude whether the bias corrected proportion estimate or the stratified and bias corrected proportion estimate is the overall best fit.

	Lilleby	Naive	Stratified	Bias Corrected	Stratified and Bias Corrected
Origin 1	0	0.017	0.023	0.011	0.016
Married	0	0.039	0.098	0.054	0.018
1 child	0	0.152	0.152	0.007	0.008
University (Higher Level)	0	0.022	0.062	0.032	0.015
Child neglected	0	0.008	0.010	0.008	0.008
Do not believe in God	0	0.069	0.059	0.017	0.021
Not disabled	0	0.019	0.025	0.007	0.012
Lonely	0	0.070	0.067	0.014	0.015

Table 5.3: The root mean squared error of proportion estimates for $m = 100$ response samples for the relevant single outcomes.

	Lilleby	Naive	Stratified	Bias Corrected	Stratified and Bias Corrected
Married / Age 20-39	0	0.030	0.044	0.044	0.018
Not child neglected / Not disabled	0	0.031	0.026	0.010	0.012
Do believe in God / Origin 5	0	0.004	0.005	0.003	0.004
High School / Male	0	0.029	0.015	0.015	0.013
Lonely / Not disabled	0	0.063	0.061	0.012	0.014
Personal Christian / Married	0	0.013	0.018	0.024	0.014

Table 5.4: The root mean squared error of proportion estimates for $m = 100$ response samples for the relevant pairs of outcomes.

Summary

The bias correction model has major impact on the centering of the proportion estimate. The centering can be further improved by stratification but on the expense of somewhat larger spread.

Chapter 6

Concluding Remarks

Lilleby is simulated with its residents and the responses to a questionnaire survey is modelled. The residents of Lilleby are distributed according to an interaction model where the variables included are inspired by Oslo Monitor 1.0. Oslo Monitor 1.0 is thoroughly discussed and revised in the beginning of the thesis. The marginal distribution with corresponding parameters are defined for each variable. The concept of copulas is used to derive the interaction model to account for the interplay among the variables. A sequential simulation algorithm is used to sample the realizations. The simulated realizations are assumed to be the true city of Lilleby. A questionnaire is then distributed to a representative and stratified sample of the Lilleby population. The data collection deals with two major types of correction. The stratification model corrects for over- or under-representation and the posterior model corrects for potential prejudices in the responses. The response sample is compared to the true Lilleby by proportion estimates.

Plots of the univariate and bivariate characteristics of the residents of Lilleby confirm that the realized Lilleby reflects the marginal distributions from Oslo Monitor 1.0. Additionally, the dependence assumed to exist between the included variables while sampling from the interaction model is reflected in Lilleby. Histograms are plotted and the centering and spread are calculated to evaluate the sensitivity to the correction models in the response sample. Bias correction has major impact on the centering of the proportion estimate. The centering can be further improved by stratification but on the expense of somewhat larger spread. The bias corrected proportion estimate compared to the stratified and bias corrected proportion estimate by their RMSE calls them even. Still, the stratified and bias corrected proportion estimate is centered closest to the true Lilleby proportion compared to the bias corrected proportion estimate. Overall, the stratification model and, especially, the bias correction model appear as effective tools to correct for skewness in a response sample and to deal with the bias caused by potential prejudices in a statistical survey including the subjectivity and unpredictable behaviour of humans.

A more extensive analysis of the interaction model can be done. The interaction model can also be modified to include discrete and continuous variables as well as categorical variables. The likelihood model can be expanded to include responses when independence cannot be assumed. In addition, a thorough study should be made on how to avoid prejudices in the responses caused by the psychological aspects of answering a question-

naire.

To be able to apply statistics to problems arising in social science we need models that can handle non-ordered, categorical variables. There is also a need for tools to correct for the unpredictable nature of humans in the collection of data. As statisticians we possess a valuable and requested knowledge that might be used to solve complex social problems. We might contribute to make—not only a city better to live in for everybody—but even a world that is better to live in for everybody.

Bibliography

Literature

Casella, G. and R.L. Berger (2002). *Statistical Inference*. Brooks/Cole.

Cont, R. and P. Tankov (2004). *Financial Modelling With Jump Processes*. Chapman & Hall/CRC: Financial Mathematics Series.

Geer, S. van der (2019). *Mathematical Statistics*. Lecture notes. Eidgenössische Technische Hochschule Zürich.

Haff, I.H. (2012). *Pair-copula constructions - an inferential perspective*. PhD dissertation. University of Oslo.

Hellevik, O. (2015). *Hva betyr respondentbortfallet i intervjuundersøkelser?* Tidsskrift for samfunnsforskning (Universitetsforlaget) 56.2, pp. 211-231.

James, G. et al. (2017). *Introduction to Statistical Learning with Applications in R*. Springer.

Johannessen, A., P.A. Tufte, and L. Christoffersen (2016). *Introduksjon til samfunnsvitenskapelig metode*. Abstrakt forlag AS.

Nocedal, J. and S.J. Wright (2006). *Numerical Optimization*. Springer.

Walpole, R.E. et al. (2012). *Probability & Statistics*. Pearson.

Data Sources and References

CINT (2020). *About us*. URL: `https://www.cint.com/about`.

Epland, J. and M.I. Kirkeberg (2017). *Ett av ti barn tilhører en husholdning med vedvarende lavinntekt*. URL: `https://www.ssb.no/inntekt-og-forbruk/artikler-og-publikasjoner/ett-av-ti-barn-tilhorer-en-husholdning-med-vedvarende-lavinntekt`.

FHI (2018). *Forventet levealder i Norge*. URL: `https://www.fhi.no/nettpub/hin/befolkning/levealder/`.

Ingebretsen, T., J. Holbæk-Hanssen, and E. Dalen (2016). *Norsk Monitor 2015/16*. Tech. rep. IPSOS.

KANTAR (2020). *Analysebasert rådgivning på grunnlag av relevante data fra mange kilder*. URL: `https://kantar.no/om/om-kantar-tns/`.

NAV (2018). *Beregning av uføretrygd*. URL: `https://www.nav.no/no/person/pensjon/uforetrygd/beregning-av-uforetrygd`.

NAV (2019). *Uføretrygd*. URL: https://www.nav.no/no/Person/Pensjon/Uforetrygd#chapter-1.
NESH (2015). *About NESH*. URL: https://www.etikkom.no/en/our-work/about-us/the-national-committee-for-research-ethics-in-the-social-sciences-and-the-humanities-nesh/about-nesh/.
Nørgaard, E. (2013). *Samfunnsspeilet 4/2013*. Tech. rep. SSB.
NSD (2018). *Nettbaserte spørreundersøkelser*. URL: https://nsd.no/personvernombud/hjelp/forskningsmetoder/nettbaserte_sporreundersokelser.html.
— (2019a). *Do I have to notify my project?* URL: https://nsd.no/personvernombud/en/notify/index.html.
— (2019b). *Hva må jeg informere om?* URL: https://nsd.no/personvernombud/hjelp/informere_om.html.
— (2020). *NSD - Norwegian Centre for Research Data*. URL: https://nsd.no/nsd/english/index.html.
NSPCC (2007). *Child protection research briefing: Child neglect*. Tech. rep. NSPCC.
Oslo Kommune (2015). *Bydelsstatistikken 2015*. URL: https://www.oslo.kommune.no/politikk-og-administrasjon/statistikk/statistiske-publikasjoner/bydelsstatistikken/bydelsstatistikken-2015/.
— (2019). *Folkemengde og endringer*. URL: https://www.oslo.kommune.no/politikk-og-administrasjon/statistikk/befolkning/folkemengde-og-endringer/#gref.
Rafoss, T.W. (2017). *Nordmenns Bibelbruk*. Tech. rep. KIFO.
Skatteetaten (2019). *Grunnbeløpet i folketrygden*. URL: https://www.skatteetaten.no/satser/grunnbelopet-i-folketrygden/.
SSB (2015). *Personer 0-17 år, etter alder og kjønn*. URL: https://www.ssb.no/a/barnogunge/2015/tabeller/befolkning/bef0003.html.
— (2018a). *05196: Befolkning, etter kjønn, statsborgerskap, alder, statistikkvariabel og år*. URL: https://www.ssb.no/statbank/table/05196/tableViewLayout1/.
— (2018b). *05703: Barn som opplevde skilsmisse, etter statistikkvariabel og år*. URL: https://www.ssb.no/statbank/table/05703/tableViewLayout1/.
— (2018c). *07459: Befolkning, etter region, kjønn, alder, statistikkvariabel og år*. URL: https://www.ssb.no/statbank/table/07459/tableViewLayout1/.
— (2018d). *Fakta om religion*. URL: https://www.ssb.no/kultur-og-fritid/faktaside/religion.
— (2018e). *Gjennomføring i videregående opplæring*. URL: https://www.ssb.no/utdanning/statistikker/vgogjen.
— (2018f). *Samboere, 2014-2016*. URL: https://www.ssb.no/befolkning/statistikker/samboer/aar/2018-01-23?fane=tabell#content.
— (2018g). *Satellittregnskap for ideelle og frivillige organisasjoner*. URL: https://www.ssb.no/nasjonalregnskap-og-konjunkturer/statistikker/orgsat/aar.
— (2019a). *Befolkning*. URL: https://www.ssb.no/befolkning/statistikker/folkemengde.

— (2019b). *Fakta om befolkningen*. URL: https://www.ssb.no/befolkning/faktaside/befolkningen.
Statistikkbanken (2019). *Statistikkbanken*. URL: http://statistikkbanken.oslo.kommune.no/webview/.
Talset, O.T. (2018). *Oslo Monitor 1.0*. Tech. rep. Tankesmien Skaperkraft.
Tønder, J.K. (2009). *Samfunnsspeilet 1/2009*. Tech. rep. SSB.

Appendix A

Questionnaire: The Social and Spiritual Situation in Lilleby

Dear resident of Lilleby. Thank you for participating in this 3-5 minutes survey and helping us collect information about the social and spiritual situation in our city. Our goal is to make Lilleby a better city to live in for everybody. Your answers will be processed anonymously and the privacy is according to the GDPR regulations. You are free to withdraw your answers from the process at any given point without any explanation. Thank you!

General Characteristics

1. **What is your gender?**
   - ☐ Male
   - ☐ Female
2. **Which age group do you belong to?**
   - ☐ 15-19
   - ☐ 20-39
   - ☐ 40-49
   - ☐ 50-69
   - ☐ 70+
3. **Which district do you currently live in?**
   - ☐ West
   - ☐ South
   - ☐ North
   - ☐ East

4. **Where was your mother raised?**
   - ☐ North of Europe including Norway and the Nordic countries as well as North America and Oceania
   - ☐ Middle of Europe, South of Europe and South America
   - ☐ East of Europe and North of Asia including Russia among others
   - ☐ The Far East including China, Korea, Japan, India and Thailand among others
   - ☐ Africa and The Near East including The Middle East, Iran, Afghanistan and Turkey among others

5. **How many children do you have?**
   - ☐ No children
   - ☐ 1 child
   - ☐ 2 children
   - ☐ 3 children or more

6. **Which of the following statements describes your current situation most accurately?**
   - ☐ I am single or a widow/widower
   - ☐ I am in a relationship but do NOT live together with my partner
   - ☐ I live together with my partner
   - ☐ I live together with my husband/wife

Social Background

7. **What is the highest level of education you have completed?**
   - ☐ Not applicable
   - ☐ Elementary school
   - ☐ High school
   - ☐ Less than 4 years of University
   - ☐ 4 years or more of University

8. **What income level in thousands of Norwegian Kroner (NOK) per year do you belong to?**
   - ☐ 0-199
   - ☐ 200-399
   - ☐ 400-599
   - ☐ 600-799
   - ☐ 800+

9. **Are you prevented from working and therefore receive a disability benefit?**
   - ☐ Yes
   - ☐ No
10. **Have you ever received help from the Child Welfare because you were neglected as a child?**
   - ☐ Yes
   - ☐ No
   - ☐ I don't know
11. **Do you often or every so often find yourself lonely?**
   - ☐ Yes
   - ☐ No
   - ☐ I don't know

Spiritual Situation

12. **How often do you read the Bible?**
   - ☐ Never
   - ☐ Only when attending church for Christian holidays or celebrations
   - ☐ Some times a year
   - ☐ Once a month
   - ☐ Once a week
   - ☐ Every day
13. **Do you attend a church at a weekly basis?**
   - ☐ Yes
   - ☐ No
14. **Do you believe in a monotheistic God?**
   - ☐ Yes
   - ☐ No
   - ☐ I don't know
15. **Do you define yourself as a Christian with a personal relationship with God?**
   - ☐ Yes
   - ☐ No
   - ☐ I don't know

NTNU

Norwegian University of
Science and Technology