UNIVERSITY OF GOTHENBURG

MASTER'S THESIS

Optimal Auxiliary Variable Assisted Two-Phase Sampling Designs

HENRIK IMBERG

Department of Mathematical Sciences
Division of Mathematical Statistics
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2016

Thesis for the Degree of Master of Science

Optimal Auxiliary Variable Assisted Two-Phase Sampling Designs

Henrik Imberg

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and University of Gothenburg
SE – 412 96 Gothenburg, Sweden
Gothenburg, May 2016

Matematiska vetenskaper
Göteborg 2016

Abstract

Two-phase sampling is a procedure in which sampling and data collection is conducted in two phases, aiming at achieving increased precision in estimation at reduced cost. The first phase typically involves sampling a large number of elements and collecting data on variables that are easy to measure. In the second phase, a subset is sampled for which all variables of interest are observed. Utilization of the information provided by the data observed in the first phase may increase precision in estimation by optimal selection of sampling design the second phase.

This thesis deals with two-phase sampling when a random sample following some general parametric statistical model is drawn in the first phase, followed by subsampling with unequal probabilities in the second phase. The method of maximum pseudo-likelihood estimation, yielding consistent estimators under general two-phase sampling procedures, is presented. The design influence on the variance of the maximum pseudo-likelihood estimator is studied. Optimal subsampling designs under various optimality criteria are derived analytically and numerically using auxiliary variables observed in the first sampling phase.

Keywords: Anticipated variance; Auxiliary information in design; Maximum pseudo-likelihood estimation; Optimal designs; Poisson sampling; Two-phase sampling.

Acknowledgements

I would like to thank my supervisors, Vera Lisovskaja and Olle Nerman, for guidance during this project. I am especially grateful to Vera for sharing thoughts and ideas from your manuscripts, on which much of the material in this thesis is based.

Henrik Imberg, Gothenburg, May 2016

Contents

1 Introduction **1**
- 1.1 Background . . . 2
- 1.2 Purpose . . . 2
- 1.3 Scope . . . 3
- 1.4 Outline . . . 3

2 Theoretical Background **5**
- 2.1 A General Two-Phase Sampling Framework . . . 5
- 2.2 Two Approaches to Statistical Inference . . . 7
  - 2.2.1 Maximum Likelihood . . . 7
  - 2.2.2 Survey Sampling . . . 13
- 2.3 Maximum Pseudo-Likelihood . . . 17
  - 2.3.1 Topics in Related Research . . . 24
- 2.4 Optimal Designs . . . 26

3 Optimal Sampling Schemes under Poisson Sampling **31**
- 3.1 The Variance of the PLE under Poisson Sampling . . . 32
  - 3.1.1 The Total Variance . . . 32
  - 3.1.2 The Anticipated Variance . . . 34
- 3.2 Optimal Two-Phase Sampling Designs . . . 37
  - 3.2.1 L-Optimal Sampling Schemes . . . 37
  - 3.2.2 D and E-optimal Sampling Schemes . . . 39
- 3.3 Some Modifications . . . 40
  - 3.3.1 Adjusted Conditional Poisson Sampling . . . 40
  - 3.3.2 Stratified Sampling . . . 41
  - 3.3.3 Post-Stratification . . . 42

4 Examples **44**
- 4.1 The Normal Distribution . . . 44
  - 4.1.1 L-Optimal Designs for (μ, σ) . . . 44

4.1.2 D and E-Optimal Designs for (μ, σ) 49
4.1.3 Optimal Sampling Schemes for μ Revisited 50
4.2 Logistic Regression 53
4.2.1 A Single Continuous Explanatory Variable 55
4.2.2 Auxiliary Information with Proper Design Model 59
4.2.3 Auxiliary Information with Improper Design Model 63
4.2.4 When the Outcome is Unknown 65

5 Conclusion **66**

Appendices **69**

A The Variance of the Maximum Pseudo-Likelihood Estimator **69**
A.1 Derivation of the Asymptotic Conditional Variance 69
A.2 On the Contributions to the Realized Variance 70

B Derivation of L-Optimal Sampling Schemes under Poisson Sampling **72**

List of Symbols

$\mathcal{P}$	Infinite target population.
S_1	First phase sample, random sample from $\mathcal{P}$. Elements denoted by k,l etc.
S_2	Second phase sample, probability sample from S_1.
I_k	Sample inclusion indicator variable, $I_k = 1$ if $k \in S_2$ and 0 else.
π_k	First order inclusion probability, $\pi_k = P(k \in S_2) = P(I_k = 1)$.
π_{kl}	Second order inclusion probability, $\pi_{kl} = P(k,l \in S_2) = P(I_k = 1, I_l = 1)$.
N, $\lvert S_1 \rvert$	Size of first phase sample.
n, $\lvert S_2 \rvert$	(Expected) size of second phase sample.
$\boldsymbol{Y}$	Outcome, response variable.
$\boldsymbol{X}$	Explanatory variable.
$\boldsymbol{Z}$	Auxiliary variable.
Y_k, $\boldsymbol{X}_k$, $\boldsymbol{Z}_k$	Study variables corresponding to element k.
$\boldsymbol{y}$, $\tilde{\boldsymbol{X}}$, $\tilde{\boldsymbol{Z}}$	Realizations of study variables.
y_k, $\boldsymbol{x}_k$, $\boldsymbol{z}_k$	Realized study variables corresponding to element k.
$f(y_k \lvert \boldsymbol{x}_k; \boldsymbol{\theta})$	Model of interest.
$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$	Parameter of interest.
$f(y_k, \boldsymbol{x}_k \lvert \boldsymbol{z}_k; \boldsymbol{\phi})$	Design model.
$\boldsymbol{\phi}$	Design (model) parameter, assumed to be known.

$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}, \tilde{\boldsymbol{X}})$	Likelihood function.
$\ell(\boldsymbol{\theta}; \boldsymbol{y}, \tilde{\boldsymbol{X}})$	Log-likelihood function.
$\boldsymbol{S}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \boldsymbol{y}, \tilde{\boldsymbol{X}})$	Score, gradient of log-likelihood function.
$\hat{\boldsymbol{\theta}}_{ML}$	Maximum likelihood estimator (MLE).
$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}$	Asymptotic variance-covariance matrix of MLE.
$\boldsymbol{I}(\boldsymbol{\theta})$	Fisher information matrix.
$\ell_{\pi}(\boldsymbol{\theta}; \boldsymbol{y}, \tilde{\boldsymbol{X}})$	Pseudo log-likelihood function.
$\hat{\boldsymbol{\theta}}_{\pi}$	Maximum pseudo-likelihood estimator (PLE).
$\boldsymbol{S}_{\pi}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \ell_{\pi}(\boldsymbol{\theta}; \boldsymbol{y}, \tilde{\boldsymbol{X}})$	π-expanded score, gradient of pseudo log-likelihood function.
$\widetilde{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}}$	Asymptotic variance-covariance matrix of PLE.

1

Introduction

In many areas of research, data collection and statistical analysis play a central role in the acquisition of new knowledge. However, collection of data is often associated with some cost, and in studies involving human subjects possibly also with discomfort and potential harm. There are often also statistical demands on the analysis, namely that the characteristics or parameters of interest should be estimated with sufficient precision. Efficient use of data is thus tractable for economical, ethical and statistical reasons. The precision in estimation depend on the number of observations available for analysis as well as on the study design, i.e. the conditions under which the study is conducted in combination with the methods used for sample selection.

A special situation arise when some information about the elements or subjects available for study is accessible prior to sampling. Incorporation of such information in design and analysis of a study can improve the precision in estimation substantially. In practice, such information is seldom available prior to study but rather obtained through a first sampling phase, collecting data of variables that are easily measured for a large number of subjects. A sampling procedure in which sampling and data collection is performed in two phases is called two-phase sampling, and could be used to meet the statistical and economical demands encountered in empirical research.

While two-phase sampling provides an opportunity to select elements that are believed to contribute with much information to the analysis, it also introduces a number of challenges. It is important to use methods of estimation that properly account for the sampling procedure, and to understand how the selection of elements influence the precision in estimation of the parameters of interest. The former is necessary in order to obtain valid inferences, the latter in order to be able to use the data available in an efficient way.

1.1 Background

Two-phase sampling as a tool to achieve increased precision in estimation in studies with economical limitations was proposed by Neyman [34] within the context of survey sampling. It is a procedure in which sampling is conducted in two phases, the first involving a large sample and collection of information that is easily obtained, the second involving a smaller sample in which the variables of interest are observed. The idea is that use of easy accessible data could aid in the collection of data from more expensive sources.

The variables observed in the first phase that are called auxiliary variables. These are not necessarily of particular interest themselves, but can be used in design and analysis of a study to increase precision in estimation. It is assumed that the variables of interest are associated with a high cost, making it unfeasible to observe these for all elements in the first phase and profitable to collect other information for a large number of elements in the first phase. The high cost could for example be due to need for interviews be carried out or measurements to be made by trained staff in order to assess or measure the variables of interest. It is also assumed that the auxiliary variables are related to the variables of interest.

The use of auxiliary variables in design, estimation and analysis is well studied within the field of survey sampling, see for example Särndal et al. [45]. It is however less frequently encountered among practitioners in other statistical disciplines. The use of two-phase sampling in case-control studies has been suggested by Walker [47] and White [48], and in clinical trials by Frangakis and Baker [15]. Another possible area of application is to naturalistic driving studies, such as the recently conducted European Field Operational Test (EuroFOT) study [1]. This study combines data from different sources, including video sequences continuously filmed in the drivers cabin as well as automatically measured data, such as speed, acceleration, steering wheel actions and GPS coordinates. The access to automatically generated data could possibly be used for efficient selection of video sequences for annotation and analysis.

Optimal subsampling designs using auxiliary information have previously been studied in the literature, see for example Jinn et al. [27], Reilly and Pepe [38,39] and Frangakis and Baker [15]. Much of the previous work in the area is however limited in the classes of estimators and models considered.

1.2 Purpose

The aim of this thesis is to derive optimal subsampling designs for a general class of estimators and statistical models, using auxiliary information obtained in the first sampling phase to optimize the sampling design in the second phase.

1.3 Scope

The work is restricted to the use of auxiliary information in the design stage, using the method of maximum pseudo-likelihood for estimation. The pseudo-likelihood is closely related to the classical likelihood, with some modifications for use under general sampling designs. In its classical form, it does not incorporate auxiliary information in estimation.

This thesis deals with two-phase sampling when a random sample following some general parametric model is drawn in the first phase, followed by Poisson sampling in the second phase. Poisson sampling is a sampling design in which elements are sampled independently of each other, possibly with unequal probabilities. Total independence in sampling of elements leads to important simplifications of the optimization problem, while the use of unequal probabilities allows for construction of flexible designs.

Some minor excursions from the above delimitation are made, introducing other designs or auxiliary information in estimation post hoc. Adjusted conditional Poisson designs and stratified sampling, as well as the use of auxiliary information in estimation by sampling weight adjustment, are mentioned.

1.4 Outline

This thesis is divided into five chapters, including the current one. Chapter 2 gives a general formulation of the two-phase sampling procedure and presents the framework for the situations considered in the thesis. The essentials in maximum likelihood estimation and survey sampling are described, and the method of maximum pseudo-likelihood estimation is presented. Some topics in optimal design theory are also covered. The aim is to present the most important topics and results in some generality without being too technical. Focus is thus on ideas and results rather than on proofs. References to some specific results are given in the text as presented, while references covering broader topics are given at the end of each paragraph under the section *Perspective and Sources*. This section also contains some historical remarks and comments on the material. Examples illustrating the theory and techniques presented is the thesis are given under the section *Illustrative Examples*. Many of these concern the normal distribution. It is chosen due to its familiar form and well known properties, which enables for focus to be on the new topics. Many of the examples are also related and it might be necessary to return to previous examples for details left out.

The main results of this thesis is presented are Chapter 3, investigating the use of auxiliary information for selection of subsampling design. This chapter is restricted to certain classes of sampling designs, for which optimal sampling schemes are derived with respect to various optimality criteria. Some post hoc adjustments of design and methods for estimation are discussed.

The performance of the subsampling designs derived in Chapter 3 are illustrated by a number of examples in Chapter 4. These include estimation of parameters of the normal distribution and in logistic regression models, with various amount of information available in the design stage. Rather simple models are considered in order to ease

interpretation and understanding.

In the last chapter, limitations and practical implications of the work is discussed. Some of the theoretical material is presented in Appendix.

2

Theoretical Background

The main ideas about two-phase sampling are presented and the framework for the situations considered in the thesis is described. The main principles of maximum likelihood estimation and survey sampling are described. Estimation under two-phase sampling, using the method of maximum pseudo-likelihood, is presented. Some topics in optimal design theory needed for comparison, evaluation and optimization of two-phase sampling design are discussed.

2.1 A General Two-Phase Sampling Framework

Consider a situation in which sampling from an infinite population $\mathcal{P}$ is conducted in two phases. In the first phase, a random sample $S_1 = \{e_1,e_2, e_3, \ldots, e_N\}$ of N elements is drawn from the target population. To simplify notation, let k represent element e_k in S_1. Associated with each element is a number or random variables, namely an outcome or response variable Y_k, explanatory variables $\boldsymbol{X}_k$ and auxiliary variables $\boldsymbol{Z}_k$. Statistical independence of the triplets $(Y_k,\boldsymbol{X}_k, \boldsymbol{Z}_k)$ between elements is assumed. Let $\boldsymbol{Y}$ be the vector with elements Y_k and denote by $\boldsymbol{X}$ and $\boldsymbol{Z}$ the matrices with rows $\boldsymbol{X}_k$ and $\boldsymbol{Z}_k$ respectively. The role of the explanatory variables are to describe the outcome through some statistical model on which inference about the target population is based. The role of the auxiliary variables are to provide information about the response and/or the explanatory variables before these are observed, which can be used in the planning of design. It is not required that $\boldsymbol{Z}$ is disjoint from $(\boldsymbol{Y}, \boldsymbol{X})$.

Conditional on the explanatory variables, Y_k are assumed to be independent and follow some distribution law with density $f(y_k|\boldsymbol{x}_k;\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ is the parameter of interest. The aim is to estimate $\boldsymbol{\theta}$, or possibly a subset or specific linear combination of its elements. As an example one may think of logistic regression, in which $f(y_k|\boldsymbol{x}_k;\boldsymbol{\theta})$ is the probability mass function of a $Bernoulli(p_k)$ distributed random variable with $p_k = 1/(1 + e^{-\boldsymbol{x}_k^T\boldsymbol{\beta}})$. The parameter of interest is the vector of regression

coefficients $\boldsymbol{\theta} = \boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ or possibly a subset or linear combination of those. One might also be interested in the simpler situation without explanatory variables. It is then assumed that all Y_k are independent and identically distributed with some density $f(y; \boldsymbol{\theta})$.

The realizations of $(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z})$, generated from the underlying population when drawing S_1, are denoted by $\boldsymbol{y}$, $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Z}}$, respectively. The k-th element of $\boldsymbol{y}$ is denoted by y_k, which is the realized value of the response variable for element k in S_1. Similarly, the rows in the matrices $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Z}}$ corresponding to element k are denoted by $\boldsymbol{x}_k$ and $\boldsymbol{z}_k$, respectively. If measurement of some components in $(\boldsymbol{y}, \tilde{\boldsymbol{X}})$ is associated with a high cost, the need for a second sampling phase is introduced by infeasibility of observing all of $(\boldsymbol{y}, \tilde{\boldsymbol{X}})$ for all elements in S_1. It is thus not possible to estimate $\boldsymbol{\theta}$ from the first sample, since the outcome or some of the explanatory variables are unknown. A second sampling phase is thus conducted.

The second phase sample, with sample size or expected sample size n, is denoted by S_2. The method of sampling can be such that elements are sampled with unequal probabilities. It turns out that the precision in estimation depend on the method of sampling, and it is desirable to find a sampling design that yields a high precision. This can be achieved by use of the auxiliary variables in the planning of design, since these introduce knowledge about $(\boldsymbol{y}, \tilde{\boldsymbol{X}})$ between the two phases of sampling. This requires some prior knowledge about the relationship between auxiliary variables and outcome and explanatory variables.

It is assumed in this thesis that a model for $(Y_k, \boldsymbol{X}_k)$ conditional on $\boldsymbol{Z}_k$, described by some density function $f(y_k, \boldsymbol{x}_k | \boldsymbol{z}_k; \boldsymbol{\phi})$ with parameter vector $\boldsymbol{\phi}$, is known to some extent prior to study. This model will be referred to as the design model and its parameter as the design parameter, and the use of this model will be restricted to determination of the sampling procedure in the second phase. The design model need not be completely known and must in practice often be guessed. However, a good agreement between guessed and true model is desirable for the methods described in this thesis to be used successfully. In the case of a continuous variable Y_k and no explanatory variables, the design model for Y_k conditional on $\boldsymbol{Z}_k$ could for example be a linear regression model, so that $Y_k | \boldsymbol{Z}_k \sim \mathcal{N}(\boldsymbol{Z}_k^T \boldsymbol{\beta}, \sigma_\varepsilon)$. If the parameter $\boldsymbol{\phi} = (\beta_1, \ldots, \beta_r, \sigma_\varepsilon)$ is known to some extent prior to study and $\boldsymbol{Z}_k$ explains some of the variation in Y_k, knowledge about $\boldsymbol{z}_k$ gives information about the distribution of Y_k. Such information can be of great importance in the choice of subsampling design in the second phase.

Once the subsample S_2 is drawn, the realizations (y_k,$\boldsymbol{x}_k$) are observed for the sampled elements. Estimation of $\boldsymbol{\theta}$ can then be carried out from the sampled elements in the second phase sample. However, the distribution of Y_k given $\boldsymbol{X}_k$ in the sample might differ from the underlying population distribution, since S_2 is not necessarily a simple random sample. The sampling procedure must be properly taken into account in the analyses in order to obtain valid inference about $\boldsymbol{\theta}$. One alternative is to use the method of maximum pseudo-likelihood, which is introduced in Section 2.3.

A flowchart presenting the two-phase sampling procedure is presented in Figure 2.1. The key feature in two-phase sampling is that some information about the elements in

S_1 is available between the two sampling phases by observation of the auxiliary variables. Efficient use of the auxiliary information in the planning of subsampling design might improve precision in estimation.

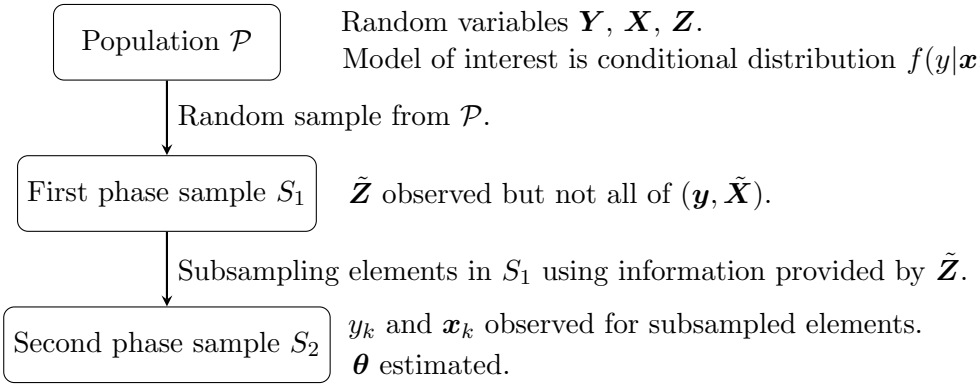

Figure 2.1: Flowchart describing the two-phase sampling procedure.

2.2 Two Approaches to Statistical Inference

Two random processes are involved in the two-phase sampling procedure considered in this thesis. In the first phase, randomness is introduced by the distribution of $(\boldsymbol{Y}, \boldsymbol{X})$ in the underlying population, which is described by some statistical model. In the second phase, randomness is introduced by subsampling of elements in S_1. This random process is fully described by the sample selection procedure. An inference procedure that properly accounts for both sources of randomness is required and will be introduced in Chapter 2.3. Before that, two different types of inference procedures, dealing with the two random processes separately, will be discussed.

2.2.1 Maximum Likelihood

Consider a random sample S_1 of N elements from an infinite population $\mathcal{P}$. Associated with each sampled element is some variables $(y_k, \boldsymbol{x}_k)$, generated from some population model for which inference is to be made. Conditional on the explanatory variables, the response variables Y_k are assumed to be independent and to have density $f(y_k|\boldsymbol{x}_k;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the parameter of interest. Estimation of $\boldsymbol{\theta}$ is often carried out using the method of maximum likelihood, which now will be described.

The Maximum Likelihood Estimator

The *maximum likelihood estimator* (MLE) of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}_{ML}$, is defined by

$$\hat{\boldsymbol{\theta}}_{ML} := \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}, \tilde{\boldsymbol{X}}) \ ,$$

where the *likelihood* $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}, \tilde{\boldsymbol{X}})$ is the joint density of $\boldsymbol{Y}$ given $\boldsymbol{X}$ seen as a function of $\boldsymbol{\theta}$. Due to independence, the likelihood function can be written as

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}, \tilde{\boldsymbol{X}}) = \prod_{k \in S_1} f(y_k | \boldsymbol{x}_k; \boldsymbol{\theta}) \ .$$

In place of the likelihood, it is often more convenient to work with the *log-likelihood*

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}, \tilde{\boldsymbol{X}}) := \log \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}, \tilde{\boldsymbol{X}}) = \sum_{k \in S_1} \log f(y_k | \boldsymbol{x}_k; \boldsymbol{\theta}) \ , \tag{2.2.1}$$

which has the same argmax as the likelihood. The argument $k \in S_1$ under the sum will be omitted from now on, simply writing sum over k.

The solution to (2.2.1) is found by solving the *estimating equation*

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \boldsymbol{y}, \tilde{\boldsymbol{X}}) = \mathbf{0},$$

where

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \boldsymbol{y}, \tilde{\boldsymbol{X}}) = \left(\frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y}, \tilde{\boldsymbol{X}})}{\partial \theta_1}, \ldots, \frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y}, \tilde{\boldsymbol{X}})}{\partial \theta_p} \right)$$

is the gradient of the log-likelihood. It is often also called the *score* and will be denoted by $\boldsymbol{S}(\boldsymbol{\theta})$ to simplify notation. Strictly speaking, finding the global maximum of (2.2.1) requires all critical points of the log-likelihood to be considered and the boundary of the parameter space to be investigated, following standard procedures in multivariate calculus. The examples presented in this thesis will however only be concerned with finding the solutions to the estimating equation (2.2.1), leaving the additional steps to the reader for verification that a global maximum is found.

Asymptotic Properties of the MLE

The asymptotic distribution of a maximum likelihood estimator is multivariate normal with

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}) \underset{a}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}) \ ,$$

using the notation "$\underset{a}{\sim}$" for the asymptotic distribution of a random variable. The variance-covariance matrix $\boldsymbol{\Gamma}$ called the *asymptotic variance* of the normalized MLE. The MLE is *asymptotically unbiased*, which we write

$$\mathrm{E}(\hat{\boldsymbol{\theta}}_{ML}) \underset{a}{=} \boldsymbol{\theta} \quad \Leftrightarrow \quad \mathrm{E}(\hat{\boldsymbol{\theta}}_{ML}) \to \boldsymbol{\theta} \text{ as } N \to \infty \ ,$$

using the notation "$\underset{a}{=}$" for equalities that hold in the limit. Furthermore, the bias of the MLE is relatively small compared to the standard error. This implies that the MLE is approximately unbiased for large samples and the bias can be neglected. Also, the MLE converge in distribution to the constant $\boldsymbol{\theta}$ as N tends to infinity, and we say that the MLE is *consistent*. That is, the distribution of $\hat{\boldsymbol{\theta}}_{ML}$ is tightly concentrated

around $\boldsymbol{\theta}$ for large samples, so that the MLE with high certainty will be within an arbitrary small neighborhood of the true parameter if N is large enough. The MLE is also *asymptotically efficient*, which roughly speaking is to say that the MLE has minimal asymptotic variance.

Note that unbiasedness and normality of MLE is guaranteed only in the limit as the sample size tend to infinity. However, for finite samples it is reasonable to think of asymptotic equalities as large sample approximations and of to use asymptotic distributions as large sample approximations of the sample distribution of an estimator.

The Fisher Information and the Variance of the MLE

The asymptotic distribution of the MLE can also be written as

$$\hat{\boldsymbol{\theta}}_{ML} \underset{a}{\sim} \mathcal{N}\left(\boldsymbol{\theta}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}\right) \ .$$

The variance-covariance matrix $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}$ of the MLE is the inverse of the so called *Fisher information* $\boldsymbol{I}(\boldsymbol{\theta})$:

$$\begin{aligned} \boldsymbol{I}(\boldsymbol{\theta}) &= \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{X}}\left[\sum_k \nabla_{\boldsymbol{\theta}} \log f(Y_k|\boldsymbol{x}_k;\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^T \log f(Y_k|\boldsymbol{x}_k;\boldsymbol{\theta})\right] \\ &= \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{X}}\left[-\boldsymbol{\partial}_{\boldsymbol{\theta}} \boldsymbol{S}(\boldsymbol{\theta})\right] \ , \end{aligned} \tag{2.2.2}$$

where $\boldsymbol{\partial}_{\boldsymbol{\theta}} \boldsymbol{S}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}}^T \ell(\boldsymbol{\theta}; \boldsymbol{y}, \tilde{\boldsymbol{X}})$ is the Hessian of the log-likelihood. The Fisher information will also be referred to as the *information matrix*. The elements of (2.2.2) are given by

$$\begin{aligned} \boldsymbol{I}(\boldsymbol{\theta})_{(i,j)} &= \sum_k \mathrm{E}_{Y_k|\boldsymbol{X}_k}\left[\frac{\partial \log f(Y_k|\boldsymbol{x}_k;\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \log f(\boldsymbol{\theta}; Y_k, \boldsymbol{x}_k)}{\partial \theta_j}\right] \\ &= \sum_k \mathrm{E}_{Y_k|\boldsymbol{X}_k}\left[-\frac{\partial^2 \log f(Y_k|\boldsymbol{x}_k;\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right] \ . \end{aligned}$$

Typically, the Fisher information depends on the values of the explanatory variables $\tilde{\boldsymbol{X}}$ as well as on the parameter $\boldsymbol{\theta}$. However, if Y_k are independent and identically distributed and there are no explanatory variables, the information matrix simplifies to

$$\begin{aligned} \boldsymbol{I}(\boldsymbol{\theta})_{(i,j)} &= N \, \mathrm{E}_Y\left[\frac{\partial \log f(Y;\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \log f(Y;\boldsymbol{\theta})}{\partial \theta_j}\right] \\ &= N \, \mathrm{E}_Y\left[-\frac{\partial^2 \log f(Y;\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right] \ . \end{aligned} \tag{2.2.3}$$

The variance-covariance matrix can be estimated by the inverse of the estimated information matrix, which provides a simple connection between the score of and the variance of the MLE. Since $\boldsymbol{\theta}$ is unknown, the information matrix must be estimated. One possibility is simply to plug in the estimate $\hat{\boldsymbol{\theta}}_{ML}$ instead of $\boldsymbol{\theta}$ in the Fisher information

$\boldsymbol{I}(\boldsymbol{\theta})$. This estimator is referred to as the *expected information*. Another commonly used estimator is

$$\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}_{ML}) = -\boldsymbol{\partial}_{\boldsymbol{\theta}}\boldsymbol{S}(\hat{\boldsymbol{\theta}}_{ML}) \ ,$$

which is called the *observed information*. It has elements

$$\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}_{ML})_{(i,j)} = -\sum_k \frac{\partial^2 \log f(y_k|\boldsymbol{x}_k;\hat{\boldsymbol{\theta}}_{ML})}{\partial\theta_i\partial\theta_j} \ .$$

Ignoring the randomness of $\hat{\boldsymbol{\theta}}_{ML}$, the first estimator $\boldsymbol{I}(\hat{\boldsymbol{\theta}}_{ML})$ is the expectation of the observed information. In practice, the observed information is often preferred before the expected information [14].

Perspective and Sources

Much of the early contributions to the development of the theory of maximum likelihood estimation is due to R. A. Fisher. The main topics in maximum likelihood theory are covered by most standard textbooks in statistics, see for example Casella and Berger [9]. The asymptotic results presented in this section are quite general and holds for most standard distributions. Necessary conditions for these to hold essentially has to do with the support and differentiability of $f(y|\boldsymbol{x};\boldsymbol{\theta})$, see Casella and Berger [9] or Serfling [42] for more details on these technical conditions.

Illustrative Examples

Example 2.2.1 (The Likelihood Function) *Suppose that Y_k are independent and identically distributed with $Y_k \sim \mathcal{N}(\mu,\sigma)$, $k = 1,\ldots,N$, where σ is known. Given the observed data $\boldsymbol{y} = (y_1,\ldots,y_N)$ the likelihood is a function of μ:*

$$\mathcal{L}(\mu;\boldsymbol{y}) = \prod_k f(y_k;\mu) = \prod_k \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y_k-\mu)^2}{\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2}\frac{\sum_k (y_k-\mu)^2}{\sigma^2}} \ ,$$

which is illustrated in Figure 2.2. The maximum likelihood estimator $\hat{\mu}_{ML}$ of μ is chosen so that $\mathcal{L}(\mu;\boldsymbol{y})$ is maximized, i.e., $\hat{\mu}_{ML}$ is the point along the x-axis for which the maximum along the y-axis is reached.

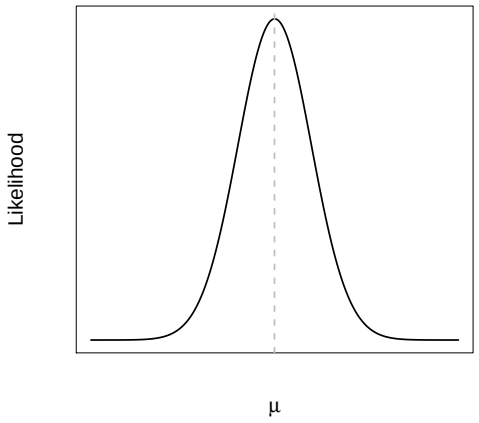

Figure 2.2: The likelihood as function of μ for a sample from a $\mathcal{N}(\mu, \sigma)$-distribution, where σ is known. The MLE is the point along the x-axis for which the maximum along the y-axis is reached, indicated by the grey line in the figure.

Example 2.2.2 (Estimating Parameters of the Normal Distribution) *Suppose that Y_k are independent and identically distributed with $Y_k \sim \mathcal{N}(\mu, \sigma)$, $k = 1, \ldots, N$, where both μ and σ are unknown. The maximum of $\mathcal{L}(\mu, \sigma; \boldsymbol{y})$ is found by maximizing the log-likelihood*

$$\ell(\mu, \sigma; \mathbf{y}) = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2} \frac{\sum_k (y_k - \mu)^2}{\sigma^2} \ .$$

The partial derivatives of the log-likelihood are

$$\frac{\partial \ell(\mu, \sigma; \mathbf{y})}{\partial \mu} = \sum_k \frac{y_k - \mu}{\sigma^2} \ ,$$

$$\frac{\partial \ell(\mu, \sigma; \mathbf{y})}{\partial \sigma} = -\frac{N}{\sigma} + \sum_k \frac{(y_k - \mu)^2}{\sigma^3} \ .$$

Solving $\boldsymbol{S}(\mu, \sigma) = \mathbf{0}$ gives the maximum likelihood estimators for μ and σ as

$$\hat{\mu}_{ML} = \frac{\sum_k y_k}{N} \ ,$$

$$\hat{\sigma}_{ML} = \sqrt{\frac{\sum_k (y_k - \hat{\mu}_{ML})^2}{N}} \ .$$

The second order partial derivatives of $\log f(Y; \mu, \sigma)$ are given by

$$\frac{\partial^2 \log f(Y; \mu, \sigma)}{\partial \mu^2} = -\frac{1}{\sigma^2} \ ,$$

$$\frac{\partial^2 \log f(Y;\mu,\sigma)}{\partial \sigma^2} = -3\frac{(Y-\mu)^2}{\sigma^4} + \frac{1}{\sigma^2} \ ,$$
$$\frac{\partial^2 \log f(Y;\mu,\sigma)}{\partial \mu \partial \sigma} = -2\frac{Y-\mu}{\sigma^3} \ .$$

According to (2.2.3), the Fisher information is thus

$$\boldsymbol{I}(\boldsymbol{\theta}) = NE_Y \begin{pmatrix} \frac{1}{\sigma^2} & 2\frac{Y-\mu}{\sigma^3} \\ 2\frac{Y-\mu}{\sigma^3} & 3\frac{(Y-\mu)^2}{\sigma^4} - \frac{1}{\sigma^2} \end{pmatrix} = \begin{pmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{2N}{\sigma^2} \end{pmatrix} \ .$$

The information matrix has inverse

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \begin{pmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{\sigma^2}{2N} \end{pmatrix} \ ,$$

which is the asymptotic or approximate variance-covariance matrix of $(\hat{\mu}_{ML},\hat{\sigma}_{ML})$*. Note that the asymptotic distribution of the sample mean is* $\hat{\mu}_{ML} \underset{a}{\sim} \mathcal{N}(\mu, \sigma^2/N)$*, which coincide with the sample distribution of* $\hat{\mu}_{ML}$ *for finite samples. Note also that* $(\hat{\mu}_{ML},\hat{\sigma}_{ML})$ *are asymptotically independent, which also holds for finite samples. Finally, the variance-covariance matrix of* $(\hat{\mu}_{ML},\hat{\sigma}_{ML})$ *can be estimated by*

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}} = \begin{pmatrix} \frac{\hat{\sigma}^2_{ML}}{N} & 0 \\ 0 & \frac{\hat{\sigma}^2_{ML}}{2N} \end{pmatrix} \ .$$

Example 2.2.3 (The Fisher Information) *A simple example illustrating the connection between the second order derivatives of the log-likelihood and the variance of an estimator is now given.*

Consider two simple random samples from a normal population with known variance, the first sample being of size 25 *and the second of size* 100*. The corresponding log-likelihoods are shown in Figure 2.3. The smaller sample has a blunt peak around the estimated value. There are thus many points that are almost equally likely given the observed data. If another sample is drawn, another value close to the current peak will probably be the most likely value. A blunt peak thus corresponds to a large variance. In terms of derivatives of the log-likelihood, this is the same as to have a small negative second derivative at* $\hat{\theta}_{ML}$*. The larger sample has peaked log-likelihood around the estimated value and a large second derivative of the log-likelihood at* $\hat{\theta}_{ML}$*, corresponding to a small set of estimates which are likely under the observed data and thus small variance of the estimator.*

The information the sample contains about μ *is summarized by the Fisher information number, which is*

$$I(\mu) = N\,\mathrm{E}_Y\left(\frac{\partial^2 \log f(Y;\mu)}{\partial \mu^2}\right) = \frac{N}{\sigma^2} \ .$$

The second sample has four times larger sample size and thus contain four times as much information about μ as the first sample, resulting in a variance reduction in $\hat{\mu}_{ML}$ by a factor 4. This example shows that increasing the sample size is one way to achieve larger information and smaller variance. It will later be shown how increased information and reduced variance can be achieved also by choice of design.

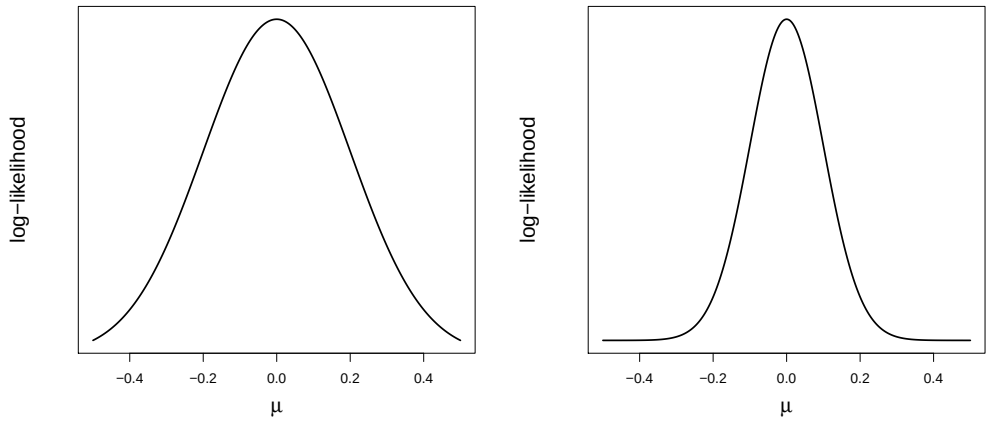

Figure 2.3: The log-likelihood as function of μ for a sample from a $\mathcal{N}(\mu, \sigma)$-distribution, where σ is known. To the left: $N = 25$. The log-likelihood has a blunt peak around the maximum, corresponding to low information and high variance. To the right: $N = 100$. The log-likelihood has a tight peak around the maximum, corresponding to high information and low variance.

2.2.2 Survey Sampling

Suppose now that S_1 is a fixed finite population of N elements. Associated with each element is a non-random but unknown quantity y_k. In this setting, interest could be in estimation of some characteristic of the finite population, such as the total or mean of y_k, or a ratio of two variables. By complete enumeration of all elements in S_1, the actual value of the population characteristic could be obtained. This is however often infeasible for practical and economical reasons, so a sample S_2 has to be selected from which the characteristic of interest can be estimated. Let us consider the total t of the variable y_k in S_1, given by

$$t = \sum_{k \in S_1} y_k \ . \tag{2.2.4}$$

In this section, various designs for sampling from a finite population will first be discussed and estimation of the total (2.2.4) will then be addressed. Even though other characteristics could be of interest, estimation of totals will be of particular interest in this thesis and other finite population characteristics will not be considered.

Sampling Designs

When drawing a sample from S_1, each element in the finite population can either be included in S_2 or not, and we introduce the indicator functions

$$I_k = \begin{cases} 1, & \text{if } k \in S_2 \\ 0, & \text{if } k \notin S_2 \end{cases}$$

for the random inclusion of an element in the sample S_2. Let $\pi_k = P(k \in S_2) = P(I_k = 1)$ be the probability that element k is included in S_2, and $\pi_{kl} = P(k,l \in S_2) = P(I_k = 1, I_l = 1)$ be the probability that element k and l are both included in S_2. π_k and π_{kl} are referred to as the *first order* and *second order inclusion probabilities*, respectively. The inclusion probabilities are typically determined using information about the elements in the population provided by auxiliary variables known for all elements in S_1.

Let $\boldsymbol{I} = (I_1, \ldots, I_N)$ be the random vector of sample inclusion indicator functions and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$ be the vector of inclusion probabilities corresponding to $\boldsymbol{I}$. Note that the indicator variables are $Bernoulli(\pi_k)$-distributed random variables, possibly dependent, with

$$\mathrm{E}(I_k) = \pi_k, \quad \mathrm{Var}(I_k) = \pi_k(1-\pi_k), \quad \mathrm{Cov}(I_k, I_l) = \pi_{kl} - \pi_k\pi_l \ .$$

The sample selection procedure is called *sampling design* or *sampling scheme*. Of particular importance are *probability sampling designs*. These are designs in which each element has a known and strictly positive probability of inclusion, i.e. $\pi_k > 0$ for all $k \in S_1$.

Many different probability sampling designs are available for sampling of elements from finite populations, of which only a few will be mentioned and considered in this thesis. Broadly speaking, sampling designs can be classified as sampling with replacement in contrast to sampling without replacement, as fixed size sampling in contrast to random size sampling, and as sampling with equal probabilities in contrast to sampling with unequal probabilities. Sampling without replacement is in general more efficient than sampling with replacement. Fixed size sampling designs are in general more efficient than sampling designs with random size. Sampling with unequal probabilities is in general more efficient that sampling with equal probabilities, if additional information is available for selection of inclusion probabilities.

Perhaps the most well known sampling design is *simple random sampling*, in which n elements are selected at random with equal probabilities. A closely related sampling procedure is *Bernoulli sampling*, in which all I_k are independent and identically distributed with $\pi_k = \pi$. In contrast to simple random sampling, the sample size under Bernoulli sampling is random and follows a $Binomial(N, \pi)$ distribution, and has expectation equal to $N\pi$. Independent inclusion of elements makes sampling from a Bernoulli design easy. It can be thought of as flipping of a biased coin N times, including element k or not in S_2 depending on the outcome of the k-th coin flip.

A generalization of Bernoulli sampling is *Poisson sampling*, in which I_k are independent but not necessarily identically distributed, so that $I_k \sim Bernoulli(\pi_k)$ with π_k possible unequal. In this case the sample size is also random with expectation

$$\mathrm{E}\left(\sum_k I_k\right) = \sum_k \pi_k \ .$$

The random sample size follow a Poisson-Binomial distribution, which for small π_k and large N can be approximated by a Poisson distribution, according to the Poisson limit theorem. Thinking of this in the coin flipping setting, each element has its own biased coin. Such a design is useful if one believe that some elements provide 'more information' about the characteristic of interest than others. Another sampling procedure that makes use of this fact is *stratified sampling*. With this procedure, elements are grouped into disjoint groups, called *strata*, according to a covariate that explains some of the variability in y. A simple random sample is then selected from each strata. Since the covariate explains some of the variability in the variable of interest, variation will be smaller within strata than in the entire population, so that the characteristic of interest can be estimated with high precision within strata. By pooling the estimates across strata, increased precision in estimation of t can be achieved. In particular, a large gain can be achieved by choosing sampling fractions within strata so that more elements are sampled from strata with high variability in y.

The Horvitz-Thompson Estimator

Let us now consider estimation of the total (2.2.4) from a probability sample S_2. A commonly used estimator of the population total (2.2.4) is the so called π-*expanded estimator*, or *Horvitz-Thompson estimator* [24], which is

$$\hat{t}_\pi = \sum_{k \in S_1} \frac{I_k}{\pi_k} y_k = \sum_{k \in S_2} \frac{y_k}{\pi_k} \ .$$

The distribution of $\hat{t}_\pi$ over iterated sampling from S_1, i.e under the distribution law of $\boldsymbol{I} = (I_1, \ldots, I_N)$, is called the *sampling distribution* of $\hat{t}_\pi$. Note that the expectation of $\hat{t}_\pi$ under the sampling distribution is

$$\mathrm{E}(\hat{t}_\pi) = \sum_k \frac{\mathrm{E}(I_k)}{\pi_k} y_k = \sum_k y_k = t \ ,$$

provided that $\pi_k > 0$ for all $k \in S_1$, and we say that $\hat{t}_\pi$ is *design unbiased* for t. The variance of the π-estimator is

$$\begin{aligned} \mathrm{Var}(\hat{t}_\pi) &= \sum_{k,l} \mathrm{Cov}\left(\frac{I_k}{\pi_k} y_k, \frac{I_l}{\pi_l} y_l\right) \\ &= \sum_k \frac{1-\pi_k}{\pi_k} y_k^2 + \sum_{k \neq l} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l} y_k y_l \ . \end{aligned} \tag{2.2.5}$$

In similarity with estimation of t, the variance of $\hat{t}_\pi$ can be estimated by π-expansion as

$$\widehat{\mathrm{Var}}(\hat{t}_\pi) = \sum_k \frac{I_k}{\pi_k}\frac{1-\pi_k}{\pi_k}y_k^2 + \sum_{k\neq l}\frac{I_k I_l}{\pi_k \pi_l}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l}y_k y_l \ .$$

The above variance estimator is design unbiased provided that $\pi_{kl} > 0$ for all k,$l \in S_1$.

The intuition behind π-expanded estimators is the following. Since fewer elements are included in S_2 than in S_1, expansion is needed in order to reach the total of y_k in S_1. As an easy example one can think of Bernoulli sampling with $\pi_k = 1/10$. Since approximately 10% of the population is sampled, the total in S_1 will be approximately ten times the total in the sample, and an expansion with a factor $1/\pi_k = 10$ is appropriate. In a general sampling scheme with unequal inclusion probabilities, the factor $1/\pi_k$ can be thought of as the number of elements in S_1 represented by element k. An element with a high inclusion probability thus represents a small number of elements, while an element with a small inclusion probability represents a large number of elements, and the contribution of each element to the estimated total is inflated accordingly.

The use of a probability sampling design is crucial for design unbiasedness and it is easy to come up with examples with π-estimators being biased when $\pi_k = 0$ for some k. For example, think of a situation where every element with y_k below the mean of y_k in S_1 is sampled with zero probability - this will always lead to overestimation the true total of t in S_1.

Note that inference about finite population characteristics is free of model assumptions on the study variables, and that the statistical properties of an estimator is completely determined by the design. Inference about finite population characteristics is consequently called *design based*, in contrast to the model based inference discussed in the previous section.

Perspective and Sources

Sample estimators for finite population characteristics rarely are unique, and optimal estimators in terms of efficiency does in general not exist [18]. It is often possible to apply more efficient estimators than the Horvitz-Thompson estimator, in particular when auxiliary information about the population is available. By incorporation of such information in estimation, substantial gain in precision can be achieved. See Särndal et al. [45] for a presentation of such methods, as well as for more details on the material presented in this section.

Even for inference about finite populations, the asymptotic properties of estimators could be of interest. Design based central limit theorems have been established, showing asymptotic normality and consistency of t_π and similar estimators. Important contributions to the study of asymptotic properties of design based estimators have been made by Hájek and Rosén, among others, and the main results are covered by Fuller [17] Chapter 1.3. Since the target population is finite, any statement about the limiting behavior of an estimator involves sequences of simultaneously increasing populations and samples, and the asymptotic properties depend on the construction of these sequences.

The requirements for convergence of sample estimators are quite technical, involving the existence of moments of the study variables and conditions on the limiting behavior of the inclusion probabilities.

Having introduced the survey sampling viewpoint on statistics, a word of clarification regarding the two-phase sampling procedure considered in this thesis might be in place. Two-phase sampling is most commonly encountered in the context of survey sampling, where the target population is a finite population. This is however quite different from the situations considered in this thesis, where the first sample is a random sample form an infinite population. The survey sampling viewpoint is to think of the study variables as fixed constants through both phases of sampling, while the viewpoint in this thesis is to think of the study variables as generated by some random process in the first phase and as constants in the second phase.

2.3 Maximum Pseudo-Likelihood

Let us now return to the two-phase sampling situation described Chapter 2.1, considering random sampling from some population model in the first phase followed by subsampling with unequal probabilities in the second phase. In contrast to the situation considered in Section 2.2.1, the conditional distribution of Y_k given $\boldsymbol{X}_k$ in S_2 might differ from the underlying population distribution, since S_2 is not necessarily a simple random sample. Classical maximum likelihood methods can thus not be applied. However, if the log-likelihood in S_1 were known, maximum likelihood could have been used to estimate $\boldsymbol{\theta}$. Now, thinking of the first phase sample S_1 as a finite population, the log-likelihood (2.2.1) can be thought of as a finite population characteristic. Inspired the methods presented in section 2.2.2, a two-step procedure for estimation of $\boldsymbol{\theta}$ can be proposed as follows. In the first step, the log-likelihood in S_1 is estimated from the observed data in S_2 using π-expansion. The second step uses classical maximum-likelihood methods to estimate $\boldsymbol{\theta}$ from the estimated log-likelihood, rather than from the log-likelihood as it appears in S_2. Doing so, the possible non-representativeness of S_2 as a sample from $\mathcal{P}$ is adjusted for in the estimation procedure. This is the idea behind maximum pseudo-likelihood estimation.

The Maximum Pseudo-Likelihood Estimator

Given the observed data $(y_k, \boldsymbol{x}_k)$, $k \in S_2$, obtained by any probability sampling design, we introduce the *π-expanded log-likelihood* or *pseudo log-likelihood* as

$$\ell_\pi(\boldsymbol{\theta}; \boldsymbol{y}, \tilde{\boldsymbol{X}}) := \sum_{k \in S_1} \frac{I_k}{\pi_k} \log f(\boldsymbol{\theta}; y_k, \boldsymbol{x}_k) = \sum_{k \in S_2} \frac{\log f(\boldsymbol{\theta}; y_k, \boldsymbol{x}_k)}{\pi_k} \ .$$

With maximum pseudo-likelihood estimation, the *maximum pseudo-likelihood estimator* (PLE) $\hat{\boldsymbol{\theta}}_\pi$ chosen to be the point satisfying

$$\hat{\boldsymbol{\theta}}_\pi := \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ \ell_\pi(\boldsymbol{\theta}; \boldsymbol{y}, \tilde{\boldsymbol{X}}) \ .$$

Denote by $\boldsymbol{S}_\pi(\boldsymbol{\theta})$ the π-expanded score, that is the gradient of the pseudo log-likelihood. The PLE can be found by solving the estimating equation

$$\boldsymbol{S}_\pi(\boldsymbol{\theta}) = \mathbf{0} \ .$$

As for the classical MLE, a more thorough investigation of the critical points of the log-likelihood and the boundary of the parameter space than indicated above might be needed, following standard procedures in multivariate calculus.

Asymptotic Properties of the PLE

If S_2 is a probability sample, we have that

$$\mathrm{E}_{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}_\pi|\boldsymbol{Y},\boldsymbol{X}) \underset{a}{=} \hat{\boldsymbol{\theta}}_{ML}(\boldsymbol{Y},\boldsymbol{X}) \ , \tag{2.3.1}$$

which means that the PLE is asymptotically design unbiased for the MLE conditional on the first phase sample. As the subscript indicates, expectation is taken with respect to the distribution law of $\boldsymbol{I}$. The limiting procedure is such that $|S_1| = N \to \infty$, $|S_2| = n \to \infty$ and $n/N \to h \in [0,1)$, implying also that $(N-n) \to \infty$. That is, the sample sizes in both phases tend to infinity, the S_1 grows faster than S_2 but the sampling fraction might be non-negligible.

The interpretation of (2.3.1) is as follows. $\hat{\boldsymbol{\theta}}_{ML} = \hat{\boldsymbol{\theta}}_{ML}(\boldsymbol{Y},\boldsymbol{X})$ is a random function of $(\boldsymbol{Y},\boldsymbol{X})$, but once the first phase sample is drawn it is determined by the realizations $(\boldsymbol{y},\tilde{\boldsymbol{X}})$ and is no longer random. One can thus think of $\hat{\boldsymbol{\theta}}_{ML}$ as a population parameter in S_1, which is unknown since $(\boldsymbol{y},\tilde{\boldsymbol{X}})$ are not fully observed. A subsample S_2 is drawn and $\hat{\boldsymbol{\theta}}_{ML}(\boldsymbol{y},\tilde{\boldsymbol{X}})$ is estimated by $\hat{\boldsymbol{\theta}}_\pi = \hat{\boldsymbol{\theta}}_\pi(\boldsymbol{y},\tilde{\boldsymbol{X}},\boldsymbol{I})$, which conditional on S_1 is a random function solely of $\boldsymbol{I}$. The asymptotic equality (2.3.1) states that the mean of the PLE under iterated subsampling is approximately equal to the MLE in S_1 for large samples. One can thus think of the PLE as an estimator of the MLE, which in turn is an estimator of the population parameter. As a consequence, we have that

$$\mathrm{E}(\hat{\boldsymbol{\theta}}_\pi) = \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{X}}\left[\mathrm{E}_{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}_\pi)|\boldsymbol{Y},\boldsymbol{X}\right] \underset{a}{=} \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{X}}(\hat{\boldsymbol{\theta}}_{ML}) \underset{a}{=} \boldsymbol{\theta} \ ,$$

using the law of iterated expectation. This is to say that the PLE is an asymptotically unbiased estimator of the parameter of interest. The expectation is taken with respect to the joint distribution of $(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{I})$.

Another way to present this result is through the expression

$$\hat{\boldsymbol{\theta}}_\pi - \boldsymbol{\theta} = (\hat{\boldsymbol{\theta}}_\pi - \hat{\boldsymbol{\theta}}_{ML}) + (\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}) \ , \tag{2.3.2}$$

where the expectation of both terms on the right hand side are null in the limit. This follows from asymptotic unbiasedness of the PLE as an estimator of the MLE conditional on S_1, and by asymptotic unbiasedness of the MLE for the population parameter. Under some technical conditions, the bias is relatively small compared to the standard error of the estimator, and can thus be neglected for large samples. It also holds that $\hat{\boldsymbol{\theta}}_\pi$ is a consistent estimator of $\boldsymbol{\theta}$ under the distribution of $\boldsymbol{I}$ and $\boldsymbol{Y}|\boldsymbol{X}$ jointly, and that the two terms in (2.3.2) are asymptotically independent [41].

Asymptotic Normality and Variance of the PLE

Under general assumptions, the asymptotic distribution of the PLE is multivariate normal with

$$\hat{\boldsymbol{\theta}}_\pi \underset{a}{\sim} \mathcal{N}\left(\boldsymbol{\theta}, \widetilde{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}}\right) .$$

The variance-covariance matrix $\widetilde{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}}$ can be found using the law of total variance:

$$\begin{aligned} \mathrm{Var}(\hat{\boldsymbol{\theta}}_\pi) &= \mathrm{Var}_{\boldsymbol{Y}|\boldsymbol{X}}\left[\mathrm{E}_{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}_\pi|\boldsymbol{Y},\boldsymbol{X})\right] + \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{X}}\left[\mathrm{Var}_{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}_\pi|\boldsymbol{Y},\boldsymbol{X})\right] \\ &\underset{a}{=} \mathrm{Var}_{\boldsymbol{Y}|\boldsymbol{X}}[\hat{\boldsymbol{\theta}}_{ML}(\boldsymbol{Y},\boldsymbol{X})] + \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{X}}\left[\mathrm{Var}_{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}_\pi|\boldsymbol{Y},\boldsymbol{X})\right] . \end{aligned} \tag{2.3.3}$$

Using the asymptotic independence of the two terms in (2.3.2), the same result can be found directly as

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}_\pi - \boldsymbol{\theta}) \underset{a}{=} \mathrm{Var}(\hat{\boldsymbol{\theta}}_\pi - \hat{\boldsymbol{\theta}}_{ML}) + \mathrm{Var}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}) . \tag{2.3.4}$$

Formulas (2.3.4) and (2.3.3) show that the variance of the PLE can be decomposed into two parts, which one can think of as the variance in the first phase plus the variance in the second phase. $\mathrm{Var}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}) = \mathrm{Var}_{\boldsymbol{Y}|\boldsymbol{X}}\left[\hat{\boldsymbol{\theta}}_{ML}(\boldsymbol{Y},\boldsymbol{X})\right] = \boldsymbol{I}(\boldsymbol{\theta})^{-1}$ is the variance of the MLE between first phase samples and $\mathrm{Var}(\hat{\boldsymbol{\theta}}_\pi - \hat{\boldsymbol{\theta}}_{ML}) = \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{X}}\left[\mathrm{Var}_{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}_\pi|\boldsymbol{Y},\boldsymbol{X})\right]$ is the expectation of the conditional variance of the PLE within first phase samples. These components will be referred to as the *first phase variance* and the *second phase variance*.

Conditional on S_1, the variance of the PLE around the MLE can be written as

$$\mathrm{Var}_{\boldsymbol{I}}[\hat{\boldsymbol{\theta}}_\pi|\boldsymbol{Y},\boldsymbol{X}] \underset{a}{=} \boldsymbol{I}(\boldsymbol{\theta})^{-1}\,\mathrm{Var}_{\boldsymbol{I}}\left(\boldsymbol{S}_\pi(\boldsymbol{\theta})\right)\boldsymbol{I}(\boldsymbol{\theta})^{-1} , \tag{2.3.5}$$

which is called the *conditional variance*. Is is obtained by a first order Taylor approximation of the score [7]. A derivation of the formula is given in Appendix A.1 and a simple illustration of the linearization technique is given in Example 2.3.2. The second phase variance is the expectation of the conditional variance. Putting the variance formulas together, the total variance of the PLE can be written as

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}_\pi) \underset{a}{=} \boldsymbol{I}(\boldsymbol{\theta})^{-1} + \boldsymbol{I}(\boldsymbol{\theta})^{-1}\,\mathrm{E}_{\boldsymbol{Y}|\boldsymbol{X}}\left(\mathrm{Var}_{\boldsymbol{I}}\left[\boldsymbol{S}_\pi(\boldsymbol{\theta})\right]\right)\boldsymbol{I}(\boldsymbol{\theta})^{-1} \tag{2.3.6}$$

The Design Influence on the Variance of the PLE

Let us now turn our attention to the middle term of the conditional variance (2.3.5), that is $\mathrm{Var}_{\boldsymbol{I}}\left(\boldsymbol{S}_\pi(\boldsymbol{\theta})\right)$. First, we introduce the notation

$$\begin{aligned} s_k^{(i)} &= \frac{\partial \log f(y_k|\boldsymbol{x}_k;\boldsymbol{\theta})}{\partial \theta_i} , \\ \boldsymbol{s}_k &= \nabla_{\boldsymbol{\theta}} \log f(y_k|\boldsymbol{x}_k;\boldsymbol{\theta}) = (s_k^{(1)},\ldots,s_k^{(p)}) . \end{aligned}$$

The π-expanded score can then be written as

$$\boldsymbol{S}_\pi(\boldsymbol{\theta}) = \sum_k \frac{I_k}{\pi_k} \boldsymbol{s}_k \ .$$

Similar to the variance of the Horvitz-Thompson estimator (2.2.5), the design variance of $\boldsymbol{S}_\pi(\boldsymbol{\theta})$ has elements

$$\begin{aligned} \mathrm{Var}_{\boldsymbol{I}} \left[\boldsymbol{S}_\pi(\boldsymbol{\theta}) \right]_{(i,j)} &= \sum_{k,l} \mathrm{Cov}_{I_k, I_l} \left(\frac{I_k}{\pi_k} s_k^{(i)}, \frac{I_l}{\pi_l} s_l^{(j)} \right) \\ &= \sum_k \frac{1-\pi_k}{\pi_k} s_k^{(i)} s_k^{(j)} + \sum_{k \neq l} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l} s_k^{(i)} s_l^{(j)} \ , \end{aligned}$$

in which the influence of the design on the variance of the PLE is made evident. The entire matrix can be written as

$$\mathrm{Var}_{\boldsymbol{I}} \left[\boldsymbol{S}_\pi(\boldsymbol{\theta}) \right] = \sum_k \frac{1-\pi_k}{\pi_k} \boldsymbol{s}_k \boldsymbol{s}_k^T + \sum_{k \neq l} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l} \boldsymbol{s}_k \boldsymbol{s}_l^T \ . \tag{2.3.7}$$

Variance Estimation

In order to estimate the variance of the PLE, both terms in (2.3.6) must be estimated. This can be done by π-expansion of their observed analogues evaluated at $\hat{\boldsymbol{\theta}}_\pi$, that is

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}) = \hat{\boldsymbol{I}}_\pi(\hat{\boldsymbol{\theta}}_\pi)^{-1} = -\boldsymbol{\partial} \boldsymbol{S}_\pi(\hat{\boldsymbol{\theta}}_\pi)^{-1} \ , \tag{2.3.8}$$

and

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_\pi - \hat{\boldsymbol{\theta}}_{ML}) = \hat{\boldsymbol{I}}_\pi(\hat{\boldsymbol{\theta}}_\pi)^{-1} \left(\sum_k \frac{I_k}{\pi_k} \frac{1-\pi_k}{\pi_k} \hat{\boldsymbol{s}}_k \hat{\boldsymbol{s}}_k^T + \sum_{k \neq l} \frac{I_k I_l}{\pi_k \pi_l} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l} \hat{\boldsymbol{s}}_k \hat{\boldsymbol{s}}_l^T \right) \hat{\boldsymbol{I}}_\pi(\hat{\boldsymbol{\theta}}_\pi)^{-1} \ , \tag{2.3.9}$$

where $\hat{\boldsymbol{s}}_k = \nabla_{\boldsymbol{\theta}} \log f(y_k | \boldsymbol{x}_k; \hat{\boldsymbol{\theta}}_\pi)$. The estimator (2.3.8) is a π-expanded estimator of the observed Fisher information, and has elements

$$\hat{\boldsymbol{I}}_\pi(\hat{\boldsymbol{\theta}}_\pi)_{(i,j)} = -\sum_k \frac{I_k}{\pi_k} \frac{\partial^2 \log f(y_k | \boldsymbol{x}_k; \hat{\boldsymbol{\theta}}_\pi)}{\partial \theta_i \partial \theta_j} \ .$$

The sum of the estimators (2.3.8) and (2.3.9) yield a consistent estimator of $\tilde{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}}$ provided that $\pi_{kl} > 0$ for all $k,l \in S_1$, under some additional technical conditions. It is also possible to estimate the variance of the PLE with resampling methods, such as the jackknife and bootstrap [17, 43].

Perspective and Sources

Among sampling statisticians, model based inference using pseudo-likelihood and similar methods is often referred to as *superpopulation modeling* [21], in contrast to design based finite population modeling. Dealing with finite populations, the superpopulation viewpoint is to think of the finite population as generated from a hypothetical infinite population through some random process, and the aim of analysis is to describe the underlying random process rather than the finite population itself. The modeling considered in this thesis is however not really the same as superpopulation modeling in its classical meaning, since S_1 truly is a random sample.

The notion of maximum pseudo-likelihood in the context of survey sampling was introduced by Skinner [44], but the method was already present in the literature by then. Conditions under which the PLE is is consistent and asymptotically normal has been established for regression models by Fuller [16], for generalized linear models by Binder [7], for logistic regression models and proportional hazards models by Chambless and Boyle [12], and for estimators defined as solutions to estimating functions by Godambe and Thompson [19]. The subject has been treated in some more generality by Rubin and Bleuer-Kratina [41], showing consistency and asymptotic normality of estimators defined as solutions to estimating equations, such as the PLE. The asymptotic results rely on convergence of the sample estimators under the design law and of the model statistics under the model law, both converging in distribution to a normally distributed random variable. In brief, these conditions concern the existence of moments of the study variables, the continuity and differentiability of the function defining the estimating equation and some conditions pertaining to the design. In particular, the design should be such that a design based central limit theorem holds, for which the use of probability sampling is a minimum requirement.

Procedures for pseudo-likelihood estimation are available in the R package 'survey' [2, 32] and in the SAS survey procedures [25]. These software provide procedures for parameter estimation and variance estimation for most standard parametric distributions and models, including generalized linear models. Other software often allow for specification of element weights as an additional argument to classical maximum likelihood estimation procedures, and the PLE can then be obtained by supplying the inverse sampling probabilities $1/\pi_k$ as element weights. Most such procedures are however not developed for the survey sampling or two-phase sampling setting, and uses variance formulas that do not take the design and sampling procedure into account. As a consequence, the variances tend to be underestimated [5].

Much of the material presented in this section is covered by Fuller [17], Chapters 1.3, 3.3, 4 and 6. An overview of the topic is also given by Chambers and Skinner [10].

Illustrative Examples

Example 2.3.1 (The Normal Distribution) *Consider a simple random sample S_1 of $Y_k \sim N(\mu, \sigma)$ with unobserved realizations y_k, $k = 1, \ldots, N$. A subsample S_2 for which y_k is observed is drawn using probability sampling. In analogy with Example 2.2.2, the*

π-expanded log-likelihood is

$$\ell_\pi(\mu, \sigma; \mathbf{y}) = \sum_k \frac{I_k}{\pi_k} \left(-\frac{\log(2\pi)}{2} - \frac{\log \sigma^2}{2} - \frac{1}{2} \frac{(y_k - \mu)^2}{\sigma^2} \right) ,$$

with partial derivatives

$$\frac{\partial \ell_\pi(\mu, \sigma; \mathbf{y})}{\partial \mu} = \sum_k \frac{I_k}{\pi_k} \frac{y_k - \mu}{\sigma^2} ,$$

$$\frac{\partial \ell_\pi(\mu, \sigma; \mathbf{y})}{\partial \sigma} = \sum_k \frac{I_k}{\pi_k} \left(\frac{(y_k - \mu)^2}{\sigma^3} - \frac{1}{\sigma} \right) .$$

The pseudo log-likelihood attains its maximum at

$$\hat{\mu}_\pi = \frac{\sum_k \frac{I_k}{\pi_k} y_k}{\sum_k \frac{I_k}{\pi_k}} ,$$

$$\hat{\sigma}_\pi = \sqrt{\frac{\sum_k \frac{I_k}{\pi_k} (y_k - \hat{\mu}_\pi)^2}{\sum_k \frac{I_k}{\pi_k}}} .$$

The asymptotic variance-covariance matrix of $\hat{\boldsymbol{\theta}}_\pi = (\hat{\mu}_\pi, \hat{\sigma}_\pi)$ is given by

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}_\pi) \underset{a}{=} \boldsymbol{I}(\boldsymbol{\theta})^{-1} + \boldsymbol{I}(\boldsymbol{\theta})^{-1} \, \mathrm{E}_{\boldsymbol{Y}} \left(\sum_k \frac{1 - \pi_k}{\pi_k} \boldsymbol{s}_k \boldsymbol{s}_k^T + \sum_{k \neq l} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l} \boldsymbol{s}_k \boldsymbol{s}_l^T \right) \boldsymbol{I}(\boldsymbol{\theta})^{-1} ,$$

where $\boldsymbol{I}(\boldsymbol{\theta})^{-1}$ is given in Example 2.2.2, and

$$\boldsymbol{s}_k = \left(\frac{Y_k - \mu}{\sigma^2}, \frac{(Y_k - \mu)^2}{\sigma^3} - \frac{1}{\sigma} \right) .$$

Estimation of $\mathrm{Var}(\hat{\boldsymbol{\theta}}_\pi)$ can be carried out as described by (2.3.8) and (2.3.9).

Note that the maximum pseudo-likelihood estimators in the example above are of the same form as the classical maximum likelihood estimators, but with π-expansions of the term in the numerator and with the 'pseudo-sample size' $\sum_k \frac{I_k}{\pi_k}$ in the denominator. Note in particular that $(\hat{\mu}_\pi, \hat{\sigma}_\pi) = (\hat{\mu}_{ML}, \hat{\sigma}_{ML})$ if S_2 is a Bernoulli sample.

Example 2.3.2 (Variance by Linearization) *A common method for deriving approximate variance formulas for non-linear functions of random variables is by linearization, sometimes called the δ-method. This is also the method used to derive the formula for the conditional variance (2.3.5). We illustrate this method by a simple example.*

Figure 2.4 illustrates the π-expanded score as a function of μ (left) and σ (right) for a sample from a normal distribution with one known parameter. Consider first the figure to the left where the score is a function of μ and σ is assumed to be known. The thick red line is the score as function of μ in S_1. If the study variable was observed for all

elements in S_1, the MLE of μ would have been the point along the x-axis where the line intersects the y-axis, in this case $\hat{\mu}_{ML} = 0$. The grey lines are the π-expanded scores for as function of μ for 50 random subsamples from S_1. For each subsample, the PLE is the point at which the corresponding grey line intersect the y-axis. Note that the PLE varies around the MLE. The variance of the PLE around the MLE is the variance of the intersections of the grey lines with the y-axis. This is the unknown variance that we want to approximate and estimate. Note that the score is a linear function of μ, so we can write

$$Score(\hat{\mu}_\pi) = Score(\hat{\mu}_{ML}) + b(\hat{\mu}_\pi - \hat{\mu}_{ML}) \Leftrightarrow Score(\hat{\mu}_{ML}) = -b(\hat{\mu}_\pi - \hat{\mu}_{ML}) \ ,$$

since the PLE is defined to satisfy $Score(\hat{\mu}_\pi) = 0$. Taking variances on both sides, we obtain

$$\text{Var}\left[Score(\hat{\mu}_{ML})\right] = b^2 \,\text{Var}(\hat{\mu}_\pi - \hat{\mu}_{ML}) \Leftrightarrow \text{Var}(\hat{\mu}_\pi - \hat{\mu}_{ML}) = \text{Var}[Score(\hat{\mu}_{ML})]/b^2 \ .$$

If we evaluate this formula at μ instead of $\hat{\mu}_{ML}$ and insert $b = I(\mu)$, we obtain the formula for the conditional variance given in (2.3.5).

A similar argument can also be used for the variance of $\hat{\sigma}_\pi$. However, the score is not a linear function of σ so the equalities must be replaced by approximations. For samples large enough it holds that $\hat{\sigma}_\pi$ will be close to σ, where a linear approximation is reasonable.

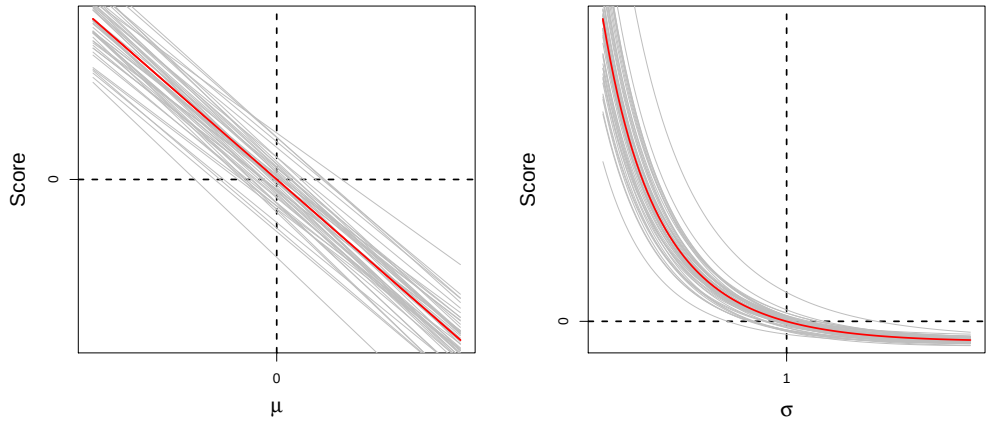

Figure 2.4: The score as function of μ (left) and σ (right) for a sample from a $\mathcal{N}(\mu, \sigma)$-distribution illustrated a thick red line. The grey lines represent the π-expanded scores in 50 random subsamples.

Example 2.3.3 (The Total Variance) *Consider $Y_k \sim (\mu, \sigma)$ where σ is known. The decomposition of* $\text{Var}(\hat{\mu}_\pi)$ *into variance between plus variance within first phase samples is illustrated in Figure 2.5. In this instance we have $N = 1000$ and n varying from 50*

to N using simple random sampling in the second phase. When a small fraction of S_1 is subsampled almost all variance is due to sampling of elements. This variance component decreases as the sampling fraction increases and vanishes as $n \to N$.

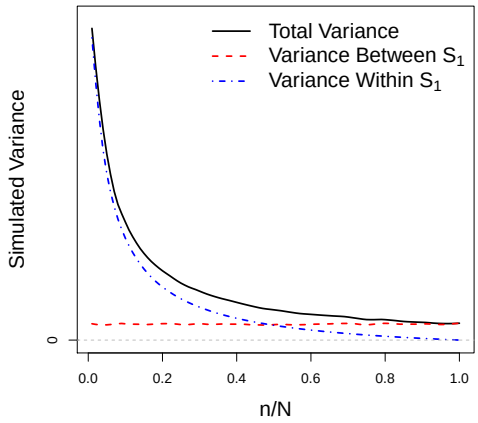

Figure 2.5: Simulated variance of $\hat{\mu}_\pi$ decomposed as variance within and variance between first phase samples based on 10^4 simulations. $N = 10^3$ observations were generated from a $\mathcal{N}(\mu,\sigma)$-distribution in the first phase, followed by subsampling of n elements in the second phase. The two variance components and the total variance are plotted against the ratio n/N.

2.3.1 Topics in Related Research

The crucial step in maximum pseudo-likelihood estimation is to obtain an unbiased estimator of the total of an estimating equation as it would have appeared in the first phase sample. This is obtained by π-expansion of the log-likelihood. This is however not the unique nor necessarily the optimal option for unbiased estimation of the log-likelihood, according to the discussion following Section 2.2.2. The PLE is in that sense not unique, and no best estimator of pseudo-likelihood type exists. Also, approximately unbiased and consistent estimators can be obtained by π-expansion of other estimating equations than the likelihood equations [44].

Criticism has been directed towards the method of maximum pseudo-likelihood for for two reasons. First, it discards the data observed in the first phase, which if incorporated in the inference procedure could lead to more efficient estimators. Second, inclusion of the sample weights $1/\pi_k$ in the estimating equation often lead to large variability of the estimator. Some alternative methods for estimation and methods for improvement of the PLE have been proposed, of which a few now will be discussed.

Even though being quite similar to the MLE, the PLE does not possess all the desirable properties of maximum likelihood estimators, such as efficiency. The PLE is in general less efficient than the MLE, when the latter is valid. It is thus important to

address the question of when and why classical maximum likelihood methods cannot be applied in two-phase sampling. The notion *informative* and *non-informative* sampling designs are central concepts when dealing with this question. A sampling design is said to be non-informative if the conditional distribution of the response variable given the explanatory variables does not depend on the sampling mechanism. That is, the distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$ should be the same in S_2 as in the underlying population. Non-informative sampling implies that the sampling design is ignorable conditional on $\boldsymbol{X}$, and maximum likelihood inference can be carried out based on $f(\boldsymbol{y}|\boldsymbol{x})$ using the data observed in the second phase. This is related to the missing data principles of Rubin [40] and Little and Rubin [31]. If the sampling mechanism is determined by some known auxiliary variables $\boldsymbol{Z}$, non-informative sampling can be achieved by including these as explanatory variables. However, the following three situations have been mentioned in the literature when this is not possible or suitable; when the variables used for selection of design are unavailable, when there are a large number of variables involved in the selection of design or when inclusion of these variables as explanatory variables lead to complex models, and in outcome-based designs. The latter two are relevant for the situations considered in this thesis. See Pfefferman [36,37] for a more thorough discussion about these topics.

Two main alternatives to the pseudo-likelihood under informative sampling have been proposed in the literature. Breckling et al. [8] propose a full maximum likelihood approach, formulating the estimation problem as a problem of estimation under incomplete data. It is closely related to the EM-algorithm, both being based on the same missing information principle of Orchard and Woodbury [35]. In two-phase sampling, complete data is available for all elements in S_2, while only the auxiliary variables are observed for the other elements in S_1. The full maximum likelihood approach makes more efficient use of data than the pseudo-likelihood by incorporation of the auxiliary variables in estimation. However, the full likelihood can be very complicated and might be sensitive to the model specification, involving the relations between outcome, explanatory variables and auxiliary variables.

The other proposed method is called sample likelihood and is due to Krieger and Pfeffermann et al. [30, 37]. They introduce the sample likelihood as the conditional distribution of the observed study variables given the auxiliary variables for all elements in S_1. A connection between the sample likelihood and the likelihood in S_1 is then found using Bayes rule. In similarity with the full maximum likelihood, it is a model based approach to analysis and uses the first phase data more efficient than the pseudo-likelihood. It can also be extended to a Bayesian setting with prior distributions on the parameters. An overview of the methods of full maximum likelihood and sample likelihood are given by Chambers and Skinner [10] and Chambers et al. [11]. Some other methods are also discussed by Pfefferman [36].

Some improvements of pseudo-likelihood have also been proposed, addressing the variability of the PLE by modification of the sampling weights in the estimating equation. The simplest modification is to replace the sampling weights $1/\pi_k$ in the estimating equation with another set of weights w_k, such that known totals in S_1 are estimated

correctly. This is called *calibration weighting*, and is well studied in the literature for sample estimators, see e.g. Särndal et al. [45]. Modifications of the sampling weights in the pseudo-likelihood by functions of the auxiliary variables has been proposed by Magee [33] and Kim and Skinner [28], among others.

Given the discussion above, argumentation for the use of PLE is in place. One of the major drawbacks of this method is the inefficient use of data, incorporating no information about about auxiliary variables in estimation. This is however also one of the major advantages of the PLE, since the validity of the inference procedure does not rely on certain assumptions made on the auxiliary variables. The inclusion of the sampling weights in the estimating equation thus simultaneously protects against informative sampling and against misspecification of the often rather complex models including $\boldsymbol{Z}$. In addition, it can quite easily be adopted to almost any kind of model and inference problem. The pseudo-likelihood might therefore be preferred due to its simplicity and general validity.

While most of the work reviewed above deals with improvement of estimators through the use of auxiliary variables in estimation, increased efficiency can also be achieved by incorporation of such information in the design. This has not been given much attention in this thesis so far, but is in fact a common feature in sampling statistics. Optimal subsampling designs in two-phase sampling has previously been addressed in the literature [15,17,27,38,39]. Much of the previous work in the area is however limited in the classes of estimators and models considered, and there is a need to address this issue in more generality. This issue will be the topic of Chapter 3.

2.4 Optimal Designs

The possibility to achieve increased precision in estimation by efficient subsampling of elements was mentioned in the previous section. In order to address this issue, a general framework for comparison of designs is needed. If estimation of a single parameter is of interest and a number of suitable designs are possible, the design giving the most precise estimate, i.e that yielding smallest variance, is often preferred. However, the notion of having 'smallest variance' does not really have a meaning for multidimensional parameters, since the variance is not a scalar but given by a variance-covariance matrix. Still, one would like to summarize the size of the variances and covariances of all or a subset of the parameters jointly in a single number. Trying to measure the size of the variance-covariance matrix, a number of *optimality criteria* have been proposed. A few of these will now be presented and motivated geometrically.

Confidence Regions

Consider an approximately unbiased and asymptotically normal estimator $\hat{\boldsymbol{\theta}}$ of the p-dimensional parameter $\boldsymbol{\theta}$, such as the MLE or PLE. The random variable W, defined by

$$W = (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \ ,$$

is approximately distributed according to a χ^2_p-distribution. Let $\chi^2_p(\alpha)$ denote the $1-\alpha$-percentile of the χ^2_p distribution. Consider the set of points $\boldsymbol{x}$ satisfying

$$(\boldsymbol{x}-\hat{\boldsymbol{\theta}})^T\boldsymbol{\Sigma}^{-1}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}-\hat{\boldsymbol{\theta}}) \leq \chi^2_p(1-\alpha)\ . \tag{2.4.1}$$

The set of points satisfying the above inequality defines an ellipsoid in p-dimensional space, which is called the $1-\alpha$ *confidence ellipsoid* or $1-\alpha$ *confidence region* for $\boldsymbol{\theta}$. It is a multivariate extension of the one-dimensional confidence interval. The *confidence level* or *coverage probability* $1-\alpha$ is the approximate probability that the confidence region cover the true parameter $\boldsymbol{\theta}$. For α sufficiently small, the confidence ellipsoid of $\hat{\boldsymbol{\theta}}$ will cover $\boldsymbol{\theta}$ with large probability. If, in addition, the size of the confidence ellipsoid is small, $\hat{\boldsymbol{\theta}}$ will be within a small neighborhood of $\boldsymbol{\theta}$ with high probability. The precision of $\hat{\boldsymbol{\theta}}$ is thus related to the size and shape of the confidence ellipsoid, which is what the different optimality criteria that soon will be described try to address.

To illustrate this, the $1-\alpha$ confidence regions for a two-dimensional parameter $\boldsymbol{\theta}=(\theta_1,\theta_2)$ in four different designs are shown in Figure 2.6. Ellipses with axes parallel to the coordinate axes correspond to designs in which $\hat{\theta}_1$ and $\hat{\theta}_2$ independent. A tilted ellipse correspond to a design in which the estimators are correlated. The projection of an ellipse on one the coordinate axis gives us the $1-\alpha$ confidence interval for the corresponding parameter. If θ_1 is to be estimated with high precision, design b) is optimal, while a) and c) are optimal with respect to the precision in θ_2. The confidence ellipsoid with minimal area is obtained with design d), but this design has the largest marginal variances along both axes. The average variance, which is proportional to the average length of the projections of the ellipses to the coordinate axes, is equal for designs a)-c) and twice as large with d). The design that minimizes the variance along the direction with maximal variance, seen as the maximal distance from the border to the center of an ellipse, is a), while the worst design in this aspect is d).

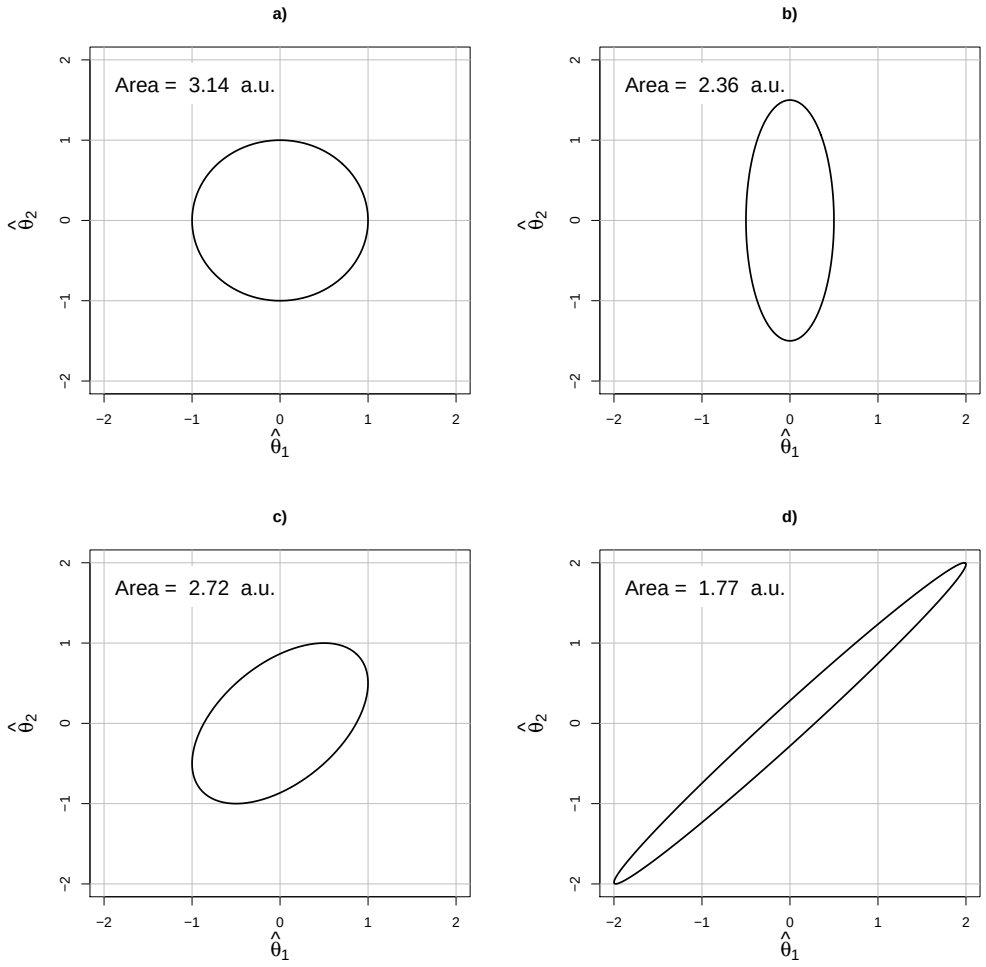

Figure 2.6: $1-\alpha$-confidence ellipses for a two-dimensional parameter $\boldsymbol{\theta} = (\theta_1, \theta_2)$. The ellipses have different shape, orientation and area.

Optimality Criteria

As the above discussion indicate, one design is seldom optimal in all aspects. The volume, shape and axis length of the confidence ellipsoid are all important features to consider. Formula (2.4.1) shows that these properties somehow depend on $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}$. In fact, the volume of the confidence ellipsoid is proportional to the square root of the determinant of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}$, denoted by $\det\left(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}\right)$. The design that minimizes the determinant of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}$, and so minimizes the volume of the confidence ellipsoid, is called *D-optimal*. The design that minimizes the average axis length, which is the same as minimizing the average variance or $\text{trace}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}})$, is called *A-optimal*. The design that minimizes the longest axis length, which is to say that minimizes the variance in the most extreme direction, is called *E-optimal*. This is the same as minimizing $\max_{|\boldsymbol{a}|=1} \text{Var}(\boldsymbol{a}^T\hat{\boldsymbol{\theta}}) = \max_{|\boldsymbol{a}|=1} \boldsymbol{a}^T\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}\boldsymbol{a}$. Comparing again the designs

in Figure 2.6, design d) is the best in terms of D-optimality while a)-c) are the best in terms of A-optimality and design a) is the best in terms of E-optimality.

These three optimality criteria can also be defined in terms of the eigenvalues of $\mathbf{\Sigma}_{\hat{\boldsymbol{\theta}}}$. A variance-covariance matrix is symmetric and has thus an eigendecomposition into real orthonormal eigenvectors. The eigenvectors are orthogonal and thus correspond to independent directions and the eigenvalues are the variances along these directions, since the eigenvectors are normalized. This means that the eigenvectors are parallel to the ellipsoid axes and the eigenvalues are proportional to the length of the axes. The volume of the ellipsoid is proportional to the product of the axes lengths, and thus proportional to the product of the eigenvalues. Minimizing the volume of the ellipsoid is thus equivalent to minimizing the product of the eigenvalues of the variance-covariance matrix. Minimizing the average variance is equivalent to minimizing the average of the eigenvalues, and minimizing the variance along the direction with largest variance is equivalent to minimizing the largest eigenvalue. A summary of A, D and E-optimality criteria formulated in terms of $\mathbf{\Sigma}_{\hat{\boldsymbol{\theta}}}$ and its eigenvalues λ_i is given below.

- A-optimality: $\min \text{trace}(\mathbf{\Sigma}_{\hat{\boldsymbol{\theta}}}) \Leftrightarrow \min \sum_{i=1}^{p} \lambda_i$
- D-optimality: $\min \det\left(\mathbf{\Sigma}_{\hat{\boldsymbol{\theta}}}\right) \Leftrightarrow \min \prod_{i=1}^{p} \lambda_i$
- E-optimality: $\min \max_{|\boldsymbol{a}|=1} \boldsymbol{a}^T \mathbf{\Sigma}_{\hat{\boldsymbol{\theta}}} \boldsymbol{a} \Leftrightarrow \min \max_{i=1,\ldots,p} \lambda_i$

It is also possible to consider a subset of the parameters rather than the entire parameter vector $\boldsymbol{\theta}$. This is useful when some components of $\boldsymbol{\theta}$ are nuisance parameters or if a specific subset of the parameters are of particular interest. A, D and E-optimality are then defined in a similar fashion on the specified subset.

Another class of optimality criteria that will be of interest in this work is linear optimality criteria. These address the average or sum of a linear combination of elements of the variance-covariance matrix. Examples of such linear combinations are linear combinations of variances and variances of linear combinations of parameters. The former include minimizing the sum of variances, which is the same as A-optimality. L-optimality is a generalization of A-optimality in that not only variances but also covariances between parameters are taken into account. A design that is optimal with respect to some linear optimality criteria is called *L-optimal*. Even though minimization of any linear combination of parameters of $\mathbf{\Sigma}_{\hat{\boldsymbol{\theta}}}$ is possible, it is often of interest to consider those linear combinations that arise from variances of linear combinations of parameters. These linear combinations are of the form $\text{Var}(\boldsymbol{a}^T\hat{\boldsymbol{\theta}}) = \boldsymbol{a}^T\mathbf{\Sigma}_{\hat{\boldsymbol{\theta}}}\boldsymbol{a}$. Furthermore, L-optimality allows not only for a single such linear combination to be minimized but for the average variance over a set of linear combinations of parameters. The objective function for minimization of the average variance over m different linear combinations $\boldsymbol{a}_i^T\hat{\boldsymbol{\theta}}$ of estimators can be written as

$$\min \sum_{i=1}^{m} \text{Var}(\boldsymbol{a}_i^T\hat{\boldsymbol{\theta}}) \Leftrightarrow \min \sum_{i=1}^{m} \boldsymbol{a}_i^T\mathbf{\Sigma}_{\hat{\boldsymbol{\theta}}}\boldsymbol{a}_i \ .$$

The objective function in A-optimality in recovered by having $\boldsymbol{a}_1 = (1,0,\ldots,0), \boldsymbol{a}_2 = (0,1,0,\ldots,0), \ldots, \boldsymbol{a}_p = (0,\ldots,0,1)$ in the above formulation.

One property that makes D-optimality preferable before A, E and L-optimality is scale invariance. Optimal designs under A, E and L-optimality are scale dependent, and hence depend on the unit of measurement, while D-optimality is scale invariant. For this reason, D-optimality is probably the most popular optimality criteria.

Finally, it shall also be mentioned that the optimal design does also depend on the true parameter, so that different designs are optimal for different values of the true parameter.

Perspective and Sources

The methods presented in this section have a natural place in controlled experiments where the experimental settings can be chosen by the experimenter, which gives an opportunity to determine the structure variance-covariance matrix of an estimator to a large extent. The theory of optimal designs is however of importance in many other areas of application, and gives the foundation for planning of studies and experiments involving multidimensional parameters. For references regarding the material presented in this section, see Atkinson and Donev [6].

3

Optimal Sampling Schemes under Poisson Sampling

The variance of the PLE under Poisson sampling is considered and the impact of the design on the variance is investigated. The anticipated variance is introduced and the role of auxiliary information in the planning of design is discussed. Optimal sampling schemes under L, D and E-optimality are derived based on the anticipated variance. Some post hoc modifications in design and estimation are presented.

The theoretical foundation needed for selection of optimal subsampling designs in two-phase sampling has been presented in the previous chapter. The impact of the design on the variance-covariance matrix of the PLE is given Formula (2.3.7) and various optimality criteria were presented in Chapter 2.4, stating the problem of finding an optimal design as a problem of minimizing some function of the variance-covariance matrix of an estimator. However, the variance of the PLE is in general a quite complicated function of the design, or more specifically of π_k and π_{kl}. Without restrictions on the designs considered it might not be possible to find an optimal design. This chapter is therefore restricted to Poisson sampling, which was introduced in section 2.2.2. Poisson sampling allows for quite general and flexible designs to be constructed, while achieving simplification in variance formulas by independence of sample inclusion indicator variables.

3.1 The Variance of the PLE under Poisson Sampling

Recall from section 2.3 that the variance of the PLE could be decomposed into

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\theta}}_\pi - \boldsymbol{\theta}) &\underset{a}{=} \mathrm{Var}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}) + \mathrm{Var}(\hat{\boldsymbol{\theta}}_\pi - \hat{\boldsymbol{\theta}}_{ML}) \\ &\underset{a}{=} \mathrm{Var}_{\boldsymbol{Y}|\boldsymbol{X}}[\hat{\boldsymbol{\theta}}_{ML}(\boldsymbol{Y},\boldsymbol{X})] + \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{X}}(\mathrm{Var}_{\boldsymbol{I}}[\hat{\boldsymbol{\theta}}_\pi|\boldsymbol{Y},\boldsymbol{X}]) \\ &\underset{a}{=} \boldsymbol{I}(\boldsymbol{\theta})^{-1} + \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{X}}\left(\boldsymbol{I}(\boldsymbol{\theta})^{-1}\,\mathrm{Var}_{\boldsymbol{I}}\left[\boldsymbol{S}_\pi(\boldsymbol{\theta})\right]\boldsymbol{I}(\boldsymbol{\theta})^{-1}\right)\ ,\end{aligned} \tag{3.1.1}$$

which is the variance between plus the variance within first phase samples. It was also shown that

$$\mathrm{Var}_{\boldsymbol{I}}\left[\boldsymbol{S}_\pi(\boldsymbol{\theta})\right] = \sum_k \frac{1-\pi_k}{\pi_k}\boldsymbol{s}_k\boldsymbol{s}_k^T + \sum_{k\neq l}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l}\boldsymbol{s}_k\boldsymbol{s}_l^T\ , \tag{3.1.2}$$

where

$$\boldsymbol{s}_k = \nabla_{\boldsymbol{\theta}}\log f(y_k|\boldsymbol{x}_k;\boldsymbol{\theta})\ .$$

These formulas will now be studied in some more detail. Some simplifications of the total variance formula (3.1.1) will first be made. The use of auxiliary information for guessing the total variance will then be discussed, and the anticipated variance introduced.

3.1.1 The Total Variance

Under Poisson sampling, Formula (3.1.2) simplifies to

$$\mathrm{Var}_{\boldsymbol{I}}\left[\boldsymbol{S}_\pi(\boldsymbol{\theta})\right] = \sum_k \frac{1-\pi_k}{\pi_k}\boldsymbol{s}_k\boldsymbol{s}_k^T\ ,$$

following from the fact that $\mathrm{Cov}(I_k, I_l) = 0$ and $\pi_{kl} = \pi_k\pi_l$ for $k \neq l$. Letting

$$\boldsymbol{W}_k = \boldsymbol{I}(\boldsymbol{\theta})^{-1}\boldsymbol{s}_k\boldsymbol{s}_k^T\boldsymbol{I}(\boldsymbol{\theta})^{-1}\ , \tag{3.1.3}$$

we can write

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{I}}[\hat{\boldsymbol{\theta}}_\pi|\boldsymbol{Y},\boldsymbol{X}] &\underset{a}{=} \boldsymbol{I}(\boldsymbol{\theta})^{-1}\left(\sum_k \frac{1-\pi_k}{\pi_k}\boldsymbol{s}_k\boldsymbol{s}_k^T\right)\boldsymbol{I}(\boldsymbol{\theta})^{-1} \\ &= \sum_k \frac{1-\pi_k}{\pi_k}\boldsymbol{W}_k\ .\end{aligned}$$

From this formula we see that an element which is certainly included in S_2, i.e. having $\pi_k = 1$, has no contribution to the second phase variance. Note now that

$$\begin{aligned}\mathrm{E}_{\boldsymbol{Y}|\boldsymbol{X}}\left(\frac{1-\pi_k}{\pi_k}\boldsymbol{W}_k\right) &= \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{X}}\left(\sum_k \frac{\boldsymbol{W}_k}{\pi_k}\right) - \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{X}}\left(\sum_k \boldsymbol{W}_k\right) \\ &= \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{X}}\left(\sum_k \frac{\boldsymbol{W}_k}{\pi_k}\right) - \boldsymbol{I}(\boldsymbol{\theta})^{-1}\sum_k \mathrm{E}_{Y_k|\boldsymbol{X}_k}\left(\boldsymbol{s}_k\boldsymbol{s}_k^T\right)\boldsymbol{I}(\boldsymbol{\theta})^{-1} \\ &= \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{X}}\left(\sum_k \frac{\boldsymbol{W}_k}{\pi_k}\right) - \boldsymbol{I}(\boldsymbol{\theta})^{-1}\ .\end{aligned}$$

From this we see that total variance matrix (3.1.1) simplifies to

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\theta}}_\pi - \boldsymbol{\theta}) &\underset{a}{=} \boldsymbol{I}(\boldsymbol{\theta})^{-1} + \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{X}}\left(\mathrm{Var}_{\boldsymbol{I}}[\hat{\boldsymbol{\theta}}_\pi|\boldsymbol{Y},\boldsymbol{X}]\right) \\ &= \boldsymbol{I}(\boldsymbol{\theta})^{-1} + \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{X}}\left(\sum_k \frac{\boldsymbol{W}_k}{\pi_k}\right) - \boldsymbol{I}(\boldsymbol{\theta})^{-1} \\ &= \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{X}}\left(\sum_k \frac{\boldsymbol{W}_k}{\pi_k}\right) . \end{aligned} \tag{3.1.4}$$

The total variance is thus the expectation of $\sum_k \frac{\boldsymbol{W}_k}{\pi_k}$ with respect to the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$, where $\boldsymbol{W}_k$ is given by (3.1.3).

Bernoulli Sampling

An interesting result is found when S_2 is a Bernoulli sample, i.e. $\pi_k = \pi$. Let $\pi = n/N$, so that the expected sample size in S_2 is n. The total variance becomes

$$\mathrm{E}_{\boldsymbol{Y}|\boldsymbol{X}}\left(\sum_k \frac{1}{n/N}\boldsymbol{W}_k\right) = \frac{N}{n}\mathrm{E}_{\boldsymbol{Y}|\boldsymbol{X}}\left(\sum_k \boldsymbol{W}_k\right) = \frac{N}{n}\boldsymbol{I}(\boldsymbol{\theta})^{-1} .$$

Note that this is the same as the approximate variance of the MLE in a simple random sample of size n. Also, the MLE and PLE coincide under Bernoulli sampling, since the weights in the pseudo log-likelihood are the same for all elements. The PLE under Bernoulli sampling is therefore essentially equivalent to the MLE under simple random sampling. There is however a difference between the two designs, namely that the sample size is random with Bernoulli sampling and fixed with simple random sampling. An additional source of randomness is introduced by the random sample size, which increase the variance under Bernoulli sampling compared to simple random sampling. This additional variance vanishes as n increase, so the the variance of the MLE under simple random sampling is essentially the same as the variance of the PLE under Bernoulli sampling for large samples.

The Realized Variance

An additional simplification of the variance of the PLE can be made, writing (3.1.4) as

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\theta}}_\pi - \boldsymbol{\theta}) &\underset{a}{=} \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{X}}\left(\sum_k \frac{\boldsymbol{W}_k}{\pi_k}\right) = \sum_k \frac{\mathrm{E}_{Y_k|\boldsymbol{X}_k}(\boldsymbol{W}_k)}{\pi_k} \\ &\underset{a}{=} \sum_k \frac{\mathrm{E}_{(Y_k,\boldsymbol{X}_k)}(\boldsymbol{W}_k)}{\pi_k} \underset{a}{=} \sum_k \frac{\boldsymbol{W}_k}{\pi_k} ,\end{aligned}$$

using the law of large numbers twice. The formula

$$\sum_k \frac{\boldsymbol{W}_k}{\pi_k} = \sum_k \frac{\boldsymbol{I}(\boldsymbol{\theta})^{-1}\boldsymbol{s}_k\boldsymbol{s}_k^T\boldsymbol{I}(\boldsymbol{\theta})^{-1}}{\pi_k} \tag{3.1.5}$$

will be referred to as the asymptotic *realized variance* of the PLE. This is similar to the observed Fisher information, which is in some sense is the inverse of the realized variance of the MLE.

Formula (3.1.5) shows that each element contributes a term $\boldsymbol{W}_k/\pi_k$ to the realized variance of the PLE. The variance contribution contribution thus depends both on the inclusion probability and on the matrix $\boldsymbol{I}(\boldsymbol{\theta})^{-1}\boldsymbol{s}_k\boldsymbol{s}_k^T\boldsymbol{I}(\boldsymbol{\theta})^{-1}$. This matrix in turn combines the precision in the first phase, as described by $\boldsymbol{I}(\boldsymbol{\theta})^{-1}$, with the influence of element k on the estimates in the second phase, as described by $\boldsymbol{s}_k\boldsymbol{s}_k^T$. We will have a closer look at this fact in Appendix A.2.

3.1.2 The Anticipated Variance

Given the matrices $\boldsymbol{W}_k$, the realized variance (3.1.5) is a function of the inclusion probabilities π_k, which can be optimized with respect to some optimality criteria. However, the matrices $\boldsymbol{W}_k$ are random and therefore unknown in the design stage, making it impossible to use the realized variance for construction of efficient subsampling designs. One might instead consider the expected variance (3.1.4), but even this might be unknown, unless all of $\tilde{\boldsymbol{X}}$ is observed in S_1. We thus need a formula for the variance of the PLE that does not require full knowledge about $\tilde{\boldsymbol{X}}$, but still can work as substitute of the true variance and ideally agree with (3.1.4) and (3.1.5) when these are evaluable. Due to the lack of information available, it is evident that any such variance formula must be guessed and to some extent is subject to personal belief.

Derivation

As described in Section 2.1, we consider a situation where some auxiliary variables $\boldsymbol{Z}$ are observed for all elements in the first phase. Conditional on $\boldsymbol{Z}$, $(\boldsymbol{Y},\boldsymbol{X})$ follow some distribution law with density function $f(y_k,\boldsymbol{x}_k|\boldsymbol{z}_k;\boldsymbol{\phi})$, called the *design model*. We want to utilize the information about $(\boldsymbol{Y},\boldsymbol{X})$ provided by $\boldsymbol{Z}$ to guess the variance of the PLE. In place of the unknown realized variance, we introduce the *anticipated variance*

$$\mathrm{Var}_a(\hat{\boldsymbol{\theta}}_\pi) := \mathrm{E}_{(\boldsymbol{Y},\boldsymbol{X})|\boldsymbol{Z}}\left(\sum_k \frac{\boldsymbol{W}_k}{\pi_k}\right) = \mathrm{E}_{(\boldsymbol{Y},\boldsymbol{X})|\boldsymbol{Z}}\left(\sum_k \frac{\boldsymbol{I}(\boldsymbol{\theta})^{-1}\boldsymbol{s}_k\boldsymbol{s}_k^T\boldsymbol{I}(\boldsymbol{\theta})^{-1}}{\pi_k}\right) , \qquad (3.1.6)$$

which is the expectation or prediction of the realized variance under the design model. Note that the anticipated variance depend on the design model for $(\boldsymbol{Y},\boldsymbol{X})$ given $\boldsymbol{Z}$ and its parameter $\boldsymbol{\phi}$ as well as on $\boldsymbol{\theta}$, which all must be guessed or known for evaluation of the anticipated variance.

Note that two different models are involved in (3.1.6), the terms $\boldsymbol{W}_k/\pi_k$ involving the model of interest, and the expectation being taken with respect to the design model. Ideally, these two models should be consistent, so that both could simultaneously be true. This is easily achieved when $\boldsymbol{Y}$ given $\boldsymbol{X}$ or $\boldsymbol{X}$ given $\boldsymbol{Y}$ is conditionally independent of $\boldsymbol{Z}$. In more complicated situations, specification of the design model so that consistency is obtained might be a difficult task. However, even though consistency between models

and a correctly specified design model is desirable, it is not needed for the validity of the inference procedure using the pseudo-likelihood approach. The design model is used only in the planning of design, as indicated by the name, and the inference procedure with pseudo-likelihood estimation is free of assumptions made about the study variables in the design stage.

Simplification

In the simple case with no explanatory variables and Y_k assumed to be independent and identically distributed, or if $\boldsymbol{Z} = \boldsymbol{X}$, we have that $\boldsymbol{I}(\boldsymbol{\theta})$ is a constant matrix conditionally on $\boldsymbol{Z}$. In the more complicated situation with explanatory variables that are not all observed in S_1, let $\boldsymbol{I}_a(\boldsymbol{\theta})$ be the *anticipated information* under the design model, given by

$$\boldsymbol{I}_a(\boldsymbol{\theta}) = \mathrm{E}_{(\boldsymbol{Y},\boldsymbol{X})|\boldsymbol{Z}} \left(-\boldsymbol{\partial} \boldsymbol{S}(\boldsymbol{\theta})\right) \ , \tag{3.1.7}$$

with elements

$$\boldsymbol{I}_a(\boldsymbol{\theta})_{(i,j)} = -\sum_k \mathrm{E}_{(Y_k,\boldsymbol{X}_k)|\boldsymbol{Z}_k} \left(\frac{\partial^2 \log f(Y_k, \boldsymbol{X}_k; \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right) \ .$$

The anticipated information is a sum of N random variables under the model for $(\boldsymbol{Y}, \boldsymbol{X})$ given $\boldsymbol{Z}$, and so is approximately constant for large N, due to the law of large numbers. We can thus move the information matrix outside the expectation in (3.1.6), and that anticipated variance simplifies to

$$\mathrm{Var}_a(\hat{\boldsymbol{\theta}}_\pi) = \boldsymbol{I}_a(\boldsymbol{\theta})^{-1} \left(\sum_k \frac{\mathrm{E}_{(Y_k,\boldsymbol{X}_k)|\boldsymbol{Z}_k}(\boldsymbol{s}_k \boldsymbol{s}_k^T)}{\pi_k} \right) \boldsymbol{I}_a(\boldsymbol{\theta})^{-1} \ . \tag{3.1.8}$$

To simplify notation further, let

$$\widetilde{\boldsymbol{W}}_k = \boldsymbol{I}_a(\boldsymbol{\theta})^{-1} \, \mathrm{E}_{(Y_k,\boldsymbol{X}_k)|\boldsymbol{Z}_k} \left(\boldsymbol{s}_k \boldsymbol{s}_k^T \right) \boldsymbol{I}_a(\boldsymbol{\theta})^{-1} \ . \tag{3.1.9}$$

The anticipated variance can then be written as

$$\mathrm{Var}_a(\hat{\boldsymbol{\theta}}_\pi) = \sum_k \frac{\widetilde{\boldsymbol{W}}_k}{\pi_k} \ ,$$

showing that each element contributes a term $\widetilde{\boldsymbol{W}}_k/\pi_k$ to the anticipated variance of the PLE. This is similar to the realized variance (3.1.5), replacing the matrices $\boldsymbol{W}_k$ by their anticipated counterparts $\widetilde{\boldsymbol{W}}_k$. These matrices can be interpreted accordingly as a combination of the anticipated precision in the first phase and the anticiapted influence in estimation, see also Appendix A.2.

Note that the anticipated information (3.1.7) coincide with the usual Fisher information in case $\boldsymbol{Z} = \boldsymbol{X}$ or in case there are no explanatory variables in the model of interest. Also, the anticipated variance (3.1.6) coincide with the expected variance (3.1.4) in case $\boldsymbol{Z} = \boldsymbol{X}$, and with the realized variance (3.1.5) in case $\boldsymbol{Z} = (\boldsymbol{Y}, \boldsymbol{X})$. The latter is of course a trivial case, where two-phase sampling is not really needed. In general, anticipated variance will differ from the variance actually realized.

Computation

According to (3.1.8), evaluation of the anticipated variance amounts to computing the expectation

$$\mathrm{E}_{(Y_k,\boldsymbol{X}_k)|\boldsymbol{Z}_k}(\boldsymbol{s}_k\boldsymbol{s}_k^T) = \mathrm{E}_{(Y_k,\boldsymbol{X}_k)|\boldsymbol{Z}_k}\left[\nabla_{\boldsymbol{\theta}} \log f(Y_k|\boldsymbol{X}_k;\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}^T \log f(Y_k|\boldsymbol{X}_k;\boldsymbol{\theta})\right] \ ,$$

for all elements in S_1, and the anticipated information

$$\boldsymbol{I}_a(\boldsymbol{\theta}) = \mathrm{E}_{(\boldsymbol{Y},\boldsymbol{X})|\boldsymbol{Z}}\left(-\boldsymbol{\partial}\boldsymbol{S}(\boldsymbol{\theta})\right) \ ,$$

which has elements

$$\boldsymbol{I}_a(\boldsymbol{\theta})_{(i,j)} = -\sum_k \mathrm{E}_{(Y_k,\boldsymbol{X}_k)|\boldsymbol{Z}_k}\left(\frac{\partial^2 \log f(Y_k,\boldsymbol{X}_k;\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right) \ .$$

Explicit formulas for the anticipated information matrix $\boldsymbol{I}_a(\boldsymbol{\theta})$ and for $\mathrm{E}_{(Y_k,\boldsymbol{X}_k)|\boldsymbol{Z}_k}(\boldsymbol{s}_k\boldsymbol{s}_k^T)$ might not exist. Numerical integration, e.g. Monte Carlo simulation, can then be used to approximate these quantities with a desired degree of accuracy.

The Design Model

As indicated in the derivation of the anticipated variance, the specification of the design model is an important planning task. If the design model describes the unknown study variables in a precise way, the anticipated variance will agree well with the true variance of the PLE and an optimal sampling scheme derived based on the anticipated variance will be close to the true optimal sampling scheme. Even though not crucial for the validity of the inference procedure, a proper specification of the design model is thus important in order to achieve as high precision in estimation as possible.

In a situation where a single study variable is unobserved in the first phase, it is often reasonable to assume some kind of regression relationship between the unobserved study variable and the observed auxiliary variables. However, if more than one variable is unobserved in the first phase, a multivariate design model must be specified. A proper specification of a multivariate model is often difficult, especially if $(\boldsymbol{Y},\boldsymbol{X})$ contains both continuous and discrete random variables. Simplification of the design model is often needed in order for the model construction to become a feasible task. Two possible simplifications are proposed.

One option is to assume a simple multivariate model for $(\boldsymbol{Y},\boldsymbol{X})$ given $\boldsymbol{Z}$, e.g. multivariate normal (MVN). Even though a MVN model in a strict sense is meaningful only for continuous variables, it can be useful even for discrete variables. It makes modeling of dependent variables easy, so that correlations between variables can be taken into account in a proper way. One could also discretize continuous variables generated from a MVN distribution in order to obtain discrete variables. This will however change the dependence structure between variables, leading to a covariance structure different from the one originally used in the MVN model.

Another alternative is to use Bayes law rewrite $f(y_k, \boldsymbol{x}_k|\boldsymbol{z}_k;\boldsymbol{\phi})$ by a sequence of conditional distributions. It is then possible to specify each conditional distribution according to any suitable model, allowing for discrete variables to be modeled correctly. A simple special case of this approach is to assume conditional independence between all unobserved study variables given the auxiliary variables, so that each unobserved variable can be modeled separately.

Perspective and Sources

The material presented in this section has not been encountered in the literature during the writing of this thesis. Similar work has however been done in the context of survey sampling, considering sample estimators of finite population characteristics, but not in the generality considered in this thesis. The nomenclature has been adopted to that of Isaki and Fuller [26]. They introduced the notion of anticipated variance as the variance of a sample estimator under an assisting parametric model, which they called the design model. With this notion it is emphasized that the anticipated variance is a prediction or expectation of the true unknown variance. Isaki and Fuller also pointed out that the anticipated variance in general is different from the realized sampling variance. Even for design based inference, the use of the model based anticipated variance in the planning of design has been proved fruitful, see for example Godfrey et al. [20] and Kott [29]. See also Fuller [17] Chapter 3 for a presentation of the use of the anticipated variance in selection of design for sampling from finite populations.

Some complications were encountered in the derivation of the anticipated variance. Essentially, this has to do with the joint consideration of two different models, the model of interest and the design model, which might not simultaneously be true. This problem is not present when considering simple finite population characteristics, since the estimator of interest is free of model assumptions.

3.2 Optimal Two-Phase Sampling Designs

Optimal subsampling designs based on the anticipated variance will now be presented. In the following sections, optimal sampling schemes under Poisson sampling are derived analytically for L-optimality, and methods for finding D and E-optimal sampling schemes numerically are discussed.

3.2.1 L-Optimal Sampling Schemes

Let us first consider L-optimality with respect to linear combinations of elements of the variance-covariance matrix of $\hat{\boldsymbol{\theta}}_\pi$ that appear as variances of linear combinations of parameters. The results will then be extended to sets of such linear combinations and to any linear combination of elements of the variance-covariance matrix.

A Single Linear Combination

Consider a linear combination $\boldsymbol{a}^T\boldsymbol{\theta} = a_1\theta_1 + \ldots + a_p\theta_p$ of parameters. According to the discussion in the previous section, we consider the anticipated variance of $\boldsymbol{a}^T\hat{\boldsymbol{\theta}}_\pi$ in place of the true unknown variance, i.e.

$$\mathrm{Var}_a(\boldsymbol{a}^T\hat{\boldsymbol{\theta}}_\pi) = \boldsymbol{a}^T\,\mathrm{Var}_a(\hat{\boldsymbol{\theta}}_\pi)\boldsymbol{a} = \sum_k \frac{\boldsymbol{a}^T\widetilde{\boldsymbol{W}}_k\boldsymbol{a}}{\pi_k}\ , \tag{3.2.1}$$

where $\widetilde{\boldsymbol{W}}_k$ is given by (3.1.9). Given $\boldsymbol{Z} = \tilde{\boldsymbol{Z}}$, the matrices $\widetilde{\boldsymbol{W}}_k$ are constant and (3.2.1) a known function of $\boldsymbol{\pi}$. We want to find sampling scheme, i.e. a set of inclusion probabilities $\boldsymbol{\pi}$, so that (3.2.1) is minimized. In order to control the sample size, we add the constraint that the expected sample size should be equal to $n \leq N$. We also require that $\pi_k \leq 1$ so that valid probabilities are obtained and $\pi_k > 0$ ensuring consistency of $\hat{\boldsymbol{\theta}}_\pi$. Let $c_k = \boldsymbol{a}^T\widetilde{\boldsymbol{W}}_k\boldsymbol{a}$. The objective function can then be written as

$$\min_{\boldsymbol{\pi}} g(\boldsymbol{\pi}) = \sum_k \frac{c_k}{\pi_k} \tag{3.2.2}$$

$$\text{subject to} \quad \begin{matrix} \pi_k \in (0,1] \\ \sum_k \pi_k = n \end{matrix}\ . \tag{3.2.3}$$

Relaxing the constraint $\pi_k \leq 1$, the optimal solution to this problem is given by

$$\pi_k^* = \frac{n}{\sum_i \sqrt{c_i}}\sqrt{c_k}\ . \tag{3.2.4}$$

For simplicity we also write

$$\pi_k^* = \kappa\sqrt{c_k}\ , \tag{3.2.5}$$

where it is understood that $\kappa = \frac{n}{\sum_i \sqrt{c_i}}$ is a constant chosen so that the expected sample size equals to n.

Choosing π_k^* according to (3.2.4) one might end up with $\pi_k > 1$ for some k, which is an infeasible solution. This is solved by letting $\pi_k = 1$ for all such elements and solve (3.2.2) subject to (3.2.3) with all elements having $\pi_k > 1$ removed in the previous step and the expected sample size reduced accordingly. Again, this might result in $\pi_k > 1$ for some k, and the same procedure must be repeated until a feasible solution is found. The procedure will stop in a finite number of iterations since $n \leq N$ and S_1 contains a finite number of elements. The solution found by this procedure is L-optimal for the linear combination of interest. A proof of the optimality of (3.2.4) and the iterative procedure described above is given in Appendix B. The resulting design is a probability sampling design provided that $c_k > 0$ for all k. In practice, having $c_k = 0$ is to say that element k is believed not to influence any of the parameters that are considered in the objective function. Dealing with continuous data and/or continuous parameters and having any uncertainty about the study variables and the true parameter in the design model does in general ensure that $c_k > 0$. In case elements with $c_k = 0$ are encountered in practice, one can set $c_k = \epsilon$ for some small positive ϵ.

General L-Optimal Sampling Schemes

As mentioned in section 2.4, more than a single linear combination of parameters can be considered with L-optimality. If the aim is to minimize the average variance over a set of m estimated linear combinations of parameters $\boldsymbol{a}_i^T\hat{\boldsymbol{\theta}}_\pi$, we replace c_k with

$$c_k' = \sum_{i=1}^{m} \boldsymbol{a}_i^T \widetilde{\boldsymbol{W}}_k \boldsymbol{a}_i$$

in the objective function (3.2.2) and consequently arrive at the same optimal solution as before, but with c_k replaced by c_k'. Even for arbitrary linear combinations of elements in the anticipated variance-covariance matrix, the objective function can be written in the form (3.2.2).

3.2.2 D and E-optimal Sampling Schemes

Let us turn our attention to D and E-optimality. Based on the anticipated variance, we can formulate the objective functions under these optimality criteria as

$$\text{D-optimality: } \min_{\boldsymbol{\pi}} \det\left(\text{Var}_a(\hat{\boldsymbol{\theta}}_\pi)\right) \Leftrightarrow \min_{\boldsymbol{\pi}} \det\left(\sum_k \frac{\widetilde{\boldsymbol{W}}_k}{\pi_k}\right),$$

$$\text{E-optimality: } \min_{\boldsymbol{\pi}} \max_{|a|=1} \text{Var}_a(\boldsymbol{a}^T\hat{\boldsymbol{\theta}}_\pi) \Leftrightarrow \min_{\boldsymbol{\pi}} \max_{|a|=1} \sum_k \frac{\boldsymbol{a}^T\widetilde{\boldsymbol{W}}_k\boldsymbol{a}}{\pi_k}$$

$$\Leftrightarrow \min_{\boldsymbol{\pi}} \max_{i=1,\ldots,p} \lambda_i(\boldsymbol{\pi}), \quad \lambda_i \text{ eigenvalues of } \sum_k \frac{\widetilde{\boldsymbol{W}}_k}{\pi_k}.$$

The non-linearity of these objective functions above have a few consequences of importance. Explicit formulas for the D and E-optimal designs do rarely exist, and numerical optimization must be used. As for L-optimality, methods based on Lagrangian multipliers are suitable. This becomes quite computer intensive for large samples, since the dimension of the problem equals the number of elements in S_1.

Comments

Note for L-optimality that the optimal design compensate large values of c_k by large values of π_k. Even though less explicit, a similar statement can be made about D and E-optimality. In general, the optimal designs compensate 'large' $\widetilde{\boldsymbol{W}}_k$ by large values of π_k, where the meaning of 'large' is determined by the optimality criteria used. Recall from (3.1.9) and the discussion that followed that $\widetilde{\boldsymbol{W}}_k = \boldsymbol{I}_a(\boldsymbol{\theta})^{-1}\,\text{E}_{(Y_k,\boldsymbol{X}_k)|\boldsymbol{Z}_k}\left(\boldsymbol{s}_k\boldsymbol{s}_k^T\right)\boldsymbol{I}_a(\boldsymbol{\theta})^{-1}$, which could be interpreted as the anticipated precision in the first phase weighted by the anticipated influence of element k in estimation. The precision in the first phase is thus taken into account in the selection of subsampling design, and the sampling schemes compensate low precision in the first phase by high precision in the second phase.

Perspective and Sources

As mentioned in Chapters 1.1 and 2.3, precision optimization by design in the context of two-phase sampling has previously been addressed in the literature [15, 27, 38, 39]. Much of the previous work in the area is however limited in the classes of estimators and models considered. Some of the previous work in the topic is summarized by Fuller [17] Chapter 3, including optimal Poisson sampling schemes. Fuller also address optimization with respect multiple finite population characteristics, each characteristic with a certain required degree of accuracy, and a cost function associated with sampling of elements. His presentation is however limited to one-dimensional parameters in the finite population setting. Optimal two-phase sampling designs for multidimensional pseudo-likelihood estimators have not been encountered in the literature during the writing of this thesis.

3.3 Some Modifications

As mentioned in Chapter 2.3, criticism had been directed towards the pseudo-likelihood approach to estimation. The PLE might be inefficient, due to the lack of incorporation of auxiliary information in estimation and due the use of the sampling weights $1/\pi_k$ in the pseudo log-likelihood. Additional variance in estimation is also introduced when using random size designs, such as Poisson sampling. Some post hoc modifications of design and estimation procedure are therefore proposed, aiming at reducing the variance of the PLE.

3.3.1 Adjusted Conditional Poisson Sampling

As already mentioned, a drawback with Poisson sampling is the random sample size. This is undesirable for a number of different reasons. First, it is often practical and convenient to have a fixed sample size. One might otherwise by chance end up with a sample size larger or smaller than desired. A too large sample increase the cost of conducting a study. If the study involves human subjects, it is for ethical reasons not appreciated to recruit more participants than necessary. Ending up with a too small sample on the other hand, the desired precision in estimation might not be reached. Also, the random sample size increase variance in estimation, since an additional random element is introduced.

Rather than considering Poisson sampling, one might ask if there is an optimal sampling design with unequal probabilities and fixed size. However, the requirement of fixed sample size introduces dependencies between the indicator variables I_k, so that variance-covariance matrix of the PLE becomes a function of both first and second order inclusion probabilities. Finding an optimal sampling scheme for such a design might not be possible. On the other hand, the second order inclusion probabilities and their influence on the design variance are probably of relatively little importance compared to the first order inclusion probabilities. Ignoring the second order inclusion probabilities, we are back in the same optimization problem as for Poisson sampling. It might thus be possible to use the optimal Poisson sampling design as basis for a design with fixed size.

The simplest way to obtain a fixed size design from a Poisson sampling design is to draw repeated Poisson samples until the desired sample size is obtained, discarding all samples of the wrong size. Such a design is called *conditional Poisson sampling*, abbreviated CP-sampling. However, the inclusion probabilities under CP-sampling differ from the ones in the underlying Poisson sampling design, but is possible to adjust the CP-design so that the desired inclusion probabilities are obtained, called *adjusted conditional Poisson sampling*. A proposal of a fixed size design with unequal probabilities is to use an adjusted CP-design with inclusion probabilities that are optimal under Poisson sampling. Even though such a design might not be optimal, it will probably have a performance similar to that of Poisson sampling, with the attractive additional property of having fixed size.

3.3.2 Stratified Sampling

Another proposal for obtaining fixed size designs is by use of stratified sampling. So far, sampling probabilities have been assigned on element level. Suppose instead that elements are grouped into H disjoint subgroups or strata G_h, each consisting of N_h elements. Suppose also that we assign sampling probabilities $\pi_k = f_h$ to all elements $k \in G_h$, so that elements are sampled with equal probabilities within groups but possibly different probabilities between groups. The anticipated variance then becomes

$$\mathrm{Var}_a(\hat{\boldsymbol{\theta}}_\pi) = \sum_{h=1}^{H} \frac{\sum_{k \in G_h} \widetilde{\boldsymbol{W}}_k}{f_h} \ .$$

For elements in $k \in G_h$, let

$$\overline{\boldsymbol{W}}_k = \frac{1}{N_h} \sum_{l \in G_h} \widetilde{\boldsymbol{W}}_l \ .$$

The anticipated variance can then be written as

$$\mathrm{Var}_a(\hat{\boldsymbol{\theta}}_\pi) = \sum_{k} \frac{\overline{\boldsymbol{W}}_k}{\pi_k} \ ,$$

from which we see that optimization can carried out as before. However, the optimal solution now satisfies $\pi_k^* = f_h^*$ for all elements in $k \in G_h$. Since elements in the same subgroup have equal probabilities of being sampled, f_h^* can be interpreted as a the optimal sampling fraction in group G_h. It is then easy to fix the sample size by using simple random sampling within strata, sampling $n_h = N_h f_h^*$ elements from group G_h. By this procedure, we arrive at stratified sampling.

The above approach restricts the inclusion probabilities to be equal within strata, so the set of feasible vectors $\boldsymbol{\pi}$ is restricted. The optimal design with this restriction is thus never better than the optimal solution without this restriction. Also, it might be necessary to round $n_h = N_h f_h^*$ to obtain integer sample sizes, so that the actual inclusion probabilities used differ from those found to be optimal. A fixed size design is finally applied to a set of sampling fraction optimized for a random size design, and the

implications of this step on the performance of the design is not evident. Even though the resulting design might not be optimal, the proposed approach have some advantages over the optimal Poisson design. First, a fixed size design is obtained. Also, when assigning inclusion probabilities on element level, it is important that the influence of each specific element on the variance of the PLE can be guessed with high certainty. If this is not the case, it might be more reasonable to assign equal probabilities to groups of similar elements, guarding against errors made on element level. It can also be more convenient to assign probabilities on groups of elements than using different inclusion probabilities for all elements.

3.3.3 Post-Stratification

Some methods for weight adjustment, such as calibration weighting, weight optimization and smoothing, were discussed in Section 2.3.1. These methods have been proposed as improvements to classical the PLE by adjusting the sampling weights in the pseudo log-likelihood. One simple such method will now be presented, called *post-stratification*. In contrast to the modifications proposed in the previous sections, post-stratification is a method for sample size correction applied in the estimation procedure rather than in design. The idea is the following. Suppose that S_1 consists of a number of subgroups of known sizes which could serve as strata, but the group membership is unknown before sampling, so that stratified sampling can not be conducted. After sampling, the group membership is observed, and it might be that the representation of the subgroups in the sample is different from what would have been obtained if groups were known and stratified sampling used. By post-stratification, this issue is addressed after sampling by adjustment of the sampling weights in the Horvitz-Thompson estimator.

As in Section 3.3.2, suppose that S_1 consists of H disjoint subgroups G_h, each with N_h elements, but that the subsampling design is not stratified sampling. Based on the selected sample, the Horvitz-Thompson estimator $N_{h,\pi}$ of the size of group G_h is

$$\hat{N}_{h,\pi} = \sum_{k \in G_h} \frac{I_k}{\pi_k} \ ,$$

which is constant and equal to N_h for a fixed size design, such as stratified sampling, but has non-zero variance for random size designs. Since the sampling weights $1/\pi_k$ enter any Horvitz-Thompson estimator of a population characteristic, having $\hat{N}_{h,\pi} > N_h$ leads to over representation of group G_h in estimation, and the opposite when having $\hat{N}_{h,\pi} < N_h$. The contribution of the H subgroups to estimation is thus not balanced in random size designs. However, by adjusting the sampling weights in group G_h by a factor

$$g_h = \frac{N_h}{\hat{N}_{h,\pi}} \ ,$$

we see that

$$\tilde{N}_{h,\pi} = \sum_{k \in G_h} \frac{I_k g_h}{\pi_k} = \frac{N_h}{\hat{N}_{h,\pi}} \hat{N}_{h,\pi} = N_h \ .$$

Knowing the group sizes N_h and replacing the sampling weights $1/\pi_k$ by g_h/π_k, $k \in G_h$, balance can be achieved even for random size designs. Note that this method actually incorporates the auxiliary information about group sizes in estimation.

The role of such weight adjustment in the context of maximum pseudo-likelihood can be understood intuitively in the same way. The number of elements in group G_h is known to be N_h. Having $\hat{N}_{h,\pi} > N_h$ means that the size of this group is overestimated, resulting in too large influence of the elements in this group on the pseudo log-likelihood. Since we know the true size of G_h, we can reduce the influence of the elements in G_h by a factor $N_h/\hat{N}_{h,\pi}$. In the same way, $\hat{N}_{h,\pi} < N_h$ correspond to underestimation of the contribution of elements in G_h to the pseudo log-likelihood, and the same adjustment applies. Estimation of the PLE using post-stratification is easily carried out by adjusting the sampling weights $1/\pi_k$ by a factor $g_h = N_h/\hat{N}_{h,\pi}$, $k \in G_h$. After sampling is conducted with to the optimal inclusion probabilities found for the classical PLE, the use of post-stratification in the estimation procedure can be proposed as an additional step for variance reduction.

Perspective and Sources

Poisson sampling and adjusted CP-sampling were both introduced by and thoroughly studied by Hájek [22, 23]. Sampling from an adjusted CP-design can be quite complicated, but efficient methods for samples of moderate size have been developed by Chen et al. [13] and Tillé [46]. Implementations are available in the R function 'UPmaxentropy' in the 'sampling' package [3].

The procedure for stratification presented in this chapter have not been encountered in the literature during the writing of this thesis. However, the use of stratified sampling based on the anticipated variance has proved fruitful for estimation of finite population characteristics, see for example Godfrey et al. [20] and Kott [29].

Post-stratification is one of the simplest methods for weight adjustment, and more involved methods are commonly applied in survey sampling, see Särndal et al. [45] for an overview. More complicated procedures for weight adjustment might also be used, such the weight optimization and weight smoothing [28, 33], but is is unclear if more complicated weight adjustments lead to any improvements when the design is optimized for the standard sampling weights.

It shall be noted that all the methods proposed in this section changes the variance of the estimator compared to the classical PLE under Poisson sampling. It is therefore unclear how application of these methods influence the performance and optimality of the sampling scheme derived for the classical pseudo-likelihood under Poisson sampling. None of the methods are claimed to be optimal among the class of adjusted CP-designs, stratified sampling designs and sampling weight adjusted pseudo-likelihood methods, but they could possible lead to some improvement of the optimal sampling scheme derived for the classical PLE under Poisson sampling.

4

Examples

A number of examples are presented, illustrating the methods for optimal subsampling designs derived in Chapter 3. Optimal designs are derived and discussed for estimation of parameters of the normal distribution in various situations with auxiliary data. Some somewhat realistic scenarios using logistic regression are studied by means of simulations.

Methods for finding L, D and E-optimal designs based on the anticipated variance were presented in Chapter 3.2. We have seen that L-optimal designs can be found explicitly, while D and E-optimality require numerical approximations. This in turn leads to computer intensive optimization problems. For computational reasons, most of this chapter will thus be concerned with optimization with respect to some linear optimality criteria.

4.1 The Normal Distribution

In this section we consider a situation without explanatory variables. Let $Y_k \sim \mathcal{N}(\mu, \sigma)$ and S_2 be a Poisson sample from S_1 with $\mathrm{E}(|S_2|) = n$.

4.1.1 L-Optimal Designs for (μ, σ)

Consider first the degenerate case where $\boldsymbol{Z} = \boldsymbol{Y}$ and $\boldsymbol{\theta}$ is known, so that complete information about the study variable and even of the population parameter is available in the selection of subsampling design. Even though this eliminates the need for a second sampling phase to be conducted and is thus of no use in practice, it is suitable for illustration and discussion of optimal sampling schemes.

Optimality Criteria

Optimal sampling schemes will be derived for a variety of linear optimality criteria, namely:

$$\begin{array}{l} 1.\ \min \mathrm{Var}(\hat{\mu}_\pi) \\ 2.\ \min \mathrm{Var}(\hat{\sigma}_\pi) \\ 3.\ \min \mathrm{trace}\, \mathrm{Var}[(\hat{\mu}_\pi, \hat{\sigma}_\pi)] \Leftrightarrow \min \mathrm{Var}(\hat{\mu}_\pi) + \mathrm{Var}(\hat{\sigma}_\pi) \\ 4.\ \min \mathrm{Var}(\hat{\mu}_\pi + z\hat{\sigma}_\pi) \Leftrightarrow \min \mathrm{Var}(\hat{\mu}_\pi) + 2z\,\mathrm{Cov}(\hat{\mu}_\pi, \hat{\sigma}_\pi) + z^2\,\mathrm{Var}(\hat{\sigma}_\pi) \end{array} \tag{4.1.1}$$

Criterion 3 is the same as the A-optimality criterion. The last criterion aims at minimizing the variance in estimation a percentile of the distribution.

Since $\boldsymbol{Y} = \boldsymbol{Z}$ and $\boldsymbol{\theta}$ is known, the anticipated variance (3.1.6) coincide with the realized variance (3.1.5), and the matrices $\boldsymbol{W}_k = \boldsymbol{I}(\boldsymbol{\theta})^{-1}\boldsymbol{s}_k\boldsymbol{s}_k^T\boldsymbol{I}(\boldsymbol{\theta})^{-1}$ are known. Depending on the optimality criteria, the optimal choice of $\boldsymbol{\pi}$ should be such that a particular linear combination of the elements of the matrix

$$\mathrm{Var}_a(\hat{\boldsymbol{\theta}}_\pi) = \sum_k \frac{\widetilde{\boldsymbol{W}}_k}{\pi_k} = \sum_k \frac{\boldsymbol{W}_k}{\pi_k} = \sum_k \frac{\boldsymbol{I}(\boldsymbol{\theta})^{-1}\boldsymbol{s}_k\boldsymbol{s}_k^T\boldsymbol{I}(\boldsymbol{\theta})^{-1}}{\pi_k}$$

is minimized. Using the formulas for $\boldsymbol{I}(\boldsymbol{\theta})^{-1}$ and $\boldsymbol{s}_k$ derived in Examples 2.2.2 and 2.3.1, we have that

$$\widetilde{\boldsymbol{W}}_k = \boldsymbol{W}_k = \frac{\sigma^2}{N^2}\begin{pmatrix} \left(\frac{y_k-\mu}{\sigma}\right)^2 & \frac{1}{2}\left(\left(\frac{(y_k-\mu}{\sigma}\right)^3 - \frac{y_k-\mu}{\sigma}\right) \\ \frac{1}{2}\left(\left(\frac{(y_k-\mu}{\sigma}\right)^3 - \frac{y_k-\mu}{\sigma}\right) & \frac{1}{4}\left(\left(\frac{(y_k-\mu}{\sigma}\right)^2 - 1\right)^2 \end{pmatrix}. \tag{4.1.2}$$

Optimal Sampling Schemes

The objective functions and the optimal sampling schemes for the four optimality criteria (4.1.1) are given in Table 4.1, where the optimal sampling schemes are given by (3.2.5). The optimal inclusion probabilities as function of the standardized distance from y_k to the mean, $(y_k - \mu)/\sigma$, is presented in Figure 4.1. The optimal designs with respect to the four optimality criteria (4.1.1) will be referred to as Design 1, Design 2, etc.

Design 1 is optimized for precision in estimation of μ. Each element contributes with a term $\frac{\sigma^2}{N^2}\frac{1}{\pi_k}(\frac{y_k-\mu}{\sigma})^2$ to the realized variance. Assuming for a moment that all elements are sampled with equal probabilities $\pi_k = \pi$, we see that elements with y_k relatively close to μ will have little influence on the variance while elements with y_k relatively far away from μ will have large influence on the variance. One might thus guess that elements with y_k in the tail of the distribution should be sampled with large probabilities, while elements close to the mean should be sampled with lower probabilities. In this way, large variance contributions due to large $(y_k - \mu)^2/\sigma^2$ are compensated by small $1/\pi_k$ and vice versa. This does also agree with he idea that the term $1/\pi_k$ can be thought of as the number of elements being represented by element k. Tail observations in a normal

distribution represent a small proportion of the population, and should have π_k large. Observations in the center of the distribution could be represented by a single element, and such elements should have π_k small. This is also seen in Table 4.1 and in the top left of Figure 4.1, presenting the optimal design for μ.

Design 2 has optimal inclusion probabilities quite different from Design 1. Elements that have y_k close to one standard deviation from μ have little impact on the objective function and are thus sampled with small probability. Elements in the tails are sampled with even higher probability than with Design 1.

By adding the variances of $\hat{\mu}_\pi$ and $\hat{\sigma}_\pi$ and completing the square, the objective function for the third criterion in (4.1.1) can be written as in Table 4.1. The resulting optimal inclusion probabilities is a second order function of the standardized deviation of y_k from the mean. Note that the two objectives of minimizing $\mathrm{Var}(\hat{\mu}_\pi)$ and minimizing $\mathrm{Var}(\hat{\sigma}_\pi)$ somehow work against each other. Those elements that barely contribute to one of the variances contribute to the other, so no element is sampled with as low probability as in Design 1 or 2.

Design 4 is optimized for estimation of a percentile of the distribution, and the optimal design depend on which percentile to estimate. Optimal sampling schemes for precision in estimation of the 5-th and 97.5-th percentile are presented Figure 4.1, corresponding to $z = -1.65$ and $z = 1.96$ in criterion 4 in (4.1.1), respectively. Design 4 reminds of Design 2 for these percentiles, but is slightly shifted. In general, elements in the left tail will be sampled with high probability if interest is in estimation of a percentile to the left of the distribution, and similarly for percentiles to the right. Note also that the first criterion is a special case the fourth, namely estimation of the 50-th percentile.

Table 4.1: Objective functions and optimal sampling schemes for estimation of parameters of the normal distribution with respect to various linear optimality criteria, namely; minimization of the variance of $\hat{\mu}_\pi$ (Criterion 1), the variance of $\hat{\sigma}_\pi$ (Criterion 2), the average variance (Criterion 3) and the variance in estimation of a percentile (Criterion 4). $w_k^{(i,j)}$ stands for the (i,j)-th element of the matrix $\widetilde{\boldsymbol{W}}_k$ in (4.1.2). The constant κ is chosen so that the expected sample size is n.

Criterion	Objective Function	Optimal Sampling Scheme
1.	$\min_{\boldsymbol{\pi}} \frac{\sigma^2}{N^2} \sum_k \frac{1}{\pi_k} \left(\frac{y_k-\mu}{\sigma}\right)^2$	$\pi_k^* = \kappa \lvert y_k - \mu \rvert$
2.	$\min_{\boldsymbol{\pi}} \frac{\sigma^2}{4N^2} \sum_k \frac{1}{\pi_k} \left(\left(\frac{y_k-\mu}{\sigma}\right)^2 - 1\right)^2$	$\pi_k^* = \kappa \lvert \left(\frac{y_k-\mu}{\sigma}\right)^2 - 1 \rvert$
3.	$\min_{\boldsymbol{\pi}} \frac{\sigma^2}{4N^2} \sum_k \frac{1}{\pi_k} \left(\left(\frac{y_k-\mu}{\sigma}\right)^2 + 1\right)^2$	$\pi_k^* = \kappa \left(\left(\frac{y_k-\mu}{\sigma}\right)^2 + 1\right)$
4.	$\min_{\boldsymbol{\pi}} \frac{\sigma^2}{N^2} \sum_k \frac{1}{\pi_k} \left(w_k^{(1,1)} + 2z w_k^{(1,2)} + z^2 w_k^{(2,2)}\right)$	$\pi_k^* = \kappa \sqrt{w_k^{(1,1)} + 2z w_k^{(1,2)} + z^2 w_k^{(2,2)}}$

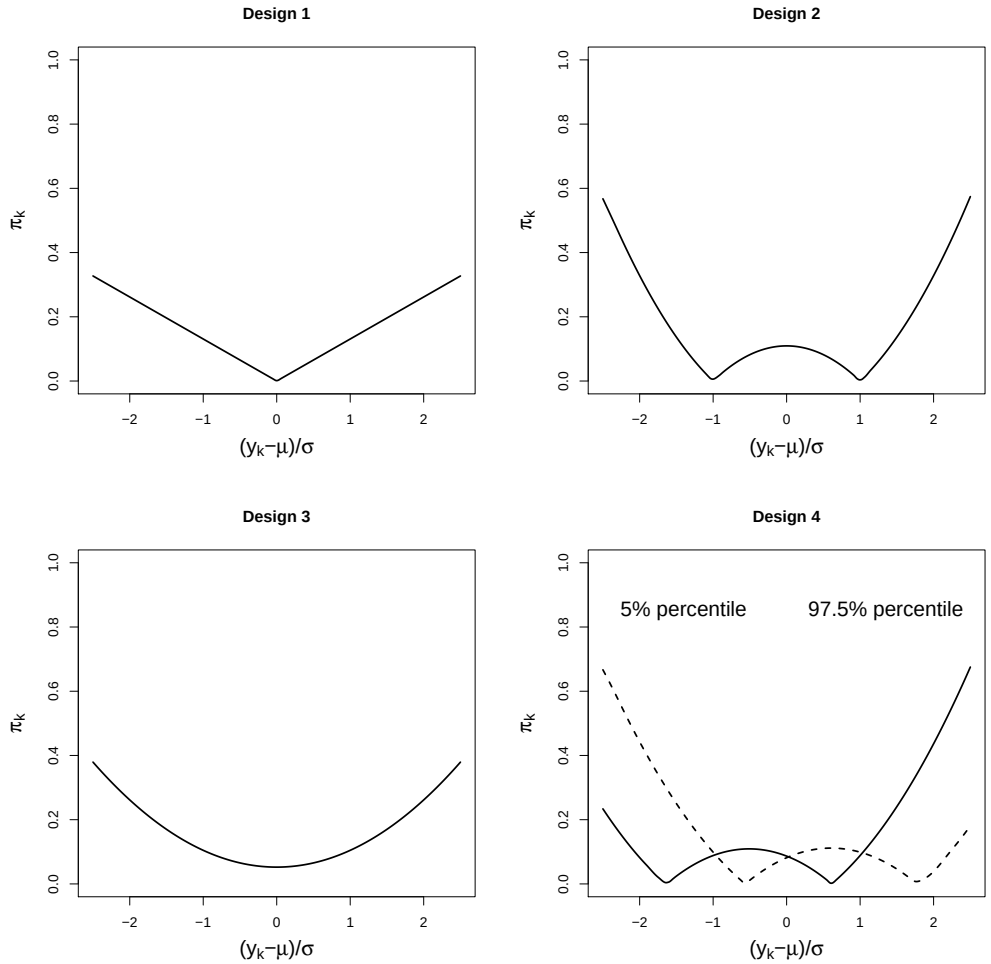

Figure 4.1: Optimal sampling schemes for estimation of parameters of the normal distribution with respect to various linear optimality criteria, namely; minimization of the variance of $\hat{\mu}_\pi$ (top left), the variance of $\hat{\sigma}_\pi$ (top right), the average variance (bottom left) and the variance in estimation of a percentile (bottom right). The optimal inclusion probabilities are presented as functions of the standardized distance from y_k to the mean. Sampling schemes are presented for $N = 500$ and $n = 50$.

Confidence Regions

The 95% confidence ellipses for the optimal designs under the four optimality criteria (4.1.1) are presented in Figure 4.2. The confidence ellipse using Bernoulli sampling is presented with dashed red lines for reference. The center of the confidence ellipses if the estimated value of $\boldsymbol{\theta}$. The projection of the confidence ellipse of the MLE on the x-axis gives us the classical 95% confidence interval for μ.

Compared to Bernoulli sampling, Designs 1 and 2 yield reduced variance in estimation

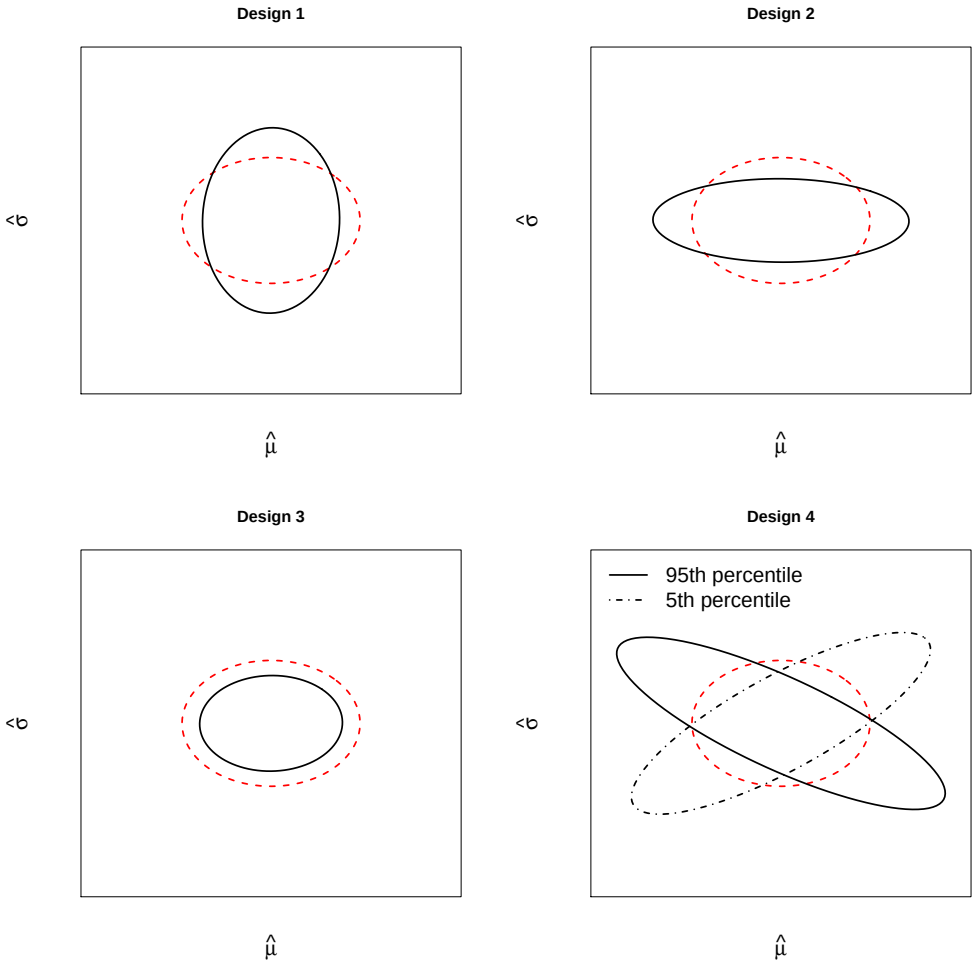

Figure 4.2: 95% confidence regions for the parameters (μ, σ) of a normal distribution under optimal designs with respect to various linear optimality criteria, namely; minimization of the variance of $\hat{\mu}_\pi$ (top left), the variance of $\hat{\sigma}_\pi$ (top right), the average variance (bottom left) and the variance in estimation of a percentile (bottom right). 95% confidence ellipses for the optimal designs are presented as black ellipses, the red dashed ellipses are 95% confidence ellipses using Bernoulli sampling and are presented for reference. The confidence ellipses are presented for a specific first phase sample with $N = 500$ and $n = 50$.

of the parameter for which the design was optimized, but the variance of the other estimator is increased. Note that the estimators are independent under these designs. The estimators are independent also in Design 3, which is the A-optimal design. It is evidently also better than Bernoulli sampling in terms of D-optimality, the entire confidence ellipse being contained within the confidence ellipse pertaining to Bernoulli sampling. It gives almost the same precision in estimation of μ as Design 1 and almost the same precision in estimation of σ as Design 2.

Design 4 is optimized for estimation of the 95-th and 5-th percentile separately, which is the same percentiles as in as in Figure 4.1. In (4.1.1), this correspond to $z = 1.96$ and $z = -1.65$ in Criterion 4. In order to estimate the 95-th percentile with high precision, the design compensates for underestimation of μ by overestimation of σ. This introduces a negative correlation between the two estimators. The same applies for designs optimized for precision in estimation of any percentile in the right tail of the distribution. Similarly, a positive correlation is seen when Design 4 is optimized estimation of the 5-th percentile, or any percentile in the left tail. Note that the variances of both estimators are larger for Design 4 with $z = -1.65$ and $z = 1.96$ compared to Bernoulli sampling, but increased precision along the direction $\mu + z\sigma$ is achieved.

4.1.2 D and E-Optimal Designs for (μ, σ)

In the same setting as in the previous section, i.e. with full knowledge of $\boldsymbol{Y}$ and $\boldsymbol{\theta}$ in the design stage, D and E-optimal sampling schemes will now be found numerically and illustrated graphically. The objective functions are

$$\text{D-optimality: } \min_{\boldsymbol{\pi}} \det\left(\sum_k \frac{\widetilde{\boldsymbol{W}}_k}{\pi_k}\right) , \tag{4.1.3}$$

$$\text{E-optimality: } \min_{\boldsymbol{\pi}} \max_{i=1,\dots,p} \lambda_i, \quad \lambda_i \text{ eigenvalues of } \sum_k \frac{\widetilde{\boldsymbol{W}}_k}{\pi_k} , \tag{4.1.4}$$

where $\boldsymbol{W}_k$ is given in (4.1.2).

The optimal inclusion probabilities as function of the standardized distance from y_k to the mean, $(y_k - \mu)/\sigma$, are presented in Figure 4.3. Optimization of (4.1.3) and (4.1.4) was carried out using an augmented Lagrangian minimization algorithm implemented in the function 'auglag' in the R package 'alabama' [4]. The A-optimal design for (μ, σ), i.e. Design 3 in (4.1.1), was used as initial guess of $\boldsymbol{\pi}^*$. 95% confidence ellipses corresponding to these designs are presented in Figure 4.4. The D-optimal design is similar to the A-optimal design in Section 4.1.1, both in terms of inclusion probabilities and confidence region. The E-optimal design remind about Design 1 in Section 4.1.1, but with much better control of the variance in $\hat{\sigma}_\pi$.

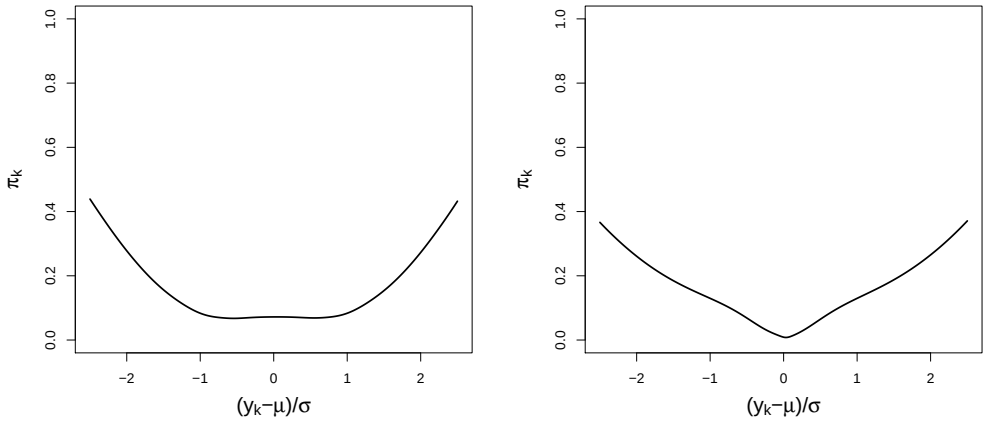

Figure 4.3: D-optimal (left) and E-optimal (right) sampling schemes for estimation of parameters of the normal distribution. The optimal inclusion probabilities are presented as functions of the standardized distance from y_k to the mean. Sampling schemes are presented for $N = 500$ and $n = 50$.

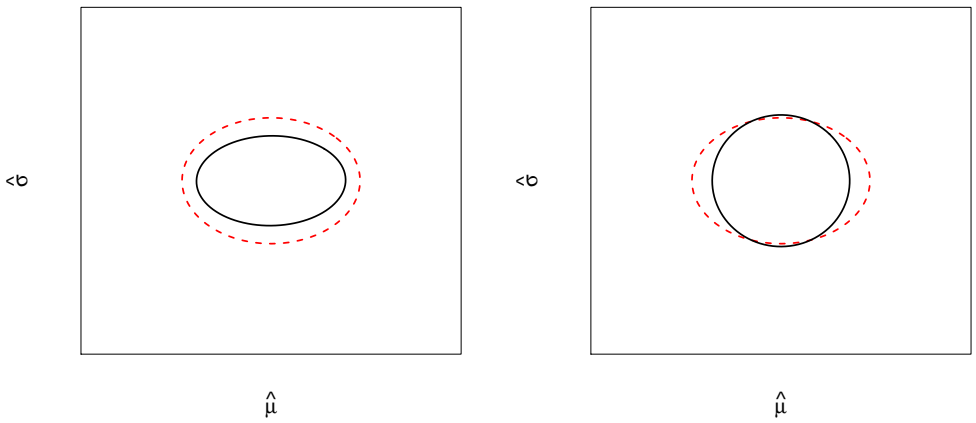

Figure 4.4: 95% confidence regions for the parameters (μ, σ) of a normal distribution under D-optimal (left) and E-optimal (right) designs presented as black ellipses. The red dashed ellipses are 95% confidence ellipses using Bernoulli sampling and are presented for reference. The confidence ellipses are presented for the same first phase sample as in Figure 4.2, with $N = 500$ and $n = 50$.

4.1.3 Optimal Sampling Schemes for μ Revisited

Even though the examples in Sections 4.1.1 and 4.1.2 are interesting and illustrative for the theory presented, they are unlikely useful in practice. Evaluation of the objective

functions and calculation of optimal inclusion probabilities require knowledge of y_k as well as of μ and σ. If such information were available there would have been no need for a second sampling phase to be carried out. Let us now consider a situation where less information is available in the design stage.

The Design Model and the Anticipated Variance

Suppose that (Y_k, Z_k) follow a bivariate normal distribution, i.e.

$$(Y_k, Z_k) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\mu} = (\mu_Y, \mu_Z), \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_Y^2 & \rho\sigma_Y\sigma_Z \\ \rho\sigma_Y\sigma_Z & \sigma_Z^2 \end{pmatrix} .$$

Conditional on Z_k, this implies that

$$Y_k|Z_k \sim \mathcal{N}\left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_Z}(Z_k - \mu_Z), \sqrt{(1-\rho^2)}\sigma_Y\right) . \tag{4.1.5}$$

Suppose now that z_k is observed in the first phase, and that the aim of the study is to estimate and draw inference about $\theta = \mu_Y$. Under the model (4.1.5), the anticipated variance (3.1.8) becomes

$$\begin{aligned}
\mathrm{Var}_a(\boldsymbol{\theta}) &= \sum_k \frac{1}{\pi_k} \boldsymbol{I}_a(\mu_Y)^{-1} \mathrm{E}_{Y_k|z_k}\left(\boldsymbol{s}_k \boldsymbol{s}_k^T\right) \boldsymbol{I}_a(\mu_Y)^{-1} \\
&= \frac{\sigma_Y^2}{N^2} \sum_k \frac{1}{\pi_k} \mathrm{E}_{Y_k|z_k}\left[\left(\frac{Y_k - \mu_Y}{\sigma_Y}\right)^2\right] \\
&= \frac{\sigma_Y^2}{N^2} \sum_k \frac{1}{\pi_k} \frac{(z_k - \mu_Z)^2\rho^2\frac{\sigma_Y^2}{\sigma_Z^2} + (1-\rho^2)\sigma_Y^2}{\sigma_Y^2} \\
&= \frac{\sigma_Y^2}{N^2} \sum_k \frac{1}{\pi_k} \left(1 - \rho^2 + \left(\frac{z_k - \mu_Z}{\sigma_Z}\right)^2 \rho^2\right) ,
\end{aligned}$$

using the fact that $\boldsymbol{I}_a(\mu_Y) = \boldsymbol{I}(\mu_Y) = N/\sigma_Y^2$.

Optimal Sampling Schemes

Since a one-dimensional parameter is of interest, all optimality criteria are equivalent and the optimal solution can be found explicitly using the methods for L-optimality. Note that $\frac{\sigma_Y^2}{N^2}$ is just a constant and can be neglected in the optimization. Plugging in $c_k = \left(1 - \rho^2 + \left(\frac{z_k - \mu_Z}{\sigma_Z}\right)^2 \rho^2\right)$ in (3.2.5), we obtain the optimal sampling scheme as

$$\pi_k = \kappa\sqrt{1 - \rho^2 + \left(\frac{z_k - \mu_Z}{\sigma_Z}\right)^2 \rho^2} , \tag{4.1.6}$$

where, as usual, κ chosen so that the expected sample size is n. It is interesting that neither of μ_Y and σ_Y need to be known, it suffices to know the standardized difference from z_k to the mean μ_Z and the correlation between $\boldsymbol{Z}$ and $\boldsymbol{Y}$. The design parameter is thus $\boldsymbol{\phi} = (\mu_Z, \sigma_Z, \rho)$. The parameters μ_Z and σ_Z can be estimated from S_1, while ρ must be guessed based on previous knowledge.

It was shown in Section 4.1.1 that elements with extreme value of y_k, i.e. with $|y_k - \mu_Y|$ large, should be sampled with high probability. Since y_k is unknown, the value of $|y_k - \mu_Y|$ must be guessed based on z_k. If the correlation between $\boldsymbol{Y}$ and $\boldsymbol{Z}$ is small, knowing z_k does not tell much about y_k, and we are not sure whether y_k is extreme or not. Elements are thus sampled with almost equal probabilities. On the other hand, knowing z_k makes it possible to guess y_k with high certainty if the correlation is large. In particular, elements that have extreme values of z_k are likely also to have extreme values of y_k, and should thus be sampled with high probability, while elements with z_k close to the mean should be sampled with low probability. In agreement with this discussion, Formula (4.1.6) shows that $\pi_k \to \kappa = \pi$ for all k as $\rho \to 0$. In particular, this means that Bernoulli sampling is optimal if $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are uncorrelated. We also see that $\pi_k \to \kappa' |z_k - \mu_Z|$ as $\rho \to \pm 1$. Note that this is the same as $\pi_k = \kappa'' |y_k - \mu_Y|$ in the limit, so for $\rho = \pm 1$ we arrive at the same optimal solution as when minimizing $\mathrm{Var}(\hat{\mu}_\pi)$ in Section 4.1.1, presented in Table 4.1.

For large correlations, the optimal inclusion probabilities are essentially proportional to $|z_k - \mu_Z|$. It is a standard method in survey sampling to have probability of inclusion proportional to the size of an auxiliary variable correlated to the variable of interest. This is called sampling with *probability proportional to size*, abbreviated πps-sampling. In this setting, we see that the proportionality should be with respect to the absolute deviation from z_k to its mean, rather than directly proportional to z_k.

Simulated Performance

We investigate the performance of the sampling scheme (4.1.6) by use of simulations. The variance of $\hat{\mu}_{Y,\pi}$ under optimal designs for different guessed values of ρ is compared to Bernoulli sampling. The relative variance of the estimator is presented as a function of the true correlation in [0,1], using the variance under Bernoulli sampling as reference. The results are presented in Figure 4.5. Guessing that $\rho = 0.1$ is essentially the same as to use Bernoulli sampling. Guessing that $\rho = 0.5$ gives variances close to Bernoulli sampling when the true correlation is small, but leads to some improvement when the true correlation is large. Assuming that $\rho = 0.9$ gives a large variance reduction when the true correlation is large, but might actually result in increased variance when the true correlation is small. The reason is that the design model disagree with the true model, so that optimization is carried out with respect to the wrong objective function. In conclusion, the sampling scheme (4.1.6) is optimal when the guessed correlation and the true correlation agree, but requires quite large correlations for substantial improvements to be made. The performance is reasonably good even if the correlation is not guessed correctly, as long as the guess is not too far from the truth.

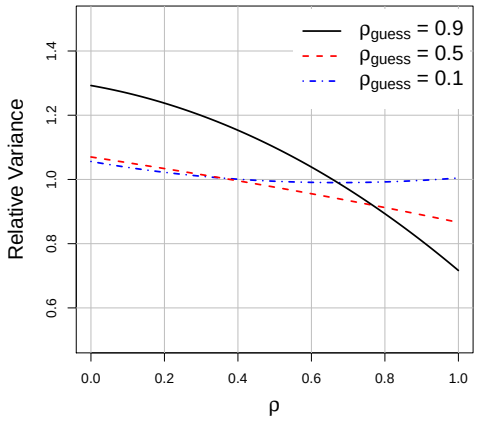

Figure 4.5: Simulated relative variance of $\hat{\mu}_{Y,\pi}$ under various designs compared to Bernoulli sampling, as function of $\rho(\boldsymbol{Y},\boldsymbol{Z})$. The designs are optimized for estimation of μ_Y by use of auxiliary information for various guessed values of $\rho(\boldsymbol{Y},\boldsymbol{Z})$. Simulations were performed with $N = 500$, $n = 50$ and (Y_k, Z_k) following a bivariate normal distribution with standard normal marginal distributions.

4.2 Logistic Regression

Let us now consider a situation models including explanatory variables. Logistic regression models are considered in this section, but the methods can be applied to other types of models in the class of generalized linear models in a similar fashion.

We start by briefly describing the logistic regression model. A binary response variable Y is explained by a set of explanatory variables as follows. First, the probability mass function of $Y \sim Bernoulli(p)$ can be written as

$$f(y;p) = p^y(1-p)^{1-y} = e^{y\log\frac{p}{1-p}+log(1-p)}, \quad y \in \{0,1\}, \quad p \in [0,1] \ .$$

With logistic regression, it is assumed that the log-odds of Y_k, $\log\left(\frac{p_k}{1-p_k}\right)$, is a linear function of some explanatory variables $\boldsymbol{x}_k = (1, x_{k,1}, \ldots, x_{k,p})$, so that

$$\log\left(\frac{p_k}{1-p_k}\right) = \beta_0 + \beta_1 x_{k,1} + \ldots + \beta_p x_{k,p} = \boldsymbol{x}^T\boldsymbol{\beta} \quad \Leftrightarrow \quad p_k = \frac{1}{1+e^{-\boldsymbol{x}_k^T\boldsymbol{\beta}}} \ ,$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ are regression coefficients. The parameter of interest is $\boldsymbol{\theta} = \boldsymbol{\beta}$. Note that we include a column of ones in $\boldsymbol{X}$, corresponding to β_0 in the logistic model.

The Pseudo-Likelihood

For an observed sample $(y_k, \boldsymbol{x}_k), k \in S_1$, the likelihood function takes the form

$$\mathcal{L}(\boldsymbol{\beta}; \boldsymbol{y}, \tilde{\boldsymbol{X}}) = \prod_k f(y_k|\boldsymbol{x}_k;\boldsymbol{\beta}) = \prod_k e^{y_k\boldsymbol{x}_k^T\boldsymbol{\beta} - \log(1+e^{\boldsymbol{x}_k^T\boldsymbol{\beta}})} \ ,$$

and the corresponding log-likelihood is

$$\ell(\boldsymbol{\beta}; \boldsymbol{y}, \tilde{\boldsymbol{X}}) = \sum_k y_k \boldsymbol{x}_k^T \boldsymbol{\beta} - \log(1 + e^{\boldsymbol{x}_k^T \boldsymbol{\beta}}) \ .$$

The pseudo log-likelihood is thus

$$\ell_\pi(\boldsymbol{\beta}; \boldsymbol{y}, \tilde{\boldsymbol{X}}) = \sum_k \frac{I_k}{\pi_k} \left(y_k \boldsymbol{x}_k^T \boldsymbol{\beta} - \log(1 + e^{\boldsymbol{x}_k^T \boldsymbol{\beta}}) \right) \ ,$$

which we use for estimation under two-phase sampling. The partial derivatives of the pseudo log-likelihood are

$$\frac{\partial \ell_\pi(\boldsymbol{\beta}; \boldsymbol{y}, \tilde{\boldsymbol{X}})}{\partial \beta_i} = \sum_k \frac{I_k}{\pi_k} \left(y_k x_{k,i} - \frac{e^{\boldsymbol{x}_k^T \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_k^T \boldsymbol{\beta}}} x_{k,i} \right) = \sum_k \frac{I_k}{\pi_k} x_{k,i} (y_k - p_k) \ .$$

The PLE $\hat{\boldsymbol{\beta}}_\pi$ is defined as the solution to the estimating equation

$$\nabla_{\boldsymbol{\beta}} \ell_\pi(\boldsymbol{\beta}; \boldsymbol{y}, \tilde{\boldsymbol{X}}) = \boldsymbol{0} \ .$$

Explicit solutions to the above equation does in general not exist and $\boldsymbol{\beta}$ must be estimated numerically. Estimation procedures for the PLE in generalized linear models are available in the 'svyglm' function in the R package 'survey' [2, 32].

The Anticipated Variance

The elements of the Fisher information of $\boldsymbol{\beta}$ are given by

$$\boldsymbol{I}(\boldsymbol{\beta})_{(i,j)} = \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{X}} \left(\sum_k x_{k,i} x_{k,j} (Y_k - p_k)^2 \right) = \sum_k x_{k,i} x_{k,j} \operatorname{Var}(Y_k) = \sum_k x_{k,i} x_{k,j} p_k (1 - p_k) \ ,$$

so the information matrix can be written as

$$\boldsymbol{I}(\boldsymbol{\beta}) = \sum_k p_k (1 - p_k) \boldsymbol{x}_k \boldsymbol{x}_k^T = \sum_k \operatorname{Var}(Y_k) \boldsymbol{x}_k \boldsymbol{x}_k^T \ .$$

According to (3.1.4), the asymptotic variance of $\hat{\boldsymbol{\beta}}_\pi$ is thus

$$\begin{aligned} \operatorname{Var}(\hat{\boldsymbol{\beta}}_\pi) &\underset{a}{=} \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{X}} \left(\sum_k \frac{\boldsymbol{W}_k}{\pi_k} \right) \\ &= \boldsymbol{I}(\boldsymbol{\beta})^{-1} \left(\sum_k \frac{\mathrm{E}_{Y_k|\boldsymbol{X}_k}(\boldsymbol{s}_k \boldsymbol{s}_k^T)}{\pi_k} \right) I(\boldsymbol{\beta})^{-1} \ , \end{aligned}$$

where

$$\boldsymbol{s}_k = \nabla_{\boldsymbol{\beta}} \log f(y_k | \boldsymbol{x}_k; \boldsymbol{\beta}) = (y_k - p_k) \boldsymbol{x}_k \ .$$

Similarly, the anticipated variance is given by

$$\mathrm{Var}_a(\hat{\boldsymbol{\beta}}_\pi) = \boldsymbol{I}_a(\boldsymbol{\beta})^{-1} \left(\sum_k \frac{\mathrm{E}_{(Y_k, \boldsymbol{X}_k)|\boldsymbol{Z}_k} \left(\boldsymbol{s}_k \boldsymbol{s}_k^T \right)}{\pi_k} \right) \boldsymbol{I}_a(\boldsymbol{\beta})^{-1} \ ,$$

where

$$\boldsymbol{I}_a(\boldsymbol{\beta}) = \sum_k \mathrm{E}_{\boldsymbol{X}_k|\boldsymbol{Z}_k} \left(p_k(1-p_k) \boldsymbol{X}_k \boldsymbol{X}_k^T \right) \ , \tag{4.2.1}$$

and

$$\mathrm{E}_{(Y_k, \boldsymbol{X}_k)|\boldsymbol{Z}_k} \left(\boldsymbol{s}_k \boldsymbol{s}_k^T \right) = \mathrm{E}_{(Y_k, \boldsymbol{X}_k)|\boldsymbol{Z}_k} \left((Y_k - p_k)^2 \boldsymbol{X}_k \boldsymbol{X}_k^T \right) \ . \tag{4.2.2}$$

Note that (4.2.1) and (4.2.2) depend on the unknown regression coefficients $\boldsymbol{\beta}$ through $p_k = 1/(1 + \exp(-\boldsymbol{x}_k^T \boldsymbol{\beta}))$.

4.2.1 A Single Continuous Explanatory Variable

The examples following in this chapter deals with subsampling when the outcome is known, as in case-control studies. Models with a single continuous explanatory variable are considered, assuming a regression like relationship between the unobserved explanatory variable a known auxiliary variable. Of particular interest in this chapter is estimation of β_1. This is the regression coefficient corresponding to $\boldsymbol{X}$, which describes the relation between $\boldsymbol{X}$ and $\boldsymbol{Y}$. We consider the simple logistic regression model

$$\log \left(\frac{p_k}{1-p_k} \right) = \beta_0 + \beta_1 x_k \ ,$$

where x is a continuous variable. Let $(\beta_0, \beta_1) = (0,1)$. The model is illustrated in Figure 4.6, showing the probabilities that $Y_k = 1$ a function of x_k.

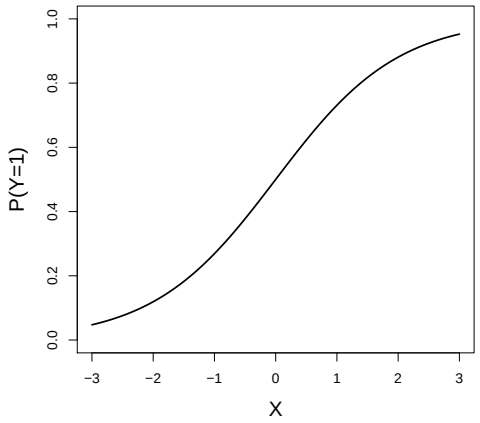

Figure 4.6: A logistic regression model with a single continuous explanatory variable X, showing the probability that $Y = 1$ as function of X.

Minimizing the Realized Variance

Assume that, in addition to the outcome y_k, also p_k, $\boldsymbol{x}_k$ and $\boldsymbol{I}(\boldsymbol{\beta})^{-1}$ are known in the planning of subsampling design. The anticipated variance(3.1.6) coincide with the realized variance (3.1.5), which can be found by plugging in the values of y_k, p_k, $\boldsymbol{x}_k$ and $\boldsymbol{I}(\boldsymbol{\beta})^{-1}$ in Formulas (4.2.1) and (4.2.2). More realistic situations with less information available in the design stage will be considered later on. The sampling scheme optimized for maximal precision in estimation of β_1 is illustrated in Figure 4.7 with inclusion probabilities both as function of x_k and p_k. Note that observations with $p_k = P(Y_k = 1) < 0.5$ but with $y_k = 1$ are sampled with high probability, and similarly for observations with $y_k = 0$ and $p_k > 0.5$. That is, observations that deviate from the model are sampled with high probability. For $\beta_0 \neq 0$ non-symmetric sampling schemes are obtained, in contrast to this example where the sampling scheme is symmetric around $P(Y = 1) = 0.5$. However, it holds in general that the inclusion probabilities increase as the deviation between observed data and model increase. In particular, if $y_k = 1, y_l = 0$ and $p_k = p_l < 0.5$ it holds that $\pi_k > \pi_l$, and vice versa.

An interesting remark can be made about optimal sampling schemes in the study of rare events, such as the outcomes studied in case-control studies. Subjects with the condition of interest, i.e with $Y_k = 1$, are often called *cases* and the rest are called *controls*. Case-control studies are typically applied for rare diseases, having p_k is relatively small for all elements since the outcome $Y = 1$ is rare. This implies that $\boldsymbol{s}_k = (y_k - \pi_k)\boldsymbol{x}_k \approx \boldsymbol{0}$ for controls and $\boldsymbol{s}_k = (y_k - \pi_k)\boldsymbol{x}_k \approx \boldsymbol{x}_k$ for cases. Cases should thus be sampled with high probability since they deviate from the underlying population model and have a relatively large contribution to the variance of the estimator. This is in agreement with standard methodology in case-control sampling.

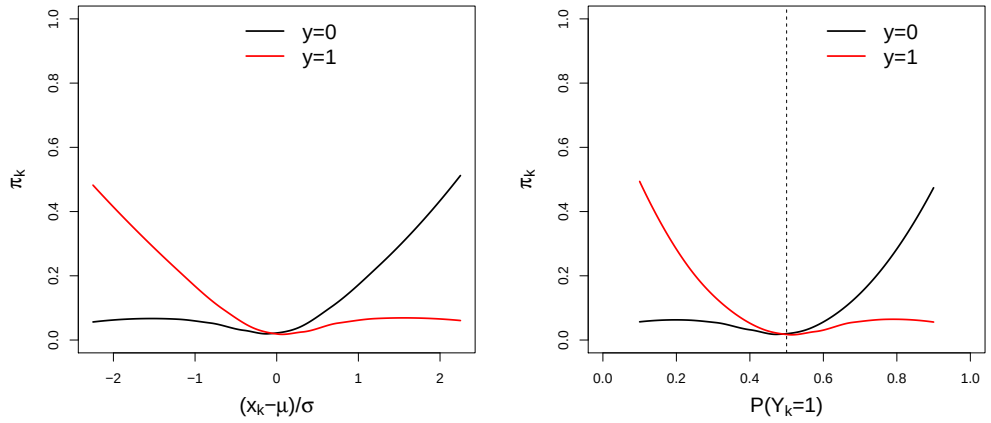

Figure 4.7: Optimal sampling schemes for estimation of β_1 in a simple logistic regression model with a single continuous explanatory variable, where β_1 is the regression coefficient pertaining to the explanatory variable. The inclusion probabilities are presented as function of the standardized distance of x_k from the mean (left) and as function of $p_k = P(Y_k = 1)$ (right). In this instance, we have $X \sim \mathcal{N}(0,1)$, $N = 500$, $n = 50$ and $\boldsymbol{\beta} = (0, 1)$.

Minimizing the Anticipated Variance

Suppose now that the explanatory variable is unknown and that some auxiliary variable is observed in the first phase. Consider the case were (X_k, Y_k) follow a bivariate normal distribution, so that

$$(X_k, Z_k) \sim \mathcal{N}\left(\begin{pmatrix}\mu_X \\ \mu_Z\end{pmatrix}, \begin{pmatrix}\sigma_X^2 & \rho\sigma_X\sigma_Z \\ \rho\sigma_X\sigma_Z & \sigma_Z^2\end{pmatrix}\right) .$$

This implies that

$$X_k|Z_k \sim \mathcal{N}\left(\mu_X + \rho\frac{\sigma_X}{\sigma_Z}(Z_k - \mu_Z), \sqrt{(1-\rho^2)}\sigma_X\right) , \tag{4.2.3}$$

which also can be expressed as

$$X_k = \alpha_0 + \alpha_1 z_k + \varepsilon_k, \quad \varepsilon_k \sim \mathcal{N}(0, \sigma_\varepsilon) \quad \Leftrightarrow \quad X_k|Z_k \sim \mathcal{N}\left(\alpha_0 + \alpha_1 Z_k, \sigma_\varepsilon\right) . \tag{4.2.4}$$

Since $\boldsymbol{y}$ and $\boldsymbol{z}$ are known for all elements in S_1, we can approximate

$$p_k = \frac{1}{1 + e^{-\boldsymbol{x}_T\boldsymbol{\beta}}}$$

by

$$\hat{p}_k = \frac{1}{1 + e^{-\boldsymbol{x}_T\hat{\gamma}_{ML}}} , \tag{4.2.5}$$

where $\hat{\gamma}_{ML}$ is the MLE of the regression coefficients when modeling $\boldsymbol{Y}$ as function of $\boldsymbol{Z}$ with logistic regression using the data available in S_1. We need to compute (4.2.1) and (4.2.2) under the design model (4.2.3) in order to find the anticipated variance. Consider first (4.2.1), which can be written as

$$\begin{aligned}
&\mathrm{E}_{(Y_k,\boldsymbol{X}_k)|\boldsymbol{Z}_k}\left((Y_k-p_k)^2\boldsymbol{X}_k\boldsymbol{X}_k^T\right) = (y_k-\hat{p}_k)^2\,\mathrm{E}_{\boldsymbol{X}_k|\boldsymbol{Z}_k}\left(\boldsymbol{X}_k\boldsymbol{X}_k^T\right)\\
&= (y_k-\hat{p}_k)^2\,\mathrm{E}_{\boldsymbol{X}_k|\boldsymbol{Z}_k}\begin{pmatrix}1 & X_k\\ X_k & X_k^2\end{pmatrix}\\
&= (y_k-\hat{p}_k)^2\begin{pmatrix}1 & \mathrm{E}_{\boldsymbol{X}_k|\boldsymbol{Z}_k}(X_k)\\ \mathrm{E}_{\boldsymbol{X}_k|\boldsymbol{Z}_k}(X_k) & \mathrm{Var}_{\boldsymbol{X}_k|\boldsymbol{Z}_k}(X_k)+\mathrm{E}_{\boldsymbol{X}_k|\boldsymbol{Z}_k}(X_k)^2\end{pmatrix},
\end{aligned} \tag{4.2.6}$$

where $\mathrm{E}_{\boldsymbol{X}_k|\boldsymbol{Z}_k}(X_k)$ or $\mathrm{Var}_{\boldsymbol{X}_k|\boldsymbol{Z}_k}(X_k)$ can be computed under the design model (4.2.3) and (4.2.4). The above matrix can be evaluated if the parameters in the model for X_k given Z_k is known or could be guessed. We have here used the fact y_k is known and p_k estimated from the data available in the first phase. The anticipated information matrix can be found similarly to the calculation in (4.2.6).

Sampling schemes optimized for precision in estimation of β_1 are illustrated for various correlations between $\boldsymbol{X}$ and $\boldsymbol{Z}$ in Figure 4.8. When $\rho = 0$, the auxiliary variable contains no information about the explanatory variable and the optimal sampling scheme is a function of the outcome alone. To be more specific, the optimal inclusion probabilities are given by

$$\pi_k^* = \kappa|y_k - \bar{y}| \ .$$

In this instance we have $\bar{y} = 0.5$, and Bernoulli sampling is optimal . As $\rho \to \pm 1$, the sampling schemes converge to the optimal sampling scheme found when $\boldsymbol{x}$ was known, compare to Figure 4.7.

In more complicated situations, it might be necessary to guess the value of $\boldsymbol{\theta}$. In this case this was not needed since $\boldsymbol{\theta}$ only enter the anticipated variance through p_k, which could be estimated directly using the auxiliary variable. Note that $\hat{p}_k$ will be close to $\bar{y}$ for all k when $\rho(\boldsymbol{X}, \boldsymbol{Z}) \approx 0$, where $\bar{y}$ is the mean of $\boldsymbol{y}$ in S_1. For strong correlations between $\boldsymbol{X}$ and $\boldsymbol{Z}$, $\hat{p}_k$ will be close to p_k, but the estimated probabilities will in general be pulled from the true probabilities towards $\bar{y}$. It is so since $\boldsymbol{Z}$ essentially is as a disturbed version of $\boldsymbol{X}$, and adding noise to an explanatory variable does in general pull the corresponding regression coefficient towards zero, thus pulling the predicted probabilities towards the mean.

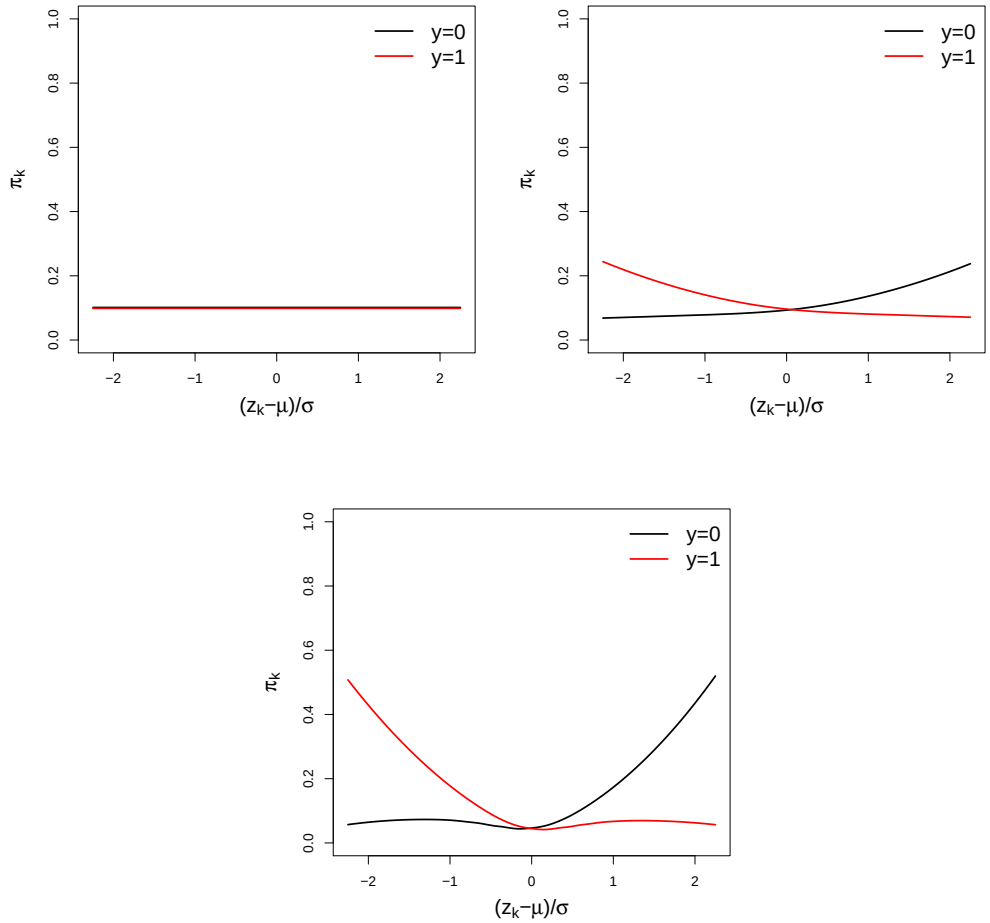

Figure 4.8: Optimal sampling schemes for estimation of β_1 in a simple logistic regression model with a single continuous explanatory variable, where β_1 is the regression coefficient pertaining to the explanatory variable. The inclusion probabilities are presented as function of the standardized distance of z_k from the mean. The designs are optimized for estimation of β_1 by use of auxiliary information under the design model (4.2.3). In this instance, we have $N = 500$, $n = 50$, $\boldsymbol{\beta} = (0, 1)$ and (X_k, Z_k) following a bivariate normal distribution with standard normal marginal distributions and correlation $\rho(\boldsymbol{X},\boldsymbol{Z}) = 0$ (top left), $\rho(\boldsymbol{X},\boldsymbol{Z}) = 0.5$ (top right) and $\rho(\boldsymbol{X},\boldsymbol{Z}) = 0.95$ (bottom).

4.2.2 Auxiliary Information with Proper Design Model

Let us now consider a somewhat more realistic situation, trying to mimic the possible use of optimal subsampling designs in practice. We continue the example in the previous section and consider logistic regression with a single continuous explanatory variable.

The outcome and an auxiliary variable are observed in the first phase, while the explanatory variable is unobserved. Since it is believed that the explanatory variable is

related both to the outcome and to the auxiliary variable, it is reasonable to assume some kind of indirect relation between the auxiliary variable and the outcome. This motivates the use of estimated probabilities from a model for $\boldsymbol{Y}$ given $\boldsymbol{Z}$ as approximations of p_k, as given by (4.2.5) in the previous section. A regression like relationship between $\boldsymbol{X}$ and $\boldsymbol{Z}$ is assumed, i.e. a model of the form

$$X_k = \alpha_0 + \alpha_1 z_k + \varepsilon_k, \quad \varepsilon_k \sim \mathcal{N}(0, \sigma_\varepsilon) \tag{4.2.7}$$

is used to describe the relationship between auxiliary and explanatory variable. Assume that some knowledge about the design model parameter $\boldsymbol{\phi} = (\alpha_0, \alpha_1, \sigma_\varepsilon)$ is available from a previous study.

Simulations were used to investigate the performance of the optimal two-phase design optimized for estimation of β_1 under these settings. A bivariate normal model for (X_k, Z_k) with standard normal marginal distributions was used. A random sample of size $N = 500$ was generated from the joint distribution of $(\boldsymbol{X}, \boldsymbol{Z})$, followed by simulation of $\boldsymbol{Y}$ according to a logistic model with explanatory variable $\boldsymbol{X}$ and parameters $(\beta_0, \beta) = (0,1)$. A separate sample of 50 observations from the joint distribution of $(\boldsymbol{X}, \boldsymbol{Z})$ was generated and used to estimate (α_0,α_1) and σ_ε in (4.2.7), mimicking existing knowledge of the distribution of $\boldsymbol{X}$ given $\boldsymbol{Z}$ from a previous study. Optimal sampling schemes were found under the estimated design model for $\boldsymbol{X}$ given $\boldsymbol{Z}$ as in the previous section. The variance of the PLE was then estimated using simulations of 5000 subsamples S_2 of size $n = 50$, estimating the total variance as the average over the simulated variances plus the variance of the MLE in the first phase.

Varying Correlation

Simulations were conducted for various correlations $\rho(\boldsymbol{X},\boldsymbol{Z})$ between 0 and 1. Four different designs were considered, namely Poisson sampling and adjusted CP-sampling, both with optimal Poisson sampling inclusion probabilities, stratified sampling with optimal sampling fractions and simple random sampling. Bernoulli sampling was also considered for reference. Strata where defined by grouping elements with $y_k = 1$ into two almost equal size groups by a split at the median of z_k for such elements, and similarly for elements with $Y_k = 0$. Stratified sampling was achieved by restricting the inclusion probabilities to be equal within strata and using the optimal inclusion probabilities as sampling fractions within strata. By the construction of strata, the relation between $\boldsymbol{X}$ and $\boldsymbol{Z}$ and the structure of true model for $\boldsymbol{Y}$ given $\boldsymbol{X}$, the four strata essentially consist of elements with $p_k < 0.5$ and $y_k = 0$, $p_k < 0.5$ and $y_k = 1$, $p_k > 0.5$ and $y_k = 0$, and $p_k > 0.5$ and $y_k = 1$.

The results of the simulations are shown in Figure 4.9, presenting the ratio of variance of the estimators under the four designs compared to the variance under Bernoulli sampling. Neither of simple random sampling and Bernoulli sampling use the auxiliary information obtained between the two sampling phases, and the variances under these designs do not depend on ρ. Note that Poisson sampling gives the same precision as Bernoulli sampling and stratified sampling the sample precision as simple random sampling when $\rho(\boldsymbol{X},\boldsymbol{Z}) = 0$. In fact the two random size designs are essentially equivalent

in this instance, since $\boldsymbol{Z}$ contains no information about $\boldsymbol{X}$ and $\bar{y} = 0.5$, and similarly for the two fixed size designs. If the outcome groups are not equally large, the knowing y_k would in general lead to improvement of the stratified sampling over simple random sampling and of Poisson sampling over Bernoulli sampling. In addition, as the correlation increase, the auxiliary variable provides more information about the unknown explanatory variable. Utilizing this information in the design stage leads to a variance reduction of 25-50% for moderate to large correlations.

While the variances under simple random sampling and adjusted CP-sampling are quite similar to the variances under their random size counterparts, giving a variance improvement of no more than a few percentages, a large gain in precision is seen when using stratified sampling. Even in the extreme case where $\rho(\boldsymbol{X},\boldsymbol{Z}) = 1$, stratified sampling is superior to Poisson sampling. By construction of homogeneous strata in combination with fixed size sampling, balance between the outcome and covariate groups in the subsamples is ensured, resulting in small variance. With more heterogeneous strata or by use of random size sampling from each strata, less variance reduction is expected.

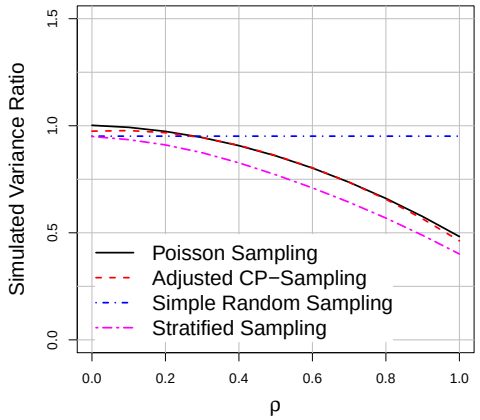

Figure 4.9: Relative variance of $\hat{\beta}_{1,\pi}$ under various designs optimized for estimation of β_1 compared to Bernoulli sampling, where β_1 is the regression coefficient pertaining to the explanatory variable in a logistic regression model with a single continuous explanatory variable. The simulated variance ratio is shown as function of the true correlation between the explanatory variable $\boldsymbol{X}$ and the auxiliary variable $\boldsymbol{Z}$. In this instance, we have $N = 500$, $n = 50$, $\boldsymbol{\beta} = (0, 1)$ and (X_k, Z_k) following a bivariate normal distribution with standard normal marginal distributions. The design model was estimated from a previous study of 50 observations.

Varying other Parameters

Only a single parameter is varied in Figure 4.9, namely $\rho(\boldsymbol{X},\boldsymbol{Z})$. Preliminary simulations were also conducted for a fixed correlation $\rho(\boldsymbol{X},\boldsymbol{Z}) = 0.5$, varying other parameters.

Figures presenting these results are omitted, but the results are discussed and motivated intuitively.

Consider first the impact on the sample sizes in the two phases of sampling on the possible gain in precision. Thinking of a situation with $N = n$, we realize that all elements in S_1 are certainly included in S_2, so that all designs are equivalent. Keeping this in mind, the results presented in Figure 4.9 shows that the variance of the PLE can be reduced when N is enlarged even though n is held fixed, since the designs are equivalent when $N = n = 50$ but Poisson sampling is superior to Bernoulli sampling when $N = 500$ and $n = 50$. This is motivated by the fact that more information becomes available when the size of S_1 increases, so that more efficient designs can be found. The sizes of both S_1 and S_2 thus influence the possible gain in precision in two-phase sampling. However, it turns out that the gain by using Poisson sampling is much larger for $n = 50$ and $N = 500$ than for $n = 500$ and $N = 5000$, even though $n/N = 0.1$ in both cases. For small and moderate sample size, the variability between samples will be high with Bernoulli sampling and the extreme observations will influence the variance of the estimators substantially. Sampling with unequal probabilities allows for the influence of extreme elements to be reduced by sampling these with high probability. However, if n and N are both sufficiently large, Bernoulli sampling will essentially always produce balanced samples that well represent the underlying population. As a consequence, the variability between samples will be small, resulting in estimators with small variances. The gain in more complex designs is thus maximal for small to moderate n and quite large N.

The strength of association between outcome and explanatory variables is also of importance for the performance of the subsampling designs. A strong relation here correspond to a large β_1 is large relatively to the variability in the explanatory variable. This gives a better separation between observations with $Y_k = 1$ from observations with $Y_k = 0$, so that the variability of $\boldsymbol{Y}$ is reduced. In particular, this enlarges the influence of the term $(y_k - p_k)^2$ on the anticipated variance for elements that deviate from the model, so that high sampling probabilities could be assigned to such elements. This in turn leads to reduced variance in estimation compared to equal probability sampling designs. Putting this slightly different, a strong association between the response and the explanatory variable implies a strong association between variables in the design model, which if utilized in the design stage could lead to high gain in precision.

As a final remark on this example, we also comment on the importance of the prior information available about the model for $\boldsymbol{X}$ given $\boldsymbol{Z}$. There are two sources of uncertainty in this model as estimated from previous studies, namely uncertainty of the estimated parameters and variability of observations around the model. The former is small if the sample size of the previous study is sufficiently large, but the latter has to do with the randomness in underlying population model and can only be reduced by finding good auxiliary variables. For construction of efficient subsampling designs, the precision in the estimated design model parameters is of secondary importance. Of main importance is the knowledge about the variability of the unknown variable in the underlying population, and to find auxiliary variables that explain a large proportion of the vari-

ance in the unobserved variables. Even though a previous study of sufficient size to be able specify the design model correctly and estimate its parameters with high precision is desirable, it is not necessary. Similar results to those presented in Figure 4.9 where obtained even if the previous study had only 20 observations. Despite the fact that this resulted in quite poor precision in estimation of $\boldsymbol{\phi}$, it gave sufficient information about the strength of association between $\boldsymbol{X}$ and $\boldsymbol{Z}$ in order for the auxiliary information to be used efficiently in selection of subsampling design.

4.2.3 Auxiliary Information with Improper Design Model

The previous example used knowledge about the design model from a hypothetical previous study. This allowed for specification of a reasonable structure of the design model, using the estimated parameters as a guess of the design model parameter $\boldsymbol{\phi}$. Consider now a situation where such knowledge is not available, but where data is available for an auxiliary variable that is believed to be strongly correlated with the explanatory variables.

Let us continue the previous example, only modifying the information available about $\boldsymbol{X}$ given $\boldsymbol{Z}$. A regression like relationship is assumed, but the parameters of the regression model are unknown. Assuming that $\boldsymbol{X}$ and $\boldsymbol{Z}$ are highly correlated, it is tempting to plug in $\boldsymbol{Z}$ direct in the anticipated variance in place of $\boldsymbol{X}$. Plugging in $\boldsymbol{X} = \boldsymbol{Z}$ and $\hat{p}_k$ as estimated in (4.2.5) into (4.2.2) gives

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{X}_k|\boldsymbol{Z}_k}\left(\boldsymbol{s}_k\boldsymbol{s}_k^T\right) &= (y_k - \hat{p}_k)^2\,\mathrm{E}_{\boldsymbol{X}_k|\boldsymbol{Z}_k}(\boldsymbol{X}_k\boldsymbol{X}_k^T) \\
&= (y_k - \hat{p}_k)^2 \begin{pmatrix} 1 & \mathrm{E}_{\boldsymbol{X}_k|\boldsymbol{Z}_k}(X_k) \\ \mathrm{E}_{\boldsymbol{X}_k|\boldsymbol{Z}_k}(X_k) & \mathrm{Var}_{\boldsymbol{X}_k|\boldsymbol{Z}_k}(X_k) + \mathrm{E}_{\boldsymbol{X}_k|\boldsymbol{Z}_k}(X_k)^2 \end{pmatrix} \\
&= (y_k - \hat{p}_k)^2 \begin{pmatrix} 1 & Z_k \\ Z_k & Z_k^2 \end{pmatrix}.
\end{aligned} \tag{4.2.8}$$

Note that Formula (4.2.8) is the same as (4.2.6) under the assumption that the correlation between $\boldsymbol{X}$ and $\boldsymbol{Z}$ is 1. Plugging in $\tilde{\boldsymbol{Z}}$ in place of $\boldsymbol{X}$ is thus to assume that the two variables are strongly related. In addition, it also has implications on the prior belief in $\boldsymbol{\beta}$, namely that $\boldsymbol{\beta} = \boldsymbol{\gamma}$, where $\boldsymbol{\gamma}$ is the regression coefficient for the logistic model for $\boldsymbol{Y}$ given $\boldsymbol{Z}$. Suppose for example that the true correlation between $\boldsymbol{X}$ and $\boldsymbol{Z}$ is 0. Two major mistakes will then be made when plugging in $\boldsymbol{X} = \tilde{\boldsymbol{Z}}$ in the anticipated variance. First, this is the same as to assume that the correlation between $\boldsymbol{X}$ and $\boldsymbol{Z}$ is 1. Second, it will be assumed that $\beta_1 = \gamma_{1,ML} \approx 0$ since $\boldsymbol{Z}$ is only associated with $\boldsymbol{Y}$ indirectly through the common association with $\boldsymbol{X}$. Simply plugging in the value of a proxy variable in the anticipated variance makes strong assumptions of the relation between $\boldsymbol{X}$ and $\boldsymbol{Z}$, and is reasonable only when $\boldsymbol{X}$ and $\boldsymbol{Z}$ are strongly correlated.

Simulated Performance

The above approach was investigated by means of simulations for various designs optimized for estimation of β_1, as in the previous example. The true correlation between $\boldsymbol{X}$ and $\boldsymbol{Z}$ was varied between 0 and 1. Three different sampling strategies were considered, namely Poisson sampling, Poisson sampling with post-stratification and stratified sampling. Bernoulli sampling was also considered for reference. Strata and post-strata where defined by grouping elements with $y_k = 1$ into two almost equal size groups by a split at the median of z_k for such elements, and similarly for elements with $Y_k = 0$. The Poisson design with post-stratification had the same inclusion probabilities as the standard Poisson design, applying sample size adjustment of the sampling weights in estimation. Stratified sampling was achieved by restricting the inclusion probabilities to be equal within strata and using the optimal inclusion probabilities as sampling fractions within strata.

The results are presented in Figure 4.10, showing the relative variance of the PLE of $\hat{\beta}_1$ under the three designs compared to Bernoulli sampling. Note that the variance under Poisson sampling is almost twice as large as for Bernoulli sampling when the correlation between $\boldsymbol{X}$ and $\boldsymbol{Z}$ is small. This is due to the strong assumptions implied by plugging in $\tilde{\boldsymbol{Z}}$ in place of $\boldsymbol{X}$ in the anticipated variance, actually assuming perfect correlation. Optimization is thus conducted with respect to the wrong objective function, resulting in large variance. As the true correlation increases, the assumption of perfect correlation becomes more plausible and eventually leads to reduced variance compared to Bernoulli sampling. As $\rho(\boldsymbol{X}, \boldsymbol{Z})$ tend to 1, the same degree of variance reduction as in Figure 4.9 in the previous example is obtained. The variance of the PLE is reduced by use of post-stratification for small to moderate correlations. Still, the variance is much higher than under Bernoulli sampling. The gain is using post-stratification is reduced when the correlation is large, since the true optimal inclusion probabilities are then found and modification of these result in increased variance.

As in the previous example, the stratified sampling design performs remarkably well. The key to success with stratified sampling is that the strata consist of groups of elements that truly are similar in terms of the combination of outcome and covariate values. Assigning equal probabilities within strata can be seen as averaging the variance contribution over elements within strata. Even though large errors are made on element level, the construction of strata has the effect that errors are canceled out by averaging the variance contribution over elements within strata. As claimed in Section 3.3, this example confirms that assigning probabilities on cluster level could be a sensible approach when large errors are believed to made on element level and when the information available on element level is restricted.

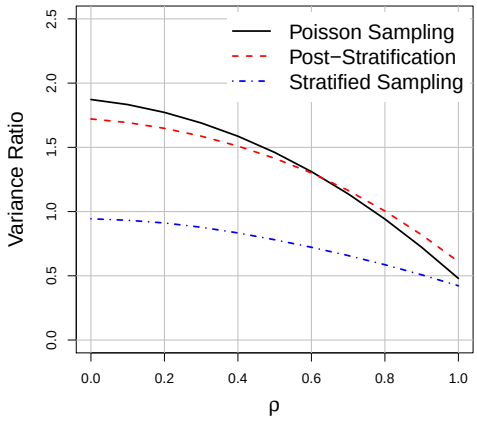

Figure 4.10: Relative variance of $\hat{\beta}_{1,\pi}$ with Poisson sampling, Poisson sampling with post-stratification and stratified sampling compared to Bernoulli sampling, as function of $\rho(\boldsymbol{X},\boldsymbol{Z})$. β_1 is the regression coefficient pertaining to the explanatory variable in a logistic regression model with a single continuous explanatory variable. The sampling schemes were optimized for precision in estimation of β_1 using $\boldsymbol{Z}$ as substitute for $\boldsymbol{X}$ in the anticipated variance.

4.2.4 When the Outcome is Unknown

The examples presented in the previous sections have all been concerned with situations where the outcome is known for all elements in S_1. This allowed for simplification of the anticipated variance, and made it possible to find and sample elements with outcome deviating from the model with high probability. However, if the outcome is unknown, it becomes more difficult to say which elements that deviate from the model. In such situations, it is important to have access to good auxiliary information for the outcome. If such information is unavailable, the best one can do is to assume that the outcome follow the underlying model. The sampling schemes will then essentially be based on the anticipated extremeness of elements in $\boldsymbol{X}$-space. So called *leverage points*, i.e. points that are extreme in $\boldsymbol{X}$-space and therefore have a large influence in estimation, should then be sampled with high probabilities.

5

Conclusion

Estimation and precision optimization under general two-phase sampling procedures have been studied. The use of auxiliary information obtained between the two phases of sampling in construction of efficient subsampling designs has been investigated. Optimal Poisson sampling designs for the PLE has been derived and presented analytically for L-optimality, and methods for numerical optimization with respect to D and E-optimality has been discussed. It was shown that elements that are believed to deviate from the population model should be sampled with high probability, which agrees with the interpretation of the inverse of the inclusion probabilities as the number of elements represented by each element in the sample.

It was shown that the variance of the PLE under the special case of Poisson sampling with $\pi_k = \pi$, i.e. Bernoulli sampling, yielded approximately the same variance as the MLE under of a simple random sampling for sufficiently large samples. The optimal Poisson design is thus expected to yield more efficient estimators than simple random sampling, provided that enough auxiliary information is available in the design stage. This motivates the use of two-phase sampling when feasible.

A large gain in terms of precision can be achieved when good auxiliary variables are available in the design stage, i.e. variables that describe a large proportion of the variability in the unobserved study variables. The gain is also expected to be large when a relatively small proportion of elements are to be sampled in the second phase. By optimal selection of design, smaller sample sizes might be needed than if simple random sampling was used, as otherwise is common practice. There could thus be an economic gain in the use of two-phase sampling and optimal subsampling designs.

In the formulation of the objective function, the anticipated variance was introduced as a substitute to the unknown realized variance of the PLE based on a heuristic argument. This requires further investigation and a more rigorous treatment. Also, since the actual variance and the anticipated variance are evaluated with respect to different models, caution is needed in the specification of design model so that inconsistency between

the two models is avoided, if possible. Inconsistencies inevitably leads to discrepancies between the anticipated and the realized variance. Still, the anticipated variance as introduced in the context of survey sampling has been proven useful, and is believed to be so even in the more complex situation with model based inference considered in this thesis. Empirical studies of this claim are requested for future study. In particular, studies of the robustness of performance of subsampling designs found under moderate inconsistencies between models is requested.

A proper specification of the design model is of great importance. Errors made in the design stage will lead to optimization of the wrong objective function, and potentially a large loss of efficiency. It is in general worse to put strong belief in the wrong model than to put weak belief in the correct model. While a design model that assumes small variation in $(\boldsymbol{Y},\boldsymbol{X})$ given $\boldsymbol{Z}$ yields optimal or close to optimal sampling schemes at the true model and true parameter, it is also sensitive to errors in the model. A design model with high uncertainty in $(\boldsymbol{Y},\boldsymbol{X})$ given $\boldsymbol{Z}$ yields less efficient designs than the former at the true model, but is in general less sensitive to errors made in the specification of the model. Ideally, the sampling schemes derived should be optimal or near optimal at the true model, while giving high precision in estimation even for moderate deviations from the true model. Further improvements of the robustness of performance of the subsampling designs could possibly be made.

Three methods were proposed for further improvement of the estimation procedure and designs, namely adjusted CP-sampling, stratified sampling and post-stratification. The first two address the variability by fixing the sample size, which also might have some practical benefits, and the last by adjusting the sampling weights in the pseudo log-likelihood. While adjusted CP-sampling and post-stratification might successfully by used to reduce variance, stratified sampling could also be a guard against design model misspecification. A large potential gain in precision was seen when using stratified sampling, while the results with post-stratification and adjusted CP-sampling were less promising. The implications of modified designs and estimation procedures on the precision in estimation is not evident and the use of these methods requires further attention.

Two other topics require more attention than given in this thesis. The first is variance estimation, which was just briefly was mentioned. Ideally, a sampling design should not only provide good estimates of the parameters of interest, but also for the variances of these parameters. Variance estimation does also require consideration in the planning and might add additional requirements on the design. The use of efficient methods for variance estimation is also a topic in itself. The second issue that requires further consideration has to do with the asymptotic properties of the PLE when the design is determined by the variables observed between the phases. In order to ensure consistency and asymptotic normality of the PLE, some conditions on the auxiliary variables and the design model are needed. Assuming that the first phase samples in the limiting procedure are independent, convergence of statistics of the study variables are ensured under general conditions, and asymptotic results and ought to hold under quite general conditions on the auxiliary variables. What is needed is essentially some conditions on

the limiting behavior of the inclusion probabilities as random functions of the auxiliary variables.

A rather strong assumption of independence between triplets $(Y_k, \boldsymbol{X}_k, \boldsymbol{Z}_k)$ has been assumed in this thesis. However, the weaker assumption that Y_k are conditionally independent given $\boldsymbol{X}_k$ is enough for the inference procedure to be valid. The additional requirements on $(Y_k, \boldsymbol{X}_k, \boldsymbol{Z}_k)$ makes important simplifications in the optimization stage. In reality, the relation between elements might be such that this does not hold. By assuming independence anyway, the methods presented in this thesis could be applied, relying on the robustness of validity of the PLE. The optimal design derived under these conditions might however be far from optimal under the true conditions. This final comment actually applies to any additional assumptions made in the design stage, and the benefit here in using the pseudo-likelihood approach for estimation shall be stressed. It allows for quite general situations and models to be considered and generous assumptions to be in the design stage, since the validity of the inference procedure is does not depend on the assumptions made in the design stage.

It shall be pointed out that the anticipated variance is computed for a specific design model with a fixed parameter $\boldsymbol{\phi}$, and also for a specific model of interest with a specific value of $\boldsymbol{\theta}$. Optimization is then conducted with respect to some optimality criteria, so that the subsampling design is optimized for a specific purpose. Still, inference from the obtained sample is not limited to the models considered in the design and optimization stage, and other models can be studied than the one for which the design was optimized.

As a final comment and suggestion, it would be interesting to study the possibility to conduct the second sampling phase in two steps, the first step using Bernoulli sampling or simple random sampling and the second using Poisson sampling. The design model and its parameters could be specified and estimated based on the sample obtained in the first step. In the second step, the so obtained design model could be used for efficient subsampling. This can be useful when too little prior information is available in the design stage for specification of the design model. Questions to be addressed are how to allocate the sampling fractions between the two steps and how to combine the information gained in each step into a single inference procedure.

A

The Variance of the Maximum Pseudo-Likelihood Estimator

The variance of the PLE will now be investigated in some more detail. First, the formula for the asymptotic variance of the PLE conditional on S_1 is derived. The influence on elements on the variance of the PLE is then studied.

A.1 Derivation of the Asymptotic Conditional Variance

The formula for the variance of $\hat{\boldsymbol{\theta}}_\pi$ around $\hat{\boldsymbol{\theta}}_{ML}$ conditional on the first phase sample will now be derived. This variance is referred to as the *conditional variance* and is according to (2.3.5) claimed to be

$$\mathrm{Var}_{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}_\pi|\boldsymbol{Y},\boldsymbol{X}) \underset{a}{=} \boldsymbol{I}(\boldsymbol{\theta})^{-1}\,\mathrm{Var}_{\boldsymbol{I}}\left[\boldsymbol{S}_\pi(\boldsymbol{\theta})\right]\boldsymbol{I}(\boldsymbol{\theta})^{-1}\ .$$

A derivation of this formula is given as follows. Using Taylor linearization of the π-expanded score in a neighborhood of $\hat{\boldsymbol{\theta}}_{ML}$, we get

$$0 = \boldsymbol{S}_\pi(\hat{\boldsymbol{\theta}}_\pi) \underset{a}{=} \boldsymbol{S}_\pi(\hat{\boldsymbol{\theta}}_{ML}) + \boldsymbol{\partial}_{\boldsymbol{\theta}}\boldsymbol{S}_\pi(\hat{\boldsymbol{\theta}}_{ML})(\hat{\boldsymbol{\theta}}_\pi - \hat{\boldsymbol{\theta}}_{ML})\ , \tag{A.1.1}$$

where the leftmost equality follows from that $\hat{\boldsymbol{\theta}}_\pi$ is chosen to satisfy $\boldsymbol{S}_\pi(\hat{\boldsymbol{\theta}}_\pi) = 0$. The approximation (A.1.1) is justified by the fact that $\hat{\boldsymbol{\theta}}_\pi$ is a consistent estimator of $\hat{\boldsymbol{\theta}}_{ML}$, so $\hat{\boldsymbol{\theta}}_\pi$ will be in an arbitrary small neighborhood of $\hat{\boldsymbol{\theta}}_{ML}$ with high certainty for large enough samples. Note that $\boldsymbol{\partial}_{\boldsymbol{\theta}}S_\pi(\hat{\boldsymbol{\theta}}_{ML})$ approximates $-\boldsymbol{I}(\boldsymbol{\theta})$ and is approximately constant for large samples, due to the law of large numbers. Rearranging and taking variances with respect to $\boldsymbol{I}$ on both sides of (A.1.1), one can write

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{I}}\left(\boldsymbol{S}_\pi(\hat{\boldsymbol{\theta}}_{ML})\right) &\underset{a}{=} \mathrm{Var}_{\boldsymbol{I}}\left(\boldsymbol{\partial}_{\boldsymbol{\theta}}S_\pi(\hat{\boldsymbol{\theta}}_{ML})(\hat{\boldsymbol{\theta}}_\pi - \hat{\boldsymbol{\theta}}_{ML})\right)\\ &\underset{a}{=} \boldsymbol{I}(\boldsymbol{\theta})\,\mathrm{Var}_{\boldsymbol{I}}\left(\hat{\boldsymbol{\theta}}_\pi - \hat{\boldsymbol{\theta}}_{ML}\right)\boldsymbol{I}(\boldsymbol{\theta})\ .\end{aligned} \tag{A.1.2}$$

A final rearrangement shows that

$$\mathrm{Var}_{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}_\pi|\boldsymbol{Y},\boldsymbol{X}) = \mathrm{Var}_{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}_\pi - \hat{\boldsymbol{\theta}}_{ML}) \underset{a}{=} \boldsymbol{I}(\boldsymbol{\theta})^{-1}\,\mathrm{Var}_{\boldsymbol{I}}\left(\boldsymbol{S}_\pi(\hat{\boldsymbol{\theta}}_{ML})\right)\boldsymbol{I}(\boldsymbol{\theta})^{-1}\ , \tag{A.1.3}$$

which concludes the derivation. See Binder [7] for a rigorous proof, dealing with the technical details.

A.2 On the Contributions to the Realized Variance

As shown in section 3.1.1, the realized variance of the PLE under Poisson sampling can be written as

$$\sum_k \frac{\boldsymbol{W}_k}{\pi_k} = \sum_k \frac{\boldsymbol{I}(\boldsymbol{\theta})^{-1}\boldsymbol{s}_k\boldsymbol{s}_k^T\boldsymbol{I}(\boldsymbol{\theta})^{-1}}{\pi_k}\ ,$$

The variance contribution thus depends both on the inclusion probability and on the matrix $\boldsymbol{W}_k = \boldsymbol{I}(\boldsymbol{\theta})^{-1}\boldsymbol{s}_k\boldsymbol{s}_k^T\boldsymbol{I}(\boldsymbol{\theta})^{-1}$. Let us have a closer look at the latter.

Note that

$$\boldsymbol{W}_k = \boldsymbol{I}(\boldsymbol{\theta})^{-1}\boldsymbol{s}_k\boldsymbol{s}_k^T\boldsymbol{I}(\boldsymbol{\theta})^{-1} = \left(\boldsymbol{I}(\boldsymbol{\theta})^{-1}\boldsymbol{s}_k\right)\left(\boldsymbol{I}(\boldsymbol{\theta})^{-1}\boldsymbol{s}_k\right)^T\ ,$$

which is the outer product of the vector $\boldsymbol{I}(\boldsymbol{\theta})^{-1}\boldsymbol{s}_k$ with itself. The (i,j)-th element of $\boldsymbol{W}_k$ is thus the product of the i-th and j-th coordinates of $\boldsymbol{I}(\boldsymbol{\theta})^{-1}\boldsymbol{s}_k$.

Since $\boldsymbol{I}^{-1}(\boldsymbol{\theta})$ is a real symmetric matrix, it has p orthonormal eigenvectors, where p is the number of parameters in $\boldsymbol{\theta}$, provided that $\boldsymbol{I}^{-1}(\boldsymbol{\theta})$ is of full rank. Let the eigenvectors of $\boldsymbol{I}^{-1}(\boldsymbol{\theta})$ be denoted by $\boldsymbol{v}_1,\ldots,\boldsymbol{v}_p$ with corresponding eigenvalues $\lambda_1,\ldots,\lambda_p$. One can write

$$\boldsymbol{s}_k = \sum_{l=1}^{p} a_l\boldsymbol{v}_l\ , \tag{A.2.1}$$

where $(a_1,\ldots,a_p)$ are the coordinates of $\boldsymbol{s}_k$ in the eigenspace. Writing the j-th coordinate of $\boldsymbol{v}_l$ as $\boldsymbol{v}_{l,j}$, the vector $\boldsymbol{s}_k$ can be written as

$$\boldsymbol{s}_{k,j} = \sum_{l=1}^{p} a_l\boldsymbol{v}_{l,j}\ ,$$

and the j-th coordinate of $\boldsymbol{I}(\boldsymbol{\theta})^{-1}\boldsymbol{s}_k$ as

$$\sum_{l=1}^{p} a_l\lambda_l\boldsymbol{v}_{l,j}\ . \tag{A.2.2}$$

From this it follows that the (i,j)-th element of the matrix $\boldsymbol{W}_k$ is of the form

$$\left(\sum_{l=1}^{p} a_l\lambda_l\boldsymbol{v}_{l,i}\right)\left(\sum_{j=1}^{p} a_k\lambda_k\boldsymbol{v}_{l,j}\right)\ , \tag{A.2.3}$$

which is the product of the i-th and the j-th coordinate of $\nabla_{\boldsymbol{\theta}} \log f(y_k|\boldsymbol{x}_k;\boldsymbol{\theta})$, weighted by the eigenvalues of $\boldsymbol{I}(\boldsymbol{\theta})^{-1}$. Recall now that the eigenvectors are the orthogonal directions in $\mathbb{R}^p$ and the corresponding eigenvalues represent the magnitude of variance along these directions. Recall also that the gradient of a function shows the direction of steepest change and its magnitude is the rate of change at that point. One can think of $\boldsymbol{s}_k$ as the direction of influence of element k on estimation, and of the size of $\boldsymbol{s}_k$ as the magnitude of influence. In words formula (A.2.3) tells us that elements with high magnitude of influence with direction of influence along directions with high uncertainty will contribute substantially to the variance in estimation. It has been shown that optimal sampling schemes compensate a large contribution to the variance due to $\boldsymbol{W}_k$ by a small contribution due to $1/\pi_k$. This means that observations that with high potential influence on the model through extreme values in unstable directions should be sampled with high probability.

B

Derivation of L-Optimal Sampling Schemes under Poisson Sampling

The optimal sampling scheme under linear optimality criteria presented in Formula (3.2.4) will now be derived. We start by presenting the solution to the relaxed problem requiring only that $\pi_k > 0$ and then solve for $\pi_k \in (0,1]$.

Result B.1 *Let $c_k > 0, k = 1, \ldots, N$. The solution to the constrained optimization problem*

$$\min_{\boldsymbol{\pi}} g(\boldsymbol{\pi}) = \sum_k \frac{c_k}{\pi_k} \ ,$$

$$\text{subject to} \quad \begin{matrix} \pi_k > 0 \\ \sum_k \pi_k = n \end{matrix} \ ,$$

where $n \leq N$, is given by

$$\pi_k^* = \frac{n}{\sum_i \sqrt{c_i}} \sqrt{c_k} \ .$$

Proof of Result B.1 *Using Lagrangian multipliers, we introduce the auxiliary function*

$$\Lambda(\boldsymbol{\pi}, \lambda) = g(\boldsymbol{\pi}) + \lambda h(\boldsymbol{\pi}) \ ,$$

where

$$g(\boldsymbol{\pi}) = \sum_k \frac{c_k}{\pi_k}, \quad h(\boldsymbol{\pi}) = \left(\sum_k \pi_k \right) - n \ .$$

Critical points to the Lagrangian are found by solving the equation system

$$\nabla \Lambda(\boldsymbol{\pi}, \lambda) = \mathbf{0} \Leftrightarrow \begin{cases} h(\boldsymbol{\pi}) = 0 \\ -\nabla g(\boldsymbol{\pi}) = \lambda \nabla h(\boldsymbol{\pi}) \end{cases} \ , \tag{B.1}$$

implying that

$$\begin{cases} \sum_k \pi_k = n \\ -\frac{\partial g(\boldsymbol{\pi})}{\partial \pi_k} = \frac{c_k}{\pi_k^2} = \lambda = \frac{\partial h(\boldsymbol{\pi})}{\partial \pi_k} \end{cases} .$$

This in turn gives

$$\frac{c_k}{\pi_k^2} = \frac{c_l}{\pi_l^2}, \quad k,l \in S_1 \ .$$

In particular, all π_k in the solution of (B.1) can be expressed in terms of π_1 as

$$\pi_k = \pi_1 \sqrt{\frac{c_k}{c_1}} \ , \tag{B.2}$$

$$\pi_1 = \frac{n}{\sum_k \sqrt{\frac{c_k}{c_1}}} \ , \tag{B.3}$$

which by insertion of (B.3) in (B.2) gives the solution

$$\pi_k^* = \frac{n}{\sum_i \sqrt{c_i}} \sqrt{c_k} \ . \tag{B.4}$$

Furthermore, the Hessian of $g(\boldsymbol{\pi})$ is positive definite in (0,1], showing that $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_N^*)$ given by (B.4) is a local minimum for $g(\boldsymbol{\pi})$ in the domain specified by the constraints. The domain and the objective function are both convex, and it follows that $\boldsymbol{\pi}^*$ is a global minimum, which proves the claim. Having $c_k > 0$ ensures that the calculations above are valid, and also that the solution is unique and that probability sampling design is obtained.

Remark B.1 *Note that the expected number of elements sampled from a specific subset $\mathcal{J}$ of S_1 is given by*

$$n_j = \mathrm{E}_{\boldsymbol{I}} \left(\sum_{k \in \mathcal{J}} \pi_k \right) \ .$$

Suppose that $\boldsymbol{\pi}^*$ is a solution to the optimization problem in Result B.1, and that we want to change the expected number of sampled elements in $\mathcal{J}$ from n_j to n_j', keeping π_k fixed for elements not contained in $\mathcal{J}$. A simple calculation shows that the optimal solution to the new problem is given by

$$\tilde{\pi}_k^* = \begin{cases} \pi_k^* \frac{n_j'}{n_j}, & k \in \mathcal{J} \\ \pi_k^*, & k \notin \mathcal{J} \end{cases} .$$

Result B.2 *If all $\pi_k^* \leq 1$, where π_k^* are defined as in Result B.1, then $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_N^*)$ is also the solution to the constrained optimization problem*

$$\begin{aligned} &\min_{\boldsymbol{\pi}} g(\boldsymbol{\pi}) = \sum_k \frac{c_k}{\pi_k} \ , \\ &\textit{subject to} \quad \begin{matrix} \pi_k \in (0,1] \\ \sum_k \pi_k = n \end{matrix} \ , \end{aligned} \tag{B.5}$$

If $\pi_k^* > 1$ for some elements, let $\mathcal{I}$ be the index set consisting of all m elements with $\pi_k^* > 1$. Update the inclusion probabilities according to

$$\begin{cases} \tilde{\pi}_k^* = 1 & k \in \mathcal{I} \\ \tilde{\pi}_k^* = \frac{n-m}{\sum_{i \notin \mathcal{I}} \sqrt{c_i}} \sqrt{c_k} & k \notin \mathcal{I} \end{cases} . \tag{B.6}$$

If this again result in some $\tilde{\pi}_k^* > 1$, add those to the index set $\mathcal{I}$, increase m accordingly and update the inclusion probabilities as in (B.6). This procedure is iterated until a feasible solution is found with $\tilde{\pi}_k^* \in (0,1]$ for all $k \in S_1$. The procedure will stop in a finite number of iterations since $n \leq N$ and S_1 contains a finite number of elements. The solution obtained in the final iteration is the solution to the constrained optimization problem (B.5).

Proof of Result B.2 *The claim is trivially true if all $\pi_k^* \in (0,1]$ according to Result B.1. Suppose that $\pi_k^* > 1$ for some k and that $\tilde{\pi}_k^* \in (0,1]$ are found according to the procedure described in Result B.2. Consider a partition of $\tilde{\boldsymbol{\pi}}^* = (\tilde{\pi}_1^*, \ldots, \tilde{\pi}_N^*)$ into two components $\tilde{\boldsymbol{\pi}}_{(1)}^*$ and $\tilde{\boldsymbol{\pi}}_{(2)}^*$, where the first component contains all $\tilde{\pi}_k^* < 1$ and the second component contains all $\tilde{\pi}_k^* = 1$. In other words, the first component correspond to elements $k \notin \mathcal{I}$ and the second component to elements in $k \in \mathcal{I}$. We show that all points $\boldsymbol{\pi}$ in an arbitrary small neighborhood of $\tilde{\boldsymbol{\pi}}^*$ within the feasible region has a larger value of the objective function than $\tilde{\boldsymbol{\pi}}^*$. In particular, we show that any valid change in the second component followed by optimal selection of inclusion probabilities in the first component results in larger value of the objective function than when evaluated at $\tilde{\boldsymbol{\pi}}^*$. This means that $\tilde{\boldsymbol{\pi}}^*$ is a local minimum and by convexity also a global minimum.*

We consider first changes in a single coordinate of $\tilde{\boldsymbol{\pi}}_{(2)}^*$ keeping the others fixed. Say that the i-th coordinate of $\tilde{\boldsymbol{\pi}}_{(2)}^*$ is decreased by some $\epsilon > 0$, i.e. let $\tilde{\pi}_{(2),i}^* = 1 - \epsilon$. Let ϵ be so small that $\frac{n_{(1)}+\epsilon}{n_{(1)}} \tilde{\pi}_k^* < 1$ for all $k \notin \mathcal{I}$, where $n_{(1)} = \sum_{k \notin \mathcal{I}} \tilde{\pi}_k^*$ is the expected number of sampled elements from the first component. We then think of $\tilde{\boldsymbol{\pi}}_{(2)}^*$ as fixed and minimize the objective function as a function of $\boldsymbol{\pi}_{(1)}$, keeping the total expected sample size fixed to be n. According to Remark B.1, this correspond to increasing all $\tilde{\pi}_{(1),j}^*$ in the first component of $\tilde{\boldsymbol{\pi}}^*$ by a factor $\frac{n_{(1)}+\epsilon}{n_{(1)}}$. Denote the new point by $\dot{\boldsymbol{\pi}}^*$.

The difference in the objective function evaluated at $\dot{\boldsymbol{\pi}}^*$ and $\tilde{\boldsymbol{\pi}}^*$ is

$$\begin{aligned} g(\dot{\boldsymbol{\pi}}^*) - g(\tilde{\boldsymbol{\pi}}^*) &= \sum_k \frac{c_k}{\dot{\pi}_k^*} - \sum_k \frac{c_k}{\tilde{\pi}_k^*} \\ &= \sum_{k:\dot{\pi}_k^*<1} \frac{c_k}{\dot{\pi}_k^*} + \sum_{k:\dot{\pi}_k^*=1} c_k - \sum_{k:\tilde{\pi}_k^*<1} \frac{c_k}{\tilde{\pi}_k^*} - \sum_{k:\tilde{\pi}_k^*=1} c_k \\ &= \sum_{k \notin \mathcal{I}} \frac{c_k}{\dot{\pi}_k^*} + \frac{\epsilon}{1-\epsilon} c_{(2),i} - \sum_{k \notin \mathcal{I}} \frac{c_k}{\tilde{\pi}_k^*} , \end{aligned} \tag{B.7}$$

since the two sums over elements with $\dot{\pi}_k^* = 1$ and $\tilde{\pi}_k^* = 1$ cancel out expect for the term $\frac{\epsilon}{1-\epsilon} c_{(2),i}$.

The difference between the two sums over elements $k \notin \mathcal{I}$ in (B.7) can be written as

$$\sum_{k\notin\mathcal{I}} c_k \left(\frac{1}{\dot{\pi}_k^*} - \frac{1}{\tilde{\pi}_k^*}\right) = \sum_{k\notin\mathcal{I}} c_k \left(\frac{n_{(1)}}{\tilde{\pi}_k^*(n_{(1)}+\epsilon)} - \frac{1}{\tilde{\pi}_k^*}\right) = -\sum_{k\notin\mathcal{I}} \frac{c_k}{\tilde{\pi}_k^*}\frac{\epsilon}{n_{(1)}+\epsilon} \ .$$

Note that $\frac{c_k}{\tilde{\pi}_k^{*2}} = \frac{c_j}{\tilde{\pi}_j^{*2}}$ for all j,k $\notin \mathcal{I}$ according to Formula (B.4). It follows that

$$\sum_{k\notin\mathcal{I}} \frac{c_k}{\tilde{\pi}_k^*}\frac{\epsilon}{n_{(1)}+\epsilon} = \frac{c_j}{\tilde{\pi}_j^{*2}}\frac{\epsilon}{n_{(1)}+\epsilon}\sum_{k\notin\mathcal{I}}\tilde{\pi}_k^* = \frac{c_j}{\tilde{\pi}_j^{*2}}\frac{n_{(1)}\epsilon}{n_{(1)}+\epsilon} \ , \tag{B.8}$$

where j is any element not in $\mathcal{I}$.

Note that we initially had $\frac{c_k}{\pi_k^{*2}} = \frac{c_l}{\pi_l^{*2}}$ for all k,l $\in S_1$ in the first optimization step, according to Result B.1. Then, by increasing π_k^* for $k \notin \mathcal{I}$ and decreasing π_l^* for $l \in \mathcal{I}$ when calculating $\tilde{\boldsymbol{\pi}}^*$, we get $\frac{c_k}{\tilde{\pi}_k^{*2}} < \frac{c_l}{\tilde{\pi}_l^{*2}} \leq c_l$ for all $k \notin \mathcal{I}, l \in \mathcal{I}$. This in turn implies that $c_k < c_{(2),i}\tilde{\pi}_k^{*2}$ for all $k \notin \mathcal{I}$. Combining this with (B.7) and (B.8), we now we have that

$$\begin{aligned} g(\dot{\boldsymbol{\pi}}^*) - g(\tilde{\boldsymbol{\pi}}^*) &= \frac{\epsilon}{1-\epsilon}c_{(2),i} - \frac{c_j}{\tilde{\pi}_j^{*2}}\frac{n_{(1)}\epsilon}{n_{(1)}+\epsilon} \\ &= \frac{c_{(2),i}\epsilon(n_{(1)}+\epsilon)\tilde{\pi}_j^{*2} - c_j n_{(1)}\epsilon(1-\epsilon)}{\textit{positive constant}} > \epsilon n_{(1)}\frac{c_{(2),i}\tilde{\pi}_j^{*2} - c_j}{\textit{positive constant}} > 0 \ . \end{aligned}$$

This shows that an arbitrary small decrease in $\tilde{\boldsymbol{\pi}}_{(2)}^*$ along a single coordinate result in larger values of the objective function.

It remains to show that the same holds when decreasing $\tilde{\boldsymbol{\pi}}_{(2)}$ along any direction $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_m)$. Doing so, the expected sample size in the second component is reduced by $\sum_{k=1}^m \epsilon_k$ and the expected sample size in the first component consequently increased by the same amount. Let $\boldsymbol{\epsilon}$ be so small that $\frac{n_{(1)}+\sum_{k=1}^m \epsilon_k}{n_{(1)}}\tilde{\pi}_k^* < 1$ for all $k \notin \mathcal{I}$. According to Remark B.1, the optimal allocation of the probability mass $\sum_{k=1}^m \epsilon_k$ added to the first component is to increase all $\tilde{\pi}_{(1),j}^*$ in the first component by a factor $\frac{n_{(1)}+\sum_{k=1}^m \epsilon_k}{n_{(1)}}$. We write the new point as

$$\begin{aligned} \dot{\pi}_{(1),j}^* &= \tilde{\pi}_{(1),j}^* \frac{n_{(1)} + \sum_{k=1}^m \epsilon_k}{n_{(1)}} \\ &= \tilde{\pi}_{(1),j}^* \frac{n_{(1)}+\epsilon_1}{n_{(1)}} \frac{n_{(1)}+\epsilon_1+\epsilon_2}{n_{(1)}+\epsilon_1} \cdots \frac{n_{(1)}+\epsilon_1+\epsilon_2+\ldots+\epsilon_m}{n_{(1)}+\epsilon_1+\epsilon_2+\ldots+\epsilon_{m-1}} \ . \end{aligned}$$

$\dot{\pi}_{(1),j}^*$ can thus be found sequentially by first increasing $\tilde{\pi}_{(1),j}^*$ by a factor $\frac{n_{(1)}+\epsilon_1}{n_{(1)}}$ and then by a factor $\frac{n_{(1)}+\epsilon_1+\epsilon_2}{n_{(1)}+\epsilon_1}$ and so on. This sequential update of $\dot{\pi}_{(1),j}^*$ is the same as when decreasing $\tilde{\boldsymbol{\pi}}_{(2)}^*$ along one coordinate at a time, and we have already shown that such changes result in increased values of the objective function. This proves the claim.

Bibliography

[1] http://www.eurofot-ip.eu/en/about_eurofot. [Online; accessed 17-March-2016].

[2] https://cran.r-project.org/web/packages/survey/index.html. [Online; accessed 12-May-2016].

[3] http://search.r-project.org/library/sampling/html/UPmaxentropy.html. [Online; accessed 12-May-2016].

[4] https://cran.r-project.org/web/packages/alabama/index.html. [Online; accessed 12-May-2016].

[5] T. Asparouhov. Sampling weights in latent variable modeling. *Structural Equation Modeling*, 12(3):411–434, 2005.

[6] A.C. Atkinson and A.N. Donev. *Optimum Experimental Designs*. Oxford Statisticl Science Series. Oxford University Press, 1992.

[7] D.A. Binder. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review / Revue Internationale de Statistique*, 51(3):279–292, 1983.

[8] J.U. Breckling, R.L. Chambers, A.H. Dorfman, et al. Maximum likelihood inference from sample survey data. *International Statistical Review / Revue Internationale de Statistique*, 62(3):349–363, 1994.

[9] G. Casella and R.L. Berger. *Statistical Inference*. Duxbury advanced series. Duxbury Thomson Learning, 2008.

[10] R.L. Chambers and C.J. Skinner, editors. *Analysis of Survey Data*. Wiley Series in Survey Methodology. Wiley, 2003.

[11] R.L. Chambers, D.G. Steel, S. Wang, and A. Welsh. *Maximum Likelihood Estimation for Sample Surveys*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 2012.

[12] L.E. Chambless and K.E. Boyle. Maximum likelihood methods for complex sample data: logistic regression and discrete proportional hazards models. *Communications in Statistics - Theory and Methods*, 14(6):1377–1392, 1985.

[13] X.H. Chen, A.P. Dempster, and J.S. Liu. Weighted Finite Population Sampling to Maximize Entropy. *Biometrika*, 81(3), 1994.

[14] B. Efron and D.V. Hinkley. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika*, 65(3):457–482, 1978.

[15] C.E. Frangakis and S.G. Baker. Compliance subsampling designs for comparative research: Estimation and optimal planning. *Biometrics*, 57(3):899–908, 2001.

[16] W.A. Fuller. Least-squares and related analyses for complex survey designs. *Survey Methodology*, 10:97–118, 1984.

[17] W.A. Fuller. *Sampling Statistics*. Wiley Series in Survey Methodology. Wiley, 2011.

[18] V.P. Godambe. A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 17(2):269–278, 1955.

[19] V.P. Godambe and M.E. Thompson. Parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Review / Revue Internationale de Statistique*, 54(2):127–138, 1986.

[20] J. Godfrey, A. Roshwalb, and R.L. Wright. Model-based stratification in inventory cost estimation. *Journal of Business & Economic Statistics*, 2(1):1–9, 1984.

[21] H.O. Hartley and R.L. Sielken. A "super-population viewpoint" for finite population sampling. *Biometrics*, 31(2):411–422, 1975.

[22] J. Hájek. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35(4):1491–1523, 1964.

[23] J. Hájek and V. Dupač. *Sampling from a finite population*. Statistics, textbooks and monographs. M. Dekker, 1981.

[24] D.G. Horvitz and D.J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685, 1952.

[25] SAS Institute. *SAS/STAT 9.3 User's Guide: Survey Data Analysis (Book Excerpt)*. ITPro collection. SAS Institute, 2011.

[26] C.T. Isaki and W.A. Fuller. Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377):89–96, 1982.

[27] J.H. Jinn, J. Sedransk, and P. Smith. Optimal two-phase stratified sampling for estimation of the age composition of a fish population. *Biometrics*, 43(2):343–353, 1987.

[28] J.K. Kim and C.J. Skinner. Weighting in survey analysis under informative sampling. *Biometrika*, 100(2):385–398, 2013.

[29] P.S. Kott. A note on model-based stratification. *Journal of Business & Economic Statistics*, 3(3):284–286, 1985.

[30] A.M. Krieger and D. Pfeffermann. Maximum likelihood from complex sample surveys. *Survey Methodology*, 18:225–239, 1992.

[31] R.J.A. Little and D.B. Rubin. *Statistical Analysis With Missing Data*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley, 1987.

[32] T. Lumley. *Complex Surveys: A Guide to Analysis Using R*. Wiley Series in Survey Methodology. Wiley, 2011.

[33] L. Magee. Improving survey-weighted least squares regression. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 60(1):115–126, 1998.

[34] J. Neyman. Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33(201):101–116, 1938.

[35] T. Orchard and M.A. Woodbury. A missing information principle: theory and applications. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics*, pages 697–715, Berkeley, Calif., 1972. University of California Press.

[36] D. Pfeffermann. The role of sampling weights when modeling survey data. *International Statistical Review / Revue Internationale de Statistique*, 61(2):317–337, 1993.

[37] D. Pfeffermann, A.M. Krieger, and Y. Rinott. Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8(4):1087–1114, 1998.

[38] M. Reilly. Optimal sampling strategies for two-stage studies. *American Journal of Epidemiology*, 143(1):92–100, 1996.

[39] M. Reilly and M.S. Pepe. A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, 82(2):299–314, 1995.

[40] D.B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[41] S. Rubin-Bleuer and I.S. Kratina. On the two-phase framework for joint model and design-based inference. *The Annals of Statistics*, 33(6):2789–2810, 2005.

[42] R.J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Statistics. Wiley, 2009.

[43] J. Shao and D. Tu. *The Jackknife and Bootstrap*. Springer Series in Statistics. Springer New York, 2012.

[44] C.J. Skinner. Domain means, regression and multivariate analysis. In C.J. Skinner, D. Holt, and T.M.F. Smith, editors, *Analysis of complex surveys*, Wiley series in probability and mathematical statistics. Wiley, 1989.

[45] C.E. Särndal, B. Swensson, and J. Wretman. *Model Assisted Survey Sampling*. Springer Series in Statistics. Springer New York, 2003.

[46] Y. Tillé. *Sampling Algorithms*. Springer Series in Statistics. Springer, 2006.

[47] A.M. Walker. Anamorphic analysis: Sampling and estimation for covariate effects when both exposure and disease are known. *Biometrics*, 38(4):1025–1032, 1982.

[48] J.E. White. A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*, 115(1):119–128, 1982.