MAGISTERARBEIT

Titel der Magisterarbeit

„Parametric (lognormal) estimation of the
at-risk-of-poverty rate“

Verfasser

Johannes Klotz Bakk.

angestrebter akademischer Grad

Magister der Sozial- und Wirtschaftswissenschaften
(Mag.rer.soc.oec.)

Wien, im April 2008

Studienkennzahl lt. Studienblatt	A 066 951
Studienrichtung lt. Studienblatt	Magisterstudium Statistik
Betreuer	Ao. Univ.-Prof. Mag. Dr. Andreas Futschik

Preface

Some years ago, while I was working in the EU-SILC team at Statistik Austria, I had to calculate the reliability of at-risk-of-poverty rate estimates for federal provinces (Bundeslaender). To my surprise, the confidence intervals were quite large, indicating little reliability of the regional estimates. Since there is large demand from the public not only for estimates for federal provinces, but even for smaller regional units, I considered the large sampling error of the regional estimates as a major problem.

Since then, during my study of statistics at the University of Vienna, I have learned a lot about mathematical statistics and estimation theory. I took particular interest in estimation methods and the comparison of estimators. I also concerned myself with methods of small area estimation. From that viewpoint, I decided to use my master thesis to try to find an improved estimator for the at-risk-of-poverty rate in small samples, without calling for unrealistic large auxiliary information or giving user-unfriendly complicated formulae.

I started working on the topic in the spring of 2007 and finished in March 2008. I spent most of the time with the derivation of formulae, and also invested a lot in theoretical considerations and in simulation studies. What I can say with certainty is that never before during my studies have I learned so much about the scientific subject of statistics. A substantial improvement in my own knowledge and also a positive outcome that might be beneficial for the public compensate for the many hours spent with paper and pencil.

I would like to extend my gratitude to my supervisor, Prof. Andreas Futschik, for all the technical support he has given me. Special thanks go to my parents Anneliese and Gottfried Klotz for financial and emotional support during my studies. Without their help, writing the current thesis would not have been possible. Last but not least, I would like to thank my longtime companion Katharina for all the love and affection I receive from her from day-to-day.

Abstract

The thesis generally deals with a parametric estimator of the at-risk-of-poverty rate, assuming lognormally distributed income data. In the first chapter, state-of-the-art nonparametric estimation of that indicator is reviewed and the motivation for developing an alternative estimator is explained. Chapter 2 deals with parametric modelling of income distributions as well as the basic properties and main advantages of the lognormal distribution. In the following three chapters formulae for parametric point and interval estimators and their asymptotic properties are derived, and in chapter 6 the applicability of the large-sample results in the small-sample case are examined in a Monte Carlo study. Simulation is also used in the subsequent chapter to compare the performance of the parametric with the nonparametric estimator applied to real life datasets. The last two chapters in the thesis deal with variance inflation caused by estimating for persons in households as well as by complex sampling designs and nonresponse.

The main finding of the thesis is that usage of the parametric estimator might substantially improve the accuracy of estimates if the sample size is small, whereas it is not recommended if the sample size is large. Concerning medium sample sizes, the performance of the parametric compared to the nonparametric estimator depends on the deviation from the theoretical lognormal assumption in the actual dataset. We investigate two empirical distributions, indicating a better performance of the parametric estimator up to a sample size of about 150 observations in the first dataset and up to a sample size of about 2,000 observations in the second dataset.

Abstract in German

Die vorliegende Magisterarbeit behandelt ein parametrisches Verfahren zur statistischen Schaetzung von Armutsgefaehrdungsquoten. Im ersten Kapitel wird der aktuelle Stand der nichtparametrischen Schaetzung wiederholt und die Motivation zur Entwicklung eines alternativen Schaetzers dargelegt. Kapitel 2 behandelt das parametrische Modellieren von Einkommensverteilungen sowie die grundlegenden Eigenschaften und Vorzuege der Lognormalverteilung. In den folgenden drei Kapiteln werden parametrische Punkt- und Intervallschaetzer und deren asymptotische Eigenschaften hergeleitet, worauf in Kapitel 6 mittels einer Monte-Carlo-Studie die Anwendbarkeit der asymptotischen Ergebnisse in endlichen Stichproben geprueft wird. Eine Simulationsstudie wird auch im darauffolgenden Kapitel verwendet, um die Effizienz des parametrischen Schaetzers im Vergleich zum nichtparametrischen Schaetzer angewandt auf empirische Datensaetze zu untersuchen. In den letzten beiden Kapiteln der Magisterarbeit wird der Einfluss des Schaetzens fuer Personen in Haushalten einerseits sowie von komplexen Stichprobenverfahren und Antwortausfall andererseits auf die Intervallschaetzer behandelt.

Das wichtigste Ergebnis der Arbeit lautet, dass die Verwendung des parametrischen Schaetzers die Genauigkeit von Schaetzwerten bei kleiner Stichprobengroesse erheblich verbessern kann, waehrend sie fuer grosse Stichprobengroessen nicht empfohlen wird. Betreffend Stichproben mittlerer Groesse haengt die relative Effizienz des parametrischen Schaetzers vom Grad der Abweichung von der theoretischen Lognormalverteilungsannahme im konkreten Datensatz ab. Bei der Untersuchung zweier empirischer Einkommensverteilungen zeigt sich ein geringerer mittlerer quadratischer Fehler des parametrischen Schaetzers im ersten Datensatz bis zu einer Stichprobengroesse von ca. 150 Beobachtungen und im zweiten Datensatz bis zu einer Stichprobengroesse von ca. 2.000 Beobachtungen.

Contents

1 Introduction

1.1 Definition of the indicator

In scientific research on relative poverty and social exclusion in developed countries, an important indicator for the prevalence of relative poverty is the proportion of elements below a given fraction of a quantile of the income distribution. For instance, EUROSTAT defines a low wage as one below 60% of the national median monthly wage (Berger and Skinner, 2003). The most frequent quantile used in this context is the median, since it is a measure of central tendency, and a deviation from the median therefore indicates deviation from the "average" or "relative standard". As Preston (1995) mentions, "Statistics on proportions of populations falling below given fractions of average income have come to play a central role in the discussion of poverty". Till (2006) points out that in social research disposable income is taken as a measure for the expected standard of living. The proportion of persons living in households below 60% of the national median annual disposable household income is one of the "Laeken indicators", i.e. common statistical indicators for social inclusion in the European Union (EUROSTAT, 2003).

It should be noted at this point that according to different requirements both "elements" and "income" may have different meanings. Elements may be persons, households, families, employment relationships, or persons in households (if income is measured at household level, but the rate is estimated for persons). Income may refer to gross income, net income, equivalised disposable income, income from certain sources only (e.g. wage income), or even expenditures. In this thesis we shall generally refer to households or persons in households and equivalised disposable annual household income, as it is the case in estimating the Laeken indicators (EUROSTAT, 2003). Furthermore, we will restrict our results to the case when the quantile is the median.

In accordance to EUROSTAT, the indicator itself will be called the "at-risk-of-poverty rate". Note that this terminology is not unique, as for instance Berger and Skinner (2003) use "low-income proportion" instead. Furthermore it should be mentioned that the at-risk-of-poverty rate is not a "rate" as it is usually used in statistics, but a proportion (Elandt-Johnson and Johnson, 1980, ch. 2). The fraction of the quantile will be called the "poverty threshold" or the "poverty line".

In a formal way, the at-risk-of-poverty rate is defined as follows. Let Y denote income and let G denote its distribution function, i.e.

$$G(y) = \Pr\{Y \leq y\}$$

The counterpart of the distribution function is the quantile function,

$$G^{-1}(p) = \inf\{y : G(y) > p\}$$

Then the at-risk-of-poverty rate θ for a certain fraction c of a certain quantile p is given by

$$\theta_{c,p} = G\left[c * G^{-1}(p)\right] \tag{1.1}$$

Particularly, if the quantile is the median and the fraction is taken as $c = 0.6$, then

$$\theta_{0.6,0.5} = G\left[0.6 * G^{-1}(0.5)\right] = G\left[0.6 * Med(Y)\right]$$

We see that the (true) at-risk-of-poverty rate is a function of the income distribution function. If we know the actual distribution function for the target universe, we are able to calculate the indicator. However, the true distribution function is usually unknown and has to be estimated from sample data. There are generally two possible ways of estimating the indicator: nonparametric and parametric. Nonparametric means "to estimate the population proportion below the population poverty line by the sample proportion below the sample poverty line" (Preston, 1995). Parametric means to first assume a certain distribution model, express the indicator as a function of the model parameters (Graf, 2007), and then to estimate the indicator via the sample parameter estimates.

Table 1.1: At-risk-of-poverty rate estimates for EU member states, 2003

Country	At-risk-of-poverty rate
Belgium	15
Denmark	12
Germany	15
Ireland	21
Greece	21
Spain	19
France	12
Italy	...
Luxembourg	10
Netherlands	12
Austria	13
Portugal	19
Finland	11
Sweden	...
United Kingdom	18

1.2 Empirical estimates

EUROSTAT (2007, ch. 4) reports nonparametric estimates for the at-risk-of-poverty rate for the 15 member states of the European Union in 2003 (table 1.1). The poverty line was set at 60% of the national median income, with "income" referring to annual equivalised disposable household income after social transfers. Values are given as percentages of persons in households.

We note that all estimates fall in the range between 10% and 21%. Relative poverty is least frequent in Luxembourg and highest in Greece and in Ireland. High rates can generally be seen in Southern European and Anglo-Saxon societies.

Estimates are calculated not only for the entire country but also for many subpopulations such as various household types, males and females, different age groups, and so forth. The poverty line is in these cases kept constant at 60% of the national median income. In Austria in 2003 the poverty line was estimated to be EUR 9,425 equivalised annual income. For domains with a very low income level this implies that the rate can

Table 1.2: At-risk-of-poverty rate estimates for various domains of study, Austria 2003

Domain	At-risk-of-poverty rate
Males	12
Females	14
0-19 years	15
20-39 years	13
40-64 years	11
65+ years	16
Austrian citizens	12
Non-Austrian citizens	27
Basic education	20
Vocational education	10
Secondary school	10
University degree	7

exceed 50%. Statistik Austria (2005) reports estimates for various domains of study in Austria in 2003 (table 1.2).

We see that relative poverty is over-proportional among females, non-nationals, children and retirees, and people with only basic education.

1.3 Nonparametric vs. parametric estimation

As already mentioned, sample estimates of the at-risk-of-poverty rate can be calculated either nonparametrically or parametrically. Nonparametric estimation - to estimate the indicator via the sample quantiles - is the usual procedure, for instance used by EUROSTAT.

Formally the nonparametric estimator with respect to c and p is given by

$$\hat{\theta}_{c,p} = \hat{G}(c * \hat{G}^{-1}(p)) \tag{1.2}$$

One advantage of (1.2) is that the procedure is intuitively clear. Furthermore, it does not rely on a distributional assumption concerning income, so it cannot be affected by any violation of a distributional assumption.

Preston (1995) derives exact small-sample and large-sample results for (1.2), with some interesting findings: The estimator is asymptotically unbiased, and an explicit formula for its sampling variance is available. The fact that the true poverty line is unknown and has to be estimated too does **not** in general increase the sampling variance of (1.2), since the two sampling errors tend to offset each other unless c is very small. The sampling distribution of (1.2) in finite samples is somewhat complicated, including the incomplete beta function.

One disadvantage of Preston's paper is that the results are restricted to simple random sampling from an infinite universe. The assumption of an infinite universe, though usually wrong, is less problematical, since the sampling fractions in that context are usually far below the conservative 5% critical level mentioned by Cochran (1963, ch. 2.5). However, the assumption of simple random sampling is a heavy constraint since it is rarely the case in practice and violations of that assumption may cause considerable deficiency in results (Kish and Fraenkel, 1974; Saerndal et al., 1992, ch. 2.10).

Berger and Skinner (2003) extend Preston's results to complex probability samples, such as stratified sampling or cluster sampling. They also consider nonresponse correction (based on known response probabilities) and the use of raking weights. One of their findings is that ignorance of unequal design weights and nonresponse may cause considerable underestimation of the sampling variance. Concerning variance estimation, they refer to the linearization and residual technique proposed by Deville (1999), a technique that was also used later to establish the official EUROSTAT formulae for variance estimation (EUROSTAT, 2005).

So if an unbiased nonparametric estimator is available and its variance is known, why should we discuss alternative estimators? The main reason is that nonparametric estimation usually gives poor estimates in small samples – "poor" in a sense that sampling errors and confidence intervals are too large to consider point estimates satisfactory. An intuitive explanation of that deficiency is that (1.2) relies on sample quantiles, and since the sample distribution function is a step function, these sample quantiles are very unreliable when it comes to small samples. Reliable small area comparisons (where "area" may refer to geographic areas as well as to other subdomains of the target universe) are difficult to obtain using (1.2). One possible way to deal with that is to use additional sample or nonsample information for model-based inference (Longford, 2005, part II). Another possible way is to use to observed sample data only but with an alternative estimator.

The idea behind parametric estimation is as follows. Given we find a suitable parametric model for the income distribution in the target universe. Then it should be possible to express the at-risk-of-poverty rate as a function of the model parameters (Graf, 2007). It might appear that the sampling variability of such model parameters is much smaller than the sampling variability of quantiles. Consequently the sampling accuracy of a parametric estimator can be higher than that of (1.2), given the distributional assumption holds. Less formal, if we know that the income distribution can be reasonably described by an established model, we may use this information to get an improved estimator for the at-risk-of-poverty rate.

Of course, in practice a parametric model will rarely fit the income distribution in the target universe exactly. Usually we will find some ranges of the distribution where the parametric fit is suboptimal. This implies that in general a parametric estimator is not unbiased. However, if the bias is moderate, in small samples a parametric estimator might still perform better in terms of the mean square error than the nonparametric estimator. This is because the decrease in sampling variance by parametric estimation may overcompensate the corresponding introduction of bias, as long as the sample size is modest. On the contrary, if the sample size becomes large and the differences in sampling variance between parametric and nonparametric estimator therefore become negligible, the nonparametric estimator might be preferred.

2 The lognormal distribution as a model for income data

2.1 Candidate distributions for income data

In 1897 economist Vilfredo Pareto stated his famous power law of income and wealth distributions (Dragulescu and Yakovenko, 2000). Since then, a large number of theoretical distributions have been proposed as models for income data. Some of them were created specifically for that purpose, others were already established distributions applied to income data.

The fact that a large number of theoretical distributions have been proposed suggests an important fact: There is no such thing as a unique distribution widely accepted as a suitable model for income data. Scientific research on the topic is far from agreement: "No unified theory of income distribution actually exists", as Atkinson and Bourguignon (2000) point out.

A large review of the topic is given by Kakwani (1980, ch. 1 and 2). He mentiones that theories concerning the shape of the income distribution have emerged from two major schools. The socio-economic school tries to explain the shape of the income distribution in a society by means of economic and institutional factors (such as sex, age, occupation, or the distribution of wealth). On the contrary, the statistical school tries to explain income distribution via stochastic processes. We shall not track that distinction further but concentrate on the resulting distributions proposed.

Kakwani (1980, ch. 2) classifies theoretical income distributions along whether or not they satisfy a weak Pareto power law (that is, a power law for higher incomes). Among the distributions satisfying the weak Pareto law we find the Burr distribution, the Fisk (or log-logistic) distribution, the Pareto distribution of first and second kind, and Champernowne's distribution. As examples of distributions not satisfying the weak Pareto law the lognormal distribution and the Gamma distribution are given.

Based on empirical data on household income in different countries Clementi and Gallegati (2005) suggest the use of the lognormal distribution for the lower and middle range of income and the Pareto distribution for the highest 1-3%. They hypothesize that the different fit for the highest income is caused by the underlying income sources: Whereas income in the lower and middle range is usually labor income, the highest incomes are typically capital incomes. They also mention a new distribution that should be applicable to the entire range of income.

The concept of a combined fit is also used for personal income by Yakovenko and Silva (2005), who propose an exponential distribution for lower und middle incomes, and again the Pareto distribution for the rich. Furthermore, as household income is the sum of personal incomes, they suggest the Gamma distribution as the corresponding model for household income in lower and middle ranges. The exponential distribution as a model for personal income data is also described by Dragulescu and Yakovenko (2001).

Much information about the lognormal distribution as a model for income data can be found in the standard monograph by Aitchison and Brown (1957, particularly ch. 11). Graf (2007) reports about the lognormal and the Fisk distribution as parametric models especially in the context of poverty indicators.

In this thesis we shall use the lognormal distribution as a model for income distribution. There are several reasons for doing so. Empirical evidence shows that the lognormal curve is an appropriate model for describing the actual distribution in a wide range, particularly in the middle part. Deviations primarily occur at the margins, though there seems to be also an asymmetric deviation, since “it tends to overcorrect for the positive skewness” (Kakwani, 1980, p. 28). The use of the lognormal distribution as a model for income data can also be theoretically justified by Girat's law of proportionate effects (Aitchison and Brown, 1957, ch. 1.2 and 3.3). However, the most important reason for choosing the lognormal distribution is that it is strongly related to the normal distribution, which greatly simplifies all procedures concerning statistical estimation and inference (Kakwani, 1980, ch. 2)

2.2 Properties of lognormal distributions

All formulae in this section – partly modified – are given by Aitchison and Brown (1957, ch. 2.2 and 2.3).

A random variable follows a lognormal distribution if and only if the logarithm of the random variable follows a normal distribution. Therefore the lognormal distribution is applicable for continuous random variables with positive values.

As the underlying normal distribution, the lognormal distribution has two parameters. Let us denote the random variable of interest by Y and its logarithm by X. We write

$$Y \sim LN(\mu, \sigma^2) \Leftrightarrow X = \log(Y) \sim N(\mu, \sigma^2)$$

We see that the parameters of the lognormal distribution are equal to those of the underlying normal. These two parameters, μ and σ^2, fully describe the shape of the lognormal curve. Concerning X, μ refers to the expectation and σ^2 to the variance. Concerning Y, μ and σ^2 are not expectation and variance, but can somewhat be interpreted as a location parameter and a scale parameter, respectively.

It should be noted at this point that the two-parameter distribution without truncation or censorship is the simplest type of a lognormal distribution. More complicated types are handled by Aitchison and Brown (1957, ch. 2 and 9) and by Thompson (1951). We shall not track them further.

For positive values of Y the probability density function is given by

$$g(y) = \frac{1}{\sqrt{2\pi\sigma^2}y} \exp\left(-\frac{1}{2\sigma^2}(\log(y) - \mu)^2\right)$$

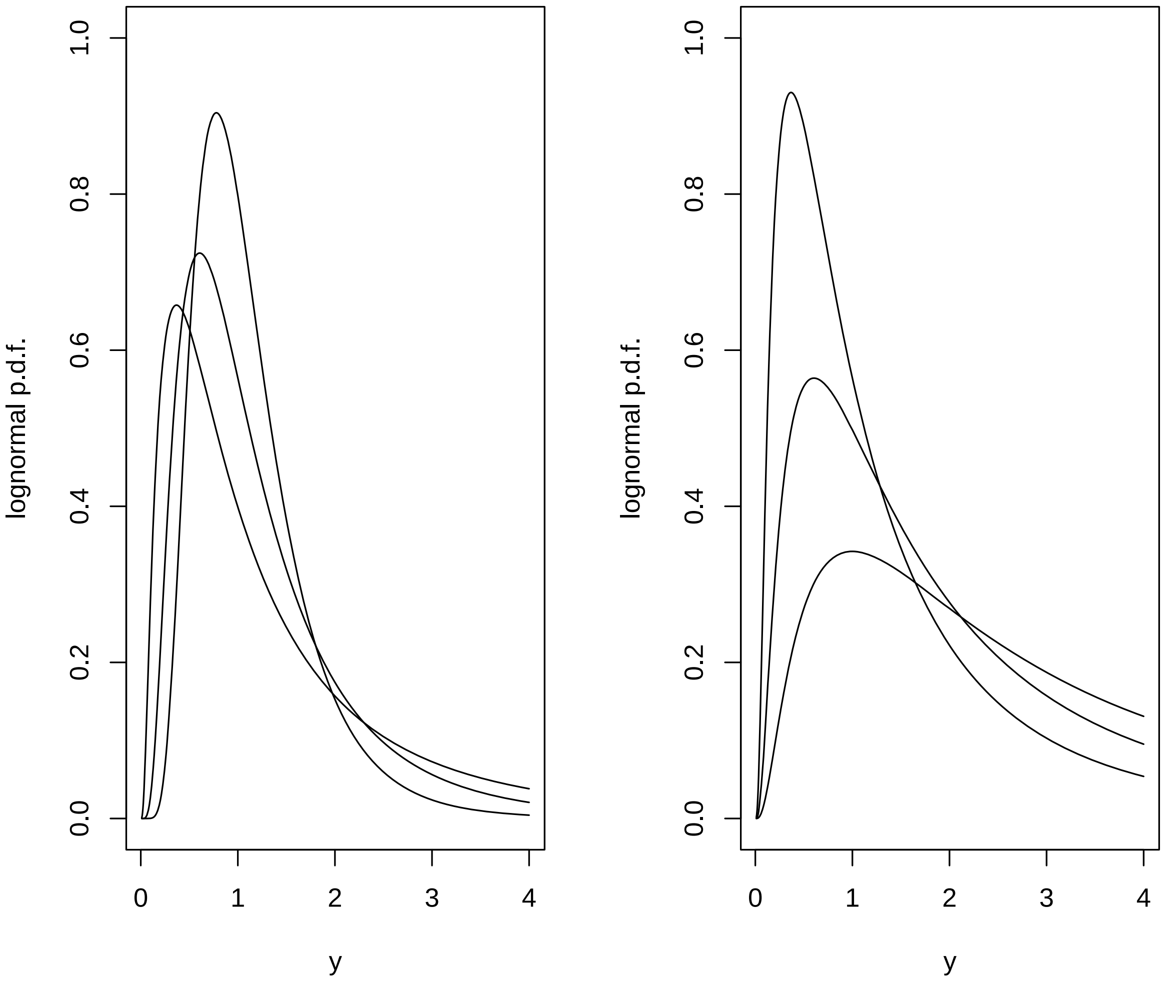
For all other values, the density is zero.

Figure 2.1 illustrates some theoretical lognormal curves. In the left panel, $\mu = 0$ is fixed and $\sigma^2 = 0.25, 0.5, 1$ varies. The magnitude reaches its highest level for $\sigma^2 = 0.25$. In the right panel, $\mu = 0, 0.5, 1$ varies and $\sigma^2 = 0.5$ is fixed. The magnitude reaches its highest level for $\mu = 0$.

The expectation of Y is

$$E(Y) = \exp(\mu + \sigma^2/2)$$

Figure 2.1: Some lognormal curves



and its variance is

$$Var(Y) = \exp\left(2\mu + \sigma^2\right)\left(\exp(\sigma^2) - 1\right)$$

The coefficient of variation is

$$CV(Y) = \frac{\sqrt{Var(Y)}}{E(Y)} = \sqrt{\exp(\sigma^2) - 1}$$

and does therefore not depend on μ.

The coefficient of skewness is

$$Skew(Y) = \frac{E(Y - E(Y))^3}{(Var(Y))^{\frac{3}{2}}} = \exp(\sigma^2 + 2)CV(Y)$$

that is, the lognormal distribution is right-skewed.

Concerning the quantiles of the lognormal curve, the following important relationship holds: If z_p is the p-quantile of the standard normal, that is,

$$z_p = \Phi^{-1}(p)$$

then the p-quantile of the lognormal distribution equals

$$G^{-1}(p) = \exp(\mu + z_p\sigma)$$

Particularly, the median of Y, which for lognormal distributions equals the geometric mean, is given by

$$Med(Y) = G^{-1}(0.5) = \exp(\mu) \tag{2.1}$$

2.3 Agreement of the model with real data

As it was pointed out in section 2.1, the lognormal model may be more or less appropriate to describe the shape of an empirical income distribution. Kakwani (1980, ch. 2) suggests that a poor fit might be observed towards the tails, Clementi and Gallegati (2005) report deviations primarily in the upper tail.

We fit a lognormal curve for two datasets, namely the SILC data and the ADMIN data. The parameters are estimated via maximum likelihood (Limpert et al., 2005;

Aitchison and Brown, 1957, ch. 5.2). Then we graphically compare the estimated lognormal quantiles with the actual quantiles using a QQ-plot. Since a function for a normal QQ-plot is by default implemented in R, version 2.6.1., we in fact compare on the logarithmic scale.

The SILC data ($n = 4,623$) refers to equivalised annual disposable household income, collected in the Austrian EU-SILC 2003 survey (Statistik Austria, 2005). Since it is a voluntary survey, some values are imputed or partly imputed. The ADMIN data ($n = 29,644$) refers to a 1% sample of matched administrative data on Austrian employee wage incomes in 2003 (Klotz, 2005).

To eliminate extremely high or low incomes in both datasets, a top-bottom recoding was carried out as follows. All incomes greater than 10 times the median income were imputed at random by an observed value greater than 2.5 times the median income. On the opposite side incomes lower than 0.1 times the median income were replaced by an observed value lower than 0.4 times the median income. It should be noted that parameter estimation of the lognormal curve is more affected by extremely low incomes. However, the top-bottom recoding was done not only to achieve a better fit but also because there is in fact reason to believe that some of the extreme incomes (especially lower incomes in the SILC data and higher as well as lower incomes in the ADMIN data) are in fact erroneous. Top-bottom recoding changed 33 low and 2 high observations in the SILC data and 173 low and 156 high incomes in the admin data, so in both cases appeared to about 1% of the sample.

Figure 2.2 shows the QQ-plots for the lognormal fit for the SILC data (the left panel refers to the original data, the right panel to the data after top-bottom recoding). We see that a huge deviation from the normal curve on the logarithmic scale (and therefore a deviation from the lognormal curve on the original scale) can be observed especially towards the lower tail of the distribution. A smaller deviation is also found for the upper tail. Good coverage can be observed for about 70-75% of the data before and about 80-85% after top-bottom recoding.

A poor fit towards the tails is not a surprise, since this phenomenon generally observed for lognormal income data fits (Kakwani 1980, p. 28). However, what is surprising is that especially the lower tail of the distribution is badly covered, since Clementi and Gallegati (2005) suggest the opposite. Possible explanations for this are e.g. systematic unit nonresponse and/or systematic item nonresponse in the survey, furthermore data collection and measurement errors, imputation errors, and of course true deviations.

Figure 2.2: Normal QQ-plot for the logarithmic SILC data

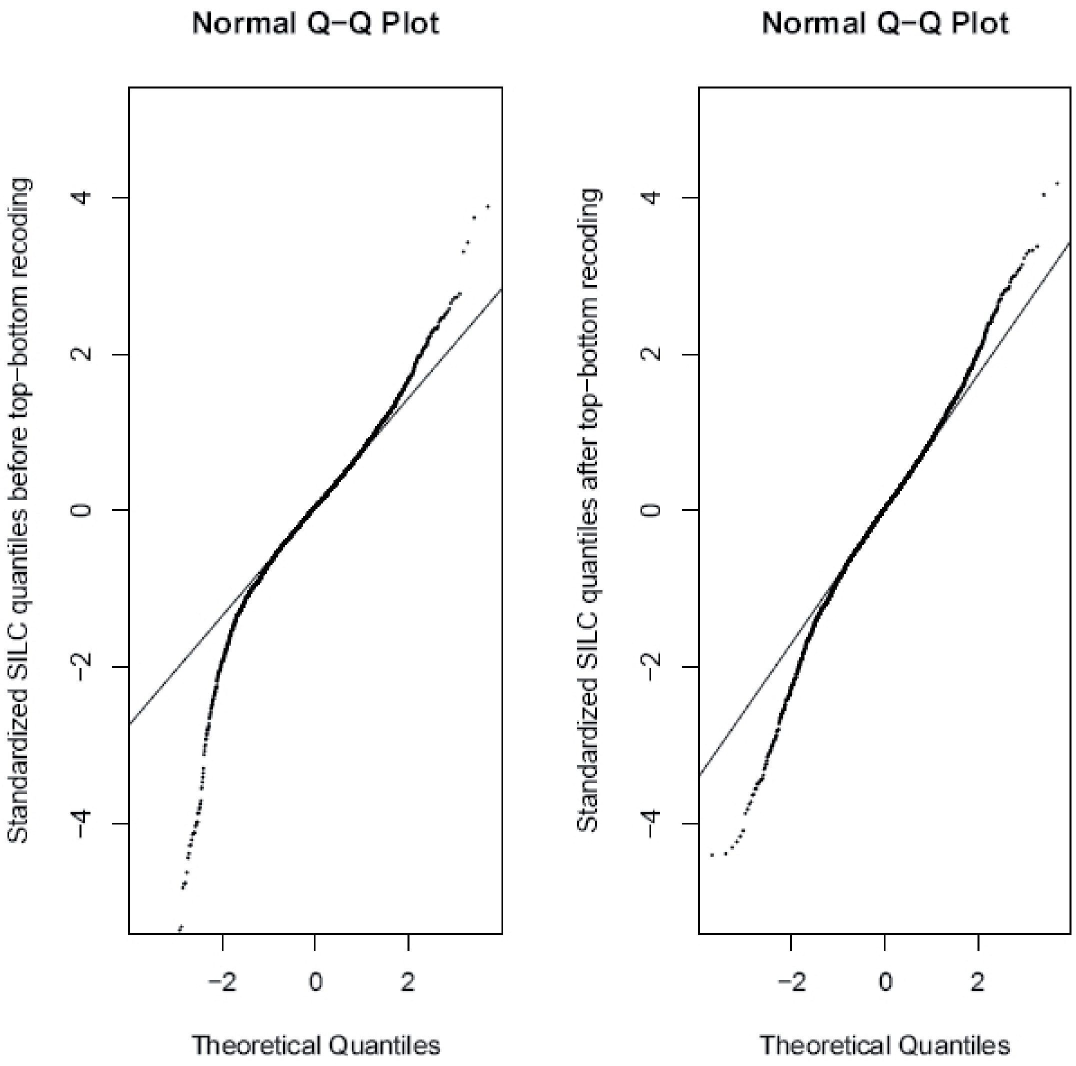


Figure 2.3: Normal QQ-plot for the logarithmic ADMIN data

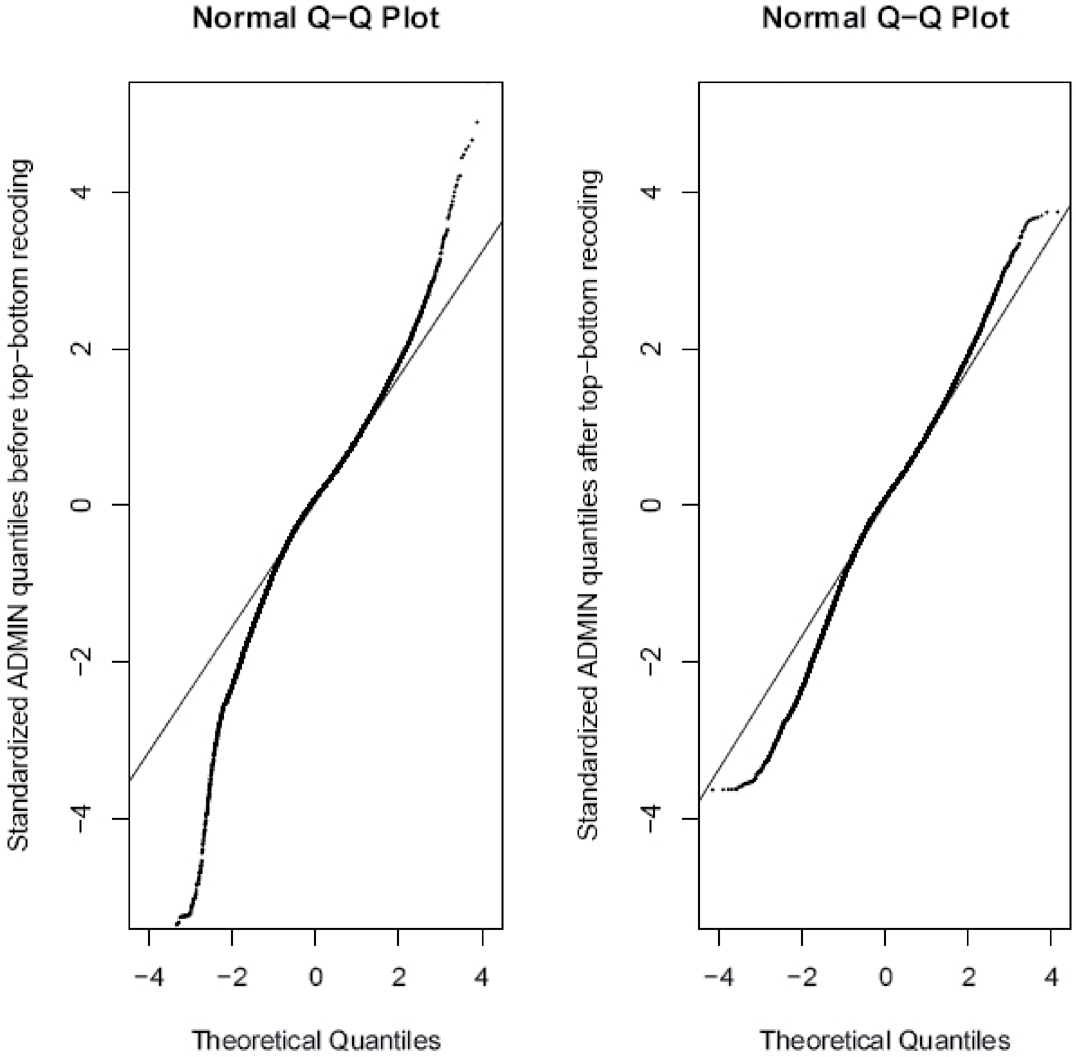


Figure 2.3 gives the QQ-plots for the ADMIN data (analogous to the SILC data). We observe a poor fit towards both tails of the distribution. Compared with the SILC the deviation is higher in the upper tail, whicht might be explained by the different data source (administrative data where nonresponse does not occur). We again see a substantial improvement in the fit by the top-bottom recoding: Before the procedure, a good fit can be observed for about 75-80% of the data, whereas this value increases to about 85-90% after top-bottom recoding (though the procedure itself appeared to only about 1% of the data).

We conclude that for both the SILC data and the ADMIN data the lognormal assumption seems justified for a large center part of the distribution, but is somewhat invalid towards the tails. More to the point, the empirical tails are “too heavy” compared with the theoretical tails. This fact should be kept in mind since the at-risk-of-poverty rate deals primarily with the lower tail of the income distribution.

3 Domain poverty threshold

In this chapter we stress estimating the at-risk-of-poverty rate when the poverty threshold is defined specifically for the domain of interest. That is, if we are interested in the at-risk-of-poverty rate of single person households, then the poverty threshold is set as 60% times the median income of single person households, or, more generally, a factor c times the median income of single person households.

Usually the factor c is chosen so that $0 < c < 1$. Common values, besides 0.6, are 0.5, 0.4, and 0.7 (Statistik Austria, 2005, p. 120). A value $c \leq 0$ will not make sense if we assume that the disposable household income is a positive number (as it is frequently the case, cf. Kakwani, 1980, p. 12). Clearly, if $c = 1$, then the at-risk-of-poverty rate equals 50% - the half of all households that lies below the median. A value $c > 1$ is also imaginable, but will be rarely used in poverty research. It is worth noting that the estimation rules that will be established now work for any $c \geq 0$.

Estimation for a domain poverty threshold is important for at least three reasons: First, what might happen is that for some domains one is in fact interested the so defined at-risk-of-poverty rate. This will be essentially the case when the domain coincides with the population, but maybe also for some subpopulations like major geographic areas. Second, the formulae for a domain poverty threshold allow for a very clear and intuitive interpretation of the target indicator and the parameters of the lognormal distribution. Particularly, we will see that the at-risk-of-poverty is solely a measure of relative income inequality. And third, the derivations of the formulae in the case of a population poverty threshold is much easier if one has once understood them for domain poverty threshold.

3.1 Derivation of the true value

The at-risk-of-poverty rate, θ, is defined as the proportion of households with an income that is lower than c times the median income (cf. (1.1)). In terms of probability theory, we can think of income as a random variable. Then θ equals the probability that a realization of that random variable is lower than c times the median realization.

Formally

$$\theta = \Pr\{Y < c * Med(Y)\}$$

where Y refers to income (as a random variable) and $Med(Y)$ to its median. Note that this definition holds in general and does not depend on any distributional assumption on Y.

Let us now further assume that Y follows a lognormal distribution with parameters μ and σ^2, respectively. That is,

$$Y \sim LN(\mu, \sigma^2) \Leftrightarrow X = \log(Y) \sim N(\mu, \sigma^2)$$

Referring to (2.1) we may write

$$\theta = \Pr\{Y < c * \exp(\mu)\}$$

Applying the formulae from 2.2 the at-risk-of-poverty can be written as

$$\begin{aligned}\theta &= \Pr\{Y < c * \exp(\mu)\}\\ &= \Pr\{X < \log[c * \exp(\mu)]\}\\ &= \Pr\{X < \log(c) + \mu\}\\ &= \Pr\{X - \mu < \log(c)\}\\ &= \Pr\left\{\frac{X-\mu}{\sigma} < \frac{\log(c)}{\sigma}\right\}\\ &= \Phi\left(\frac{\log(c)}{\sigma}\right) \square\end{aligned}$$

where $\Phi(\cdot)$ refers to the distribution function of the standard normal distribution (Sahai and Ageel, 2000, p. 605 et seq.).

Note that the eventual formula

$$\theta = \Phi\left(\frac{\log(c)}{\sigma}\right) \tag{3.1}$$

gives θ as a proportion, that is, a value between 0 and 1. If we want to express θ as a percentage – as it is usually the case when publishing results – we have to multiply that proportion by 100%.

3.2 Discussion of the formula

In the previous section we have derived the formula for the true at-risk-of-poverty rate for a domain poverty threhold, provided that the lognormal assumption holds. Let us now briefly discuss the eventual formula (3.1).

First, as θ is a value of a cumulative distribution function, it lies in the range between 0 and 1 (Hogg and Craig, 1978, ch. 1.7). For estimation purposes this means that if we estimate θ by distribution function value, the resulting estimate will be an admissible value between 0 and 1.

Next note that the formula for θ does not contain the location parameter of the lognormal distribution, μ, but only the scale parameter, σ. (Strictly speaking, it contains the square root of σ^2.) This indicates that the at-risk-of-poverty rate is a measure of relative poverty, based on the extent of relative inequality in the income distribution. Berger and Skinner (2003) refer to the at-risk-of-poverty rate as the "low-income proportion".

Independence of μ can also be shown for other indicators of income inequality (Graf, 2007; Aitchison and Brown, 1957, ch. 11.5, with a special reference to the Gini coefficient, there called the Lorenz measure). Kakwani (1980, p. 85) mentions the standard deviation of log-income as a "widely used, single measure of inequality".

A practical application of the independence of μ is that if the scale parameters of two or more domains are equal, then also the at-risk-of-poverty rates are equal.

Formally,

$$\sigma_A = \sigma_B \Leftrightarrow \theta_A = \theta_B \tag{3.2}$$

For instance, consider the case that all incomes within a society double. We can express this by introducing a new random variable, $\check{Y}$, so that

$$\check{Y} = 2Y$$

Then the scale parameter for $\check{Y}$ is the same as for Y:

$$\begin{aligned} Var(\log(\check{Y})) &= Var(\log(2Y)) = Var(\log(2) + \log(Y)) \\ &= Var(\log(Y)) = Var(X) = \sigma^2 \end{aligned}$$

Table 3.1: Theoretical values of θ

	c			
σ	0.4	0.5	0.6	0.7
0.35	0.4	2.4	7.2	15.4
0.40	1.1	4.2	10.1	18.6
0.45	2.1	6.2	12.8	21.4
0.50	3.3	8.3	15.3	23.8
0.55	4.8	10.4	17.7	25.8
0.60	6.3	12.4	19.7	27.6

and so from (3.2) we know that also the at-risk-of-poverty rate is the same. In general, any multiplicative transformation $\check{Y} = bY$ does not affect the at-risk-of-poverty rate. For instance, if we measure income in Cents instead of Euros, θ remains unchanged.

Let us now discuss some special cases of c. If $c = 0$, then $\log(c) = -\infty$ and consequently $\theta = 0$. On the other hand, if $c = 1$, then $\log(c) = 0$ and so $\theta = 0.5$. For any $0 < c < 1$, $\log(c)$ is a negative value, so is $\log(c)/\sigma$, and therefore $0 < \theta < 0.5$.

For a given c an increase in σ decreases $\log(c)/\sigma$ in absolute value. So if $0 < c < 1$ the at-risk-of-poverty rate θ is the higher the higher σ is. In the (rare) case that $c > 1$ an increase in σ coincides with a decrease in θ.

Table 3.1 gives some hypothetical θ values, depending on c and σ, respectively. The values were calculated in MS EXCEL and are expressed as percentages. Given $c = 0.6$, we see that for $\sigma = 0.45$ the at-risk-of-poverty rate equals 12.8%, what is approximately the (person-weigthed) nonparametric estimate for EU-SILC-Austria in 2003 (Statistik Austria, 2005, p. 13).

3.3 Point estimation

Given (3.1), if we knew the true σ, then we could calculate the true θ. Usually σ is unknown and has to be estimated from sample data. Given an admissible estimator for σ, we may insert this estimator into (3.1) to get an admissible estimator for θ:

$$\hat{\theta} = \Phi\left(\frac{\log(c)}{\hat{\sigma}}\right) \tag{3.3}$$

Note that c is not estimated since it is not random.

The most widely used estimator for σ is the sample standard deviation, that is, the square root of the sample variance (Hogg and Craig, 1978, ch. 4.1). For simple random sampling from an infinite universe it is

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where x_i refers to the logarithmic income of the i'th household in the sample, $\bar{x}$ to the sample mean of the x_i's and n to the sample size.

The resulting estimate is then plugged into (3.3).

To evaluate an estimator we investigate its sampling distribution. Desirable properties of an estimator are unbiasedness, consistency, efficiency, and sufficiency (Bortz, 1999, ch. 3.3).

Calculating the small-sample distribution of $\hat{\theta}$ would be very laborious, since the sampling distribution of $\hat{\sigma}$ is quite complicated and $\hat{\theta}$ is a complex transformation of $\hat{\sigma}$. However, concerning the large-sample case, we can state the following important result:

Since $\hat{\sigma}^2$ is a consistent estimator for the true σ^2, and $\hat{\sigma}$ is a continuous function of $\hat{\sigma}^2$, we know from Slutsky's theorem that also $\hat{\sigma}$ is a consistent estimator of σ (Judge et al., 1988, p. 85). Correspondingly, as $\hat{\theta}$ is a continuous function of $\hat{\sigma}$ (because it refers to the cumulative distribution function of the standard normal), $\hat{\theta}$ is a consistent estimator of θ as well.

In other words, given a sufficiently large sample size, (3.3) will produce an estimate that lies arbitrarily close to the true at-risk-of-poverty rate, provided that the lognormal assumption holds. We also say that for a lognormal distribution $\hat{\theta}$ is asymptotically unbiased (ibid.):

$$E(\hat{\theta}) = \theta \quad n \to \infty$$

for lognormally distributed income.

Note that this result holds for any consistent $\hat{\sigma}$ for a given sampling design. Concerning estimation of $\hat{\sigma}$ in case of other designs than simple random sampling from an infinite universe the reader may refer to Saerndal et al. (1992, ch. 5.9). Some basic ideas are also given in chapter 9 of this thesis.

In finite samples, (3.3) is in general not unbiased. A Monte Carlo study on the amount of bias in finite sample is presented in chapter 6, indicating that for a lognormal distribution the bias is modest, even for small samples. This implies that the large-sample results might serve as good approximations for the small-sample case. Given the lognormal assumption holds, (3.3) will produce an estimate that on average lies close to the true value. Of course, if the target income distribution is far from lognormal, (3.3) in general is biased, no matter how large the sample size is. A study on the bias of (3.3) in real world examples is presented in chapter 7.

So far we have stressed the expectation of $\hat{\theta}$. Another important characteristic of the sampling distribution of $\hat{\theta}$, its variance, is discussed in the following section.

3.4 Interval estimation

Concerning the variance of (3.3), we again use the fact that $\hat{\theta}$ is a continuous (and twice differentiable) function of $\hat{\sigma}$. Therefore we can calculate the asymptotic variance of $\hat{\theta}$ by the δ-method. This method relies on a first order (linear) Taylor series expansion and is described in detail by Wolter (1985, ch. 6). The general formula for the asymptotic variance of a function of a statistic $\hat{\zeta}$ is

$$Var(f(\hat{\zeta})) = (f'(\zeta))^2 * Var(\hat{\zeta})$$

(ibid.), and this large-sample results is frequently found to be a useful approximation for the finite-sample case (Saerndal et al., 1992, ch. 5.5).

Applied to (3.3) we get

$$Var(\hat{\theta}) = Var(g(\hat{\sigma})) \approx (g'(\sigma))^2 * Var(\hat{\sigma})$$

The derivative of (3.3) with respect to $\hat{\sigma}$ is

$$\frac{d\hat{\theta}}{d\hat{\sigma}} = \phi\left(\frac{\log(c)}{\hat{\sigma}}\right) * \log(c) * \left(\frac{-1}{\hat{\sigma}^2}\right)$$

where $\phi(\cdot)$ refers to the ordinate of the standard normal distribution (Aitchison and Brown, 1957, p. 162).

The variance of $\hat{\sigma}$ depends on the sampling design. For simple random sampling from an infinite universe it is

$$Var(\hat{\sigma}) = \sigma^2/2n$$

(Bortz, 1999, p. 92)

So the asymptotic variance of $\hat{\theta}$, given simple random sampling from an infinite lognormal population, is

$$\begin{aligned} Var(\hat{\theta}) &\approx \left[\phi\left(\frac{\log(c)}{\sigma}\right) * \frac{\log(c)}{-\sigma^2}\right]^2 * \frac{\sigma^2}{2n} \\ &= \left[\phi\left(\frac{\log(c)}{\sigma}\right)\right]^2 * \left(\frac{\log(c)}{\sigma}\right)^2 /2n\square \end{aligned}$$

Since the exact value of σ is unknown, we replace it by the sample estimate and get an estimator of the variance of (3.3):

$$\widehat{Var}(\hat{\theta}) = \left[\phi\left(\frac{\log(c)}{\hat{\sigma}}\right)\right]^2 * \left(\frac{\log(c)}{\hat{\sigma}}\right)^2 /2n \tag{3.4}$$

The estimated standard error of $\hat{\theta}$ is the square root of (3.4), namely

$$\widehat{SE}(\hat{\theta}) = \phi\left(\frac{\log(c)}{\hat{\sigma}}\right) * \left|\frac{\log(c)}{\hat{\sigma}}\right| /\sqrt{2n} \tag{3.5}$$

Since the sampling distribution of $\hat{\theta}$ converges to a normal distribution as $n \to \infty$ (Bortz, 1999, ch. 3.2), we can calculate an asymptotic (1-α)*100% confidence interval for $\hat{\theta}$ as

$$CI_{1-\alpha}(\hat{\theta}) = \left[\hat{\theta} - z_{1-\alpha/2} * \widehat{SE}(\hat{\theta}), \hat{\theta} + z_{1-\alpha/2} * \widehat{SE}(\hat{\theta})\right]$$

where z_p refers to the p-quantile of the standard normal distribution (Sahai and Ageel, 2000, p. 607).

For $n \to \infty$ such an interval will cover the true θ except for a $\alpha * 100\%$ chance of error, given the lognormal assumption holds. It should be mentioned that for moderate deviations from the lognormal distribution and a consequently moderately biased $\hat{\theta}$, the actual coverage probability of a calculated confidence interval will still be very close to the desired (1-α)*100%, especially in small samples (Cochran, 1963, 1.7).

The behavior of the asymptotic estimator (3.5) in finite samples is investigated in chapter 6. We will see that the asymptotic standard error is a fairly good approximation for the finite sample standard error, even for modest sample sizes.

Interval estimation for sampling designs other than simple random sampling from an infinite universe is discussed in chapter 9. Since only the variance of $\hat{\sigma}$, but not the derivative of $\hat{\theta}$ with respect to $\hat{\sigma}$ is affected by the sampling design, it makes sense to use the results from the current section as a point of reference.

4 Population poverty threshold: Equal within-domain dispersion

In the previous section we stressed the case when the poverty threshold is set (and estimated) specifically for any domain of study. This is not usually the case, except perhaps in international comparisons. The more important case is when the poverty threshold is estimated at the population level and equally applied to all domains.

We use the term "population" to indicate the superdomain that is divided into the domains of study. In most practical settings this superdomain will be a country. On the contrary, we use "universe" to indicate the target population we want to make inferences about. According to circumstances we can think of a universe referring to the domain of study (e.g. all households in Vienna) as well as a universe referring to the population (e.g. all households in Austria).

In this chapter we discuss the case when the within-domain dispersion of income, measured by σ, is equal among all domains. A more general case, when the within-domain dispersion of incomes can vary among domains, is treated in the next chapter.

4.1 Derivation of the true value

The poverty threshold is set for a population denoted by U. The income in U is denoted by Y, and its logarithm by X.

Suppose that U can be fully divided into q disjoint subpopulations. These disjoint subpopulations - the domains of interest - are denoted by $U_1, U_2, \ldots, U_q$.

"Disjoint" means that

$$U_1 \cup \ldots \cup U_q = U$$

and that

$$U_j \cap U_{j'} = \emptyset \qquad j \neq j'$$

Famous examples for disjoint subpopulations are geographic area, family type (on household level), or age group (on personal level). Note that the sizes of the subpopulations (and the corresponding sample sizes) may differ substantially. Furthermore note that the sample sizes are usually unknown a priori, except for the case of stratified sampling (Cochran, 1963, ch. 5; Saerndal et al., 1992, ch. 3.7). The assumption of a full division into disjoint subpopulation can be made without any loss of generality, since it is always possible to create a residual category covering all elements that do not belong to a specified domain.

Now assume that the income distributions within the domains are lognormal distributions with different location parameters but equal dispersion parameter.

That is,

$$\begin{aligned} &Y_1 \sim LN(\mu_1, \sigma^2) \\ &\quad\vdots \\ &Y_q \sim LN(\mu_q, \sigma^2) \end{aligned}$$

Practical situations are found to agree with that assumption. For instance, Aitchison and Brown (1957, ch. 11.3) report similar variation of log-incomes among different economic sectors.

Note that on the logarithmic scale we have

$$\begin{aligned} &X_1 \sim N(\mu_1, \sigma^2) \\ &\quad\vdots \\ &X_q \sim N(\mu_q, \sigma^2) \end{aligned}$$

and that this is equivalent to the assumptions of the famous one-way analysis of variance (ANOVA) procedure (Sahai and Ageel, 2000, ch. 2.2).

What can we say about the distribution of X? Since the number of domains is finite, we have a finite mixture of q normals. Substantial information on the topic can be found in Everitt and Hand (1981, ch. 2) or Titterington et al. (1985). An important finding is that “a ’non-degenerate’ finite mixture of normal distributions cannot itself be normal” (Titterington et al., 1985, p. 39). Special reference to infinite mixtures of normals is given by Teichroew (1957) and by Barndorff-Nielsen et al. (1982).

The general formula for the density function of X is

$$f(x) = \sum_{i=1}^{j} f(x_j)\frac{N_j}{N} \tag{4.1}$$

If we think of the μ_j’s as realizations of a random variable M with expectation μ_0 and variance σ_0^2, then the following results hold:

1. $E(X) = \mu_0$
2. $Var(X) = \sigma_0^2 + \sigma^2$
3. if the number of domains, q, is large, and M is approximately normally distributed, then X follows a normal distribution.

The proof for 3. is analogous to that given in Aitchison and Brown (1957, p. 110.). The proofs for 1. and 2. are given below.

$$\begin{aligned}
E(X) &= \int_X xf(x)dx \\
&= \int_X x\left(\sum_j f(x|\mu_j)h_M(\mu_j)\right)dx \\
&= \sum_j \left(\int_X xf(x|\mu_j)dx\right)h_M(\mu_j) \\
&= \sum_j \mu_j h_M(\mu_j) \\
&= \mu_0 \quad \square
\end{aligned}$$

$$
\begin{aligned}
Var(X) &= \int_X (x-\mu_0)^2 f(x)dx \\
&= \int_X (x-\mu_0)^2 \left(\sum_j f(x|\mu_j) h_M(\mu_j) \right) dx \\
&= \sum_j \left(\int_X (x-\mu_0)^2 f(x|\mu_j) dx \right) h_M(\mu_j) \\
&= \sum_j \left(\int_X (x-\mu_j+\mu_j-\mu_0)^2 f(x|\mu_j) dx \right) h_M(\mu_j) \\
&= \sum_j (\sigma^2+\sigma_0^2) h_M(\mu_j) \\
&= (\sigma^2+\sigma_0^2) \sum_j h_M(\mu_j) \\
&= \sigma^2+\sigma_0^2 \quad \Box
\end{aligned}
$$

Note that in terms of the ANOVA model, we can regard to σ_0^2 as the between-domain variation and σ^2 the within-domain variation of log-incomes (Sahai and Ageel, 2000, ch. 2.3 and 2.12).

So if 3. is fulfilled, we have

$$Y \sim LN(\mu_0, \sigma_0^2 + \sigma^2) \tag{4.2}$$

The assumption of normally distributed domain location parameters is not unrealistic. A practical interpretation is that when the number of domains is large, we will find a majority of domains with average incomes around the overall average, and some domains with average incomes considerably higher than the general average (Statistik Austria, 2007, p. 238 et sqq.).

In practice the applicability of (4.2) will depend on the situation. For the remainder of this thesis we shall assume that (4.2) approximately holds. This will be especially the case if the between-domain variation is relatively small compared with the within-domain variation (Aitchison and Brown, 1957, p. 111).

Now we write the at-risk-of-poverty rate for domain U_j in terms of probability theory:

$$\theta_j = \Pr\{Y_j < c * Med(Y)\}$$

Analogous to the derivation in chapter 3, we get

$$\begin{aligned}
\theta_j &= \Pr\left\{Y_j < c * \exp(\mu_0)\right\} \\
&= \Pr\left\{X_j < \log\left[c * \exp(\mu_0)\right]\right\} \\
&= \Pr\left\{X_j < \log(c) + \mu_0\right\} \\
&= \Pr\left\{X_j - \mu_j < \log(c) + \mu_0 - \mu_j\right\} \\
&= \Pr\left\{\frac{X_j - \mu_j}{\sigma} < \frac{\log(c) + \mu_0 - \mu_j}{\sigma}\right\} \\
&= \Phi\left(\frac{\log(c) + \mu_0 - \mu_j}{\sigma}\right) \square
\end{aligned}$$

4.2 Discussion of the formula

For interpretational purposes, let us re-write the eventual formula

$$\theta_j = \Phi\left(\frac{\log(c) + \mu_0 - \mu_j}{\sigma}\right) \tag{4.3}$$

as

$$\theta_j = \Phi\left(\frac{\log(c) + d_j}{\sigma}\right)$$

where $d_j = \mu_0 - \mu_j$ equals the difference between the expected log-income in the entire population and the expected log-income in the study domain.

If the expected log-income in the study domain is lower than in the entire population, then $d_j > 0$ and consequently $\log(c) + d_j > \log(c)$. For practical purposes this means that the at-risk-of-poverty rate is higher than if it was calculated for the same domain but with a domain poverty threshold.

In other words, given equal dispersion within the study domains, the at-risk-of-poverty rate is the higher the lower the median income in the study domain is. This property of (4.3) does well agree with what we demand of a poverty measure.

Since d_j is a difference of expectations on the logarithmic scale, because of (2.1) it corresponds to a ratio of medians on the original scale. For instance, $d_j = 0.3$ implies that $Med(Y)/Med(Y_j) = \exp(0.3) \approx 1.35$. Note that such a ratio of medians is scale-free, so we can interpret the at-risk-of-poverty rate as a combined measure of income

Table 4.1: Theoretical values of θ_D

	$d_j = \mu_0 - \mu_j$								
σ	0.70	0.40	0.20	0.10	-	-0.10	-0.20	-0.40	-0.70
0.35	70.6	37.6	18.7	12.0	7.2	4.0	2.1	0.5	0.0
0.40	68.2	39.1	21.9	15.2	10.1	6.3	3.8	1.1	0.1
0.45	66.3	40.3	24.5	18.1	12.8	8.7	5.7	2.1	0.4
0.50	64.7	41.2	26.7	20.6	15.3	11.1	7.8	3.4	0.8
0.55	63.5	42.0	28.6	22.8	17.7	13.3	9.8	4.9	1.4
0.60	62.4	42.7	30.2	24.7	19.7	15.4	11.8	6.5	2.2

inequality and the relative income level in the study domain compared to the entire population.

For a low income domain, where μ_j is much smaller than μ_0, the at-risk-of-poverty rate can reach or even exceed 50% also if $c < 1$. If $d_j = -\log(c)$, then the indicator is always 50%, regardless of σ.

Table 4.1 gives some illustrative population values of the at-risk-of-poverty rate, dependent on the arguments d_j and σ, respectively. The factor c is constant 0.6. All values were calculated in MS EXCEL and are expressed as percentages. We see that for a given σ, already a moderate difference in the median incomes of about 10% considerably influences θ_j.

4.3 Point estimation

Let us now turn to the situation when μ_0, μ_j, and σ have to be estimated from sample data.

Again, we use the plug-in technique by inserting admissible estimators $\hat{\mu}_0$, $\hat{\mu}_j$, and $\hat{\sigma}$, respectively, into (4.3):

$$\hat{\theta}_j = \Phi\left(\frac{\log(c) + \hat{\mu}_0 - \hat{\mu}_j}{\hat{\sigma}}\right) \tag{4.4}$$

As we divide the entire population U into domains $U_1, \ldots, U_q$, let us denote the corresponding domain sample sizes by $n_1, \ldots, n_q$. If the sample is a simple random sample, then the n_j's are random numbers unknown before sampling. On the contrary, if the sample is a stratified random sample with the domains serving as strata, then the n_j's are constants. In any case,

$$n = \sum_{j=1}^{q} n_j$$

Random sample sizes increase the sampling error of a statistic (Saerndal et al., 1992, ch. 2.10). For the remainder of this chapter we shall assume that we are able either to draw a stratified random sample with proportional allocation (ibid., ch. 3.7.4), or, in case of a simple random sample, to calculate expected domain sample sizes. The latter is achieved by using auxiliary information, particularly the universe totals of the domain sizes:

$$E(n_j) = \frac{N_j}{N} \tag{4.5}$$

Auxiliary information for (4.5) may be available e.g. from census data.

We estimate μ_j by the arithmetic mean of log-income in the study domain:

$$\hat{\mu}_j = \bar{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_i \tag{4.6}$$

If income is lognormally distributed, then log-income is normally distributed and consequently the arithmetic mean is the (most) efficient estimator for the expectation (Judge et al., 1988, p. 73 et seq.). Its sampling variance is approximately

$$Var(\hat{\mu}_j) \approx \sigma^2 / E(n_j) \tag{4.7}$$

If n_j is random and its expectation is very small, then the approximation in (4.7) may be crude. However, if $E(n_j)$ is very small, it might be anyway preferable to use an alternative estimator for the domain mean. Such alternative estimators ("small-area estimators") are treated by Longford (2005, part II). We shall not track this point further.

The population mean of log-income, μ_0, is estimated by the (post-)stratified mean of the domain means, that is,

$$\hat{\mu}_0 = \left(\sum_{j=1}^{q} E(n_j) \mu_j \right) / n \tag{4.8}$$

In the case of a simple random sample with poststratification its sampling variance is

$$Var(\hat{\mu}_0) = \sigma^2/n \tag{4.9}$$

(Bortz, 1999, p. 91).

In the case of stratified sampling (4.9) usually overestimates the actual sampling variance of $\hat{\mu}_0$ (Saerndal et al., 1992, ch. 3.7), so using (4.9) we might play it safe.

If poststratification of a simple random sample is not possible, then μ_0 is estimated by the crude sample mean. In such a case the variance of $\hat{\mu}_0$ equals $(\sigma_0^2 + \sigma^2)/n$. The additional term, σ_0^2/n, can be interpreted as additional variation caused by the randomness of domain sample sizes (ibid.). We shall not track this case further since poststratification is almost always possible.

The estimator of σ is the square root of the (post-)stratified mean square within,

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{q}\frac{E(n_j)}{n_j}\sum_{i=1}^{n_j}(x_i - \hat{\mu}_j)^2} \tag{4.10}$$

and its sampling variance is approximately

$$Var(\hat{\sigma}) \approx \sigma^2/2n \tag{4.11}$$

(Bortz, 1999, p. 92).

All three estimators, (4.6), (4.8), and (4.10), are consistent. Referring to Slutsky's theorem (Judge et al., 1988, p. 85; already mentioned in 3.3), we know that also (4.3) is consistent.

4.4 Interval estimation

To estimate the asymptotic variance of (4.3), we use again the δ-method. The general formula for the asymptotic variance of a function of several statistics $\hat{Z} = (\hat{\zeta}_1, \ldots, \hat{\zeta}_p)$ is

$$Var(f(\hat{Z})) = \sum_{s=1}^{p}\sum_{t=1}^{p} \left.\frac{\partial f}{\partial \hat{\zeta}_s}\right|_Z \left.\frac{\partial f}{\partial \hat{\zeta}_t}\right|_Z Cov(\hat{\zeta}_s, \hat{\zeta}_t)$$

(Wolter, 1985, ch. 6).

Applied to (4.3) we get

$$\begin{aligned}
Var(\hat{\theta}_j) &= Var(g(\hat{\mu}_0, \hat{\mu}_j, \hat{\sigma})) \\
&\approx \left(\frac{\partial \hat{\theta}_j}{\partial \hat{\mu}_0} \bigg|_{\mu_0, \mu_j, \sigma} \right)^2 Var(\hat{\mu}_0) \\
&\quad + \left(\frac{\partial \hat{\theta}_j}{\partial \hat{\mu}_j} \bigg|_{\mu_0, \mu_j, \sigma} \right)^2 Var(\hat{\mu}_j) \\
&\quad + \left(\frac{\partial \hat{\theta}_j}{\partial \hat{\sigma}} \bigg|_{\mu_0, \mu_j, \sigma} \right)^2 Var(\hat{\sigma}) \\
&\quad + 2 \frac{\partial \hat{\theta}_j}{\partial \hat{\mu}_0} \bigg|_{\mu_0, \mu_j, \sigma} \frac{\partial \hat{\theta}_j}{\partial \hat{\mu}_j} \bigg|_{\mu_0, \mu_j, \sigma} Cov(\hat{\mu}_0, \hat{\mu}_j) \\
&\quad + 2 \frac{\partial \hat{\theta}_j}{\partial \hat{\mu}_0} \bigg|_{\mu_0, \mu_j, \sigma} \frac{\partial \hat{\theta}_j}{\partial \hat{\sigma}} \bigg|_{\mu_0, \mu_j, \sigma} Cov(\hat{\mu}_0, \hat{\sigma}) \\
&\quad + 2 \frac{\partial \hat{\theta}_j}{\partial \hat{\mu}_j} \bigg|_{\mu_0, \mu_j, \sigma} \frac{\partial \hat{\theta}_j}{\partial \hat{\sigma}} \bigg|_{\mu_0, \mu_j, \sigma} Cov(\hat{\mu}_j, \hat{\sigma})
\end{aligned}$$

The variances of the estimators are given by (4.7), (4.9), and (4.11), respectively. Further required are the estimators' covariances and the partial derivatives.

Since the estimated mean and the estimated variance of a normal distribution are stochastically independent (Hogg and Craig, 1978, ch. 4.8) and independence holds for functions random variables too (ibid., ch. 4.3), we conclude that

$$Cov(\hat{\mu}_0, \hat{\sigma}) = Cov(\hat{\mu}_j, \hat{\sigma}) = 0$$

On the contrary, the covariance between $\hat{\mu}_0$ and $\hat{\mu}_j$ is not zero since the sampling elements in U_j contribute to both $\hat{\mu}_0$ and $\hat{\mu}_j$. To derive the covariance, we first re-write $\hat{\mu}_0$ as a weighted sum of two independent domain means:

$$\hat{\mu}_0 = w_j \hat{\mu}_j + w_{-j} \hat{\mu}_{-j}$$

where $U_{-j} = U \setminus U_j$ and the weights correspond to the expected domain fractions (that is, the expected domain sample sizes divided by the overall sample size). The derivation is now straightforward.

$$\begin{aligned}
Cov(\hat{\mu}_0, \hat{\mu}_j) &= Cov(w_j\hat{\mu}_j + w_{-j}\hat{\mu}_{-j}, \hat{\mu}_j) \\
&= E(w_j\hat{\mu}_j^2 + w_{-j}\hat{\mu}_{-j}\hat{\mu}_j) - (w_j\mu_j + w_{-j}\mu_{-j})\mu_j \\
&= w_jE(\hat{\mu}_j^2) + w_{-j}\mu_{-j}\mu_j - w_j\mu_j^2 - w_{-j}\mu_{-j}\mu_j \\
&= w_j\left[E(\hat{\mu}_j^2) - \mu_j^2\right] \\
&= w_jVar(\hat{\mu}_j) \\
&\approx \frac{E(n_j)}{n}\frac{\sigma^2}{E(n_j)} \\
&= \sigma^2/n \quad \square
\end{aligned}$$

One can discover that

$$Corr(\hat{\mu}_0, \hat{\mu}_j) = \sqrt{\frac{E(n_j)}{n}}$$

Finally we have to calculate the partial derivatives. For simplicity we define

$$\kappa = \frac{\log(c) + \mu_0 - \mu_j}{\hat{\sigma}} \tag{4.12}$$

The partial derivatives are given by

$$\begin{aligned}
\left.\frac{\partial\hat{\theta}_j}{\partial\hat{\mu}_0}\right|_{\mu_0,\mu_j,\sigma} &= \phi(\kappa)\frac{1}{\sigma} \\
\left.\frac{\partial\hat{\theta}_j}{\partial\hat{\mu}_j}\right|_{\mu_0,\mu_j,\sigma} &= \phi(\kappa)\frac{-1}{\sigma} \\
\left.\frac{\partial\hat{\theta}_j}{\partial\hat{\mu}_0}\right|_{\mu_0,\mu_j,\sigma} &= \phi(\kappa)\kappa\frac{-1}{\sigma}
\end{aligned}$$

So the asymptotic variance of (4.3) is

$$\begin{aligned}
Var(\hat{\theta}_j) &\approx [\phi(\kappa)]^2\left[\frac{1}{\sigma^2}\frac{\sigma^2}{n} + \frac{1}{\sigma^2}\frac{\sigma^2}{E(n_j)} + \kappa^2\frac{1}{\sigma^2}\frac{\sigma^2}{2n} + 2\left(\frac{-1}{\sigma^2}\right)\frac{\sigma^2}{n}\right] \\
&= [\phi(\kappa)]^2\left[\kappa^2/2n + \left(\frac{1}{E(n_j)} - \frac{1}{n}\right)\right] \square
\end{aligned}$$

with κ given by (4.12).

The variance estimator for (4.3) is

$$\widehat{Var}(\hat{\theta}_j) = \left[\phi\left(\frac{\log(c) + \hat{\mu}_0 - \hat{\mu}_j}{\hat{\sigma}}\right)\right]^2 \tag{4.13}$$

$$* \left[\left(\frac{\log(c) + \hat{\mu}_0 - \hat{\mu}_j}{\hat{\sigma}}\right)^2 /2n + \left(\frac{1}{E(n_j)} - \frac{1}{n}\right)\right] \tag{4.14}$$

It is easy to verify that if the study domain coincides with the entire population, that is, $\mu_j = \mu_0$ and $E(n_j) = n$, then (4.13) reduces to (3.4).

If there is not actually between-domain variation of incomes and the median income in the study domain U_j equals the median income in the entire population, then (3.3) and (4.4) produce the same estimates, but their variances differ. More precisely, the variance of (4.4) exceeds that of (3.3) since

$$\left(\frac{1}{E(n_j)} - \frac{1}{n}\right) > 0$$

5 Population poverty threshold: Unequal within-domain dispersion

In the previous chapter we examined the at-risk-of-poverty rate in the case of different domain location parameters but equal within-domain dispersion parameters. Let us now discuss the general case when both the expectation and the standard deviation of log-income differ among the domains of interest.

5.1 Derivation of the true value

Formally we consider the case

$$\begin{aligned} &Y_1 \sim LN(\mu_1, \sigma_1^2) \\ &\quad\vdots \\ &Y_q \sim LN(\mu_q, \sigma_q^2) \end{aligned}$$

or, equivalently,

$$\begin{aligned} &X_1 \sim N(\mu_1, \sigma_1^2) \\ &\quad\vdots \\ &X_q \sim N(\mu_q, \sigma_q^2) \end{aligned}$$

Note that the model in the previous chapter can be seen as the special case with

$$\sigma_1^2 = \ldots = \sigma_q^2$$

The compound distribution of X is again a finite mixture of normal distributions (Everitt and Hand, 1981, ch. 2). The general definition is given by (4.1).

We may think of (μ_j, σ_j^2) as realizations of a random vector (M, Σ) with expectation $(\mu_0, \bar{\sigma}^2)$. If the two components of the random vector are statistically independent and the distribution of M is symmetric with variance σ_0^2, then the following results hold:

1. $E(X) = \mu_0$

2. $Var(X) = \sigma_0^2 + \bar{\sigma}^2$

The proofs are given below.

$$\begin{aligned}
E(X) &= \int_X x f(x) dx \\
&= \int_X x \left(\sum_j \sum_j f(x|\mu_j) f(x|\sigma_j^2) h_\Sigma(\sigma_j^2) h_M(\mu_j) \right) dx \\
&= \sum_j \left[\sum_j \left(\int_X x f(x|\mu_j) dx \right) h_M(\mu_j) \right] h_\Sigma(\sigma_j^2) \\
&= \sum_j \left[\sum_j \mu_j h_M(\mu_j) \right] h_\Sigma(\sigma_j^2) \\
&= \sum_j \mu_0 h_\Sigma(\sigma_j^2) \\
&= \mu_0 \sum_j h_\Sigma(\sigma_j^2) \\
&= \mu_0 \quad \square
\end{aligned}$$

$$
\begin{aligned}
Var(X) &= \int_X (x-\mu_0)^2 f(x)dx \\
&= \int_X (x-\mu_0)^2 \left(\sum_j \sum_j f(x|\mu_j) f(x|\sigma_j^2) h_\Sigma(\sigma_j^2) h_M(\mu_j) \right) dx \\
&= \sum_j \left[\sum_j \left(\int_X (x-\mu_0)^2 f(x|\mu_j)dx \right) h_M(\mu_j) \right] h_\Sigma(\sigma_j^2) \\
&= \sum_j \left[\sum_j \left(\int_X (x-\mu_j+\mu_j-\mu_0)^2 f(x|\mu_j)dx \right) h_M(\mu_j) \right] h_\Sigma(\sigma_j^2) \\
&= \sum_j \sum_j (\sigma_j^2+\sigma_0^2) h_M(\mu_j) h_\Sigma(\sigma_j^2) \\
&= \sum_j \left(\sigma_0^2 + \sum_j \sigma_j^2 h_\Sigma(\sigma_j^2) \right) h_M(\mu_j) \\
&= \sum_j (\sigma_0^2 + \bar{\sigma}^2) h_M(\mu_j) \\
&= (\sigma_0^2 + \bar{\sigma}^2) \sum_j h_M(\mu_j) \\
&= \sigma_0^2 + \bar{\sigma}^2 \quad \square
\end{aligned}
$$

Note that in terms of the ANOVA model, we can regard to $\bar{\sigma}^2$ as the expected mean square within (Sahai and Ageel, 2000, ch. 2.5 and 2.20).

If the distribution of M is normal and the variation in Σ tends towards zero, then as the limiting case we get (4.2).

For practical purposes

$$
Y \sim LN(\mu_0, \sigma_0^2 + \bar{\sigma}^2) \tag{5.1}
$$

can be a useful approximation.

The actual shape of the compound distribution depends on the number of domains, the variation in M and Σ, and the statistical dependence of M and Σ. If the latter are correlated, then the distribution of X is asymmetric. A practical interpretation of positive (negative) correlation between M and Σ is that domains with over-proportional average income tend to have over-proportional (under-proportional) within-domain dispersion.

For the remainder of this thesis we shall assume that (5.1) approximately holds. In section 6.3 we exemplify and discuss the effect of statistical dependence between M and Σ on the accuracy of the estimators.

If (5.1) holds, then the derivation of the at-risk-of-poverty rate in domain U_j is straightforward.

$$\begin{aligned}\theta_j &= \Pr\left\{Y_j < c * \exp(\mu_0)\right\} \\ &= \Pr\left\{X_j < \log\left[c * \exp(\mu_0)\right]\right\} \\ &= \Pr\left\{X_j < \log(c) + \mu_0\right\} \\ &= \Pr\left\{X_j - \mu_j < \log(c) + \mu_0 - \mu_j\right\} \\ &= \Pr\left\{\frac{X_j - \mu_j}{\sigma_j} < \frac{\log(c) + \mu_0 - \mu_j}{\sigma_j}\right\} \\ &= \Phi\left(\frac{\log(c) + \mu_0 - \mu_j}{\sigma_j}\right) \square\end{aligned}$$

5.2 Discussion of the formula

Let us re-write the eventual formula

$$\theta_j = \Phi\left(\frac{\log(c) + d_j}{\sigma_j}\right) \tag{5.2}$$

as

$$\theta_j = \Phi\left(\frac{\log(c) + \mu_0 - \mu_j}{\sigma_j}\right) = \Phi\left(\frac{\log(c) + d_j}{\bar{\sigma}}\frac{\bar{\sigma}}{\sigma_j}\right) \tag{5.3}$$

where $\bar{\sigma}$ equals the square root of the expected mean square within:

$$\bar{\sigma} = \sqrt{\frac{1}{N}\sum_{j=1}^{q} E(n_j)\sigma_j^2}$$

The domain scale parameter σ_j in (5.2) can therefore be seen as an average parameter $\bar{\sigma}$ times an inflation factor, where the latter is given by the ratio $\bar{\sigma}/\sigma_j$. If $\sigma_j > \bar{\sigma}$, that is, income distribution is more unequal in the domain of interest than on the average, then $\bar{\sigma}/\sigma_j < 1$ and so the at-risk-of-poverty rate is larger (except for the rare case when

Table 5.1: Theoretical values of θ_D

		$d_j = \mu_0 - \mu_j$						
σ_j	$\bar{\sigma}/\sigma_j$	0.40	0.20	0.10	-	-0.10	-0.20	-0.40
0.30	1.50	35.6	15.0	8.5	4.4	2.1	0.9	0.1
0.40	1.13	39.1	21.9	15.2	10.1	6.3	3.8	1.1
0.45	1.00	40.3	24.5	18.1	12.8	8.7	5.7	2.1
0.50	0.90	41.2	26.7	20.6	15.3	11.1	7.8	3.4
0.60	0.75	42.7	30.2	24.7	19.7	15.4	11.8	6.5
0.70	0.64	43.7	32.9	27.9	23.3	19.1	15.5	9.7

the argument of $\Phi(\cdot)$ in (5.3) is positive). As a hypothetical special case, if $\sigma_j \to \infty$, then $\theta_j = 0.5$ generally holds.

Note that both the deviation in the location parameters, d_j, and the ratio of the actual dispersion parameter over the expected parameter, $\bar{\sigma}/\sigma_j$, have an impact on θ_j. The practical message is that a low income level (measured by d_j) may be counterbalanced by little income inequality (measured by $\bar{\sigma}/\sigma_j$). We may observe that for a domain with median income below average the at-risk-of-poverty rate is **lower** than for the entire population, and vice versa. This feature of the indicator is illustrated by table 5.1. We hold $c = 0.6$ and $\bar{\sigma} = 0.45$ fixed and compute the true θ_j's for various pairs of values of d_j and σ_j. Again, the θ_j-values were calculated in MS EXCEL and are expressed as percentages.

5.3 Point estimation

Applying the plug-in technique, we insert admissable estimators for μ_0, μ_j, and σ_j and get an admissable estimator for (5.2):

$$\hat{\theta}_j = \Phi\left(\frac{\log(c) + \hat{\mu}_0 - \hat{\mu}_j}{\hat{\sigma}_j}\right) \tag{5.4}$$

The estimators for μ_0 and μ_j are given by (4.6) and (4.8), respectively. Adapting the variances estimators (4.7) and (4.9) to the case of unequal within-domain dispersion, we get

$$Var(\hat{\mu}_j) \approx \sigma_j^2/E(n_j) \tag{5.5}$$

and

$$Var(\hat{\mu}_0) = \bar{\sigma}^2/n \tag{5.6}$$

The domain scale parameter σ_j is estimated by the square root of the estimated domain variance, that is,

$$\hat{\sigma}_j = \sqrt{\frac{1}{n_j - 1}\sum_{i=1}^{n_j}(x_i - \hat{\mu}_j)^2} \tag{5.7}$$

and its sampling variance is approximately

$$Var(\hat{\sigma}_j) \approx \sigma_j^2/2E(n_j) \tag{5.8}$$

If the (expected) domain sample size is relatively small in relation to the overall sample size, then (5.8) is considerably larger than (4.11). It might be the case that using (4.10) as an estimator for σ_j, though in general biased, leads to a smaller mean square sampling error than (5.7). The practical consequence is that if we have reason to believe that the within-domain scale parameters do not differ tremendously, it might be preferable to use (4.10) or a tailored estimator or a combined estimator (Longford, 2005, ch. 6) instead of (5.7). Another approach would be to build clusters of domains for which it can be assumed that the within-domain scale parameters are similar and to estimate σ_j clusterwise. Such a clustering could be based e.g. on theoretical considerations, on a pilot survey, or on auxiliary information. For the remainder of this thesis we shall assume that (5.8) is a suitable estimator for σ_j.

5.4 Interval estimation

The derivation of the variance of (5.4) is analogous to the procedure in section 4.4. We skip the detailed calculation and just present the results.

The covariance of $\hat{\mu}_0$ and $\hat{\mu}_j$ equals

$$Cov(\hat{\mu}_0, \hat{\mu}_j) = \frac{E(n_j)}{n} Var(\hat{\mu}_j) \approx \sigma_j^2/n$$

and the corresponding correlation is

$$Corr(\hat{\mu}_0, \hat{\mu}_j) = \sqrt{\frac{E(n_j)}{n} \frac{\sigma_j^2}{\bar{\sigma}^2}}$$

The covariances between $\hat{\mu}_0$ and $\hat{\sigma}_j$ and between $\hat{\mu}_j$ and $\hat{\sigma}_j$, respectively, are zero (Hogg and Craig, 1978, ch. 4.3 and 4.8).

The asymptotic variance of (5.4) is

$$\begin{aligned}
Var(\hat{\theta}_j) &\approx \left[\phi\left(\frac{\log(c) + \mu_0 - \mu_j}{\sigma_j}\right)\right]^2 \\
&\quad * \left[\frac{1}{\sigma_j^2}\frac{\bar{\sigma}^2}{n} + \frac{1}{\sigma_j^2}\frac{\sigma_j^2}{E(n_j)} + \left(\frac{\log(c) + \mu_0 - \mu_j}{\sigma_j}\right)^2 \frac{1}{\sigma_j^2}\frac{\sigma_j^2}{2E(n_j)} - \left(\frac{2}{\sigma_j^2}\right)\frac{\sigma_j^2}{n}\right] \\
&= \left[\phi\left(\frac{\log(c) + \mu_0 - \mu_j}{\sigma_j}\right)\right]^2 \\
&\quad * \left[\left(\frac{\log(c) + \mu_0 - \mu_j}{\sigma_j}\right)^2 /2E(n_j) + \left(\frac{1}{E(n_j)} - \frac{2 - \bar{\sigma}^2/\sigma_j^2}{n}\right)\right]
\end{aligned}$$

The variance estimator of (5.4) is therefore

$$\begin{aligned}
\widehat{Var}(\hat{\theta}_j) &= \left[\phi\left(\frac{\log(c) + \hat{\mu}_0 - \hat{\mu}_j}{\hat{\sigma}_j}\right)\right]^2 \\
&\quad * \left[\left(\frac{\log(c) + \hat{\mu}_0 - \hat{\mu}_j}{\hat{\sigma}_j}\right)^2 /2E(n_j) + \left(\frac{1}{E(n_j)} - \frac{2 - \hat{\bar{\sigma}}^2/\hat{\sigma}_j{}^2}{n}\right)\right]
\end{aligned}$$

6 Finite sample behavior of the estimators in chapters 3-5

In chapters 3 to 5 we developed parametric point and interval estimators of the at-risk-of-poverty rate under different conditions. We saw that the point estimators are asmyptotically unbiased and derived large-sample formulae for their variances. Whenever we deal with real life data, the sample is finite and consequently we should know about the estimators' behavior in finite samples. Particularly we are interested in how the large-sample formulae approximate the behavior in finite samples. One way of examining this is the Monte Carlo technique.

The approach is as follows. Given an artificial lognormal population (in fact a sample of 100,000 random numbers), we draw 10,000 samples of a specified sample size (with replacement). To each sample the parametric point and interval estimator are applied. We are interested in the distribution of the respective 10,000 pairs of estimates.

The question is how good do the asymptotic results on expectation and variance approximate the behavior in finite samples. From a "good" approximation we would expect that

- the mean of the 10,000 point estimates should conform to the true population value (which is calculated nonparametrically),
- the theoretical (asymptotic) standard error for the given sample size and parameters should conform to the standard deviation of the 10,000 point estimates, and
- about 9,500 of the 10,000 estimated 95% confidence intervals should cover the true value.

Note that the last item does not only depend on correct point and interval estimation but also on the degree of deviation of the statistic's distribution from normality.

6.1 Domain poverty threshold

We draw repeated samples of size n = 30, 50, 100, 200, 500, and 1,000. As parameter values we choose first $\sigma = 0.40, c = 0.6$, then $\sigma = 0.60, c = 0.6$, and finally $\sigma = 0.60, c = 0.4$. In all cases $\mu = 9.00$ is held fixed. The computations were done in R, version 2.6.1. The results are given in table 6.1. The true value was computed nonparametrically.

The mean estimate is in general lower than the true value, so the estimator is biased downwards in finite samples. However, the size of bias is relatively small and decreases with increasing sample size. For n = 1,000, the absolute value of bias is about 0.1 percentage points and its relative size about 2% of the true value, what is a fairly acceptable deviation (Cochran, 1963, ch. 1.7).

The asymptotic standard error agrees very well with the empirical standard deviation of the point estimates, even for modest sample sizes. We conclude that the asymptotic formula is highly valid in finite samples.

The actual coverage frequency of the asymptotic confidence intervals increases first with increasing sample size, as the bias decreases. However, from a sample size of 500 onwards the coverage frequency does not seem to increase further, but is in any case at least 93%. For small samples, the value is usually smaller, probably as a result of bias.

6.2 Population poverty threshold: Equal within-domain dispersion

The entire population consists of 10 domains each of equal size 100,000. We use stratified random sampling with proportional allocation, that is, the total sample size is 10 times the domain sample size. The domain location vector is

$$\vec{\mu} = (8.70, 8.80, 8.88, 8.94, 8.98, 9.02, 9.06, 9.12, 9.20, 9.30)$$

so $\mu_0 = 9.00$ and $\sigma_0^2 \approx 0.03$. The within domain variance of the log-incomes, σ^2, is taken first 0.16 and then 0.36. The factor c is fixed at 0.6. As domains of interest, first domain j=1 with location parameter 8.70 (so $\mu_0 - \mu_j = 0.30$) and then domain j=8 with location parameter 9.12 (so $\mu_0 - \mu_j = -0.12$) were chosen.

Table 6.1: Finite sample behavior of $\hat{\theta}$

n	Mean of estimate	True value	Asymptotic S.E.	S.D. of estimate	C.I. coverage probability
			$\sigma = 0.40, c = 0.6$		
30	9.80	10.22	2.91	2.87	91.4
50	9.90	10.22	2.25	2.21	92.9
100	9.96	10.22	1.59	1.58	93.7
200	10.02	10.22	1.13	1.12	94.1
500	10.03	10.22	0.71	0.71	94.0
1,000	10.05	10.22	0.50	0.51	93.0
			$\sigma = 0.60, c = 0.6$		
30	19.27	19.84	3.05	3.16	93.2
50	19.43	19.84	2.36	2.41	93.9
100	19.61	19.84	1.67	1.71	94.4
200	19.68	19.84	1.18	1.19	94.5
500	19.73	19.84	0.75	0.75	94.6
1,000	19.74	19.84	0.53	0.53	94.4
			$\sigma = 0.60, c = 0.4$		
30	6.25	6.45	2.45	2.37	90.2
50	6.29	6.45	1.90	1.87	91.8
100	6.32	6.45	1.34	1.34	92.9
200	6.33	6.45	0.95	0.95	93.7
500	6.34	6.45	0.60	0.59	94.3
1,000	6.36	6.45	0.42	0.43	93.9

What makes estimation in the case of a domain poverty threshold different from estimation in the case of a population poverty threshold is that in the latter case we do not actually know whether income in the entire population follows a lognormal distribution. In chapter 4 we used (4.2) as an approximation. The goodness of the estimators also depends on the goodness of that approximation.

Opposite to the previous results, the mean decreases with increasing sample size and the bias does not in general tend towards zero (table 6.2). However, the relative bias in the worst case is less than 3%.

The theoretical standard errors are almost equal to the empirical standard deviations in repeated samples, so we conclude that the asymptotic variance estimator is well suitable for variance estimation in finite samples. The actual coverage frequencies of the asymptotic 95% confidence intervals are very close to the desired level. All values were far beyond 90%, but not too large.

6.3 Population poverty threshold: Unequal within-domain dispersion

We apply the same sampling procedure and the same $\vec{\mu}$ as in the previous section. For the vector of the within-domain variances we first use a random permutation of the sequence 0.16 (0.02) 0.34, resulting in $\vec{\sigma^2} = (.32, .26, .20, .16, .24, .30, .18, .34, .28, .22)$. A random permutation indicates that $\vec{\mu}$ and $\vec{\sigma^2}$ are stochastically independent (the sample correlation is -0.07). For the average mean square within we have $\bar{\sigma^2} = 0.25$ and $\hat{\sigma} = 0.50$. Again c is fixed at 0.6. We investigate three domains of interest, $j = 1, 5, 9$, with parameters (8.70,0.32), (8.98,0.24), and eventually (9.20,0.28).

Note that for estimation in the case of a population poverty threshold and unequal within-domain dispersion we used (5.1) as an approximation. For the case of stochastically independent $\vec{\mu}$ and $\vec{\sigma^2}$ this assumption is more justified than if the two parameter vectors are not statistically independent.

We see (table 6.3) that the means of the point estimates are again very close to the true values. In detail, no relative bias is larger than 2%. However, analogous to the previous section, convergence of the mean estimate to the true value with increasing sample size cannot be observed.

Table 6.2: Finite sample behavior of $\hat{\theta}_D$

n_j ($n = 10n_j$)	Mean of estimate	True value	Asymptotic S.E.	S.D. of estimate	C.I. coverage probability
	$\mu_0 - \mu_j = 0.30, \sigma = 0.40$				
30	29.98	29.68	6.06	6.07	93.6
50	29.96	29.68	4.69	4.64	94.7
100	29.87	29.68	3.32	3.33	94.8
200	29.84	29.68	2.35	2.37	94.8
500	29.78	29.68	1.48	1.47	95.0
1,000	29.80	29.68	1.05	1.05	95.0
	$\mu_0 - \mu_j = 0.30, \sigma = 0.60$				
30	36.45	36.52	6.52	6.48	93.9
50	36.44	36.52	5.05	4.99	94.6
100	36.47	36.52	3.57	3.54	94.7
200	36.46	36.52	2.52	2.56	94.2
500	36.39	36.52	1.60	1.60	94.7
1,000	36.38	36.52	1.13	1.13	94.8
	$\mu_0 - \mu_j = -0.12, \sigma = 0.40$				
30	5.98	5.83	2.13	2.18	92.4
50	5.88	5.83	1.65	1.66	93.0
100	5.82	5.83	1.16	1.18	93.6
200	5.77	5.83	0.82	0.83	93.6
500	5.75	5.83	0.52	0.52	94.1
1,000	5.75	5.83	0.37	0.37	93.3
	$\mu_0 - \mu_j = -0.12, \sigma = 0.60$				
30	14.95	14.77	4.10	4.15	92.8
50	14.90	14.77	3.17	3.20	94.1
100	14.73	14.77	2.24	2.25	94.3
200	14.68	14.77	1.59	1.58	94.3
500	14.67	14.77	1.00	1.01	94.1
1,000	14.66	14.77	0.71	0.71	94.6

Table 6.3: Finite sample behavior of $\hat{\theta}_D$

n_j ($n = 10n_j$)	Mean of estimate	True value	Asymptotic S.E.	S.D. of estimate	C.I. coverage probability
	$\mu_0 - \mu_j = 0.30, \sigma_j^2 = 0.32$				
30	35.44	35.64	6.61	6.68	93.8
50	35.50	35.64	5.12	5.15	94.6
100	35.49	35.64	3.62	3.64	94.5
200	35.57	35.64	2.56	2.58	94.7
500	35.54	35.64	1.62	1.63	94.8
1,000	35.54	35.64	1.15	1.14	94.6
	$\mu_0 - \mu_j = 0.02, \sigma_j^2 = 0.24$				
30	15.89	15.98	5.23	5.19	92.0
50	15.85	15.98	4.05	4.04	92.6
100	15.83	15.98	2.86	2.86	93.8
200	15.86	15.98	2.03	2.01	94.8
500	15.87	15.98	1.28	1.28	94.6
1,000	15.87	15.98	0.91	0.90	94.9
	$\mu_0 - \mu_j = -0.20, \sigma_j^2 = 0.28$				
30	9.16	9.05	3.95	3.92	90.5
50	9.13	9.05	3.06	3.11	91.7
100	9.06	9.05	2.17	2.18	93.2
200	9.01	9.05	1.53	1.53	94.1
500	9.02	9.05	0.97	0.96	95.1
1,000	9.01	9.05	0.68	0.70	94.0

The theoretical asymptotic standard error again conforms very well to the empirical standard deviation of the point estimates. As a result, the actual coverage frequencies of the asymptotic confidence intervals are very good, especially for $n \geq 100$ (all values between 94.0% and 95.1%).

As a second case we consider now stochastical dependence between domain location and within-domain dispersion, namely $\vec{\sigma^2} = (.16, .18, .20, .22, .24, .26, .28, .30, .32, .34)$, so $\vec{\mu}$ and $\vec{\sigma^2}$ are positively correlated (the sample correlation is 0.99). In this case the approximation (5.1) is not very realistic. Rather the log-incomes in the entire population are positively skewed.

What we observe here (table 6.4) is far different from the previous results. In particular, the point estimators are considerably biased, up to a relative amount of 8%. For large sample sizes the bias even exceeds the standard deviation of the point estimates. This clearly results in much lower coverage frequencies (Cochran, 1963, ch. 1.7). For instance, the actual coverage frequencies for sample size 1,000 lie between 66.7% and 85.0%. On the contrary, for small samples the procedure still gives good results.

We conclude that the validity of the parametric estimator highly depends on the validity of (5.1). Furthermore we see — what is somewhat surprising — that the actual coverage frequency of the asymptotic confidence intervals in general does not become better with increasing sample size.

Table 6.4: Finite sample behavior of $\hat{\theta}_D$

n_j ($n = 10n_j$)	Mean of estimate	True value	Asymptotic S.E.	S.D. of estimate	C.I. coverage probability
	$\mu_0 - \mu_j = 0.30, \sigma_j^2 = 0.16$				
30	29.88	28.16	6.63	6.69	93.6
50	29.78	28.16	5.14	5.15	93.9
100	29.91	28.16	3.63	3.61	92.7
200	29.88	28.16	2.57	2.58	90.3
500	29.90	28.16	1.62	1.64	81.5
1,000	29.92	28.16	1.15	1.14	66.7
	$\mu_0 - \mu_j = 0.02, \sigma_j^2 = 0.24$				
30	15.75	14.93	5.23	5.17	93.0
50	15.87	14.93	4.05	4.00	94.3
100	15.87	14.93	2.86	2.88	94.3
200	15.83	14.93	2.03	2.01	94.1
500	15.84	14.93	1.28	1.27	90.1
1,000	15.85	14.93	0.91	0.90	84.0
	$\mu_0 - \mu_j = -0.20, \sigma_j^2 = 0.32$				
30	10.55	9.80	4.27	4.18	92.6
50	10.58	9.80	3.31	3.28	94.0
100	10.57	9.80	2.34	2.30	94.7
200	10.54	9.80	1.65	1.64	94.3
500	10.53	9.80	1.05	1.04	90.8
1,000	10.53	9.80	0.74	0.73	85.0

7 A comparison of parametric and nonparametric estimation

7.1 Data and methods

In chapter 6 we examined the properties of the parametric estimators in finite samples. These samples were drawn from artificial lognormal universes. Real life datasets do not exactly follow a theoretical curve. Indeed, the lognormal distribution is a model that may be more or less appropriate to describe an empirical income distribution. Consequently, the parametric estimators may be less or more biased. It is widely agreed that a moderately biased estimator can be accepted if the estimator performs well in sampling variance (Cochran 1963, ch. 1.7; Saerndal et al., 1992, ch. 5.2). On the other hand, it is generally agreed that greatly biased estimators should be avoided, since "...greatly biased estimates are poor no matter what other properties they have" (Saerndal et al., 1992, p. 164).

In this chapter we inspect the performance of the new parametric estimator in real life datasets. We restrict to the situation of estimating the at-risk-of-poverty rate for the entire universe (cf. chapter 3). As a point of reference we take the established nonparametric estimator (Preston, 1995).

We use the SILC data and the ADMIN data after top-bottom recoding (cf. chapter 2). The entire SILC data and the entire ADMIN data, respectively, as regarded as target universes, that is, as fictive populations for which we want to make inferences about. Consequently, the nonparametric estimates for the entire datasets are regarded as true values.

To evaluate the performance of the parametric estimator compared to the nonparametric estimator for different sample sizes, we investigate samples of size 30, 50, 100, 200, 500, 1,000, and 2,000. For each sample size, a number of 10,000 repeated samples is drawn (without replacement) from the imaginary target universe. Then for each repeated sample, we compute both the parametric and the nonparametric point estimator. The comparison of the estimators is then done on characterisitics of the corresponding point estimate distribution.

To judge the goodness of an estimator, we compute the following measures:

- mean estimate, bias and squared bias,
- standard deviation, variance and mean square error, and
- the empirical 2.5% and 97.5% quantiles (that is, the 250th smallest and the 9750th smallest estimate).

Note that as we are sampling without replacement, the nonparametric estimate per definition coincides with the true value if the sample size equals the universe size.

7.2 SILC data

Table 7.1 gives the results for the SILC data, computed with R, version 2.6.1. Values are given as percentages, respectively percentage points and squared percentage points (squared bias, variance, and mean square error). Note that for $n = 500$ onwards, the sampling fraction exceeds 5%, so the reduction of the sampling variance by doubling the sample size is more than 50% (Cochran 1963, ch. 2.4 and 2.5).

The "true value" equals 14.38%.

We see that both the parametric and the nonparametric estimator are biased in small samples. However, the bias of the nonparametric estimator is much smaller in absolute value. Additionally, it tends to zero as the sample size increases, whereas the bias of the parametric estimator does not only not vanish with increasing sample size but even increases in absolute value. That phenomenon is observed also for the ADMIN data in the next subsection and cannot readily be explained.

As expected, the standard deviation of the parametric point estimates is smaller than of the nonparametric estimates. Regardless of the sample size the reduction in sampling variance is about 50%.

Based on the empirical mean square error, the parametric estimator performs 20% better for a sample of size 100 and 20% worse for a sample of size 200. Concerning the parametric estimates, we see that for $n = 500$ onwards the empirical 2.5% quantile is larger than the true value.

7.3 ADMIN data

The results for the ADMIN data are given in table 7.2. Interpretation is analogous to the previous section, but the sampling fraction exceeds 5% only for $n = 2,000$ (about 7%).

The "true value" equals 19.83%.

Again we see that both estimators are biased in small samples, and that the bias of the nonparametric estimator vanishes with increasing sample size, whereas the bias of the parametric estimator increases. However, compared with the SILC data, the excess bias of the parametric estimator is much smaller in the ADMIN data. We conclude that for this dataset the lognormal model fits the data better.

The reduction in sampling variance because of parametric estimation is larger in the ADMIN data than in the SILC data: about 2/3. One possible explanation is that because of the good agreement of the model with the data, the theoretical gains in accuracy (e.g. because of estimating the expectation by the arithmetic mean) can be better achieved.

The empirical mean square error of the parametric estimator is smaller in all considered sample sizes, with approximate equivalence for $n = 2,000$. For $n = 200$, the mean square error of the parametric estimator is more than 60% smaller than of the nonparametric estimator. Remember from the previous section that for the SILC data for $n = 200$ we observed a higher mean square error for parametric estimation.

We conclude that the use of parametric estimators may result in substantial improvement of the estimates if the lognormal curve fits the data quite well and the sample size is small or medium. On the contrary, if the agreement of the data with the distributional assumption is poor, nonparametric estimation does substantially outperform parametric estimation, except for small samples. Therefore in small samples the parametric estimator should be generally preferred, whereas it should be avoided in large samples and is in general not recommended in medium samples.

Table 7.1: Comparison of parametric and nonparametric estimator for the SILC data

Estimator	Mean	Bias	Sq.Bias	S.D.	Variance	MSE	2.5% quantile	97.5% quantile
	$n = 30$							
Parametric	15.72	1.34	1.79	4.02	16.13	17.92	7.78	27.46
Nonparam.	13.95	-0.43	0.19	5.85	34.18	34.37	3.33	40.00
	$n = 50$							
Parametric	15.98	1.60	2.56	3.14	9.83	12.39	9.55	26.62
Nonparam.	14.18	-0.20	0.04	4.54	20.61	20.66	6.00	32.00
	$n = 100$							
Parametric	16.19	1.80	3.26	2.25	5.08	8.33	11.66	24.09
Nonparam.	14.27	-0.11	0.01	3.22	10.36	10.37	8.00	31.00
	$n = 200$							
Parametric	16.28	1.89	3.58	1.57	2.47	6.05	13.14	21.17
Nonparam.	14.33	-0.05	0.00	2.24	5.03	5.04	10.00	24.00
	$n = 500$							
Parametric	16.36	1.98	3.92	0.95	0.91	4.83	14.48	19.75
Nonparam.	14.33	-0.05	0.00	1.38	1.91	1.91	11.60	20.40
	$n = 1,000$							
Parametric	16.38	1.99	3.96	0.64	0.41	4.37	15.12	19.02
Nonparam.	14.33	-0.06	0.00	0.92	0.84	0.84	12.50	17.90
	$n = 2,000$							
Parametric	16.39	2.01	4.02	0.38	0.15	4.17	15.64	17.62
Nonparam.	14.33	-0.05	0.00	0.55	0.31	0.31	13.25	16.20

Table 7.2: Comparison of parametric and nonparametric estimator for the ADMIN data

Estimator	Mean	Bias	Sq.Bias	S.D.	Variance	MSE	2.5% quantile	97.5% quantile
				$n = 30$				
Parametric	19.78	-0.04	0.00	3.63	13.19	13.19	11.92	30.34
Nonparam.	19.32	-0.51	0.26	6.17	38.03	38.29	6.67	40.00
				$n = 50$				
Parametric	20.08	0.25	0.06	2.82	7.96	8.03	13.99	28.02
Nonparam.	19.49	-0.33	0.11	4.82	23.20	23.31	10.00	38.00
				$n = 100$				
Parametric	20.26	0.44	0.19	1.94	3.76	3.95	16.19	26.46
Nonparam.	19.72	-0.11	0.01	3.43	11.76	11.77	13.00	34.00
				$n = 200$				
Parametric	20.32	0.50	0.25	1.36	1.86	2.11	17.51	25.16
Nonparam.	19.75	-0.08	0.01	2.44	5.95	5.96	15.00	31.00
				$n = 500$				
Parametric	20.39	0.56	0.32	0.86	0.75	1.06	18.64	23.51
Nonparam.	19.81	-0.01	0.00	1.55	2.40	2.40	16.80	25.80
				$n = 1,000$				
Parametric	20.41	0.59	0.36	0.60	0.36	0.71	19.21	22.53
Nonparam.	19.82	-0.01	0.00	1.10	1.21	1.21	17.70	24.80
				$n = 2,000$				
Parametric	20.42	0.60	0.42	0.38	0.17	0.53	19.61	22.02
Nonparam.	19.84	0.02	0.75	0.55	0.56	0.56	18.40	22.75

8 Estimation for persons in households

8.1 Principal considerations

In the previous chapters we have stressed the at-risk-of-poverty rate for households, that is, the proportion of households below a fraction of the median household income, and how to estimate that indicator from sample data. An important complement is the at-risk-of-poverty rate for **persons in households**. This means that income is measured at household level, but the proportion is measured at personal level, so rate describes how many persons live in households below a given fraction of median household income.

It is easy to see that in general the rate for households and the rate for persons in households are not be the same. More precisely, if household size (that is, the number of persons in a household) and income are correlated, then the rate for households will be different from the rate for persons in households. For instance, if the equivalized disposable household income decreases, on average, with increasing household size, then the rate for persons in households will be higher than the rate for households, since the households below the poverty line are on average larger than the households above the poverty line. This is not an unrealistic assumption, as one could think e.g. of households of migrant families. So if we have income data for households but are in fact interested in the prevalence of poverty among persons, we have to take household size into account.

Be ware that it is **not** allowed to simply adapt the formulae from chapters 3 to 5 as if persons were equal to households (for instance, by replacing the number of households, n, with the number of persons). Why is this the case? The reason is that so far we have assumed that the household incomes are independent of each other. If we estimate at personal level, this is by far not true. Quite the contrary, the intra-household correlation of equivalized disposable household income is 100%, since the observed household income in equally applied to all household members.

One possibility to handle the problem is to interpret the household size as an estimation weight in estimation at household level. This technique is in principal also used by EUROSTAT (2005) in variance estimation. The advantage of doing so is that we can still apply the formulae for households and have to adapt only for estimation weights.

8.2 Estimation of a location parameter

Given n sample households, let us denote with $h_1, h_2, \ldots, h_n$ the corresponding household sizes. We interpret these household sizes, or weights, as realizations of a random variable H. Then the sample total of persons is

$$\tau_H = \sum_{i=1}^{n} h_i$$

and the sample average household size is

$$\bar{h} = \tau_H / n$$

The persons-in-households-estimator of a location parameter μ_H in case of simple random sampling is then the person-weighted mean of the logarithmic household incomes, that is,

$$\hat{\mu}_H = \frac{\sum_{i=1}^{n} h_i x_i}{\sum_{i=1}^{n} h_i} \tag{8.1}$$

Note that this definition for a weighted mean holds not only for weighting by the number of persons but for any comparable weigthing procedure. Estimation on household level can be seen as the special case where all weights equal 1.

Since both the x_i's as well as the h_i's are random variables, we can interpret $\hat{\mu}_H$ as a ratio of two random sums. The general formula for the asymptotic simple random sampling variance of a ratio of two random variables is a special application of the linearization method (Wolter, 1985, ch. 6) and given by

$$Var\left(\frac{S}{T}\right) = \frac{1}{T^2}\left[Var(T) + \left(\frac{S}{T}\right)^2 Var(S) - 2\left(\frac{S}{T}\right) Cov(S,T)\right]$$

(ibid., p. 229)

Applied to (8.1) we get

$$Var(\hat{\mu}_H) \approx \frac{1}{\tau_H^2}\left[Var\left(\sum_i h_i\right) + \hat{\mu}_H^2 Var\left(\sum_i h_i x_i\right) - 2\hat{\mu}_H Cov\left(\sum_i h_i x_i, \sum_i h_i\right)\right]$$

Since the sample households are independent of each other, we may write

$$\begin{aligned} Var(\hat{\mu}_H) &\approx \frac{1}{\tau_H^2}\left[\sum_i Var(h_i) + \hat{\mu}_H^2 \sum_i Var(h_i x_i) - 2\hat{\mu}_H \sum_i Cov(h_i x_i, h_i)\right] \\ &= \frac{1}{\tau_H^2}\left[nVar(H) + \hat{\mu}_H^2 nVar(HX) - 2\hat{\mu}_H nCov(HX, H)\right] \\ &= \frac{1}{\tau_H \bar{h}}\left[Var(H) + \hat{\mu}_H^2 Var(HX) - 2\hat{\mu}_H Cov(HX, H)\right] \end{aligned}$$

So the variance estimator for (8.1) is

$$\widehat{Var}(\hat{\mu}_H) = \frac{1}{\tau_H \bar{h}}\left[\widehat{Var}(H) + \hat{\mu}_H^2 \widehat{Var}(HX) - 2\hat{\mu}_H \widehat{Cov}(HX, H)\right] \tag{8.2}$$

We see that the variance of (8.1) is influenced by both the variance of the log-income and the variance of the household size. Furthermore, it is influenced by the algebraic sign of the correlation between HX and H. In case of a negative correlation the variance increases, in case of a positive correlation it decreases. A practical interpretation is that the sampling variance in the number of persons below the poverty line is larger if poverty primary affects large households than smaller if it primary affects single-person households.

Simulation shows that if income and household size are uncorrelated, so the true rate for persons in households equals the true rate for households, then the variance of the person-weighted estimator is higher than the variance of the unweighted estimator. An informal explanation for that phenomenon is that the introduction of needless weights causes additional sampling variability. By contrast, situations can be simulated where the correlation between weights and income leads to a sampling variance of the weighted estimator compared with the unweighted estimator.

8.3 The concept of weighting effect and variance inflation

As mentioned above, the sampling variance of the persons-in-households-estimator of a location parameter, given by (8.2), may be higher or lower than the sampling variance of the corresponding unweighted estimator at household level, $Var(\hat{\mu})$.

Let us now introduce a (scale-free) relative measure of sampling variability of the weighted mean, the weighting effect:

$$W_{eff}(\hat{\mu}_H) = \frac{Var(\hat{\mu}_H)}{Var(\hat{\mu})} \tag{8.3}$$

The idea is analogous to the design effect for various sampling designs (Kish and Fraenkel, 1974; Saerndal et al., 1992, ch. 2.10): The weighting effect describes the ratio of the variance of the weighted mean to the variance of the unweighted mean. A value of 1 stands for no variance inflation caused by weighting, whereas a value ¿1 indicates variance inflation. As mentioned before, if income and the number of persons are uncorrelated, then $W_{eff}(\hat{\mu}_P)$ is greater than 1.

The reason for introducing the weighting effect is as follows. Whenever we have to estimate for persons in households a more complicated sampling variation, for instance the sampling variance of a person-weighted standard deviation or the sampling covariance between a person-weigthed domain mean and a person-weighted population mean, it might be a good practice to calculate the corresponding sampling variability for estimation at household level and inflate it by (8.3). An analogous procedure for variance inflation caused by complex sampling designs and/or nonresponse is proposed by Kish and Fraenkel (1974), discussed by Saerndal et al. (1992, ch. 13.3), and reviewed in this thesis in chapter 9.

For instance, in case of simple random sampling the persons-in-households-estimator of a scale parameter σ_H is the person-weighted standard deviation of the logarithmic household incomes,

$$\hat{\sigma}_H = \sqrt{\frac{1}{\tau_H - 1} \sum_{i=1}^{n} h_i (x_i - \hat{\mu}_H)^2}$$

and its sampling variance can be approximated by

$$Var(\hat{\sigma}_H) \approx \frac{\sigma^2}{2n} W_{eff}(\hat{\mu}_H)$$

9 Other sources of variance inflation

9.1 Overview

So far we have assumed that the data came from a simple random sample from an infinite universe, if need be a stratified random sample with proportional allocation. This assumption is meaningful in as much formulae for point and interval estimators can be derived easily. However, in real life, the data frequently do not come from a simple random sample from an infinite universe. This is particularly the case in a survey.

The assumption of an infinite universe is discussable. On the one hand, sampling elements are drawn from a finite frame. On the other hand, the target population does not necessarily coincide with the sampling frame (or some modification of the sampling frame) because it is possible to interpret the sampling frame as a random sample of a fictive infinite target universe (Saerndal et al., 1992, ch. 13.6). However, this discussion is mostly of little importance since in poverty research the sampling fraction is usually quite small. We shall not track this point further.

The more serious problem is the validity of point and especially interval estimators when the sample differs from a simple random sample. Reasons for such a deviation that frequently occur are

- complex sampling designs,
- unit nonresponse,
- item nonresponse, and
- the use of auxiliary variables in estimation (e.g. post-stratification and raking)

An extensive discussion of the topic is beyond the scope of this thesis. The reader may refer to the standard reference by Saerndal et al. (1992). A very useful textbook, particularly for introductory purposes, is also Cochran (1963). A brief overview may be found in Wolter (1985, ch. 1.3 and 1.4). We shall review some basic results concerning complex sampling designs and unit nonresponse.

9.2 Probability sampling, the π estimator, and the design effect

Let us assume that the class of possible sampling designs is restricted to probability sampling (Saerndal et al., 1992, 1.3). Informally speaking, probability sampling means that all elements in the finite target universe have known positive probabilities of being included in the sample. These (first order) inclusion probabilities are not necessarily equal for all elements, but they are all different from zero (ibid.).

Famous examples of probability sampling designs are simple random sampling, stratified random sampling, cluster sampling, two-stage sampling, and systematic sampling (ibid., ch. 3 and 4). An example of a non-probability sampling design is quota sampling (ibid., ch. 14.4).

Let us denote the finite target universe by U and the sample set of elements by S. Given probability sampling we can for any element k in U and for any pair of elements k, l in U define first order respectively second order inclusion probabilities:

$$\pi_k = \Pr(k \in S) \tag{9.1}$$

$$\pi_{kl} = \Pr(k \& l \in S) \tag{9.2}$$

The calculation of the inclusion probabilities is given by Saerndal et al. (1992, ch. 2.4). In general the second order inclusion probability is **not** equal to the product of the corresponding first order inclusion probabilities.

Then the total of a random variable Y, which we shall denote t_Y, can be estimated via the π estimator (sometimes referred to as the Horvitz-Thompson estimator):

$$\hat{t}_Y = \sum_{k \in S} \frac{y_k}{\pi_k} \tag{9.3}$$

and its estimated sampling variance is

$$\widehat{Var}(\hat{t}_Y) = \sum_{k \in S} \sum_{l \in S} \left(1 - \frac{\pi_k \pi_l}{\pi_{kl}}\right) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \tag{9.4}$$

(Deville, 1999).

The general forms of (9.3) and (9.4) can be applied to the specific sampling design to get more explicit formulae. For instance, the π estimator for simple random sampling is

$$\hat{t}_Y|SRS = N\bar{y}$$

and its sampling variance may be estimated by

$$\widehat{Var}(\hat{t}_Y|SRS) = N^2\frac{1-f}{n}\widehat{Var}(Y) \tag{9.5}$$

where N refers to the size of the target universe, n to the sample size and f to the sampling fraction, that is, n/N (Saerndal et al., 1992, p.46).

The concept of the design effect (ibid., ch. 2.10; Kish and Fraenkel, 1974) is now as follows. Given a certain probability sampling design Δ, we estimate the variance $\widehat{Var}(\hat{t}_Y|\Delta)$ according to (9.4). Then we compare this variance estimate with the variance estimate for a simple random sample of equal size and refer to the ratio as the design effect:

$$\widehat{D_{eff}}(\Delta) = \frac{\widehat{Var}(\hat{t}_Y|\Delta)}{\widehat{Var}(\hat{t}_Y|SRS)} \tag{9.6}$$

The design effect of a specific sampling design may be greater or less unity. The importance of (9.6) is that whenever we have to estimate the variance (strictly speaking, the design variance) of a more complex statistic that is difficult to calculate explicitly for the given sampling design, it is usually a good procedure to first calculate the variance for simple random sampling and then inflate it by the design effect. Kish and Fraenkel (1974) report, based on a large Monte Carlo study, that the resulting implicitly calculated variances for various complex statitics were mostly conservative approximations of the actual Monte Carlo variances, so using the variance inflation technique we might play it safe without becoming unrealistic (in their paper they calculate the design effect with regard to the arithmetic mean, what makes little difference to the total). A summary is given by Saerndal et al. (1992, ch. 13.4) who also report some theoretical design effects.

In a formal way, the variance of a complex statistic $\hat{\psi}$ under a specific sampling design Δ can be approximated as

$$\widehat{Var}(\hat{\psi}|\Delta) \approx \widehat{Var}(\hat{\psi}|SRS) * \widehat{D_{eff}}(\Delta) \tag{9.7}$$

9.3 Correction for unit nonresponse

An important phenomenon in surveys that has considerable impact on the accuracy of sample estimates is unit nonresponse. Unit nonresponse means that for some of the sample elements (say, households) no data is available, e.g. because the household is not accessible or refuses to answer. Particularly when the survey is voluntary, unit nonresponse plays a central role in the validity of estimators because in general we cannot assume independence between response behavior and the target variable (e.g. Cochran, 1963, p. 356).

Advanced statistical models to correct for unit nonresponse had not been established for a long time. For instance Cochran (1963, p. 357) reports that "We are left in the position of relying some guess about the size of bias..." However, since the 1970's statisticians have developed more useful techniques to deal with the problem, particularly because in several countries nonresponse proportions in surveys increased sharply around 1970 (Saerndal et al., 1992, p. 552).

A general model to correct for unit nonresponse is the introduction of response probabilities. Given U the target universe and S the set of included sampling units, let us denote by R the set of responding sampling units (that is, R is a subset of S). Let π_k and π_{kl} be defined according to (9.1) and (9.2), respectively. Then we define the first and second order response probabilities:

$$p_k = \Pr(k \in R | k \in S) \tag{9.8}$$

$$p_{kl} = \Pr(k\&l \in R | k\&l \in S) \tag{9.9}$$

Let us first assume that these response probabilities are known. Then the expanded π estimator of the total of Y is

$$\hat{t}_Y = \sum_{k \in R} \frac{y_k}{\pi_k p_k} \tag{9.10}$$

and its estimated sampling variance is

$$\begin{aligned}\widehat{Var}(\hat{t}_Y) = & \sum_{k \in R} \sum_{l \in R} \left(1 - \frac{\pi_k \pi_l}{\pi_{kl}}\right) \frac{1}{p_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \\ & + \sum_{k \in R} \sum_{l \in R} \left(1 - \frac{p_k p_l}{p_{kl}}\right) \frac{y_k}{\pi_k p_k} \frac{y_l}{\pi_l p_l}\end{aligned}$$

(Deville, 1999).

In general the true response probabilites are unknown and have to be estimated either from the current sample (Saerndal et al., 1992, ch. 15.6.2) and/or using auxiliary information (Berger and Skinner, 2003).

A useful model is to partition S into G groups for which response behavior is likely to be similar within the groups. An important criterion for partitioning may be the geographic location (e.g. urban or rural; cf. Statistik Austria, 2005). Then let us assume that the response probability within a group is constant, whereas different groups have different response probabilities, and that sampling elements respond independent of each other. This model is known as the **response homogeneity group (RHG) model** and described in detail by Saerndal et al. (1992, ch. 15.6.2).

Let us denote the achieved response size in group g by a_g. The natural estimator for the response probability in group g is the response fraction,

$$\hat{p}_k | k \in S_g = a_g / n_g$$

and because of independent response we have

$$\hat{p}_{kl} = \hat{p}_k \hat{p}_l \tag{9.11}$$

Note that the a_g's are random, even if the n_g's are fixed.

The overall sampling process is then modeled as a kind of two-phase sampling: The first phase is equivalent to the respective sampling design, and a second phase with stratified Bernoulli sampling serves as a model for response (Saerndal et al., 1992, ch. 9.8).

The expanded π estimator for the RHG model is

$$\hat{t}_Y = \sum_{g=1}^{G} \frac{1}{\hat{p}_g} \sum_{k \in R_g} \frac{y_k}{\pi_k}$$

and its sampling variance may be estimated by

$$\begin{aligned}\widehat{Var}(\hat{t}_Y) = & \sum_{k \in R} \sum_{l \in R} \left(1 - \frac{\pi_k \pi_l}{\pi_{kl}}\right) \frac{1}{\hat{p}_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \\ & + \sum_{g=1}^{G} \left(\frac{1}{\hat{p}_g^2} - \frac{1}{\hat{p}_g}\right) \sum_{k \in R_g} \left(\frac{y_k}{\pi_k}\right)^2\end{aligned}$$

(ibid.; EUROSTAT, 2005)

Saerndal (1992, ch. 15.6.3) also proposes a modification applied especially to estimated response probabilities. In a nutshell, the approach is to condition on the achieved response sizes $a_1, \ldots, a_G$ and interpret sampling phase two as stratified random sampling. Doing so, (9.11) does not hold.

The concept of the design effect, introduced in 9.2, can be extended to account for both the sampling design and unit nonresponse. In a somewhat comparable application Berger and Skinner (2003) use the term “misspecification effect”.

Bibliography

[1] Aitchison, J., and Brown, J. (1957): *The lognormal distribution.* Cambridge University Press.

[2] Atkinson, A., and Bourguignon, F. (2000): *Income distribution and economics.* In Atkinson, A., and Bourguignon, F. (eds.): Handbook of income distribution, vol. 1. Elsevier Science, Amsterdam.

[3] Barndorff-Nielsen, O., et al. (1982): *Normal variance-mean mixtures and z-distributions.* International Statistical Review, **50**, 145-159.

[4] Berger, Y., and Skinner, C. (2003): *Variance estimation for a low-income proportion.* Social Statistics Research Centre Methodology Working Paper M03/03, University of Southampton.

[5] Bortz, J. (1999): *Statistik fuer Sozialwissenschaftler.* Springer-Verlag, Berlin.

[6] Clementi, F., and Gallegati, M. (2005): *Pareto's law of income distribution: evidence for Germany, the United Kingdom, and the United States.* In Chatterjee, A., et al. (eds.): Econophysics of wealth distributions. Springer-Verlag, Milan.

[7] Cochran, W. (1963): *Sampling techniques.* 2nd ed., John Wiley, New York.

[8] Deville, J. (1999): *Variance estimation for complex statistics and estimators: linearization and residual techniques.* Survey Methodology, **25**, 193-203.

[9] Dragulescu, A., and Yakovenko, V. (2001): *Evidence for the exponential distribution of income in the USA.* The European Physical Journal B, **20**, 585-589.

[10] Elandt-Johnson, R., and Johnson, N. (1980): *Survival models and data analysis.* John Wiley, New York.

[11] EUROSTAT (2003): *Laeken indicators - Detailed calculation methodology.* Room Document E2/IPSE/2003 for the Working Group on "Statistics on Income, Poverty & Social exclusion", 28-29 April 2003.

[12] EUROSTAT (2005): *Variance estimation methodology.* In: Statistik Austria (2006): Einkommen, Armut und Lebensbedingungen. Ergebnisse aus EU-SILC 2004 fuer Oesterreich (p. 80-82). Vienna.

[13] EUROSTAT (2007): *Europa in Zahlen - Eurostat-Jahrbuch 2006-07.* Eurostat, Luxembourg.

[14] Everitt, B., and Hand, D. (1981): *Finite Mixture Distributions.* Chapman and Hall, London.

[15] Graf, M. (2007): *Use of Distributional Assumptions for the Comparison of four Laeken Indicators on EU-SILC Data.* Bulletin of the International Statistical Institute, 56th Session, Lisbon, Vol. 56.

[16] Hogg, R., and Craig, A. (1978): *Introduction to mathematical statistics.* 4th ed., Macmillan, New York.

[17] Judge, G., et al. (1988): *Introduction to the theory and practice of econometrics.* 2nd ed., John Wiley, New York.

[18] Kakwani, N. (1980): *Income inequality and poverty.* Published for the World Bank, Oxford University Press.

[19] Kish, L., and Fraenkel, M. (1974): *Inference from Complex Samples.* Journal of the Royal Statistical Society, Series B, 1-22.

[20] Klotz, J. (2005): *Verdienstunterschiede zwischen den Bundeslaendern - eine Folge von Struktureffekten?* Statistische Nachrichten, **60**, 1003-1008.

[21] Limpert, E., et al. (2001): *Log-normal distributions across the sciences: keys and clues.* BioScience, **51**, 341-352.

[22] Longford, N. (2005): *Missing data and small-area estimation.* Springer-Verlag, New York.

[23] Preston, I. (1995): *Sampling Distributions of Relative Poverty Statistics.* Journal of the Royal Statistical Society, Series C - Applied Statistics, **44**, 91-99.

[24] Saerndal, C., et al. (1992): *Model assisted survey sampling.* Springer-Verlag, New York.

[25] Sahai, H., and Ageel, M. (2000): *The Analysis of Variance. Fixed, Random and Mixed Models.* Birkhaeuser, Boston.

[26] Statistik Austria (2005): *Einkommen, Armut und Lebensbedingungen. Ergebnisse aus EU-SILC 2003 fuer Oesterreich.* Vienna.

[27] Statistik Austria (2007): *Statistik der Lohnsteuer 2006.* Vienna.

[28] Teichroew, D. (1957): *The Mixture of Normal Distributions with Different Variances.* The Annals of Mathematical Statistics, **28**, 510-512.

[29] Thompson, H. (1951): *Truncated Lognormal Distributions: I. Solution by Moments.* Biometrika, **38**, 414-422.

[30] Till, M. (2006): *Jahreseinkommen und erwartete Lebensstandardpositionen von Personen in Privathaushalten auf Grundlage von EU-SILC 2004.* Statistische Nachrichten, **61**, 250-260.

[31] Titterington, D., et al. (1985): *Statistical analysis of finite mixture distributions.* John Wiley, Chichester, UK.

[32] Wolter, K. (1985): *Introduction to variance estimation.* Springer-Verlag, New York.

[33] Yakovenko, V., and Silva, A. (2005): *Two-class structure of income distribution in the USA: exponential bulk and power-law tail.* In Chatterjee, A., et al. (eds.): Econophysics of wealth distributions. Springer-Verlag, Milan.

JOHANNES KLOTZ BAKK.

LEBENSLAUF

Geburtsdatum:	5.9.1981
Geburtsort:	Innsbruck, Österreich

Ausbildung & beruflicher Werdegang

1988–1992	Volksschule Axams
1992–2000	Bundesrealgymnasium Innsbruck, Matura mit Auszeichnung
2000–2001	Präsenzdienst (Hochgebirgsjäger)
seit 2001	Angestellter bei Statistik Austria
seit WS 2002/03	Studium der Statistik an der Universität Wien • Abschluss des Bakkalaureats im Oktober 2006 • Tätigkeit als Tutor im WS 2007/08

Bakkalaureatsarbeiten

- Imputation in Surveys (2004).
- Determinanten von Ernährungsverhalten und Ernährungswissen von Wiener Lehrlingen (2006).

Ausgewählte berufliche Publikationen

- Verbrauchsausgaben – Sozialstatistische Ergebnisse der Konsumerhebung 1999/2000.
- Kleinräumige Bevölkerungsprognose für Wien 2005-2035 (Ko-Autor).
- Verdienstunterschiede zwischen den Bundesländern – eine Folge von Struktureffekten? Statistische Nachrichten, 60(11), 1003-1008.
- Soziale Unterschiede in der Sterblichkeit. Bildungsspezifische Sterbetafeln 2001/2002. Statistische Nachrichten, 62(4), 296-311.
- Soziale Unterschiede in der todesursachenspezifischen Sterblichkeit 2001/2002. Statistische Nachrichten, 62(11), 1010-1022.