Median estimation of chemical constituents for sampling on two occasions under a log-normal model

Athanassios Kondylis*

Product Development, Philip Morris International, Rue des Usines 56, 2000 Neuchâtel, Switzerland

Received 5 May 2014; revised 2 March 2015; accepted 21 March 2015

Sampling from a finite population on multiple occasions introduces dependencies between the successive samples when overlap is designed. Such sampling designs lead to efficient statistical estimates, while they allow estimating changes over time for the targeted outcomes. This makes them very popular in real-world statistical practice. Sampling with partial replacement can also be very efficient in biological and environmental studies where estimation of toxicants and its trends over time is the main interest. Sampling with partial replacement is designed here on two occasions in order to estimate the median concentration of chemical constituents quantified by means of liquid chromatography coupled with tandem mass spectrometry. Such data represent relative peak areas resulting from the chromatographic analysis. They are therefore positive-valued and skewed data, and are commonly fitted very well by the log-normal model. A log-normal model is assumed here for chemical constituents quantified in mainstream cigarette smoke in a real case study. Combining design-based and model-based approaches for statistical inference, we seek for the median estimation of chemical constituents by sampling with partial replacement on two time occasions. We also discuss the limitations of extending the proposed approach to other skewed population models. The latter is investigated by means of a Monte Carlo simulation study.

Keywords: Chemical compounds; Composite median; Model-assisted inference; Regression estimator; Sampling with partial replacement.

1 Introduction

Modern instrumentation, such as chromatography coupled with tandem mass spectrometry, and advances in analytic chemistry have substantially improved the identification and quantification of chemical constituents in real-life samples. These include, for instance, aquatic samples from water resources, samples from marketed food products, as well as air samples from an industrial site. Samples need to be selected for their chemical constituents to be identified and quantified. This is essential for toxicological assessment and related environmental purposes. Such studies focus on estimating a statistic of interest, such as the average or a quantile of the distribution of the chemical constituents.

Sample surveys are conducted in order to get reliable estimates for a statistic of interest, such as the average level of a chemical compound in a product, by studying a small number of products in the market. The ensemble of products in a market is called the population. Analyzing all elements in a population is known as census and, if not practically impossible, it may be extremely expensive. Sample surveys use probability sampling techniques to select samples, these are denoted by \( s \). The sample size is assumed fixed and is denoted by \( n \); this may be much smaller than the population size \( N \). The characteristic of interest (\( Y \)) is measured on the units selected in the sample in order to draw conclusions on the population level. This is done by means of statistical inference. It mainly includes

*Corresponding author: e-mail: thanasios.kondylis@pmi.com, Phone: +41-58-2423069
point estimation and confidence intervals. The width of the confidence intervals reflects the uncertainty on the provided estimates and depends upon the sampling design. Two different approaches are used in survey sampling. The first one is the design-based approach. According to this approach, variation in the outcome is exclusively due to sampling. Stochasticity, therefore, stems from the sampling design, denoted as $p(s)$, which defines the probability for sample $s$ to be selected. The most commonly used sampling designs include simple random sampling with or without replacement, stratified random sampling, sampling proportional to size, and cluster sampling (see Särndal et al., 1992). The second approach assumes that the population values are generated from a superpopulation. In this model-based approach, the variation in the outcome stems from the fact that $Y$ is itself a random variable following a probability distribution model.

Surveys are sometimes designed on two or more occasions. The reason for this is either the need to follow-up units to detect trends and changes in the population over time, or the fact that given the available resources, the sample selected in any one occasion is not by itself sufficient to provide reliable estimates. Examples of surveys on multiple occasions include maritime and forest surveys for environmental monitoring. A large amount of literature is available in this domain, using either design-based or model-based approach (see Duncan and Kalton, 1987; Fuller, 1990; Urquhart and Kincaid, 1999; as well as Steel and McLaren, 2008, and the references therein). Depending on the goal of the survey overlap between the successive occasions may be included. If there is no overlap, each sample is selected independently in the first and second time occasions, respectively. The population units included in the samples are selected only once, either on the first or second time occasion. When all the units selected in one time occasion are kept across time, there is a complete overlap. In the latter case, interest lies on estimating changes in the population over time. These two schemes are two extremes. One may decide to replace some of the selected units from one occasion with new ones. In such cases, a fraction of samples are kept in the study across time, while freshly sampled units are added on each time occasion. This sampling is known as sampling with partial replacement, and it has been presented in Patterson (1950) and further elaborated in Newton et al. (1969, 1974), Scott (1981), Scott and Köhl (1994), Raj (1965), and Särndal et al. (1992). The estimates resulting from such designs are called composite estimates.

Interest lies here on composite estimation for the median while sampling with partial replacement on two time occasions. The median for the characteristic of interest $Y$, denoted as $Med(y)$, corresponds to the $Y$-value that splits the population on two equal halves. The median equals to the mean when the characteristic of interest is symmetrically distributed. Yet, when the characteristic of interest is skewed distributed the median is preferred to the mean. This is the case for most chemical and biological data; they are positive-valued and skewed. Such data are usually fitted well by the log-normal distribution (see Limpert et al., 2001). We propose a composite median estimator for chemical constituents by sampling with partial replacement on two time occasions, assuming a log-normal model for the outcome. We combine thus design-based and model-based approaches. The basic statistical features of the log-normal model are provided in Section 2. The sampling design and composite median estimator are presented in Section 3. The proposed estimator is used for median estimation of chemical constituents quantified in the mainstream cigarette smoke of brands sold in an Asian market. The sample survey was conducted on two occasions, in 2010 and 2012. Details and results on the survey are given in Section 4. A Monte Carlo simulation is used to investigate the performance of the proposed estimator in Section 5. Discussion and conclusions are finally given in Section 6.

## 2 Log-normal model

A random variable $Y$ is log-normally distributed, if its logarithm follows a normal distribution, that is:

$$Y \sim LN(\mu(y), \sigma(y)) \text{ if } \log(Y) \sim N(\mu_{\log}(y), \sigma_{\log}(y)).$$
The Greek letters $\mu$ and $\sigma$ denote the mean and square root of the variance of $Y$, respectively. The log subscript is used hereafter to highlight parameters in the logarithmic scale, while no subscript indicates parameters in the original scale. Log-normal data are skewed to the right and fit well positive-valued data. The mean $\mu(y)$ and median $Med(y)$ for $Y$ are given by:

$$\mu(y) = \exp\left\{ \mu_{\log}(y) + \frac{1}{2} \sigma_{\log}^2(y) \right\}$$

and

$$Med(y) = \exp\left\{ \mu_{\log}(y) \right\},$$

accordingly. It is easy to verify from Expressions (1) and (2) that within the log-normal model

$$Med(y) \leq \mu(y),$$

which is a common finding in positively skewed data. The median corresponds to the $q$-th quantile, denoted as $Q_q(y)$, for $q = 0.5$. The $q$-th quantile of a log-normally distributed $Y$ is given by:

$$Q_q(y) = \exp\left\{ \mu_{\log}(y) + t_q \sigma_{\log}(y) \right\},$$

where $t_q$ is defined such that $q = \Pr(Z \leq t_q)$ for $Z$ having a standard normal distribution. We have

$$Med(y) = Q_{0.5}(y).$$

Log-normal theoretical properties and its various applications are presented in Crow and Shimizu (1988). A very good overview on the use of log-normal models in biosciences is given in Limpert et al. (2001). The use of the logarithm transformation in biostatistics is well illustrated in Rousson (2008). The use of the log-normal distribution in survey sampling is mostly found in the model-based framework (see, for instance, Karlberg, 2000; Chambers and Dorfman, 2003). This is due to the fact that design-based survey practitioners assume that variation is only due to sample selection, not from variation in the outcome. Yet, in many cases, introduction of information on the distribution of $Y$ may lead to more efficient estimates. Consider, for instance, the case where $Y$ results from instrumental readings quantifying chemical concentrations. Chemical concentrations are typical examples of positive-valued data that are very often ratios. They result from dividing integrated chromatographic peaks of targeted chemical compounds by a reference peak. Therefore, chemical data are commonly skewed and are well fitted by the log-normal model. This favors a combined approach for the median estimation, which includes design-based estimator under a log-normal model for the outcome.

A very convenient feature of the log-normal model is that it allows working within the normal distribution framework. This is achieved by using the log transformation. Working on a logarithmic scale normalizes the data distribution and stabilizes their variance. Moreover, the median estimate results directly by transforming back in the original scale the mean estimated in the logarithmic scale, see Expression (2). Note that

$$\mu_{\log}(y) = Med_{\log}(y) = \log[Med(y)],$$

with the first equality on the left side coming from the fact that $Y$ follows a log-normal model; therefore, the mean and median are identical in the logarithmic scale. Note that the second equality on the right side in (4) stems directly from the invariance property of quantiles.

Confidence intervals for the median estimate may be obtained either after back-transformation of the endpoints of the log scale based confidence interval for $\mu_{\log}(y)$, or by using a normal approximation.
for the median estimator and Taylor expansion. In the latter case, the standard deviation of \( Y \) in the logarithmic scale (\( \sigma_{\log}(Y) \)) is approximately equal to the coefficient of variation of \( Y \) in the original scale, that is, \( CV(Y) \). To see that, use a Taylor series expansion of \( f(y) = \log(y) \) around its mean value, say \( \alpha \), to approximate \( f(y) \) as:

\[
  f(y) \approx f(\alpha) + f'(\alpha)(y - \alpha),
\]

with \( f'(z) \) denoting the derivative of function \( f \) at \( z \). Rearranging and taking expectation on both sides we get:

\[
  E(f(y) - f(\alpha))^2 \approx \frac{1}{\alpha^2} E(y - \alpha)^2.
\]

That is:

\[
  \sigma^2(f(y)) \approx \left( \frac{\sigma(y)}{\mu(y)} \right)^2 = CV(y)^2.
\]

and the standard deviation for log-transformed data is approximately equal to the coefficient of variation for data in the original scale. The approximation will work better for large sample sizes; the larger the sample size, the better the approximation will be. Taylor series methods for statistical inference are described in Casella and Berger (2002, Chapter 5). The use of Taylor series methods for variance estimation in sample surveys is described in Wolter (1985, Chapter 6).

3 Sampling design and median estimation

3.1 Sampling with partial replacement on two occasions

Consider sampling on two occasions from a finite population \( U \) of size \( N \). We assume that the population is composed of the same elements in the two occasions. Following notation from Särndal et al. (1992), we use \( X \) to denote the variable of interest in the population when measured at the first occasion, and \( Y \) when measured at the second time occasion. A sample \( s_a \) is drawn from the population at time 1 with sampling design \( p_a(.) \). The sample size at time 1 equals \( n_a \). At the second time occasion, complete overlap, no overlap, or partial overlap may be considered depending on the survey goal. Sampling with partial replacement implies partial overlap between the two samples. Consequently, at time 2, the following two samples are drawn:

(i) The matched sample \( s_m \) of size \( n_m \) selected from \( s_a \) with sampling design \( p_m(.|s_a) \).

(ii) The unmatched sample \( s_u \) of size \( n_u \) drawn from the complement sample \( s'_a = U \setminus s_a \) with sampling design \( p_u(.|s'_a) \).

At the second occasion, \( s = s_m \cup s_u \) and the total sample size is \( n = n_m + n_u \). The probability sampling designs for samples \( s_m \) and \( s_u \) at the second occasion are conditional on \( s_a \) and \( s'_a \), respectively. Note that the subscripts \( a, m, u \) are used hereafter to indicate the sample from which estimates are derived. We employ the \( Var(\hat{\theta}) \) notation to denote the variance for an estimator, for instance \( \hat{\theta} \). The covariance between two estimators, say \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \), is denoted by \( Cov(\hat{\theta}_1, \hat{\theta}_2) \). Their estimates are then denoted using the hat notation.

When sampling on more than one occasion, composite estimators are commonly used. They have been introduced in Patterson (1950). Let \( \hat{\theta}_r(.) \) generally denote an estimator of a statistic of interest \( \theta \).
computed on sample \( s_r \), \( r \in \{a, m, u\} \). Under a general sampling design, the composite estimator of \( \theta \) for sampling on two occasions has the following form:

\[
\hat{\theta} = w_1 \hat{\theta}_1 + w_2 \hat{\theta}_2,
\]

where the coefficients \( w_1 \) and \( w_2 \) are nonnegative weights that satisfy \( w_1 + w_2 = 1 \) and will be discussed in what follows. The estimator \( \hat{\theta}_1 \) in Expression (5) relies uniquely on \( s_u \), and it is given as:

\[
\hat{\theta}_1 = \hat{\theta}_u(Y).
\]

Let \( \bar{Y} = BX \), for \( B \) a known constant, and \( \Delta = Y - \bar{Y} \). The estimator \( \hat{\theta}_1 \) in (5) combines both \( s_m \) and \( s_u \), and it is given according to

\[
\hat{\theta}_1 = \hat{\theta}_u(\bar{Y}) + \hat{\theta}_m(\Delta) \quad \text{with} \quad \hat{\theta}_m(\Delta) = \hat{\theta}_m(Y) - \hat{\theta}_m(\bar{Y}).
\]

The estimator \( \hat{\theta}_1 \) combines what is learned from \( s_u \) updated on time 2, through the constant shift \( B \), plus an estimator on the difference (\( \Delta \)) of the outcome between the first and second occasions as evaluated in \( s_m \). The term \( \hat{\theta}_m(\Delta) \) in (6) provides the estimator for \( \theta \) on time occasion 2 based on the units sampled in \( s_m \). The terms \( \hat{\theta}_u(\bar{Y}) \) and \( \hat{\theta}_m(\bar{Y}) \) represent an approximation of \( \hat{\theta}_u(X) \) and \( \hat{\theta}_m(X) \) in the second time occasion, respectively. They are equal to

\[
\hat{\theta}_u(\bar{Y}) = B \cdot \hat{\theta}_u(X) \quad \text{and} \quad \hat{\theta}_m(\bar{Y}) = B \cdot \hat{\theta}_m(X).
\]

### 3.2 Median composite estimation under a log-normal model

We focus on estimating the composite median for sampling on two occasions using simple random sampling without replacement for \( p_s(.) \), \( p_m(.) \), and \( p_u(.) \). A composite median estimator would be a very difficult task even for a simple random sampling design. Let \( F_y(y) = \frac{1}{N} \#A_y \) with \( A_y = \{k \in U, y_k \leq y\} \) and \#\( A_y \) denoting the number of elements in \( A_y \). Given that the median of \( Y \) is

\[
Med(y) = \inf\{y : F_y(y) \geq 0.5\},
\]

with the \( N \) subscript highlighting a finite population setting, the composite median estimation for sampling on two occasions would require the estimation of several distribution functions. These include \( F_y(x) \), \( F_u(y) \), and \( F_m(z) \), where \( F_z(z) \) generally denotes the estimate of the cumulative distribution function \( F_y(z) \) based on the sample \( s_r \), with \( r \in \{a, m, u\} \), while \( z \) can be either \( X \) or \( Y \).

Up to our knowledge, very limited research on the composite median estimator for sampling on two occasions is available (Rueda et al., 2007; Chadysás, 2009).

We tackle the median estimation problem while working under a log-normal model. This allows us estimating the composite median in a much more convenient fashion. Under the log-normal model and with a simple random sample (SRS) without replacement on each time occasion, the composite median estimator is given by exponentiating the composite mean estimator in the logarithmic scale. The latter will be denoted as \( \hat{\theta}^{\text{log}} \), and is given by Expression (5) with the two terms, \( \hat{\theta}^{\text{log}}_1 \) and \( \hat{\theta}^{\text{log}}_2 \), given as follows:

\[
\hat{\theta}^{\text{log}}_1 = \frac{\log \hat{Y}_m}{\log \hat{X}_m} + B \left( \frac{\log \hat{X}_a}{\log \hat{X}_m} - \frac{\log \hat{X}_m}{\log \hat{X}_m} \right),
\]

and

\[
\hat{\theta}^{\text{log}}_2 = \frac{\log \hat{Y}_u}{\log \hat{X}_u},
\]

where \( \frac{\log \hat{Y}_m}{\log \hat{X}_m} = \frac{1}{n_m} \sum_{k \in s_m} \log (y_k) \) and \( \frac{\log \hat{Y}_u}{\log \hat{X}_u} = \frac{1}{n_u} \sum_{k \in s_u} \log (y_k) \), respectively. In a similar manner, \( \hat{X}_a^{\text{log}} \) denotes the estimated mean using log-transformed data at the first time occasion based on sample \( s_a \).
Expression (9) is straightforward. Expression (8) is based on Expression (6) with the value of $B$ being a known constant. As mentioned in Särndal et al. (1992, p. 370), the value for $B$ may be suggested by previous studies or subject matter theory. Särndal et al. (1992, p. 373) additionally show that under an SRS the value of $B$ that minimizes the final variance is given by:

$$B = \rho(X, Y) \cdot \frac{\sigma(Y)}{\sigma(X)},$$  

(10)

where $\rho(X, Y)$ is the finite population correlation coefficient between $X$ and $Y$, and $\sigma(X)$ and $\sigma(Y)$ are the finite population standard deviations of $X$ and $Y$, respectively. Expression (10) is based on population terms. When variances, covariance, and correlation are unknown, sample analogues derived from the matched sample $s_m$ are used. It is worth noting that Expression (8) reassembles to a regression estimator where $s_m$ acts as the population and $s_m$ acts as the selected sample from this population.

The variance of the composite estimator, while working with log-transformed data, is given by:

$$\text{Var}(\bar{\theta}^{\log}) = w_1^2 \text{Var}(\bar{\theta}_1^{\log}) + w_2^2 \text{Var}(\bar{\theta}_2^{\log}) + 2 w_1 w_2 \text{Cov}(\bar{\theta}_1^{\log}, \bar{\theta}_2^{\log}).$$  

(11)

The covariance term $\text{Cov}(\bar{\theta}_1^{\log}, \bar{\theta}_2^{\log})$ has the potential to reduce the total variance and may lead to an accurate final estimate. Following Särndal et al. (1992), under simple random sampling without replacement and $B$ given by Expression (10), the variance and covariance terms in (11) are given as:

$$\text{Var}(\bar{\theta}_1^{\log}) = \frac{\sigma_{\log}^2(Y)}{\kappa n_d} \cdot [(1 - \rho_{\log}^2(X, Y)) + \kappa (\rho_{\log}^2(X, Y) - f)],$$  

(12)

$$\text{Var}(\bar{\theta}_2^{\log}) = \frac{\sigma_{\log}^2(Y)}{\kappa n_d} \cdot \frac{\kappa (1 - (1 - \kappa) f)}{1 - \kappa},$$  

(13)

$$\text{Cov}(\bar{\theta}_1^{\log}, \bar{\theta}_2^{\log}) = \frac{\sigma_{\log}^2(Y)}{\kappa n_d} \cdot (-\kappa f) = - \frac{\sigma_{\log}^2(Y)}{N},$$  

(14)

for $f = n_m/N$ the sampling fraction and $\kappa = n_m/n_d$ the fraction of matched data. Moreover, $\sigma_{\log}^2(Y)$ denotes the sample variance of $\log(Y)$, and $\rho_{\log}(X, Y)$ stands for the squared correlation coefficient between $\log(X)$ and $\log(Y)$. Sample variance and correlation estimates replace the population analogues in (12)–(14) to get $\text{Var}(\bar{\theta}_1^{\log})$, $\text{Var}(\bar{\theta}_2^{\log})$, and $\text{Cov}(\bar{\theta}_1^{\log}, \bar{\theta}_2^{\log})$, respectively. These are used to provide an estimate for the variance in (11) denoted as $\text{Var}(\bar{\theta}^{\log})$. Note that the formulae above rely on assuming equal variance on both time occasions. The variance $\sigma_{\log}^2(Y)$ is estimated using data on the second time occasion.

The variance estimate in (11) depends on the weights $w_1$ and $w_2$. Särndal et al. (1992) give expressions for optimal weights that achieve minimum variance (see also Brewer, 2002, p. 236). The optimal weights, denoted as $w_1^*$ and $w_2^*$, are given according to

$$w_1^* = 1 - w_2^* = \frac{\text{Var}(\bar{\theta}_2^{\log}) - \text{Cov}(\bar{\theta}_1^{\log}, \bar{\theta}_2^{\log})}{\text{Var}(\bar{\theta}_1^{\log}) + \text{Var}(\bar{\theta}_2^{\log}) - 2 \text{Cov}(\bar{\theta}_1^{\log}, \bar{\theta}_2^{\log})}.$$  

(15)

Combining design-based and model-based approaches, the final composite median estimator is given according to

$$\hat{\theta} = \exp[\bar{\theta}^{\log}] = \exp\left(\bar{w}_1^{\log} \bar{\theta}_1^{\log} + \bar{w}_2^{\log} \bar{\theta}_2^{\log}\right).$$  

(16)
where \( \overline{\delta}_1 \) and \( \overline{\delta}_2 \) are computed using \( \overline{\text{Var}}(\overline{\delta}_1^{\log}) \), \( \overline{\text{Var}}(\overline{\delta}_2^{\log}) \), and \( \overline{\text{Cov}}(\overline{\delta}_1^{\log}, \overline{\delta}_2^{\log}) \) in (15) instead of the population analogues. Assuming that \( \overline{\delta}^{\log} \) is normally distributed, and for sufficiently large samples, a (1 − \( \alpha \))\% confidence interval for the median \( \hat{\theta} \) is given according to

\[
\text{CI}_{(1-\alpha)\%} = \left( \exp \left( \widehat{\theta}^{\log} - z_{1-\alpha/2} \cdot \sqrt{\overline{\text{Var}}(\widehat{\theta}^{\log})} \right); \exp \left( \widehat{\theta}^{\log} + z_{1-\alpha/2} \cdot \sqrt{\overline{\text{Var}}(\widehat{\theta}^{\log})} \right) \right),
\]

with \( z_\alpha \) denoting the \( \alpha \)-th quantile of the standard normal distribution. Using Taylor series, as described in Section 2, the variance of \( \hat{\theta} \) is estimated as:

\[
\overline{\text{Var}}(\hat{\theta}) = \overline{\text{Var}}(\hat{\theta}^{\log}) \exp \left( \hat{\theta}^{\log} \right)^2,
\]

and an approximate (1 − \( \alpha \))\% confidence interval for \( \theta \) is given according to

\[
\text{CI}'_{(1-\alpha)\%} = \left( \hat{\theta} - z_{1-\alpha/2} \cdot \sqrt{\overline{\text{Var}}(\hat{\theta})}; \hat{\theta} + z_{1-\alpha/2} \cdot \sqrt{\overline{\text{Var}}(\hat{\theta})} \right).
\]

The validity of this approximation into a finite population setting is not, yet, guaranteed. Wolter (1985, Chapter 6) discusses Taylor series methods for variance estimation in finite populations. He states that: at issue is whether the Taylor series converges, and if so, at what range does it converge? adding a few sentences later that: for the finite population model, no such results are possible without also assuming a superpopulation model. Applying Taylor series approximation as in (18) into a finite population setting requires additional assumptions. In particular, one needs to assume that the finite population is an instance of an infinite superpopulation model. As the size of the former goes to infinity, the bivariate distribution for the outcome on the two occasions can be safely approximated by a continuous distribution with marginal densities \( f_x \) and \( f_y \) for the first and second time occasions, respectively. In the simulation study in Section 5, the Taylor series approximation is empirically tested by building confidence intervals around the composite median estimator using both (17) and (18).

### 3.3 Estimating changes over time

Sampling at two or more occasions allows estimating the change in the population of the characteristic under study. This is a special feature of such sampling designs and is due to the inclusion of the matched sample (\( s_y \)) in the sampling design. If one is only interested to estimate changes over time, there would be no need for new sampled units, and the same sampled units would be measured on both or all time occasions. Sampling with partial replacement is a compromise between obtaining new samples to update the median estimate, and using information from the matched sample to check for changes across time.

We follow Särndal et al. (1992, Chapter 9.9.3) and present the main results for estimating the mean difference. Working under a log-normal model, the mean difference is estimated on log-transformed data, and it is denoted as \( \Delta^{\log} \). To estimate the change \( \Delta^{\log} \) between occasions 1 and 2, under a general sampling design, the following statistic is used:

\[
\hat{\Delta}^{\log} = d_1 \hat{\Delta}_1^{\log} + d_2 \hat{\Delta}_2^{\log},
\]

for \( \hat{\Delta}_1^{\log} = \hat{\theta}_1^{\log} - \hat{\theta}_0^{\log}(X) \) and \( \hat{\Delta}_2^{\log} = \hat{\theta}_2^{\log} - \hat{\theta}_0^{\log}(X) \). Expression (19) provides an approximately unbiased estimator for \( \Delta^{\log} \) with approximate variance given by:

\[
\text{Var}(\hat{\Delta}^{\log}) \approx d_1^2 \text{Var}(\hat{\Delta}_1^{\log}) + d_2^2 \text{Var}(\hat{\Delta}_2^{\log}) + 2d_1d_2 \text{Cov}(\hat{\Delta}_1^{\log}, \hat{\Delta}_2^{\log}).
\]
The coefficients $d_1$ and $d_2$ are nonnegative and they sum up to 1. Their optimal values depend upon the variances $\text{Var}(\Delta_1^{\log})$ and $\text{Var}(\Delta_2^{\log})$. Similar to the weights $w_1^*$, $w_2^*$, the optimal coefficients for $d_1^*$ and $d_2^*$ are given according to

$$d_1^* = 1 - d_2^* = \frac{\text{Var}(\hat{\Delta}_2^{\log}) - \text{Cov}(\hat{\Delta}_1^{\log}, \hat{\Delta}_2^{\log})}{\text{Var}(\Delta_1^{\log}) + \text{Var}(\Delta_2^{\log}) - 2 \text{Cov}(\Delta_1^{\log}, \Delta_2^{\log})}.$$  

Under simple random sampling without replacement, and assuming equal variance on both time occasions, that is, $\sigma_{\log}^2(X) = \sigma_{\log}^2(Y) = \sigma_{\log}^2$, the variance and covariance terms in (20) are given according to

$$\text{Var}(\hat{\Delta}_1^{\log}) = \frac{\sigma_{\log}^2}{n_s} \left[ \frac{1}{\kappa} (1 - \nu \rho_{\log}(X, Y)) + 1 - 2 \rho_{\log}(X, Y) - 2f (1 - \rho_{\log}(X, Y)) \right],$$  

$$\text{Var}(\hat{\Delta}_2^{\log}) = \frac{\sigma_{\log}^2}{n_s} \left[ \frac{1}{\nu} + 1 - 2f (1 - \rho_{\log}(X, Y)) \right],$$  

$$\text{Cov}(\hat{\Delta}_1^{\log}, \hat{\Delta}_2^{\log}) = \frac{\sigma_{\log}^2}{n_s} \left[ 1 - \rho_{\log}(X, Y) - 2f (1 - \rho_{\log}(X, Y)) \right],$$

where $f = n_s/N$, $\kappa = n_m/n_s$, as before, while $\nu = 1 - \kappa$. Sample estimates replace the population analogues in (21)-(23) to obtain variance and covariance estimates. The samples $s_m$ on the second occasion and $s_s$ are used to estimate $\sigma_{\log}^2$. The variance and covariance estimates are then plugged into Expression (20) to get the final approximate variance estimate for the change.

The difference in (19) is estimated on the logarithmic scale. The final change estimate results from exponentiating the latter, so that mean differences are back-transformed from the logarithmic scale to the original scale. This shifts the interpretation of the results from absolute changes to relative changes. The latter are symmetric around 1 in the ratio scale. Confidence intervals of the change are given by exponentiating the endpoints of the confidence intervals constructed on the log-transformed data.

### 4 Median estimation for smoke constituents

#### 4.1 Sampling design and modeling aspects

The proposed method for median estimation is illustrated with a real case study. This is a brand survey that has been conducted by Philip Morris International in Asia in 2010 and 2012. The goal of the survey has been to estimate the median yields of the two tobacco-specific nitrosamines: N-nitrosonornicotine (NNN) and N-(N-nitroso-methylamino)-1-(3-pyridyl)-1-butanol (NNK) quantified in mainstream smoke emissions of cigarettes. To estimate the median yields, the market under study has been sampled twice: in 2010 and 2012; sampling with partial replacement has been employed. In 2010, $n_s = 55$ brands have been sampled and analyzed for NNN and NNK. Out of this sample, $n_m = 17$ matched brands have been resampled and analyzed in 2012. In addition to these brands, a new sample of $n_u = 37$ products has been selected in 2012.

The sampling unit in the survey is the brand. This is a cigarette brand name registered and commercialized within a market by a tobacco manufacturer. A brand is accessible in the market through cigarette packs an adult consumer can buy at points of sales. Cigarette packs contain cigarettes of the same dimension (e.g., length and diameter), same blend type, and similar other characteristics, such as filter type, menthol inclusion, etc. The market size is the total number of brands. The selected brands...
Figure 1  NNN concentration in smoke (ng/mg SN) versus NNK concentration in smoke (ng/mg SN) for all sampled brands. NNN and NNK yields are plotted in the original scale (left panel), and in the logarithmic scale (right panel). NNN and NNK yields for brands selected in 2010 are depicted in nonfilled diamonds, while NNN and NNK yields for products sampled in 2012 are depicted using filled diamonds.

Machine-smoking has been replicated three times, so that three analytical results were available for each sampled brand. Aggregated NNN and NNK yields were then used for the median estimation. The relative variation for NNN and NNK raw yields was around 10%. The relative variation between brands based on aggregated yields is 70% for NNN and 50% for NNK. NNN and NNK yields are strongly correlated; their correlation coefficient equals 0.70. Plotting the NNN versus NNK yields (see Fig. 1) provides an empirical justification for using a log-normal model. Yields are strongly heteroskedastic on the original scale. The variation is stabilized in the logarithmic scale while the joint distribution of NNN and NNK is fitted much better by a bivariate normal distribution.

1 Note that three smoking replicates do not necessarily correspond to three cigarette sticks. For each smoking replicate, cigarette sticks are placed in the smoking machine channels and smoked according to specified smoking conditions. The latter are described as a combination of the puff volume, its duration, puff frequency, and ventilation or not of the cigarette filter. The Health Canada smoking conditions include 55 mL puff volume, every 30 s, lasting 2 s each, and no filter ventilation. More information is provided in the document Health-Canada, 1999.
Figure 2  NNN concentration in smoke (ng/mg SN) for matched brands in 2010 and 2012. The best line fit is depicted by the dashed line. The solid line corresponds to the line of equal yields. The graph is drawn in the logarithmic scale. Concentrations are also given in original units on the right-side Y-axis and the top-level X-axis.

4.2 Matched data

The matched sample lists 17 brands produced by different cigarette manufacturers. Their smoke NNN and smoke NNK yields are plotted in Figs. 2 and 3, respectively. Smoke yields for products sold in 2010 are given on the X-axis, while yields for products sold in 2012 are given on the Y-axis. Smoke NNN and smoke NNK are measured in nanograms per milligram of smoke nicotine (ng/mg SN) and are plotted in the logarithmic scale. The quantified yields are also reported in the original scale (see the top and right axes labels).

The solid lines in Figs. 2 and 3 correspond to the line of equal yields in 2010 and 2012. Points lying on these solid lines correspond to brands for which smoke constituents did not change between 2010 and 2012. The dashed lines correspond to the linear fits of the 2010 yields regressed on the 2012 yields. They allow observing shifts between the 2 years. For instance, Fig. 2 shows an upward shift in NNN yields in 2012 compared to 2010 for the matched brands. This is observed for low as well as for high NNN levels. This is not the case for NNK as seen in Fig. 3. Smoke NNN yields in 2010 and 2012 are strongly correlated. The estimated correlation coefficient is equal to 0.97. The corresponding estimated correlation coefficient for NNK is 0.45.

When sampling on two occasions is designed, the correlation coefficient of the targeted outcome between the two events plays a very important role. The larger the correlation coefficient, the larger is the gain in the precision of the final estimates. This is especially true for the estimates of change, as seen from Expressions (21)–(23), where the correlation coefficient comes with a negative sign. Expression (12) shows that a high correlation coefficient may lead to a substantial gain in precision, even when the fraction of matched data (κ) is small.
Figure 3  NNK concentration in smoke (ng/mg SN) for matched brands in 2010 and 2012. The best line fit is depicted by the dashed line. The solid line corresponds to the line of equal yields. The graph is drawn in the logarithmic scale. Concentrations are also given in original units on the right-side $Y$-axis and the top-level $X$-axis.

Table 1  NNN and NNK composite median estimates together with the 2010 and 2012 simple random sample median estimates.

| Expression (18) | NNN Lower | Estimate | Upper | NNK Lower | Estimate | Upper |
|-----------------|-----------|----------|-------|-----------|----------|-------|
|                 | 70.79     | 78.87    | 86.96 | 43.90     | 49.12    | 54.35 |
| Expression (17) | 71.19     | 78.87    | 87.38 | 44.17     | 49.12    | 54.63 |
| Expression (7) for 2010 | 59.39 | 74.45    | 81.86 | 43.44     | 44.95    | 53.06 |
| Expression (7) for 2012 | 69.78 | 81.87    | 90.67 | 46.74     | 60.08    | 64.28 |

Note: All estimates are reported together with their associated 95% lower and upper confidence limits.

4.3 Median and change estimates

Table 1 provides the composite median estimate for NNN (left) and NNK (right) together with the associated 95% lower and upper confidence limits. The second row in Table 1 reports median estimate and confidence limits resulting from directly exponentiating the endpoints of the 95% confidence interval for the mean in the logarithmic scale (see Expression (17)). The first row in Table 1 reports the composite median estimate and the associated 95% lower and upper confidence limits using a normal approximation with standard deviation estimated via the coefficient of variation (see Expression (18)). The difference between the two is rather small; most probably due to the relatively large sample sizes.
on both years, 2010 and 2012. Furthermore, Table 1 provides the SRS median estimates, for 2010 and 2012, resulting from inverting the empirical cumulative density function at 0.5 (see Expression (7)). Estimates are given for each time occasion separately. Lower and upper confidence limits are estimated following Särndal et al. (1992, paragraph 5.11).

The final composite median estimates have around 10% error for both NNN and NNK. The relative error for NNN equals 10.3% while for NNK it is 10.6%. This relative error results from dividing the interval half-length by the median level. The 2010 and 2012 simple random sample median estimates for NNN have 15.2% and 17.1% relative errors, respectively. This is far larger than the 10.3% relative error of the composite median estimate. The 2010 and 2012 simple random sample median estimates for NNK have 10.6% and 14.6% relative errors, respectively. The latter is larger than the 10.6% error of the composite median estimate. The former, that is the 2010 simple random sample median estimate, is of equal precision to the composite median estimate.

Composite estimation benefits from designing sample overlap and matching data. The gain in precision is expected to be larger when the correlation of the studied variable between the two occasions is large; this is the case for NNN as compared to NNK. The larger gain for NNN compared to NNK is justified by the fact that the correlation coefficient between NNN yields across time equals 0.97, while for NNK it is equal to 0.45.

Designing surveys on two time occasions with sample overlap allows estimating changes across time. Changes on the NNN and NNK market medians across time are estimated according to the results given in Section 3.3. The estimate for the change in smoke NNN median concentration for 2012 relative to 2010 is 1.19. The associated 95% confidence interval is [1.12, 1.27]. For smoke NNK, the change is estimated at 1.07, and the associated 95% confidence interval is [0.93, 1.23]. The results for the NNN and NNK 2012 over 2010 yield ratio are plotted in Fig. 4.

Despite the increase in NNN and NNK yields in 2012 relative to 2010, the observed relative changes are on average less than 1.25. The uncertainty associated with the observed estimates is large; especially for NNK. This is partly due to the fact that both NNN and NNK in smoke correlate with the NNN and NNK concentration in tobacco. The concentration in tobacco generally depends on the tobacco type, climatic conditions, and agricultural practices.

5 Monte Carlo simulation study

5.1 Simulation setting

The development of the proposed composite median estimator assumes a log-normal distribution for the superpopulation model. In this framework, working on the logarithmic scale transforms data to normal and facilitates the median estimation problem. This is achieved by transforming back in the original scale the mean obtained in the logarithmic scale. The variance of the estimator in the logarithmic scale is expected to impact the properties of the median estimator, especially for large values of $\text{Var}(\hat{\theta})$. The latter, as seen in Section 3.2, depends strongly on the variance $\sigma_{log}(y)$, as well as the sampling fractions $f, \kappa$, and the correlation $\rho_{log}(X, Y)$. Given that $\sigma_{log}(y) \approx CV(y)$, and
the latter impacts the skewness of the log-normal distribution, a limited Monte Carlo simulation is performed to control the properties of the proposed estimator employing different skewness levels under the log-normal model.

We simulate from a log-normal population with size equal to \( N = 500 \). The population is generated from a bivariate log-normal random variable centered at 50 on both dimensions (time occasions); no change has been designed between the time occasions. The simulation study allows different skewness levels by varying the coefficient of variation from 15% to 100%. This is achieved by properly defining the diagonal elements of the variance-covariance matrix for the generated bivariate distribution. The off-diagonal elements of this matrix determine the correlation coefficient of the result between the two occasions; this is set to \( \rho_{xy} = 0.85 \). Two positively skew distributions are also used in the simulations; a gamma and a chi-square (\( \chi^2 \)). The gamma distribution has shape and scale parameters equal to 20 and 0.2, respectively. The \( \chi^2 \) distribution is generated with degrees of freedom equal to 20. No change is designed between the two time occasions. The correlation coefficient \( \rho_{xy} \) is set equal to 0.85, as before.

For each generated population \( U \), we use \( R = 20,000 \) runs following the steps described below:

(i) a simple random sample without replacement is drawn from \( U \); its size equals \( n_a = 100 \),
(ii) a simple random sample without replacement is selected from \( n_a \); we let two sample sizes, that is, \( n_a = 30 \) and \( n_m = 50 \),
(iii) a simple random sample without replacement is selected from the complementary set of \( U \); its sample size equals \( n_u = n_a - n_m \), so that \( n_u + n_m = 100 \).

At the end of the \( R = 20,000 \) runs, we compute:

(i) the relative bias of \( \hat{\theta} \):

\[
\text{RB}(\hat{\theta}) = \frac{1}{\theta} \left\{ \frac{1}{R} \sum_{j=1}^{R} (\hat{\theta}^j - \theta) \right\},
\]

where \( \hat{\theta}^j \) is the median estimate at simulation run \( j \), and \( \theta \) is the known median value in \( U \),
(ii) the relative mean squared error of \( \hat{\theta} \):

\[
\text{RMSE}(\hat{\theta}) = \frac{1}{\theta} \left\{ \frac{1}{R} \sum_{j=1}^{R} (\hat{\theta}^j - \theta)^2 \right\}^{1/2},
\]

(iii) the empirical coverage rate of a 95% nominal-level confidence interval, given by:

\[
\text{ECR}_{95\%}(\hat{\theta}) = \frac{1}{R} \left\{ \# \left[ \theta \in CI_\theta^j(95\%) \right] \right\},
\]

where \( CI_\theta^j(95\%) \) is the 95% confidence interval for \( \theta \) constructed at simulation run \( j \). Expression (26) computes the proportion of the true population median being included in the constructed 95% confidence intervals.

\[ \textbf{5.2 Simulation results} \]

The simulation results are given in Table 2. For the different simulation parameter combinations (in rows), Table 2 reports the relative bias and root mean square error of the proposed composite median estimator, as well as the empirical coverage rate of the derived 95% confidence intervals. The corresponding empirical confidence interval coverage rates for both the symmetric (see Expression (18)) and direct (see Expression (17)) estimators are reported in columns 7 and 8 of Table 2. The bias and root mean square error results of a nonparametric composite median estimator, presented in Rueda et al. (2007), are finally reported in the last two columns in Table 2.
Table 2  Simulation results: relative bias (100 × RB), root mean square error (RMSE), and empirical coverage rates for the 95% confidence intervals of the proposed median estimator under a log-normal, a gamma, and a χ² model for varying matched sample sizes (nₘ) and skewness levels (given via the skewness coefficient γ). The coefficients of variation are reported in the third column (CV).

| Model  | γ   | CV(%) | nₘ | 100 × RB | RMSE    | ECR  | ECR* | 100 × RB_R | RMSE_R |
|--------|-----|-------|----|----------|---------|------|------|------------|--------|
| Log-normal | 0.5 | 15    | 30 | −0.032   | 0.016   | 0.957| 0.956| 0.226      | 0.114  |
|         | 0.5 | 50    | −0.04| 0.02    | 0.96    | 0.958| 0.028| 0.105      |        |
|         | 1   | 30    | −0.196| 0.097  | 0.947   | 0.944| 0.19 | 0.194      |        |
|         | 1   | 50    | −0.174| 0.087  | 0.954   | 0.951| 0.141| 0.837      |        |
|         | 1.5 | 50    | −0.642| 0.294  | 0.937   | 0.931| 0.145| 0.164      |        |
|         | 1.5 | 50    | −0.372| 0.17   | 0.949   | 0.943| 0.144| 0.152      |        |
|         | 3   | 100   | −0.634| 0.29   | 0.934   | 0.928| 1.176| 0.530      |        |
|         | 3   | 50    | −0.393| 0.18   | 0.949   | 0.944| 1.102| 0.692      |        |
| Gamma   | 0.5 | 30    | −2.595| 1.22   | 0.836   | 0.821| −0.61| 0.287      |        |
|         | 0.5 | 50    | −2.607| 1.226  | 0.857   | 0.847| −0.807| 0.38       |        |
| χ²     | 0.75| 30    | −0.906| 0.173  | 0.934   | 0.926| 1.13 | 0.609      |        |
|         | 0.75| 50    | −0.791| 0.151  | 0.952   | 0.946| 1.076| 0.603      |        |

Notes: The ECR and ECR* columns report the empirical coverage rates for the 95% nominal-level confidence intervals of the composite median and the symmetric composite median estimator, respectively. The bias and the RMSE results indexed by R report the bias and the precision for the composite median estimator presented in Rueda et al. (2007).

The proposed estimator is negatively biased. The amount of the observed bias is generally small as long as we are under a log-normal population, and increases for larger skewness levels. A similar trend is observed for the root mean square error, as well. To reduce the bias and improve the precision, one should design a larger matched sample. The bias and root mean square error increase when the population model is not log-normal. The amount of bias and root mean square error are reasonable for a χ² population, yet, for the gamma population they both increase substantially. In such cases, the use of a nonparametric composite median estimator, such as the one presented in Rueda et al. (2007), would reduce the bias and root mean square error (see the last two columns in Table 2). Comparing the proposed estimator to the composite median estimator in Rueda et al. (2007) confirms that under a log-normal population model, the two estimators perform similarly in terms of bias. The root mean square error of the quantile estimator in Rueda et al. (2007) is often larger.

Columns seven and eight in Table 2 suggest that under a log-normal model, the empirical coverage rates tend toward the nominal level for increasing nₘ regardless the skewness of the population. For all log-normal populations and all coefficients of variation, the nominal 95% confidence level is almost reached for nₘ = 50. For smaller matching (nₘ = 30), the empirical coverage rates lie below the 95% nominal level for coefficients of variation equal or above 50%. The empirical coverage rate resulting from the symmetric median estimator (see Expression (18)) are comparable to the empirical coverage rate of the direct median estimator given by Expression (17). Their difference is not impacted either by the skewness of the distribution or the matching fraction. The empirical coverage rates resulting from a χ² population are not far from the nominal level (this is almost reached for nₘ = 50), yet, for a gamma population the resulting empirical coverage rates are far below the 95% nominal level.

This is a limited simulation study and its results are also limited to this restricted set of simulation parameters under investigation. For larger sample sizes, large sampling fractions, and stronger correlations, one would expect better performance for the median estimator, even for small departures from the log-normal distribution. This would not be the case, of course, for small sample sizes, moderate sampling fractions, and weak correlations.
6 Discussion and conclusions

The motivation behind the proposed composite estimator for the median comes from a real brand survey. The objective of the survey was to estimate the market median for two tobacco-specific nitrosamines, NNN and NNK, in mainstream cigarette smoke in an Asian market. Sampling with partial replacement on two occasions has been employed to estimate the market median, and cigarette brands have been sampled and analyzed in 2010 and 2012. Designing the brand survey on two occasions allowed us to efficiently plan and conduct the brand survey, while obtaining median estimates updated for changes between the years. The resulting composite estimate for NNN was much more precise than the individual-years simple random sample median estimates. For NNK, the corresponding gain in precision has been less, due to the small correlation of the NNK levels between the years.

The quantification of the NNN and NNK in the cigarette's mainstream smoke was done by means of LC-MS/MS. The data resulting from such instruments are commonly skewed, and are fitted very well by the log-normal distribution; this has been justified in our data, as well. Indeed, dealing with log-normal data is a common practice in analytical chemistry. Working under a log-normal model allowed us working on a logarithmic scale in a very convenient fashion: normality has been assumed on the logarithmic scale rather than the original scale, and the composite estimator for the median and its variance resulted directly from known theoretical results for the mean estimator. Models other than the log-normal do not provide an estimator for the median, therefore, the log-normality assumption should be scrutinized before applying the presented composite estimator.

The bias and precision of the proposed estimator has been tested through a limited Monte Carlo simulation study using different superpopulations and different design strategies. The former included log-normal populations of different skewness. The latter focused on the matched sampling fraction. The simulation results confirmed the importance of the matched sampling fraction on the efficiency of the composite estimator for the median in terms of relative bias and root mean square error. Under the log-normal model, the proposed median estimator performed rather well even under highly skewed populations. The empirical coverage rate of the 95% nominal-level confidence intervals remained close to their nominal level. For highly skewed populations and small matching between the two occasions, the observed coverage rates were slight below the 95% nominal level. The provided simulation study and presented results are limited. Different sampling strategies in terms of sampling fractions, as well as different correlations between the two time occasions, will undoubtedly influence the performance of the median estimator, in the same manner this would influence the composite estimator for the mean when sampling on two occasions. The larger the sampling fractions and the stronger the correlation between the two time occasions, the better the estimator will perform; this is expected to be the case even for slight departures from the log-normal distribution.

Variance estimation focused mainly on the sampling variance; the measurement variance resulting from analytical variability has not been explicitly taken into account. The use of repeated measurements provides the means to estimate and include measurement variance on the total variance estimation. This is undoubtedly a subject for further investigation. Throughout the analysis we have assumed that the total number of brands in the market has not changed between 2010 and 2012. The observed market changes across these 2 years have been limited, at least in terms of the main structural features of the sampled market, such as blend types and cigarette designs used in the products. We focused on median estimation of toxicants present in mainstream cigarette smoke. The estimation of quantiles other than the median, especially for higher quantiles, can also be very interesting for this kind of surveys. Expression (3) in Section 2 provides the means to directly estimate quantiles other than the median under a log-normal model. Yet, this was out of the focus of the present work. For estimating quantiles under sampling on two occasions under general sampling designs, see Rueda et al. (2007).

Note, finally, that optimally designing sampling with partial replacement is not an easy task. It includes selecting three samples, $s_a$, $s_m$, and $s_u$, with special focus given on $s_m$ and its size. The latter will definitely impact the gain in precision by employing a composite estimator. Available prior knowledge on the correlation for the characteristic of interest across time will largely improve the
sampling design. Yet, when more than one characteristics of interest are targeted, no one sampling
design can be optimal for all outcomes under study. Finally, one needs to take also into account other
cost restrictions, making the design of such studies very challenging.

Acknowledgments The author would like to thank the Associate Editor and a referee for their comments and
suggestions that greatly improved the quality of this manuscript. The author would also like to thank Alina Matei
for fruitful discussions on survey sampling. This research would not be possible without the support of Ferruccio
Gadani and Gebhard Durner, to whom the author is very thankful.

Conflict of interest
The authors have declared no conflict of interest.

References
Brewer, K. R. W. (2002). Combined Survey Sampling Inference. Oxford University Press, New York, NY.
Casella, G. and Berger, L. R. (2002). Statistical Inference. Duxbury Advanced Series, New York, NY.
Chadyšas, V. (2009). Estimation of a distribution function under sampling on two occasions. Numerical Analysis:
Modeling and Control 14, 315–331.
Chambers, R. L. and Dorfman, A. H. (2003). Transformed variables in survey sampling. S3RI Methodology
Working Paper M03/21. Southampton Statistical Sciences Research Institute, University of Southampton.
Crow, E. L. and Shimizu, K. (1988). Log-normal Distributions: Theory and Application. Dekker, New York, NY.
Duncan, G. J. and Kalton, G. (1987). Issues of design and analysis of surveys across time. International Statistical
Review 55, 97–117.
Health-Canada (1999). Official Method T-115 – Determination of “Tar”, Nicotine and Carbon Monoxide
in Mainstream Tobacco Smoke, Tobacco Control Programme, Modified:2007-11-28 ed. Health Canada,
Ottawa, Canada.
Fuller, W. A. (1990). Analysis of repeated surveys. Survey Methodology 16, 167–180.
Karlberg, F. (2000). Population total prediction under a lognormal superpopulation model. Metron 58, 53–80.
Limpert, E., Stahel, W. A. and Abbt, M. (2001). Log-normal distributions across the sciences: keys and clues.
BioScience 51, 341–452.
Newton, C. M., Cunia, T. and Bickford, C. A. (1969). Sampling with partial replacement on three or more
occasions. Forest Science 15, 205–224.
Newton, C. M., Cunia, T., and Bickford, C. A. (1974). Multivariate estimators for sampling with partial replace-
ment on two occasions. Forest Science 20, 107–116.
Patterson, R. D. (1950). Sampling on successive occasions with partial replacement of units. Journal of the Royal
Statistical Society, Series B 12, 241–255.
Raj, D. (1965). On sampling over two occasions with probability proportionate to size. Annals of Mathematical
Statistics 36, 327–331.
Rousson, V. (2008). On the logarithm transformation in biostatistics. Seminar in Pharmacology. Institute of Social
and Preventive Medicine, CHUV, Lausanne, CH.
Rueda, M., Munoz, J., Gonzalez, S. and Arcos, A. (2007). Estimating quantiles under sampling on two occasions
with arbitrary sample design. Computational Statistics & Data Analysis 51, 6596–6613.
Särrndal, C. E., Swensson, B. and Wretman, J. (1992). Model Assisted Survey Sampling. Springer Series in Statistics,
New York, NY.
Scott, C. T. (1981). Simplified Estimators for Sampling with Partial Replacement on Multiple Occasions. Paper
Series 23. Department of Forest Resources, University of Minnesota, Minneapolis, MN.
Scott, C. T. and Köhl, M. (1994). Sampling with partial replacement and stratification. Forest Science 40, 30–46.
Steel, D. and McLaren, C. (2008). Design and Analysis of Repeated Surveys. Paper Series 11-08. Centre for
Statistical and Survey Methodology, University of Wollongong, AU.
Urquhart, S. N. and Kincaid, T. M. (1999). Designs for detecting trend from repeated surveys of ecological
resources. Journal of Agricultural, Biological, and Environmental Statistics 4, 404–414.
Wolter, K. M. (1985). Introduction to Variance Estimation. Springer Series in Statistics, New York, NY.