Efficient Design and Analysis of Biospecimens with Measurements Subject to Detection Limit

Albert Vexler[*1], **Aiyi Liu**[1], and **Enrique F. Schisterman**[1]

[1] Division of Epidemiology, Statistics and Prevention Research, National Institute of Child Health and Human Development, NIH/DHHS, 6100 Executive Blvd., Rockville, MD 20852, U.S.A.

Received , revised
Published online

Summary

Pooling biospecimens is a well accepted sampling strategy in biomedical research to reduce study cost of measuring biomarkers, and has been shown in the case of normally distributed data to yield more efficient estimation. In this paper we examine the efficiency of pooling, in the context of information matrix related to estimators of unknown parameters, when the biospecimens being pooled yield incomplete observations due to the instruments' limit of detection. Our investigation of three sampling strategies shows that, for a range of values of the detection limit, pooling is the most efficient sampling procedure. For certain other values of the detection limit, pooling can perform poorly.

Key words: Detection limit; Information matrix; Pooling design; Random sampling; Truncated/Censored data.

1 Introduction

Often in epidemiological or environmental studies dealing with a large population or expensive measurements of biomarker assays the laboriousness or the cost of a study can be reduced by examining only a sample of the population. One of the sampling strategies is the well-accepted pooling design (e.g. Dorfman, 1943; Sterrett, 1957; Weinberg and Umbach, 1999). The basic idea is, under homogeneity assumption about the population, to pool together biological samples (e.g., urine, sera, or plasma) from a number of individuals and then to represent the data by measured assays of the synthetic personalities. As a sampling strategy that avoids ignoring individual biological samples, under certain parametrical models of the measurements, pooling design can have high efficiency (e.g. Faraggi, *et al.*, 2003). Thus, if N is the total size of the population, by applying the pooling sampling strategy, we collect only $n \equiv N/p$ measured biological samples from the synthetically pooled specimens, where each pool has the same number p of individuals. Although the strategy of pooling specimens has been used in practice, methods for analysis of set-based data from such experiments have not been fully and well developed in the lite rature, except for certain special cases (normally distributed data or in the logistic regression context). This is, perhaps, partly because for a general distribution of individual biomarker values, the likelihood methods based on the pooled data may not be feasible, since the distribution of the pooled biomarker values involves the complex convolution of p individual biomarker values.

The sample size limitation is not the only restriction in biomarker development and evaluation. Instrument sensitivity is another aspect due to which a portion of study participants have levels at or below some experimentally determined detection limit and hence can not be observed; the value of the biomarker can be completely observed only if this value is not below the detection threshold. Biomarker quantification may be compromised if instrumentation cannot detect low levels. Although the problem can be considered as a special case of fragmentary sampling and truncated/censored data (e.g. Wilks, 1932; Gupta, 1952;

* Corresponding author: e-mail: vexlera@mail.nih.gov, Phone: +301 435 6944, Fax: +301 402 2084

Cohen, 1955; Chapman, 1956 etc.), due to the importance of instrument sensitivity in many areas such as occupational medicine and epidemiology, the topic of limit of detection has been extensively dealt with in the biostatistical literature with recent examples including Finkelstein and Verma, 2001; Lynn, 2001; Helsel, 2005; Schisterman, *et al.*, 2006.

The growing use of biomarkers in exposure assessment strengthens the need to jointly address the issues of pooling biospecimens and the limit of detection related to their measurements. In the present paper we examine the efficiency of pooling biospecimens whose measurements are subject to a detection limit of the instrument. We address issues concerning the pooling design and a detection limit such as 1) does the pooling strategy relax or aggravate the detection limit problem, as compared to other sampling approaches? and 2) if and how the detection limit affects the performance of a sampling strategy? as well as 3) can data obtained by applying a specific sampling strategy lead to more efficient inference than the full data?

In Section 2, we present a general likelihood based method. Sections 3 and 4 address the problem for data that follow a normal or gamma distribution, respectively. Section 5 presents a simulated example based on real data. We give some concluding remarks in Section 6.

2 A General Methodology

Let $X^{(F)} = \{X_i,\, i = 1, \ldots, N\}$, referred to as the "full data", denote the set of independent identically distributed random variables corresponding to the measurements from a population of size N. In order to make inference on the distribution of the population, ideally we would want to observe all the Xs.When this is not feasible, possibly due to cost restrictions, the conventional approach, referred to as the "random sampling strategy", is to randomly select n subjects and subsequently obtain, say, a random sample of $X^{(r)} = \{X_i,\, i = 1, \ldots, n\}$. In contrast, the pooling strategy, aiming at fully utilizing all the available subjects without increasing the study cost, randomly groups the N subjects into n sets, each of size p. (For convenience, we assume $n = N/p$ is an integer.) Subsequently, instead of $X^{(F)}$ or $X^{(r)}$, the pooling design yields observations, say, of $X^{(p)} = \left\{X_j^{(p)} \equiv \sum_{i=p(j-1)+1}^{jp} X_i/p,\, j = 1, \ldots, n\right\}$, where the average of the Xs in each set is the result of the measuring process of a pooled biological sample (e.g. Faraggi, *et al.*, 2003; Liu and Schisterman, 2003).

Often in many practical situations the measurements of a biomarker are subject to a detection limit, say d. Thus, X_i (or $X_i^{(p)}$) is observed only if $X_i \geq d$ (or $X_i^{(p)} \geq d$). We write $Z^{(F)}$, $Z^{(p)}$ and $Z^{(r)}$ as the observable portion of $X^{(F)}$, $X^{(p)}$ and $X^{(r)}$, respectively. Assume that the random variable X has a density function $f_X(u;\theta)$ dependent on an unknown vector of parameters $\theta = (\theta_1, \ldots, \theta_m)$. Our primary interest is to make inference on θ based on $Z^{(p)}$ and compare the efficiency with that based on $Z^{(F)}$ and $Z^{(r)}$. (Here and henceforth, we say that an estimator is more efficient if it has smaller mean squared error, or variance if the estimators under consideration are asymptotically unbiased.) Note that, since $X_j^{(p)} = \sum_{i=p(j-1)+1}^{jp} X_i/p$, the convolution represented as

$$f_{X^{(p)}}(u;\theta) \;=\; \int f_X(pu - t_1 - \ldots - t_{p-1};\theta) f_X(t_1;\theta) \cdots f_X(t_{p-1};\theta) dt_1 \cdots dt_{p-1} \qquad (1)$$

yields the density function of $X_j^{(p)}$ for $j = 1, \ldots, n$.

The likelihood functions based on $Z^{(F)}$, $Z^{(p)}$ and $Z^{(r)}$ can be expressed in the form of

$$L(Z;\theta) \;=\; \frac{N_Z!}{R_Z\,!\,(N_Z - R_Z)!} \prod_{i:X_i^{(Z)} \in Z} f_Z(X_i^{(Z)};\theta) \times \left(\int_{-\infty}^{\hat{d}_Z} f_Z(u;\theta) du\right)^{R_Z}, \qquad (2)$$

where $Z = Z^{(F)}, Z^{(p)}, Z^{(r)}$; $N_{Z^{(F)}} = N, N_{Z^{(p)}} = N_{Z^{(r)}} = n$; $R_{Z^{(F)}}, R_{Z^{(p)}}, R_{Z^{(r)}}$ are the differences between N, n, n and the number of observations of $Z^{(F)}, Z^{(p)}, Z^{(r)}$, respectively (i.e. R is the the number

of observations from the corresponding population that are below the detection limit d); $f_{Z^{(F)}} = f_{Z^{(r)}} = f_X, f_{Z^{(p)}} = f_{X^{(p)}}$; $\left\{X_i^{(Z^{(F)})}, X_i^{(Z^{(p)})}, X_i^{(Z^{(r)})}\right\} = \left\{X_i, X_i^{(p)}, X_i\right\}$ and

$$\hat{d}_Z \quad = \quad \begin{cases} d, & \text{if } d \text{ is known;} \\ \min_{X_i^{(Z)} \in Z} X_i^{(Z)}, & \text{if } d \text{ is unknown.} \end{cases}$$

Thus, the maximum likelihood estimator of θ based upon the observed sample Z is given by

$$\left\{\hat{\theta}_1, \ldots, \hat{\theta}_m\right\} \quad = \quad \arg \max_{a_1, \ldots, a_m} L(Z; a_1, \ldots, a_m).$$

The accuracy of the estimators $\hat{\theta}_1, \ldots, \hat{\theta}_m$ depends on the number of observations of Z. Since the observations are independent identically distributed, the probability $P\left\{X_1^{(Z)} \geq d\right\}$ is the expected proportion of observations in set Z. In the case of $Z = Z^{(p)}$, $X_i^{(Z)} = X_i^{(p)}$ is the average of the Xs with $EX_i^{(p)} = EX_1$ and $var\left(X_i^{(p)}\right) = var\left(X_1\right)/p$ $(\leq var\left(X_1\right))$. Therefore, we can assume that the density function $f_{X^{(p)}}$ is more concentrated around some point d_0 than f_X (i.e. in some neighborhood about d_0 the area under function $f_{X^{(p)}}$ is larger than that under function f_X); and hence if the detection limit $d \leq d_0$ (obviously, e.g., if X_1 is normally distributed then $d_0 = EX_1$) we have

$$P\left\{X_1^{(p)} < d\right\} = \int_{-\infty}^{d} f_{X^{(p)}}(u;\theta)du \quad \leq \quad \int_{-\infty}^{d} f_X(u;\theta)du = P\left\{X_1 < d\right\},$$

because

$$\int_{-\infty}^{\infty} f_{X^{(p)}}(u;\theta)du = \int_{-\infty}^{\infty} f_X(u;\theta)du = 1 \quad \text{and} \quad f_{X^{(p)}}, f_X \geq 0.$$

Therefore, for some values of d and p, the sample $Z^{(p)}$ can have more observations than $Z^{(r)}$ (and perhaps $Z^{(F)}$). This will be further reflected in the next two sections.

3 Normally Distributed Data

In this case $f_X(u;\theta_1 = \mu, \theta_2 = \sigma) = \varphi\left((u-\mu)/\sigma\right)$, where $\varphi(.)$ is the standard normal density function, $\mu = EX_1$ and $\sigma^2 = var(X_1)$ are the unknown parameters. It is clear that, for the pooled observations, $f_{X^{(p)}}(u;\theta_1 = \mu, \theta_2 = \sigma) = \varphi\left(p^{1/2}(u-\mu)/\sigma\right)$. From (2) the likelihood function is

$$L(Z;\mu,\sigma) \quad \propto \quad (\sigma)^{R_Z - N_Z} \exp\left(-\frac{p_Z}{2\sigma^2} \sum_{X_i^{(Z)} \in Z} \left(X_i^{(Z)} - \mu\right)^2\right) \left(\int_{-\infty}^{\frac{\hat{d}_Z - \mu}{\sigma p_Z^{-1/2}}} \varphi(u)du\right)^{R_Z}, \quad (3)$$

$$Z = Z^{(F)}, Z^{(p)}, Z^{(r)};\ p_{Z^{(F)}} = p_{Z^{(r)}} = 1, p_{Z^{(p)}} = p,$$

and therefore the maximum likelihood estimators $(\hat{\mu}(F), \hat{\sigma}(F)), (\hat{\mu}(p), \hat{\sigma}(p)), (\hat{\mu}(r), \hat{\sigma}(r))$ of (μ, σ) based on samples $Z^{(F)}$, $Z^{(p)}$ and $Z^{(r)}$, respectively, are solutions of the equations:

$$\begin{cases} \frac{\partial \ln L(Z;\hat{\mu},\hat{\sigma})}{\partial \hat{\mu}} = & \frac{p_Z}{(\hat{\sigma})^2} \sum_{X_i^{(Z)} \in Z} \left(X_i^{(Z)} - \hat{\mu}\right) - \phi\left(\frac{\hat{d}_Z - \hat{\mu}}{\hat{\sigma} p_Z^{-1/2}}\right) \frac{R_Z\, p_Z^{1/2}}{\hat{\sigma}} = 0 \\ \frac{\partial \ln L(Z;\hat{\mu},\hat{\sigma})}{\partial \hat{\sigma}} = & \frac{R_Z - N_Z}{\hat{\sigma} p_Z^{-1/2}} + \frac{p_Z^{3/2}}{(\hat{\sigma})^3} \sum_{X_i^{(Z)} \in Z} \left(X_i^{(Z)} - \hat{\mu}\right)^2 - \frac{\hat{d}_Z - \hat{\mu}}{(\hat{\sigma})^2} \phi\left(\frac{\hat{d}_Z - \hat{\mu}}{\hat{\sigma} p_Z^{-1/2}}\right) R_Z\, p_Z = 0, \end{cases} \quad (4)$$

$$Z = Z^{(F)}, Z^{(p)}, Z^{(r)}; \quad \phi(u) = \frac{\varphi(u)}{\int_{-\infty}^{u} \varphi(u)du}.$$

(E.g. Gupta, 1952 as well as Harter and Moore, 1966.)

One can show that if $d \to -\infty$ (i.e. the detection limit is not in effect), (4) yields simple forms of the estimators

$$\hat{\mu}(F) = \frac{1}{N}\sum_{i=1}^{N} X_i \quad = \quad \frac{1}{n}\sum_{i=1}^{n} X_i^{(p)} = \hat{\mu}(p), \quad \hat{\mu}(r) = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$\hat{\sigma}(F)^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \hat{\mu}(F)\right)^2, \; \hat{\sigma}(p)^2 = \frac{p}{n}\sum_{i=1}^{n}\left(X_i^{(p)} - \hat{\mu}(p)\right)^2, \; \hat{\sigma}(r)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \hat{\mu}(r)\right)^2,$$

where $\hat{\sigma}(p)$ is distributed as $\hat{\sigma}(r)$ (Liu and Schisterman, 2003). Thus, in this case, $var\left(\hat{\mu}(F)\right) = var\left(\hat{\mu}(p)\right) \leq var\left(\hat{\mu}(r)\right)$ and $var\left(\hat{\sigma}(F)\right) \leq var\left(\hat{\sigma}(p)\right) = var\left(\hat{\sigma}(r)\right)$.

Hence, when d is small enough (i.e. $d < 0$ and $|d| \gg 0$), the statistical characteristics of the μ-estimator from pooling sample $Z^{(p)}$ are similar to the statistical properties of the μ-estimator based on the full sample $Z^{(F)}$; and $\hat{\mu}(p)$ is more efficient than $\hat{\mu}(r)$. Since the normal probabilities $P\left\{X_1^{(p)} \geq d\right\} \geq P\{X_1 \geq d\}$ with $d \leq \mu$, the expected number $n\,P\left\{X_1^{(p)} \geq d\right\}$ of observations in $Z^{(p)}$ decreases as the detection limit d moves from $-\infty$ to μ, in a rate that is smaller than that of the expected number $N\,P\{X_1 \geq d\}$ of observations in $Z^{(F)}$. Therefore, we can expect that there are values of $d \leq \mu$ at which $\hat{\mu}(p)$ is more efficient even than $\hat{\mu}(F)$, which is based on a larger number of samples. Similarly, if $d \geq \mu$, then $P\left\{X_1^{(p)} \geq d\right\} \leq P\{X_1 \geq d\}$; thus we expect a large d at which $\hat{\mu}(p)$ is least efficient. To give a simple example, set $\mu = 0$, $\sigma = 1$, $N = 300$, $p = 2$ and hence $n = 150$. Figure 1 depicts the expected sizes of sets $Z^{(F)}$, $Z^{(p)}$ and $Z^{(r)}$ for different values of d. Thus, around $d = -1/2$ (where $-1/2 \simeq \max \arg_q \int_q^\infty \exp\left(-u^2\right) du / \int_q^\infty \exp\left(-u^2/2\right) du$) the pooling design yields the most efficient estimator of the normal mean μ.

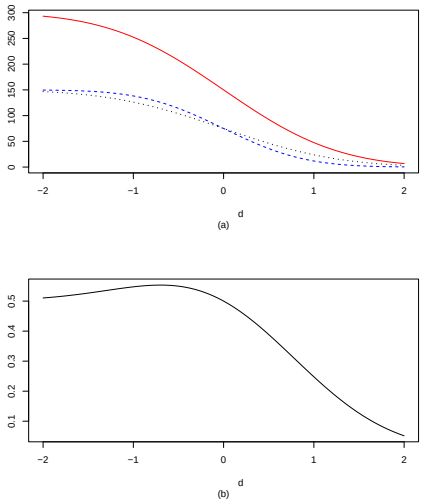

Fig. 1 *(a):* the expected numbers of observations of $Z^{(F)}$ (curve —), $Z^{(p)}$ (- - -) and $Z^{(r)}$ ($\cdots$); *(b):* the ratio of the expected size of set $Z^{(p)}$ to the expected size of set $Z^{(F)}$ plotted against d (the axis of abscissae); $p = 2$.

Similarly, we consider the estimators of the standard deviation σ. When d is small enough, the statistical characteristics of the σ-estimator based on the pooled sample $Z^{(p)}$ are similar to that of the σ-estimator based on the random sample $Z^{(r)}$; for $d \leq \mu$, $\hat{\sigma}(p)$ is more efficient than $\hat{\sigma}(r)$.

For finite d and p, the vectors $N^{1/2}[\hat{\mu}(F)-\mu, \hat{\sigma}(F)-\sigma]^T$, $N^{1/2}[\hat{\mu}(p)-\mu, \hat{\sigma}(p)-\sigma]^T$ and $N^{1/2}[\hat{\mu}(r)-\mu, \hat{\sigma}(r)-\sigma]^T$ have asymptotically (as $N \to \infty$) a bivariate normal distribution with zero expectation and covariance matrices $\Delta(F), \Delta(p)$ and $\Delta(r)$, respectively (Persson and Rootzen, 1977). The covariance matrices have the form $\Delta = \lim_{N\to\infty} N\Delta_N$, where

$$\Delta_N = \begin{bmatrix} var(\hat{\mu}) & cov(\hat{\mu}, \hat{\sigma}) \\ cov(\hat{\mu}, \hat{\sigma}) & var(\hat{\sigma}) \end{bmatrix} \tag{5}$$

which is the inverse of the Fisher information matrix

$$\begin{bmatrix} -E\frac{\partial^2 \ln L(Z;\mu,\sigma)}{\partial\mu^2} & -E\frac{\partial^2 \ln L(Z;\mu,\sigma)}{\partial\mu\partial\sigma} \\ -E\frac{\partial^2 \ln L(Z;\mu,\sigma)}{\partial\mu\partial\sigma} & -E\frac{\partial^2 \ln L(Z;\mu,\sigma)}{\partial\sigma^2} \end{bmatrix}.$$

By applying the asymptotic results of Gupta (1952) as well as Harter and Moore (1966), we obtain

$$\begin{aligned}
\lim_{N\to\infty} -E\frac{\partial^2 \ln L(Z;\mu,\sigma)}{N\partial\mu^2} &= \frac{\omega(Z)}{\sigma^2}\left(\int_{\eta_Z}^{\infty}\varphi(u)du + \varphi(\eta_Z)(\eta_Z + \phi(\eta_Z))\right), \qquad (6)\\
\lim_{N\to\infty} -E\frac{\partial^2 \ln L(Z;\mu,\sigma)}{N\partial\mu\partial\sigma} &= \frac{\omega(Z)}{\sigma^2 p_Z^{1/2}}\left(\varphi(\eta_Z) + \eta_Z\varphi(\eta_Z)(\eta_Z + \phi(\eta_Z))\right),\\
\lim_{N\to\infty} -E\frac{\partial^2 \ln L(Z;\mu,\sigma)}{N\partial\sigma^2} &= \frac{\omega(Z)}{\sigma^2 p_Z}\left(2\int_{\eta_Z}^{\infty}\varphi(u)du + \eta_Z\varphi(\eta_Z) + \eta_Z^2\varphi(\eta_Z)(\eta_Z + \phi(\eta_Z))\right),\\
\eta_Z = \frac{d-\mu}{\sigma p_Z^{-1/2}}; &\qquad Z = Z^{(F)}, Z^{(p)}, Z^{(r)};\ \omega(Z^{(F)}) = \omega(Z^{(p)}) = 1, \omega(Z^{(r)}) = 1/p.
\end{aligned}$$

Setting $\mu = 0$, $\sigma = 1$ and utilizing (5) and (6), we calculate the asymptotic variances of the estimators as functions of d. Depending on the sampling strategy and the detection limit d, the values of $\lim_{N\to\infty}(N\,var(\hat{\mu}))$ and $\lim_{N\to\infty}(N\,var(\hat{\sigma}))$ are presented in Figure 2. In accordance with the graphs (b) and (c) (or (b') and (c')), for $d \le -1.5$ we have $var(\hat{\mu}(F)) \simeq var(\hat{\mu}(p)) < var(\hat{\mu}(r))$ and $var(\hat{\sigma}(F)) < var(\hat{\sigma}(p)) \simeq var(\hat{\sigma}(r))$. The graphs (a) and (a') correspond to the probabilities to observe X and $X^{(p)}$ (i.e., for example, $P\{X_1 \ge d\} = 1-$ "the area under the curve — from $-\infty$ to d"). Thus, for $d \le 0\,(=\mu)$: $P\{X_1 \ge d\} \le P\left\{X_1^{(p)} \ge d\right\}$ and therefore, as was expected, $var(\hat{\mu}(F)) \ge var(\hat{\mu}(p))$ and $var(\hat{\sigma}(p)) \le var(\hat{\sigma}(r))$. When $d \ge 0$, we have the inverse situation. This example again illustrates the same conclusions regarding the efficiency of the estimators.

4 Gamma Case

In certain practical situation the distribution of the biospecimen values is skewed and the normality assumption is not reasonable. In this case assumption of the gamma distribution, which takes a variety of skewed shapes, provides a reasonable alternative (e.g. Faraggi *et al.*, 2003).

In this section we consider the probability density function of the random variable $X_1 \ge 0$ having a gamma distribution with scale parameter α and shape parameter β:

$$f_X(u;\theta_1=\alpha,\theta_2=\beta) = \frac{u^{\alpha-1}}{\beta^\alpha\Gamma(\alpha)}\exp\left(-\frac{u}{\beta}\right), \quad \alpha,\beta>0,\ u\ge 0.$$

Following Faraggi, *et al.* (2003), the pooled variate $X_1^{(p)}$ has the gamma density

$$f_{X^{(p)}}(u;\theta_1=\alpha,\theta_2=\beta) = \frac{u^{p\alpha-1}}{(\beta/p)^{p\alpha}\Gamma(p\alpha)}\exp\left(-\frac{up}{\beta}\right).$$

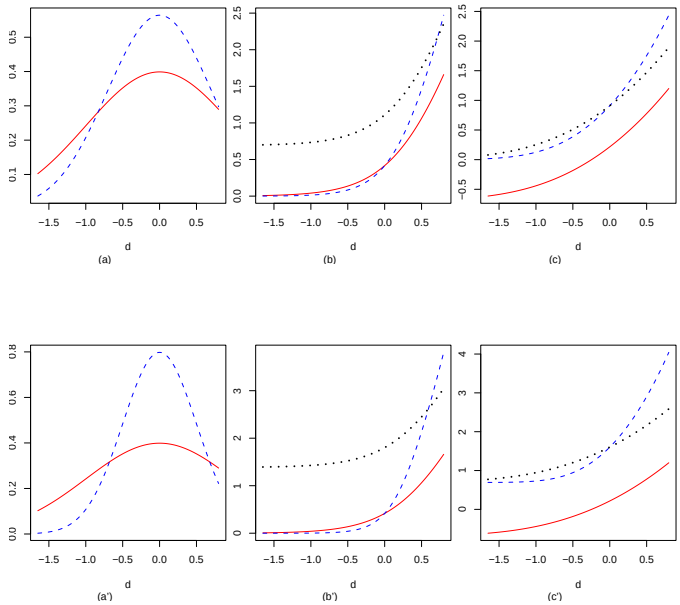

Fig. 2 (a) (or (a')): the density functions $\varphi\left((d-\mu)/\sigma\right)$ and $\varphi\left(p^{1/2}(d-\mu)/\sigma\right)$ of X_1 (curve —) and $X_1^{(p)}$ (- - -), where $p = 2$ (or $p = 4$); (b) (or (b')): the curve "—" is $\ln\left(\lim_{N\to\infty} N\, var(\hat{\mu}(F))\right)$, "· · ·" presents $\ln\left(\lim_{N\to\infty} N\, var(\hat{\mu}(r))\right)$, "- - -" depicts $\ln\left(\lim_{N\to\infty} N\, var(\hat{\mu}(p))\right)$ for $p = 2$ (or $p = 4$); (c) (or (c')): the curve "—" is $\ln\left(\lim_{N\to\infty} N\, var(\hat{\sigma}(F))\right)$, "· · ·" presents $\ln\left(\lim_{N\to\infty} N\, var(\hat{\sigma}(r))\right)$, "- - -" depicts $\ln\left(\lim_{N\to\infty} N\, var(\hat{\sigma}(p))\right)$ for $p = 2$ (or $p = 4$).

By (2), the natural logarithm of the likelihood function for observed samples $Z = Z^{(F)}, Z^{(p)}, Z^{(r)}$ is

$$\begin{aligned} \ln L(Z;\alpha,\beta) \;=\; & \ln\frac{N_Z!}{R_Z!\,(N_Z-R_Z)!} - (N_Z-R_Z)\left(\ln\Gamma(p_Z\alpha) + p_Z\alpha\ln\frac{\beta}{p_Z}\right) \\ & +(p_Z\alpha-1)\sum_{X_i^{(Z)}\in Z}\ln X_i^{(Z)} - \frac{p_Z}{\beta}\sum_{X_i^{(Z)}\in Z} X_i^{(Z)} \\ & +R_Z\ln\left(\frac{1}{\Gamma(p_Z\alpha)}\int_0^{\frac{p_Z\hat{d}_Z}{\beta}} t^{p_Z\alpha-1}e^{-u}du\right), \quad p_{Z^{(F)}} = p_{Z^{(r)}} = 1, p_{Z^{(p)}} = p. \end{aligned} \tag{7}$$

As in the case of a normally distributed population, solving the maximum likelihood equations $\partial\ln L(Z;\hat{\alpha},\hat{\beta})/\partial\hat{\alpha} = 0$, $\partial\ln L(Z;\hat{\alpha},\hat{\beta})/\partial\hat{\beta} = 0$, based on samples $Z = Z^{(F)}, Z^{(p)}, Z^{(r)}$, yields estimators $(\hat{\alpha}(F), \hat{\beta}(F))$, $(\hat{\alpha}(p), \hat{\beta}(p))$ and $(\hat{\alpha}(F), \hat{\beta}(F))$ of (α,β), respectively (e.g. Chapman, 1956). The estimators are asymptotically unbiased and normally distributed, as $N\to\infty$. The asymptotic covariance matrix of $N^{1/2}[\hat{\alpha}-\alpha, \hat{\beta}-\beta]^T$ can be found by inverting the asymptotic Fisher information matrix divided by N, as $N\to\infty$. By utilizing propositions of Harter and Moore (1967), we obtain the limiting values of

the elements of the information matrix (multiplied by $1/N$), which are

$$\lim_{N\to\infty} -E\frac{\partial^2 \ln L(Z;\alpha,\beta)}{N\partial\alpha^2} = \omega(Z)p_z\left(\frac{\Gamma(p_z\alpha)\Gamma''(p_z\alpha)-(\Gamma'(p_z\alpha))^2}{\Gamma(p_z\alpha)^2}\right. \tag{8}$$
$$\left.-\frac{\Gamma(p_Z\alpha;\eta_Z)\Gamma''(p_Z\alpha;\eta_Z)-(\Gamma'(p_Z\alpha;\eta_Z))^2}{\Gamma(p_Z\alpha)^2\int_0^{\eta_Z} f_X(u;p_Z\alpha,1)du}\right),$$
$$\lim_{N\to\infty} -E\frac{\partial^2 \ln L(Z;\alpha,\beta)}{N\partial\alpha\partial\beta} = \frac{\omega(Z)}{\beta}\left(\int_d^{\infty} f_X(u;p_Z\alpha,\beta/p_Z)du\right.$$
$$\left.+\eta_Z\frac{f_X(\eta_Z;p_Z\alpha,1)}{\int_0^{\eta_Z} f_X(u;p_Z\alpha,1)du}\left(\ln(\eta_Z)\int_0^{\eta_Z} f_X(u;p_Z\alpha,1)du-\Gamma'(\alpha;\eta_Z)/\Gamma(\alpha)\right)\right),$$

$$\lim_{N\to\infty} -E\frac{\partial^2 \ln L(Z;\alpha,\beta)}{N\partial\beta^2} = \frac{\omega(Z)}{p_z\beta^2}\left(-p_Z\alpha\int_d^{\infty} f_X(u;p_Z\alpha,\beta/p_Z)du\right.$$
$$+\frac{2}{\Gamma(p_Z\alpha)}\left(\Gamma(p_Z\alpha+1)-\Gamma(p_Z\alpha+1;\eta_Z)\right)+\eta_Z\frac{f_X(\eta_Z;p_Z\alpha,1)}{\int_0^{\eta_Z} f_X(u;p_Z\alpha,1)du}$$
$$\left.\times\left((\eta_Z-p_Z\alpha-1)\int_0^{\eta_Z} f_X(u;p_Z\alpha,1)du+\eta_Z f_X(\eta_Z;p_Z\alpha,1)\right)\right),$$

where

$$\omega(Z^{(F)})=\omega(Z^{(p)})=1,\omega(Z^{(r)})=1/p;\ p_{Z^{(F)}}=p_{Z^{(r)}}=1,p_{Z^{(p)}}=p;$$
$$\Gamma(a;b)=\int_0^b u^{a-1}e^{-u}du;\ \eta_Z=p_Z d/\beta.$$

Similarly to Section 3, we start with examining the efficiency of the estimators $(\hat{\alpha}(p),\ \hat{\beta}(p))$ and $(\hat{\alpha}(r),\hat{\beta}(r))$. When $d=0$, (8) leads to the following information matrices

$$V^{(F)}=\begin{bmatrix}\ln''(\Gamma(\alpha)) & \frac{1}{\beta}\\ \frac{1}{\beta} & \frac{\alpha}{\beta^2}\end{bmatrix},\ V^{(p)}=\begin{bmatrix}p\ln''(\Gamma(p\alpha)) & \frac{1}{\beta}\\ \frac{1}{\beta} & \frac{\alpha}{\beta^2}\end{bmatrix},\ V^{(r)}=\begin{bmatrix}\frac{\ln''(\Gamma(\alpha))}{p} & \frac{1}{p\beta}\\ \frac{1}{p\beta} & \frac{\alpha}{p\beta^2}\end{bmatrix}, \tag{9}$$

corresponding to the full data, pooling, and random sampling strategies, respectively. Thus, the pooling strategy yields the same information related to β-estimation as does the full data. However, $\hat{\alpha}(p)$ does not have the information density that $\hat{\alpha}(F)$ has (or, for some (p,α), $\hat{\alpha}(r)$ has). This fact is stipulated by the equation $\hat{\beta}(p)=\hat{\beta}(F)$, for a given α; nevertheless statistics based on averages $X^{(p)}$ are not close to a sufficient statistic for the parameter α based on individual X. (Since $\hat{\alpha}(F):\ \ln(\hat{\alpha}(F))-\Gamma'(\hat{\alpha}(F))=\ln\left(\sum_{i=1}^N X_i/N\right)-\sum_{i=1}^N\ln(X_i)/N$, $\hat{\alpha}(F)$ can not be represented by the particular sums of Xs.) Define $A(\alpha,p)\equiv\beta^2\left(\det\left(V^{(p)}\right)-\det\left(V^{(r)}\right)\right)$. Figure 3 plots the function $A(\alpha,p)$, for different values of $\alpha>0$ and $p\geq 2$.

Hence, $\det\left(V^{(p)}\right)\geq\det\left(V^{(r)}\right)$, and therefore it is clear that $var(\hat{\beta}(p))\leq var(\hat{\beta}(r))$. However, by using (9), we can easily define α,β and p such that $var(\hat{\alpha}(p))\geq var(\hat{\alpha}(r))$.

Now, if the detection limit $d>0$ is below some point d_0 such that

$$\int_0^{d_0}\frac{u^{p\alpha-1}}{(\beta/p)^{p\alpha}\Gamma(p\alpha)}\exp\left(-\frac{up}{\beta}\right)du\leq\int_0^{d_0}\frac{u^{\alpha-1}}{(\beta)^{\alpha}\Gamma(\alpha)}\exp\left(-\frac{u}{\beta}\right)du,$$

the pooled sample yields more observations than the random sample, and hence the difference in the amount of information, which are obtained by the pooling design and the random sampling strategy, increases. (Existence of the point d_0 such that for all $d\leq d_0$: $P\left\{X_1^{(p)}<d\right\}\leq P\{X_1<d\}$ can be

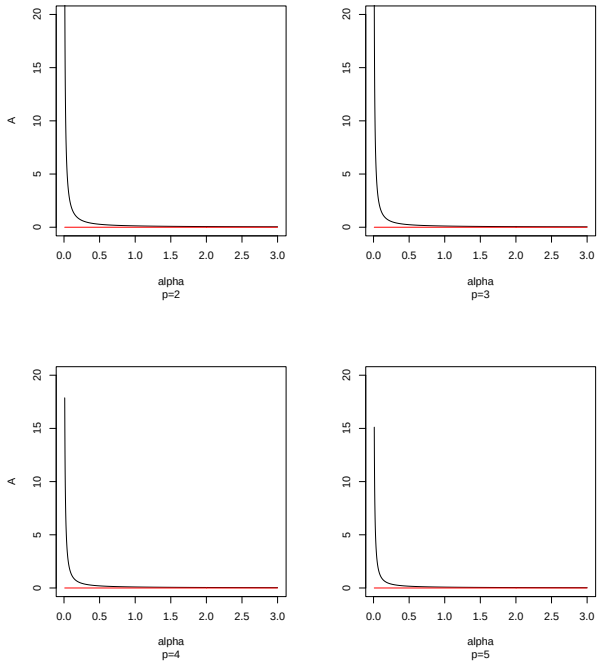

Fig. 3 The differences, multiplied by β^2, between the amounts of information that are obtained by the pooling design and the random sampling strategy, for $p = 2, 3, 4, 5$ and different $\alpha > 0$.

graphically displayed easily.) However, in the case where $d \gg d_0$ the pooling design is an unreasonable method, as illustrated in the following examples. In the figures we specify $\alpha = 1/4$ and $\beta = 4$ for Figure 4(a,b,c) with $p = 2$ and for Figure 5(a,b,c) with $p = 4$; similarly, $\alpha = 4$ and $\beta = 1/4$ for graphs (a', b', c') of Figures 4 and 5. When $p = 2$, α is relative small and β is relative large ($\alpha = 1/4,\ \beta = 4$); for $d \leq 1.25$ the pooling strategy is obviously more efficient than the random sampling. However when $p = 2$, $\alpha = 4$ and $\beta = 1/4$, for $d \in (0, 1)$ the variances of the estimators based on the pooling sample are slightly smaller than that based on the random sample (note that following Figure 4 (a'): $P\left\{X_1^{(p)} < d\right\} \leq \{X_1 < d\}$, for $d \leq 1$). Moreover, Figure (5) (b) indicates that, in the case $(p = 4, \alpha = 1/4, \beta = 4)$, if estimation of α is the main focus, pooling is not recommended.

5 An Example

We exemplify the presented issue concerning the pooling design using data from a study of biomarkers of coronary heart disease. In this study cholesterol level measured in mg/dl was obtained from $N = 40$ controls who had a normal rest electro cardiogram, were free of symptoms and had no previous cardiovascular procedures or myocardial infarctions. In order to investigate the effectiveness of pooling, blood specimens were randomly pooled in groups of $p = 2$ for the controls, and cholesterol level was re-measured and treated as the average of the corresponding individual cholesterol levels. The data were examined by Faraggi *et al.* (2003) and Schisterman *et al.* (2005), assuming that cholesterol levels follow a normal distribution. We also suppose that the data is normally distributed with the mean and the standard deviation estimated by $(\hat{\mu}(F) = 205.53, \hat{\sigma}(F) = 42.29)$ based on the full sample ($N = 40$, $p = 1$); and

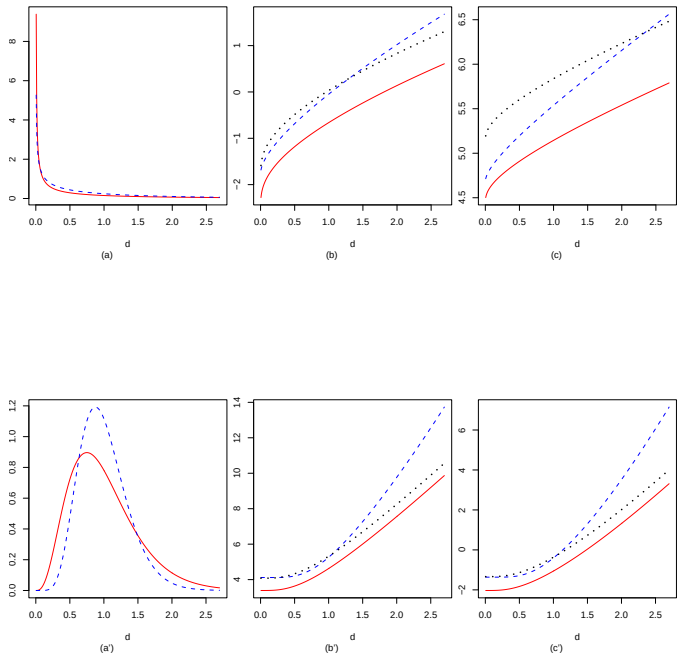

Fig. 4 For $p = 2$, graphs (a, b, c) and (a', b', c') correspond to ($\alpha = 1/4$, $\beta = 4$) and ($\alpha = 4$, $\beta = 1/4$), respectively. (a) and (a') the density functions of X_1 (curve —) and $X_1^{(p)}$ (- - -); (b) and (b'): the curves "—", "- - -" and "· · ·" depict $\ln\left(\lim_{N\to\infty} N\, var(\hat{\alpha}(F))\right)$, $\ln\left(\lim_{N\to\infty} N\, var(\hat{\alpha}(p))\right)$ and $\ln\left(\lim_{N\to\infty} N\, var(\hat{\alpha}(r))\right)$, respectively; (c) and (c'): the curves "—", "- - -" and "· · ·" present $\ln\left(\lim_{N\to\infty} N\, var(\hat{\beta}(F))\right)$, $\ln\left(\lim_{N\to\infty} N\, var(\hat{\beta}(p))\right)$ and $\ln\left(\lim_{N\to\infty} N\, var(\hat{\beta}(r))\right)$, respectively.

$(\hat{\mu}(p) = 207.88, \hat{\sigma}(p) = 48.51)$ based on the pooled data ($n = N/p = 20$, $p = 2$). In the context of the detection limit, we simulate putative $d = 120, \ldots, 250$ and, by applying results from Section 3, estimate $\mu, \sigma, var(\hat{\mu})$ and $var(\hat{\sigma})$, for each value of d. The results are presented by Table 1.

In addition, it is natural to assume that $var(\hat{\mu}(r))$ and $var(\hat{\sigma}(r))$ (where a random sample has size $n = 20$) are about two times $var(\hat{\mu}(F))$ and $var(\hat{\sigma}(F))$, respectively. As is shown by the theoretical part of this paper, $\hat{var}(\hat{\mu}(p)) < 2\,\hat{var}(\hat{\mu}(F))(\approx var(\hat{\mu}(r)))$, for $d \le 210$; $\hat{var}(\hat{\mu}(p)) < \hat{var}(\hat{\mu}(F))$, for $d = 200, 210$ (note that these values of d are close to the estimated mean of the data) and $\hat{var}(\hat{\mu}(p)) > 2\,\hat{var}(\hat{\mu}(F))(\approx var(\hat{\mu}(r)))$, for $d > 210$. Perhaps, since the data do not exactly follow a normal distribution, for small d, variance $\hat{var}(\hat{\sigma}(p))$ does not possess a value that is very close to $2\,\hat{var}(\hat{\sigma}(F))(\approx var(\hat{\sigma}(r)))$. However, for $d = 200, 210$, we have $\hat{var}(\hat{\sigma}(p)) < 2\,\hat{var}(\hat{\sigma}(F))(\approx var(\hat{\sigma}(r)))$. Certainly, the pooling sample is the most efficient, if $d = 200, 210$. On the other hand pooling design is undesirable for $d = 250$.

6 Discussion

In the present paper no attention is paid to the possibility of technical monitoring of the instrument limitations. However, we show that utilizing pooling design can relax or aggravate the influence of the detection

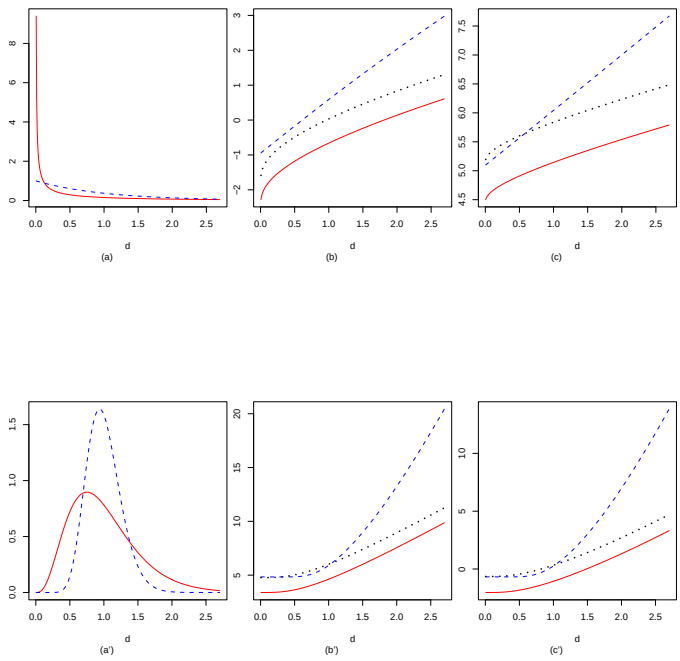

Fig. 5 For $p = 4$, graphs (a, b, c) and (a', b', c') correspond to ($\alpha = 1/4$, $\beta = 4$) and ($\alpha = 4$, $\beta = 1/4$), respectively. (a) and (a') the density functions of X_1 (curve —) and $X_1^{(p)}$ (- - -); (b) and (b'): the curves "—", "- - -" and "$\cdots$" depict $\ln\left(\lim_{N\to\infty} N\, var(\hat{\alpha}(F))\right)$, $\ln\left(\lim_{N\to\infty} N\, var(\hat{\alpha}(p))\right)$ and $\ln\left(\lim_{N\to\infty} N\, var(\hat{\alpha}(r))\right)$, respectively; (c) and (c'): the curves "—", "- - -" and "$\cdots$" present $\ln\left(\lim_{N\to\infty} N\, var(\hat{\beta}(F))\right)$, $\ln\left(\lim_{N\to\infty} N\, var(\hat{\beta}(p))\right)$ and $\ln\left(\lim_{N\to\infty} N\, var(\hat{\beta}(r))\right)$, respectively.

limit on the measurement of biological samples. Hence, arriving at a decision to apply the pooling strategy is not a trivial determination. Since the efficiency of the pooling strategy is a function of unknown parameters, a two-stage sampling design can be recommended. In order to evaluate the efficiency of the sampling strategies, the first pilot sample can be executed. In this context, a consideration of risk-function $T = \sum_{i=1}^{m} w_{ii} var(\hat{\theta}_i) + \sum_{i\neq j}^{m} w_{ij} cov(\hat{\theta}_i, \hat{\theta}_j)$ is appealing, where θ_i, $i = 1, \ldots, m$ are unknown parameters and $\{w_{ij}\}$ is a set of weights. Moreover, for fixed N and d, an optimal size p of pooling, which minimizes T with $pn \leq N$, can be obtained. (T can be interpreted as the generalized variance of the vector parameters' estimator. For example, setting $w_{ij} = 1$, if $i = j$ and $w_{ij} = 0$, if $i \neq j$ leads to the trace of the dispersion matrix.) Note that, even if a detection limit is not in effect, the efficiency of the pooling design is dependent on the distribution of the data. In particular, by basing on a gamma distribution, Section 4 presents the method to obtain the pooling sample efficiency depending on the parameters of the distribution function. The paper considers measurements that are subject to left-censoring due to values below the assay detection limit. However, for right-censoring or doubly-censoring data, the contemplations are similar.

Table 1 Applying the maximum likelihood method (Section 3) to the real data with the simulated detection limits d. $p = 1$ and $p = 2$ correspond to the estimation based on the unpooled (full) and pooled data, respectively. (In accordance with (5) and (6), we assume that $var(\hat{\mu}(r))$ and $var(\hat{\sigma}(r))$ are about two times $var(\hat{\mu}(F))$ and $var(\hat{\sigma}(F))$, respectively.)

d	p	R_Z	$\hat{\mu}$	$\hat{var}(\hat{\mu})$	$\hat{\sigma}$	$\hat{var}(\hat{\sigma})$
120	1	1	205.13	45.69	42.69	23.66
120	2	1	207.41	63.60	50.42	64.30
150	1	3	205.23	45.93	42.46	26.27
150	2	1	208.87	51.40	45.24	54.03
170	1	7	205.68	46.88	42.04	30.33
170	2	3	208.59	53.80	45.81	63.10
200	1	18	203.62	70.46	44.54	57.09
200	2	6	211.21	48.12	40.19	72.31
210	1	24	197.03	127.00	50.37	102.98
210	2	10	207.73	81.84	44.95	134.12
250	1	34	199.47	544.44	49.63	289.55
250	2	18	196.16	1563.28	59.59	1320.00

Acknowledgements The authors are grateful to the editor, associate editor, and referee for their helpful comments that clearly improved this paper.

References

Chapman, D. G. (1956). Estimating the parameters of a truncated gamma distribution. *Annals of Mathematical Statistics* **27**, 498–506.

Cohen, A. C., Jr. (1955). Restriction and selection in samples from bivariate normal distributions. *Journal of the American Statistical Association* **50**, 884–893.

Dorfman, R. (1943). The detection of defective members of large populations. *Annals of Mathematical Statistics* **14**, 436–440.

Faraggi, D., Reiser, B. and Schisterman, E. F. (2003). ROC curve analysis for biomarkers based on pooled assessments. *Statistics in Medicine* **15**, 2515–2527.

Finkelstein, M. M. and Verma, D. K. (2001). Exposure estimation in the presence of nondetectable values: another look. *American Industrial Hygiene Association Journal* **62**, 195–198.

Gupta, A. K. (1952). Estimation of the mean and standard deviation of a normal population from a censored sample. *Biometrika* **39**, 260–273.

Harter, H. L. and Moore, A. H. (1966). Iterative maximum-likelihood estimation of the parameters of normal populations from singly and doubly censored samples. *Biometrika* **53**, 205–213.

Harter, H. L. and Moore, A. H. (1967). Asymptotic variances and covariances of maximum-likelihood estimators, from censored samples, of the parameters of Weibull and Gamma populations. *Annals of Mathematical Statistics* **38**, 557–570.

Helsel, D. (2005). *Nondetects and data analysis: Statistics for censored environmental data*. Wiley, New Jersey.

Liu, A. and Schisterman, E. (2003). Comparison of diagnostic accuracy of biomarkers with pooled assessments. *Biometrical Journal* **45**, 631–644.

Lynn, H. S. (2001). Maximum likelihood inference for left-censored HIV RNA data. *Statistics in Medicine* **20**, 33–45.

Persson, T. and Rootzen, H. (1977). Simple and highly efficient estimators for Type I censored normal sample. *Biometrika* **64**, 123–128.

Schisterman, E. F., Perkins, N., Liu, A. and Bondell, H. (2005). Optimal Cut-point and its Corresponding Youden Index to Discriminate Individuals Using Pooled Blood Samples. *Epidemiology* **16**, 73–81.

Schisterman, E. F., Vexler, A., Whitcomb, B. W. and Liu, A. (2006). The limitations due to exposure detection limits for regression models. *American Journal of Epidemiology* **163**, 374–383.

Sterrett, A. (1957). On the detection of defective members of large populations. *Annals of Mathematical Statistics* **28**, 1033–1036.

Wilks, S. S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. *Annals of Mathematical Statistics* **3**, 163–195.

Weinberg, C. R. and Umbach, D. M. (1999). Using pooled exposure assessment to improve efficiency in case-control studies. *Biometrics* **55**, 718–726.