Ann. Inst. Statist. Math.
Vol. 54, No. 4, 861–878 (2002)

ESTIMATION OF THE COMMON MEAN OF A BIVARIATE NORMAL POPULATION*

PHILIP L. H. YU[1], YIJUN SUN[2]** AND BIMAL K. SINHA[2]

[1]*Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, CHINA*

[2]*Department of Mathematics and Statistics, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21228-5398, U.S.A.*

(Received December 14, 2000; revised June 1, 2001)

Abstract. In this paper we discuss the problem of estimating the common mean of a bivariate normal population based on paired data as well as data on one of the marginals. Two double sampling schemes with the second stage sampling being either a simple random sampling (SRS) or a ranked set sampling (RSS) are considered. Two common mean estimators are proposed. It is found that under normality, the proposed RSS common mean estimator is always superior to the proposed SRS common mean estimator and other existing estimators such as the RSS regression estimator proposed by Yu and Lam (1997, *Biometrics*, **53**, 1070–1080). The problem of estimating the mean Reid Vapor Pressure (RVP) of regular gasoline based on field and laboratory data is considered.

Key words and phrases: Ranked set sampling, relative precision, REML, simple random sampling.

1. Introduction

The problem discussed in this paper is motivated by the following practical issue in the context of the attempt by the Environmental Protection Agency (EPA) of the United States to evaluate the gasoline quality based on what is known as Reid Vapor Pressure (RVP). Occasionally, an EPA inspector would visit gas pumps in a city, take samples of gasoline of a particular brand, and measure RVP right at the spot which produces cheap and quick measurements. Once in a while, the inspector after measuring RVP at the spot will ship a gasoline sample to the laboratory for a measurement of presumably higher precision at a higher cost, thus getting the pair (field, lab). Since usually laboratory measurements (Y) are much more expensive than field measurements (X) because of special packaging to be used to ship a gasoline sample from a field to a laboratory, not all the gasoline samples will be shipped to the laboratory and hence the resulting data would consist of many field measurements with occasional paired measurements obtained from both the field and laboratory. Therefore, it never happens at least in our context that we have lab data without field data.

*The research of Philip L. H. Yu was substantially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project No. HKU7171/97H). The research of Bimal K. Sinha was funded by a grant under 'Presidential Research Professorship' at UMBC.

**Now at Quintiles Inc., 1801 Rockville Pike, Suite 300, Rockville, MD 20852, U.S.A.

As both field measurement X and lab measurement Y are referred to the same chemical (RVP), it is reasonable to assume that the measurements X and Y have the common mean, denoted by μ, but with possibly unequal variances σ^2 and η^2. Moreover, when a paired measurement (X, Y) is observed, it is natural that X and Y are correlated so that (X, Y) is distributed with mean vector $\mu \mathbf{1}_2$ and variance-covariance matrix Σ, where

$$\Sigma = \begin{bmatrix} \sigma^2 & \xi \\ \xi & \eta^2 \end{bmatrix}, \quad \xi = \rho\sigma\eta, \quad \mathbf{1}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Here, ρ is the correlation coefficient between X and Y. Of course, when only a field measurement X is observed, X is marginally distributed with mean μ and variance σ^2. The goal here is to efficiently estimate the mean RVP μ in gasoline consumed by the public when X and Y follows a bivariate normal distribution.

In practice, a two-phase or double sampling is usually used to collect the above data. This involves the drawing of a random sample of gas pumps in the first phase, in which a crude RVP measurement X is obtained from each gas pump (field); and the drawing of a subsample from the original units in the second phase, in which a more precise RVP measurement Y is obtained from the laboratory. In this case, this is a classical double sampling scheme. Recently, Yu and Lam (1997) demonstrated that the regression estimator is always more efficient when the data are collected using a double sampling with its second-phase sampling being a ranked set sampling (RSS) rather than a simple random sampling (SRS). Therefore, it is worthwhile to consider the problem of point estimation of the common mean μ under two double sampling methods where the first-phase sampling is always simple random sampling and the second-phase sampling is either simple random sampling or ranked set sampling. Hereafter, we refer these two sampling methods as SRS-SRS double sampling and SRS-RSS double sampling.

In this paper, we first consider the case of SRS-SRS double sampling scheme. In Section 2, we discuss the problem of estimating μ when Σ is known. When Σ is unknown, various estimators for Σ are proposed. In Section 3, we discuss the problem of estimating μ when the data are collected using a SRS-RSS double sampling scheme. Other plausible estimators are proposed in Section 4. Numerical comparisons of the relative precision of the proposed common mean estimators under the two sampling schemes and other estimators are discussed in Section 5. We apply the proposed methods to the above practical EPA problem in Section 6. Section 7 gives some concluding remarks.

2. Estimation of μ using SRS-SRS double sampling

In this section we discuss the problem of estimation of μ based on the data collected using a SRS-SRS double sampling scheme. Suppose that a simple random sample of size $n + m$ is drawn in the first phase (field level) and a subsample of size m is drawn in the second phase (lab level). After collecting the measurements at the field and lab, we have two sets of data: the "field only" data $\{z_i, i = 1, \ldots, n\}$, and the paired "(field,lab)" data $\{\boldsymbol{w}_j = (x_j y_j)', j = 1, \ldots, m\}$. They are summarized by a vector $\boldsymbol{t} = (z_1, z_2, \ldots, z_n, x_1, y_1, x_2, y_2, \ldots, x_m, y_m)'$. It is easily seen that the vector $\boldsymbol{t}$ has mean $\mu \mathbf{1}_{n+2m}$ and variance-covariance matrix V, where

$$V = \begin{bmatrix} \sigma^2 \boldsymbol{I}_n & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_m \otimes \Sigma \end{bmatrix}.$$

Here $\otimes$ denotes the Kronecker product, $\boldsymbol{I}_m$ and $\boldsymbol{I}_n$ are identity matrices of orders m and n, respectively.

2.1 *Estimation of μ when Σ is known*

When Σ is known, V is also known. Without any distribution assumption, a natural estimator for μ is to use the generalized least squares (GLS) method which minimizes $(\boldsymbol{t}-\mu\mathbf{1}_{n+2m})'V^{-1}(\boldsymbol{t}-\mu\mathbf{1}_{n+2m})$, leading to the GLS estimator $\hat{\mu}_{srs}$:

$$\begin{aligned}\hat{\mu}_{srs} &= \frac{\mathbf{1}'_{n+2m}V^{-1}\boldsymbol{t}}{\mathbf{1}'_{n+2m}V^{-1}\mathbf{1}_{n+2m}} \\ &= \frac{\frac{n}{\sigma^2}\bar{z}+m\mathbf{1}_2'\Sigma^{-1}\bar{\boldsymbol{w}}}{\frac{n}{\sigma^2}+m\mathbf{1}_2'\Sigma^{-1}\mathbf{1}_2} \\ &= \frac{\frac{n}{\sigma^2}\bar{z}+m\left(\frac{\eta^2-\xi}{\sigma^2\eta^2-\xi^2}\bar{x}+\frac{\sigma^2-\xi}{\sigma^2\eta^2-\xi^2}\bar{y}\right)}{\frac{n}{\sigma^2}+m\frac{\sigma^2+\eta^2-2\xi}{\sigma^2\eta^2-\xi^2}} \qquad (2.1)\end{aligned}$$

where $\bar{z}=n^{-1}\sum_{i=1}^n z_i$, $\bar{\boldsymbol{w}}=(\bar{x},\bar{y})'$ with $\bar{x}=m^{-1}\sum_{j=1}^m x_j$ and $\bar{y}=m^{-1}\sum_{j=1}^m y_j$.

Obviously $\hat{\mu}_{srs}$ is also the MLE of μ under normality assumption, and is always unbiased with variance

$$\mathrm{Var}(\hat{\mu}_{srs})=\frac{1}{\frac{n}{\sigma^2}+m\mathbf{1}_2{}'\Sigma^{-1}\mathbf{1}_2}=\frac{1}{\frac{n}{\sigma^2}+m\frac{\sigma^2+\eta^2-2\xi}{\sigma^2\eta^2-\xi^2}}. \qquad (2.2)$$

2.2 *Estimation of Σ*

Let $s_z^2=\frac{\sum_{i=1}^n(z_i-\bar{z})^2}{(n-1)}$ and

$$A=\sum_{j=1}^m(\boldsymbol{w}_j-\bar{\boldsymbol{w}})(\boldsymbol{w}_j-\bar{\boldsymbol{w}})'=\begin{bmatrix}a_{11} & a_{12}\\ a_{12} & a_{22}\end{bmatrix}=(m-1)\begin{bmatrix}s_x^2 & s_{xy}\\ s_{xy} & s_y^2\end{bmatrix}$$

where

$$s_x^2=\frac{\sum_{i=1}^m(x_i-\bar{x})^2}{(m-1)},\quad s_y^2=\frac{\sum_{i=1}^m(y_i-\bar{y})^2}{(m-1)}\quad\text{and}\quad s_{xy}=\frac{\sum_{i=1}^m(x_i-\bar{x})(y_i-\bar{y})}{(m-1)}.$$

A simple unbiased estimator for Σ is the sample variance-covariance matrix based only on the paired data $\boldsymbol{w}$'s, i.e. $A/(m-1)$. As both s_z^2 and $s_x^2(=a_{11}/(m-1))$ are unbiased for σ^2, a natural unbiased estimator for Σ based on all the data is given by

$$\hat{\sigma}_1^2=\frac{(n-1)s_z^2+(m-1)s_x^2}{n+m-2},\quad \hat{\eta}_1^2=s_y^2,\quad \hat{\xi}_1=s_{xy}. \qquad (2.3)$$

However, it does not guarantee that the resulting $\hat{\Sigma}_1$ is always nonnegative definite(nnd). Below we provide some other estimators for Σ.

2.2.1 *REML and ML estimators of* Σ

Under normality, the well known REML of Σ is obtained by maximizing the marginal likelihood of s_z^2 and A, which is given by:

$$L_1 \propto \frac{1}{(\sigma^2)^{n-1/2}|\Sigma|^{m-1/2}} \exp\left\{-\frac{1}{2}\operatorname{tr}(A\Sigma^{-1}) - \frac{(n-1)s_z^2}{2\sigma^2}\right\}. \tag{2.4}$$

Equating the first derivatives of $\ln L_1$ with respect to the components of Σ to zero and solving the resultant equations lead to the following REML estimator $\hat{\Sigma}_2$ for Σ:

$$\hat{\sigma}_2^2 = \frac{(n-1)s_z^2 + (m-1)s_x^2}{n+m-2} \tag{2.5}$$

$$\hat{\eta}_2^2 = \hat{\sigma}_2^2 \frac{s_{xy}^2}{s_x^4} + \frac{s_x^2 s_y^2 - s_{xy}^2}{s_x^2} \tag{2.6}$$

$$\hat{\xi}_2 = \hat{\sigma}_2^2 \frac{s_{xy}}{s_x^2}. \tag{2.7}$$

Now we discuss the ML estimator for Σ. Let

$$Z(\mu) = \sum_{i=1}^{n}(z_i - \mu)^2, \qquad B(\mu) = \sum_{j=1}^{m}(\boldsymbol{w}_j - \mu\mathbf{1}_2)(\boldsymbol{w}_j - \mu\mathbf{1}_2)' = \begin{bmatrix} b_{11} & b_{12} \\ b_{12} & b_{22} \end{bmatrix}$$

where $b_{11} = \sum_{i=1}^{m}(x_i - \mu)^2$, $b_{12} = \sum_{i=1}^{m}(x_i - \mu)(y_i - \mu)$ and $b_{22} = \sum_{i=1}^{m}(y_i - \mu)^2$. Then under normality, the likelihood function is given by

$$L_2 \propto \frac{1}{(\sigma^2)^{n/2}|\Sigma|^{m/2}} \exp\left\{-\frac{1}{2}\operatorname{tr}(B(\mu)\Sigma^{-1}) - \frac{Z(\mu)}{2\sigma^2}\right\}. \tag{2.8}$$

Note that if we replace n, m, $Z(\mu)$ and $B(\mu)$ in (2.8) by $(n-1)$, $(m-1)$, $(n-1)s_z^2$ and A respectively, L_2 in (2.8) becomes L_1 in (2.4). So, applying the same steps for L_1 to L_2, we obtain the following equations:

$$\hat{\sigma}_3^2 = \frac{Z(\mu) + b_{11}}{n+m} \tag{2.9}$$

$$\hat{\eta}_3^2 = \hat{\sigma}_3^2 \frac{b_{12}^2}{b_{11}^2} + \frac{b_{11}b_{22} - b_{12}^2}{mb_{11}} \tag{2.10}$$

$$\hat{\xi}_3 = \hat{\sigma}_3^2 \frac{b_{12}}{b_{11}}. \tag{2.11}$$

Equations (2.9), (2.10), (2.11) along with the solution to $\frac{\partial \ln L_2}{\partial \mu} = 0$, i.e. (2.1), are the final equations to be used for solving the ML estimators of μ and Σ. To obtain the MLE, we may plug (2.9)–(2.11) into (2.1) to obtain the MLE for μ first, then obtain the MLE of Σ. However, by doing so, it will result in a complicated fifth degree polynomial in μ. Thus closed form expression for the MLE of μ seems impossible and subsequent inference based on it is indeed a difficult task. Hence, we will not consider the ML estimator for Σ in this paper.

2.2.2 *Properties of the REML Estimator for* Σ

In the following we discuss some properties of the REML estimator $\hat{\Sigma}_2$ for Σ and compare it with the ad hoc estimator $\hat{\Sigma}_1$ in (2.3).

Property 1. Validity. It is easy to see from (2.5)–(2.7) that $\hat{\sigma}_2^2$, $\hat{\eta}_2^2$ and $|\hat{\Sigma}_2| = \hat{\sigma}_2^2 \frac{s_x^2 s_y^2 - s_{xy}^2}{s_x^2}$ are positive with probability 1, thus making $\hat{\Sigma}_2$ a valid estimator for Σ.

Property 2. Bias. Clearly, $\hat{\sigma}_2^2$ and $\hat{\xi}_2$ are unbiased but $\hat{\eta}_2^2$ is not. Using the properties of normality and applying simple algebra, it can be shown that

$$E(\hat{\eta}_2^2) = \eta^2 + \eta^2 \frac{2(n-1)(1-\rho^2)}{(n+m-2)(m-1)(m-3)}, \qquad m > 3.$$

Therefore, the bias of $\hat{\eta}_2^2$ and hence the bias of $\hat{\Sigma}_2$ will tend to zero for large m.

Property 3. Mean squared error (MSE). To derive the MSE of $\hat{\Sigma}_2$, we first represent Σ, $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ in vectorized forms:

$$\begin{aligned}
\boldsymbol{\Theta} &= (\theta_1, \theta_2, \theta_3)' = (\sigma^2, \eta^2, \xi)' \\
\hat{\boldsymbol{\Theta}}_1 &= (\hat{\theta}_{11}, \hat{\theta}_{12}, \hat{\theta}_{13})' = (\hat{\sigma}_1^2, \hat{\eta}_1^2, \hat{\xi}_1)' \\
\hat{\boldsymbol{\Theta}}_2 &= (\hat{\theta}_{21}, \hat{\theta}_{22}, \hat{\theta}_{23})' = (\hat{\sigma}_2^2, \hat{\eta}_2^2, \hat{\xi}_2)'.
\end{aligned}$$

The MSE of $\hat{\Sigma}_1$, denoted by $MSE(\hat{\boldsymbol{\Theta}}_1)$, is defined as $E[(\hat{\boldsymbol{\Theta}}_1 - \boldsymbol{\Theta})(\hat{\boldsymbol{\Theta}}_1 - \boldsymbol{\Theta})']$ and the expression for MSE of $\hat{\Sigma}_2$ is similar. It is shown in Appendix I that

$$MSE(\hat{\boldsymbol{\Theta}}_2) = \frac{2}{n+m-2} \begin{bmatrix} \sigma^4 & \xi^2 & \sigma^2\xi \\ \xi^2 & \frac{(n+m-2)-(n-1)d(\rho)}{m-1}\eta^4 & (1+\frac{(n-1)(1-\rho^2)}{m-3})\xi\eta^2 \\ \sigma^2\xi & (1+\frac{(n-1)(1-\rho^2)}{m-3})\xi\eta^2 & \frac{(n+m-4)\sigma^2\eta^2+(m-n-2)\xi^2}{2(m-3)} \end{bmatrix} \tag{2.12}$$

where

$$d(\rho) = \rho^4 - \frac{4}{m-3}\rho^2(1-\rho^2) - \frac{7(m-1)(n+m-2) - 4(n+4m-5)}{(n+m-2)(m-1)(m-3)(m-5)}(1-\rho^2)^2.$$

Of course, it is assumed that $m > 5$.

Comparison of REML estimator $\hat{\Sigma}_2$ *with ad hoc estimator* $\hat{\Sigma}_1$

We now compare the MSE of the REML estimator $\hat{\Sigma}_2$ with that of the ad hoc estimator $\hat{\Sigma}_1$. Although $\hat{\Sigma}_1$ is not always a *valid* estimator in the sense of not being nnd, component-wise comparison however makes sense. It is shown in Appendix I that the MSE of $\hat{\Sigma}_1$ is given by

$$MSE(\hat{\boldsymbol{\Theta}}_1) = \frac{2}{n+m-2} \begin{bmatrix} \sigma^4 & \xi^2 & \sigma^2\xi \\ \xi^2 & \frac{n+m-2}{m-1}\eta^4 & \frac{n+m-2}{m-1}\xi\eta^2 \\ \sigma^2\xi & \frac{n+m-2}{m-1}\xi\eta^2 & \frac{n+m-2}{m-1}(\sigma^2\eta^2+\xi^2) \end{bmatrix}$$

and hence

$$MSE(\hat{\Theta}_1) - MSE(\hat{\Theta}_2) = \frac{2}{n+m-2}\begin{bmatrix} 0 & 0 & 0 \\ 0 & \gamma_{22} & \gamma_{23} \\ 0 & \gamma_{23} & \gamma_{33} \end{bmatrix}$$

where

$$\gamma_{22} = \frac{(n-1)d(\rho)}{m-1}\eta^4, \qquad \gamma_{23} = \frac{(n-1)\xi\eta^2}{m-3}\left(\rho^2 - \frac{2}{m-1}\right),$$
$$\gamma_{33} = \frac{(n-1)(m-2)\sigma^2\eta^2}{(m-1)(m-3)}\left(\rho^2 - \frac{1}{m-2}\right).$$

Therefore, for $m > 5$, we have the following observations:

(a) If $\rho^2 > \frac{1}{m-2}$, then $\gamma_{33} > 0$, i.e., $MSE(\hat{\xi}_1) > MSE(\hat{\xi}_2)$, implying that the REML estimator of ξ is preferred to s_{xy}. Note that both are unbiased for ξ.

(b) If $\rho^2 > \Delta/(1+\Delta)$, where

$$\Delta = \frac{2}{m-3}\left\{1 + \sqrt{\frac{11}{4} + \frac{1}{m-5}\left(\frac{7}{2} - \frac{m-3}{m-1} - \frac{3(m-3)}{n+m-2}\right)}\right\},$$

then $d(\rho) > 0$ and $\rho^2 > \frac{1}{m-2}$, i.e., $\gamma_{22} > 0$, $\gamma_{33} > 0$, implying that the REML estimators of ξ and η^2 are better than s_{xy} and s_y^2, respectively.

(c) To have $MSE(\hat{\Theta}_1) - MSE(\hat{\Theta}_2)$ as nnd, $\gamma_{22}\gamma_{33} - \gamma_{23}^2$ should be positive. It can be shown that

$$\gamma_{22}\gamma_{33} - \gamma_{23}^2 = \frac{4\sigma^2\eta^6(m-2)}{(n+m-2)^2(m-1)^2(m-3)}\left[\rho^6 + O\left(\frac{1}{m}\right)\right].$$

Hence, for large m, we expect it to be positive.

In conclusion, we note that, for large m, the REML estimator $\hat{\Sigma}_2$ for Σ has a smaller MSE compared to the ad hoc estimator $\hat{\Sigma}_1$. Therefore in our subsequent analysis, we will use the REML estimator $\hat{\Sigma}_2$ with its subscript dropped for notational simplicity.

2.3 *Estimation of μ when Σ is unknown*

When Σ is unknown, substituting the REML estimator $\hat{\Sigma} = \hat{\Sigma}_2$ into (2.1) gives

$$\tilde{\mu}_{srs} = \frac{\frac{n}{\hat{\sigma}^2}\bar{z} + m\mathbf{1}_2'\hat{\Sigma}^{-1}\bar{\boldsymbol{w}}}{\frac{n}{\hat{\sigma}^2} + m\mathbf{1}_2'\hat{\Sigma}^{-1}\mathbf{1}_2} = \frac{\frac{n}{\hat{\sigma}^2}\bar{z} + m\left(\frac{\hat{\eta}^2 - \hat{\xi}}{\hat{\sigma}^2\hat{\eta}^2 - \hat{\xi}^2}\bar{x} + \frac{\hat{\sigma}^2 - \hat{\xi}}{\hat{\sigma}^2\hat{\eta}^2 - \hat{\xi}^2}\bar{y}\right)}{\frac{n}{\hat{\sigma}^2} + m\frac{\hat{\sigma}^2 + \hat{\eta}^2 - 2\hat{\xi}}{\hat{\sigma}^2\hat{\eta}^2 - \hat{\xi}^2}}. \tag{2.13}$$

Since $\hat{\Sigma}_2$ is independent of $\bar{z}$ and $\bar{\boldsymbol{w}}$, $\tilde{\mu}_{srs}$ is unbiased for μ with variance given by

$$\mathrm{Var}(\tilde{\mu}_{srs}) = E\{\Psi(\hat{\Theta}, \Theta)\}$$

where

$$\Psi(\hat{\Theta}, \Theta) = \mathrm{Var}(\tilde{\mu}_{srs} \mid \hat{\Sigma}) = \frac{\frac{n\sigma^2}{\hat{\sigma}^4} + m\mathbf{1}_2'\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\mathbf{1}_2}{\left(\frac{n}{\hat{\sigma}^2} + m\mathbf{1}_2'\hat{\Sigma}^{-1}\mathbf{1}_2\right)^2}. \tag{2.14}$$

An exact expression for $\mathrm{Var}(\tilde{\mu}_{srs})$ is usually very difficult to obtain. However, for inference purpose what is really needed is an estimate of $\mathrm{Var}(\tilde{\mu}_{srs})$, for which some approximation methods described below can be used.

Method 1. A naive estimator for $\mathrm{Var}(\tilde{\mu}_{srs})$ is obtained by plugging an estimator $\hat{\Sigma}$ of Σ in $\mathrm{Var}(\tilde{\mu}_{srs} \mid \hat{\Sigma})$ given by (2.14), leading to

$$\dot{M}(\hat{\Sigma}) = \frac{1}{\frac{n}{\hat{\sigma}^2} + m\mathbf{1}_2'\hat{\Sigma}^{-1}\mathbf{1}_2}. \tag{2.15}$$

As pointed out by many investigators, this method is likely to underestimate $\mathrm{Var}(\tilde{\mu}_{srs})$, a phenomenon discussed later in this section.

Method 2. Here we first approximate $\Psi(\hat{\boldsymbol{\Theta}}, \boldsymbol{\Theta})$ by a second-order Taylor expansion:

$$\Psi(\hat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}) \approx \Psi(\boldsymbol{\Theta}, \boldsymbol{\Theta}) + (\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})' \left(\frac{\partial \Psi}{\partial \hat{\boldsymbol{\Theta}}}\right)\Big|_{\hat{\boldsymbol{\Theta}}=\boldsymbol{\Theta}} + \frac{1}{2}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})'\Phi(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})$$

where $\Phi = \left(\frac{\partial^2 \Psi}{\partial \hat{\boldsymbol{\Theta}} \partial \hat{\boldsymbol{\Theta}}'}\right)|_{\hat{\boldsymbol{\Theta}}=\boldsymbol{\Theta}}$, the matrix of second derivatives of Ψ with respect to $\hat{\boldsymbol{\Theta}}$ evaluated at Θ. It can be shown by direct derivation that

$$\Psi(\boldsymbol{\Theta}, \boldsymbol{\Theta}) = \dot{M}(\Sigma) = \frac{1}{\frac{n}{\sigma^2} + m\mathbf{1}_2'\Sigma^{-1}\mathbf{1}_2}, \qquad \left(\frac{\partial \Psi}{\partial \hat{\boldsymbol{\Theta}}}\right)|_{\hat{\boldsymbol{\Theta}}=\boldsymbol{\Theta}} = 0$$

and $\Phi = (\alpha_{ij})_{3\times 3}$ where

$$\begin{aligned}
\alpha_{11} &= \frac{2m}{h(\boldsymbol{\Theta})}\{n[\sigma^2(\sigma^2\eta^2 - \xi^2)(\sigma^2\eta^2 - 2\xi^2) + \xi^4(\sigma^2 + \eta^2 - 2\xi)] + m\sigma^6(\eta^2 - \xi)^2\} \\
\alpha_{12} &= \frac{2m\sigma^2(\xi - \sigma^2)}{h(\boldsymbol{\Theta})}\{n[\sigma^4\eta^2 - 2\sigma^2\xi^2 + \xi^3] + m\sigma^4(\eta^2 - \xi)\} \\
\alpha_{13} &= \frac{2m\sigma^2(\xi - \eta^2)}{h(\boldsymbol{\Theta})}\{n[\sigma^4\eta^2 - 3\sigma^2\xi^2 + 2\xi^3] + m\sigma^4(\eta^2 - \sigma^2)\} \\
\alpha_{22} &= \frac{2m(n+m)\sigma^6(\sigma^2 - \xi)^2}{h(\boldsymbol{\Theta})} \\
\alpha_{23} &= \frac{2m\sigma^4(\sigma^2 - \xi)}{h(\boldsymbol{\Theta})}\{n[\sigma^2\eta^2 - 2\sigma^2\xi + \xi^2] + m\sigma^2(\eta^2 - \sigma^2)\} \\
\alpha_{33} &= \frac{2m\sigma^4}{h(\boldsymbol{\Theta})}\{n[\sigma^2\eta^2(\sigma^2 + \eta^2 - 6\xi) + \xi^2(3\sigma^2 + 3\eta^2 - 2\xi)] + m\sigma^2(\sigma^2 - \eta^2)^2\} \\
h(\boldsymbol{\Theta}) &= [n(\sigma^2\eta^2 - \xi^2) + m\sigma^2(\sigma^2 + \eta^2 - 2\xi)]^3.
\end{aligned}$$

Thus we get

$$\mathrm{Var}(\tilde{\mu}_{srs}) \approx \dot{M}(\Sigma) + \frac{1}{2}E\{(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})'\Phi(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\} = \dot{M}(\Sigma) + \frac{1}{2}\mathrm{tr}\{\Phi[MSE(\hat{\boldsymbol{\Theta}})]\}. \tag{2.16}$$

It is obvious that (2.16) will always give a larger estimator for $\mathrm{Var}(\tilde{\mu}_{srs})$ than $\dot{M}(\Sigma)$. In fact in a general mixed linear model setup, which covers our linear model for $\boldsymbol{t}$ as

a special case, Kackar and Harville (1984) proposed a similar approximation expression for the variance of estimators of fixed and random effects. It is evidenced by their simulations that (2.16) approximates well the actual variance of $\tilde{\mu}_{srs}$. Therefore, $\mathrm{Var}(\tilde{\mu}_{srs})$ is estimated by

$$\widehat{\mathrm{Var}}(\tilde{\mu}_{srs}) = \dot{M}(\hat{\Sigma}) + \frac{1}{2}\,\mathrm{tr}\{\hat{\Phi}[\widehat{MSE}(\hat{\Theta})]\} \tag{2.17}$$

where $\dot{M}(\hat{\Sigma})$ is given by (2.15), $\hat{\Phi}$ and $\widehat{MSE}(\hat{\Theta})$ respectively refer to Φ and $MSE(\hat{\Theta}_2)$ in (2.12), with elements of Σ replaced by $\hat{\Sigma} = \hat{\Sigma}_2$.

3. Estimation of μ using SRS-RSS double sampling

In this section we explore the use of a ranked set sampling (RSS) in place of a simple random sampling in the second-phase of a double sampling. RSS, originally introduced by McIntyre (1952) for efficient estimation of a population mean in a purely nonparametric setup, has been found to be fairly effective in various problems of parametric estimation (see Chuiv and Sinha (1998) and Patil *et al.* (1994) and references therein). In our specific problem, we propose to use the field-only data and paired (field, lab) data in a modified form described as follows.

For a simple random sample of size r units (gas pumps), we collect X-values (field) from all the units. We identify the unit with the smallest X-value and send the corresponding sample to the laboratory to record the Y-value (lab). We next draw another simple random sample of r units, and collect their X-values (field). We identify the unit with the second smallest X-value and send the corresponding sample to the laboratory to record the Y-value (lab). This process is continued in r steps and at the last stage after collecting X-values (field) from all the r units, we identify the unit with the largest X-value and send it to the laboratory to record the Y-value (lab). At the end of this process, what we have collected is a sample of r^2 field measurements and a suitably selected RSS of r lab measurements. The entire process is now repeated N cycles to yield eventually a sample of Nr^2 field measurements and a suitably selected RSS of Nr lab measurements. Denote the measurements recorded in the i-th cycle by

$$\begin{aligned}
&X^{(i)}_{(11)}, \ldots, X^{(i)}_{(1r)}, Y^{(i)}_{[11]}\\
&X^{(i)}_{(21)}, \ldots, X^{(i)}_{(2r)}, Y^{(i)}_{[22]}\\
&\ldots\\
&X^{(i)}_{(r1)}, \ldots, X^{(i)}_{(rr)}, Y^{(i)}_{[rr]}
\end{aligned}$$

where $X^{(i)}_{(jk)}$ is the k-th order statistic (field measurement) in a simple random sample of size r arising out of the j-th sample in the i-th cycle, $i = 1, \ldots, N$, $j, k = 1, \ldots, r$ and $Y^{(i)}_{[kk]}$ is the lab measurement corresponding to the field measurement $X^{(i)}_{(kk)}$ obtained in the i-th cycle, $i = 1, \ldots, N$, $k = 1, \ldots, r$.

Denote the overall mean of X by $\bar{X} = \sum_{i=1}^{N}\sum_{j=1}^{r}\sum_{k=1}^{r} X^{(i)}_{(jk)}/(Nr^2)$, and the sample means of X and Y based on the ranked set sample obtained in the second phase by $\bar{X}_{rss} = \sum_{i=1}^{N}\sum_{j=1}^{r} X^{(i)}_{(jj)}/(Nr)$ and $\bar{Y}_{rss} = \sum_{i=1}^{N}\sum_{j=1}^{r} Y^{(i)}_{[jj]}/(Nr)$ respectively. Note that $\bar{X}$ and $\bar{X}_{rss}$ are always unbiased for μ but $\bar{Y}_{rss}$ may be biased. Suppose X and Y follows the linear model (see David (1973) and Stokes (1977)):

$$Y = \mu + \beta(X - \mu) + \varepsilon \tag{3.1}$$

where $\beta = \rho\eta/\sigma = \xi/\sigma^2$ and ε has zero mean and variance $\eta^2(1-\rho^2)$ and is independent of X. It is easy to show that $\bar{Y}_{rss}$ is unbiased for μ. Since X and Y follow a bivariate normal distribution, the linear model in (3.1) is satisfied and hence $\bar{Y}_{rss}$ is unbiased.

3.1 *Estimation of μ when Σ is known*

As discussed in Section 2, the SRS-SRS double sampling involves drawing of a large random sample of size $n+m$ and a subsample of size m. Based on this sampling method, we derived the MLE for μ when Σ is known as shown in (2.1). Under a SRS-RSS double sampling setting, $n = Nr(r-1)$ and $m = Nr$. After making some obvious changes:

$$\bar{z} = \frac{\sum_{i=1}^{N}\sum\sum_{j\neq k} X_{(jk)}^{(i)}}{Nr(r-1)} = \frac{Nr^2\bar{X} - Nr\bar{X}_{rss}}{Nr(r-1)}, \qquad \bar{x} = \bar{X}_{rss} \quad \text{and} \quad \bar{y} = \bar{Y}_{rss},$$

we propose the following estimator for μ when Σ is known:

$$\begin{aligned}\hat{\mu}_{rss} &= \frac{\dfrac{Nr^2\bar{X} - Nr\bar{X}_{rss}}{\sigma^2} + Nr\left[\dfrac{\eta^2-\xi}{\sigma^2\eta^2-\xi^2}\bar{X}_{rss} + \dfrac{\sigma^2-\xi}{\sigma^2\eta^2-\xi^2}\bar{Y}_{rss}\right]}{\dfrac{Nr(r-1)}{\sigma^2} + Nr\dfrac{\sigma^2+\eta^2-2\xi}{\sigma^2\eta^2-\xi^2}} \\ &= \frac{\dfrac{r}{\sigma^2}\bar{X} + \dfrac{\sigma^2-\xi}{\sigma^2\eta^2-\xi^2}\left(\bar{Y}_{rss} - \dfrac{\xi}{\sigma^2}\bar{X}_{rss}\right)}{\dfrac{1}{\sigma^2}\left(r + \dfrac{(\sigma^2-\xi)^2}{\sigma^2\eta^2-\xi^2}\right)}. \qquad (3.2)\end{aligned}$$

Of course, the above estimator for μ is far from being the MLE under a SRS-RSS sampling. Interestingly enough, it is shown in Appendix II that when (3.1) is satisfied, $\hat{\mu}_{rss}$ given by (3.2) is the best linear unbiased estimator (BLUE) for μ based on $\bar{X}$, $\bar{X}_{rss}$ and $\bar{Y}_{rss}$, and the variance of $\hat{\mu}_{rss}$ is given by

$$\mathrm{Var}(\hat{\mu}_{rss}) = \frac{\sigma^2}{Nr}\cdot\frac{1}{r + \dfrac{(\sigma^2-\xi)^2}{\sigma^2\eta^2-\xi^2}}. \qquad (3.3)$$

Therefore when Σ is known, $\hat{\mu}_{rss}$ is more efficient than $\bar{X}$, $\bar{X}_{rss}$ and $\bar{Y}_{rss}$.

3.2 *Estimation of μ when Σ is unknown*

When Σ is unknown, a standard practice is to start from $\hat{\mu}_{rss}$ given in (3.2) and use a suitable estimator for Σ. In the context of SRS-SRS double sampling discussed in Section 2, it is found that the REML estimator of Σ has some nice properties than other estimators. It is clear that in our context, due to the complicated nature of the likelihood function (due primarily to RSS nature), it is extremely difficult to derive the REML estimator for Σ. In what follows, we adopt the REML estimator for Σ even in our context. Define

$$S_z^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{r}\sum_{k=1\,j\neq k}^{r}(X_{(jk)}^{(i)} - \bar{X})^2}{Nr(r-1)-1}$$

$$S_x^2 = \frac{\sum_{i=1}^{N}\sum_{k=1}^{r}(X_{(kk)}^{(i)} - \bar{X}_{rss})^2}{Nr-1}$$

$$S_y^2 = \frac{\sum_{i=1}^{N}\sum_{k=1}^{r}(Y_{[kk]}^{(i)} - \bar{Y}_{rss})^2}{Nr-1}$$
$$S_{xy} = \frac{\sum_{i=1}^{N}\sum_{k=1}^{r}[(X_{(kk)}^{(i)} - \bar{X}_{rss})(Y_{[kk]}^{(i)} - \bar{Y}_{rss})]}{Nr-1}.$$

Then our proposed estimator for Σ is given by $\hat{\Sigma}_{rss}$ where

$$\hat{\sigma}_{rss}^2 = S_z^2$$
$$\hat{\eta}_{rss}^2 = \hat{\sigma}_{rss}^2 \frac{S_{xy}^2}{S_x^4} + \frac{S_x^2 S_y^2 - S_{xy}^2}{S_x^2}$$
$$\hat{\xi}_{rss} = \hat{\sigma}_{rss}^2 \frac{S_{xy}}{S_x^2}.$$

It may be noted that these estimates are well-defined and valid in the sense of the estimated dispersion matrix being nnd, irrespective of the underlying model. After substituting $\hat{\Sigma}_{rss}$ into (3.2), the resultant estimator of μ is denoted by $\tilde{\mu}_{rss}$.

To prove the unbiasedness of $\tilde{\mu}_{rss}$, we first notice that $\tilde{\mu}_{rss}$ can be expressed as $\mu + \tilde{\mu}^*$ where $\tilde{\mu}^*$ is the $\tilde{\mu}_{rss}$ with X and Y replaced by $X^* = X - \mu$ and $Y^* = Y - \mu$, respectively. Since $\hat{\Sigma}_{rss}$ is an even function of X^* and Y^*, replacing X^* and Y^* by $-X^*$ and $-Y^*$ in $\tilde{\mu}^*$ implies $E[\tilde{\mu}^*] = E[-\tilde{\mu}^*]$. It follows that $E[\tilde{\mu}^*] = 0$ and hence $\tilde{\mu}_{rss}$ is unbiased.

It is clear that the exact variance of $\tilde{\mu}_{rss}$ is difficult to obtain, and in what follows we therefore employ the variance of $\hat{\mu}_{rss}$ given in (3.3) as a large sample approximate of $\mathrm{Var}(\tilde{\mu}_{rss})$ for large N.

4. Other estimators for μ

Note that when the data are collected using a double sampling scheme, a regression estimator is usually used to estimate the population mean of Y based on a covariate X no matter X and Y have common mean or not. Recently, Yu and Lam (1997) proposed a RSS regression estimator based on a SRS-RSS double sampling scheme mentioned in Section 3:

$$\tilde{\mu}_{reg} = \bar{Y}_{rss} + \hat{\beta}(\bar{X} - \bar{X}_{rss}) \tag{4.1}$$

where

$$\hat{\beta} = \hat{\xi}_{rss}/\hat{\sigma}_{rss}^2 = \frac{S_{xy}}{S_x^2} \tag{4.2}$$

is an estimator for the slope β in (3.1). If (3.1) is satisfied and hence normality holds, Yu and Lam (1997) showed that $\tilde{\mu}_{reg}$ is unbiased and its variance is given by:

$$\mathrm{Var}(\tilde{\mu}_{reg}) = \frac{\sigma^2\eta^2 - \xi^2}{\sigma^2 Nr}[1 + \Delta] + \frac{\xi^2}{\sigma^2 Nr^2} \tag{4.3}$$

where

$$\Delta = E\left[\frac{Nr(\bar{X}_{rss} - \bar{X})^2}{(Nr-1)S_x^2}\right] \tag{4.4}$$

and we take $\mu = 0$ and $\sigma = 1$ in the computation of Δ. Obviously, Δ is a fixed constant depending only on N and r.

Of course, a similar SRS regression estimator based on a SRS-SRS double sampling scheme can also be proposed here. However, Yu and Lam (1997) found that under normality, the RSS regression estimator is always superior to the SRS regression estimator for all ρ.

Finally, since $\bar{X}$, $\bar{X}_{rss}$ and $\bar{Y}_{rss}$ do not utilize all the available data and they are inferior than $\hat{\mu}_{rss}$ when Σ is known, we do not intend to consider these estimators although they are unbiased.

In next section, we will compare the two proposed common mean estimators with the RSS regression estimator.

5. Numerical comparisons

Assuming that (X, Y) follows a bivariate normal distribution with common mean $\mu = 0$, we compute the variances of the two proposed common mean estimators $\tilde{\mu}_{srs}, \tilde{\mu}_{rss}$ and the RSS regression estimator $\tilde{\mu}_{reg}$. Since these three estimators are unbiased, we use the variance ratio as a measure of relative precision (RP). The set size examined is $r = 3$, the numbers of cycles are $N = 5, 10$, and the values of ρ are $0, 0.1, 0.2, \ldots, 0.9$. It is easy to see that the RP can be expressed as a function of η/σ and ρ. Without loss of generality, we assume $\sigma = 1$ and consider various choices of $\theta = \eta/\sigma$. As the lab data is expected to be more precise than the field data, θ is usually less than 1. Here, we consider four values of θ: 0.9,0.7,0.3,0.1. The variance of $\tilde{\mu}_{reg}$ is evaluated using (4.3). Because the variances of $\tilde{\mu}_{srs}$ and $\tilde{\mu}_{rss}$ have no exact analytical expressions, their variances are evaluated by a simulation of size 100,000.

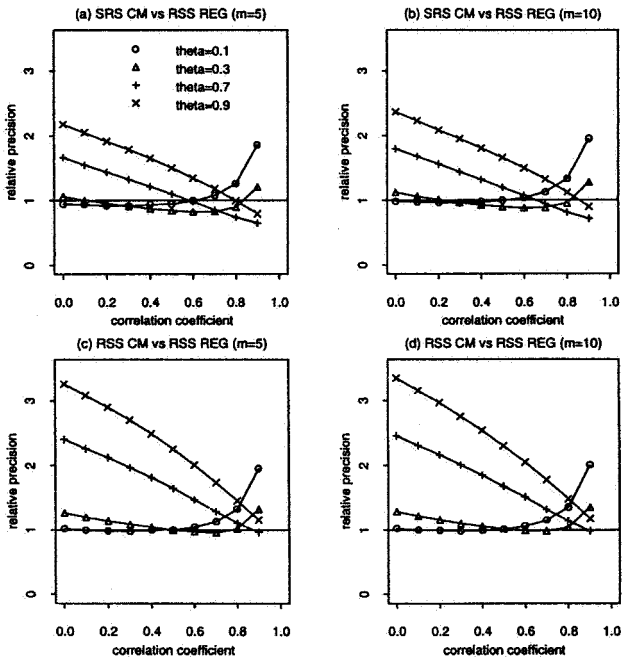

Fig. 1. The relative precision of SRS and RSS common mean estimators relative to RSS regression estimator.

5.1 *Comparison of common mean estimators with RSS regression estimator*

Figure 1 depicts the relative precisions of the two proposed common mean estimators $\tilde{\mu}_{srs}$, $\tilde{\mu}_{rss}$ relative to the RSS regression estimator $\tilde{\mu}_{reg}$. It can be seen that the RSS common mean estimator is almost superior to the RSS regression estimator but not for the SRS common mean estimator. However when θ is large (≥ 0.7 say) and ρ is not too large, both common mean estimators perform significantly better than the RSS regression estimator.

It is not surprising that the RPs of $\tilde{\mu}_{rss}$ to $\tilde{\mu}_{reg}$ are close to 1 when θ is close to 0. Note that when θ is close to 0, X is too variable and becomes nearly useless in estimating μ. Therefore the RSS regression estimator, which aims to estimate the mean of Y, will perform similarly to the RSS common mean estimator. In fact, it can be shown that the RSS common mean estimator $\tilde{\mu}_{rss}$ can be rewritten as a weighted sum of two unbiased estimators $\bar{X}$ and $\tilde{\mu}_{reg}$ with random weights:

$$\tilde{\mu}_{rss} = (1-\hat{a})\bar{X} + \hat{a}\tilde{\mu}_{reg} \quad \text{where} \quad \hat{a} = \frac{1-\hat{\rho}\hat{\theta}}{r\hat{\theta}^2(1-\hat{\rho}^2)+(1-\hat{\rho}\hat{\theta})^2} \tag{5.1}$$

with $\hat{\theta} = \hat{\eta}_{rss}/\hat{\sigma}_{rss}$ and $\hat{\rho} = \hat{\xi}_{rss}/(\hat{\eta}_{rss}\hat{\sigma}_{rss}) = \hat{\beta}/\hat{\theta}$. Note that $\hat{a} = 1$ if and only if $\hat{\theta} = 0$ or $\hat{\theta} = \hat{\rho}/[r(1-\hat{\rho}^2)+\hat{\rho}^2] \equiv \theta_0$. Table 1 lists the values of θ_0 for various choices of $\hat{\rho}$ and $r = 3$. Thus if $\hat{\theta}$ is close to θ_0, $\hat{a}$ is close to 1 and hence the RSS regression estimator is approximately equivalent to the RSS common mean estimator.

As analogy to $\tilde{\mu}_{rss}$ in (5.1), $\tilde{\mu}_{srs}$ can also be expressed as a weighted sum of $\bar{X}$ and the SRS regression estimator with weight $\hat{b}$ having the similar form to $\hat{a}$. Therefore, when θ is close to 0, $\hat{b}$ is likely close to 1 and hence the SRS common mean estimator is close to the SRS regression estimator. Since Yu and Lam (1997) showed that the SRS regression estimator is always less precise than the RSS regression estimator, the SRS common mean estimator perform poorer than the RSS regression estimator when θ is close to 0.

Table 1. The values of θ_0 for various choices of $\hat{\rho}$ and $r = 3$.

$\hat{\rho}$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.99
θ_0	0	0.034	0.068	0.106	0.149	0.200	0.263	0.347	0.465	0.652	0.795	0.952

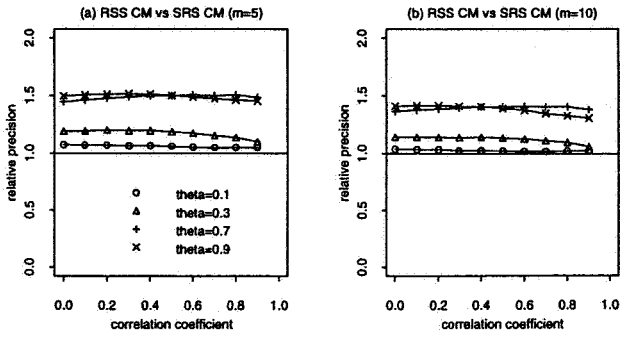

Fig. 2. The relative precision of RSS common mean estimator relative to SRS common mean estimator.

Table 2. The ratios of the approximate variance to the actual variance of $\tilde{\mu}_{rss}$.

r	ρ	$N=5$				$N=10$				$N=15$			
		θ				θ				θ			
		0.1	0.3	0.7	0.9	0.1	0.3	0.7	0.9	0.1	0.3	0.7	0.9
3	0.0	0.97	0.96	0.95	0.95	0.99	0.98	0.98	0.98	0.99	0.99	0.99	0.99
	0.3	0.97	0.95	0.95	0.95	0.98	0.97	0.97	0.97	0.99	0.98	0.99	0.99
	0.6	0.96	0.95	0.95	0.95	0.98	0.98	0.98	0.98	0.99	0.98	0.99	0.99
	0.9	0.96	0.96	0.95	0.96	0.99	0.98	0.97	0.98	1.00	0.98	0.98	0.99
5	0.0	0.98	0.97	0.97	0.98	0.99	0.98	0.99	0.99	0.99	0.99	0.99	1.00
	0.3	0.99	0.98	0.98	0.99	1.00	0.99	0.99	0.99	1.00	1.00	1.00	1.00
	0.6	0.98	0.97	0.98	0.99	1.00	0.99	0.99	1.00	0.99	0.99	0.99	1.00
	0.9	0.99	0.98	0.98	0.99	0.98	0.99	1.00	1.00	1.00	0.99	0.99	1.00

5.2 *Comparison of RSS common mean estimator with SRS common mean estimator*

Figure 2 depicts the relative precision of the RSS common mean estimator $\tilde{\mu}_{rss}$ relative to the SRS common mean estimator $\tilde{\mu}_{srs}$. It is easily seen that the RSS common mean estimator always performs better than the SRS common mean estimator. It should be noted that the values of RPs mainly depends on the value of θ only and they are significantly greater than 1 for large θ. This indicates that when the variances of X and Y are close, a double sampling scheme with its second stage being a ranked set sampling can provide a more precise common mean estimator than the one with its second stage being a simple random sampling.

5.3 *Comparison of the approximate variance and the actual variance for RSS common mean estimator*

Table 2 presents the ratios of the approximate variance to the actual variance for the RSS common mean estimator $\tilde{\mu}_{rss}$ for various combinations of θ and ρ. The approximate variance is computed by using (3.3) while the actual variance is obtained from the above-mentioned simulation based on a bivariate normal distribution. The set size examined is $r = 3, 5$ and the number of cycles is $N = 5, 10, 15$. It can be seen from Table 1 that although the ratios are all less than 1, they vary in a very narrow range from 0.95 to 1.00. This indicates that the approximate variance a little bit underestimates the actual variance of $\tilde{\mu}_{rss}$. The ratios are very close to 1 when the ranked set sample size is moderately large, says $Nr > 30$. This concludes that the approximate variance expression given in (3.3) provides a robust and close-form expression for the variance of $\tilde{\mu}_{rss}$ even the ranked set sample is of moderate size.

6. Application to an EPA data set

In this section, we return to the practical problem of estimating the mean of Reid Vapor Pressure (RVP) of the new reformulated gasoline in the U.S. Since the laboratory analyses are costly, a SRS-RSS double sampling scheme is adopted to reduce the quantity of laboratory analyses and hence save cost. Here a SRS-RSS double sampling scheme with set size $r = 3$ and number of cycles $N = 5$ is used to draw the sample and the field (X) and lab measurements (Y) in the sample are then collected. Table 3 presents the data on X and Y and their summary statistics are shown in Table 4.

Table 3. The field and lab data on RVP for new reformulated gasoline* (bold numbers indicate the selected X in the second phase).

X			Y
8.03	8.09	8.46	8.28
7.37	**8.64**	8.80	8.63
7.59	8.62	**9.14**	9.28
7.86	7.88	7.98	7.85
7.47	**8.70**	8.90	8.62
8.51	8.69	**9.28**	9.14
7.86	7.93	7.96	7.86
7.45	**7.83**	8.02	7.90
7.32	7.45	**8.60**	8.52
7.83	7.86	7.88	7.92
7.39	**7.88**	8.03	7.89
7.31	7.44	**8.56**	8.48
7.83	7.95	7.92	7.95
7.53	**7.99**	8.01	8.32
7.16	7.31	**7.56**	7.60

* Data Source: Private Communication

Table 4. Summary statistics for the crude RVP measurement X and the accurate RVP measurement Y.

r	N	$\bar{X}$	$\bar{Y}_{rss}$	$\bar{X}_{rss}$	S_z^2	S_x^2	S_y^2	S_{xy}
3	5	7.997	8.283	8.239	0.252150	0.284778	0.245392	0.256838

Table 5. Point estimates, standard errors and relative precisions of estimators for μ.

	Benchmark estimators			RSS regression	RSS common mean
	$\bar{X}_{rss}$	$\bar{Y}_{rss}$	$\bar{X}$	estimator, $\tilde{\mu}_{reg}$	estimator, $\tilde{\mu}_{rss}$
Point estimate	8.239	8.283	7.997	8.064	8.035
Standard error	0.0937	0.0898	0.0749	0.0741	0.0727
RP*	100%	109.0%	156.8%	160.0%	166.0%

* RP = relative precision with $\bar{X}_{rss}$ as the base

Using the summary statistics in Table 4, we have $\hat{\sigma}_{rss}^2 = 0.252$, $\hat{\eta}_{rss}^2 = 0.219$, $\hat{\xi}_{rss} = 0.227$, $\hat{\beta} = 0.902$, $\hat{\theta} = 0.932$ and $\hat{\rho} = 0.968$. Based on these statistics, we can compare the performance of RSS common mean estimator $\tilde{\mu}_{rss}$ and the RSS regression estimators $\tilde{\mu}_{reg}$. Three unbiased estimators $\bar{X}$, $\bar{Y}_{rss}$ and $\bar{X}_{rss}$ are also considered as benchmarks. Table 5 shows their point estimates, standard errors, and relative precisions.

It can be seen from Table 5 that the RSS common mean estimator $\tilde{\mu}_{rss}$ attains the smallest precisions (about 66% increase over the worst benchmark and 6% increase over the best benchmark). This result is not surprising because since in this example

$\hat{a} = 0.566$, $\tilde{\mu}_{rss}$ is approximately an average of $\bar{X}$ and $\tilde{\mu}_{reg}$. Simply using either $\bar{X}$ or $\tilde{\mu}_{reg}$ cannot beat $\tilde{\mu}_{rss}$.

7. Concluding remarks

In this paper, we proposed two common mean estimators and showed that the proposed RSS common mean estimator is more precise than the other estimators including Yu and Lam's (1997) RSS regression estimator, McIntyre's (1952) RSS naive estimator and the proposed SRS common mean estimator. Simulation study performed in Section 4 shows that the approximate variance expression given in (3.3) provides a robust estimate for the actual variance of the RSS common mean estimator even when the sample size is moderate large.

Apart from the problem of estimating the common mean μ, it is also of interest to consider the problems of constructing hypotheses testing and a confidence interval (CI) for μ. As long as the tests and confidence intervals based separately on the 'field-only' data and the paired data are available, we can adopt various combination techniques described in Yu *et al.* (1999) to combine the tests and hence construct a confidence interval for μ by converting the acceptance region of the combined test. For example using the sample drawn by a SRS-SRS double sampling scheme as in Section 3, it is well known that based on $(\bar{z}, s_z^2)$ only, we can use the one-sample t-test to test for $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$, where μ_0 is a given constant, and its test statistic is given by

$$t_1 = \frac{\bar{z} - \mu_0}{\sqrt{s_z^2/n}}$$

which follows a t distribution with $n-1$ d.f. under H_0 and its associated $100(1-\alpha)\%$ CI for μ is

$$\{\mu_0 : |t_1| < t_{\alpha/2,n-1}\} = \left(\bar{z} - t_{\alpha/2,n-1}\sqrt{\frac{s_z^2}{n}}, \bar{z} + t_{\alpha/2,n-1}\sqrt{\frac{s_z^2}{n}}\right)$$

where $t_{\alpha/2,n-1}$ is the upper $\alpha/2$-point of the t_{n-1} distribution. Similarly, based on the paired data (x_i, y_i)'s, we can derive a likelihood ratio test (LRT) for H_0 and the equivalent test statistic is given by

$$t_2 = \frac{\bar{u} - \mu_0 - \dfrac{\bar{v}s_{uv}}{s_v^2}}{\sqrt{h}}$$

where

$$\bar{u} = \frac{\bar{x}+\bar{y}}{2}, \quad \bar{v} = \frac{\bar{y}-\bar{x}}{2}, \quad s_u^2 = \frac{1}{4}(s_x^2 + 2s_{xy} + s_y^2), \quad s_v^2 = \frac{1}{4}(s_x^2 - 2s_{xy} + s_y^2),$$
$$s_{uv} = \frac{1}{4}(s_x^2 - s_y^2) \quad \text{and} \quad h = \frac{1}{m-2}\left(\frac{m-1}{m} + \frac{\bar{v}^2}{s_v^2}\right)\frac{(s_u^2 s_v^2 - s_{uv}^2)}{s_v^4},$$

which follows a t distribution with $m-2$ d.f. under H_0 and its associated the $100(1-\alpha)\%$ CI of μ is

$$\{\mu_0 : |t_2| < t_{\alpha/2,m-2}\} = \left(\bar{u} - \frac{\bar{v}s_{uv}}{s_v^2} - t_{\alpha/2,m-2}\sqrt{h}, \bar{u} - \frac{\bar{v}s_{uv}}{s_v^2} + t_{\alpha/2,m-2}\sqrt{h}\right).$$

Let $F_1 = t_1^2$ and $F_2 = t_2^2$ so that $F_1 \sim F_{1,n-1}$ and $F_2 \sim F_{1,m-2}$. Define the p-values based on the two F-statistics as $P_1 = \int_{F_1}^{\infty} f_{1,n-1}(x)dx$ and $P_2 = \int_{F_2}^{\infty} f_{1,m-2}(x)dx$, where $f_{1,k}$ denotes the pdf of the $F_{1,k}$ distribution. Following Yu *et al.* (1999), we can combine these two t_i's or F_i's or P_i's to test for H_0 and hence construct confidence intervals for μ.

Appendix I: MSEs of $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$

MSEs of $\hat{\Sigma}_1$

Following the notations in Section 2, we first note that

$$\hat{\boldsymbol{\Theta}}_1 = \begin{bmatrix} \hat{\sigma}_1^2 \\ \hat{\eta}_1^2 \\ \hat{\xi}_1 \end{bmatrix} = \begin{bmatrix} \dfrac{(n-1)s_z^2}{n+m-2} \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} \dfrac{a_{11}}{n+m-2} \\ \dfrac{a_{22}}{m-1} \\ \dfrac{a_{12}}{m-1} \end{bmatrix}.$$

Since s_z^2 is independent of A and $\hat{\boldsymbol{\Theta}}_1$ is unbiased,

$$MSE(\hat{\boldsymbol{\Theta}}_1) = \mathrm{Var}(\hat{\boldsymbol{\Theta}}_1) = \mathrm{Var}\left(\begin{bmatrix} \dfrac{(n-1)s_z^2}{n+m-2} \\ 0 \\ 0 \end{bmatrix}\right) + \mathrm{Var}\left(\begin{bmatrix} \dfrac{a_{11}}{n+m-2} \\ \dfrac{a_{22}}{m-1} \\ \dfrac{a_{12}}{m-1} \end{bmatrix}\right).$$

Using the result from Muirhead ((1982), p. 90) that if $H = (h_{ij}) \sim W_p(\Sigma, m)$, then $\mathrm{Cov}(h_{ij}, h_{kl}) = m(\sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk})$, where σ_{ij} is the ij-th element of Σ and the fact that $\mathrm{Var}(s_z^2) = 2\sigma^4/(n-1)$ and $A \sim W_2(\Sigma, m-1)$, the expression for $MSE(\hat{\boldsymbol{\Theta}}_1)$ is then obtained.

MSEs of $\hat{\Sigma}_2$

Using Theorem 3.2.10 of Muirhead (1982) and basic properties on conditional moments, we first obtain some preliminary results:

(a) $a_{12} \mid a_{11} \sim N(\frac{\rho\eta}{\sigma} a_{11}, \eta^2(1-\rho^2)a_{11})$.

(i) $E(a_{12}^2 \mid a_{11}) = a_{11}\eta^2(1-\rho^2) + a_{11}^2 \dfrac{\rho^2\eta^2}{\sigma^2}$.

(ii) $\mathrm{Var}(a_{12}^2 \mid a_{11}) = 2a_{11}^2\eta^4 \left[(1-\rho^2)^2 + 2a_{11}\dfrac{\rho^2(1-\rho^2)}{\sigma^2}\right]$.

(b) $a_{22} - \frac{a_{12}^2}{a_{11}} \sim \eta^2(1-\rho^2)\chi_{m-2}^2$ and is independent of a_{11} and a_{22}.

(i) $E(a_{22} \mid a_{11}) = (m-1)\eta^2(1-\rho^2) + a_{11}\dfrac{\rho^2\eta^2}{\sigma^2}$.

(ii) $\mathrm{Cov}(a_{12}^2, a_{22} \mid a_{11}) = \mathrm{Var}(a_{12}^2 \mid a_{11})/a_{11} = 2a_{11}\eta^4 \left[(1-\rho^2)^2 + 2a_{11}\dfrac{\rho^2(1-\rho^2)}{\sigma^2}\right]$.

(iii) $\mathrm{Var}(a_{22} \mid a_{11}) = 2\eta^4 \left[(m-1)(1-\rho^2)^2 + 2a_{11}\dfrac{\rho^2(1-\rho^2)}{\sigma^2}\right]$.

To derive the MSE of $\hat{\Sigma}_2$, we first note that

$$E\hat{\sigma}_2^2 = \sigma^2, \quad E\hat{\xi}_2 = \xi, \quad E\hat{\eta}_2^2 = \eta^2\left(1 + \frac{2(n-1)(1-\rho^2)}{(n+m-2)(m-1)(m-3)}\right).$$

So,

- $MSE(\hat{\sigma}_2^2) = \mathrm{Var}(\hat{\sigma}_2^2) = \mathrm{Var}(\hat{\sigma}_1^2) = \frac{2\sigma^4}{n+m-2}$.
- $MSE(\hat{\eta}_2^2) = \mathrm{Var}(\hat{\eta}_2^2) + [E(\hat{\eta}_2^2) - \eta^2]^2$.
- $MSE(\hat{\xi}_2) = \mathrm{Var}(\hat{\xi}_2)$.
- $E(\hat{\sigma}_2^2 - \sigma^2)(\hat{\eta}_2^2 - \eta^2) = E\hat{\sigma}_2^2\hat{\eta}_2^2 - \sigma^2 E\hat{\eta}_2^2$.
- $E(\hat{\sigma}_2^2 - \sigma^2)(\hat{\xi}_2 - \xi) = E\hat{\sigma}_2^2\hat{\xi}_2 - \sigma^2\xi$.
- $E(\hat{\eta}_2^2 - \eta^2)(\hat{\xi}_2 - \xi) = \mathrm{Cov}(\hat{\eta}_2^2, \hat{\xi}_2)$.

The rest of the derivation of $MSE(\hat{\boldsymbol{\Theta}}_2)$ follows by using the previous preliminary results.

Appendix II: The BLUE of μ based on $\bar{X}$, $\bar{X}_{rss}$ and $\bar{Y}_{rss}$

Consider a linear estimator of μ:

$$L = a\bar{X} + b\bar{X}_{rss} + c\bar{Y}_{rss} \quad \text{with} \quad a+b+c = 1. \tag{A.1}$$

We can write $\mathrm{Var}(L) = \mathrm{Var}[E(L \mid X)] + E[\mathrm{Var}(L \mid X)]$. Under (3.1), we get (taking $\mu = 0$ without any loss of generality)

$$\begin{aligned} E[L \mid X] &= a\bar{X} + b\bar{X}_{rss} + c\beta\bar{X}_{rss} \\ &= a\bar{X} + (b + c\beta)\bar{X}_{rss} \\ \mathrm{Var}[L \mid X] &= c^2\eta^2(1-\rho^2)/(Nr). \end{aligned} \tag{A.2}$$

To compute $\mathrm{Var}(E[L \mid X])$, we first condition on all X, denoted as S, and treat the selection of RSS as random and then uncondition on X. Since, given S, $\bar{X}$ is fixed, we get $\mathrm{Var}(E[L \mid X] \mid S) = (b + c\beta)^2\,\mathrm{Var}(\bar{X}_{rss} \mid S)$ and hence

$$\begin{aligned} \mathrm{Var}(E[L \mid X]) &= \mathrm{Var}\{E(E[L \mid X] \mid S)\} + E[\mathrm{Var}(E[L \mid X] \mid S)] \\ &= (1 - c + c\beta)^2\frac{\sigma^2}{Nr^2} + (b + c\beta)^2 E[\mathrm{Var}(\bar{X}_{rss} \mid S)]. \end{aligned} \tag{A.3}$$

Combining (A.2) and (A.3), we get $\mathrm{Var}(L)$. Clearly, for a given c, $\mathrm{Var}(L)$ is minimized when $b = -c\beta$ and $\mathrm{Var}(L)$ becomes

$$\mathrm{Var}(L) = (1 - c + c\beta)^2\frac{\sigma^2}{Nr^2} + c^2\frac{\eta^2(1-\rho^2)}{Nr}. \tag{A.4}$$

The optimum value of c is easily obtained by minimizing the above quadratic function in c and turns out to be $c_{opt} = \frac{1-\beta}{(1-\beta)^2 + r(\theta^2 - \beta^2)}$. Substituting $a = 1 - b_{opt} - c_{opt}$, $b_{opt} = -c_{opt}\beta$ and c_{opt} into (A.1) and (A.4), the resulting optimum linear unbiased estimator of μ is precisely $\hat{\mu}_{rss}$ with variance as shown in (3.3).

References

Chuiv, N. and Sinha, B. K. (1998). On some aspects of ranked set sampling in parameter estimation, *Handbook of Statistics 17* (eds. N. Balakrishnan and C. R. Rao), 337–377, Elsevier, North-Holland, Amsterdam.

David, H. A. (1973). Concomitants of order statistics, *Bulletin of the International Statistical Institute*, **45**(1), 295–300.

Kackar, R. N. and Harville, D. A. (1984). Approximations for standard errors of fixed and random effects in mixed linear models, *J. Amer. Statist. Assoc.*, **79**, 853–862.

McIntyre, G. A. (1952). A method of unbiased selective sampling using ranked sets, *Australian Journal of Agricultural Research*, **3**, 385–390.

Muirhead, R. J.(1982). *Aspects of Multivariate Statistical Theory*, Wiley, New York.

Patil, G. P., Sinha, A. K. and Taillie, C. (1994). Ranked set sampling, *Handbook of Statistics 12* (eds. G. P. Patil and C. R. Rao), 167–200, Elsevier, North-Holland, Amsterdam.

Stokes, S. L. (1977). Ranked set sampling with concomitant variables, *Comm. Statist. Theory Methods*, **6**, 1207–1211.

Yu, P. L. H. and Lam, K. (1997). Regression estimator in ranked set sampling, *Biometrics*, **53**, 1070–1080.

Yu, P. L. H., Sun, Y. and Sinha, B. K. (1999). On exact confidence intervals for the common mean of several normal populations, *J. Statist. Plann. Inference*, **81**(2), 263–277.