Bootstrap in Common Mean Estimation – a Case Study

B. Arendacká

Institute of Measurement Science, Slovak Academy of Sciences, Dúbravská cesta 9,
841 04 Bratislava, Slovakia
Email: barendacka@gmail.com

Abstract. *It is known that in small samples the asymptotic variance of a maximum likelihood estimator for the common mean in random effects model underestimates the true variance and leads to too short confidence intervals for the true consensus value. This is illustrated e.g. in Rukhin, Metrologia 46(323-31), 2009. We look at two concrete small sample situations to investigate the possibility of improving the confidence intervals by employing bootstrap.*

Keywords: Bootstrap, Common Mean, Confidence Intervals, Maximum Likelihood

1. Introduction

Estimation of the common mean, determination of its uncertainty and of a confidence interval for its true value are tasks arising in certification of reference materials or in interlaboratory studies, see e.g. [1], [2]. The model used is very often the random effects model allowing for different within-laboratory variances and the estimates of the common mean are various weighted means of the estimates supplied by the different laboratories. Procedures of this kind are discussed at length in [1] and it is stressed that they differ in how their variance is estimated. However, this is of crucial importance and influences in turn the quality of the derived confidence intervals. In situations when there is no simple formula for the variance of a procedure or for the distribution of an (approximate) pivot underlying the construction of confidence intervals, bootstrap (see [3]) is a generic method that may be used to overcome the difficulties. By nature, it is an asymptotic method; however, in reality (when e.g. certifying a reference material) it is not uncommon to have a relatively small number of observations coming from only a few laboratories. In this paper we will study performance of bootstrap confidence intervals for the common mean in two such situations. The model and methods used are described in detail in Section 2. Section 3 summarizes results of our simulation study and Section 4 offers some concluding remarks.

2. Model and Methods

The model used for describing measurements of essentially the same quantity obtained in k laboratories is

$$y_{ij}=\mu+ b_i+ e_{ij}\ ,\ i=1,...,k>1, j=1,...,n_i>1 \qquad (1)$$

where y_{ij} denotes the j-th observation in the i-th laboratory, μ is the unknown common mean, $b_i \sim N(0,\tau^2)$ is the random effect of the i-th laboratory and $e_{ij} \sim N(0,\sigma_i^2)$ are random errors. All b_is and e_{ij}s are assumed to be mutually independent, $\tau^2 \geq 0,\ \sigma_i^2 >0,\ i=1,...,k,$ are unknown nuisance parameters. The resulting model for the laboratory means $y_i=\Sigma_j y_{ij}/n_i$ is $y_i \sim N(\mu,\tau^2+\theta_i^2),$ where $\theta_i^2=\sigma_i^2/n_i$. An unbiased estimator of θ_i^2 is $u_i^2=\Sigma_j(y_{ij}\ -\ y_i)^2/[n_i(n_i-1)]\sim\theta_i^2\chi_i^2/(n_i-1)$, where χ_i^2 denotes a χ^2distribution with $n_i\,-1$ degrees of freedom.

Estimators for μ of the form $\Sigma_i w_i y_i$, $\Sigma_i w_i=1$ were considered in [1]. We will use two of them: the maximum likelihood (ML) estimator, $\mu_{ML},$ and the DerSimonian-Laird estimator, μ_{DL}. For details on ML estimators of the unknown parameters in model (1) see [4]. Denoting them $\mu_{ML},$

$\tau^2{}_{ML}$, $\theta^2{}_{iML}$, $i=1,...,k$, asymptotic considerations lead to the variance of μ_{ML} being estimated as

$$Var_A(\mu_{ML})=(\Sigma_i\, 1/(\tau^2{}_{ML}+\theta^2{}_{iML}))^{-1} \tag{2}$$

and the corresponding 95% confidence interval for μ being

$$\mu_{ML}\pm q_{0.975}\sqrt{}Var_A(\mu_{ML}), \tag{3}$$

where $q_{0.975}$ denotes the 97.5th quantile of $N(0,1)$. This interval was found in [1] to be sometimes too short, i.e. its coverage was lower than the nominal level. The interval based on μ_{DL} suggested in [1], which performed quite well in the simulations therein, is of the form

$$\mu_{DL} \pm t_{0.975,k-1}\sqrt{}Var_w(\mu_{DL}), \tag{4}$$

where $t_{0.975,k-1}$ denotes the 97.5th quantile of the t-distribution with $k-1$ degrees of freedom and

$$Var_w(\mu_{DL})=\Sigma_i w^2{}_{iDL}(y_i-\mu_{DL})^2/(1-w_{iDL}) \tag{5}$$

where $w_{iDL}=v_{iDL}/\Sigma_i v_{iDL}$, $v_{iDL}=1/(\tau^2{}_{DL}+u_i^2)$ and $\tau^2{}_{DL}=max(0,[\Sigma_i u_i^{-2}(y_i-y_0)^2-k+1]/[\Sigma_i u_i^{-2}-\Sigma_i u_i^{-4}(\Sigma_i u_i^{-2})^{-1}])$, $y_0=\Sigma_i u_i^{-2}y_i/\Sigma_i u_i^{-2}$.

Since $\mu_{ML}=\Sigma_i w_{iML}y_i$, with $w_{iML}=v_{iML}/\Sigma_i v_{iML}$, $v_{iML}=1/(\tau^2{}_{ML}+\theta^2{}_{iML})$ (see [4]), its variance may be estimated similarly to the variance (5) of μ_{DL}, see also [1], p. 327, so that

$$Var_w(\mu_{ML})=\Sigma_i w^2{}_{iML}(y_i-\mu_{ML})^2/(1-w_{iML}). \tag{6}$$

However, what quantile should be used in combination with this estimator to form a confidence interval for μ is not clear. This difficulty can be avoided by employing bootstrap. In this paper we will consider only bootstrap t-intervals, which are suited especially for location parameters, see [3], p. 161. A 95% bootstrap t-interval is derived as follows:

1. Based on y_is, $u^2{}_i$s estimate the unknown parameters μ_{est}, $\tau^2{}_{est}$, $\theta^2{}_{i\,est}$, $i=1,...,k$ and $Var(\mu_{est})$.

2. Generate N_B bootstrap samples from model (1) with the unknown parameters replaced by their estimates from the step 1.

3. For each of the N_B bootstrap samples, estimate the unknown common mean and its variance, μ_{estB}, $Var(\mu_{estB})$, and compute $T=(\mu_{estB}-\mu_{est})/\sqrt{}Var(\mu_{estB})$.

4. The interval for μ *is* $[\mu_{est} - q_{T,0.975}\sqrt{}Var(\mu_{est}),\ \mu_{est} - q_{T,0.025}\sqrt{}Var(\mu_{est})]$, where $q_{T,0.025}$ $(q_{T,0.975})$ denotes the 2.5th (97.5th) quantile of the distribution of T (estimated from the N_B values of T).

In our simulation study we considered model (1) with $k=3$, $n_1=10$, $n_2=10$, $n_3=12$, $\mu=0$, $\theta^2{}_1=2.7$, $\theta^2{}_2=1.9$, $\theta^2{}_3=0.5$ (case I) and $\theta^2{}_3=2.1$ (case II). τ^2 was 0, $0.25*m$, m, $(1+M)/2$, M, $4*M$, $m=min(\theta^2{}_i)$, $M=max(\theta^2{}_i)$, a choice inspired by [5]. For each scenario, we simulated 1000 sets of observations of (1) and constructed appropriate confidence intervals for μ. Based on these 1000 intervals we estimated the coverage of the respective procedures. We considered interval (4) and its bootstrap version (using μ_{DL}, $Var_w(\mu_{DL})$), interval (3) and its bootstrap version (μ_{ML}, $Var_A(\mu_{ML})$), as well as a bootstrap interval based on μ_{ML}, $Var_w(\mu_{ML})$. *F*or obtaining bootstrap t-intervals $N_B=1500$ was used. Computations were done in R.

τ^2	*0*	*.25m*	*m*	*.5(1+M)*	*M*	*4M*		*0*	*.25m*	*m*	*.5(1+M)*	*M*	*4M*
$\mathrm{DL_A}$	4.08	4.39	4.94	7.30	8.15	14.0		6.01	6.53	8.32	8.07	9.02	14.4
$\mathrm{DL_B}$	5.05	5.65	6.30	9.92	11.7	19.5		6.91	7.66	9.68	9.42	10.8	16.9
$\mathrm{ML_{Bw}}$	4.71	4.95	5.42	7.08	7.92	23.1		6.76	7.13	8.80	8.62	9.62	18.0

Table 1: Median lengths of the simulated 95% confidence intervals in case I (left) and II (right). Notation is the same as in Figure 1.

3. Results

Figure 1 shows simulated probabilities of coverage of the different intervals in the two cases of model (1) as described in the previous Section. We see that the approximate interval (4) based on the DerSimonian-Laird estimator may have lower probability of coverage than the nominal level, especially when the within-laboratory variances are substantially different (case I). It is also clear that the employment of bootstrap improves the performance of the interval. In case of the ML estimator and the associated intervals, interval (3), as expected, does not maintain the desired probability of coverage. Bootstrap results in an improved behaviour of the interval, but a real improvement appears only in combination with the modified estimator of the variance (6). A comparison of the length of the different intervals makes sense only in cases when the nominal confidence level is preserved. In Table 1 we state median lengths obtained in the simulations for 3 of the considered intervals for which the actual probability of coverage was roughly satisfactory. We see that the improvement in the probability of coverage resulting from employing bootstrap in case of interval (4) does not result in a too dramatic increase in the length.

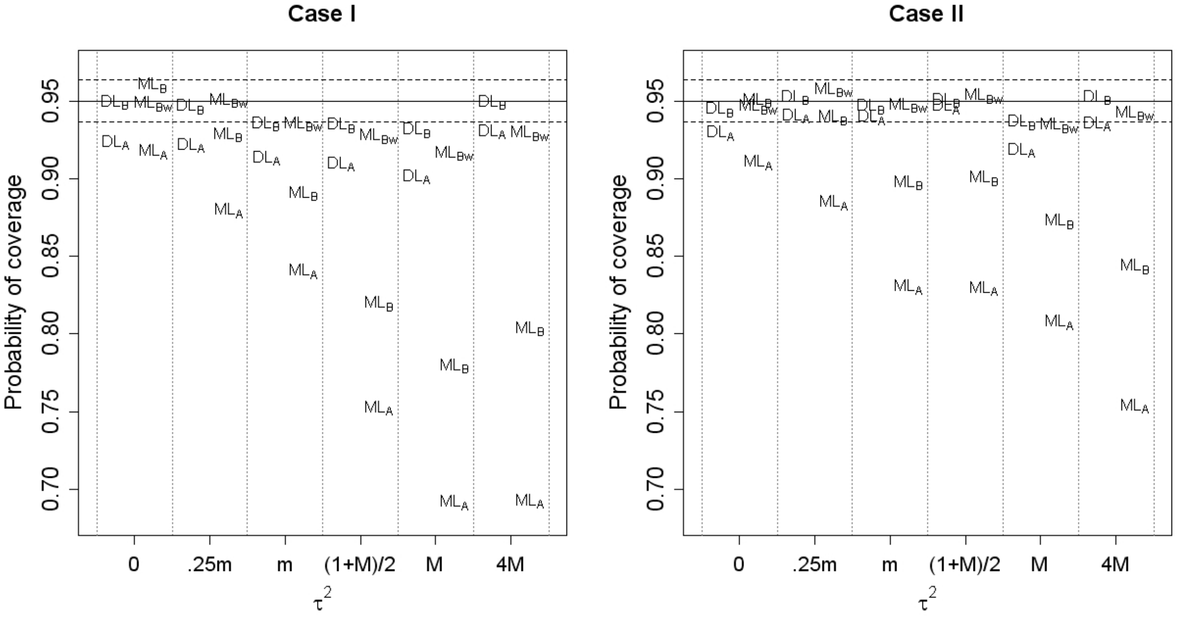

Fig. 1: Simulated probabilities of coverage for the two cases of model (1). DL$_A$ denotes interval (4), DL$_B$ its bootstrap version, ML$_A$ denotes interval (3), ML$_B$ its bootstrap version and ML$_{Bw}$ a bootstrap interval using (6). Shown is the nominal level 0.95 together with limits which the simulated probability of coverage should fall into with probability 0.95 if the true coverage is 95%.

4. Discussion

Even though further investigation is needed to clarify the matter, the studied cases show a potential for an improvement in the probability of coverage of confidence intervals for the common mean in small samples when bootstrap is employed. However, a naive application of bootstrap may not be of help, as can be seen from the case of the ML estimator, when only bootstrap combined with a modified estimator of the variance led to a meaningful increase in the probability of coverage. The assumption for the bootstrap t-intervals to work well is that the quantity $T^*=(\mu_{est} -\mu)/\sqrt{Var(\mu_{est})}$ is an approximate pivot, i.e. its distribution is (approximately) independent of the unknown parameters. In the considered model, this means independence not only of the parameter of interest, the common mean, but also of the nuisance parameters τ^2, θ^2_i, $i=1,...,k$. For μ_{ML} this condition seems to be better satisfied with (6) than with (2) as can be seen from Figure 2 comparing the distributions of the respective T^* when $\tau^2=0$ and when $\tau^2=10$ in case I considered in our simulations.

Although not reported, we examined also bootstrap percentile intervals (such an interval is formed by the lower and upper quantiles of μ_{estB} estimated from the N_B values of μ_{estB}), but except for the case when $\tau^2=0$ (and sometimes $\tau^2=0.25*m$), the obtained probability of coverage was unsatisfactory.

Currently, further investigation into the performance of bootstrap in combination with weighted means estimators of μ is being undertaken.

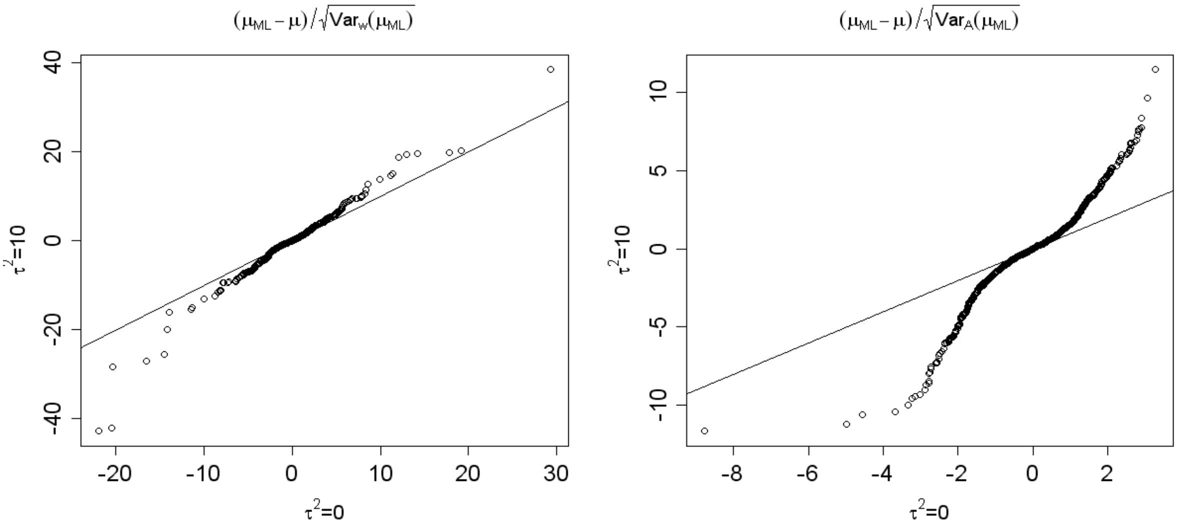

Fig. 2: QQplots comparing the distributions of T* with (6) (left) and with (2) (right) for two different values of τ^2 in case I of model (1). The closer to the identity line the points lie, the more alike the two distributions

Acknowledgements

The author wishes to thank Dr. Antonio Possolo for an introduction to the topic. The work was initiated during the author's visit to Statistical Engineering Division at NIST and was further supported by the Slovak Research and Development Agency under the contract No. LPP-0388-09, and by the grants VEGA 1/0077/09 and VEGA 2/0019/10.

References

[1] Rukhin AL. Weighted means statistics in interlaboratory studies. *Metrologia,* 46: 323-331, 2009.

[2] Rukhin AL, Sedransk N. Statistics in Metrology: interlaboratory key comparisons and interlaboratory studies. *Journal of Data Science*, 5: 393-412, 2007.

[3] Efron B, Tibshirani RJ. An Introduction to the Bootstrap. Chapman & Hall, Inc., New York, 1993.

[4] Vangel MG, Rukhin AL. Maximum likelihood analysis for heteroscedastic one-way random effects ANOVA in interlaboratory studies. *Biometrics,* 55: 129-136, 1999.

[5] Iyer HK, Wang CMJ, Mathew T. Models and confidence intervals for true values in interlaboratory trials. *JASA*, 99: 1060-1071, 2004.