



Improved non-negative estimation of variance components for exposure assessment

CHAVA PERETZ^a AND DAVID M. STEINBERG^b

^aThe Department of Epidemiology, Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel

^bThe Department of Statistics and Operations Research, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel

Hygiene surveys of pollutants exposure data can be analyzed by analysis of variance (ANOVA) model with a random worker effect. Typically, workers are classified into homogeneous exposure groups, so it is very common to obtain a zero or negative ANOVA estimate of the between-worker variance (σ_B^2). Negative estimates are not sensible and also pose problems for estimating the probability (θ) that in a job group, a randomly selected worker's mean exposure exceeds the occupational exposure standard. Therefore, it was suggested by Rappaport et al. to replace a non-positive estimate with an approximate one-sided 60% upper confidence bound. This article develops an alternative estimator, based on the upper tolerance interval suggested by Wang and Iyer. We compared the performance of the two methods using real data and simulations with respect to estimating both the between-worker variance and the probability of overexposure in balanced designs. We found that the method of Rappaport et al. has three main disadvantages: (i) the estimated σ_B^2 remains negative for some data sets; (ii) the estimator performs poorly in estimating σ_B^2 and θ with two repeated measures per worker and when true σ_B^2 is quite small, which are quite common situations when studying exposure; (iii) the estimator can be extremely sensitive to small changes in the data. Our alternative estimator offers a solution to these problems. *Journal of Exposure Analysis and Environmental Epidemiology* (2001) 11, 414–421.

Keywords: ANOVA estimator, bias adjustment, exposure assessment, hygiene surveys, repeated measures, variance component.

Introduction

Recently, analysis of variance (ANOVA) random effects models have been applied to data sets consisting of repeated measurements of pollutants within factories in order to identify determinants of exposure and estimate within- and between-worker variance components. The within-worker variance in these studies reflects day-to-day variations in the levels of exposure to pollutants, which often vary greatly. Between-worker variance, on the other hand, is often rather small due to the use of homogeneous exposure groups. Thus, the variance ratio $\lambda (= \sigma_B^2 / \sigma_W^2)$ may be quite small. As a result, when analyzing data using ANOVA random effects models, it is very common to obtain a zero or negative estimate of the between-worker variance. In many applications, it is common practice to report such negative values as zeros.

The occurrence of negative or zero between-worker ANOVA variance estimates causes a number of problems. First, zero between-worker variance appears to be an unrealistic result since it implies that all workers have the

same mean exposure. This contradicts common industrial hygiene experience. Furthermore, in exposure assessment in epidemiological studies and for hazard control, the probability θ of overexposure is often of more interest than the variance components themselves. This is the probability that in a job group, a randomly selected worker's mean exposure exceeds the occupational exposure standard, where the worker's mean exposure is relevant to the risk of chronic adverse health effects (Rappaport et al., 1995). The probability of overexposure depends on both σ_B^2 and σ_W^2 . Common practice is to adopt a "plug in" approach in which σ_B^2 and σ_W^2 are estimated and their estimates are inserted into the formula for θ . This approach is impossible to employ when the estimate of σ_B^2 is zero or negative. Finally, the variance ratio should have implications for planning future sampling design. Small variance ratios imply that it may be advantageous to sample fewer individuals but at more time points.

The estimation of the probability of overexposure (point estimator) becomes meaningless when a zero or negative between-worker variance estimate appears. Therefore, it was suggested by Rappaport et al. (1995) to replace a negative or zero estimate with an approximate one-sided 60% upper bound, as derived from formulas of Williams and cited in Searle et al. (1992). This practice is based on empirical evidence that such a procedure has minimal impact on significance levels and statistical power. This

1. Address all correspondence to: Chava Peretz, The Department of Epidemiology, Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel. Tel.: +972-3-640-9867. Fax: +972-3-641-0555. E-mail: cperetz@post.tau.ac.il

Received 24 July 2001.

proposal does have some drawbacks. Many negative ANOVA estimates are not adjusted to positive values and the estimator is very sensitive to small changes in the data.

This article develops an alternative — the bias-corrected variance component estimator — based on the upper tolerance interval suggested by Wang and Iyer (1994) to deal with the problem of negative variance component estimates. We compare the performance of the two methods using real data and simulations, focusing on the estimation of probabilities of overexposure (beyond standards) in balanced designs.

ANOVA method

We briefly review the ANOVA, or least squares (LS), method for estimating variance components in a balanced one-way random effects model. We denote: k =number of subjects in a group; n =number of repeated measurements obtained from each subject in the group:

$$MSW = SSW/(k(n-1)); MSB = SSB/(k-1);$$

$$F = MSB/MSW$$

The estimators of the between-subject (σ_B^2) and within-subject (σ_W^2) variance components are:

$$\hat{\sigma}_B^2 = [MSB - MSW]/n ; \hat{\sigma}_W^2 = MSW$$

For more details, see Searle et al. (1992).

An example from real data: lead exposure

Nineteen workers at two Car Battery Producers in Israel were repeatedly measured to study their annual exposure to lead. They were randomly selected — 9 workers in the first factory and 10 in the second — to represent those exposed to the main processes (details can be found elsewhere; Peretz et al., 1997). Ten hygiene surveys, with intervals of 3–7 weeks, were performed in each factory over the course of a year. Due to missing data (absence of workers, etc.), each worker had 6–10 repeated measures. We have taken the first six measures of each worker, and estimated the variance components σ_B^2 and σ_W^2 factory. According to Israel's regulations for factories with exposure to lead, it is mandatory to conduct two hygiene surveys each year.

In order to highlight the sensitivity of the σ_B^2 estimator, we have created new data sets, each including just two repeated measures out of the six. In total, we had 15 sets of data with two repetitions for each factory. The exposure level was taken as a log transformation of the TLV¹ fraction (=log(concentration/TLV)) (Peretz et al., 1997). The

TLV–TWA² standard for occupational lead exposure according to Israel's Regulations is 0.1 mg/m³.

Table 1A shows summary measures of the estimators in each factory, in comparison to the original estimators (=“accurate”) based on six repetitions. It can be seen that a negative σ_B^2 estimate resulted from 40% of the series in the first factory (with true $\lambda=.17$) and from 20% of the series in the second factory (with true $\lambda=.09$). In addition, the ANOVA estimators for λ were quite poor. This reinforces the importance of performing more than two repeated surveys per year. In practice, though, many surveys are limited to two measurements as mentioned above for lead exposure. So the example also highlights the need for statistical methods that can cope small samples. Table 1B shows summary measures of the estimators if four repeated surveys were performed in each factory. One can see the improvement in the estimation when doubling the number of repeated measurements per subject. The MSE [= (mean $\hat{\sigma}_B^2 - \sigma_B^2$)² + var $\hat{\sigma}_B^2$] is reduced by about 75% in the two factories.

Estimating θ in the presence of a negative ANOVA estimate of σ_B^2

Overexposure

For hazard control, the probability θ of overexposure is very important. We present here the basic equations for overexposure as derived by Rappaport et al. (1995). They followed the common assumption that the exposure x_{ij} of worker i on day j follows a log normal distribution with:

$$y_{ij} = \ln(x_{ij}) = \mu_y + \alpha_i + \varepsilon_{ij}$$

where μ_y is the mean of the overall logged exposure distribution in the group, α_i is a random effect for the i th worker and ε_{ij} is the within-worker random error.

It was furthermore assumed that: $\alpha_i \sim N(0, \sigma_B^2)$, where α_i 's are all independent; $\varepsilon_{ij} \sim N(0, \sigma_W^2)$, ε_{ij} 's are all independent. $\sigma^2 = \sigma_W^2 + \sigma_B^2$; σ_B^2 =variance between workers; σ_W^2 =variance within workers.

This model is applied to homogeneous work groups consisting of workers who perform similar tasks and therefore should have similar exposures. A worker is considered overexposed if his mean value μ_{xi} (conditional on α_i) exceeds a standard limit (S). The probability θ that a randomly selected person from a work group is overexposed is thus:

$$\theta = p\{\mu_{xi} > S\} = p\left\{Z > \frac{\ln(S) - \mu_y - 0.5\sigma_w^2}{\sigma_B} = Z_{1-\theta}\right\} \quad (1)$$

¹TLV=threshold limit value; a health-based concentration to which nearly all workers may be exposed without adverse effect.

²TLV–TWA=threshold limit value, with respect to 8-h time-weighted average, that should not be exceeded during any part of the working day.

Table 1. Summary measures of ANOVA (LS) estimators, mean, SD (min,max) on semi-simulated data sets.

	Series	Number of series	$\hat{\sigma}_B^2$	$\hat{\lambda}(=\hat{\sigma}_B^2/\hat{\sigma}_W^2)$
<i>(A) n^a=2</i>				
Factory 1	$k^b=9$	Accurate	.09	.17 (=09/.53)
	Total	15	.10, .23 (-.23, .46)	.35, .57 (-.33, 1.35)
	Positive	9	.26, .14 (.00, .46)	.72, .42 (.01, 1.35)
	Negative	6	-.14, .08 (-.23, -.04)	-.20, .11 (-.33, -.05)
Factory 2	$k^b=10$	Accurate	.11	.09 (=11/1.29)
	Total	15	.13, .37 (-.59, .65)	.24, .36 (-.27, .88)
	Positive	12	.29, .20 (.03, .65)	.36, .30 (.03, .88)
	Negative	3	-.48, .10 (-.59, -.42)	-.23, .04 (-.27, -.20)
<i>(B) n^a=4</i>				
Factory 1	$k^b=9$	Accurate	.09	.17 (=09/.53)
	Total	15	.09, .08 (-.03, .24)	.19, .18 (-.06, .58)
	Positive	12	.12, .07 (.03, .24)	.25, .16 (.06, .58)
	Negative	3	-.02, .01(-.03, -.01)	-.04, .02 (-.06, -.02)
Factory 2	$k^b=10$	Accurate	.11	.09 (=11/1.29)
	Total	15	.12, .12 (-.09, .29)	.11, .12 (-.05, .38)
	Positive	12	.16, .09 (.00, .29)	.14, .10 (.00, .38)
	Negative	3	-.05, .03 (-.09, -.02)	-.03, .02 (-.05, -.01)

^an=Number of repetitions.
^bk=Number of workers.

The relationship between σ_B^2 and θ for different values of $C=\mu_x/S$ (for $\sigma_W^2=.5$) is presented in Figure 1. It can be seen that when $.5 \leq c \leq 1.0$, θ has a maximum and then decreases, with little sensitivity to σ_B^2 . Therefore, as θ is calculated based on an estimate of σ_B^2 , the estimate of θ is quite stable for σ_B^2 values, which are slightly larger than zero. There is a problem in the estimation of θ when σ_B^2 is near zero because θ is sensitive to σ_B^2 in that region and because the ANOVA estimate of σ_B^2 may be negative.

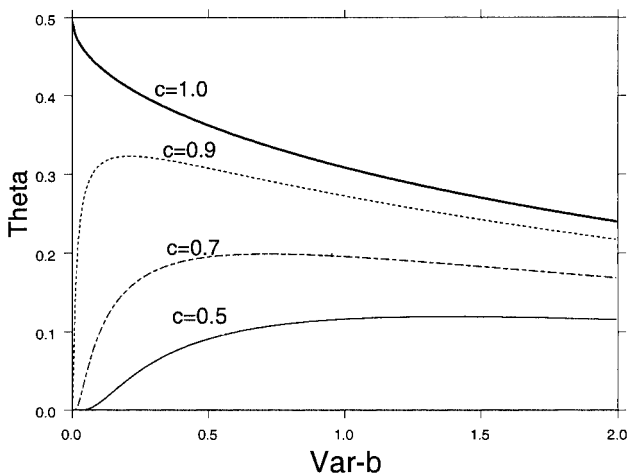


Figure 1. Relationship of the between-workers variance (Var-b) and the probability of overexposure (θ) for different values of $c(=\mu_x/\text{Standard})$ for the group of workers ($\sigma_W^2=.50$).

Since nowadays there is an emphasis on making the exposure groups as homogeneous as possible, we may be faced with applications that have small values of $\lambda(=\sigma_B^2/\sigma_W^2)$.

Rappaport et al. Method

Rappaport et al. (1995) recognized the problems of negative between-worker variance component estimates for estimating overexposure probabilities and for testing for compliance to standards. They proposed the following alternative estimator.

Use the ANOVA estimate if $\hat{\sigma}_B^2$ is positive. Otherwise, substitute $\hat{\sigma}_{B,1-\alpha}^2$ for $\hat{\sigma}_B^2$ where $\hat{\sigma}_{B,1-\alpha}^2$ is an approximate 100(1- α)% upper confidence bound for σ_B^2 , namely $P(\sigma_B^2 < \hat{\sigma}_{B,1-\alpha}^2) \approx 1-\alpha$. The upper bound derived by Williams (1962), and cited in Searle et al. (1992), is:

$$\hat{\sigma}_{B,1-\alpha}^2 = \frac{(k-1)(MSB - F_L MSW)}{n\chi_{k-1,L}^2}$$

where

$$P\{F_{k(n-1)}^{(k-1)} \leq F_L\} = 1 - \alpha/2$$

$$P\{\chi_{k-1}^2 \leq \chi_{k-1,L}^2\} = 1 - \alpha/2,$$

where $F_{k(n-1)}^{(k-1)}$ and χ_{k-1}^2 represent random variables distributed as F with $(k-1)$ numerator and $k(n-1)$ denominator degrees of freedom and χ^2 with $(k-1)$

degrees of freedom, respectively. Rappaport et al. (1995) suggested using a 60% confidence bound. In a subsequent article, Lyles et al. (1997) used the same basic approach but with a 95%, rather than a 60%, approximate upper confidence bound for a negative $\hat{\sigma}_B^2$ ANOVA estimate. Although the latter article dealt only with hypothesis testing, the 95% upper bound could also be used in estimating θ .

Some Drawbacks to the Rappaport et al. Estimator

We note here two problems with the between-worker variance component estimator proposed by Rappaport et al. First, the adjustment made to negative ANOVA estimates is often insufficient to produce a positive estimate. We illustrate this feature later in a simulation study.

Second, the fact that Rappaport et al.'s estimator only corrects negative ANOVA estimates makes it very sensitive to small changes in the data. According to Rappaport et al., when $\lambda(\sigma_B^2/\sigma_W^2) \approx 0.1-0.2$, negative estimated values of σ_B^2 could be observed as much as 30-40% of the time when $k=10$ and $2 \leq n \leq 4$. This probability can be reduced by increasing the sample size; however, in reality, many occupational hygiene groups are of this order of magnitude, having two to four repeated measurements (Kromhout et al., 1993).

We illustrate the sensitivity of Rappaport et al.'s estimator with a simple example using simulated data with $k=10$, $n=2$, $\sigma_B^2=.1$ and $\sigma_W^2=1$. First, a random set was generated and, gradually, eight slight changes were made to create eight further sets, each with the same worker averages but with increasingly larger within worker residuals.

Table 2 presents the variance component estimates according to ANOVA and Rappaport et al.'s method with the 60% confidence bound.

From step 7 on, the ANOVA estimate, $\hat{\sigma}_B^2$ ANOVA, was negative. The Rappaport et al. estimate for σ_B^2 makes a sudden jump from very small to very large values at step 7

and thus is quite sensitive to small changes in the study data. A slight increase in the within-workers mean square could change a positive ANOVA estimate to a negative one, thus sharply increasing Rappaport et al.'s estimate. This change could lead to a much larger estimate of θ . Since the error term in exposure measurements is already known to vary greatly over time (contributing to the within worker variability), measuring the same exposure group at different times can easily produce negative $\hat{\sigma}_B^2$ ANOVA estimates.

Bias-Adjusted Variance Component Estimation (BAVCE)

We suggest an alternative estimate to overcome some of the limitations of the estimator proposed by Rappaport et al. Our method, which we call BAVCE, is based on the upper tolerance interval suggested by Wang and Iyer (1994).

It takes account of the fact that an upper confidence bound will typically be biased high and multiplies by a factor that attempts to adjust for this bias. The estimator is defined as follows:

$$\hat{\sigma}_B^2 = \omega^2(MSB - F_LMSW)/n$$

where

$$P\{F_{k(n-1)}^{(k-1)} \leq F_L\} = \eta \text{ for } \eta = (1 - \gamma)/n,$$

where γ is the confidence level (which we have taken to be .95), $\omega^2 = \tilde{\phi}/[1 - (1 - \tilde{\phi})F_L]$ and

$$\tilde{\phi} = \max(0, 1 - F_LMSW/MSB).$$

The BAVCE estimator, like the others (Rappaport et al., 1995; Lyles et al., 1997), reduces the frequency of negative or zero estimates by subtracting less than the full value of MSW from MSB. However, their use of an upper confidence bound as an estimator almost guarantees an overestimate of σ_B^2 . The factor ω^2 in the BAVCE attempts to correct the upward bias. To see how the bias correction

Table 2. Sensitivity of σ_B^2 , σ_W^2 estimators to small changes in values of a set of data.

Set	Percent inflation of residuals	$\hat{\sigma}_W^2$ ANOVA	$\hat{\sigma}_B^2$ ANOVA	$\hat{\sigma}_{B-1}^2$ ^a
1	random set	.72	.23	.23
2	5	.80	.20	.20
3	10	.87	.16	.16
4	15	.96	.12	.12
5	20	1.04	.07	.07
6	25	1.13	.03	.03
7	30	1.22	-.02	3.47
8	35	1.32	-.06	3.45
9	40	1.42	-.11	3.43

$n=2$; $k=10$, original: $\sigma_B^2=.1$, $\sigma_W^2=1$.

$\hat{\sigma}_{B-1}^2$ according to Rappaport et al.'s method, based on upper bound of 60%.

works, we present an approximation to the expected value of $\hat{\sigma}_B^2$ by assuming that $\phi = 1 - [\sigma_W^2 / (n\sigma_B^2 + \sigma_W^2)]$ is known. Then:

$$\omega^2 = \frac{n\sigma_B^2 / (n\sigma_B^2 + \sigma_W^2)}{1 - F_L \sigma_W^2 / (n\sigma_B^2 + \sigma_W^2)} = \frac{n\sigma_B^2}{n\sigma_B^2 + (1 - F_L)\sigma_W^2}$$

and

$$\begin{aligned} E\{\hat{\sigma}_B^2\} &= \omega^2 E\{MSB - F_L MSW\} / n \\ &= \omega^2 [n\sigma_B^2 + (1 - F_L)\sigma_W^2] / n \\ &= \frac{n\sigma_B^2}{n\sigma_B^2 + (1 - F_L)\sigma_W^2} [n\sigma_B^2 + (1 - F_L)\sigma_W^2] / n \\ &= \sigma_B^2 \end{aligned}$$

The bias correction is implemented by using a “plug-in” estimator of ϕ in which the observed mean squares replace their expected values.

Comparison of estimators on simulated data

Simulated Data

Simulations were run to compare the different estimators of σ_B^2 and θ . The estimators of σ_B^2 were the ANOVA estimator, the estimator of Rappaport et al. with a 60% bound (method 1) and with a 95% bound (method 1A) and the BAVCE proposed here (method 2). The estimators of θ were generated by plugging the estimators

Table 3. Comparison of estimation methods based on simulations of 1000 sets for 10 workers with 2,3,4 repetitions.

A. Results for negative LS estimators of σ_B^2		$\hat{\sigma}_B^2$ (original=.20)				$\hat{\theta}$ (original=.34)	
Repetitions (number of series)		LS method	Method 1	Method 1A	Method 2	Method 1	Method 2
2^a							
(473)	mean, SD	-.23, .19	.06, .23	.51, .37	.16, .11	.32, .03	.31, .06
	min, max	-1.13, .00	.24, 6.14	-.55, 1.83	.00, .58	.00, .34	.00, .34
3^b							
(478)	mean, SD	-.13, .10	.05, .13	.32, .23	.11, .07	.31, .05	.32, .05
	min, max	-.61, .00	-.47, .43	-.33, 1.21	.00, .41	.04, .34	.00, .34
4^c							
(413)	mean, SD	-.09, .07	.04, .10	.24, .16	.09, .05	.30, .06	.31, .06
	min, max	-.41, .00	-.30, .26	-.19, .67	.00, .24	.00, .34	.00, .34

LS method=least squares method; method 1=Rappaport et al.'s methods based on upper bound of 60%; method 1A=Rappaport et al.'s methods based on upper bound of 95%; method 2=our method based on a modified upper bound.

B. Results for positive LS estimators of σ_B^2		$\hat{\sigma}_B^2$ (original=.20)		$\hat{\theta}$ (original=.34)	
Repetitions (number of series)		Method 1 (LS method)	Method 2	Method 1	Method 2
2					
(527)	mean, SD	.27, .21	.49, .21	.31, .05	.32, .02
	min, max	.00, 1.14	.11, 1.30	.00, .34	.25, .34
3					
(522)	mean, SD	.17, .14	.34, .14	.31, .06	.33, .01
	min, max	.00, .76	.10, .93	.00, .34	.28, .34
4					
(587)	mean, SD	.14, .11	.28, .11	.30, .06	.33, .01
	min, max	.00, .53	.08, .63	.00, .34	.31, .34

LS method=least squares method; method 1=Rappaport et al.'s methods; method 2=our method based on a modified upper bound.

^a298/473 positive according to method 1; 437/473 positive according to method 1; 463/473 positive according to method 2.

^b318/478 positive according to method 1; 445/478 positive according to method 1; 462/478 positive according to method 2.

^c293/413 positive according to method 1; 386/413 positive according to method 1; 408/413 positive according to method 2.

Table 4. Comparison of estimation methods based on simulations of 1000 sets for 10 workers with 2,3,4 repetitions.

Repetitions (number of series)		$\hat{\sigma}_B^2$ (original=.05)				$\hat{\theta}$ (original=.31)	
		LS method	Method 1	Method 1A	Method 2	Method 1	Method 2
2^a							
(505)	mean, SD	-.25, .19	.04, .24	.47, .37	.15, .11	.32, .05	.31, .06
	min, max	-1.11, .00	-.87, .72	-.52, 1.82	.00, .59	.00, .34	.00, .34
3^b							
(538)	mean, SD	-.14, .10	.03, .14	.29, .22	.10, .07	.31, .06	.31, .06
	min, max	-.62, .00	-.48, .40	-.33, 1.15	.00, .39	.00, .34	.00, .34
4^c							
(510)	mean, SD	-.09, .07	.03, .10	.22, .15	.08, .05	.30, .07	.31, .06
	min, max	-.39, .00	-.28, .24	-.18, .63	.00, .22	.00, .34	.00, .34

LS method=least squares method; method 1=Rappaport et al.'s method based on upper bound of 60%; method 1A= Rappaport et al.'s method based on upper bound of 95%; method 2=our method based on a modified upper bound.

B. Results for positive LS estimators of σ_B^2

Repetitions (number of series)		$\hat{\sigma}_B^2$ (original=.05)		$\hat{\theta}$ (original=.31)	
		Method 1 (LS method)	Method 2	Method 1	Method 2
2					
(495)	mean, SD	.24, .20	.46, .20	.31, .05	.32, .02
	min, max	.00, 1.05	.10, 1.16	.00, .34	.27, .34
3					
(462)	mean, SD	.15, .12	.32, .12	.31, .05	.33, .01
	min, max	.00, .83	.11, .93	.00, .34	.28, .34
4					
(490)	mean, SD	.12, .10	.25, .10	.30, .07	.33, .01
	min, max	.00, .50	.08, .60	.00, .34	.31, .34

LS method=least squares method; method 1=Rappaport et al.'s method; method 2=our method based on a modified upper bound.

^a303/505 positive according to method 1; 462/505 positive according to method 1; 485/505 positive according to method 2.

^b336/538 positive according to method 1; 485/538 positive according to method 1; 521/538 positive according to method 2.

^c326/510 positive according to method 1; 469/510 positive according to method 1; 501/510 positive according to method 2.

of σ_B^2 along with the ANOVA estimator of σ_W^2 and the sample average into Eq. (1) in Section 4. The simulations covered three different practical settings defined by the number of repetitions (n) and the number of subjects (k):

- (i) 1000 data sets for $k=10, n=2$ (20000 observations);
- (ii) 1000 data sets for $k=10, n=3$ (30000 observations); and
- (iii) 1000 data sets for $k=10, n=4$ (40000 observations).

In addition, we examined several different values of σ_B^2 . The within-subject variance σ_W^2 was held constant at 1 in all the simulations. Original values for ϕ for $n=2,3,4$ were computed from Eq. (1). When the least squares estimate of the between-workers variance component σ_B^2 was negative,

method 1 modified it to a larger value. The method 2 estimator increased all the σ_B^2 estimates, not just the negative ones.

Comparison of Estimators

Tables 3 and 4 present the estimators based on the simulated data for $n=2,3,4$, when original $\sigma_B^2=.2$ (Table 3) or $\sigma_B^2=.05$ (Table 4), which are representative of the results that we found for all the values of σ_B^2 .

Tables 3A and 4A relate to the estimators when negative ANOVA estimates were found. Tables 3B and 4B relate to the estimators when positive ANOVA estimates were found. As was found previously, the ANOVA estimator of σ_B^2 was often negative for the cases we studied. In Table 3A ($\sigma_B^2=.2, \lambda=.2$), we can see that more than 40% of the data sets for $n=2,3,4$ resulted in a negative σ_B^2 ANOVA estimate and in

Table 5. Parameter estimation according to the different methods (mean±SD).

	<i>k</i>	<i>n</i>	$\hat{\sigma}_W^2$	$\hat{\sigma}_B^2$	$\hat{\theta}$				
			LS method	LS method	Method 1	Method 1A	Method 2	Method 1	Method 2
All farmers	153	2	.64	.13 ($\lambda=0.20$)				.32 (LS method)	
Subsamples negative $\hat{\sigma}_B^2$ LS (21 series)	8–20	2	.73±.14	-.09±.06	.09 ^b ±.11	.34±.21	.11±.07	.29 ^b ±.21	.27±.18
Subsamples positive $\hat{\sigma}_B^2$ LS (79 series)	5–25	2	.59±.15	.19±.15	.19±.15	.19±.15	.34±.15	.32±.17	.32±.09

LS method=least squares method; method 1=Rappaport et al.’s method based on upper bound of 60%; method 1A=Rappaport et al.’s method based on upper bound of 95%; method 2=our method based on a modified upper bound.

^aRandomuni function returns a number from the uniform distribution on the interval (0,1). The function was applied 100 times on the original data set, each time with another seed. In each time, observations with a random number less than .1 were included in the new subsample.

^bFour negative values for $\hat{\sigma}_B^2$ according to method 1a; consequently, four missing values for $\hat{\theta}$ according to method 1.

Table 4A ($\sigma_B^2=.05, \lambda=.05$), we can see that the percentage was higher, over 50%.

A serious problem with Rappaport et al.’s method is that many negative estimates of σ_B^2 remained negative. The problem was especially acute with the 60% confidence bound. Even with $\sigma_B^2=.20$ and four replications per subject, almost 30% of the negative ANOVA estimates remained negative with this method. Using their method with a 95% confidence bound reduced the problem but did not eliminate it, with 7–10% of the negative ANOVA estimates remaining negative. Our method was much more successful in this regard. Negative estimates are automatically adjusted to 0 and these occurred in less than 4% of the cases with negative ANOVA estimates in all the settings we examined.

In conclusion, there is an estimation problem using method 1 when $n=2$ or 3 and σ_B^2/σ_W^2 is less than .20.

An example

Survey on Pig Farmers’ Exposure to Inhalable Endotoxins

In a study of 200 pig farmers from the south of the Netherlands, exposure to inhalable dust and endotoxins was monitored by personal sampling. Exposure was measured during one work shift on a randomly chosen day of the week, 1 day during the summer of 1991 and 1 day during the winter of 1992. Outdoor temperature was obtained from a monitoring station in the south of the Netherlands. Task activity patterns on the day of measurement and farm characteristics were also recorded (Preller et al., 1995). For the purpose of this paper, only the exposure data on endotoxins will be used on 153 farmers out of the 200 who had two measurements (the rest had some failure in the measuring process for one measurement). For the whole study population ($n=153$), the following estimates were calculated and they were considered to be the accurate parameters for the pig farmers: $\sigma_B^2=.13, \sigma_W^2=.64, \theta=.32, \mu_y=7.81$. We have taken the standard to be 8.29 (Standard= $\log(4000 \text{ ng/m}^3)=8.29$) for this example.

We compared the different estimators of σ_B^2 by generating 100 subsamples. Each farmer was included/excluded from a particular sample by drawing a binomial random variable with probability 0.1 for inclusion.

For the 100 subsamples, mean±SD of the μ_y values= 7.81 ± 0.15 . The same parameters were estimated while σ_B^2 was estimated by the different methods (see Table 5).

In this example, only about 20% of the series resulted in negative ANOVA estimates. Thus, one might expect that our method, which always corrects $\hat{\sigma}_B^2$, might be less successful. Nonetheless, for θ , our estimate performed better than that of Rappaport et al., with a smaller SD especially for the samples with a positive ANOVA estimate. For $\hat{\sigma}_B^2$, Rappaport et al.’s estimate over the 100 samples seemed to perform better than our method.

This conclusion differs from our previous conclusion regarding the simulated data due to the different sample sizes. Here, on average, 15 subjects were included in each sample while in our previous samples, we had only 10 subjects per each sample.

Rappaport et al.’s estimator of σ_B^2 was more accurate with a 60% bound than with a 95% bound.

Discussion

The use of Rappaport et al.’s approach for assessing compliance for hazard control is a new application. It has inherent statistical considerations and takes into account the variance components of the hazardous exposure based on real-life data sets and should be recommended for use. However, since it is a new tool, caution and further study are needed. In exposure data sets, ANOVA estimators for between-variance components are quite often negative (see sensitivity analysis).

The common practice of changing such negative values to zeroes prevents the application of popular plug-in estimators in compliance assessment and it also appears to be an unrealistic result since it implies that all workers have the same mean exposure.

The modified variance component estimator for negative values proposed by Rappaport et al. (1995) and Lyles et al. (1997) has three main disadvantages:

1. It remains negative for some data sets.
2. It performs poorly in estimating σ_B^2 and also θ when $n=2$ and when original σ_B^2 is quite small and $0.5 \times \text{Standard} \leq \mu_x \leq 1.0 \times \text{Standard}$, which are quite common situations when studying exposure.
3. Discontinuous behavior: small changes in the data set can make the ANOVA estimator negative, resulting in the use of the modification, which may cause a large change in the conclusions of a study.

In this paper, we have proposed an alternative variance component estimator, the BAVCE, to cope with the problem of negative and zero between-worker ANOVA estimates. Our modification seems to react better than the estimator of Rappaport et al. as can be seen in the tables from our simulations and the simulated subsets of data.

We think that further thought should be given to analysis of data from unbalanced designs, which are common in real-life exposure data sets due to absence of workers and changes in work practices.

Here, exposure was measured in industry and agriculture. The same ideas can be applied to environmental exposure within the community.

Acknowledgment

We acknowledge Liesbeth Preller from Wageningen University, the Netherlands, for providing the pig farmers exposure data set.

References

- Kromhout H., Symanski E., and Rappaport S.M. A comprehensive evaluation of within- and between- worker components of occupational exposure to chemical agents. *Ann Occup Hyg* 1993; 37: 253–270.
- Lyles R.H., Kupper L.L., and Rappaport S.M. Assessing regulatory compliance via the balanced one-way random effects ANOVA model. *J Agric, Biol Environ Stat* 1997; 2: 64–86.
- Peretz C., Goldberg P., Kahan E., Grady S., and Goren A. The variability of exposure over time: a prospective longitudinal study. *Ann Occup Hyg* 1997; 41: 485–500.
- Preller L., Kromhout H., Heederik D., and Tielen M. Modeling long-term average exposure in occupational exposure-response analysis. *Scand J Work, Environ Health* 1995; 21: 504–512.
- Rappaport S.M., Lyles R.H., and Kupper L.L. An exposure assessment strategy accounting for within- and between-worker sources of variability. *Ann Occup Hyg* 1995; 39: 469–495.
- Searle S.R., Casella G., and McCulloch C.E. *Variance Components*. Wiley, New York, 1992.
- Wang C.M., and Iyer H.K. Tolerance intervals for the distribution of true values in the presence of measurement errors. *Technometrics* 1994; 36: 162–170.
- Williams J.S. A confidence interval for variance components. *Biometrika* 1962; 49: 278–281.