Sampling Strategies for Occupational Exposure Assessment under Generalized Linear Model

CHU-CHIH CHEN1*, CHENG-LIN CHUANG1, KUEN-YUH WU2 and CHANG-CHUAN CHAN2

1Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Zhunan Town, Miaoli County 350, Taiwan; 2Institute of Occupational Medicine and Industrial Hygiene, National Taiwan University, College of Public Health, Taipei 100, Taiwan

Received 12 August 2008; in final form 23 April 2009; published online 20 May 2009

Objectives: Occupational exposure assessment is a major task in industrial hygiene studies. Although statistical analyses for magnitudes and variations of exposures to various types of working populations based on existing data sets are extensive, relatively few discussions on study designs appear in the literature, especially for sample size determination and number of repeated measurements.

Methods: In this paper, we propose a general framework of sampling strategies on sample size requirement together with the number of repeated measurements using the mixed-effects generalized linear model (GLM). As illustrative examples, we discuss sampling strategies separately under the log-normal assumption for hypotheses testing on (i) mean exposure differences of multiple worker groups and (ii) presence of a long-term exposure trend.

Results: Given a specified alternative hypothesis, the desired significance level and statistical power, the number of repeated measurements, within-worker and between-worker variances, and a correlation structure, we have derived and tabulated an explicit sample size requirement of the two hypothetical cases under log-normal distribution assumption.

Conclusions: On the basis of the tabulated outcomes, the sample size requirement is much more dominant than the number of repeated measurements for a group exposure comparison. Thus, in this case, recruiting more workers with fewer repeated measurements may be more economical than the opposite approach. For testing the presence of a long-term exposure trend, the sample size required decreases substantially with the number of repeated measurements. Also, equally spaced sampling times would be optimal because the effect of between-worker variance is algebraically cancelled out in sample size calculations.

Keywords: exponential family distribution; gamma distribution; generalized estimating equation; log-normal distribution; long-term trend; mixed model

INTRODUCTION

Workplace exposures are known to inherently have lots of variation among individual workers or groups of workers who may perform similar or different job tasks (between-worker variance) and to have within-worker or day-to-day variance. To clarify these variance components, a carefully designated statistical model that best describes the nature of the distribution of the exposures is crucial for exposure assessment. For example, a one-way random-effects analysis of variance (ANOVA) model is feasible for a homogeneous working group (Kromhout et al., 1993; Rappaport et al., 1993, 1995; Lyles et al., 1997a,b; Tornero-Velez et al., 1997) and a two-way mixed model or a nested mixed model is required to distinguish fixed effects from random effects (Rappaport et al., 1995, 1999; Symanski et al., 1996; 2001a,b; van Tongeren et al., 2000; Weaver et al., 2001). In addition to common ANOVA mixed models, a temporal exposure trend may also need to be adjusted in the model for a long-term occupational exposure study (Symanski et al., 1998a,b; Burstyn et al., 2000; Kromhout and Vermeulen, 2000; van Tongeren et al., 2000; Vermeulen et al., 2000). Furthermore, ignoring the trend may yield substantial bias in estimating the variance component (Symanski et al., 1996).

*Author to whom correspondence should be addressed.
Tel: +886-37-246-166 (ext. 36186); fax: +886-37-586-467;
e-mail: ccchen@nhri.org.tw
Depending on the study’s objective and specified statistical model of an occupational exposure assessment, a carefully planned sampling strategy is desirable so that measurements can be collected efficiently and resources spent wisely. However, there appear to be no guidelines on sample size and repeated measurements determination except some discussions on design of efficient measurement and optimal measurement strategies (Kromhout, 2002; Burdof and van Tongeren, 2003; Lampa et al., 2006). Furthermore, it has been assumed almost exclusively that the exposure measurements in industrial hygiene literature have a log-normal distribution, whereas a more general class of the generalized linear model (GLM) has been well recognized and is commonly used in epidemiological and medical studies. Statistical analysis based on the GLM is available with the PROC GENMOD of the SAS commercial software (SAS Institute, Cary, NC, USA) and is applicable for a broad class of exponential family of distributions, including normal, log-normal (normal after log-transformation), and gamma. Owing to the nature of positive exposure measurements, most often they are assumed to have a log-normal distribution, whereas they may actually belong to a gamma distribution, which is similar to log-normal distribution in shape but is more flexible in the degree of skewness and more efficient in some cases (Firth, 1988; McCullagh and Nelder, 1989).

In this paper, we propose a general framework of sampling design strategies for occupational exposure assessment based on the GLM approach. As illustrative examples, we present specific strategies that assume log-normal distribution for hypotheses testing: (i) K-sample group exposure differences and (ii) presence of a long-term exposure trend. Given a specified alternative hypothesis, a desired significance level and statistical power, the number of repeated measurements, within-worker and between-worker variances, and a correlation structure, we derived and tabulated an explicit sample size requirement following Liu and Liang (1997). Three correlation structures were considered: (i) independent (IND), (ii) compound symmetric (CS), and (iii) first-order autoregressive [AR(1)]. These structures represent that the correlation between measurements on the same worker is independent, the same irrespective of time interval, and exponentially decreases with time interval, respectively.

**METHODS**

In this section, we present the major statistics for the general framework of the sampling strategy using the GLM approach. Let $y_{ij}$ be the $j$th observation of the $i$th subject and the corresponding vector of observations $\mathbf{y}_i = (y_{i1}, \ldots, y_{im})', i = 1, \ldots, n; j = 1, \ldots, m$. A general form of the mixed-effects GLM may be written as

$$g(u_{ij}) = \mathbf{b}_i + \mathbf{x}_{ij}' \beta + \mathbf{z}_{ij}' \gamma,$$

where $g(\cdot)$ is the link function relating the conditional mean $E(Y_{ij} | \mathbf{b}_i) = u_{ij}$ to the vector of covariates $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$, $\mathbf{b}_i$ represents a $q \times 1$ vector of random effects, $\beta$ is a $K \times 1$ vector of the parameters of interest, and $\gamma$ is a $r \times 1$ vector of nuisance parameters. And the conditional variance of $y_{ij}$ is $\text{var}(Y_{ij} | \mathbf{b}_i) = h(u_{ij}) \phi$ for some function $h(\cdot)$ with $\phi$ a scale dispersion parameter (Zeger et al., 1988).

Under the GLM framework, for log-normal distribution, one may take $y_{ij} = \log w_{ij}$ with identity link $g(u) = u$, where $w_{ij}$ is the $j$th exposure measurement of the $i$th subject. The corresponding conditional variance is $\text{var}(Y_{ij} | \mathbf{b}_i) = \sigma^2_w$. For gamma distribution with parameters $\eta$ and $\lambda$, one may take log-link $g(u) = \log u$ to the untransformed measurement ($y_{ij} = w_{ij}$). The conditional mean and variance are thus $E(Y_{ij} | \mathbf{b}_i) = \eta \lambda = u$ and $\text{var}(Y_{ij} | \mathbf{b}_i) = \eta \lambda^2 = u^2 / \eta \rho_i$, respectively.

By taking different values of the covariates $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$, the GLM (1) may serve as the role model for various statistical hypothesis tests in exposure assessment. The following two examples illustrate some commonly encountered situations in exposure assessment.

**Example 1. K-sample group differences:** In the case of testing for mean exposure differences among multiple working groups, the model (1) may be expressed as a mixed-effects ANOVA model:

$$g(u_{ij}) = \mathbf{b}_i + v_k \times I\{i \in S_k\} + \mathbf{z}_{ij}' \gamma,$$

$$i = 1, \ldots, n, \ j = 1, \ldots, m, \ k = 1, \ldots, K,$$

where $I\{\cdot\} = 1$ if the statement in the parenthesis is true or is 0 otherwise, and $S_k$ is the set of the study subjects belonging to the $k$th working group. Let $\beta = (v_1, \ldots, v_K)'$ be the vector of mean exposure differences, $\sum_{k=1}^K v_k = 0$. One may write $\mathbf{x}_{ij} = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$ a $K \times 1$ vector with 1 on the $k$th item if the $j$th subject belongs to the $k$th group, $k = 1, \ldots, K$. The hypothesis testing that all exposure groups have the same mean is equivalent to testing for the null $H_0 : \mathbf{v} = \mathbf{0}$ versus the alternative $H_1 : \mathbf{v} \neq \mathbf{0}$.

**Example 2. Long-term exposure trend:** In the presence of a long-term exposure trend, assuming that the repeated measurement time is equally spaced, one may take $x_{ij} = j$ in the univariate case. The corresponding simple mixed-effects GLM is

$$g(u_{ij}) = \mathbf{b}_i + \beta j + \mathbf{z}_{ij}' \gamma,$$

where $\beta$ is the exposure trend with repeated measurement at time $j$. And the hypothesis testing is the null $H_0 : \beta = 0$ versus the alternative $H_1 : \beta \neq 0$.

Given that $\Psi = (\beta' \gamma)'$ and $\mu_i = E(Y_i) = (\mu_{i1}, \ldots, \mu_{im})'$, one can obtain the quasi-likelihood...
estimate of $\psi$ by solving the generalized estimating equation (GEE)

$$U(\psi) = \sum_{i=1}^{n} \left( \frac{\partial \mu_i}{\partial \psi} \right)' V_i^{-1} (Y_i - \mu_i) = 0,$$

where

$$V_i = \text{cov}(Y_i) = \text{cov}[E(Y_i|b_i)] + \text{cov}(Y_i|b_i)]$$

(Liang and Zeger, 1986; Zeger et al., 1988). The corresponding quasi-score test of the null hypothesis $H_0 : \theta = 0$ is

$$T = U(0, \hat{\gamma})' \Sigma_0^{-1} U(0, \hat{\gamma}),$$

where $\hat{\gamma}$ is the estimator of $\gamma$ obtained by solving

$$\sum_{i=1}^{n} \left( \frac{\partial \mu_i}{\partial \gamma} \right)' V_i^{-1} (Y_i - \mu_i) = 0,$$

under $H_0$ and

$$\Sigma_0 = \text{cov}_{H_0}(U(0, \hat{\gamma}))$$

(Liu and Liang, 1997). The test statistic $T$ has asymptotically a $\chi^2_{r-1}$ and a $\chi^2_1$ distribution under $H_0$ as $n \to \infty$ in cases of Examples 1 and 2, respectively, and has an asymptotic noncentral chi-square distribution with noncentral parameter $\zeta = \xi \Sigma_1^{-1} \xi$ under $H_1$, where $\xi = E_{H_1}[U(0, \hat{\gamma})]$ and $\Sigma_1 = \text{cov}_{H_1}[U(0, \hat{\gamma})]$.

 Depending on the desired study purpose for exposure assessment, the null hypothesis $H_0$ and the specified alternative hypothesis $H_1$ may be specified accordingly for statistical hypothesis testing. The corresponding sampling strategy is then to determine the required sample size $n$ and the repeated measurements $m$ for each individual to achieve the desired statistical significance $\alpha$ under $H_0$ and power $1 - \omega$ under $H_1$.

**SAMPLING STRATEGIES IN THE CASE OF LOG-NORMAL DISTRIBUTION**

In this section, we discuss the special case of the general methodological framework proposed above when exposures have a log-normal distribution, which is the commonly adopted distributional assumption in the occupational exposure literature. For illustrative purposes, we propose specific sampling strategies for two hypothesis tests: (i) differences in multiple exposure groups and (ii) presence of a long-term exposure trend, which are illustrated in Examples 1 and 2, respectively.

Under the log-normal distributional assumption, the mixed-effect GLM (1) becomes

$$Y_{ij} = b_i + x_{ij}'\beta + z_{ij}'\gamma + \epsilon_{ij},$$

where $y_{ij} = \log w_{ij}$ is a logarithm of the $j$th exposure measurement $w_{ij}$ of the $i$th subject, and $\epsilon_{ij}$ is independently of $b_i$ and is normally distributed with a mean of 0 and a variance of $\sigma^2$. Without loss of generality, in the following, we assume that the random effects $b_i$ is univariate (i.e. $q = 1$) and is normally distributed with a mean of 0 and a variance of $\sigma^2_b$. In practice, this single term of random effects $b_i$ may be regarded as an integrated term of all the sources of random effects. For example, let $b_i$ represent the summation of the random effects of workers ($w_i$), job groups ($g_i$), and plants ($p_i$) with variances $\sigma^2_w$, $\sigma^2_g$, and $\sigma^2_p$, respectively. Then $\sigma^2_b$ may be written as $\sigma^2_b = \sigma^2_w + \sigma^2_g + \sigma^2_p$. Depending on practical situations, they may also alternatively be expressed in the form of a random-effects vector $b_i$ by adding more subscripts to the observations $y_{ij}$ (see for example, Symanski et al., 2001a). Also, for simplicity, we assume here that the nuisance covariate is univariate with scale 1 (i.e. $z_{ij} = 1$). The corresponding parameter $\gamma$ is thus the overall mean exposure.

Because measurements on the same subject are dependent to a certain degree, their correlations need to be taken into account in sample size calculations. Three covariance structures are considered here: (i) IND, (ii) CS, and (iii) AR(1). Selection of a specific correlation structure may depend on time intervals and the nature of the relationship between repeated measurements (Symanski et al., 2001b). The covariance matrix $V_j$ of $Y_i = (Y_{i1}, \ldots, Y_{im})$ and its inverse under different correlation assumptions are summarized in the Appendix 1. Notice that the covariance structure IND is referred to as CS elsewhere in the literature (e.g. Symanski et al., 2001a). We regard here, however, the covariance structure CS as a symmetric correlation between measurements that is independent of the between-worker variance component $\sigma^2_{w}$.

Prior to an exposure assessment study, the length of the study period or the frequency of sample measurement is often predetermined. Thus, the number of repeated measurements $m$ and the time interval between two visits may be regarded as given information, and it is necessary only for the sampling strategy to determine the sample size $n$. We derive explicitly the sample size formulas for the multiple group comparison and the long-term exposure trend tests under a specified alternative hypothesis and correlation assumption.

**Multiple exposure group comparisons**

Often it is desirable to compare workers from various job categories with regard to their mean occupational exposure differences (Rappaport et al., 1995, 1999; Weaver et al., 2001) or heterogeneity in exposure variability (Symanski et al., 2001b; Weaver et al., 2001; van Tongeren et al., 2006). Using the above setting and Example 1 to compare $K$ exposure groups for their mean differences, we can rewrite model (2) as
\begin{equation}
Y_{ij} = b_i + \gamma + \nu_k \times I\{i \in S_k\} + \epsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m, \quad k = 1, \ldots, K.
\end{equation}

This model is essentially the same as the models that appear in the literature, but it is expressed in a different form (e.g. Rappaport et al., 1999; Weaver et al., 2001). The hypothesis testing of multiple group exposure differences is therefore \(H_0 : \mathbf{v} = \mathbf{0}\) versus \(H_1 : \mathbf{v} \neq \mathbf{0}\). If we let \(\pi_k = n_k / n\) be the proportion of the sample size \(n_k\) of the \(k\)th group among the comparison groups, \(k = 1, \ldots, K\), we have \(\sum_{k=1}^{K} \pi_k = 1\). The noncentrality parameter under the specified alternative \(H_1 : \mathbf{v} \neq \mathbf{0}\) is \(\xi = \mathbf{\Sigma}_{1}^{-1}\mathbf{v}\). Following the approximations of the items in the noncentrality parameter (given in the Appendix 1), one can show that

\begin{equation}
\mathbf{\Sigma}_1 = \text{cov}_{H_1}[\mathbf{U}(0, \mathbf{v})] = n \sum_{k=1}^{K} (\mathbf{u}_k - \mathbf{1}\mathbf{R}_{\mathbf{v}}^{-1}) \mathbf{v}_k^{-1} (\mathbf{u}_k - \mathbf{1}\mathbf{R}_{\mathbf{v}}^{-1})^\top,
\end{equation}

where \(\mathbf{V}_k\) is the covariance matrix of \(\mathbf{Y}_i\) when the \(i\)th subject belongs to the \(k\)th group, \(\mathbf{u}_k\) is an \(m \times K\) matrix with \(1\) for the \(k\)th column and \(0\) for the rest of the columns, \(\mathbf{R}_{\mathbf{v}} = \sum_{k=1}^{K} \pi_k \mathbf{u}_k \mathbf{V}_k^{-1} \mathbf{u}_k^\top\), \(\mathbf{I}_{\mathbf{v}} = \sum_{i=1}^{K} \pi_i \mathbf{V}_i^{-1}\), \(1 = (1, \ldots, 1)\), and \(\xi = \mathbf{\Sigma}_{1}\mathbf{v}\). Thus, the noncentrality parameter under \(H_1\) may be rewritten as

\begin{equation}
\xi = \mathbf{v}^\top \mathbf{\Sigma}_{1}\mathbf{v}.
\end{equation}

The covariance matrix \(\mathbf{V}_k\) for the \(k\)th working group may be different if the group exposure variances are assumed to be heterogeneous (Symanski et al., 2001b; Weaver et al., 2001; van Tongeren et al., 2006). For hypothesis testing of mean exposure difference \(\mathbf{v} \neq \mathbf{0}\) together with heterogeneity of group exposure variance that not all \(\mathbf{V}(1), \ldots, \mathbf{V}(K)\) are equal, one may simply plug in the assumed constants \(\mathbf{v}, \mathbf{V}(1), \ldots, \mathbf{V}(K)\), in calculating the noncentrality parameter. However, because the noncentrality parameter would be zero in the case of simple heterogeneity of group exposure variance with the same group mean (\(\mathbf{v} = \mathbf{0}\)), the proposed sampling strategy framework does not apply in this case. Likelihood ratio test with \(H_0 : \mathbf{V}(1) = \cdots = \mathbf{V}(K) = \mathbf{V}\) versus the alternative hypothesis of heterogeneity \(\mathbf{V}(1), \ldots, \mathbf{V}(K)\), similar to the methodologies of Symanski et al. (2001b), Weaver et al. (2001), and van Tongeren et al. (2006), may be applied for sampling strategy in the latter situation.

To achieve significance level \(\alpha\) and statistical power \(1 - \omega\), one can use the following relationship between the noncentrality parameter and the critical value of chi-square statistic \(X_{K-1}^2\):

\begin{equation}
1 - \omega = P(X_{K-1}^2(\xi) > \chi_{K-1,1-\alpha}^2).
\end{equation}

A special case is the two-sample exposure comparison when \(K = 2\). With sample proportions \(\pi_0\) and \(\pi_1\) for the baseline group and the comparison group, respectively, we are interested in testing whether the two groups have the same mean or whether the latter has a mean exposure difference of \(\beta\). With significance level \(\alpha\) and statistical power \(1 - \omega\), the relationship between sample size \(n\) and the critical values is

\begin{equation}
(Z_{1-\alpha/2} + Z_{1-\alpha})^2 = \xi^2/\Sigma_1 = n\beta^2\pi_0\pi_1\mathbf{V}^{-1}1,
\end{equation}

where \(\mathbf{V}\) is the common covariance matrix of \(\mathbf{Y}\) assuming homogeneity of group exposure variance.

**Long-term exposure trend**

To test for the presence of a long-term exposure trend, as in Example 2, one can use a direct approach to test whether the time-dependent trend of the model is different from zero. The simple mixed-effects model (3) in the case of log-normal exposure distribution may be written as

\begin{equation}
Y_{ij} = b_i + \gamma + \beta \times j + \epsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m,
\end{equation}

which is essentially the same as the models given in Symanski et al. (2001a,b) but is a simpler form. It can be shown that under this model,

\begin{equation}
\xi = \Sigma = \sum_{i=1}^{n} \left( \mathbf{X}_i - \frac{m+1}{2} \mathbf{1} \right)^\top \mathbf{V}_{ij}^{-1} \left( \mathbf{X}_i - \frac{m+1}{2} \mathbf{1} \right),
\end{equation}

where \(\mathbf{X}_i = (1, 2, \ldots, m)\). For the case of unequally spaced sampling measurements, one may plug in the desired sampling times in the vector \(\mathbf{X}_i\) of equation (14). The relationship between noncentrality parameter and critical values may be written as follows:

\begin{equation}
(Z_{1-\alpha/2} + Z_{1-\alpha})^2 = \xi^2/\Sigma_1 = \xi.
\end{equation}

When heterogeneity of exposure trends in individual workers is of study interest, one may incorporate individual random slope deviation in the model, i.e.

\begin{equation}
Y_{ij} = \gamma + b_{0i} + b_{1i} \times j + \beta \times j + \epsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m,
\end{equation}

where

\begin{equation}
\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{b_0}^2 & \rho_{b_0} \sigma_{b_1} \\ \rho_{b_0} \sigma_{b_1} & \sigma_{b_1}^2 \end{pmatrix} \right)
\end{equation}

and are independent with \(\epsilon_{ij}\). The covariance between \(y_{ij}\) and \(y_{ij'}\) in this case becomes

\begin{equation}
\text{cov}(y_{ij}, y_{ij'}) = \sigma_{b_0}^2 + 2j \times \rho_{b_0} \sigma_{b_1} + j^2 \times \sigma_{b_1}^2 + \sigma_{\epsilon}^2, \quad j \neq j'.
\end{equation}
\[
\text{cov}(y_{ij}, y_{i'j'}) = \sigma_b^2 + (j + j') \times \rho_b \sigma_b \sigma_{b1} + j \times j' \times \sigma^2_b, \quad j \neq j'.
\]

Here, the correlation \( \rho_b \) between the random intercept \( b_{0i} \) and the random slope \( b_{1i} \) is nonzero to allow possible dependence of the two terms from the same subject. The correlation \( \rho_b \) would simply be zero when the two terms are independent with each other. The covariance matrix \( \mathbf{V}_i \) of equation (14) will need to be changed accordingly. Because the main statistical inference would be for the population’s averaged long-term trend, we focused the calculations for sampling strategy here, however, on the former long-term trend, we focused the calculations for repeated measurements \( n \), repeated measurements \( m \) and statistical power \( 1 - \omega \). We may obtain the corresponding results for \( n = 5, 10, 15 \) for multiple- and two-sample working group exposure comparisons, and long-term trend, respectively. As a reference, the given number of repeated measurements \( m \), Table 1 summarizes the formula of sample size \( n \) for testing a two-sample exposure difference and presence of a long-term exposure trend under various correlation structures. The detailed expressions of \( \xi, \Sigma \) in each case are summarized in the Appendix.

NUMERICAL RESULTS

For illustrative purpose, we tabulate the required sample size to achieve a testing power \( 1 - \omega = 0.90 \) with Type-I error rate \( \alpha = 0.05 \) under the specific null hypothesis \( H_0 : \beta = 0 \) versus the alternative hypothesis \( H_1 : \beta = \beta_1, \gamma = \gamma_1 \) for the two-sample exposure difference \( (K = 2) \) and the long-term trend tests. In addition, the case of multiple group exposure differences with four comparison groups \( (K = 4) \) is tabulated under the null hypothesis \( H_0 : \nu = 0 \) versus the alternative hypothesis \( H_1 : \nu = (0.5, 0.5, -0.5, -0.5)' \). Of note, the specified nuisance parameter \( \gamma_1 \) in the alternative hypothesis does not affect sample size calculations in Table 1. The variance components of between-worker variance \( \sigma_b^2 \) and within-worker variance \( \sigma^2 \) were set between 0 and 2 to conform to existing data sets analytic outcomes (Kromhout et al., 1993; Weaver et al., 2001; Symanski et al., 2006). Depending on the study purpose, the maximum repeated measurement number \( m \) varies between 4 and 15 representing yearly measurements (long-term exposure trend) and workday shift exposures (multiple group comparisons) in the following tabulations, respectively. The correlation coefficient \( \rho \) indicates the dependence between repeated measurements, with \( \rho = 0 \) to \( \rho = 0.8 \) representing that the relationship is null to highly correlated. For sample size calculations other than the tabulated settings, one may refer to the equations (9)–(12) for multiple group or two-sample exposure difference and (15) for long-term trend or the general test statistic (5) for sampling strategy of specific study setting.

**Multiple group mean exposure differences**

Considering various combinations of the correlation coefficient \( \rho (= 0, 0.3, 0.5, 0.8) \), the number of repeated measurements \( m = 5, 10, 15 \), within-worker variance \( \sigma^2_b (= 0.5, 1, 2) \), correlation structures IND, CS, AR(1), and between-worker variance \( \sigma^2 (= 0, 0.5, 1, 2) \), we tabulated the required sample sizes for \( K = 2 \) and \( K = 4 \) separately in Table 2 (the rows with \( \rho = 0 \) correspond to the IND case). For conciseness, we list only the outcomes with equal
Table 2. Sample sizes for the multiple group mean exposure difference test: (a) $K = 2, \beta_1 = 1$; (b) $K = 4$, \( \psi = (0.5, 0.5, -0.5, -0.5) \)

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>( m )</th>
<th>( \sigma^2_\varepsilon )</th>
<th>( \sigma^2_\varepsilon )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CS</td>
<td>AR(1)</td>
</tr>
<tr>
<td>(a) $K = 2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>5</td>
<td>0.5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>9</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>10</td>
<td>0.5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>5</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>9</td>
<td>51</td>
</tr>
<tr>
<td>15</td>
<td>0.5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>6</td>
<td>48</td>
</tr>
<tr>
<td>0.3</td>
<td>5</td>
<td>0.5</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>14</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>37</td>
<td>27</td>
<td>79</td>
</tr>
<tr>
<td>10</td>
<td>0.5</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>8</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>15</td>
<td>74</td>
</tr>
<tr>
<td>15</td>
<td>0.5</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>5</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>10</td>
<td>72</td>
</tr>
<tr>
<td>0.5</td>
<td>5</td>
<td>0.5</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>19</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>51</td>
<td>37</td>
<td>93</td>
</tr>
<tr>
<td>10</td>
<td>0.5</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>11</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>47</td>
<td>22</td>
<td>89</td>
</tr>
<tr>
<td>15</td>
<td>0.5</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>8</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>45</td>
<td>15</td>
<td>87</td>
</tr>
<tr>
<td>0.8</td>
<td>5</td>
<td>0.5</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>36</td>
<td>30</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>71</td>
<td>59</td>
<td>113</td>
</tr>
<tr>
<td>10</td>
<td>0.5</td>
<td>18</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>35</td>
<td>22</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>69</td>
<td>43</td>
<td>111</td>
</tr>
<tr>
<td>15</td>
<td>0.5</td>
<td>18</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>35</td>
<td>17</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>69</td>
<td>33</td>
<td>111</td>
</tr>
<tr>
<td>(b) $K = 4$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>5</td>
<td>0.5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>9</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>18</td>
<td>40</td>
</tr>
<tr>
<td>10</td>
<td>0.5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>5</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>9</td>
<td>31</td>
</tr>
<tr>
<td>15</td>
<td>0.5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>6</td>
<td>28</td>
</tr>
<tr>
<td>0.3</td>
<td>5</td>
<td>0.5</td>
<td>10</td>
</tr>
</tbody>
</table>
proportions for the comparison groups (i.e. $p_0 = p_1 = 0.5$ for $K = 2$ and $p_i = \ldots = p_4 = 0.25$ for $K = 4$), which is optimal for the log-normal distribution case (Table 1). The tabulated sample sizes conformed to those in the literature (e.g. Table 1 of Rappaport et al., 1995 and Table 2 of Rappaport et al., 1999) given the same conditions under the IND correlation structure. However, more sample sizes might be required to ensure statistical power if the assumed correlation structure was CS or AR(1) in the above analyses. Following the tabulated outcomes, even though the length of study period with $m = 10$ is twice that with $m = 5$, the savings in sample sizes are very limited. For example, given $\rho = \sigma_b^2 = \sigma_e^2 = 0.5$ in the AR(1) case ($K = 2$), the required sample size decreases only from 20 to 16 with $m$ increases from 5 to 10, similarly as one increases $m$ from 5 to 15 (Table 2a). The outcomes for the case of $K = 4$ are similar (Table 2b). This result is consistent with the conclusion of Lampa et al. (2006) that one should include as many workers as possible regardless of the size of the variance component for optimization of the measurement strategy of a one-way random-effects ANOVA model. Given that all other conditions are the same, the sample size under the CS correlation in general is larger than that under the AR(1) correlation. As expected, sample size increases with $\sigma_b^2$ and $\sigma_e^2$. Fig. 2 depicts the relationships between required sample size $n$ and $p_{0.5}$ (Fig. 2a). Also, sample size increases with $\rho$ (Fig. 2b).

Long-term exposure trend

Similar to preceding tables, Table 3 lists the required sample sizes for different combinations of $\rho, m, \sigma_b^2$ under various correlation structures IND, CS, and AR(1). Because often one would expect a slight long-term decrease in exposures of the working environment, here we tabulated the sample sizes required in the case of $\beta_1 = -0.2$. Depending on the scale of the study as well as the exposure trend, the sample sizes required may be small as those given in Table II of Symanski et al. (1996) for relatively simple comparisons between repeated surveys.
(m = 2) within the same plant with substantial exposure trend. The tabulated outcomes with slope $β_1 = -0.2$ are consistent with those in Tables 1 and 2 of Symanski et al. (2001a) for more complicated situations of occupational sectors involving multiple plants, buildings, job groups, etc. with a small scale of slopes ($-0.013 \sim -0.079$). Unlike the case for multiple group exposure differences, sample size $n$ required for the long-term trend test decreases drastically as the number of repeated measurement $m$ increases from 2 to 4. Given that all other conditions are the same, in contrast with the multiple group comparison case, sample size required under the CS correlation is in general smaller than that under the AR(1) correlation. Figure 3 depicts the relationship between the sample size $n$ and the correlation coefficient $ρ$ under the CS and AR(1) correlation structures ($σ_ε^2 = 1$), which shows that $n$ decreases with $ρ$.

**Table 3. Sample sizes for the long-term exposure trend test ($β_1 = -0.2$)**

<table>
<thead>
<tr>
<th>$ρ$</th>
<th>$m$</th>
<th>$σ_ε^2$</th>
<th>1 CS</th>
<th>2 CS</th>
<th>3 CS</th>
<th>4 CS</th>
<th>1 AR(1)</th>
<th>2 AR(1)</th>
<th>3 AR(1)</th>
<th>4 AR(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>526</td>
<td>526</td>
<td>1052</td>
<td>1052</td>
<td>1577</td>
<td>1577</td>
<td>3</td>
<td>132</td>
<td>132</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>53</td>
<td>53</td>
<td>106</td>
<td>106</td>
<td>158</td>
<td>158</td>
<td>4</td>
<td>132</td>
<td>132</td>
</tr>
<tr>
<td>0.3</td>
<td>2</td>
<td>368</td>
<td>368</td>
<td>736</td>
<td>736</td>
<td>1104</td>
<td>1104</td>
<td>3</td>
<td>92</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>92</td>
<td>120</td>
<td>184</td>
<td>240</td>
<td>276</td>
<td>359</td>
<td>4</td>
<td>37</td>
<td>56</td>
</tr>
<tr>
<td>0.5</td>
<td>2</td>
<td>263</td>
<td>263</td>
<td>526</td>
<td>526</td>
<td>789</td>
<td>789</td>
<td>3</td>
<td>66</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>66</td>
<td>99</td>
<td>132</td>
<td>198</td>
<td>198</td>
<td>296</td>
<td>4</td>
<td>27</td>
<td>51</td>
</tr>
<tr>
<td>0.8</td>
<td>2</td>
<td>106</td>
<td>106</td>
<td>211</td>
<td>211</td>
<td>316</td>
<td>316</td>
<td>3</td>
<td>27</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>11</td>
<td>29</td>
<td>57</td>
<td>32</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**DISCUSSION AND CONCLUSIONS**

We have proposed a general framework of sampling strategies for exposure assessment under the GLM approach, which could be adapted accordingly to various hypotheses for testing problems for a specific study purpose. The GLM approach is known to be more flexible in distributional assumption and has been used widely in epidemiological and medical studies. As shown in Fig. 1, it is likely that a gamma-distributed exposure has been mistakenly assumed as having a log-normal distribution. Sample size required in this case may be substantially larger than that obtained when assuming a log-normal distribution (not shown here). It is thus preferable to have a sampling strategy that covers exposure distributions other than the common log-normal distributional assumption in industrial hygiene practice. To analyze collected data using the GLM, one may apply the statistical software SAS 9.1 procedure PROC GENMOD by specifying a distributional assumption together with a link function in the model option statement. For example, one may specify the distribution dist = normal and the link function link = log for log-normal distribution and dist = gamma for gamma distribution using this procedure. For hypothesis testing using the GEE analysis, one may also include an additional command line in the same procedure by specifying the assumed correlation structure for the repeated measurements. One important feature of the GEE analysis is that the analytical outcomes are fairly insensitive even when the correlation structure is misspecified (Zeger and Liang, 1986). Thus, model selection to determine either CS or AR(1) as the more appropriate covariance structure (see for example, Symanski et al., 2001a) might not have much effect on the results of a GEE analysis.

As a practical guide for field workers in occupational exposure assessment, we have proposed sampling strategies that assume a log-normal distribution as the exposure distribution and then use the GLM approach with a log-normal distributional assumption. The sample size required for the long-term trend test decreases drastically as the number of repeated measurements increases from 2 to 4. This is in contrast with the multiple group comparison case, where sample size required under the CS correlation is in general smaller than that under the AR(1) correlation. Figure 3 depicts the relationship between the sample size $n$ and the correlation coefficient $ρ$ under the CS and AR(1) correlation structures ($σ_ε^2 = 1$), which shows that $n$ decreases with $ρ$.
Fig. 2. Relationships between sample size $n$ and baseline sample proportion $\pi_0$ and correlation coefficient $\rho$ under CS and AR(1) for the two-sample mean exposure difference test: (a) sample size $n$ and $\pi_0$; (b) sample size $n$ and $\rho$ ($\sigma^2_b = 1, \sigma^2_e = 0.5$).

Fig. 3. Relationship between sample size $n$ and correlation coefficient $\rho$ under CS and AR(1) for the long-term exposure trend test ($\sigma^2_e = 1$).
distribution and specifically follow the two examples of multiple exposure group comparisons and long-term exposure trend. Although the sample size calculations were based on the test statistic (5) under the special case of log-normal distribution, one can show that the derived formulas given in Table 1 are the same as those obtained using the maximum likelihood estimation equations (not shown here). Therefore, at the data analysis stage, one may either employ the SAS procedures PROC MIXED by assigning the log-transformed measurements as the model response variable or PROC GENMOD by specifying a normal distribution and the log-link function given above. Given a desired significant level and statistical power under the null and the alternative hypothesis, respectively, one may determine, in part, the choice between increasing the sample size \( n \) and increasing the number of repeated measurements \( m \) for optimal sampling strategy by the ratio of the between-worker variance \( \sigma_w^2 \) and the within-worker variance \( \sigma_e^2 \) (Burdorf and van Tongeren, 2003). For multiple group exposure differences, a prolonged study period (by increasing the number of repeated measurements) may not be beneficial in terms of sample size savings, which is similar to the conclusion of Lampia et al. (2006). However, for a long-term exposure trend, a 5- to 10-fold sample size savings is realized by increasing the number of repeated measurements from \( m = 2 \) to \( m = 4 \). In light of the property that the inter-individual variance term \( \sigma_i^2 \) is algebraically cancelled out for the long-term exposure trend test, it would be desirable to have equally spaced repeated measurements to achieve the optimal design. In the case of unequally spaced sampling times, the between-worker variance \( \sigma_w^2 \) would have a significant effect on the sample size requirement. We leave the corresponding sampling strategy, however, for future research work.

We have presented the sampling strategies mainly considering the number of repeated measurements \( (m) \) per subject is the same for ease of planning at the design stage. During the sampling or data collection stage, unbalanced data may be resulted from situations such as losing follow-up or missing at random during some of the repeated measurements for some study subjects. Statistical analysis for group mean differences or trend may simply replace the subscript \( j = 1, \ldots, m \) of \( f_d \) equation (1) by \( j = 1, \ldots, m_i \), similar to the setting in \( f_d \) equation (1) of Liu and Liang (1997). The covariance matrix \( V_i \) of \( Y_i \), and the test statistic mainly based on the asymptotic property of the noncentral parameter \( \xi = \xi^T \Sigma^{-1} \xi \) will change accordingly [equations (4)–(6)]. The SAS procedure PROC GENMOD may also perform the corresponding statistical analysis for the unbalanced data set. Also, the essentially approximated results would not be affected much with small variations in the number of repeated measurements per subject. Therefore, it appears reasonable assuming equal number of repeated measurements for sampling strategies at the design stage.

The proposed general framework for sampling strategies obtains a required sample size by calculating the noncentral parameter of the asymptotic chi-square distribution of the test statistic under a specified alternative hypothesis. The procedure is suitable for testing mean exposure differences or trend existence among working groups. Some important occasions such as testing for heterogeneity of exposure variability among groups may not be applicable with this approach, unless there are simultaneous group mean exposure differences so that the noncentral parameter is nonzero under the alternative hypothesis. The heterogeneity hypothesis testing may be achieved by using the likelihood ratio of the test score (5) under the null hypothesis of homogeneous variance versus the alternative hypothesis of heterogeneity similarly as proposed in the literature (Symanski et al., 2001b; Weaver et al., 2001; van Tongeren et al., 2006). However, further investigations in sample size calculations for the corresponding sampling strategy are required; these investigations were outside the scope of this paper.

Correlations structures between repeated measurements obtained from the same worker also affect the required sample size. As shown in Tables 2 and 3, the sample size required under the CS correlation is generally larger than that under the AR(1) for multiple group exposure comparisons, whereas the opposite holds for presence of a long-term exposure trend. For the multiple group exposure difference case, the sample size required under the IND correlation is the smallest, as expected, because the information gained from each additional repeated measurement is optimal. In contrast, for the long-term trend test, the required sample sizes for the IND correlation structure were larger than those for the CS and AR(1) correlations and decreased with the correlation \( \rho \). This might be due to the nature of the trend test—the higher the correlation the more information one could obtain from additional repeated measurement. Thus, the number of repeated measurements is more efficient in required sample size. As shown in Table 3, sample sizes decreased almost 10 and 4 times for the CS and AR(1) cases, respectively, as \( m \) increased from 2 to 4 with additional information obtained from the number of repeated measurements.

The tabulated outcomes are by no means complete and attempting to cover all the general cases would be exhausting. Thus, for sampling strategies other than the specified hypotheses and alternatives, one may need to refer to formulas given in equations (9)–(12) and Table 1 or to derive formulas from the general test statistic (5). Also, with some extra work, the proposed sampling strategies could be extended to other cases such as nested mixed-effects ANOVA model with more complicated random-effects
variables. For more information on sampling design issues under the general framework of GLM, one may refer to Chapter 2 of Diggle et al. (1994). Finally, the proposed sampling strategies were aimed at the study design stage. To ensure the statistical significance level and testing power of hypothesis testing in cases of missing data, we recommend that additional samples that are proportional to the ratio of missing data be recruited at the planning stage.

**FUNDING**

Institute of Occupational Safety and Health, Council of Labor Affairs of Taiwan (IOSH95-A101); National Health Research Institutes (BS-096-PP-11).

**Acknowledgements**—The authors thank Ms Jia-Yan Yu for help with part of the calculations.

**APPENDIX 1**

**Covariance and inverse covariance matrix of** $y_i = (y_{i1}, \ldots, y_{im})'$:

(i) IND

$$\text{cov}(y_{ij}, y_{ij}) = \sigma_b^2 + \sigma_e^2, \ j = j;$$

$$\text{cov}(y_{ij}, y_{ij'}) = \sigma_b^2, \ j \neq j'.$$

$$V_i^{-1} = \left(\frac{(m\sigma_b^2 + \sigma_e^2)}{\sigma_e^2(m\sigma_b^2 + \sigma_e^2)}\right) \mathbf{I} - \left(\frac{\sigma_b^2}{\sigma_e^2(m\sigma_b^2 + \sigma_e^2)}\right) \mathbf{J},$$

where $\mathbf{I}$ is an $m \times m$ diagonal matrix with entries 1 on the diagonal and 0 elsewhere, and $\mathbf{J}$ is an $m \times m$ with all entries equal to 1.

(ii) CS

$$\text{cov}(y_{ij}, y_{ij}) = \sigma_b^2 + \sigma_e^2, \ j = j;$$

$$\text{cov}(y_{ij}, y_{ij'}) = \sigma_b^2 + \rho\sigma_e^2, \ j \neq j'.$$

$$V_i^{-1} = \left(\frac{(m\sigma_b^2 + [1 + (m - 1)\rho]\sigma_e^2)}{(1 - \rho)\sigma_e^2(m\sigma_b^2 + [1 + (m - 1)\rho]\sigma_e^2)}\right) \mathbf{I} - \left(\frac{\sigma_b^2}{(1 - \rho)\sigma_e^2(m\sigma_b^2 + [1 + (m - 1)\rho]\sigma_e^2)}\right) \mathbf{J},$$

(iii) AR(1)

$$\text{cov}(y_{ij}, y_{ij}) = \sigma_b^2 + \sigma_e^2, \ j = j;$$

$$\text{cov}(y_{ij}, y_{ij'}) = \sigma_b^2 + \rho|j-j'|\sigma_e^2, \ j \neq j'.$$

$$V_i^{-1} = \frac{1}{(1 + \rho)\sigma_e^2} \left\{ \frac{1}{1 - \rho} F - \frac{\sigma_b^2}{m - (m - 2)\rho|\sigma_b^2| + (1 + \rho)\sigma_e^2} G \right\},$$

where

$$F = \begin{pmatrix} 1 & -\rho & 0 & \ldots & 0 \\ -\rho & 1 + \rho^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 1 + \rho^2 & -\rho \\ 0 & 0 & \ldots & 1 & 1 \end{pmatrix}$$

$$G = \begin{pmatrix} 1 & 1 - \rho & \ldots & 1 - \rho & 1 \\ 1 - \rho & (1 - \rho)^2 & \ldots & (1 - \rho)^2 & 1 - \rho \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 - \rho & (1 - \rho)^2 & \ldots & 1 + \rho^2 & 1 - \rho \\ 1 & 1 - \rho & \ldots & 1 - \rho & 1 \end{pmatrix}$$

**Inverse matrix formula in deriving** $V_i^{-1}$:

$$(A + BDB')^{-1} = A^{-1} - A^{-1}B(A^{-1}B + D^{-1})^{-1}B'A^{-1} + A^{-1}BE(E + D^{-1})^{-1}EBA' - 1,$$

where $A$ and $D$ are nonsingular matrices of orders $m$ and $n$, and $B$ is an $m \times n$ matrix, $E = (B'A^{-1}B)^{-1}$ (Rao, 2001, p. 33).

**Noncentrality parameter approximation in the case of log-normal distribution:**

$$\xi = E_{H_i}[U(0, \hat{\gamma})] \approx \sum_{i=1}^{n} P_i V_i (\mu_i^1 - \hat{\gamma}^1),$$

and

$$\Sigma_i = \text{cov}_{H_i}[U(0, \hat{\gamma})] \approx \sum_{i=1}^{n} P_i V_i^{-1} P_i,$$

where

$$\mu_i^1 = E_{H_i}[Y_i],$$

$$P_i = \left(\frac{\partial \mu_i}{\partial \beta}\right)' - \mathbf{I}_{p_{i1}}^{-1} P_i^{-1},$$

$$\mathbf{I}_{p_{i1}} = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta}\right)' V_i^{-1} \left(\frac{\partial \mu_i}{\partial \beta}\right).$$
\[ I_{yy} = \sum_{i=1}^{n} \mathbf{1}' \mathbf{V}_i^{-1} \mathbf{1}, \]

and \( \hat{\gamma}^* \) is the solution of

\[ \lim_{n \to \infty} n^{-1} \mathbb{E}_H \left\{ \sum_{i=1}^{n} \mathbf{1}' \mathbf{V}_i^{-1} (\mathbf{Y}_i - \gamma \mathbf{1}) \right\} = 0. \]

**Expressions of \( \xi, \Sigma_1 \) under the two hypothesis testing:**

Two-sample mean exposure difference

**IND**

\[ \xi = \Sigma_1 = \frac{mn \pi_0 \pi_1}{m\sigma_\pi^2 + \sigma_\varepsilon^2} \]

**CS**

\[ \xi = \Sigma_1 = \frac{mn \pi_0 \pi_1}{m\sigma_\pi^2 + \sigma_\varepsilon^2 \{1 + (m - 1)p\}^2} \]

**AR(1)**

\[ \xi = \Sigma_1 = \frac{mn \pi_0 \pi_1 [m - (m - 2)p]}{[m - (m - 2)p] \sigma_\pi^2 + (1 + p) \sigma_\varepsilon^2 \} \]

A long-term exposure trend

**IND**

\[ \xi = \Sigma_1 = \frac{mn(m^2 - 1)}{12 \sigma_\varepsilon^2} \]

**CS**

\[ \xi = \Sigma_1 = \frac{mn(m^2 - 1)}{12(1 - \rho) \sigma_\varepsilon^2} \]

**AR(1)**

\[ \xi = \Sigma_1 = \frac{mn(m - 1)}{12(1 - \rho^2) \sigma_\varepsilon^2} \left\{ \frac{m(m + 1)(1 - \rho)^2}{6p[m + 1 - (m - 1)p]} \right\}. \]

**REFERENCES**


