Title

On Choice of Subject and Replicates in Estimating Among and Within Subject Variation

Authors

Weng-Kee Wong
Peter A. Lachenbruch
Philip J. Clements

Publication Date

2011-10-24

Biom. J. 38 (1996) 7, 767–777 Akademie Verlag

On Choice of Subjects and Replicates in Estimating Among and Within Subject Variation

WENG KEE WONG

Department of Biostatistics
University of California at Los Angeles, USA

PETER A. LACHENBRUCH

Center for Drug Evaluation and Research
Food and Drug Administration
Rockville, MD, USA

PHILIP J. CLEMENTS

School of Medicine
University of California at Los Angeles, USA

Summary

We consider the problem of choosing the number of replicates and number of subjects in a components of variance problem which optimizes various criteria. The case study here involves patients suffering from systemic sclerosis (Scleroderma), a form of rheumatic disease that is potentially disabling. Under the physical constraints imposed on the study, we find that using 2 or 3 replicates with as many patients as possible is the optimal strategy for several criteria.

Key words: Cost; Nested design; Power; Random effects model; Sample size.

1. Introduction

In designing a study to evaluate D-penicillamine therapy in the treatment of Scleroderma, a basic response variable is a skin score which is based on the number and severity of affected areas of the skin (CLEMENTS et al., 1993). Since this measure may depend on observer, patient and a random component measurement we needed to estimate variation due to these sources of error prior to conducting the study. A special study in which several participating observers or physicians measured the same patients one time provided an estimate of among observer variation which could be compared to estimates derived from other studies (CLEMENTS et al., 1993). After this was completed, we wanted to perform a study in which

the intra-observer error measurements could be estimated, as well as variation due to patients.

There are several practical constraints imposed on this study. The observers are Rheumatologists with a special interest in Scleroderma, who have been trained in skin scoring and are located far apart in the U.S. They are in practice throughout the U.S. and so it was not possible to have repeated measurements with multiple observers. Because the skin score may change rapidly during the early phase of the disease (averages about 1 or 2 units per month), it was important to obtain the replicate measurements in a short time, but not so short that the observer's memory of the patient would bias the score. Generally, observation periods of longer than 6 weeks were discouraged. Past experience with this rare disease suggests that it would be difficult for each observer to recruit more than 5 patients as subjects. Limitations on observer's time is also an important consideration in this study.

The overall goal of this research was to study differences in therapies in a multicenter study using the skin score as an endpoint. It is important to quantify the components of variation of this score which are due to among-patient variation, within-patient variation, and among-observer variation. Since a patient will be seen only by one observer, it is necessary to quantify the among-observer variation and to attempt to reduce it as much as possible. A previous study had shown that such variation could be quite large, but this was conducted at a common site, so observers were not as familiar with the physical arrangement as in their own offices. The patients in the earlier study were read by all observers. In the proposed study, the observers would obtain patients from their local practice, and no multiple observations from different observers would be possible. The problem became one of estimating the variation using the nesting of observer, patient within observer, replication within patient within observer. Hence, the goal was to determine how many replicates, and patients would be needed to conduct the validation study prior to the study comparing treatments. For example, a key design question is, with the 18 observers we have in this study, would having three patients measured at three times (closely spaced) be a reasonable allocation of resources? Or using 5 patients with 2 readings or 2 patients and 5 readings be a superior allocation of resources? The ultimate aim was to ensure that (a) the observers had the same mean and a small variance and (b) the larger study would have only a single reading from each observer. This paper discusses the optimal allocation of patients and replicates under these restrictions and provides some solutions.

2. A Statistical Model

We consider a nested design, with patients nested within observers, and reading nested within patients. Patients are clearly random samples; the observers may not be if we wish to infer only to this group of observers. If we hope to infer to what

would happen for a group of observers, some of whom are not included here, we can push toward observers being random as well. We first consider the case when observers are random and analyzed the data (y_{ijk}) using the statistical model

$$y_{ijk} = A + D_i + P_{j(i)} + e_{ijk},$$
$$i = 1, 2, \ldots, d, \quad j = 1, 2, \ldots, p, \quad k = 1, 2, \ldots, r.$$

Here A is an overall constant, D_i is the effect of the ith observer and $P_{j(i)}$ is the effect of the jth patient observed by the ith observer. Both D_i and $P_{j(i)}$ are random variables with means zero and variances σ_D^2 and σ_P^2 respectively. The unobservable errors are $e'_{ijk}s$ each with mean 0 and variance σ^2. The choice of this balanced nested design model lies in its simplicity and at the same time, can provide us estimates of among patient variability, and measurement error, subject to these assumptions: (a) the underlying "true" disease score is constant during the period of observation; and (b) the intra-observer variation is constant, although a well trained observer might have smaller variation than an inexperienced observer would. All random variables are assumed to be independent and normally distributed.

The data were analyzed using Analysis of Variance and expected mean squares provide us with information on variability among observers (σ_D^2), variability among patients within observer (σ_P^2) and measurement error (σ^2). The ANOVA table is given below with

$$\bar{y}_{i..} = \sum_{j=1}^{p} \sum_{k=1}^{r} y_{ijk}/rp, \qquad \bar{y}_{ij.} = \sum_{k=1}^{r} y_{ijk}/r$$

and

$$\bar{y}_{...} = \sum_{i=1}^{d} \sum_{j=1}^{p} \sum_{k=1}^{r} y_{ijk}/drp.$$

Table 1 gives the analysis for a subset of the observers participating in the intra-observer variability study by CLEMENTS et al. (1995) based on $d = 17$ observers, $p = 3$ patients and $r = 3$ readings per patient. The data from the 18th observer is not available. From the table, we obtain point estimates of the variance components to be

$$\hat{\sigma}^2 = 5.50, \qquad \hat{\sigma}_P^2 = (S_2^2 - S_3^2)/r = (137.18 - 5.50)/3 = 43.90,$$

and

$$\hat{\sigma}_D^2 = (S_1^2 - S_2^2)/rp = (570.96 - 137.18)/9 = 48.20.$$

In determining the optimal values of the number of patients and the number of replications to produce good estimates of the variance of the estimated variance components, it is helpful to define $\varrho = \sigma_P^2/\sigma^2$ and $\omega = \sigma_D^2/\sigma^2$. Here, ϱ is the relative variance of patients to measurement error and the ratio ω is the relative variance of physicians to measurement error. These ratios of variance components in the random effect model have been used in various settings for a long time. For instance in animal sciences, these ratios are frequently used to estimate the genetic

Table 1
Analysis of Variance for Skin Score

source	df	MS	EMS
observers	$d-1$	$S_1^2 = rp\sum_{i=1}^{d}\sum_{j=1}^{p}(\bar{y}_{i..}-\bar{y}_{...})^2/(d-1)$	$\sigma^2 + r\sigma_P^2 + rp\sigma_D^2$
patients (observers)	$n_2 = d(p-1)$	$S_2^2 = r\sum_{i=1}^{d}\sum_{j=1}^{p}(\bar{y}_{ij.}-\bar{y}_{i..})^2/d(p-1)$	$\sigma^2 + r\sigma_P^2$
residual (patient)	$n_3 = dp(r-1)$	$S_3^2 = r\sum_{i=1}^{d}\sum_{j=1}^{p}\sum_{k=1}^{r}(\bar{y}_{ijk}-\bar{y}_{ij.})^2/dp(r-1)$	σ^2

source	partial *SS*	df	MS	*F*	Prob > *F*
observers	9135.35	16	570.96	4.16	0.0002
patients (observers)	4664.22	34	137.18	24.96	0.0000
residual (patient)	560.67	102	5.50		
total	14360.24	152			

heritability of a certain trait of livestock breeders (GRAYBILL et al., 1956). Their ANOVA estimates are obtained directly from the above estimates:

$$\hat{\varrho} = 43.90/5.50 = 7.99 \quad \text{and} \quad \hat{\omega} = 48.20/5{:}50 = 8.77\,.$$

Minimum variance unbiased estimators are more complicated than ANOVA estimators and are given by

$$\frac{1}{r}\left\{\frac{S_2^2}{S_3^2}\left(1-\frac{2}{dp(r-1)}\right)-1\right\} \quad \text{and} \quad \frac{1}{rpS_3^2}\left(1-\frac{2}{dp(r-1)}\right)(S_1^2-S_2^2)$$

respectively (BURDICK and GRAYBILL, 1992, pp. 115). Four our data, these numerical values are 7.82 and 8.60 respectively and so they are very close to the ANOVA estimates.

We can easily find a confidence interval for σ^2 using the χ^2-distribution (MONTGOMERY, 1991, pp. 85). The next goal is to obtain an estimate for σ_P^2. This is done by subtracting S_3^2 from S_2^2 and dividing by r. However, there is no exact confidence interval for σ_P^2 since the estimate is the (scaled) difference of two χ^2 variables which does not have a simple form. However, a confidence interval for ϱ can be found using the F-distribution (MONTGOMERY, 1991, pp. 86). This is discussed more fully in Section 3.3.

3. Choice of the Number of Patients and Replicates

The general design problem here is to determine the optimal values of r and p so that we have a "good" estimate for the variance component σ_P^2. There are d observers, with p patients and r readings per patient. In the context of the medical

practices involved in this study described in section 1, we have the restriction $rp \leq 12$. Three definitions of "good" are suggested here:

1. minimum length confidence interval for σ_P^2
2. minimum variance of the variance component estimators
3. maximum power for testing hypotheses of the form H_0: $\varrho \leq \varrho_0$ vs H_a: $\varrho > \varrho_0$ for some nonnegative constant ϱ_0 or H_0: $\sigma_P^2 = \sigma_{P0}^2$ vs H_a: $\sigma_P^2 \neq \sigma_{P0}^2$ for some nonnegative constant σ_{P0}.

These hypotheses relate to finding precise estimates for σ_P^2 and are generally of interest in later phases of the study.

3.1 Minimizing length of confidence interval for σ_P^2

The procedure given in BURDICK and GRAYBILL (1992, pp. 81) may be used to construct an approximate $100(1-2\alpha)\%$ confidence interval for the variance component σ_P^2. The formula for this interval is

$$\left\{\frac{S_2^2 - S_3^2 - \sqrt{V_L}}{r}, \frac{S_2^2 - S_3^2 + \sqrt{V_U}}{r}\right\},$$

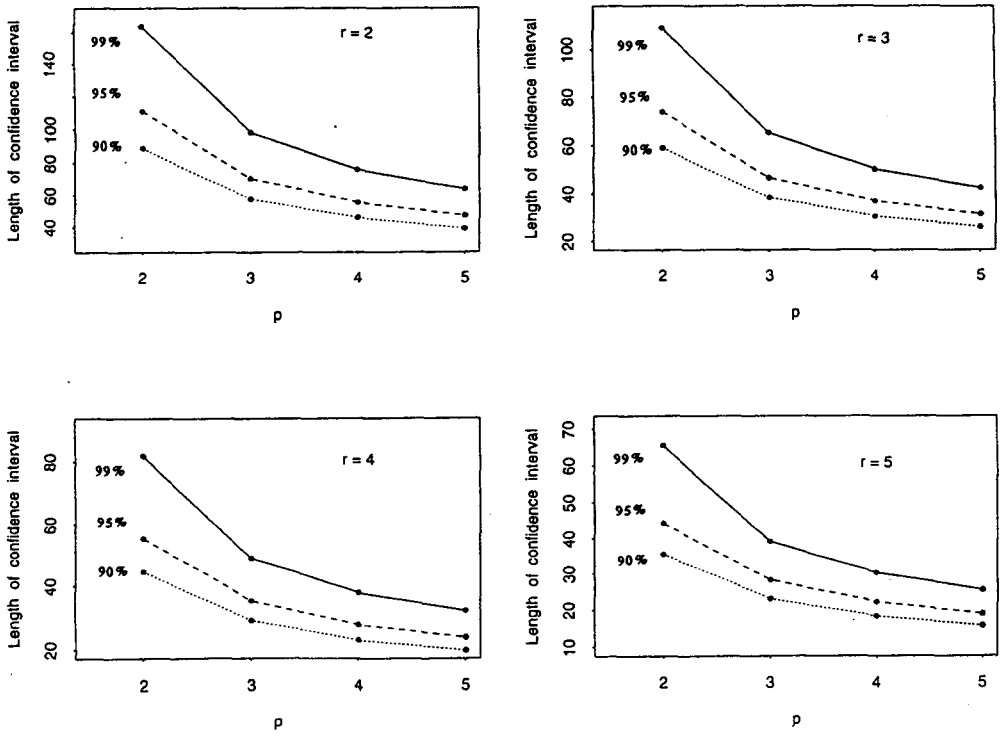

Fig. 1a: Plots of length of the 90.95 and 99% confidence intervals for σ_p^2 fixed r and various p.

where $V_L = G_2^2 S_2^4 + H_3^2 S_3^4 + G_{23} S_2^2 S_3^2$, $V_U = H_2^2 S_2^4 + G_3^2 S_3^4 + H_{23} S_2^2 S_3^2$

$$G_i = 1 - 1/F_{\alpha, n_i, \infty} \qquad H_i = 1/F_{1-\alpha, n_i, \infty} - 1, \quad i = 2, 3,$$

$$G_{23} = \{(F_{\alpha, n_2, n_3} - 1)^2 - G_2^2 F_{\alpha, n_2, n_3}^2 - H_3^2\}/F_{\alpha, n_2, n_3}$$

and

$$H_{23} = \{(F_{1-\alpha, n_2, n_3} - 1)^2 - H_2^2 F_{1-\alpha, n_2, n_3}^2 - G_3^2\}/F_{1-\alpha, n_2, n_3}.$$

Here and throughout, $F_{\alpha, a, b}$ denotes the upper α percentage point of the F-distribution with numerator a degrees of freedom and denominator b degrees of freedom. From the ANOVA table in Section 2, we have $n_2 = 34$, $n_3 = 102$, $S_2^2 = 137.18$ and $S_3^2 = 5.50$. The 90%, 95% and 99% confidence intervals are computed using estimates for σ_D^2, σ_p^2, σ^2 found earlier. The length of the 90%, 95% and 99% confidence intervals for different choices of p, r and α are displayed in Figure 1, where Figure 1a is for a fixed number of replicates and Figure 1b is for a fixed number of patients. These plots suggest that for each p and each α-level we look at, the most dramatic gain per additional measurement occurs when $r = 3$. This is because the slope of the line in the bottom figure is steepest between $r = 2$ and $r = 3$. Thus, the change in the length of the confidence interval has the greatest proportional reduction as r changes from 2 to 3. Between $r = 4$ and $r = 5$, there is not too much change in the length of the confidence intervals for $p = 2$, 3, 4 and 5.

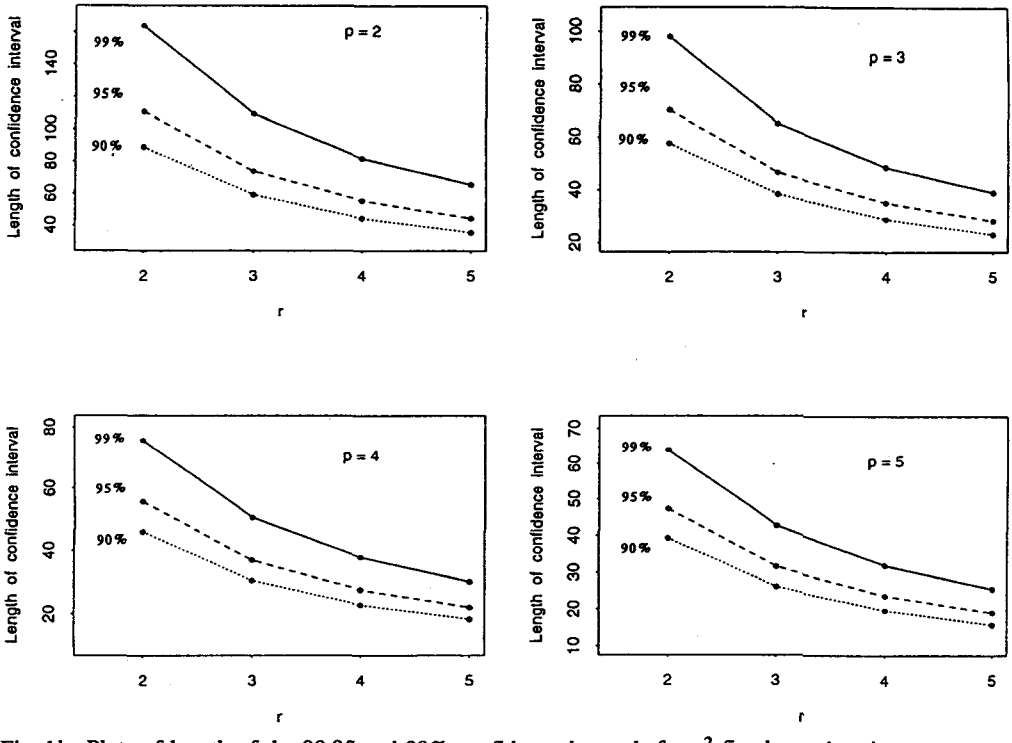

Fig. 1b. Plots of length of the 90.95 and 99% confidence intervals for σ_p^2 fixed p and various r.

3.2 Minimizing variance of the variance component estimators

Alternatively, one may seek to minimize the variance of the variance component associated with the patient factor. Using results from SEARLE (1970) for a two way nested design with random effects factors, the variances of $\hat{\sigma}_P^2$ and $\hat{\sigma}^2$ are respectively given by

$$\text{var}(\hat{\sigma}_P^2) = \frac{2\sigma^4}{r^2}\left\{\frac{(r\varrho+1)^2}{d(p-1)} + \frac{1}{dp(r-1)}\right\} \tag{3.1}$$

and

$$\text{var}(\hat{\sigma}^2) = \frac{2\sigma^4}{dp(r-1)}.$$

If we assume the total number of observations is fixed and equal to $n = drp$, this is equivalent to fixing rp since d is fixed in this problem. Consequently, the argument given in SCHEFFÉ (1959, pp. 237) for a one-way random effects model applies and the optimal choice of r is 2 because the estimated ϱ is 7.99 (≥ 1). Alternatively if the sample size n is unrestricted as in some studies, one may search over the range of values of r and p to minimize (3.1.). In this case, it is clear using the maximal permissible values of r and p are the obvious choices. However, in practice the gain resulting from the increased number of replicates is frequently minimal (as in this study) and may not be worthwhile, especially when taking an additional measurement becomes costly. We are thus led to studying the reduction in var $(\hat{\sigma}_P^2)$ per unit increase in r for fixed p and d. To do this, note that for fixed positive integers d, $p > 1$, $\varrho \geq 0$, the reduction in var $(\hat{\sigma}_P^2)$ due to an additional replicate is proportional to

$$\frac{-1 - 3r + 2pr + 2pr^2 - 2pr\varrho + 2pr^3\varrho}{r^2(r-1)(1+r)^2(p-1)}, \qquad r > 1$$

which can be shown to be a decreasing function in r using straightforward but tedious algebra with ϱ set equal to 7.99 and any positive integer $p > 1$. Thus $r = 2$ replicates with as many patients as possible is a reasonable strategy for estimating var $(\hat{\sigma}_P^2)$ whether n is restricted or not. Table 2 below shows changes in the values of the estimated var $(\hat{\sigma}_P^2)$ when p, r and ϱ are varied using estimates from our data. Note that in general, the precision of $\hat{\sigma}_P^2$ is more affected by changes in p than changes in r, increases as ϱ increases and as expected, decreases if p or r is increased.

3.3 Power of the Testing Procedures

Excessive among-patient variation may reduce the effectiveness of treatment comparisons in a study. Indeed, if such variation is sufficiently large, the tests of treat-

Table 2
Estimated variance of the variance component estimator for σ_P^2.

p	r	$\varrho = 0.10$	$\varrho = 1$	$\varrho = 7.986$	$\varrho = 10$
2	2	1.63	7.97	242.16	370.54
3	2	0.88	4.06	121.15	185.34
4	2	0.61	2.73	80.79	123.58
5	2	0.47	2.06	60.60	92.70
2	3	0.72	6.06	232.43	358.56
3	3	0.38	3.05	116.23	179.30
4	3	0.26	2.04	77.49	119.54
5	3	0.20	1.53	58.12	89.65
2	4	0.45	5.28	227.74	352.74
3	4	0.23	2.65	113.88	176.38
4	4	0.16	1.77	75.92	117.59
5	4	0.12	1.33	56.94	88.19
2	5	0.32	4.85	224.97	349.29
3	5	0.16	2.43	112.49	174.65
4	5	0.11	1.62	74.99	116.43
5	5	0.08	1.22	56.25	87.33

ment effect in the study may have so low a power as to make the study not worth doing. Thus, in a study such as this, it may be useful to test if the among-patient variance is less than a certain multiple of the measurement error. The hypothesis may be stated as H_0: $\varrho \leq \varrho_0$ vs H_0: $\varrho > \varrho_0$ for some nonnegative constant ϱ_0. It is well known that the power of this α-sized test involves only the central F-distribution and is given by

$$\beta(p, r) = Pr\{F_{d(p-1), dp(r-1)} \geq F_{\alpha, d(p-1), dp(r-1)}(1 + r\varrho_0)/(1 + r\varrho)\}.$$

The table below computes the power of this test for selected values of $\varrho = 8$ and $\varrho_0 = 5$ with $d = 18$ observers. As part of a sensitivity analysis, the power when $\varrho = 7$ and 9 also computed. The corresponding powers for $d = 10$ are given in parentheses.

From the table, it is clear that the power of the test is more affected by the number of patients than by the number of replicates. For a fixed number of patients, the difference between having 2 or 5 replicates averages about 17% difference, which increases with a larger number of patients. While this difference may appear appreciable, the greatest gain in power is attained moving from 2 to 3 replicates. In addition, note that the power always increases when ϱ or d increases. The main reason for having a larger number of observers will rest on non-statistical considerations such as credibility of the study. See also DONNER and ELIASZIW (1987) for a related power calculation for the intraclass correlation coefficient given by $1 - 1/(1 + \varrho)$.

Alternatively, if there is interest in testing the hypothesis H_0: $\sigma_P^2 = \sigma_{P0}^2$ vs H_0: $\sigma_P^2 \neq \sigma_{P0}^2$ for some nonnegative constant σ_{P0} the test statistic for this α-sized test is $F^* = S_2^2/\{(1 + r\tilde{\varrho})\, S_3^2\}$ where $\tilde{\varrho} = \sigma_{P0}^2/\sigma^2$. The null hypothesis is rejected in favour

Table 3

Power of tests with various r, p and ϱ values with $d = 18$ and $\varrho_0 = 5$. (Similar results for $d = 10$ are in parentheses).

r					
	ϱ	2	3	4	5
2	7	0.19 (0.15)	0.23 (0.17)	0.25 (0.28)	0.26 (0.19)
	8	0.29 (0.20)	0.35 (0.25)	0.39 (0.27)	0.40 (0.28)
	9	0.39 (0.27)	0.47 (0.51)	0.51 (0.36)	0.52 (0.37)
3	7	0.27 (0.19)	0.33 (0.23)	0.36 (0.26)	0.38 (0.27)
	8	0.42 (0.28)	0.52 (0.36)	0.56 (0.39)	0.58 (0.41)
	9	0.56 (0.38)	0.67 (0.48)	0.72 (0.52)	0.74 (0.54)
p					
4	7	0.34 (0.23)	0.42 (0.29)	0.46 (0.32)	0.48 (0.34)
	8	0.53 (0.36)	0.64 (0.45)	0.69 (0.50)	0.72 (0.52)
	9	0.69 (0.48)	0.80 (0.60)	0.84 (0.64)	0.86 (0.67)
5	7	0.40 (0.27)	0.50 (0.34)	0.55 (0.38)	0.57 (0.40)
	8	0.62 (0.42)	0.74 (0.53)	0.79 (0.58)	0.81 (0.61)
	9	0.79 (0.57)	0.88 (0.69)	0.92 (0.74)	0.93 (0.77)

of the alternative if F^* exceeds $F_{\alpha, d(p-1), dp(r-1)}$. It follows that for a fixed number of observers (d), the power of this test can be maximized by appropriate choice of r and p subject to a fixed number of observations, i.e. $n = drp$. Thus, assuming $\tilde{\varrho}$ is known apriori, r^* and p^* can be found by using the equation

$$Pr\{F^* > (1 + r^*\tilde{\varrho})^{-1} F_{\alpha, d(p^*-1), dp^*(r^*-1)} \mid \tilde{\varrho}\}$$
$$= \max_{drp=n} Pr\{F^* > (1 + r\tilde{\varrho})^{-1} F_{\alpha, d(p-1), dp(r-1)} \mid \tilde{\varrho}\}.$$

This is generally straightforward since we usually are interested in small values of r or p. If $\tilde{\varrho}$ is unknown (which is a more realistic situation), a prior distribution $f(\tilde{\varrho})$ may be specified on $\tilde{\varrho}$ over a certain region r. A suitable choice for $f(\tilde{\varrho})$ may be a chi-squared distribution. Proceeding as in PIGNATIELLO (1987), we now seek to find r^* and p^* so that

$$\int_R Pr\{F^* > (1 + r^*\tilde{\varrho})^{-1} F_{\alpha, d(p^*-1), dp^*(r^*-1)} \mid \tilde{\varrho}\} f(\tilde{\varrho})\, d\tilde{\varrho}$$
$$= \max_{drp=n} \int_R Pr\{F^* > (1 + r\tilde{\varrho})^{-1} F_{\alpha, d(p-1), dp(r-1)} \mid \tilde{\varrho}\} f(\tilde{\varrho})\, d\tilde{\varrho}.$$

This technique is typically complicated since the optimal values of r^* and p^* depend on the prior density $F(\tilde{\varrho})$, which in itself may be problematic to specify. We will not pursue this approach further in this paper.

3.4 *Observers as a Fixed Factor*

We now consider the case when the observers are considered to be a fixed factor. This may be appropriate when we have a higly motivated and skilled set of physicians, and we want to make inference on them and the set of possible patients. Under this situation, the information matrix M can be written as

$$M(\text{fixed, random}) = \begin{pmatrix} F_1 & 0 \\ 0 & F_2 \end{pmatrix},$$

SEARLE (1970). The submatrices F_1 and F_2 represent components from the fixed effects and random effects respectively and are functions of r and p. The objective now is to make inferences on the variance components and the fixed effects parameters by minimizing their asymptotic generalized variances. This may be accomplished by finding r to maximize $\lambda \log |F_1| + (1-\lambda) \log |F_2|$ for each given value of p. Clearly, $\lambda = 0$, $\lambda = 1$ and $\lambda = 1/2$ are of special interest. The optimal choice of r (GIOVAGNOLI and SEBASTIANI, 1989, 1990) for a given ϱ turns out to be

$$r = 1 + \frac{(1-\lambda)(1+\varrho)}{\varrho(1-\lambda+\lambda p)} \quad 0 \leq \lambda \leq 1.$$

Note that as λ increases, r decreases as expected. In the extreme when $\lambda = 1$, only one replication is needed. This corresponds to the case in which interest is in estimating the effects of observers. If $\lambda = 0$, the above formula implies $r = 2 + 1/\varrho$ and if $\lambda = 1/2$, we have $r = 1 + \frac{1+\varrho}{\varrho(1+p)}$ for all ϱ. In either case, since $\varrho \geq 1$ in our study, we conclude $r \leq 3$ or $r \leq 2$. This again supports the choice of $r = 2$ or 3.

4. Conclusions

We have discussed the problem of estimating σ_P^2 in a one way nested model (and to a lesser extent, estimating σ_D^2) where patients are nested within observers and replicates within patients. We consider 3 criteria for selecting the number of patients and replicates. In all cases for estimating σ_P^2, increasing the number of patients leads to better estimates than increasing replications. Except for the second criterion, where $r = 2$ seems best, the optimal number of replicates for our study here is three. The general strategy is to select small values of replicates for estimating σ_P^2.

We have omitted cost considerations in the design of this study because they are not important for our study. When costs become a major factor in the design of the study, such as when measuring radiographic outcomes to monitor progression of arthritis, the method described in BLOCH (1986) is directly applicable.

Acknowledgements

Supported in part by awards from United Scleroderma Foundation, Scleroderma Federation, FDA Orphan Drug Program FD-R-00863, Arthritis Foundation Clinical Science Grant Program, NIH UCLA Multipurpose Arthritis and Musculoskeletal Diseases Center grant P60 AR-36834. We wish to acknowledge the investigators in the High-Dose vs Low-Dose D-penicillamine in Systemic Scloerosis Trial who requested this design development and then employed the recommendations in their Trial.

References

BLOCH, D. A., 1986: Sample size requirements and the cost of a randomized clinical trial with repeated measurements. *Stat. Med* **5**, 663–667.

BURDICK, R. K. and GRAYBILL, F. A., 1992: *Confidence intervals on variance components.* Statistics. textbooks and monographs, Vol. **127**, Marcel Dekker Inc., New York.

CLEMENTS, P. J., LACHENBRUCH, P. A., SEIBOLD, J. R., ZEE, B., STEEN, V. D., BRENNAN, P., SILMAN, A., ALLEGAR, A., VARGA, J., MASSA, M., WIGLEY, F. M., INGENITO, F., WEISMAN, M., NEEDLEMAN, B., MARTIN, R. W., MCCLOSKEY, D., MORELAND, L. W., MAYES, M., LALLY, E. V., UNANUE, M., COLLIER, D. H., WEINER, S., WEINSTEIN, A., MEDSGER JR., T. A., ANDREWS, B., DIXON, M., FURST, D. E., 1993: Skin thickness score in systemic sclerosis (SSc): An assessment of inter-observer variability in three independent studies. *J. Rheumatol.* **20**, #11, 1892–1896.

CLEMENTS, P. J., LACHENBRUCH, P. A., SEIBOLD, J. R., WHITE, B., WEINER, S., MARTIN, R. W., WEINSTEIN, A., WEISMAN, M., MAYES, M., COLLIER, D., WIGLEY, F. M., MEDSGER, T. A., STEEN, V., MORELAND, L. W., DIXON, M., MASSA, M., LALLY, E. V., MCCLOSKEY, D. A., VARGA, J., INGENITO, F., and FURST, D. E., 1995: Inter and Intraobserver Variability of Total Skin Thickness Score (Modified Rodnan TSS) in Systemic Sclerosis. *J. Rheumatol.* **22**, 1281–1285.

DONNER, A. and ELIASZIW, M., (1987): Sample size requirements for reliability studies. *Statistics in Medicine* **6**, 441–448.

GIOVAGNOLI, A. and SEBASTIANI, P., 1989: Experimental designs for mean and variance estimation in variance Components models. *Computational Statistics and Data Analysis* **8**, 21–28.

GIOVAGNOLI, A. and SEBASTIANI, P., 1990: Correction to "Experimental designs for mean and variance estimation in variance Components models". *Computational Statistics and Data Analysis* **10**, 93.

GRAYBILL, F. A., MARTIN, F. and GODFREY, G., 1956: Confidence intervals for variance ratios specifying genetic heritability. *Biometrics* **12**, 99–109.

MONTGOMERY, D. C., 1991: *Designs and analysis of experiments*, 3rd Ed., John Wiley & Sons, New York.

PIGNATIELLO, J. J., 1987: Allocating observations in the random effects balanced one-way ANOVA. *Journal of Quality Technology* **19**, 216–228.

SCHEFFÉ, H., 1959: *The analysis of variance.* John Wiley & Sons, New York.

SEARLE, S. R., 1970: Large sample variance of MLE of variance components using unbalanced data. *Biometrics* **26**, 505–524.

Received, September 1995
Revised, July 1996
Accepted, July 1996

WENG KEE WONG
Department of Biostatistics
School of Public Health
10833 Le Conte Ave
Los Angeles, CA 90095-1772
USA