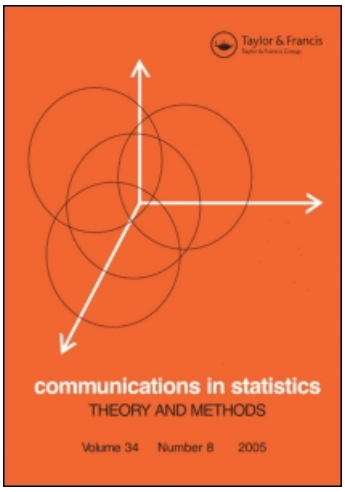

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597238

Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance

A. Lawrence Gould [a]; Weichung Joseph Shih [a]
[a] Biostatistics and Research Data Systems, Merck, Sharp, and Dohme Research Laboratories,

Online Publication Date: 01 January 1992

To cite this Article Gould, A. Lawrence and Shih, Weichung Joseph(1992)'Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance',Communications in Statistics - Theory and Methods,21:10,2833 — 2853
To link to this Article: DOI: 10.1080/03610929208830947
URL: http://dx.doi.org/10.1080/03610929208830947

COMMUN. STATIST.—THEORY METH., 21(10), 2833-2853 (1992)

SAMPLE SIZE RE-ESTIMATION WITHOUT UNBLINDING FOR NORMALLY DISTRIBUTED OUTCOMES WITH UNKNOWN VARIANCE

A. Lawrence Gould Weichung Joseph Shih

Biostatistics and Research Data Systems
Merck, Sharp, and Dohme Research Laboratories
West Point, PA 19486 Rahway, NJ 07065-914

Key Words and Phrases: *sample size adjustment; interim analysis; clinical trial; EM algorithm*

ABSTRACT

Monitoring clinical trials in nonfatal diseases where ethical considerations do not dictate early termination upon demonstration of efficacy often requires examining the interim findings to assure that the protocol-specified sample size will provide sufficient power against the null hypothesis when the alternative hypothesis is true. The sample size may be increased, if necessary to assure adequate power. This paper presents a new method for carrying out such interim power evaluations for observations from normal distributions without unblinding the treatment assignments or discernably affecting the Type 1 error rate. Simulation studies confirm the expected performance of the method.

1. INTRODUCTION

As a rule, group sequential methods (e.g., Pocock, 1977, 1982; O'Brien and Fleming, 1979; Gould and Pecore, 1982; Lan and DeMets, 1983, Geller and Pocock, 1987), allow early rejection (or, sometimes, acceptance) of the null hypothesis if warranted by the interim findings. These methods often are used in clinical trials in cancer, heart disease, and other life-threatening conditions where ethical considerations require terminating the trial if there is compelling early evidence of efficacy.

Double-blinded trials should remain so until completion if the null hypothesis will not be accepted or rejected at an interim stage, to prevent conscious or unconscious bias. However, the ability to check the assumptions made in determining the sample size without unblinding would be useful, to assure that the trial has adequate power. Gould (1992) described a means for doing so when the outcomes were binomially distributed. Normally distributed outcomes with unknown within-group variances require a different approach because estimating the within-group variances requires group mean information unneeded for binomially distributed outcomes.

Sample size readjustment for normally distributed data has been studied previously, most recently by Lohr (1988) and by Wittes and Brittain (1990). Lohr obtained the asymptotic properties of estimates of the mean and covariance matrix of a multivariate normal distribution when the sample size can be adjusted on the basis of one or two interim analyses of the data. Lohr's method is based on the usual corrected cross-product estimator of the sample covariance matrix, and so would require knowing the

individual group means in the hypothesis-testing situation considered here. Wittes and Brittain studied by simulation a procedure for adjusting the sample size in finite samples that also requires knowing the individual group means at the interim examination. The approach considered here does not require knowing the group means at the interim examination, and applies for finite samples from univariate normal distributions.

Section 2 below describes the method and how it affects the Type 1 error rate in finite samples. Section 3 discusses estimating σ^2, the common within-group variance. Section 4 provides the findings from a series of simulation studies that confirm the anticipated performance of the method. Section 6 addresses briefly several issues arising in its application.

2. METHOD

2.1 Description The method is analogous to Stein's (1945) method for obtaining a sample large enough to provide a specified-width confidence interval, but differs in (a) not requiring identification of the treatment assignments, and (b) using *all* of the information in the combined sample. Suppose that N observations are to be drawn, ΘN from a $\mathcal{N}(\mu_1, \sigma^2)$ distribution and $(1-\Theta)$N from a $\mathcal{N}(\mu_2, \sigma^2)$ distribution, σ^2 unknown, $0 < \Theta < 1$. For simplicity, assume $\Theta = 0.5$, although this is not essential. The null hypothesis H_0: $\mu_1 = \mu_2$ ordinarily would be tested against the alternative H_1: $\mu_1 \neq \mu_2$ using a Student t test. Given Type 1 and Type 2 error rates α and β, respectively, the total sample size would be determined from

$$N = 4\tilde{\sigma}^2(z_{\alpha/2} + z_{\beta})^2/(\mu_1 - \mu_2)^2 \qquad (1)$$

where $\mu_1 - \mu_2$ is determined by a specific alternative hypothesis H_1: $\mu_1 - \mu_2 = \Delta$ (a known value), $\tilde{\sigma}^2$ is an assumed value for σ^2, and z_γ is the value at which the standard normal cdf equals γ. If $\tilde{\sigma}^2$ underestimates σ^2, the actual likelihood of rejecting H_0 when H_1 is true will be less than the power specified for the trial.

Now suppose that the sample size will be reconsidered after n (< N) observations (e.g., $n \doteq N/2$) *without knowing the treatment assignments.* With a reasonable estimate, $\hat{\sigma}^2$, of the within-group variance, σ^2, one can determine via (1) the actual sample size

$$N' = N(\hat{\sigma}^2/\tilde{\sigma}^2) \tag{2}$$

needed to provide $100(1-\beta)\%$ power for rejecting the null hypothesis. If N′ is "sufficiently larger" than N, additional patients would be obtained to bring the final sample size up to N′; otherwise, the trial would be completed as planned. For example, requiring $N'/N > 1.25$ means that the sample will be increased only if the "correct" sample size is more than 25% larger than the original sample size. To keep the final sample size within reasonable limits, N′ might be limited to no more than some multiple of N (e.g., $N' \leq 2N$). The options when $N' > \omega N$ are discussed in Section 6.

2.2 Effect on Type 1 Error Rate Let the random variable $\bar{Z}_1$ denote the difference between the means of the initial samples based on a total of n_1 observations, and let the random variable $\bar{Z}_2$ denote the difference between the subsequent sample means, based on a total of n_2 observations. $\bar{Z}_1$ and $\bar{Z}_2$ both estimate $\delta = \mu_1 - \mu_2$; neither $\bar{Z}_1$ nor $\bar{Z}_2$ actually would be observed in practice because the group membership of the data

remains blinded. Suppose for simplicity that equal numbers of observations are drawn from each distribution. Combining the two samples yields

$$N = n_1 + n_2,\ m = m_1 + m_2,\ \bar{Z} = (n_1\bar{Z}_1 + n_2\bar{Z}_2)/N,$$

and

$$s^2 = (m_1 s_1^2 + m_2 s_2^2)/m$$

where s_i^2 denotes an estimator of σ^2 based on m_i degrees of freedom from the initial (i = 1) or subsequent (i = 2) sample. Assume that $m_i s_i^2/\sigma^2$ has a chi-square distribution with m_i degrees of freedom, at least approximately. Values for s_1^2 and s_2^2 are required in practice. The hypothesis H_0: $\delta = 0$ will be tested using the statistic $t = \sqrt{N}\,\bar{Z}/s$.

The probability of wrongly rejecting H_0 when n_2 does not depend on s_1^2 is provided by the integral of a central t density with m degrees of freedom over the set of values $|t| > t_c$, an appropriate critical value. The probability cannot be computed in this way when n_2 depends on s_1^2.

The joint density of the mean and sample variance from the initial sample is essentially the product of a normal and a chi-square density. *Conditional on* s_1^2, the same is true of the joint density of the mean and sample variance from the second sample. Consequently, the joint density of the statistics from both samples is the product of these densities. The joint density of $\bar{Z}$ and the sample variances can be written as

$$f(\bar{z}, s_1^2, s_2^2 ; n_1, \delta, \sigma^2)$$

$$\propto \left(\frac{m_1 s_1^2}{\sigma^2}\right)^{\frac{m_1}{2} - 1} \left(\frac{m_2 s_2^2}{\sigma^2}\right)^{\frac{m_2}{2} - 1} \exp\left\{-\frac{1}{2}\left(\frac{ms^2}{\sigma^2} + \frac{N\left(\bar{z} - \delta\right)^2}{2\sigma^2}\right)\right\}$$

Since n_1 and σ^2 are fixed quantities, this expression can be simplified with no loss of generality by the transforms $v_i = m_i s_i^2/\sigma^2$, $i = 1, 2$. With the additional transformation $\bar{Z} \to t = \sqrt{N}\,\bar{Z}/s$, the density becomes

$$f(t, v_1, v_2 ; m_1, m_2)$$
$$\propto v_1^{\frac{m_1}{2} - 1} v_2^{\frac{m_2}{2} - 1} (v_1 + v_2)^{\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(v_1 + v_2)(1 + t^2/m)\right\} \tag{3}$$

The probability of rejecting H_0 is the integral of (3):

$$\mathcal{P}(\text{Reject } H_0 \mid \delta) =$$
$$\int_{v_1=0}^{\infty} \int_{v_2=0}^{\infty} \left\{ \int_{t=-\infty}^{-t_c(v_1)} + \int_{t=t_c(v_1)}^{\infty} \right\} f(t, v_1, v_2 ; m_1, m_2)\, dt\, dv_2\, dv_1 \tag{4}$$

The quantity $t_c(v_1)$ depends on v_1 because the distribution of t and v_2 depends on n_2, which is determined by s_1^2 and, therefore, by v_1. Consequently, the order of integration in (4) cannot be interchanged, as the usual derivation of the Student t density would require.

To illustrate the effect of the dependence, suppose that n_2 depends on v_1 in the following way: $v_1 \le v_1^* \Rightarrow n_2 = n_{21}$; $v_1 > v_1^* \Rightarrow n_2 = n_{22}$. With the transformation $v_1, v_2 \to v\ (= v_1 + v_2)$, $w\ (= v_1/v)$, (4) can be written as

$$\mathcal{P}(\text{Reject } H_0 \mid \delta = 0) = 2 \int_{t=-\infty}^{-t_c^{(1)}} \int_{v_1=0}^{\infty} \int_{v_2=0}^{\infty} f(t, v_1, v_2 ;\ m_1, m_2)\, dv_2\, dv_1 dt$$
$$- \int_{v=v_1^*}^{\infty} 2\,\Phi\left(-t_c^{(1)}\sqrt{v/m^{(1)}}\right) I_{1-v_1^*/v}\left(\tfrac{m_{21}}{2}, \tfrac{m_1}{2}\right) v^{\frac{m^{(1)}}{2} - 1} f_{\chi^2}(v ; m^{(1)})\, dv \tag{5}$$

$$+ \int_{v=v_1^*}^{\infty} 2\,\Phi\left(-t_c^{(2)}\sqrt{v/m^{(2)}}\right) I_{1-v_1^*/v}\left(\frac{m_{22}}{2}, \frac{m_1}{2}\right) f_{\chi^2}(v\,;\, m^{(2)})\, dv$$

where $I_X(\cdot,\cdot)$ denotes the usual incomplete Beta function, $f_{\chi^2}(\cdot;\, m)$ denotes a central chi-square density with m degrees of freedom, $\Phi(\cdot)$ denotes the standard normal cdf, $t_c^{(1)}$ denotes the critical value for a central t distribution with $m^{(1)} = m_1 + m_{21}$ degrees of freedom, $m_{21} = n_{21} - 2$, etc. The first integral in (5) is α, the nominal Type 1 error rate. The remaining terms of (5) represent the perturbation of the Type 1 error rate due to the sequential sampling scheme. These latter two terms cancel if $n_{21} = n_{22}$.

The magnitude of the perturbation can be calculated easily. Thus, suppose that $n_1 = 20$, so that $m_1 = 18$. This is not a large initial sample. At the interim stage, decide to obtain $n_2 = 20$ more observations (10 from each group) if $s_1^2 \leq 1.5$, or $n_2 = 40$ more observations if $s_1^2 > 1.5$. Suppose that the test is to be at a nominal 5% level, so that the critical t value would be $t_c = 2.03$ ($n_2 = 20$) or 2.00 ($n_2 = 40$). Assume that $\sigma = 1$. Then the lower integration limit in (5) is $v_1^* = m_1 s_1^2/\sigma^2 = 18 \times 1.5/1 = 27$. Figure 1 plots the values of the algebraic sum of the second and third terms of (5). The net value of this sum is -0.0002, which represents the negligible difference between the true and nominal Type 1 error rates in this example. The simulation findings presented below also support the assertion that this approach has a negligible effect on the Type 1 error rate.

The sample size re-estimation approach described here does not rule out the possibility that the interim estimate of σ^2 might be small enough so that no further observations would be required to assure the desired power,

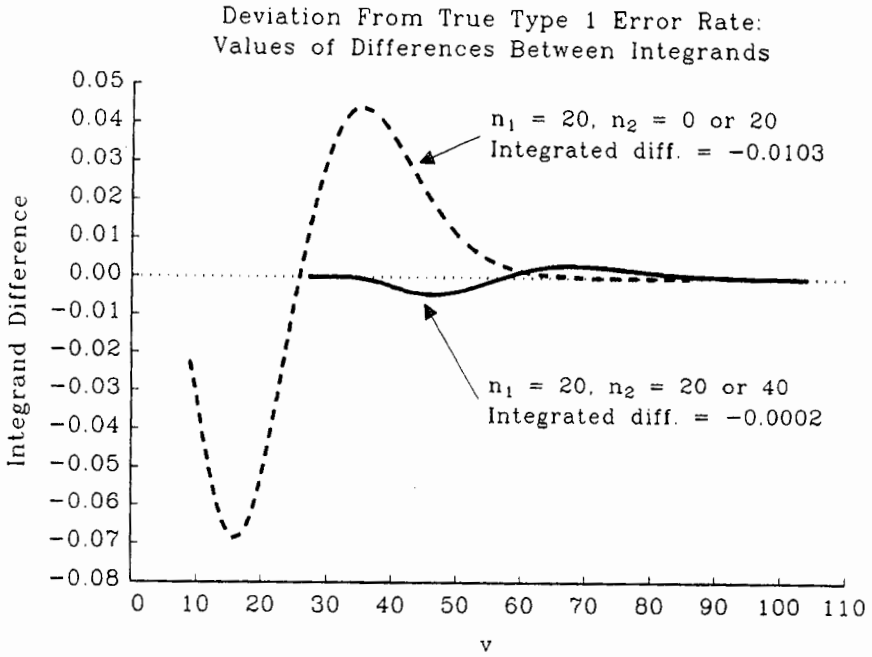

Figure 1. Deviation from True Type 1 Error Rate: Values of Differences Between Integrands

i.e., that $n_2 = 0$. Essentially the same argument used to obtain (5) establishes the following result:

$$\mathcal{P}(\text{Reject } H_0 \mid \delta = 0) - \alpha$$

$$= \int_{v=v_1^*}^{\infty} \left\{2\,\Phi\left(t_c^{(2)}\sqrt{v/m}\right) - 1\right\} I_{1-v_1^*/v}\left(\frac{m_2}{2}, \frac{m_1}{2}\right) f_{\chi^2}(v\,;\,m)\,dv \tag{6}$$

$$- \int_{v=v_1^*}^{\infty} \left\{2\,\Phi\left(t_c^{(1)}\sqrt{v/m_1}\right) - 1\right\} f_{\chi^2}(v\,;\,m_1)\,dv$$

Here, $t_c^{(1)}$ refers to the critical value for a t distribution with $m_1 = n_1-2$ degrees of freedom and $t_c^{(2)}$ refers to a t distribution with $m = m_1+m_2$ d.f.

To illustrate the effect of possible early termination on the Type 1 error rate, suppose that $n_1 = 20$. At the interim stage, obtain $n_2 = 20$ more observations if $s_1^2 \geq 0.5$, or call the trial complete if $s_1^2 < 0.5$. For a nominal 5% level test, the critical t value would be $t_c = 2.10$ ($n_2 = 0$) or 2.03 ($n_2 = 20$). If $\sigma = 1$ then the lower integration limit in (6) is $v_1^* = m_1 s_1^{*2}/\sigma^2 = 18 \times 0.5/1 = 9$. Figure 1 also displays the results of the calculations for this case. Even with the small sample size (10 or 20 observations per group), the RHS of (6) is -0.01, a small and conservative effect on the Type 1 error rate.

3. ESTIMATING σ^2

If the treatment assignments were known, $\hat{\sigma}^2$ could be computed by pooling the within-group sample variances. Since the assignments are unknown, σ^2 must be estimated some other way. We consider two ways to estimate σ^2 : a simple adjustment of the pooled sample variance based on the difference between the means presumed by H_1; and the EM algorithm, which does not depend on H_1.

3.1 Simple adjustment Suppose the interim sample contains θn observations from group 1 and $(1-\theta)$n observations from group 2; n is known, θ is unknown. Let x_{ij} denote the j-th observation from group i. The overall estimate of σ^2 based on the pooled sample can be computed without unblinding and written formally as

$$(n-1)s^2 = \sum_{i,j}(x_{ij} - \bar{x})^2 = \sum_{i,j}(x_{ij} - \bar{x}_i)^2 + n\theta(1-\theta)\,(\bar{x}_1 - \bar{x}_2)^2$$

$$= (n-2)\,\hat{\sigma}^2 + n\theta(1-\theta)\,(\bar{x}_1 - \bar{x}_2)^2$$

where $\hat{\sigma}^2$ denotes the unknown within-group estimate of σ^2. Since the interim sample is blinded, θ and the group sample means $\bar{x}_1$, $\bar{x}_2$ will be unknown, as will both terms of this last expression. However, *if the alternative hypothesis* H_1: $\mu_1 - \mu_2 = \Delta$ *is true* and if n is large enough so that $\bar{x}_1 - \bar{x}_2$ is reasonably close to Δ, then

$$n\theta(1-\theta)\,(\bar{x}_1 - \bar{x}_2)^2 \approx \Theta(1-\Theta)(n-1)\Delta^2,$$

so that if $\Theta = 0.5$,

$$\hat{\sigma}^2 \approx \frac{n-1}{n-2}\,(s^2 - \Delta^2/4). \tag{7}$$

When a blocked randomization scheme is used to assign subjects to treatments, θ will be very nearly known and very close to Θ. This will be true especially if the block size is 1× or 2× the number of treatments. The effect will be to improve the approximation immediately preceding (7).

3.2 EM Algorithm Since the treatment identifications are unknown, any of the interim observations x_i, $i = 1, \ldots, n$ could be in either treatment group, so that the treatment assignments are "missing at random" (Rubin, 1976). Let τ_i denote the treatment group membership indicator:

$\tau_i = 1$ (0) if sample member i is in treatment group 1 (group 2)

$\tau_1, \ldots, \tau_n$ are independent random variables with $\mathcal{P}(\tau_i = 1) = \theta$. Given τ_i, x_i $(i = 1, \ldots, n)$ has a normal distribution with density

$$f(x_i \mid \tau_i, \mu_1, \mu_2, \sigma) \propto \tfrac{1}{\sigma} \exp\{-\frac{1}{2\sigma^2}[\tau_i(x_i - \mu_1)^2 + (1 - \tau_i)(x_i - \mu_2)^2]\} \tag{8}$$

TABLE 1

Accuracy of EM algorithm estimate of sigma (100 iterations per case)

25 obs/gp	|True Mean Difference| / True σ							
	0		0.5		1		2	
True σ	Mean	S. D.	Mean	S. D.	Mean	S.D.	Mean	S.D.
0.5	0.479	0.051	0.474	0.061	0.505	0.088	0.535	0.130
1.0	0.964	0.099	1.112	1.294	1.032	0.115	1.104	0.259
2.0	1.932	0.221	1.947	0.206	2.014	0.316	2.224	0.492
4.0	3.902	0.491	3.849	0.467	3.972	0.647	4.324	1.094

50 obs/gp	|True Mean Difference| / True σ							
	0		0.5		1		2	
True σ	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
0.5	0.481	0.035	0.494	0.038	0.511	0.041	0.576	0.091
1.0	0.991	0.074	0.971	0.070	1.046	0.081	1.160	0.194
2.0	1.974	0.138	1.948	0.140	2.054	0.158	2.356	0.276
4.0	3.871	0.293	3.930	0.277	4.161	0.385	4.731	0.650

Notes: (1) Each recursive computation of $\hat{\sigma}$ continued until convergence (successive estimates differing by 0.01 or less) or until 50 cycles had been reached, whichever came first.

(2) The tabulated quantities are the estimated values of σ and the corresponding standard deviations among the 100 repetitions of each case.

The expression for the conditional probability (or expectation) of τ_i given x_i therefore is

$$\mathcal{P}(\tau_i = 1 \mid x_i) = \mathcal{E}(\tau_i \mid x_i)$$

$$= 1/\{1 + \tfrac{1-\theta}{\theta}\exp[\,(\mu_1-\mu_2)(\mu_1+\mu_2 - 2x_i)/2\sigma^2]\,\} \qquad (9)$$

The log likelihood of the interim observations follows from (8),

$$\mathcal{L} = (n/2) \log \sigma^2 + \{\sum_{i=1}^{n} [\tau_i(x_i - \mu_1)^2 + (1 - \tau_i)(x_i - \mu_2)^2]\}/2\sigma^2 \quad (10)$$

The EM algorithm (Dempster, Laird, and Rubin, 1977) for estimating σ proceeds as follows. Assume $\theta = \Theta$. The "E" step consists of substituting "current" estimates of μ_1, μ_2, and σ into (9) to obtain provisional values for the expectations of the τ_i. The "M" consists of obtaining maximum likelihood estimates of μ_1, μ_2, and σ^2 after replacing the τ_i in (10) with their provisional expectations. The "E" and "M" steps are repeated until the value of σ stabilizes; the resulting value is the estimate, $\hat{\sigma}^2$ of σ^2 required in (2). Table 1 provides the results of a small simulation study investigating the performance of this algorithm. Although σ^2 was estimated accurately, $(\mu_1 - \mu_2)/\sigma$ was not estimated well. The averages over the iterations of the values of $(\hat{\mu}_1 - \hat{\mu}_2)/\hat{\sigma}$, based on maximum likelihood estimators, ranged from 0.3 to 0.5 in 29 of the 32 cases shown in Table 1, in no particular pattern; the exceptional values were 0.6, 0.7, and 0.8. This is consistent with Fowlkes's (1979) assertion that the accuracy of the estimates of μ_1 and μ_2 cannot be assured due to their sensitivity to the starting values (Fowlkes, 1979).

3.3 Initial values for EM algorithm We adapt a suggestion of Fowlkes (1979) for finding initial parameter estimates for the EM algorithm. Let $z_{(1)} < z_{(2)} < \ldots < z_{(n)}$ denote the ordered data at the interim evaluation. Let $p_i = (i - 0.5)/n$ for i=1, ..., n and calculate $q_i = \Phi^{-1}(p_i)$, where Φ^{-1} denotes the inverse of the standard normal distribution function. Fit a simple linear regression by least squares to the points $\{(q_i, z_{(i)}), i=1, \ldots, n\}$; let b denote the slope of the fitted

line, and let a denote its intercept:

$$b = \frac{\sum q_i z_{(i)} - n\bar{q}\bar{z}}{\sum q_i^2 - n\bar{q}^2}, \quad a = \bar{z} - b\bar{q}.$$

The initial values of σ, μ_1, and μ_2 then are

$$\hat{\sigma}_0 = b, \quad \hat{\mu}_{1,0} = a - b/c, \quad \hat{\mu}_{2,0} = a + b/c$$

where c is some chosen constant. The choice of c influences the estimation of the means, but not the variance. Ideally, we would like c $= 2\sigma/(\mu_2-\mu_1)$; however, although b estimates σ, there is no good estimate of $(\mu_2-\mu_1)$. We get around this problem in the following way. In most clinical trials that use a normal approximation for estimating the sample size, the inverse of the coefficient of variation $\lambda = (\mu_1-\mu_2)/\sigma$ usually ranges between 0.20 and 0.50 (which correspond to about 430 and 70 patients per group, respectively, for power = .90, one-sided α = 0.05). We suggest taking the middle value in this range, 0.35, and converting it to c = 2 × (1/0.35) = 5.71.

4. SIMULATION STUDIES

4.1 Design Simulation studies explored the behavior of the procedure over a range of parameter values likely to occur in practice. The values of sigma assumed by the design ($\tilde{\sigma}$) and the true value of sigma (σ) were set at 0.707, 1, 1.414, 2, 2.828, and 4. All combinations of σ and $\tilde{\sigma}$ values were considered. The design always assumed $\Delta = 1$, and the sample size was selected to provide 90% power for rejecting the null

hypothesis when the alternative was true. Equal samples were taken from each distribution ($\Theta = 0.5$). For the simulation, the true mean differences were set at 0 (null hypothesis true), 0.5, 1, and 2. The effects of evaluating the sample size after obtaining 25% and 50% of the initially planned data were considered, as were the effects of two rules for deciding to increase the sample size (increase if $N'/N > 1.33$ or 1.05). In all cases, $N' \leq 2N$, reflecting a practical limitation on increasing the size of ongoing studies. The effect of the algorithm used to estimate σ (simple or EM) also was evaluated. In all, 864 cases (36 combinations of σ and $\tilde{\sigma}$, 3 nonzero true mean difference values, 2 examination time values, 2 values of sample size increase rule, 2 algorithms) were run. Each case included a test with a zero mean difference and a nonzero true mean difference, so there were 864 tests of the null hypothesis when it was true. Each case was replicated 1000 times, and statistics were collected about the number of rejections of the null hypothesis when it was true and when it was false, and the distributions of the final sample size under either hypothesis.

4.2 Results The probability of rejecting H_0 when it was true did not depend materially on any of the factors defining the cases, because none of the coefficients differed significantly from 0 in a logistic regression relating the probability of wrongly rejecting H_0 to these factors for each algorithm. Therefore, the 864 rejection frequency values should be distributed like Binomial variates with $n = 1000$ and $p = 0.05$. Figure 2 displays the distributions of the rejection frequencies for the two algorithms. The results agree closely with expectation.

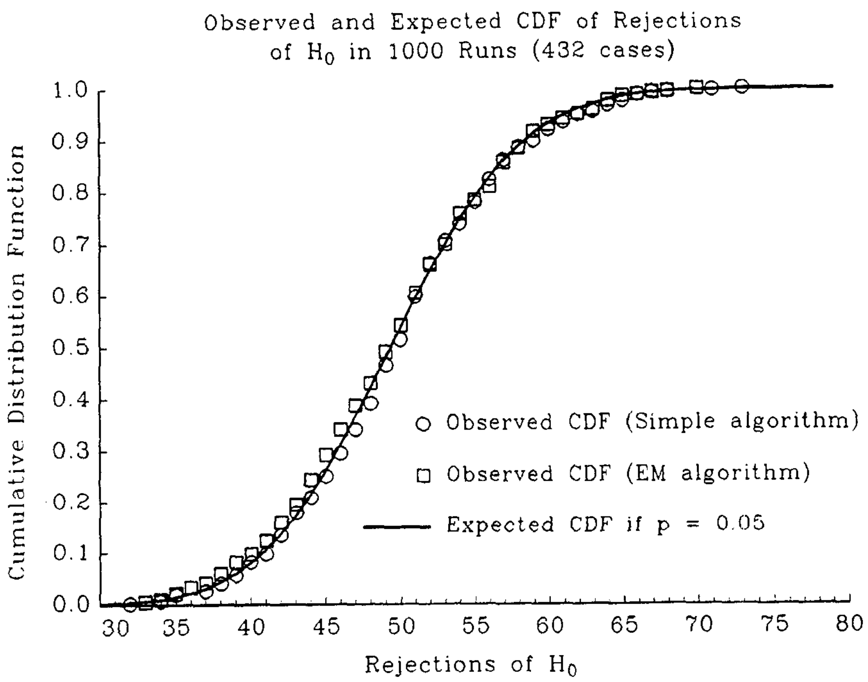

Figure 2. Observed and Expected CDF of Rejections of H_0 in 1000 Runs (432 cases for each way of estimating σ^2)

Figure 3 displays the effects of correctly and incorrectly specifying the true mean difference and the true variance on the likelihood of rejecting H_0 when $\Delta \neq 0$. The two algorithms for estimating σ behaved essentially identically. This probably reflects the range of Δ/σ values used in the simulations (which covers most of the situations in clinical trials that use a normal approximation for sample size calculations). Overspecifying the true mean difference or underspecifying the true variance caused a loss in power, as expected. However, when the true mean difference and variance were correctly

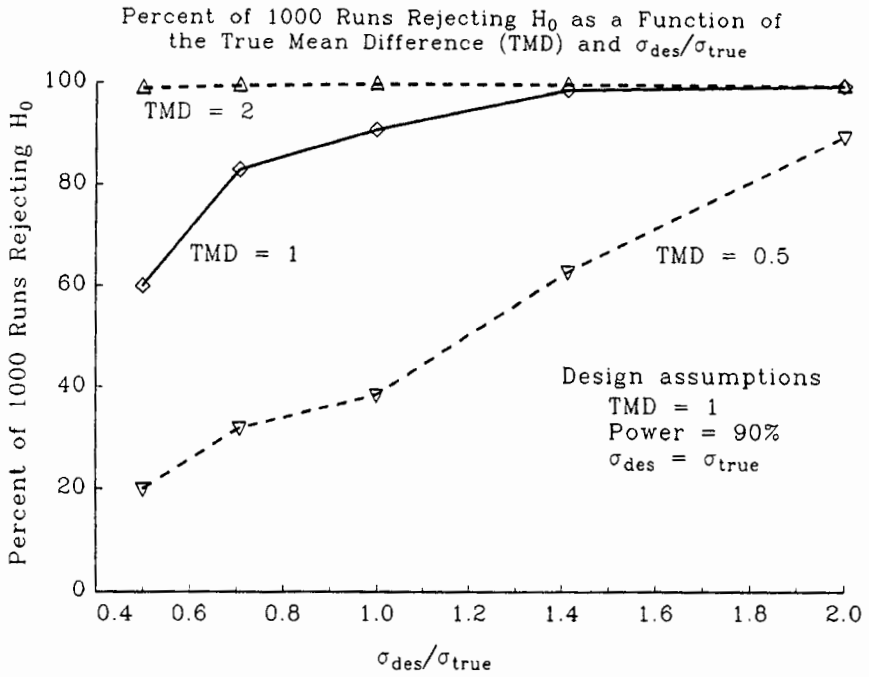

Figure 3. Percent of 1000 Runs Rejecting H_0 as a Function of the True Mean Difference (TMD) and $\sigma_{des}/\sigma_{true}$

specified, the power was very close to the assumed value of 90%, usually exceeding it slightly. Since the EM procedure does not depend on $\tilde{\sigma}$, the value assumed for σ in calculating the sample size, the loss of power when $\tilde{\sigma}$ ($= \sigma_{des}$ in Fig. 1) is less than σ_{true} actually was due to requiring that $N' \leq 2N$.

5. EXAMPLE

Suppose that a difference $\Delta = 0.30$ is to be detected with 90% power using a 1-sided 5% level test ($\alpha = .05$). A design taking $\tilde{\sigma} = 1.5$

would require 430 patients per group; a design with $\tilde{\sigma} = 0.80$ would require 120 patients in each group. If the (unknown) true value of σ actually were 1, then the trial should contain 190 patients per group. In practice an interim examination might be carried out after observing 100 patients, 50 from each group, and might suggest that the final sample should contain 200 patients in each group. If the trial had been designed with $\tilde{\sigma} = 0.80$, this would mean that 160 more patients than planned needed to be entered into the trial and assigned at random to the two groups. If the trial had been designed with $\tilde{\sigma} = 1.5$, no further patients beyond those planned would need to be recruited for the trial.

6. DISCUSSION

The method described here does not estimate reliably the true difference between the treatment means (Fowlkes, 1979), and so does not provide a way to ascertain the actual magnitude of $\mu_1 - \mu_2$. The average and median "mean difference/σ" values estimated from the 100 repetitions of each case summarized in Table 1 did not depend materially on the true "mean difference/σ" values.

The statistical power specified at the planning stage and checked at the interim stage corresponds to a fixed alternative hypothesis that the true mean difference equals Δ, a quantity specified by the researcher. In the context of a clinical trial, Δ would be the least clinically meaningful difference worth detecting, identified a priori. The method provides a given level of assurance for detecting a specified

difference if it is present. It is not designed to enhance the likelihood of detecting the difference that appears to be present (which cannot be estimated).

The method ordinarily needs to applied only once, when enough data are available to provide a reasonably reliable estimate of σ^2. Table 1 suggests that as few as 25 observations per group should suffice. From (2), N' is a random variable with a heavy tail to the right; when the assumed and true σ values happen to be close, then overly large N' values become disproportionately more likely with smaller values of N. Thus, an interim look with fewer than 25 observations per group may lead to too large a final sample size. The procedure does not have to be repeated after obtaining a reliable estimate of σ^2 because the estimate and, therefore, the sample size, will not change materially with further looks. Moreover, adding new patients to a multicenter clinical trial brings up many administrative issues, e.g., changes in contracts, funding, perhaps number of centers, etc. The fewer of these that have to be made, and the earlier, the better.

When $N' > \omega N$, there are two options. The trial may be terminated immediately and its results summarized without testing the hypotheses. Such a trial would be regarded as uninformative about the hypotheses, and reexamination of the assumptions about the variability of the responses or the relevance of the target population would be appropriate. Alternatively, the trial could be continued to completion with the additional observations, accepting the possibility that the

actual power may be less than desired. Less power does not mean zero power, so rejection of the null hypothesis still could occur on completion of the trial.

The reestimated sample size could turn out to be much smaller than the planned size (e.g., 180 vs. 430 patients per group as in Section 5), suggesting that the trial could be terminated after obtaining the initial observations. This is unlikely to affect the Type 1 error rate materially, as shown in section 2.2. However, unless ethical considerations dictate otherwise, the trial should not be terminated because demonstrating efficacy with respect to a single variable seldom is the only objective of a trial.

The EM algorithm always reasonably estimates σ, regardless of the true and assumed values of Δ and σ. This certainly is useful for designing additional trials in the same indication before completing the current trial. More importantly, however, the value of N' provided by (2) is the value likely to provide the required power for rejecting H_0 *in favor of the specified alternative.* This is not necessarily true for the simple method. The simple estimate of σ assumes a value for Δ and, from (7), may understate or overstate the true value of σ depending on whether this assumed value overestimates or underestimates the true value. Overestimating the true value of Δ causes underestimation of σ, so N' is insufficient to provide the required power against H_0 in favor of the specified alternative. This guards against an inflated sample size when H_0 is true, but the power loss may be excessive when the true

value of Δ is only a little less than the value set by H_1. The converse is true when the assumed value of Δ exceeds the true value, so that the simple method has the undesirable property of moving the sample size away from clinical reality (Spiegelhalter, Freeman, and Blackburn, 1986).

BIBLIOGRAPHY

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society,* B **39**, 1-38.

Fowlkes, E.B. (1979). Some methods for studying the mixture of two normal (lognormal) distributions. *Journal of the American Statistical Association* **74**, 561 - 575.

Geller, N. L. & Pocock, S. J. (1987). Interim analyses in randomized clinical trials: Ramifications and guidelines for practitioners. *Biometrics* **43**, 213-223.

Gould, A. L. (1992). Interim analyses for monitoring clinical trials that do not affect the type I error rate. *Statistics in Medicine* **11**, 55-66.

Gould, A.L. and Pecore, V.J. (1982). Group sequential methods for clinical trials allowing early acceptance of H_0 and incorporating costs. *Biometrika* **69**, 75-80.

Lan, K.K.G. and DeMets, D.L. (1983). Design and analysis of group sequential tests based on the Type 1 error spending rate function. *Biometrika* **74**, 149-154.

Lohr, S. L. (1988). Accurate multivariate estimation using double and triple sampling. University of Minnesota Technical Report No. 505, February 1988.

O'Brien, P. C. & Fleming, T. R. (1979). A mutiple testing procedure for clinical trials. *Biometrics* **35**, 549-556.

Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191-199.

Pocock, S. J. (1982). Interim analyses for randomized clinical trials: The group sequential approach. *Biometrics* 38, 153-162.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581-592.

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics* **16**, 243-258.

Wittes, J. and Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine* **9**, 65-72.

Received November 1991; Revised May 1992