DOKTORI (PhD) ÉRTEKEZÉS

SINGER JÚLIA

STATISZTIKAI PRÓBÁKHOZ SZÜKSÉGES MINTAELEMSZÁMOK BECSLÉSE: MATEMATIKAI ÉS SZÁMÍTÓGÉPES MÓDSZEREK BIOLÓGIAI ÉS KLINIKAI ALKALMAZÁSOKKAL

SAMPLE SIZE ESTIMATION FOR DIFFERENT STATISTICAL TESTS: MATHEMATICAL AND COMPUTERISED METHODS WITH BIOLOGICAL AND CLINICAL APPLICATIONS

Témavezető:
Dr. Győri István
tanszékvezető egyetemi tanár

Veszprémi Egyetem, Matematikai és Számítástechnikai Tanszék

Veszprém - 2001

Contents

1. Background

Reading the vaste literature of power and sample size analysis techniques in biostatistics one might have the impression that this issue is overemphasized. The feeling is strengthened by the phenomenon that if a certain number of trials were run in the same indication or if a certain number of discovery studies were performed based on a common pharmacological methodology then any trialist knows what is the recommended size of the trial. Yet the biostatistician (influenced by the existing harmonized guidelines and the internal position papers of each company dealing with trial and experiment planning) usually cannot assume the responsibility of proposing a sample size without any statistical argumentation, based only on the "common sense".

However, the increasing number of publications about power and sample size analysis does not necessarily mean a change in the conception itself. It's true that the continuous diversification of statistical methodologies implies a parallel diversification process of the sample size estimation formulae (since every widely used statistical method should have a corresponding sample size estimation procedure), but the main need is now a kind of summarizing activity leading to the change of the basic notions and preconceptions.

In fact, our view about the sample size itself has to be changed. The sample size needed to achieve a certain power varies in fact randomly, since its calculation is based on some observed values of random variables (usually the sample mean and the variance). Thus it is preferable to assess a whole distribution of the sample size instead of considering it a single number.

The confidence interval approach is now generally accepted and recommended by all the guidelines as an added information to the point estimate of each parameter. The same recommendation is seldom met for power or sample size assessment in spite of the fact that the necessary techniques are already existing for a long time. And even if in some cases no closed form expressions exist for the confidence limits, the computer-intensive methods serving this purpose are now widely available.

The two type of errors (type I and type II), the so-called “primitive inputs” of the sample size analysis may also be reconsidered as perhaps they are less “primitive” than it is usually thought and their choice does not need to be automatized. The levels of 5% for the type I and 20% for the type II error are rarely changed and even if they are, the reason is usually a retrospective justification of the initially fixed sample size rather than a conscious process of choice. The choice of these levels should depend on the aim and type of the study (e.g. to prove efficacy one needs lower type I errors than in safety studies, in a screening experiment it is more important to not lose any efficient compound, in early phases of preclinical studies usually the type II error is important, while the type I error may be higher since the experiments giving “significant” results are repeated anyway).

And finally: a new interpretation of the effect size might come. The anticipated value of the effect size is usually rather carefully chosen because of the belief that a well chosen effect size may assure the success of the trial. But sometimes the effect size might change in time (as the overall health condition of the population and the incidence of the different diseases is changing). This change is usually neglictable for short trials but it might be considerable for longer ones, involving the concept of effect size as a function of the time.

The main purpose of this study is to present a strategy for sample size estimation that subsumes these new principles.

In what follows, a frequentist approach is applied. A broad range of methods subsumed by the classical normal-theory models are treated, taking the unbalanced designs to be the norm rather than the exception. To characterize the precision of a sample size estimation method belonging to a given statistical procedure the term "relative discrepancy" was introduced (defined as the relative departure between the obtained and the nominal power of the test compared to the standard error of the actual power).

Reviewing the studies in biology, psychology, medicine and other fields relying on statistical inference one can conclude that many of them are too small (in respect to their size) to ensure enough statistical power to confirm meaningful effects [1], [2]. This might be partly due to the fact that the homogeneity of variances is usually assumed, ignoring the principle that the

sample size calculation method and the statistical test used later must rely on the same hypotheses about the population distribution. Chapter 2. presents a generalisation (made by the author of this study [3]) for unequal variances of the usual t-test method of sample size estimation, using Satterthwaite's correction. The relative discrepancy between the old and the new methodology is computed.

Chapter 3 contains an extension of the "allocation ratio" (the proportion of two group sizes) to more than two groups by introducing the term "set of allocation ratios" and presents a sample size calculation method (elaborated by the author of the present study [4]) for more than two parallel groups with unequal sizes.

When no closed-form expressions or approximation methods exist, Monte Carlo simulation techniques are used. The principles of such a simulation are worked out in Chapter 4.

In Chapter 5 the sample size is not regarded any more as a constant but a random variable the distribution of which is estimated. Knowing the distribution of the sample size enables the computation of the sample size with a certain confidence. The most common confidence interval methods of the sample size (for given power) are presented. These methods are then compared in terms of the coverage rates and mean relative widths. Examples of power and sample size calculations are also presented for the case when the effect size depends on time.

In Chapter 6 a SAS-application (written by the author of this study and presented at the SAS Conference of the European Users [5]) is described which treats 14 different simple models of sample size estimation (for a given power). Its primary aim is to visualize the different methods, the dependence of the parameters on each other by using simple graphical objects (in the SAS-terminology they are called widgets) as sliders, list boxes, check boxes, radio boxes, icons, help-entries, etc.

The Appendix contains the source code of the most important SAS programs, a short description about the SAS functions used by them and some screens of the SAS application.

2. Two independent groups, population variances are unequal

The usual methods of sample size determination assume that within-group variances are the same for each group to be compared. However, there are some situations when this assumption fails to be true (even if the tests for homogeneity of variances do not show significance). It happens rather frequently that variances tend to change with the changes of the mean. One of the possible solutions is to find a variance-stabilising transformation and to apply the usual sample size formulas to the transformed data. This method has the disadvantage of requiring previous estimates of means and variances for the transformed data. The other solution is to find a generalisation of the old sample size formulae which is valid for unequal variances. Satterthwaite's correction for a linear combination of independent variances [7] enables such a generalisation.

There are more methods to plan the sample size [8],[9] in case of the comparison of two means and equal within-group variances. Each of them uses normal deviates, with some correction applied when the t-test replaces the normal approximation. One such formula which doesn't need any iteration is given by Machin et al [9] :

$$n = \frac{2\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \sigma^2}{\left|\mu_1 - \mu_2\right|^2} + \frac{z_{1-\alpha/2}^2}{4} \qquad (1)$$

where n is the group size (supposing equal number of subjects in the two groups), α is the significance level, $1-\beta$ is the desired power, the anticipated population means are μ_1 and μ_2, z_p denotes the p^{th} percentile of the standard normal distribution, while the population variances are equal to σ^2.

For unequal group sizes with an allocation ratio of $A=n_1/n_2$, group sizes are given by:

$n_1 = \frac{A+1}{2} n$ and $n_2 = \frac{A+1}{2A} n$, where n is the sample size computed for equal-sized groups.

Guenther [10] derived a formula for the two-sample t-test with pooled variance estimate when equal group sizes are assumed. This approximate method was generalised by Schouten [11] for unequal group sizes. Schouten's recent paper gives the following estimation:

$$n_1 \geq \left(z_{1-\alpha/2} + z_\beta\right)^2 . \frac{(\tau+\gamma)\sigma_1^{\,2}}{\gamma\left(\mu_1-\mu_2\right)^2} + \frac{\left(\tau^2+\gamma^3\right).z_{1-\alpha/2}^2}{2\gamma\left(\tau+\gamma\right)^2} \quad \text{and} \quad n_2 \geq \gamma . n_1$$

where σ_1^2, σ_2^2 are the variances of the two groups, $\tau = \sigma_2^2/\sigma_1^2$ is the proportion of the variances and $\gamma = n_2/n_1$ is the allocation ratio.

Another possibility (giving almost the same result as (1) except some rounding differences) is to use Student's t-values in an iterative way, that is, to increase *n* until the

$$n \geq \frac{2\left(t_{df,1-\alpha/2} + t_{df,1-\beta}\right)^2 \sigma^2}{\left|\mu_1-\mu_2\right|^2}$$ [where *df=2(n-1)* for equal within-group variances]

inequality becomes true.

The least value of *n* which satisfies the above inequality is the sample size/group. The use of this method has the advantage that it can be easily extended to the case of unequal variances.

According to Satterthwaite's formula, the variance of the difference between means is estimated by:

$$s^2 = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2},$$

and it follows a χ^2-distribution with the degree of freedom:

$$df = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$$

where s_1^2 and s_2^2 are the estimated group variances and n_1, n_2 are the group sizes.

As Singer showed [3] it can be noted that *df* depends in fact only on 3 variables (θ, A and n) and it can be expressed in terms of $\theta = s_1^2/s_2^2$, $A = n_1/n_2$ and $n = n_2$ as:

$$df = \frac{(\theta + A)^2}{\theta^2/(An-1) + A^2/(n-1)}$$

With the above notations we have $s^2 = \frac{1}{n}\left(\frac{s_1^2}{A} + s_2^2\right)$.

For $A=1$ and $\theta=1$, $df=2(n-1)$ is reobtained.

Using the above s^2 and df, the sample size can be computed by increasing n until the inequality

$$n \geq \frac{\left(t_{df,1-\alpha/2} + t_{df,1-\beta}\right)^2 .\left(s_1^2 / A + s_2^2\right)}{\left|\mu_1 - \mu_2\right|^2} \qquad (2)$$

becomes true. The least value of n which fulfils the above criterion is the size of the second group, n_2, while the size of the first group is $n_1=n.A$.

Differences between the results obtained assuming equal variances (using the normal approximation method (1) and the pooled variance) and those obtained by assuming unequal ones (and using the iterative method (2) with Satterthwaite's correction) were evaluated for different values of θ and A. Sample sizes computed with the two methods (keeping constant $\alpha=0.05$, $1-\beta=0.8$, $|\mu_1-\mu_2|=10$ and $s_2=10$) are contained by Table 2.1. For the assumptions of Table I the sample sizes computed with Schouten's approximation [11] were exactly the same as those computed using the iterative method.

The sample sizes were then tested on simulated data. Normally distributed data were generated with group sizes n_1 and n_2, variances $\theta.s_2^2$ and s_2^2 and $|\mu_1-\mu_2|=10$ to test for the power, and with $|\mu_1-\mu_2|=0$ to test for the significance. The two groups were compared by a t-test with and without Satterthwaite's correction. 1000 datasets were generated for each power and significance estimation.

Table 2.1

Group sizes for different values of θ and A *. The simulated power and the significance of the t-tests (with and without Satterthwaite's correction).

Var. rat. (θ)	Alloc rat. (A)	Assuming $s_1=s_2$ (Method (1))		t-test assuming equal variances		t-test with Satterthwaite's correction		Assuming $s_1 \neq s_2$ (Method (2))		t-test assuming equal variances		t-test with Satterthwaite's correction	
		n_1	n_2	Power	Sign.	Power	Sign.	n_1	N_2	Power	Sign.	Power	Sign.
1/3	1/3	8	24	77.8	1.6	91.6	5.2	6	18	64.3	0.7	78.6	4.3
	1/2	9	18	79.0	3.0	88.3	4.5	8	16	75.2	1.9	83.9	4.0
	1	11	11	80.0	4.5	75.5	4.2	12	12	83.5	4.7	80.1	4.2
	2	18	9	85.2	11.2	69.5	5.4	22	11	91.0	8.7	83.7	4.7
	3	24	8	88.1	12.1	66.6	5.0	33	11	94.9	12.9	81.5	5.8
1/2	1/3	9	27	80.6	2.4	85.5	5.8	8	24	73.8	2.0	83.0	5.3
	1/2	10	20	80.7	3.6	85.7	4.6	9	18	75.5	3.3	80.9	4.4
	1	13	13	79.5	4.8	78.9	4.7	13	13	79.5	4.8	78.9	4.7
	2	20	10	83.8	6.7	77.7	4.2	24	12	90.5	7.6	82.1	4.6
	3	27	9	84.6	11.5	71.7	6.1	33	11	92.9	10.1	80.4	5.8
1	1/3	12	36	82.9	4.9	78.0	5.5	12	36	82.9	4.9	78.0	5.5
	1/2	13	26	80.4	5.7	79.9	5.3	13	26	80.4	5.7	79.9	5.3
	1	17	17	79.5	5.8	80.1	5.8	17	17	79.5	5.8	80.1	5.8
	2	26	13	83.8	5.1	78.1	4.4	26	13	83.8	5.1	78.1	4.4
	3	36	12	83.2	4.9	80.1	5.1	36	12	83.2	4.9	80.1	5.1
2	1/3	16	48	81.2	10.0	71.3	5.4	20	60	89.1	10.1	78.8	5.0
	1/2	18	36	80.6	6.7	74.2	4.7	21	42	86.3	8.4	79.4	5.4
	1	24	24	79.8	6.7	75.6	6.7	25	25	79.6	5.8	81.8	5.8
	2	36	18	75.3	3.7	83.7	5.8	34	17	73.5	3.5	81.1	5.8
	3	48	16	76.5	2.0	85.5	4.9	42	14	69.0	2.3	80.0	4.4

Table 2.1 (continued)

Var. rat. (θ)	Alloc rat. (A)	Assuming $s_1=s_2$ (Method (1))		t-test assuming equal variances		t-test with Satterthwaite's correction		Assuming $s_1 \neq s_2$ (Method (2))		t-test assuming equal variances		t-test with Satterthwaite's correction	
		n_1	n_2	Power	Sign.	Power	Sign.	n_1	N_2	Power	Sign.	Power	Sign.
3	1/3	21	63	84.0	13.8	63.4	5.0	28	84	89.2	13.3	80.7	4.9
	1/2	23	46	80.9	10.1	70.7	5.7	29	58	89.7	11.5	78.2	6.0
	1	31	31	76.7	6.4	78.4	6.3	33	33	80.8	5.4	79.6	5.1
	2	46	23	74.2	2.3	83.9	4.8	42	21	69.2	2.6	81.1	5.9
	3	63	21	72.9	1.3	86.3	6.8	51	17	63.6	1.4	81.4	4.9

* For method (1) results were rounded up to the nearest integer. For method (2) rounding was necessary only for n_1.

Methods (1) and (2) give the same result for $\theta=1$. For equal group sizes, that is, for $A=1$, the results of the two methods are close to each other. Table 2.1 shows that for $\theta<1$ the old method overestimates the sample sizes when $A<1$ (when more subjects are assigned to the group with larger variance), and underestimates them when $A\geq 1$. The reversal of this happens for $\theta>1$. Sample sizes are underestimated when $A\leq 1$, while they are overestimated when $A>1$. Intuitively this can be explained as follows: having a better estimate of the smallest variance and a less precise estimate of the largest one means fewer information (thus, less power) than needed. While estimating the larger variance using more subjects means a gain in power compared to the design of equal allocations.

Simulation results from Table 2.1 also reflect that Satterthwaite's approximation gives reasonable estimates of the type I and type II errors. Student's t-test performs well when $A=1$ or $\theta=1$, but shows a considerable departure from the nominal significance level and power when A and θ differ from 1. The t-test with Satterthwaite's correction has even larger departures from the nominal significance level and power as Student's t-test does when the sample size is computed with method (1).

The t-type tests with different adjustments for the inhomogeneity of variances will have different power values for the same sample sizes. For instance, using the data from the last line of Table 2.1 ($n_1=51$, $n_2=17$, $\mu_1=20$, $\mu_2=10$, $\sigma_1^2=300$, $\sigma_2^2=100$, $\alpha=0.05$), the power of Satterthwaite's t-test is 81.4, the Cochran and Cox approximation [12] has a power of 80.2, while the Welch's t-test [13] has a power of 81.8 %. Differences between the power values of the various tests increase when decreasing the significance level. Changing the significance level to α=0.01, for the previous stream of data we obtain: Student 29.6, Satterthwaite 59.6, Cochran 55.2 and Welch 59.8%.

Table 2.2 shows that for greater sample sizes the discrepancy between the approach which ignores inhomogeneity of variances and the one which takes it into account becomes larger. Increasing the power and decreasing the significance level does not change considerably the relative difference between the two methods. Table 2.2 contains the sample sizes and their relative differences $(n_1 - n_1')/n_1$ for different power and significance levels, assuming $s_2^2=10$, $\theta=1/4$, $A=1/4$ and $|\mu_1-\mu_2|=1$ (n_1 is the sample size of group 1 computed with method (1), while n_1' is the same sample size estimated with method (2)).

Table 2.2

The influence of the choice of power and significance level on the relative difference between the two methods

Power (%)	Sign. (%)	n_1	n_1'	$(n_1 - n_1')/n_1$
80	5	384	246	0.561
80	1	572	366	0.563
90	5	514	329	0.562
90	1	728	466	0.562

The above examples underline that in fact each statistical procedure needs his own sample size estimation method. However, there are some robust sample size approximation methods which can be applied in many circumstances. This arises the following question: how far one

can go with the violation of some of the assumptions. Or, in other terms, how can one define the robustness of a statistical method.

We introduce the term "relative discrepancy" between the statistical procedure and the sample size estimation method (for a given significance level) which can be defined as follows: $d = \frac{p_{actual} - p_{nominal}}{se(p_{act})}$. The value $|p_{actual} - p_{nominal}|$ is a measure of the appropriateness of the sample size estimation method assuming that the statistical procedure itself is adequate (that is, its nominal and actual significance levels do not differ). Table 2.3 shows the relative discrepancy between the ordinary t-test (assuming equal variances) and the traditional sample size estimation method (Machin [9]) for different allocation ratios and different variance ratios, for $\alpha=0.05$ (p_{actual} computed from 1000 simulations).

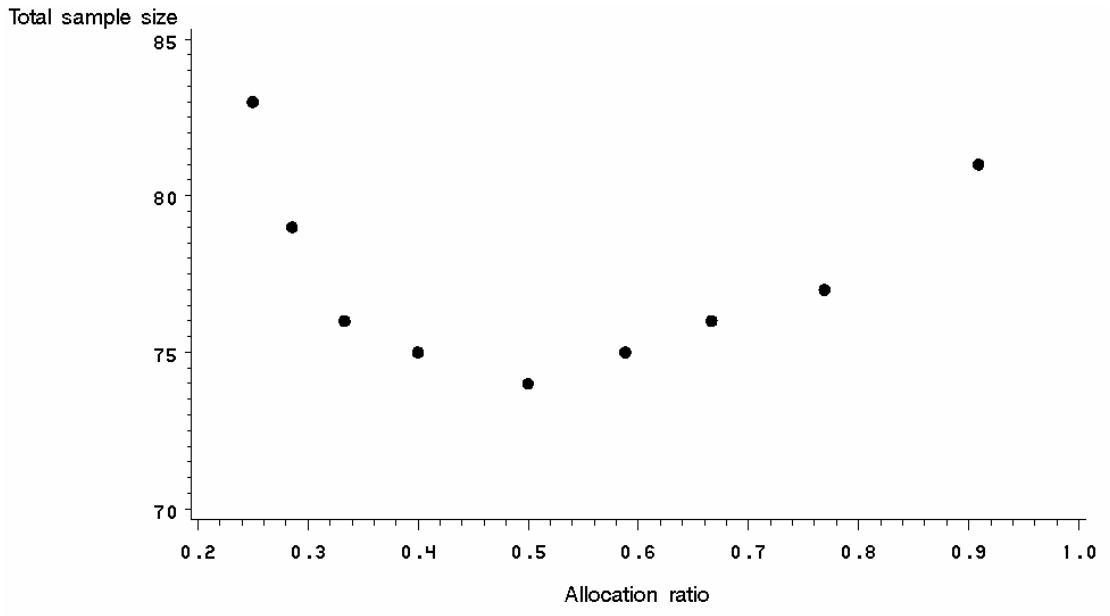
Table 2.3 Relative discrepancy values between the ordinary t-test (assuming equal variances) and the traditional sample size estimation method

Var. ratio	Alloc. ratio	n_1	n_2	Actual power(%)	SE(p_{actual}) (%)	Nominal power(%)	d
1/3	1/3	8	24	77.8	1.31	80	-1.67
	1/2	9	18	79.0	1.29	80	-0.77
	1	11	11	80.0	1.26	80	0.00
	2	18	9	85.2	1.12	80	4.63
	3	24	8	88.1	1.02	80	7.91
1/2	1/3	9	27	80.6	1.25	80	0.47
	1/2	10	20	80.7	1.25	80	0.56
	1	13	13	79.5	1.28	80	-0.39
	2	20	10	83.8	1.17	80	3.26
	3	27	9	84.6	1.14	80	4.03
1	1/3	12	36	82.9	1.19	80	2.44
	1/2	13	26	80.4	1.26	80	0.32
	1	17	17	79.5	1.28	80	-0.39
	2	26	13	83.8	1.17	80	3.26
	3	36	12	83.2	1.18	80	2.71

Table 2.3 shows (similarly to Table 2.1) that the discrepancy for a given variance ratio is the smallest when the allocation ratio equals to 1.

For $\theta=1$ it is known [9] that the optimal allocation ratio (that which minimises the total sample size, n_1+n_2) is $A=1$. Schouten's method offers the possibility of computing the optimal allocation ratio for a given θ. Computing the first-order derivative of the total sample size as a function of A (and keeping θ constant), it equals 0 for $A=\sqrt{\theta}$ and analysing the sign of the derivative it follows that $A=\sqrt{\theta}$ is a minimum point for the total sample size $N(A)$. Figure 2.1 shows that for $\alpha=0.05$, $1-\beta=0.8$, $|\mu_1-\mu_2|=1$, $s_1=1$ and $s_2=2$, the minimum of the total sample size is attained when $A=1/2$.

Figure 2.1 *The minimum of the total sample size is attained for A=0.5 when s_1/s_2 is supposed to equal 0.5*



Using Satterthwaite's correction for sample size determination [3] may improve the planning of experiments when unequal variances are anticipated and no variance-stabilising transformation is available. The iterative character of the algorithm presented here enables the substitution of Satterthwaite's approximation with other methods (Cochran and Cox, Welch, etc.).

Example 1. (real example)

In a pharmacokinetic study assessing food effect in two parallel groups, there were 12 subjects fed and 12 subjects fasted. After a single-dose administration of the active compound to both groups, the mean area under the curve was 120.6 mg.h/l (s.d.=127.8 mg.h/l) in the fasted and 1119.4 mg.h/l (s.d.=340.2 mg.h/l) in the fed group. After the administration of another formulation of the same compound to the same subjects, the mean area under the curve was 84.1 mg.h/l (s.d.=96.8 mg.h/l) in the fasted and 968.6 mg.h/l (s.d.=371.2 mg.h/l) in the fed group. Planning a new study based on these data, a coefficient of variation of 100% in the fasted and of 33.3% in the fed group can be assumed. If the fasted mean area under the curve is 300 mg.h/l and the least clinically relevant difference is of 300 mg.h/l, the group size needed in case of equal group sizes is of 65 subjects/group for a 1% significance level and a power of 90%.

If the AUCs (areas under the curve) are deemed to be log-normally distributed (as it is usually recommended), the sample size estimation can also be performed on the transformed data, using the transformations [12] which make the transition between the mean and variance of the initial and the log-transformed data:

$$l\_mean = (4\log(mean) - \log(mean * mean + sd * sd)) / 2$$

and

$$l\_sd = \sqrt{\log(mean * mean + sd * sd) - 2\log(mean)}$$

(where mean and sd are the statistics for untransformed data, while l_mean and l_sd are the statistics for the log-transformed data).

The transformed means are 5.6511 and 6.0504, while the transformed standard deviations are 0.3246 and 0.8326, respectively. So on the log scale the variances are still very different. Recomputing the sample size with these assumptions, 77 subjects/group are needed.

Example 2. (Snedecor and Cochran [8])

This is a constructed example in which two methods of estimating the concentrations of a chemical are compared. The standard method, a precise but slow one, is anticipated to give a mean of 25 and a variance of 0.67. The new method, which is quick but less precise and which is suspected to systematically over- or underestimate concentrations, is supposed to have a mean of 21 and a variance of 17.71. Analysing the sample sizes needed to have a test of 80% power for the above data and for an allocation ratio of 1/4, method (1) yields 5 and 20. Group sizes obtained with method (2) are 3 and 12.

3. More than two parallel groups with unequal sizes

In case of two groups when the overall sample size is kept constant the maximum power is achieved for equal number of elements in each group. However, it may happen that for some reasons (a rare disease, a very expensive compound) unequal group sizes are preferred. Adjustments for two groups are known. For more than two groups the noncentral F distribution can be used. It needs only a natural extension of the relation which defines the noncentrality parameter. The maximum efficiency is not any more attained when group sizes are equal.

The method described in Fleiss [14] (pp.371-376) is briefly the following:
For g parallel groups supposed to have a normal distribution and equal means $(\mu_1 = \mu_2 = \ldots = \mu_g)$ with n subjects in each, the ratio of mean squares from the analysis of variance table (mean square between groups/mean square within groups) follows a central F distribution having two parameters, the degrees of freedom of the numerator and that of the denumerator.
When the null hypothesis is rejected, thus there are at least two different group means, the ratio mentioned above has a noncentral F-distribution, which also depends on a third parameter, the noncentrality parameter. This parameter is defined as being:

$$\delta = \sqrt{n\sum_{1}^{g}\left(\mu_i - \overline{\mu}\right)^2} \Big/ \sigma \qquad (1)$$

where $\overline{\mu} = \sum_{i=1}^{g} \mu_i \Big/ g$ and σ is the common variance of the groups. To achieve a power of $1-\beta$ while the significance level is α, we have to increase n until we find that $\Pr\left(F^{nonct}{}_{\nu_1,\nu_2,\delta} > F_{\nu_1,\ \nu_2,\alpha}\right) \geq 1-\beta$ $(\nu_1 = g-1,\ \nu_2 = ng-g)$. Scheffé [15] (pp. 38-42) gives a formal definition of the noncentral F distribution, while Laubscher [16] gives a good normal approximation to the square root of a noncentral F variate.

In case of unequal group numbers let $n_1, n_2, \ldots, n_g$ denote the number of elements in each group. We define the set of allocation ratios as $(r_1, r_2, \ldots, r_g)$ with $n_1 = r_1 n_g$, $n_2 = r_2 n_g, \ldots, n_{g-1} = r_{g-1} n_g$ and $r_g = 1$ [4]. Let $\overline{\mu}_w$ denote the overall mean of the whole sample, that is the weighted

mean of the group means $\mu_1,\mu_2,...\mu_g$ with weighting factors $r_1,r_2,...,r_g$. Scheffé's definition for the noncentrality parameter says: "If in the sum of squares in the numerator of F each observation is replaced by its expected value under Ω (Ω denotes the underlying assumptions), the result is $\sigma^2\delta^2$".

Thus the noncentrality parameter can be calculated as $\delta = \sqrt{\sum_{i=1}^{g} n_i(\mu_i - \overline{\mu}_w)^2} \Big/ \sigma$ (2).

All the rest of the rationale remains the same as described for the equal group numbers. For $n_1=n_2=...=n_g=n$ (that is, $r_1=r_2=...=r_g=1$), (2) reduces to (1).

Keeping the overall sample size constant ($\sum_{i=1}^{g} n_i = k$) the maximum power is achieved when δ is as great as possible. To obtain the maximum point of δ the partial derivatives of its numerator under the above condition were calculated.

The numerator of the squared noncentrality parameter is

$$f_g = f\left(n_1, n_2, ..., n_g\right) = \sum_{i=1}^{g} n_i \left(\mu_i - \frac{\sum_{j=1}^{g} n_j \mu_j}{\sum_{j=1}^{g} n_j} \right)^2$$

From the condition $\sum_{i=1}^{g-1} n_i = k$ we substitute : $n_g = k - \sum_{i=1}^{g-1} n_i$

$$f_{g-1} = \sum_{i=1}^{g-1} n_i \left(\mu_i - \frac{\sum_{j=1}^{g-1} n_j \mu_j + (k - \sum_{j=1}^{g-1} n_j)\mu_g}{k} \right)^2 + (k - \sum_{j=1}^{g-1} n_j) \left(\mu_g - \frac{\sum_{j=1}^{g-1} n_j \mu_j + (k - \sum_{j=1}^{g-1} n_j)\mu_g}{k} \right)^2$$

Then $f_{g-1} = \sum_{i=1}^{g-1} n_i \left(\mu_i - \mu_g - \varphi \right)^2 + \left(k - \sum_{j=1}^{g-1} n_j \right) \varphi^2 \left(n_1, n_2, ..., n_{g-1} \right)$

where $\varphi = \sum_{j=1}^{g-1} n_j \left(\mu_j - \mu_g \right) \Big/ k$.

So $f_{g-1} = \sum_{i=1}^{g-1} n_i \left(\mu_i - \mu_g \right)^2 - k\varphi^2 \left(n_1, n_2, ..., n_{g-1} \right)$

Thus for any $m=1,2,...,g-1$ we have

$$\frac{\partial f}{\partial n_m} = \left(\mu_m - \mu_g\right)^2 - 2k\varphi\frac{\partial \varphi}{\partial n_m} \qquad \text{and} \qquad \frac{\partial \varphi}{\partial n_m} = \frac{\mu_m - \mu_g}{k}$$

Hence

$$\frac{\partial f}{\partial n_m} = \left(\mu_m - \mu_g\right)\left[\left(\mu_m - \mu_g\right) - 2\frac{\sum_{i=1}^{g-1} n_i\left(\mu_i - \mu_g\right)}{k}\right]$$

This function equals 0 if $\mu_m=\mu_g$ for all *m=1,2,...,g-1* (which is in fact the null hypothesis and is a minimum point for δ) or if $\sum_{i=1}^{g-1} n_i\left(\mu_i - \mu_g\right) = \frac{k\left(\mu_m - \mu_g\right)}{2}$ (3) . Since *m* is arbitrary, the expression on the right hand side of the equation is constant. Replacing μ_m-μ_g=*c* in (3) for *m=1,2,...,g-1* we obtain $\sum_{i=1}^{g-1} n_i c = \frac{kc}{2}$, that is $\sum_{i=1}^{g-1} n_i = \frac{k}{2}$ and $n_g = \frac{k}{2}$.

Thus as Singer pointed out in [4] - if no constraints are added - δ has an extreme if the anticipated means of *(g-1)* groups are equal and one of them is different. In this case the size of the group with different mean has to be a half of the total sample size to obtain the maximum of the noncentrality parameter (thus, maximum power). This is not the only solution, as we can see in Finney's [17] example in which pairwise comparisons are assessed instead of global significance:

More test preparations are compared to the same standard in a parallel line assay and "the possibility of gaining in efficiency by unequal distribution must be borne in mind". Finney [17] searches for the optimal allocation minimizing the variance of log potency (the horizontal distance between the control and a test preparation line when dose-response curves are assessed) which is approximately proportional to the variance of the difference between means, while making no hypothesis on the anticipated group means. For a fixed number of subjects where only one response per subject can be measured and *(g-1)* of the *g* group sizes are equal (n_1+*(g-1).*n_2=*k* and $n_2=n_3=\ldots=n_g$), the variance of the difference between the control and a test mean is proportional to $1/n_1+1/n_2$. This expression has a minimum if $n_1 = n_2\sqrt{g-1}$ or if n_1=*k/2* and n_2=*k/2(g-1)* (this latter solution is the same as the one mentioned above).

For $g=2$ both cases mean exactly $n_1=n_2$. Thus equal group numbers constitute indeed an optimal design for two groups, but not for more than two groups.

Coming back to the general case, without loss of generality $\mu_1 \le \mu_2 \le .. \le \mu_g$ can be assumed. The maximum of the noncentrality parameter will then be $\delta_{\max} = k\left(\frac{\mu_g - \mu_1}{2}\right)^2 \Big/ \sigma$, exactly the same as that obtained for allocation ratios *1,0,...,0,1* (but this set of ratios involves only two groups, so it is not a sensible design). This approach answers only the question of global significance and does not solve the problem of different contrasts. For further investigations Scheffé [15] recommends "to ask whether the confidence intervals for the quantities of interest will be sufficiently narrow".

Examples

For given α, β, μ_is and r_is ($i=1,2,...,g$) n_g was increased until the inequality $\Pr\left(F^{nonct}{}_{\nu_1,\nu_2,\delta} > F_{\nu_1,\ \nu_2,\alpha}\right) \ge 1-\beta$ was fulfilled (ν_1=g-1, $\nu_2 = \sum_{i=1}^{g} n_i - g$).

1. The first example, presented by Day and Graham [18] (1991), compares the mean diastolic blood pressure of three groups. The anticipated means are 100 mmHg, 95 mmHg and 85 mmHg, the common standard deviation is 15mmHg. The significance level is 0.01, the power is planned to be 0.9 . The sample size obtained on the proposed nomogram is of 35 patients/group (equal group numbers).

Trying the above method with the same means and common standard deviation and the allocation ratios 2, 1 and 1 we obtain $\delta=4.32$ and a total sample size of 112 distributed in three groups as 56, 28 and 28. In case of equal numbers a total sample size of 105 would have been sufficient. If we choose the allocation ratios to be 1,1,2, the sample sizes are 25,25,50 - a total of 100 which is less than that obtained for equal group numbers (105). Putting the condition of $n_i \ge 3$ for i=1,2,3 the minimum of total sample size is 77 (group sizes: 37, 3, 37 or 36, 3, 38 and a noncentrality of 4.31), but this is a rather disproportionate design. (Without the

condition $n_i \geq 3$ for i=1,2,3 the minimal size is 38+0+38=76 and this is in accordance with the result "felt" intuitively: the larger are the weights put on the extremes of the group means the greater is the value of the noncentrality.) To obtain a sensible optimal design some constraints are needed.

Submitting the test program for 1000 stream of data the following results (Table 3.1) were obtained:

Table 3.1

Group sizes			Total sample size	Group means (mmHg)			Number of signinficant ANOVAs
35	35	35	105	100	95	85	905
35	35	35		100	100	100	11
56	28	28	112	100	95	85	897
56	28	28		100	100	100	9
25	25	50	100	100	95	85	928
25	25	50		100	100	100	13
37	3	37	77	100	95	85	908
37	3	37		100	100	100	9

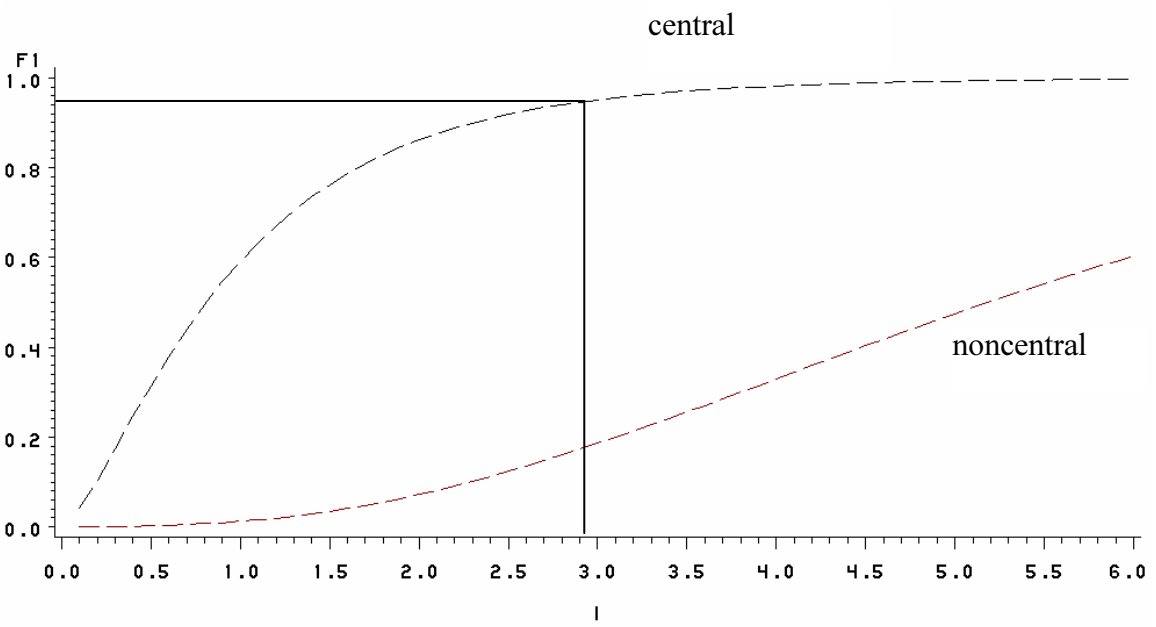
(These numbers were obtained for seed=13426,13427,...,14425. Due to the definition of the rannor procedure by choosing the same positive seed the same streams of variates can be generated again.)

2. An example with four groups is given by Fleiss [14] . The four anticipated means are 9.775, 12, 12 and 14.225. The standard deviation within each group is expected to be 3, the significance level is 0.05 and the power of the test 0.8. For equal group sizes the required sample size per group is 11. Under the condition that each group has at least three observations the optimal allocation ratios are 4:1:1:4 (for group sizes 12, 3, 3, 12 the noncentrality is 3.633 and the power is 0.82) - the total size of 44 could be reduced to 30. The cumulative central and noncentral F distributions for this latter case (ν_1=3 and ν_2=26,

$\delta=3.633$) are shown on Fig. 3.1 . Calculating the F values corresponding to the dashed lines we find $F_{3,26}(2.975)=0.95$ and $F^{nonct}{}_{3,26,3.633}(2.975)=0.183$.

Figure 3.1.

Cumulative central and noncentral (noncentrality=3.633) F distributions with degrees of freedom 3 and 26. For x values less than 2.975 the null hypothesis is not rejected because $F_{central}(x)<0.95$. The corresponding area under the noncentral F distribution is 0.183, thus the power is 0.817.



3. For two groups we have to obtain the same result which is given by the formulas $n_1 = \frac{A+1}{2A}n$ and $n_2 = \frac{A+1}{2}n$, where $\frac{n_2}{n_1} = A$ is the allocation ratio and n is the sample size for equal groups calculated by $n = \frac{2\sigma^2\left(z_{\alpha/2} + z_{\beta}\right)^2}{\left(\mu_1 - \mu_2\right)^2}$. Such a worked example (due to Godfrey et al) is presented by Campbell, Julious and Altman [19] (1995). The blood pressure of people who have no whorls on their finger is compared with blood pressure of people who have at least one whorl. Let 5 Hgmm be the minimum clinically accepted difference and the standard deviation within each group is assumed to be 17 Hgmm. We would expect to recruit two people with whorls for every one person with no whorls. The sample sizes to detect this

difference with a two-sided significance level of 5% and a power of 80% will be 137 and 274, respectively (calculated by the original formula for two normal distributions mentioned before). The approximate sample sizes for the standardised difference $d=5/17\approx0.3$ read from the table presented in the same paper are 132 and 264. While using the method based on the F distribution we obtain a noncentrality of 2.81 and sample sizes of 137 and 274.

Thus for two groups this extension gives the same result as the known formulae. In case of two groups and for fixed overall sample size the maximum power is attained when groups are equally sized, while in case of more than two groups a general solution for the same problem cannot be found (but there exist allocation ratios which give greater power than the equal ones). Sensible constraints are necessary to obtain feasible sample sizes.

4. Simulation-based sample size investigations

There are occasional situations where power and sample size equations or software do not exist. In such situations, the power or sample size can be computed with Monte Carlo simulation techniques. This section describes some situations where simulation-based power and sample size calculations are a necessary or useful alternative, and provides guidelines for conducting such calculations.

Simulation-based methods are a useful alternative whenever the power equation for the data analysis/decision procedure of interest is not analytically tractable, so no closed-form expressions and no approximation-based software exist. Such situations occur in case of: 1) multi-step data analysis/decision procedures (such as some multiple comparison tests, and some analytical tests for drug manufacture, release, and stability), 2) multi-part data analysis/decision procedures (such as multiple endpoints like AUC and Cmax, or inference about both means and variances), and 3) various ad hoc data analysis/decision procedures (such as those based on percent conforming of individual observations in bioanalytical validation, and also in some analytical tests for drug manufacture, release, and stability).

Secondly, simulation-based methods are a necessary alternative whenever the assumptions of the data analysis/decision procedure of interest are likely to be violated. In such situations, the usual sample size estimation methods and softwares may give results far from the actual values. Additionally, the robustness of the type I error against deviations from the assumptions can be determined. Examples of such assumption deviations include 1) small/moderate sample sizes for asymptotic statistical procedures, 2) departures from the distribution form (such as non-normality), and 3) departures from variance structure assumptions (such as non-homogeneity, non-independence, and hierarchical variance components).

Thirdly, simulation-based methods may give better results than the traditional methods when the latter ones are based on approximation formulae, which may not be very good in the situation of interest (such as small or moderate sample sizes).

After presenting so many applications of the simulation one may ask whether it has any contraindications. In fact, a situation when the simulation-based method gives worse result than the traditional one is hard to imagine. However, there are many situations when it does not have any advantage compared to the other methods and so it is not worth applying.

Steps to conduct a simulation study for power or sample size calculations

Situations where simulation-based sample size calculations are used can be quite complex. Although each situation has its own unique complexities, a general set of basic recommended steps can be specified.

Planning

Select the primary variables, the experimental design, the statistical method and the parameters (variability, effects, error risks α and β) according to the study plan, like for any traditional sample size estimation.

Programming

1. Specify the probability distribution of the outcome variable for each treatment population, and the program function to generate these random errors (e.g., the SAS RANNOR function to generate random N(0,1) errors).

2. Specify the statistical model and its parameter values (fixed effects, random effects, variance structure).

3. Specify the sample sizes for each group (these values should be regarded as a minimum size).

4. Generate random data according to the conditions in steps 1-3, and repeat in a DO loop for a total of N experiments. N should be chosen depending on the desired standard error of the power, using the formula:

S. E. = $\sqrt{\frac{p(1-p)}{N}}$. Example: For $p=0.8$ and $S.E.<0.01$ we have $N>\frac{p(1-p)}{S.E.^2}$, or $N > \frac{0.8 \times 0.2}{0.0001} = 1600$). Use random seeds which allow verification of your results at a later time (e.g., in SAS use the RANNOR function initialized with a positive seed rather than the 0 (time clock) seed - see SAS® Language, Reference, Chapter 12).

5. Perform the statistical analysis of the data generated, and save the *p* values (for each of the *N* streams of data). Compute the power (= proportion of the *N* simulated experiments where the desired conclusion was reached).

6. If the computed power is less than the prescribed level, then increase the group size accordingly, and repeat steps 3-5.

7. If the power has attained the prescribed level, then the sample size is sufficient.

8. If the power exceeds considerably the prescribed level, then the group sizes can be decreased, and steps 3-5 repeated.

SAS-specific pieces of advice for programming:

1. Store each of the *N* experiments in the same dataset
2. When possible, code statistical procedures within the DO loops rather than executing formal PROCs
3. When possible, generate necessary statistics instead of actual observation (a performance analysis was done and it yielded that generating directly the statistics instead of individual values - where possible - means a considerably less runtime)
4. Suppress output to LOG and OUTPUT windows (otherwise one gets the error message ‘Window is full’)

Example: Blend Uniformity Analysis

A typical situation where simulation-based sample size is required is the blend uniformity analysis, where - according to the guidelines - a "composite decision" procedure is needed. A blend is declared to be uniform when the following two criteria are met for 3 independent batches:

1. Strength (the actual mean as a percent of the theoretical mean) between 90% and 110%
2. Relative standard deviation (RSD, or - in other words - the coefficient of variation) less than 5%

The question is then how many samples have to be taken from each batch to have a high probability (e.g. >99%) of passing the test when all the three batches fulfil the uniformity criteria.

Probabilities of passing the test were computed by simulation for different sample sizes (sample size was increased by a step of 10) for a theoretical mean of 100% and RSD=4%.

The results are given in the following table (Table 4.1):

Table 4.1 The probability of passing the test for different sample sizes

N	Probability of passing the test for	
	1 batch	3 batches
10	89.5	69.1
20	94.1	83.3
30	97.2	90.7
40	>99	97.1
50	>99	98.3
60	>99	98.6
70	>99	>99
80	>99	>99
90	>99	>99
100	>99	>99

Thus, a probability greater than 99% can be obtained for *N=70* (in case of 3 batches).

5. The power and the sample size as random variables (sensitivity analysis)

The simplest definition of the sensitivity analysis would be: "the study of robustness of the sample size analysis under different scenarios when the initial assumptions are not met". However, more sophisticated interpretations can be given and the notion of “initial assumptions” can be extended by not considering them simply a set of initial values but some random variables with given distributions.

Most power analyses are prospective, that is, they assess the power of a future study. The mean and variance estimates used in the planning phase are anticipated, usually based on a pilot study or some previous trials. Hence the power for a given sample size, or the sample size needed to achieve a certain power, vary in fact randomly, since their calculation is based on some observed values of random variables (usually the sample mean and the variance). Thus it is preferable to study a whole range of possible values instead of a single point estimate. In fact, this is one of the possible definitions of the sensitivity analysis: to study the distribution of the sample size (or that of the power) for different distributions of the starting values.

Another role of the sensitivity analysis might arise when a longitudinal study covers a long period of time and the endpoint itself is time-dependent. Such situation might occur for instance in a cohort study where the endpoint is an event rate the frequency of which is decreasing with time in the control group due to new and efficient prevention methods [21]. Moyé [21] gives an example of a clinical trial in post myocardial infarction where new cholesterol reducing therapies become available which are expected to decrease the incidence of infarction. In such situations the assumption of a constant control group event rate may lead to severely underpowered trials.

According to the above definitions, different types of sensitivity analyses can be conducted. Some examples of sensitivity analysis are given below:

5.1 Confidence intervals

The most frequently used sample size formulae are based on the relationship between the variance of the estimator of a parameter of interest and the sample size. Usually the input of these formulae has a "primitive" part (e.g. type I and type II errors that may also be reconsidered as perhaps they are less “primitive” than it is usually thought and their choice does not need to be automatized - but this would be the subject of another paper) and a "sophisticated" part which needs qualified guesswork (e.g. population variances, population event rates, etc.). Currently, the routine practice is a kind of "plug-in principle", that is, to take the estimated value of a parameter from a previous study (or to combine in some way the results of more studies) and to substitute it in the formula of choice as being the "theoretical" value of the respective parameter. As some authors already pointed out [6],[32],[33] this practice is rather risky.

In the present chapter the continuous outcome and two parallel group design is analysed. We suppose equal variances, a given least clinically relevant difference and fixed error levels. If data from a previous study with one group of size *N* are available and the estimated variance is $\hat{\sigma}$, then $(N-1)\hat{\sigma}^2/\sigma^2$ is chi-square distributed with *N-1* degrees of freedom. Supposing the equality of the estimated variance and the theoretical one (applying the "plug-in method") means in fact to pick up that γ percentile of the chi-square distribution for which the equality $N-1=\chi^2_{\gamma,df}$ holds. Thus, one can be $(1-\gamma)100\%$ confident that the "true" variance is less than our estimated one. Table 5.1 shows the values of $(1-\gamma)$ corresponding to different values of *N*.

If data from a previous study with two parallel groups, each of size *N/2* are available an estimate for the population variance can be obtained by taking the usual pooled estimate, $\hat{\sigma}$. In this case $(N-2)\hat{\sigma}^2/\sigma^2$ is chi-square distributed with *N-2* degrees of freedom. Supposing the equality of the estimated variance and the theoretical one (that is, to use the plug-in method) means in fact to pick up that γ percentile of the chi-square distribution for which the equality $N-2=\chi^2_{\gamma,df}$ holds. Thus, the equation is similar to that obtained for one group. The results are shown in Table 5.1.

Table 5.1 The probability that the true variance is less than the estimated one

	1-γ	
N	one-group pilot study	two-group pilot study
3	0.368	0.317
4	0.392	0.368
5	0.406	0.392
6	0.415	0.406
10	0.437	0.433
15	0.450	0.448
20	0.457	0.456
50	0.473	0.473

Table 5.1 shows that the confidence level remains always below 50% (for both the one- and two-group pilot study). As χ^2 is asymmetric (converging slowly to the normal distribution $N(\mu,2\mu)$), the confidence level of 50% is never attained. This means that the probability of attaining the desired power is less than 50%.

To avoid the “surprisingly low powers” due to the random variation of the standard deviation one solution would be the use of internal pilots [34], [35], [36]. But this method can be applied only for long clinical trials and might arise problems with the unblinding of the trial.

The other solution is the use of a confidence interval method. A sensitivity analysis is usually performed together with the sample size estimation (computing the sample size for the confidence limits of $\hat{\sigma}^2$). However, in most of the cases the chosen final sample size is that belonging to $\hat{\sigma}^2$ and the results of the sensitivity analysis are not used further.

The situation is worsened by the fact that - as it is shown by O'Brien [6] - "unless the data are strictly Normal, the traditional χ^2-based confidence intervals for $\hat{\sigma}^2$ cannot be trusted to have their nominal confidence levels"[2] .

The aim of this study is to compare the coverage rates of different confidence interval methods.

The assumption of equal variances is supposed to be true, a given least clinically relevant difference (d) and fixed error levels (α and β) are supposed. The least clinically relevant difference is assumed to be equal to the difference between the population means. Since the lower limit of the confidence interval does not have too much importance for the sample size, only one-sided confidence intervals are considered. For a balanced two-group design with equal group variances and two-sided test the sample size formula:

$$n = \frac{2\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \sigma^2}{\left|\mu_1 - \mu_2\right|^2} + \frac{z_{1-\alpha/2}^2}{4} \qquad (1)$$

performs well [3], [9], so this formula is used throughout the present study. The methods evaluated are as follows:

a. χ^2-based confidence interval method

Given the results of a pilot study, the upper *(1-γ).100%* confidence limit of the estimated variance is computed by the formula:

$$(N-2)\,\hat{\sigma}^2 / \chi^2_{\gamma,df} \qquad (2)$$

The sample size corresponding to this variance is taken as the upper *(1-γ).100%* confidence limit of the sample size.

b. Parametric bootstrap confidence interval

A parametric bootstrap [20] (with 2000 bootstrap samples) is performed on the results of a given pilot study. The sample size is estimated for each bootstrap sample. The *(1-γ)* percentile of the sample size distribution is taken as the upper *(1-γ).100%* confidence limit of the sample size.

c. Noncentrality-based confidence interval

This method is proposed by O'Brien [6]. The t-value t_0 is computed for the given pilot study of size *N*. The *(1-γ).100%* upper confidence limit of the noncentrality δ_γ is computed by the formula:

$$\Pr(t^{nonct}_{df,\delta\gamma} \geq t_0) = \gamma \qquad (3)$$

The noncentrality value δ_γ' of a similar study from the same population but with a size *N'* is then

$$\delta_\gamma' = \frac{\sqrt{N'}}{\sqrt{N}}\delta_\gamma \qquad (4)$$

The sample size of the new study has to be increased until the

$$\text{Power} = \Pr(t^{nonct}_{df',\delta'} \geq t_{crit}) \geq 1-\beta \qquad (5)$$

becomes true, where $t_{crit} = t_{df',\alpha}$ is the *(1-α/2)* percentile of the central t-distribution.

In order to compare the three methods, the coverage rates of the confidence intervals and the relative distances of their upper limit from the theoretical value were computed. The theoretical means of the two populations and the common standard deviation were supposed to be known and 2000 two-group pilot studies of size 10 were drawn (normal distributions were deemed and the *rannor* procedure of SAS was used). The upper limit of the confidence interval for sample size estimation was then computed for each method and it was compared to the "theoretical" sample size, the one obtained for an "exemplary" pilot study having the estimated standard deviation equal to the theoretical one. The coverage rate was the proportion of cases when the theoretical sample size didn't exceed the upper limit of the confidence interval. The relative distance was the absolute value of the difference between the upper limit of the confidence interval and the theoretical mean expressed as a percentage of the theoretical mean ($rel.dist = \frac{abs(upper\,limit - theoretical\,value)x100}{theoretical\,value}$).

The fixed values used for this simulation study were as follows: $\mu_1=10$, $\mu_2=18$, $\sigma=10$, $N=10$ (the total size of the pilot study), $\alpha=0.05$, $\beta=0.2$, $\gamma=0.1$, *0.2* and *0.3*. The "theoretical" sample size computed by (1) for the "exemplary" study is

$$n = \frac{2\left(z_{0.975} + z_{0.8}\right)^2 10^2}{|8|^2} + \frac{z^2_{0.975}}{4} \approx 26$$

Table 5.2 contains the coverage rates and the mean relative differences obtained for each method.

Table 5.2. The coverage rates (%) and mean relative distances (%) of the different methods for the confidence levels of 70%, 80% and 90%

Nominal confidence (1-γ).100 (%)	Method					
	a		b		c	
	coverage rate	relative dist.	coverage rate	relative dist.	coverage rate	relative dist.
70	72.69	57.1	62.0	40.1	54.3	64.4
80	82.20	78.8	71.6	50.9	79.3	225.6
90	91.90	125.6	83.4	79.5	84.5	721.3

4. Example (constructed)

A placebo-controlled parallel group study was planned in obesity. The outcome variable was the body weight. The clinically relevant difference was presumed to be 5 kg. A pooled standard deviation of 15 kg was obtained from a pilot study of size 12. A two-sided significance level of 5% and a power of 80% was required. Substituting in equation (1) we have

$$n = \frac{2\left(z_{0.975} + z_{0.8}\right)^2 15^2}{|5|^2} + \frac{z_{0.975}^2}{4} \approx 142$$

Hence, about 142 completing patients per treatment arm were required when using the usual "substitution" method.

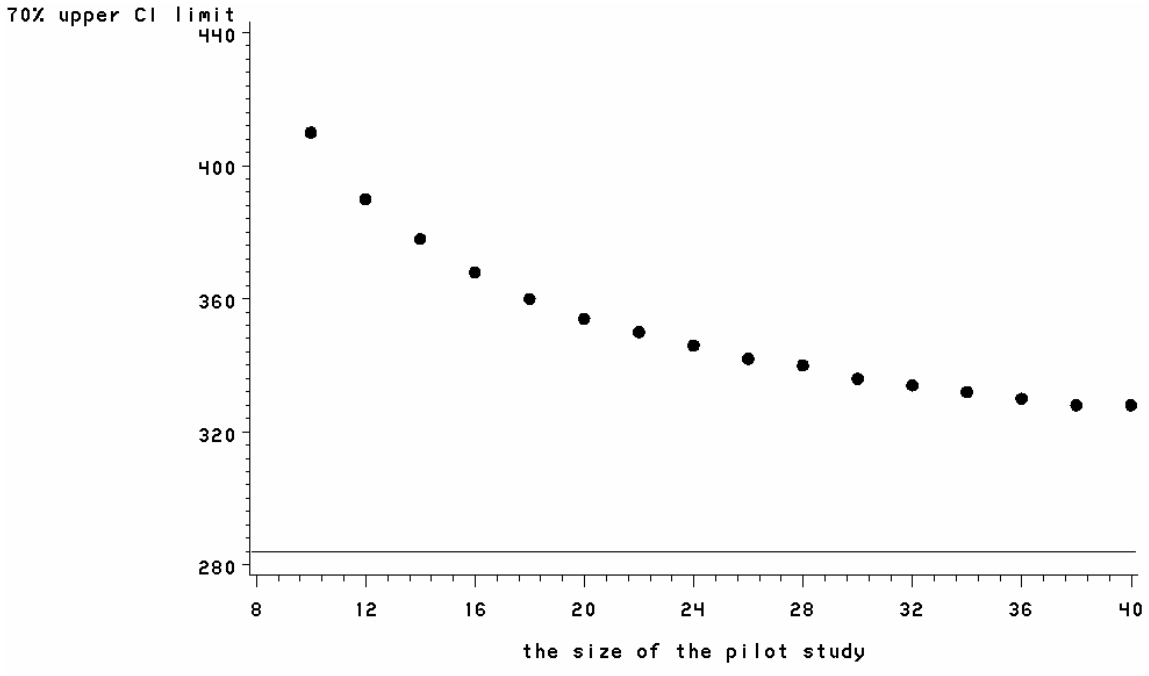
The 90% confidence interval of the standard deviation was [11.1, 23.9]. The width of the confidence interval reflected that the size of 142 per group was not at all reassuring. For example if in the future study the sample standard deviation would attain the former upper confidence limit of 23.9, the sample size of 142 yielded only a power of 42% . For a power of 80% and a standard deviation of 23.9 the sample size/group of 360 would be required which

was a total sample size 2.5 times larger than the initially planned one. It was decided to use a one-sided confidence level of 70%.

The first two methods yielded upper confidence limits of 195 and 166, respectively, while the third one had a result greater than 1000. The final sample size was chosen to be 166.

Increasing the size of the pilot study enables better planning precision and the gain in the final sample size may be considerable. Figure 5.1 presents the upper 70% confidence limit of the sample size as a function of the size of the pilot study. This figure shows that under the conditions of the example the increase of the pilot study size above 24 is of little use.

Figure 5.1. *The precision of the sample size estimation depending on the size of the pilot study*



The χ^2-based confidence interval method gives the best coverage rates (the closest to the nominal confidence levels) but rather large relative distances (and thus large sample sizes). The noncentrality-based confidence intervals are extremely large and without having a highly

reassuring the coverage rates. The bootstrap confidence intervals yield usually smaller coverage rates than the nominal confidence level but with lower relative distances and lower sample sizes. The one-sided confidence level of 70% can be recommended since the other levels result in unfeasibly wide ranges.

In most cases the variance estimates used in the planning phase are anticipated, usually based on a pilot study or some previous trials. Hence the sample size needed to achieve a certain power varies in fact randomly and the "plug-in principle" may lead sometimes to surprisingly low power values. The confidence-interval approach enables the improvement of the current practice. In case of a variance the estimated value corresponds to a confidence level less than 50%. The one-sided bootstrap upper 70% confidence limit (or the χ^2-based confidence limit) can be used instead.

5.2 Time-dependent effect size

An osteoporosis-prevention trial aimed to prove the long-term effect of a drug slowing down the rate of loss of bone mineral density (BMD). The study was a placebo-controlled, randomized clinical trial with two parallel groups. The primary endpoint was the rate of change in BMD after a 36-month treatment. The standardized effect size was supposed to be 0.55 . There was not taken into account that during the study a new diet containing calcium and vitamin D became available and the BMD of the control group lowered with a smaller rate due to this diet. The effect size was only 0.49 and this resulted in an underpowered trial.

The time-dependence of the effect size can be modelled with different functions. The simplest one is the linear one, supposing that the standardized difference is decreasing with time linearly. Let's suppose that in the above trial the effect size had a linear relationship with $\sqrt{T}$, thus the effect size was given by the function:

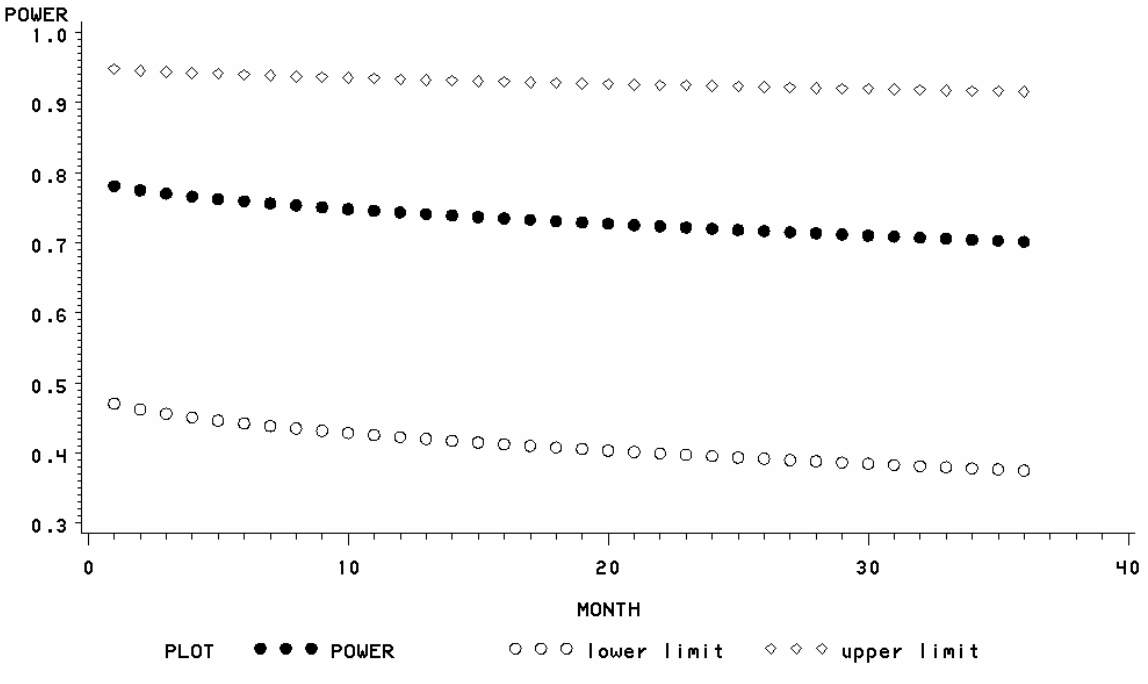
$$e(T) = e_0 - b.\sqrt{T}$$

If t is the time measured in months then b is given by the equation

$0.49 = 0.55 - b.\sqrt{36}$, hence $b=0.01$

For a balanced design with group sizes of 40 and for a significance level of 5% the power of the t-test and its 60% confidence interval depending on T is shown in Figure 5.2

Figure 5.2 *Time-dependent power and its 60% confidence interval*



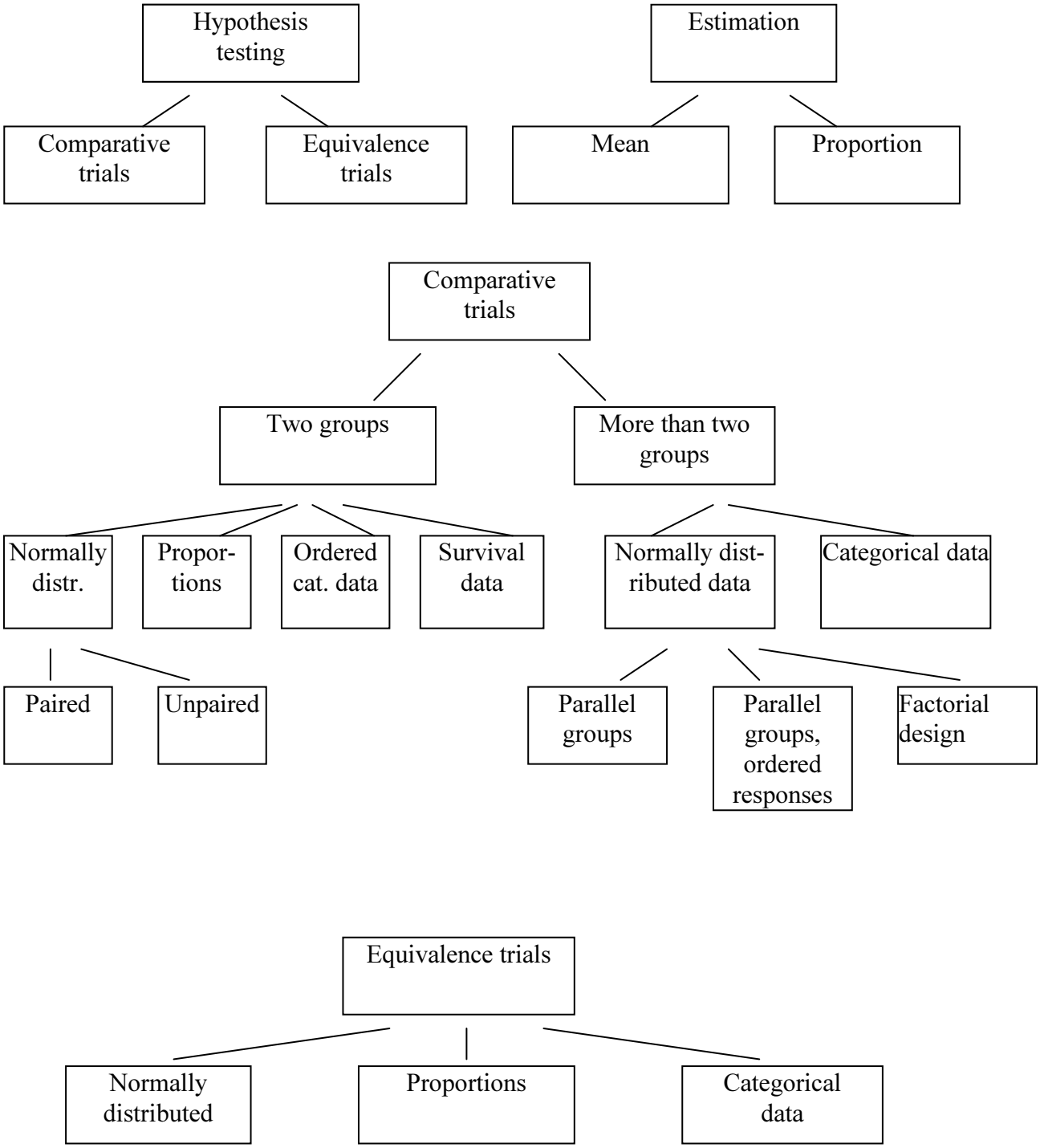
6. A SAS application for trial sample size determination

The sample size estimation in prospective clinical and preclinical trials is much more than a statistical task. It needs an agreement between researchers on the anticipated value of the parameters (the design, the treatment effect, the level of the type I and type II errors and the sidedness as well). Therefore the basic notions of the sample size estimation have to be known by all the researchers participating in the study planning. "Going through the process of determining and justifying the sample size also has an important ancillary effect: it catalyzes the synergism between science and statistics at the study's conception" - formulated concisely O'Brien in [6].

Thus the need of an interactive teaching software, a kind of tutorial for researchers was a stringent need. Our aim was to visualise by our application (written in SAS® 6.08, transposed later on to SAS® 6.11) the different models in order to help the physicians in using the program and understanding the underlying models. To achieve this we used frame entries and help entries. We associated certain widgets for certain variable types (for instance sliders for proportions, list boxes for choosing power, confidence and significance levels, check box for choosing one- or two-sided test and a graphical text box for showing always the actual sample size per group). While the majority of sample size determination problems involves the underlying distributions, the great variety of distribution functions in SAS® 6.11 (PROBIT, PROBNORM, PROBT, TINV, PROBF, FINV, TNONCT and FNONCT) helped much in simplifying the original algorithms (which used previously rather complicated approximations for these functions) and saved several steps and much time of SCL program writing. The results can be printed and each screen has a separate help (the help contains the definition and explanation of screen variables and the references concerning the respective model).

At present fourteen screens correspond to the fourteen different statistical methods contained in the application. New models and new screens can be easily added. They are classified by the type of statistical problem (hypothesis testing or estimation), by the number of groups involved and by the nature of the primary endpoint (continuous and normally distributed,

categorical, binary). They can be reached passing through a set of icons which has the following structure:



Theoretical background

The formulae which constitute the theoretical background of the program and the references:

1. Estimation of a population mean ([22]).

$$n = \frac{8\sigma^2 * z_{\alpha/2}^{\ 2}}{l^2}$$

where l=the length of the confidence interval

σ=the anticipated standard deviation

z_α=the 100αth upper percentage of the standard normal distribution

2. Estimation of a proportion (source: [25])

$$n = \frac{p\ *\left(1-p\ \right)* z_{\alpha/2}^{\ 2}}{(w/2)^2}$$

where p=the anticipated proportion

w=the width of interval

3. Comparative trials, normally distributed, two groups of equal size, related samples (two-sided) ([25]).

$$n = \frac{\sigma^2 * \left(z_{\alpha/2} + z_\beta\right)^2}{\delta^2}$$

where σ=the anticipated standard deviation

$\delta=\mu_1-\mu_2$ the least clinically significant difference between the means of the groups

β=the power of the test

4. Comparative trials, normally distributed, two groups of equal size, unrelated samples (two-sided) ([25]).

$$n = \frac{2 * \sigma^2 * \left(z_{\alpha/2} + z_\beta\right)^2}{\delta^2}$$

where σ=the anticipated standard deviation

δ=the least clinically significant difference

In case of unequal groups and if A is the allocation ratio (the ratio $\frac{size.of.group2}{size.of.group1} = \frac{n_2}{n_1} = A$):

$n_1 = \frac{A+1}{2A} n$ and $n_2 = \frac{A+1}{2} n$ [14].

5. Comparative trials, ordered categorical data, two groups, unrelated ([27])

The model is the so called "proportional odds" model. Let c_i denote the anticipated proportion in the *i*th category of the control group and let q_i denote the corresponding cumulative proportion. According to this model we suppose that the odds ratios between the cumulative proportions of the two groups are constant in each category. The equivalent for the least clinically significant difference will be the reference improvement, θ_R , which equals to the constant odds ratio mentioned before. We define the allocation ratio A which is the proportion of the sample sizes of the two groups. Then we can calculate using θ_R the anticipated proportions for all categories in the treatment group. Let p_i be the mean of the two proportions (i.e. the control group and the treatment group proportion in the *i*th category). The following formulas give the sample sizes in the two groups:

$$n_1 = \frac{3(A+1)\ (z_{\alpha/2} + z_\beta)^2}{A\theta_R^2(1 - \sum p_i^{\ 3})}$$

$$n_2 = \frac{3(A+1)\ (z_{\alpha/2} + z_\beta)^2}{\theta_R^2(1 - \sum p_i^{\ 3})}$$

6. Comparative trials, binomial data, two unrelated groups of equal size([23])

$$n = \frac{[z_{\alpha/2}\sqrt{2\overline{p}(1-\overline{p})} + z_\beta\sqrt{p_1 * (1-p_1) + p_2 * (1-p_2)}]^{\ 2}}{(p_2 - p_1)^2}$$

where p_1=the proportion anticipated in the *1*st group

p_2=the proportion anticipated in the *2*nd group

$\overline{p}$ =the mean of p_1 and p_2

In case of unequal groups and if A is the allocation ratio (the ratio $\frac{size.of.group2}{size.of.group1} = \frac{n_2}{n_1} = A$):

$n_1 = \frac{A+1}{2A} n$ and $n_2 = \frac{A+1}{2} n$ [19].

7. Survival data, two groups of equal size, up-front accrual ([26])

This is a model which supposes that the daily failure rate in each group is constant (exponential model) and that the period of the accrual is relatively short compared with the follow-up time ("up-front" accrual). Then the following relation determines the sample size per group:

$$\frac{B^2 + C^2}{n} = \frac{(Q-T)^2}{(z_{\alpha/2} + z_{\beta})^2}$$

where B and C may be calculated using the formulas:

$B^2 = Q^2(1 - e^{-Qt})^{-1}$

$C^2 = T^2(1 - e^{-Tt})^{-1}$

Q and T are the daily failure rates while t is the duration of the study (in days). The relation between the daily failure rates and the failure rates at the end of the trial (p_1, p_2) is given by the equation:

$p_1 = e^{-Qt}$

$p_2 = e^{-Tt}$

8. Normally distributed data, more than two parallel groups of equal size, unrelated (analysis of variance, completely random design) [14]

Let g denote the number of groups, μ_i the mean of the *i*th group and σ the anticipated standard deviation within each group. We define λ as:

$$\lambda = \frac{\sum(\mu_i - \bar{\mu})^2}{(g-1)\sigma^2}$$

Then the required sample size per group is such that the equation

$$z_\beta = \frac{1}{\sqrt{(g-1)(1+n\lambda)F + g(n-1)(1+2n\lambda)}} x\{\sqrt{g(n-1)[2(g-1)(1+n\lambda)^2 - (1+2n\lambda)]} - \\ -\sqrt{f(g-1)(1+n\lambda)[2g(n-1)-1]}\}$$

is approximately satisfied. So we increase *n* until the desired power is attained.

This method described by Fleiss [14] and Day [18] can be simplified in SAS because the value of noncentral F distribution with given noncentrality parameter δ can be calculated using the FINV and PROBF functions directly, without using the approximation shown above.

9. Normally distributed data, more than two parallel groups of equal size, unrelated, ordered responses [18]

The method described in 8. may be applied but the alternative hypothesis will no longer be that some of the group means are unequal. It will be that there is a certain (fixed) order between them. We also suppose that the class variable is quantitative, so a linear regression model may be applied. The formula from 8. is valid with $\lambda' = (g-1)*\lambda$ (where *g* is the number of groups) and with one degree of freedom for the sum of squares between groups.

10. The 2x2 factorial design

It is also based on the method described in 8. When applied for the difference beteween the means of the two factors the sample size obtained is the total for the two levels, so it has to be divided by 2 to obtain the sample size per group. When testing the interaction of the two factors, the estimate of the size of interaction can be obtained in the following way: we calculate for each of the factors the difference between the means of two levels and then we take the difference of the two results. The sample size per group obtained with formula 8

should be doubled because the interaction effect is determined by a contrast between four means, thus the standard error of the interaction is $\sigma\sqrt{4/n}$.

11. Categorical data, more than two parallel groups of equal size, unrelated (Kruskal-Wallis test) [14]

Introducing the concept of relative efficiency of two tests, RE , we can apply the formulas from point 8. using the relation $n=n^{*}.RE$, where n^{*} is the sample size for the Kruskal Wallis test, n is the sample size for the one-way ANOVA and the asymptotic efficiency of the Kruskal Wallis test related to the one-way ANOVA is always greater than $3/\pi$ (Hollander and Wolfe: Nonparametric statistical methods, New York, Wiley, 1972).

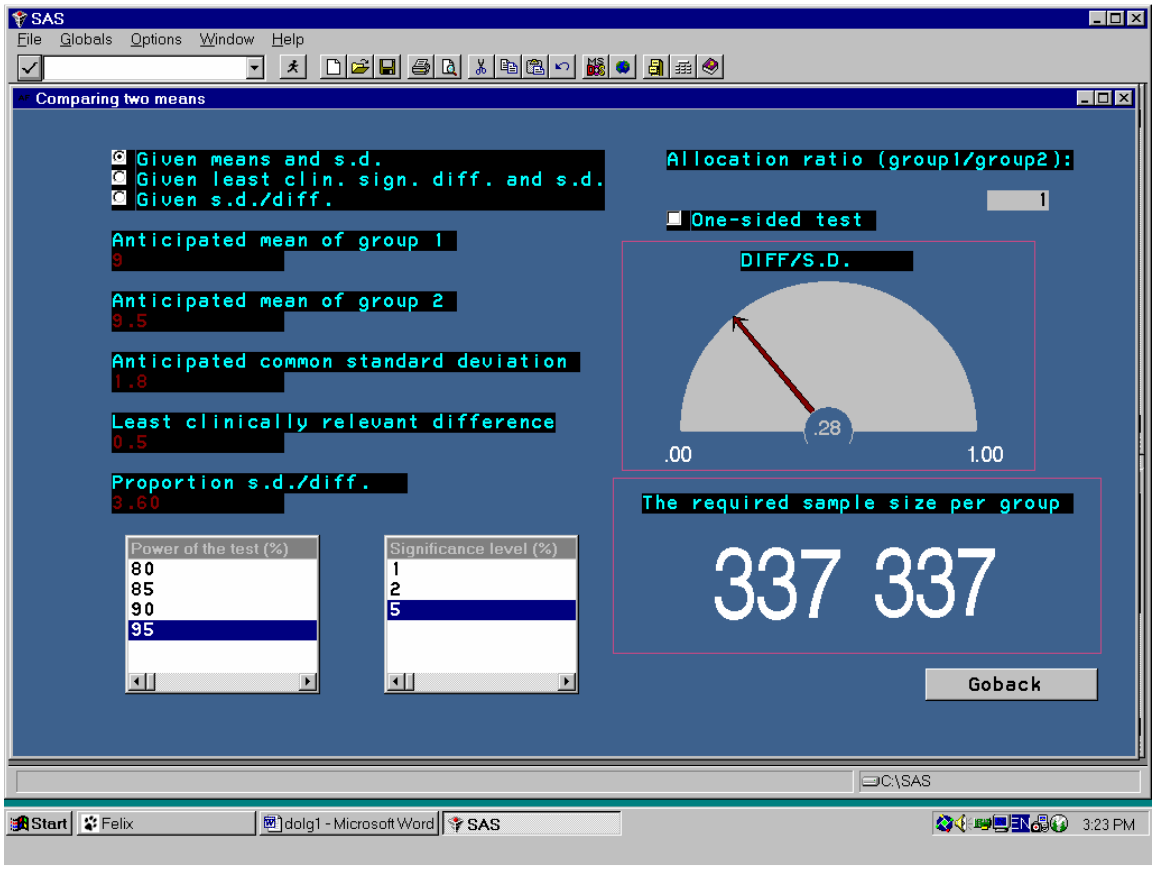
12. The equivalence trials ([23], [28], [29]) have the purpose of showing the "negative" result that two treatments are equally effective. These models use in fact the formulae already described as estimations of population means or proportions with given length of the confidence interval, applied for the difference between the two groups. The terminology used is slightly modified: the limits within which the confidence limits has to lie are called equivalence acceptance limits.

Worked examples

1. Vitamin D trial (Pocock, 1983)

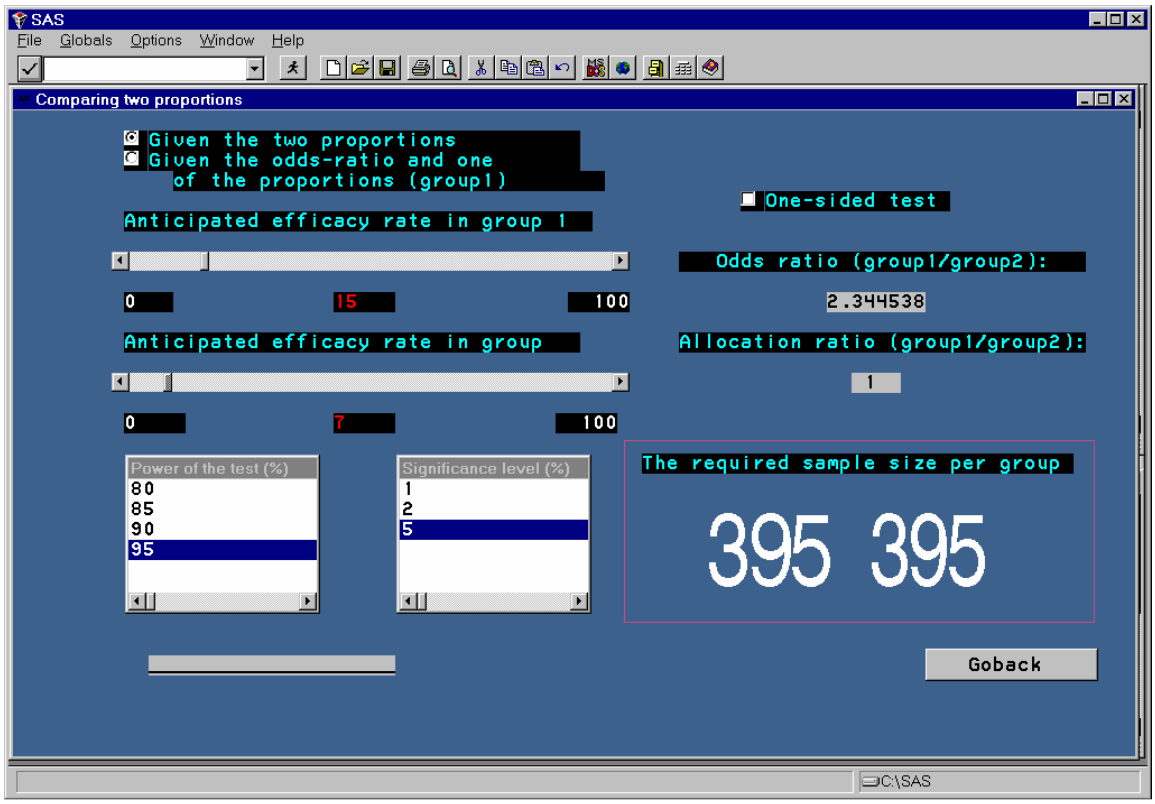
The effect of Vitamin D on the prevention of neonatal hypocalcaemia is going to be examined on pregnant women. Vitamin D group is compared to placebo group and women will be randomized to one of these groups. The main efficacy parameter is the infant's serum calcium level one week after the birth. Previous knowledge about untreated women shows that the serum calcium level has μ_1=9.0 mg/100ml and σ=1.8 mg/100ml. We consider an increase of 0.5 mg/100ml to be clinically relevant and we choose a significance level of 0.05 and a power of 0.95. Then we apply the method described in model 4. and we obtain n=337 patients per group (Fig.6.1).

Figure 6.1. *The screen for comparing two independent means*



Sometimes the value of the standard deviation is difficult to anticipate. In this case these data can be assessed in an other way applying model 6. However, dichotomizing a continuous variable means always a loss in power [30]. Let's fix the criterion of hypocalcaemia to be a serum calcium level less than 7.4 mg/100ml. The proportion of hypocalcaemic newborns in the placebo group is approximately 15% and we anticipate a decrease of 8%. Keeping the significance level and power unchanged we obtain a number of 395 patients per group (Fig.6.2).

Figure 6.2 *The screen for comparing two independent proportions*



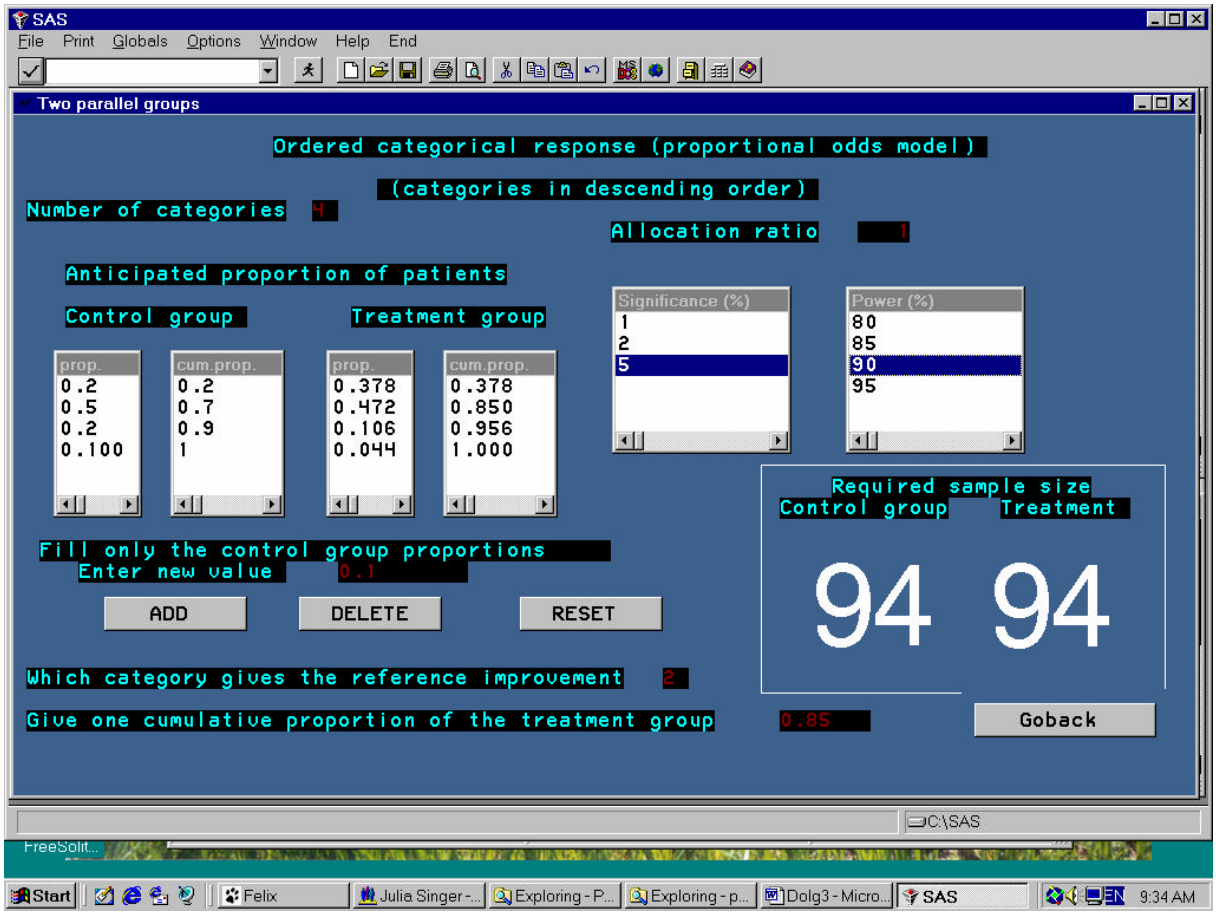
2. A trial with categorical outcome variable (Whitehead, 1993)

In a placebo-controlled, three-month study the condition of patients is classified at the end of study to be very good, good, moderate or poor. In the placebo group the four proportions corresponding to these categories are 0.2, 0.5, 0.2 and 0.1. Thus the probability of a good or very good outcome is 0.2+0.5=0.7 (these are the so-called cumulative proportions and at this

point the ordering between categories becomes necessary). As a clinically significant improvement we have to choose one of the categories and to establish the respective cumulative proportion in the treated group. Let's say that we anticipate an increase of 0.15 in the "good or very good" category, that is, a cumulative proportion of 0.85. In this case applying the proportional odds model 5. for equal group numbers (significance level: 5%, power: 90%) the result is 94 patients/group.

The example above is a refinement of grouping patients into two categories, for instance unifying category 1 with 2 as "good" and 3 with 4 as "poor". If we do this, we can apply model 6 with proportion 0.7 increasing to 0.85. Keeping the significance level of 5% and the power of 90% unchanged the result is a group size of 161 (each group representing two from the former categories, that would mean 81 patients/category for the former example).

Figure 6.3 *Categtorical outcome*



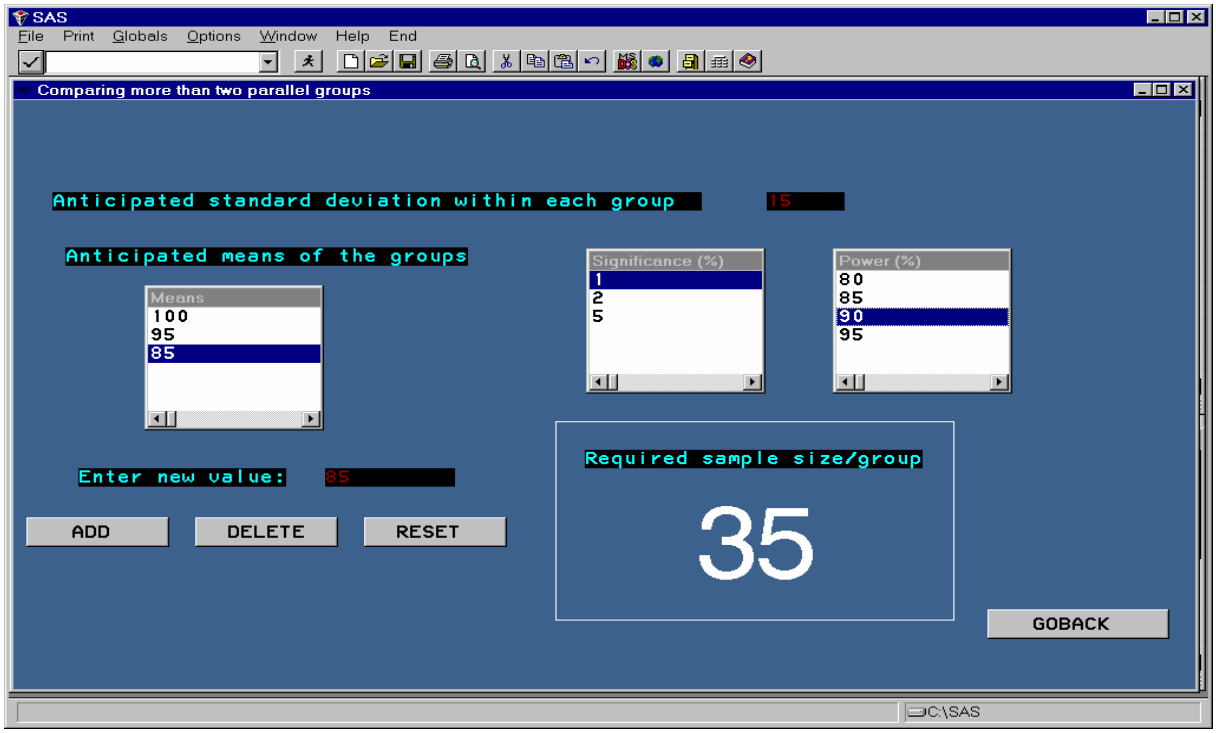
3. Survival Analysis (Shuster, 1990)

The primary endpoint of a comparative, two-group clinical trial is the time until the patient needs to take sedatives regularly. The duration of the study is one year. At the end of the trial we presume a survival rate of 10% in the placebo group and of 25% in the treated group. Considering a significance level of 2% (two-sided test) and a power of 80% we shall need 102 patients per group to prove the effect (if it exists) of the new drug, applying model 7 . This problem can also be solved like an ordered categorical one (model 5), in this case the solution is a patient number of 116, or like a simple comparison of proportions (model 4) which gives the result of 128.

4. A three-group comparison (Day and Graham, 1991.)

The diastolic blood pressure of three patient groups is to be compared. The hypothesized means are 100 mmHg, 95 mmHg and 85 mmHg. The common standard deviation of the groups is supposed to be 15 mmHg. The significance level is set to 1% and the power to 90%. Applying model 8 the sample size per group is 35 .

Figure 6.4 *Comparison of more than two groups (continuous outcome, equal group sizes)*



If the same groups are compared but the alternative hypothesis is changed from “some of the group means are unequal” to “there is a certain order between the group means”, we can decrease this number from 35 to 31.

In case of two groups this method gives the same result as 4.

All the examples above show that sample sizes depend much on the model applied, so it has to be chosen carefully. This application does not contain the model described in Chapter 2, the comparison of two groups assuming unequal variances, since this method was elaborated later. As mentioned, new models can be easily added to the program, this is among our further plans.

The modules used were SAS/BASE®, SAS/AF® and SAS/STAT®.

7. References

[1] Freiman, JA, Chalmers, TC, Smith, H, Kuebler, RR: 'The importance of beta, the type II error, and sample size in the design and interpretation of the randomized clinical trial', New England Journal of Medicine, **299,** 690-694 (1979). [Updated and reprinted as Chapter 19 in Bailar, JC and Mostseller, F. *Medical uses of statistics*. NEJM Books, Boston, 1992].

[2] Sedlemeier P and Gigerenzer G: 'Do studies of statistical power have an effect on the power of studies?', Psychological Bulletin, **105**, 309-316, 1989.

[3] Singer J: A simple procedure to compute the sample size needed to compare two independent groups when the population variances are unequal. *Statistics in Medicine,* **20,** 1089-1095, 2001.

[4] Singer J: Estimating sample size for continuous outcomes, comparing more than two parallel groups with unequal sizes. *Statistics in Medicine,* **16**, 2805-2811, 1997.

[5] Singer J: A SAS application for trial sample size determination. In: *Proceedings of the SAS European User's Group International*, Hamburg, Germany, 1996.

[6] O'Brien, RG, Muller, KE: *Unified Power Analysis for t-tests through Multivariate Hypotheses*. In: Applied Analysis of Variance in the Behavioral Sciences, Marcel Dekker, New York, 1993. Chapter 8.

[7] Satterthwaite, FE: 'An approximate distribution of estimates of variance components', Biometrics Bulletin, **2**, 110-114 (1946).

[8] Snedecor, GW and Cochran, WG: *Statistical Methods*, The Iowa State University Press, Iowa, 1979. Sixth Edition, Tenth printing, Section 4.14.

[9] Machin, D; Campbell, MJ; Fayers, PM and Pinol, APY: *Sample Size Tables for Clinical Studies*, Blackwell Science, Oxford, 1997. Second edition.

[10] Guenther, WC: 'Sample size formulas for normal theory t-tests', American Statistician, **35**, 243-244 (1981).

[11] Schouten, HJA: 'Sample size formula with a continuous outcome for unequal group sizes and unequal variances', Statistics in Medicine, **18**, 87-91 (1999).

[12] *SAS Language. Version 6. Reference*. First edition. SAS Institute Inc., Cary, NC, 1990.

[13] Welch, BL: 'The Generalisation of Student's Problems when Several Different Population Variances are Involved', Biometrika, **34**, 28-35 (1947)

[14] Fleiss, J. L. *The Design and Analysis of Clinical Experiments*. Wiley, New York, 1986.

[15] Scheffé, H. *The Analysis of Variance*. Wiley, New York, 1959.

[16] Laubscher, N. F. 'Normalizing the noncentral t and F distributions'. *Annals of Mathematical Statistics.* 31., 1105-1112 (1960).

[17] Finney, D.J. *Statistical Method in Biological Assay*. Charles Griffin & Company Ltd, London, 1964.

[18] Day, S.J. and Graham, D.F. 'Sample Size Estimation for Comparing Two or More Treatment Groups in Clinical Trials'. *Statistics in Medicine*. 10, 33-43 (1991).

[19] Campbell, M.J., Julious, S.A., Altman, D.G. 'Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons'. *British Medical Journal.* 311, 1145-1148 (1995).

[20] Efron B, Tibshirani RJ : An introduction to the bootstrap. *Chapman & Hall*, New York, 1993.

[21] Moyé, L.A. 'Sizing clinical trials with variable endpoint event rates'. *Statistics in Medicine,* 16, 2267-2282 (1997).

[22] Armitage P.: *Statistical Methods in Medical Research*. pp. 184-188. Blackwell, 1971.

[23] Pocock SJ: *Clinical Trials*. pp. 123-130, Wiley, 1983.

[24] Noether GE: *Sample Size Determination for Some Common Nonparametric Tests. Journal of the American Statistical Association*. Vol.82, pp. 645-647. 1987.

[25] Bolton S: *Pharmaceutical Statistics*. pp. 187-202. 2nd ed. Marcel Dekker, Inc. ,1990.

[26] Shuster JJ: *Handbook of Sample Size Guidelines for Clinical Trials*. pp. 51-130. CRC Press, 1990.

[27] Whitehead J: *Sample Size Calculations for Ordered Categorical data. Statistics in Medicine*, Vol. 12, pp.2257-2271, 1993.

[28] Tie-Hua NG: *A Specification of Treatment Difference in the Design of Clinical Trials with Active Controls*. Drug Information Journal, Vol.27. pp.705-719. 1993.

[29] Lin SC: *Sample Size for Therapeutic Equivalence Based on Confidence Interval*. Drug Information Journal, Vol.29, pp.45-50, 1995.

[30] Deyi BA, Kosinski AS, Snapinn SM 'Power considerations when a continuous outcome variable is dichotomized'. *Journal of Biopharmaceutical Statistics*, Vol. 8, pp. 337-352, 1998.

[31] Taylor DJ, Muller KE. 'Bias in linear model power and sample size calculation due to estimating noncentrality'. Commun. Statist. - Theory Meth., 25(7), pp. 1595-1610 (1996)

[32] Dudewicz, EJ: 'Confidence Intervals for Power, with Special Reference to Medical Trials', *Austral. J. Statist.*, **14**, 211-216 (1972)

[33] Joseph, LR, Berger, R du, Bélisle, P: 'Bayesian and Mixed Bayesian/Likelihood Criteria for Sample Size Determination', *Statistics in Medicine*, **16**, 769-782 (1997)

[34] Sandvik, L.; Erikssen, J.; Mowinckel, P. and Rødland, E. A. 'A method for determining the size of internal pilot studies', *Statistics in Medicine,* **15**, 1587-1590 (1996).

[35] Birkett, M. A. and Day, S. J. 'Internal pilot studies for estimating sample size', *Statistics in Medicine,* **13**, 2455-2463 (1994).

[36] Singer J: A method for determining the size of internal pilot studies. *Statistics in Medicine,* **18**, 1151-1153, 1999.

8. Appendix

SAS programs

8.1 Macro computing the group sizes when the group variances are unequal

```
/* given sd1,sd2,d,alfa,beta,allocation ratio */
%macro size(sd1,sd2,d,alfa,beta,ar);

option mprint;
data result;
  diff=&d;
  s1=&sd1;
  s2=&sd2;
  tau=s2*s2/s1/s1;
  a=1/&ar;
  aa=(100-&alfa/2)/100;
  ba=&beta/100;
  if a<1 then
        b=ceil(1/a)+1;
  else b=3;
  do i=b to 10000;
     teta=s1*s1/s2/s2;
     df=(teta/a+1)*(teta/a+1)/(teta*teta/a/a/(a*i-1)+1/(i-1));
     t=tinv(aa,df)+tinv(ba,df);
     s=s1*s1/a+s2*s2;
     fract=t*t*s/diff/diff;
     if (i>=fract) then do;
            m1=(a+1)*i/2;
            m2=ceil(m1/a);
            m1=ceil(m1);
            output;
        stop;
     end;
  end;
run;

data _null_;
  file print;
  set result;
  put 'tau=' tau;
  put 'allocation ratio (n2/n1) = ' a 5.3;
  put 'Size of group1:' m1;
  put 'Size of group2:' m2;
  s=m1+m2;
  put 'N=' s;
run;

%mend size;
```

8.2 Program computing the group sizes for more than two groups with unequal sizes

```
data _null_;
    array m[25];          /* group means */
    array r[25];          /* set of allocation ratios */
    array sz[25];         /* nuisance variables */
    array nnn[25];        /* group sizes */
    m[1]=100;
    m[2]=95;
    m[3]=85;
    m[4]=.;
    r[1]=1.5;
    r[2]=1.5;
    r[3]=1;
    rr=r[1]+r[2]+r[3];
    s=15;
    file print;
    alfa=0.99;
    beta=0.9;
    do i=1 to 25;
       sz[i]=r[i]*m[i];
    end;
    zlimit=probit(beta);

mean=sum(sz[1],sz[2],m[3],m[4],m[5],m[6],m[7],m[8],m[9],m[10],m[11],m[12],
m[13],m[14],m[15],m[16],m[17],m[18],m[19],m[20],m[21],m[22],m[23],m[24],
m[25])/rr   ;

n=n(m[1],m[2],m[3],m[4],m[5],m[6],m[7],m[8],m[9],m[10],m[11],m[12],m[13],
m[14],m[15],m[16],m[17],m[18],m[19],m[20],m[21],m[22],m[23],m[24],m[25]);

 stop=0;
 n1=n-1;
 nn=3;
 do while (stop=0);
     do i=1 to 25;
       nnn[i]=nn*r[i];
     end;

n2=sum(nnn[1],nnn[2],nnn[3],nnn[4],nnn[5],nnn[6],nnn[7],nnn[8],nnn[9],nnn[1
0],nnn[11],nnn[12],nnn[13],nnn[14],nnn[15],nnn[16],nnn[17],nnn[18],
nnn[19],nnn[20],nnn[21],nnn[22],nnn[23],nnn[24],nnn[25])-n;

     delta=0;
     do i=1 to n;
        delta=delta+nnn[i]*(m[i]-mean)*(m[i]-mean)/s/s;  /* noncentralitasi
parameter */
     end;
     f=finv(alfa,n1,n2);
     b=probf(f,n1,n2,delta);
     if b>1-beta then
             nn=nn+1;
           else do;
             stop=1;
             size1=ceil(nn*r[1]);
             size2=ceil(nn*r[2]);
             size3=nn*r[3];
             put n1 ' , ' n2;
             put 'b=' b;
             delt=sqrt(delta);
             put 'delta= ' delt;
             put 'mean= ' mean;
             put 'allocation ratios' r[1]' , ' r[2] ' , ' r[3];
             put 'number of patients in group1 = ' size1;
```

```
            put 'number of patients in group2 = ' size2;
            put 'number of patients in group2 = ' size3;
         end;
     if (nn>5000) then do;
            stop=1;
            put 'number of patients/group > 5000' ;
     end;

 end;

run;
```

8.3 Program simulating 2000 pilot studies and estimating the sample size and its confidence interval by parametric bootstrap

```
%macro invoke;

%macro boot(pilot);

data a;
    seed=&pilot+3;
    do i=1 to 5;
      call rannor(seed,n);
      value=n*10+10;
      group=1;
      output;
    end;
    do i=5 to 10;
      call rannor(seed,n);
      value=n*10+18;
      group=2;
      output;
    end;
run;

proc univariate data=a noprint;
  by group;
  var value;
  output out=ki
    std=var;
run;

data _null_;
   set ki;
   if group=1 then call symput('var1',left(put(var,6.3)));
   if group=2 then call symput('var2',left(put(var,6.3)));
run;

data d;
  seed=1225;
  do j=1 to 2000;
    do i=1 to 5;
      call rannor(seed,n);
      value=n*&var1+10;
      group=1;
      output;
    end;
    do i=6 to 10;
      call rannor(seed,n);
      value=n*&var2+18;
      group=2;
      output;
    end;
  end;
run;

proc univariate data=d noprint;
  by j group;
  var value;
  output out=ki
    var=var;
run;
```

```
   data b(rename=(var=var1));
      set ki;
      where group=1;
   run;

   data c(rename=(var=var2));
      set ki;
      where group=2;
   run;

   data bc;
      merge b c;
      by j;
   run;

   data abc;
      set bc;
      d=8;
      z=probit(0.975)+probit(0.8);
      z=z*z;
      s2=(var1+var2)/2;
      n=ceil(2*z*s2/d/d+probit(0.975)/4);
      output;
    run;

    proc univariate data=abc noprint;
      var n;
      output out=kiki
        pctlpts=70 80 90
        pctlpre=p_
        ;
    run;

    data a.egybe;
      set a.egybe kiki;
    run;

  %mend boot;

  data a.egybe; run;

  %do pilot=1 %to 2000;

     %boot(&pilot);

     dm log 'clear';

  %end;

%mend invoke;

libname a 'd:\cikk1\';

%invoke;

proc means data=a.egybe n;
   where p_70>=24;
run;

proc means data=a.egybe n;
   where p_80>=24;
run;

proc means data=a.egybe n;
   where p_90>=24;
run;
```