Published in: *The American Statistician*, 53, 52-55 (1999).

Searching for the Right Sample Size

Dhammika Amaratunga
The R W Johnson Pharmaceutical Research Institute, Raritan, NJ 08869-0602.

Dhammika Amaratunga is Research Fellow, Preclinical Biostatistics, The R W Johnson Pharmaceutical Research Institute, P. O. Box 300, Route 202, Raritan, NJ 08869-0602. Email: damaratu@prius.jnj.com. The author thanks Gordon Pledger, Allan Sampson, and the referees for constructive suggestions that led to a much improved manuscript.

ABSTRACT

Statisticians are often asked to recommend an appropriate sample size for an experiment. In many cases, it can be quite difficult to derive a suitable formula to address this problem. An automated search procedure that uses computer simulation is suggested as an alternative versatile approach. A key component of this procedure is an algorithm that accelerates the simulation and thereby greatly reduces the otherwise huge computational burden.

KEYWORDS: simulation, search algorithm, stochastic approximation, QuickSize, Fisher's exact test, logrank test, bioequivalence.

1. INTRODUCTION

A question frequently asked of consulting statisticians is "how large a sample should I use in this experiment I am planning to do?" Let us assume that the experiment is being performed to test a null hypothesis H_0 and that the sample size is to be determined so that the test has a stated amount of power at a specified alternative hypothesis H_a. Then, to answer the sample size question, one has to study the power of the test at H_a as a function of sample size and then solve, for sample size, the stochastic equation that sets the power function equal to the required power.

For a few cases, an exact solution to this equation can be determined. For most cases, however, only an approximate solution, at best, can be found. Often the difficulty is that the distribution, F_a, of the test statistic under H_a tends to be one that is hard to work with.

There is a vast literature providing exact and approximate sample size formulas or power functions for specific problems. A few general references are Cohen (1987), Desu and Raghavarao (1990), Lachin (1981), and Pearson and Hartley (1951). Software packages that implement such methods for commonly encountered situations are available (STPLAN, SOLO, SPSS SamplePower, PASS 6.0, nQuery Advisor, to name a few).

Given the general difficulty of working out an exact answer or a good approximation algebraically, it would be useful, particularly when faced with a new or nonstandard situation, to have an alternative approach that is simple to use yet versatile enough to give an exact solution for a broad range of problems. We show how, with the current availability of inexpensive high speed computing, simulation, augmented by a search algorithm called *QuickSize*, is just such a tool.

2. SAMPLE SIZE SELECTION BY SIMULATION

Although the procedure can handle more general situations, we shall, for ease of exposition, describe it in the simple hypothesis testing framework, in which it is planned to test H_0:θ=θ_0 vs H_1:θ≠θ_0 where θ is a univariate parameter of interest, with a sample of size *N*, that is just large enough to pick up a difference in θ of Δ, with power of at least 1-β using an α-level test. The required sample size, $N=N(\Delta, \theta_0, \xi, \alpha, \beta)$, is the smallest integer *n* satisfying $\pi(n\,/\,\Delta, \theta_0, \xi, \alpha)\geq 1-\beta$, where π is the power function of the test and ξ denotes a vector of nuisance parameters. We assume that the mechanism that would produce the sample data is known once the values of Δ and ξ are given, as this allows one to draw a *pseudosample* from a population with parameters Δ and ξ. The idea is to use such pseudosampling to avoid having to deal with a complicated F_a.

There is an obvious "brute force" pseudosampling-based simulation approach. Choose a range of values within which *N* is expected to lie. For each value of *n* in this range, randomly generate a large number of pseudosamples of size *n* based on Δ and ξ; the proportion, $P(n\,/\,\Delta, \theta_0, \xi, \alpha)$, of these pseudosamples in which a difference is detected at level α, is then an estimate of $\pi(n\,/\,\Delta, \theta_0, \xi, \alpha)$. Choose as the required sample size, *N*, the *n* corresponding to $P(n\,/\,\Delta, \theta_0, \xi, \alpha)$ that is equal to or just above $1-\beta$.

This approach is not very appealing, as in practice it would typically involve much trial-and-error and heavy computing. However, both the trial-and-error component and the computational effort can be considerably reduced by "automating" and "accelerating" the search for *N* by means of a suitable algorithm. Since $\pi(n\,/\,\Delta, \theta_0, \xi, \alpha)$ is a bounded monotonically increasing function of *n,* we suggest an algorithm we call *QuickSize*, inspired by sequential ED50 estimation procedures, such as the up-down method of Dixon and Mood (1948) and the stochastic approximation algorithm of Robbins and Monro (1951). Govindarajulu (1988), Morgan (1992) and other books on binary and bioassay data give general overviews of these methods; Garthwaite and Buckland (1992) and Garthwaite (1996) give recent novel applications of the Robbins-Monro process.

QuickSize works by "searching" through different values of *n*. To start the search, one specifies a starting value, n_0, for *n*. Then, at the *i*-th ($1\leq i\leq I$) iteration:

- Let n_i be the "current" value of *n*. Randomly generate *B* pseudosamples of size n_i each and count how many of the *B* pseudosamples lead to rejection of H_0.

- Let $B_i(n_i)$ be the total number of pseudosamples of the same size n_i drawn in all iterations up to and including the *i*th iteration. Suppose that $B_i{*}(n_i)$ of them led to rejection of H_0. At this stage, $P_i{*}(n_i)=B_i{*}(n_i)/B_i(n_i)$ serves as the "current" estimate of $\pi(n_i\,/\,\Delta, \theta_0, \xi, \alpha)$. If $P_i{*}(n_i)<(1-\beta)$, set $n_{i+1}=n_i+m_i$; else if $P_i{*}(n_i)>(1-\beta)$, set $n_{i+1}=n_i-m_i$; else if $P_i{*}(n_i)=(1-\beta)$, set $n_{i+1}=n_i$. Proceed to the $(i+1)$th iteration.

At convergence, the iteration will tend to oscillate between two consecutive values of *n*, corresponding to powers just below and just above $1-\beta$. The required sample size, *N*, is the larger of the two. The appropriateness of this sample size can be simply verified by means of a separate *confirmatory simulation*, in which two large independent sets of pseudosamples of sizes $N-1$ and N respectively are generated, and it is checked that $P{*}(N-1)$ and $P{*}(N)$ bracket $1-\beta$ as required.

Although at first the process may look rather complicated, it is actually quite simple and is easy to program. Example 1 below shows the first few steps of the iterative process in the context of an example. For a potential sample size *n*, let $B_+(n)$ denote the total number of pseudosamples of size *n* that were drawn over the *I* iterations; let $B_+{*}(n)$ denote the number of these $B_+(n)$ pseudosamples that led to H_0 being rejected; and let $P_+{*}(n)=B_+{*}(n)/B_+(n)$.

Implementation of QuickSize requires specification of n_0, m_i, ξ, B, and I. The value of n_0 can be guided by approximation or, if none is available, by a few trial-and-error runs. We usually set m_i=1, although, without a good n_0 and when the power function does not have big jumps, somewhat faster convergence may be achieved by setting $m_i=1+[c(|P_i^*(n_i)-(1-\beta)|)]$ with a suitable c. The value of ξ, as in any sample size selection situation, would be based on pilot or historical information. We have found B=10 and I=500 to be generally adequate for the problems we have encountered, as long as we were able to start at a fairly decent n_0. Occasionally, we may redo the simulation with a fresh start, trying to reach, as a rough guideline, a $B_+(n)$ of at least 1500 at N and $N-1$ for 80% power and of at least 900 for 90% power (since about 1500 pseudosamples are needed to estimate a π of 0.80 to $\pm$ 0.02 with 95% confidence and about 900 pseudosamples are needed to estimate a π of 0.90 to $\pm$ 0.02 with 95% confidence).

Modifications to this basic procedure, such as setting a more formal stopping rule or those of Wetherill (1963) for the up-down method, did not offer significant improvement but affected the simplicity of the procedure.

We make no optimality claims on the process, beyond emphasizing that, in many instances, it significantly accelerates the search. Another plus is that it is quite simple to program, and once a macro has been set up, it can be used for many different problems with only changes to the pseudosampling and hypothesis testing steps. This can be of much value in a consulting environment where there is demand for a quick answer. Even if an algebraic solution to a problem exists, one may not have time to seek it out or work it out! In such an event, QuickSize can be invaluable.

3. APPLICATIONS

We now demonstrate how QuickSize could be applied in practice.

Example 1: *Comparing two response rates via Fisher's Exact Test*

At the planning stage of a cancer clinical trial, it is known that roughly 10% of all patients respond to standard therapy and it is expected that over 50% of patients would respond to a new experimental therapy. It is required to determine the number n of patients that should be assigned to each treatment for a 5% level one-sided Fisher's Exact Test to detect a difference at least as large as this with 80% or more power.

To get an exact answer, one has to study the distribution at the alternative, a complex computation involving hypergeometric probabilities (Bennett and Hsu (1960), Gail and Gart (1973)). Hence this problem is usually addressed in practice by using an approximation such as that proposed by Casagrande, Pike, and Smith (1978), although it is thought to be rather conservative. This approximation suggests N=20, which can be used as n_0 for the simulation. Each pseudosample consists of two independent samples, $X_r \sim Binomial(n, p_r)$, $r=1,2$, with p_1=0.1, p_2=0.5. QuickSize was applied with B=10, I=500, m_i=1. The first few iterations proceeded as follows:

i	n_i	P_{obs}[a]	$B_i^*/B_i=P_i$	*Action*
1	20	9/10	9/10=0.90	decrease n_i
2	19	8/10	8/10=0.80	do not change n_i
3	19	9/10	17/20=0.85[b]	decrease n_i
4	18	9/10	9/10=0.90	decrease n_i
5	17	7/10	7/10=0.70	increase n_i
6	18	6/10	15/20=0.75	increase n_i
7	19	9/10	26/30=0.87	decrease n_i

Notes: (a) P_{obs} is the proportion of B=10 pseudosamples at the *i*th iteration that led to rejection of H_0. (b) This is the 9/10 from the 3rd iteration combined with the 8/10 from the 2nd iteration as they both correspond to *n*=19.

After *I*=500 iterations, the situation was as follows:

n	$B_+*(n)$	$B_+(n)$	$P_+*(n)$
17	370	500	0.740
18	1959	2490	0.787
19	1643	2000	0.822
20	9	10	0.900

In this case, convergence was remarkably fast! The required sample size is identified as *N* =19 (also obtained by Gail and Gart (1973) and Haseman (1978) using probability calculations). With 2000 or more pseudosamples already at both *n*=18 and 19, a separate confirmatory simulation was deemed unnecessary.

It was not really necessary to use QuickSize for this example since exact and approximate solutions are available. However, with the algebraic approach, altering the problem slightly could mean having to completely redo a very difficult derivation, while QuickSize will only need a small change. For instance, if the control group were to be double the size of the test group, rerunning QuickSize, with a trivial modification to the code, indicates that the required total exact sample size is 51 (=34+17).

Example 2: *Comparing two survival distributions via a logrank test*

The times-to-event of an experimental treatment group are to be compared with those of a control group. Instead of a parametric distributional assumption, a weaker assumption, that the ratio of the hazard functions is constant, i.e., $\lambda_T(t)/\lambda_0(t)=\theta$, is made, thereby enabling the groups to be compared via a logrank test. Although the event here is not death, the terminology is simplified by using survival analysis terms.

A control survival distribution, $S_0(t)=P(T\geq t)$, has been generated from the "survival" times of a cohort of 128 untreated patients. Since we do not ascribe a parametric model to this data, $S_0(t)$ will remain piecewise linear. This data indicates that the median survival time for untreated patients is around 65 days. The experimental treatment would be considered efficacious if this were at least doubled, so that here $\theta_0=1$ (since the null hypothesis is that the hazard functions are equal) and $\Delta=0.5$ (since this leads to a doubling of the median as required). It is planned to test this, using a 5% level logrank test with 90% power at the desired alternative, in a clinical trial with *n* patients in each group.

The survival distribution for the experimental treatment group is $S_T(t)=S_0(t)^\theta$. A pseudosample consists of two random samples of size *n* each drawn from the two survival distributions, $S_T(t)$ (with θ=0.5 since Δ=0.5) and $S_0(t)$ (Rubinstein (1991) describes how to sample from piecewise linear distributions). Patient accrual information can be incorporated into the pseudosampling process but was not done here.

This is an ideal situation in which to apply QuickSize as it allows a sample size evaluation to be made without a parametric assumption about $S_0(t)$ and $S_T(t)$. The most common approach is to assume that they are exponential, since survival distributions are often close to this form, and to calculate a sample size under this assumption (George and Desu (1974), Lachin (1981)); here it yields 44. QuickSize with n_0=44, *B*=10, *I*=1000, m_i=1 gives *N*=67. This discrepancy is not too surprising as $S_0(t)$ looked far from exponential.

Example 3: Comparing two measuring devices via an F-test

A new measurement device is to be compared to an old one by measuring n items with each. Bradley and Blackwood (1989) and Blackwood and Bradley (1991) model Y_{rs}, the observed measurement when the sth item is measured with the rth device, as $Y_{rs} = X_s + \varepsilon_{rs}$, where $X_s \sim N(\mu_X, \sigma_X^2)$, $\varepsilon_{rs} \sim N(\alpha_r, \sigma_r^2)$, $\Sigma\alpha_r=0$, $r=1,2$, $s=1,...,n$. They show that the hypothesis of no difference in devices, H$_0$: $\alpha_1=\alpha_2$, $\sigma_1^2=\sigma_2^2$, can be tested via the F-test for a significant regression (significant intercept and slope) of $D_r=Y_{r1}-Y_{r2}$ on $S_r=Y_{r1}+Y_{r2}$.

Prior experience suggests $\mu_1=\mu_X+\alpha_1=10$, $\sigma_X=2$, $\sigma_1=0.5$. It is required to determine how large n should be to detect a unit increase in bias ($\mu_2=\mu_X+\alpha_2=11$) and a twofold increase in standard deviation ($\sigma_2=1$) using a 5% test with 80% power.

This again is a situation where QuickSize comes in useful since to determine a sample size algebraically for this problem would be very difficult. Setting $\mu_X=10$, $\alpha_1=0$, $\alpha_2=1$, pseudosamples are generated by first generating $\{X_s\}$ and $\{\varepsilon_{rs}\}$, then summing them appropriately to get $\{Y_{rs}\}$. QuickSize with B=10, I=500, m_i=1, n_0=10 found the required sample size to be N=15.

Example 4: Testing equivalence via the two one-sided tests procedure

Chow and Liu (1992) give an example in which the bioequivalence of a reference (R) and a test (T) pharmaceutical formulation is to be tested via the two one-sided tests procedure (Schuirmann (1987)). Here, $H_0:|\mu_T-\mu_R|\geq \theta_0$ vs $H_1:|\mu_T-\mu_R|\leq\theta_0$ where θ_0 is typically set to $0.20\mu_R$. In this example, μ_R was set to 82.559 and it was desired to have 80% power at $0.05\mu_R$ with a 5% level test. The variance, σ^2, was set to 83.623. Chow and Liu's approach yielded N=8, which we used as n_0=8. We again set B=10, I=500, m_i=1, pseudosampled the sufficient statistics, $\bar{X}_R \sim N(\mu_R, \sigma^2/n)$, $\bar{X}_T \sim N(\mu_R, \sigma^2/n)$, and $s^2 \sim (\sigma^2/(2n-2))\chi_{2n-2}^2$, and proceeded to use QuickSize; N=8 was confirmed as the required sample size.

Pseudosampling the sufficient statistics, as we have done here, helps reduce the required computational effort; this is particularly useful when large pseudosamples have to be drawn.

4. CONCLUDING REMARKS

QuickSize is a valuable and versatile simulation-based easy-to-use tool for determining an exact sample size in essentially any hypothesis testing situation. It can also be modified for use in confidence interval situations. The computing requirements are modest by today's standards; all the computing for this paper was done using SPlus (version 4) on a 100 MHz Pentium running Windows 95. Note: an SPlus function that implements QuickSize will be placed in STATLIB or is available from the author.

REFERENCES

Bennett, B.M. and Hsu, P. (1960), On the power function of the exact test for the 2x2 contingency table, *Biometrika*, 47, 393-398.

Blackwood, L.G. and Bradley, E.L. (1991), An omnibus test for comparing two measuring devices, *Journal of Quality Technology*, 23, 12-16.

Bradley, E.L. and Blackwood, L.G. (1989), Comparing paired data: a simultaneous test for means and variances, *American Statistician*, 43, 234-235.

Casagrande, J.T., Pike, M.C. and Smith P.G. (1978), An improved approximation for calculating sample sizes for comparing two binomial distributions, *Biometrics*, 34, 483-486.

Chow, S.C. and Liu, J.P. (1992), *Design and Analysis of Bioavailability and Bioequivalence Studies*, New York: Marcel Dekker.

Cohen, J. (1987), *Statistical Power Analysis*, 2nd edition, Hillsdale(NJ):Lawrence Erlbaum Associates, Inc.

Desu, M.M. and Raghavarao, D. (1990), *Sample Size Methodology*, New York:Academic Press.

Dixon, W.J. and Mood, A.M. (1948), A method for obtaining and analyzing sensitivity data, *Journal of the American Statistical Association*, 43, 109-126.

Gail, M. and Gart, J.J. (1973), The determination of sample sizes for use with the exact conditional test in 2x2 comparative trials, *Biometrics*, 29, 441-448.

Garthwaite, P.H. (1996), Confidence intervals from randomization tests, *Biometrics,* 52, 1387-1393.

Garthwaite, P.H. and Buckland, S.T. (1992), Generating Monte Carlo confidence intervals by the Robbins-Monro process, *Applied Statistics,* 41, 159-171.

George, S.L. and Desu, M.M. (1974), Planning the size and duration of a clinical trial studying the time to some critical event, *Journal of Chronic Diseases*, 27, 15-29.

Govindarajulu, Z. (1988), *Statistical Techniques in Bioassay*, New York: Karger.

Haseman, J.K. (1978), Exact sample sizes for use with the Fisher-Irwin test for 2x2 tables, *Biometrics*, 34, 106-109.

Lachin, J. M. (1981), Introduction to sample size determination and power analysis for clinical trials, *Controlled Clinical Trials*, 2, 93-113.

Morgan, B.J.T. (1992), *Analysis of Quantal Response Data*, New York: Chapman and Hall.

Pearson, E.S. and Hartley, H.O. (1951), Charts of the power function for analysis of variance tests derived from the noncentral F-distribution, *Biometrika*, 38, 112-130.

Robbins, H. and Monro, S. (1951), A stochastic approximation method, *Annals of Mathematical Statistics*, 22, 400-407.

Rubinstein, R.Y. (1981), *Simulation and the Monte Carlo Method*, New York: John Wiley.

Schuirmann, D. J. (1987), A comparison of two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability, *Journal of Pharmacokinetics and Biopharmaceutics*, 15, 657-680.

Wetherill, G.B. (1963), Sequential estimation of quantal response curve, *Journal of the Royal Statistical Society*, *B*, 25, 1-48.