HOW CONSERVATIVE ARE POPULAR SAMPLE SIZE FORMULAS?

by

Lawrence L. Kupper and Kerry B. Hafner

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1839

November 1987

How Conservative Are Popular Sample Size Formulas?

Lawrence L. Kupper and Kerry B. Hafner*

Department of Biostatistics
School of Public Health
7400 Rosenau Hall
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599–7400

* Lawrence L. Kupper is Professor, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599−7400. Kerry B. Hafner is Ph.D. Student, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599−7400. This research was partially supported by N.I.E.H.S. training grant #2 T32 ES07018.

Abstract

One concern in the early stages of study planning and design is the minimum sample size needed to provide statistically credible results. This minimum sample size is usually determined via the use of simple formulas, or equivalently, from tables. However, the more popular formulas involve large-sample approximations and hence may be too conservative. This article provides empirical evidence indicating that this conservatism is drastic for certain sample size formulas based on confidence interval width. Common sample size formulas that consider statistical power are also discussed; these are shown to perform quite well, even for small sample size situations.

1. INTRODUCTION

For a variety of experimental and observational studies, it is of interest to estimate the mean μ of a random sample from a $N(\mu, \sigma^2)$ population. In this situation, one may wish to specify the maximum $100(1-\alpha)\%$ confidence interval width so that μ is estimated to within a *tolerance* of δ units. The minimum sample size n_m needed to achieve this precision is frequently recommended (*e.g.*, see Armitage 1971, p. 185; Koopmans 1987, p. 239; Ott 1977, p. 240) to be the smallest positive integer satisfying the inequality

$$\sigma(n_m)^{-1/2} Z_{1-\alpha/2} \leq \delta \quad \text{or} \quad n_m \geq [(\sigma/\delta)\, Z_{1-\alpha/2}]^2 , \tag{1}$$

where $\text{pr}(Z > Z_{1-\alpha/2}) = \alpha/2$ when $Z \sim N(0, 1)$.

Alternatively, suppose it is of interest to conduct a one-tailed size α test of H_0: $\mu=\mu_0$ versus H_1: $\mu>\mu_0$ for a random sample selected from a $N(\mu, \sigma^2)$ population. A popular inequality (Ott 1977, p. 241; Rosner 1986, p. 209) used to calculate the minimum sample size n_m necessary to achieve a power of at least $(1-\beta)$ when $\mu=\mu_1$ $(>\mu_0)$ is

$$n_m \geq [(Z_{1-\alpha} + Z_{1-\beta})/\theta]^2, \tag{2}$$

where $\theta = (\mu_1 - \mu_0)/\sigma$.

More generally, consider taking random samples of the same size from $N(\mu_0, \sigma^2)$ and $N(\mu_1, \sigma^2)$ populations to make inferences about $(\mu_1-\mu_0)$. Then, the inequality analogous to (1) is (Armitage 1971, p. 185; Ott 1977, p. 241)

$$\left(2\sigma^2/n_m\right)^{1/2} Z_{1-\alpha/2} \leq \delta,$$

or

$$n_m \geq 2\,[(\sigma/\delta)\, Z_{1-\alpha/2}]^2 . \tag{3}$$

Similarly, the two-population analogue of expression (2) for testing H_0: $\mu_1=\mu_0$ versus H_1: $\mu_1>\mu_0$ is (*e.g.*, see Fleiss 1986, p. 5; Kleinbaum, Kupper, and Muller 1987, p. 31; Meinert 1986, p. 84; Pocock 1983, p. 128)

$$n_m \geq 2\,[(Z_{1-\alpha} + Z_{1-\beta})/\theta]^2 \,. \tag{4}$$

During the planning phases of various types of research studies, expressions (1)-(4) are used by both statisticians and non-statisticians to provide guidelines for the numbers of experimental units to be sampled. For example, consider a randomized clinical trial designed to measure the efficacy of a new antihypertensive drug. Formulas (3) and (4), and analogous ones for proportions, are often used to help decide on the number of subjects to be allocated to the treatment and control groups (Freiman, Chalmers, Smith, and Kuebler 1978; McHugh and Le 1984).

Most users of expressions (1)-(4) probably appreciate that these inequalities involve large-sample approximations. Thus, their use in small-sample situations may lead to an underestimation of the sample sizes required to achieve specified inference-making goals. To our knowledge, no published literature seems to address the magnitude or the potential seriousness of this conservatism. The purpose of this paper is to quantify this sample size underestimation phenomenon, and to assess numerically whether it should be a cause for concern. It is shown that inequalities (2) and (4) perform amazingly well even for very small sample sizes, while inequalities (1) and (3) behave so poorly in all instances that their future use should be strongly discouraged. Extensions to other situations will also be discussed briefly.

2. ONE-SAMPLE METHODOLOGY

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample of size n from a $N(\mu, \sigma^2)$ population, and let

$$\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i \quad \text{and} \quad S^2 = (n-1)^{-1}\sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

For a one-tailed size α t-test of H_0: $\mu = \mu_0$ versus H_1: $\mu > \mu_0$, the power to reject H_0 in favor of H_1 when $\mu = \mu_1 > 0$ is

$$\text{pr}\left\{ \frac{\bar{Y}-\mu_0}{S/\sqrt{n}} > t_{n-1,1-\alpha} \Big/ \mu = \mu_1 > \mu_0 \right\},$$

where $t_{n-1,1-\alpha}$ is the $100(1-\alpha)$-th percentile of the central t_{n-1} distribution; equivalently, we have

$$\text{pr}\left\{ T'_{n-1}(\sqrt{n}\theta) > t_{n-1,1-\alpha} \Big/ \mu = \mu_1 > \mu_0 \right\}, \tag{5}$$

where $T'_{n-1}(\sqrt{n}\theta)$ has a non-central t distribution with $(n-1)$ degrees of freedom and noncentrality parameter $\sqrt{n}\theta = \sqrt{n}\,(\mu_1-\mu_0)/\sigma$.

For specified values of α and θ, expression (5) can be used (*e.g.*, see Guenther 1973) to find the minimum sample size n_m^* needed to achieve a power of at least $(1-\beta)$. It is interesting to note that the *actual* power attained with the sample size n_m computed using (2) is generally quite close in value to the desired power $(1-\beta)$. A noticeable power loss (5% or more) only occurs for small values of n_m (roughly, values less than 20). As expected, such a loss increases with decreasing α. A general rule is to increase *any* sample size n_m obtained via inequality (2) by two or three to achieve approximately the desired power $(1-\beta)$.

In contrast to inequality (2), the use of expression (1) always leads to a serious underestimation of the required sample size. This surprising conservatism can be illustrated by appealing to an overlooked, but nevertheless important, result due to Guenther (1965). Under the same assumptions which led to (5), the appropriate confidence interval for μ is

$$\bar{Y} \pm t_{n-1,1-\alpha/2}\, S/\sqrt{n}.$$

Following Guenther, we define n_m^* to be the smallest sample size such that

$$\text{pr}\left\{ S(n_m^*)^{-1/2}\, t_{n_m^*-1,1-\alpha/2} \leq \delta \right\} \geq (1-\gamma). \tag{6}$$

In contrast to expression (1), expression (6) accounts for the stochastic nature of the random variable S^2 via the *tolerance probability* $(1-\gamma)$. Expression (6) is easily shown to be equivalent to the probability statement

$$\text{pr}\left\{ (n_m^*-1)S^2/\sigma^2 \leq n_m^*(n_m^*-1)\,\delta^2/\sigma^2\, F_{1,n_m^*-1,1-\alpha} \right\} \geq (1-\gamma),$$

from which it follows that n_m^* is the smallest positive integer satisfying the inequality

$$n_m^*\,(n_m^*-1) \geq (\sigma/\delta)^2\, \chi^2_{n_m^*-1,1-\gamma}\, F_{1,n_m^*-1,1-\alpha}.$$

Since $(\sigma/\delta)^2 \doteq n_m/\chi^2_{1-\alpha}$ from (1), the expression relating n_m and n_m^* is

$$n_m^*\,(n_m^*-1)/n_m \geq \chi^2_{n_m^*-1,1-\gamma}\, F_{1,n_m^*-1,1-\alpha}/\chi^2_{1,1-\alpha}. \tag{7}$$

Appendix A provides the corresponding values of n_m^* which satisfy inequality (7) for various combinations of values for α, $(1-\gamma)$, and n_m. The entries in Appendix A clearly

illustrate the inappropriateness of inequality (1) for sample size determination in this context. As an example, if $n_m=40$ based on the use of (1), the exact sample size n_m^* needed to insure reasonably precise [say, $(1-\gamma) = 0.90$] estimation of μ with a 95% confidence interval ($\alpha = 0.05$) is $n_m^*=53$. It is quite disturbing that the actual tolerance probability $(1-\gamma')$ based on using a sample size of 40 is only 0.42, which is less than half of the desired value! Such large discrepancies should convince users that their n_m values so determined from the popular formula (1) should be corrected via Appendix A.

3. TWO-SAMPLE METHODOLOGY

For the two-sample situation, expressions analogous to (5) and (7) can be similarly developed. For i = 0 and 1, let $Y_{i1}, Y_{i2},\ldots, Y_{in}$ constitute a random sample of size n from $N(\mu_i, \sigma^2)$; define

$$\bar{Y}_i = n^{-1}\sum_{j=1}^{n} Y_{ij},\quad S_i^2 = (n-1)^{-1}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_i)^2 \text{ and } S_p^2=(S_0^2 + S_1^2)/2.$$

For a one-tailed size α test of H_0: $\mu_1=\mu_0$ versus H_1: $\mu_1>\mu_0$, the power to reject H_0 in favor of H_1 when $(\mu_1-\mu_0)$ has a specified positive value $\sigma\theta$ is

$$\text{pr}\left\{ T'_{2(n-1)}\left[(n/2)^{1/2}\theta\right] > t_{2(n-1),1-\alpha} \Big/ (\mu_1-\mu_0) = \sigma\theta > 0 \right\}, \tag{8}$$

where

$$T'_{2(n-1)}\left[(n/2)^{1/2}\theta\right] = (\bar{Y}_1-\bar{Y}_0)/S_p(2/n)^{1/2}$$

has a non-central t-distribution with 2(n−1) degrees of freedom and noncentrality parameter $(n/2)^{1/2}\theta$.

A comparison between n_m values based on (4) and corresponding n_m^* values based on (8) shows that the degree of agreement here is at least as good as that seen in the one-sample case. Moreover, exact equality often holds for sample sizes as little as 10. The excellence of the approximation (4) in the small-sample situation has also been noticed by Fleiss (1986, p. 369).

Under the stated assumptions, the appropriate $100(1-\alpha)\%$ confidence interval for $(\mu_1-\mu_0)$ is

$$(\bar{Y}_1-\bar{Y}_0) \pm t_{2(n-1),1-\alpha/2}\, S_p\, (2/n)^{1/2}$$

and hence the two-sample analogue of (6) is

$$\mathrm{pr}\left\{ S_p (2/n_m^*)^{1/2}\, t_{2(n_m^*-1),1-\alpha/2} < \delta \right\} \geq (1-\gamma).$$

Using arguments identical to those leading to (7), we find the inequality relating n_m^* in the above expression to n_m in expression (3) to be

$$2\, n_m^*\, (n_m^*-1)/n_m \geq \chi^2_{2(n_m^*-1),1-\gamma}\, F_{1,2(n_m^*-1),1-\alpha}/\chi^2_{1,1-\alpha}. \qquad (9)$$

The body of Appendix B contains n_m^* values calculated via inequality (9) for specified combinations of values for α, $(1-\gamma)$, and n_m; the entries clearly document the inappropriateness of inequality (3) for sample size determination in this situation. We strongly recommend that users of expression (3) take note of its extreme conservatism, and that they use Appendix B to correct any sample size estimates based on the use of inequality (3).

4. DISCUSSION

As stated earlier, the *specific* goal of this paper was to determine situations, if any, where the popular sample size formulas (1)-(4) could be misleading. We have found that inequalities (2) and (4), which are typically used to calculate the smallest sample size needed to achieve a specified minimum power, were quite reliable in all instances considered. In contrast, expressions (1) and (3), which are commonly employed to estimate the minimum sample size required to obtain a $100(1-\alpha)\%$ confidence interval with a specified maximum width, were seen to be uniformly inappropriate.

The fact that inequalities (1) and (3) perform so poorly is disturbing, especially since the use of such confidence interval-based sample size estimation formulas for the design of both randomized clinical trials and observational epidemiologic studies is becoming quite common. The reason for the increased popularity of formulas like these, relative to ones like (2) and (4), is that the goal of such research efforts is more often to estimate as accurately as possible the *magnitude* of the effect of interest, rather than to decide whether or not a finding is statistically significant (see Rothman 1986). Our results further suggest that the use of popular sample size formulas for estimating other parameters (*e.g.*, differences in proportions, odds ratios, etc.) to within specified tolerances may also be providing sample size estimates which are much too low.

When using confidence interval-based sample size estimation formulas for study design, what steps can be taken to correct for their anticipated conservatism? Appendices A and B can, of course, be used to adjust sample size estimates obtained via inequalities (1) and (3). For well-known confidence interval-based sample size formulas where the parameter of interest is a proportion π or a difference in proportions $(\pi_1-\pi_0)$, we recommend that, when economically feasible, researchers use that *maximum* sample size computed assuming that the population proportions are equal to 1/2. Not only will this simple approach avoid the type of

conservatism considered in this paper, it will also help to provide additional subjects so that subsequent more complicated data analyses may have reasonably good statistical properties. This is an important consideration since users of basic sample size estimation formulas like (1)-(4), and analogous ones involving proportions, often seem to ignore the fact that the sample sizes so computed are appropriate only for very simple statistical analyses. It is invariably the case that much more complicated statistical methods (*e.g.*, regression procedures) are employed at the data analysis stage; and, the sample sizes required to insure that such multivariable procedures have adequate precision and/or power will generally be considerably larger than those based on formulas like (1)-(4).

Based on the discussion above, we wish to stress to users of standard sample size estimation formulas that, for all of the reasons cited above, the sample sizes so obtained will generally be inadequate for the desired analysis goals. Even so, researchers will continue to employ formulas like (1)-(4) because of their simplicity and popularity. We hope that this paper will help to make them aware of some of the problems associated with the use of these formulas.

APPENDIX A. One-sample tolerance probability comparisons between n_m^* and n_m; $(1-\gamma)$ is the tolerance probability using n_m^*, and $(1-\gamma')$ is the tolerance probability using n_m.

$\alpha=0.10$

		$(1-\gamma)=0.70$	$(1-\gamma)=0.80$	$(1-\gamma)=0.90$	$(1-\gamma)=0.95$	$(1-\gamma)=0.99$
n_m	$(1-\gamma')$	n_m^*	n_m^*	n_m^*	n_m^*	n_m^*
5	0.33	8	9	11	12	13
10	0.39	14	15	17	19	21
15	0.41	20	21	23	25	28
20	0.42	25	27	29	32	35
25	0.43	30	33	35	38	42
30	0.44	36	38	41	44	49
35	0.44	41	44	47	50	55
40	0.45	46	49	53	56	61
45	0.45	52	55	58	62	68
50	0.45	57	60	64	67	74
55	0.45	62	65	70	73	80
60	0.46	67	71	75	79	86
65	0.46	73	76	81	85	92
70	0.46	78	81	86	90	98
75	0.46	83	87	92	96	104
80	0.46	88	92	97	102	110
85	0.46	93	97	103	107	116
90	0.46	99	103	108	113	122
95	0.46	104	108	114	119	127
100	0.47	109	113	119	124	133

APPENDIX A: (*continued*)

$\alpha = 0.05$

		$(1-\gamma)=0.70$	$(1-\gamma)=0.80$	$(1-\gamma)=0.90$	$(1-\gamma)=0.95$	$(1-\gamma)=0.99$
n_m	$(1-\gamma')$	n_m^*	n_m^*	n_m^*	n_m^*	n_m^*
5	0.26	9	10	11	12	14
10	0.34	15	16	18	19	22
15	0.37	20	22	24	26	29
20	0.39	26	27	30	32	36
25	0.40	31	33	36	38	43
30	0.41	36	39	42	44	49
35	0.42	42	44	48	50	55
40	0.42	47	50	53	56	62
45	0.43	52	55	59	62	68
50	0.43	57	60	65	68	74
55	0.43	63	66	70	74	80
60	0.44	68	71	76	80	86
65	0.44	73	77	81	85	92
70	0.44	78	82	87	91	98
75	0.44	84	87	92	97	104
80	0.44	89	93	98	102	110
85	0.45	94	98	103	108	116
90	0.45	99	103	109	114	122
95	0.45	104	109	114	119	128
100	0.45	110	114	120	125	134

APPENDIX A: (*continued*)

$\alpha = 0.01$

		$(1-\gamma)=0.70$	$(1-\gamma)=0.80$	$(1-\gamma)=0.90$	$(1-\gamma)=0.95$	$(1-\gamma)=0.99$
n_m	$(1-\gamma')$	n_m^*	n_m^*	n_m^*	n_m^*	n_m^*
5	0.13	10	11	12	13	15
10	0.23	16	17	19	20	23
15	0.27	21	23	25	27	30
20	0.30	27	29	31	33	37
25	0.32	32	34	37	39	44
30	0.34	38	40	43	46	50
35	0.35	43	45	49	52	57
40	0.36	48	51	55	58	63
45	0.37	53	56	60	63	69
50	0.38	59	62	66	69	75
55	0.38	64	67	72	75	82
60	0.39	69	73	77	81	88
65	0.39	74	78	83	87	94
70	0.39	80	83	88	92	100
75	0.40	85	89	94	98	106
80	0.40	90	94	99	104	112
85	0.40	95	99	105	109	117
90	0.41	101	105	110	115	123
95	0.41	106	110	116	120	129
100	0.41	111	115	121	126	135

APPENDIX B. Two-sample tolerance probability comparisons between n_m^* and n_m; $(1-\gamma)$ is the tolerance probability using n_m^*, and $(1-\gamma')$ is the tolerance probability using n_m.

$\alpha=0.10$

		$(1-\gamma)=0.70$	$(1-\gamma)=0.80$	$(1-\gamma)=0.90$	$(1-\gamma)=0.95$	$(1-\gamma)=0.99$
n_m	$(1-\gamma')$	n_m^*	n_m^*	n_m^*	n_m^*	n_m^*
5	0.38	7	8	9	10	11
10	0.42	13	14	15	16	18
15	0.44	18	19	21	22	25
20	0.45	23	25	27	28	31
25	0.45	29	30	32	34	37
30	0.46	34	36	38	40	43
35	0.46	39	41	44	46	49
40	0.46	44	46	49	51	55
45	0.46	50	52	55	57	61
50	0.47	55	57	60	62	67
55	0.47	60	62	65	68	73
60	0.47	65	68	71	74	78
65	0.47	70	73	76	79	84
70	0.47	75	78	82	85	90
75	0.47	81	83	87	90	96
80	0.47	86	89	92	95	101
85	0.47	91	94	98	101	107
90	0.48	96	99	103	106	112
95	0.48	101	104	108	112	118
100	0.48	106	109	114	117	124

APPENDIX B: (*continued*)

α=0.05

		$(1-\gamma)$=0.70	$(1-\gamma)$=0.80	$(1-\gamma)$=0.90	$(1-\gamma)$=0.95	$(1-\gamma)$=0.99
n_m	$(1-\gamma')$	n_m^*	n_m^*	n_m^*	n_m^*	n_m^*
5	0.33	8	8	9	10	11
10	0.38	13	14	15	16	18
15	0.41	18	20	21	23	25
20	0.42	24	25	27	28	31
25	0.43	29	31	33	34	37
30	0.44	34	36	38	40	44
35	0.44	39	41	44	46	50
40	0.44	45	47	49	51	55
45	0.45	50	52	55	57	61
50	0.45	55	57	60	63	67
55	0.45	60	63	66	68	73
60	0.45	65	68	71	74	79
65	0.46	71	73	77	79	84
70	0.46	76	78	82	85	90
75	0.46	81	84	87	90	96
80	0.46	86	89	93	96	101
85	0.46	91	94	98	101	107
90	0.46	96	99	103	107	113
95	0.46	101	104	109	112	118
100	0.46	107	110	114	117	124

APPENDIX B: (*continued*)

α=0.01

		$(1-\gamma)$=0.70	$(1-\gamma)$=0.80	$(1-\gamma)$=0.90	$(1-\gamma)$=0.95	$(1-\gamma)$=0.99
n_m	$(1-\gamma')$	n_m^*	n_m^*	n_m^*	n_m^*	n_m^*
5	0.21	8	9	10	11	12
10	0.30	14	15	16	17	19
15	0.34	19	20	22	23	25
20	0.36	24	26	28	29	32
25	0.37	30	31	33	35	38
30	0.38	35	37	39	41	44
35	0.39	40	42	44	46	50
40	0.40	45	47	50	52	56
45	0.41	51	53	55	58	62
50	0.41	56	58	61	63	68
55	0.42	61	63	66	69	74
60	0.42	66	68	72	74	79
65	0.42	71	74	77	80	85
70	0.42	76	79	83	85	91
75	0.43	82	84	88	91	96
80	0.43	87	89	93	96	102
85	0.43	92	95	99	102	108
90	0.43	97	100	104	107	113
95	0.44	102	105	109	113	119
100	0.44	107	110	115	118	125

REFERENCES

Armitage, P. (1971), *Statistical Methods in Medical Research*, New York: John Wiley.

Fleiss, J. L. (1986), *The Design and Analysis of Clinical Experiments*, New York: John Wiley.

Freiman, J. A., Chalmers, T. C., Smith, H., and Kuebler, R. R. (1978), "The Importance of Beta, the Type II Error and Sample Size in the Design and Interpretation of the Randomized Control Trial," *The New England Journal of Medicine*, 299, 690-694.

Guenther, W. C. (1965), *Concepts of Statistical Inference*, New York: McGraw-Hill.

----- (1973), "Determination of Sample Size for Tests Concerning Means and Variances of Normal Distributions," *Statistica Neerlandica*, 27, 103-113.

Kleinbaum, D. G., Kupper, L. L., and Muller, K. E. (1987), *Applied Regression Analysis and Other Multivariable Methods* (*Second Edition*), Boston: PWS-Kent.

Koopmans, L. H. (1987), *Introduction to Contemporary Statistical Methods* (*Second Edition*), Boston: Duxbury.

McHugh, R. B., and Le, C. T. (1984), "Confidence Estimation and the Size of a Clinical Trial," *Controlled Clinical Trials*, 5, 157-163.

Meinert, C. L. (1986), *Clinical Trials: Design, Conduct, and Analysis*, New York: Oxford.

Ott, L. (1977), *An Introduction to Statistical Methods and Data Analysis*, Boston: Duxbury.

Pocock, S. J. (1983), *Clinical Trials: A Practical Approach*, New York: John Wiley.

Rosner, B. (1986), *Fundamentals of Biostatistics* (*Second Edition*), Boston: Duxbury.

Rothman, K. J. (1986), *Modern Epidemiology*, Boston: Little-Brown.