SAMPLE SIZES FOR SOCIAL EXPERIMENTS[1]

George Cave, Manpower Demonstration Research Corporation
Three Park Avenue, New York, New York 10016

Greenberg and Robins (1986) have documented the frequent use since the 1970s of social experiments for program evaluation. Designing such experiments presents special problems. The chief source of difficulty is that very often, the program effects to be detected are quite modest in size, while survey followup is very costly per case. Other problems arise as well. For example, especially when evaluating ongoing programs, minimizing the size of the control group makes implementation easier, but unbalancing the sample raises the required total number of cases.

The purpose of this paper is to derive and interpret simple closed-form results to use in designing social experiments. For two special cases, it is shown precisely how the sample size necessary to detect the effect of assignment to a social program using a one-factor experimental design depends on several parameters. In general, a larger sample is necessary--

- --the smaller the actual effect of the program on its target population;
- --the greater the variance of the outcome for which an impact is measured;
- --the smaller the desired probability of a false positive or "Type I error;"[2]
- --the greater the desired power to detect effects which do in fact exist; that is, the smaller the chance of a false negative or "Type II error;" and
- --the more unbalanced the sample split among treatment groups.

The precise numerical relationship among these four parameters and the required sample size depends on several design factors--

- --the number of levels for the experimental factor; for example, whether there is only one treatment plus the control group or there is more than one treatment;
- --the way the sample is split among levels; for example, whether 25% are controls or whether there is a 50-50 split between control and treatment groups;
- --the statistical distribution of the outcome variable; for example, whether the outcome is discrete or continuous;
- --the statistical model used to infer the population effect from the sample; for example, whether ANOVA or ANCOVA;
- --the optimal design theory used; for example, whether classical or Bayesian; and
- --the way the theory is tailored to handle special problems anticipated, such as nonparticipation among subjects assigned to a treatment.

The rest of this paper contains brief discussions of each of these design factors and two examples of large sample normal theory sample size formulas which are relatively simple yet general enough to use in most practical applications in social experimentation.

Number of treatment groups. The simplest classical field experiment has two groups of subjects--those randomly assigned to treatment in "the program", and those randomly assigned to control. When there is interest in the effects of particular program components, more than one treatment group is required. For example, in an evaluation of an employment program, there might be random assignment of subjects to one of three groups: control, "job search only", and "job search plus" other program components.[3]

Sample split. A larger sample is necessary the more unequal the sizes of the treatment and control groups. For example, with one treatment and one control group, assigning 25% of the sample to control necessitates a sample a third larger than would be necessary with 50% assigned to control. That is, 250 controls and 750 treatment group members provide no more power to detect impacts of a given size at a given significance level than 375 controls and 375 treatment group members.[4]

Distribution of outcome variable. Different techniques are necessary to deal with discrete data, such as whether or not subjects were employed, and with continuous data, such as their earnings.[5] Different techniques are necessary for univariate outcomes, such as earnings, and for multivariate outcomes, such as earned and unearned income considered simultaneously.

Statistical model. The simplest fixed effect statistical model to use is ANOVA. Cohen (1977) is a standard reference for sample size calculations for this case. Analysis of covariance extends ANOVA to take account of the reduction in outcome variance when covariates are used to control for pre-treatment differences among subjects. Pitcher (1979) and Conlisk (1979) are frequently-cited references for sample size determination in this case. ANOVA sample sizes are more conservative (larger) than ANCOVA sample sizes.

Optimal design theory. The simplest approach to deciding how large a sample to recruit uses classical frequentist statistical theory to derive a relationship between sample size and parameters of the experiment, including the presumed true size of the treatment effect. This effect size may be the average impact estimated in prior studies, or it may be the amortized cost of the program in question. The decision-theoretic approach[6] starts with the expected cost per sample point (for example, $500) and attempts to locate the sample size at which the marginal expected value of the information to be gleaned from the experiment just drops below the marginal expected cost per sample point. With a randomized block evaluation design, the classical approach may result in allocating more sample to a cheaper or presumed weaker treatment (such as job search) and less to a presumed stronger treatment (such as training) than would the other approach.

Tailoring to handle special problems. Textbook approaches to experimental design are based on fifty years of field

experience in an agricultural or engineering context which may be quite different from the program evaluation context for a social experiment. Human behavior makes program evaluation more complicated. Those assigned to treatment may not show up for treatment, even when they face punishment for not showing up. Service providers may find ways to ensure that those assigned to control actually get the treatment, perhaps from other providers outside the evaluation contract. Beyond these fundamental differences between social experiments and laboratory experiments, sample size calculations may have to reflect survey design effects due to stratified sampling or other complications.[7] There is no hard-and-fast solution to these problems suitable for use in every design. However, as illustrated below for the case of nonparticipation among treatment assignees, parameters of available sample-size formulas may be re-interpreted to handle some problems.

A binomial ANOVA sample size formula. To measure the effect of assignment to a program on one binomial outcome with a completely randomized two-group experiment, classical optimal design theory for an ANOVA statistical model leads to a comparatively simple sample size formula. Where

a = significance level;

B = 1 - statistical power;

D = population effect of treatment, the difference between binomial proportions for treated and untreated population members;

M = the midpoint of the population effect, a simple average of the two binomial proportions;

c = the fraction of the sample which belongs to the control group;

z(x) = the inverse of the cumulative standardized normal distribution function (for example, z(0.5) = 0 and z(0.975) = 1.95996); and

asn(x) = the arcsine or inverse sine of x;

the total sample size required is given by[8]

$$(1)\quad n \geq 4\left\{\frac{z(1-a)+z(1-B)}{h}\right\}^2 \frac{1}{4c(1-c)},$$

where n is a positive integer and

$$(2)\quad h = 2asn\left(\sqrt{M+(D/2)}\right) - 2asn\left(\sqrt{M-(D/2)}\right).$$

Tables based on this relationship are available in Cohen (1977, p. 205). The multiplicative factor involving c is a sample split inflation factor with a value of unity when c = 0.5. This factor grows larger as c gets farther away from 0.5. For example, c = 1/3 yields an inflation factor of 1.125, while c = 0.25 increases the required sample to 1.333 times the size needed when controls and treatment assignees are split evenly. The required sample size is sensitive to the value specified for the midpoint of the population effect. A worst-case analysis, producing the most conservative sample size, would use one-half for this parameter.

Parameter n is the total number of usable data points required for analysis. If it is believed that, due to attrition, only fraction r of those subjects assigned to the sample will yield usable data, then the number of subjects assigned should be increased to n/r.

A continuous ANCOVA sample size formula. To measure the effect of assignment to a program on one continuous outcome with a completely randomized two-group experiment, classical optimal design theory for an ANCOVA statistical model leads to a slightly more complicated result. Just as in the previous case, let

a = significance level;

B = 1 - statistical power;

c = the fraction of the sample which belongs to the control group; and

z(x) = the inverse of the cumulative standardized normal distribution function.

However, the other variables of the previous case are not applicable. Their places are taken by

y_i = the continuous outcome variable;

Z_i = a k-vector of covariates measured just before assignment to treatment or control; and

S_i = a dummy variable which is zero for those assigned to control status and unity for those assigned to treatment.

In the fitted regression equation $\hat{y}_i = S_i\delta_0 + Z_i'\delta_{1:k} + \delta_{k+1}$, which explains fraction R-bar-square of sample outcome variance, the first coefficient is interpreted as the sample impact of assignment to the treatment. Its expected value is the population effect of treatment, δ_0.

An expression for its variance, derived in the Appendix below, may be manipulated as outlined there to yield the sample size formula

$$(3)\quad n \geq 4\left\{\frac{z(1-a)+z(1-B)}{\delta_0}\right\}^2 \frac{(1-\bar{R}^2)Var(y)}{1-R^2_{SZ}} \frac{1}{4c(1-c)} + 1,$$

where n is a positive integer, Var(y) is the sample variance of the outcome, and R^2_{SZ} is the proportion of variation in S_i explained by a regression of S_i on Z_i and a constant. R^2_{SZ} has expected value zero[9] if assignment to treatment truly is random; the multiplicative factor $(1/(1-R^2_{SZ}))$ is a sample-size-increasing randomization design effect analogous to a survey design effect.

Just as in the previous case, the multiplicative factor involving c is a sample split inflation factor which takes the value unity when c = 0.5, and n is the total number of usable data points required for analysis.

To use this ANCOVA sample size formula, estimates of the population effect, Var(y), and R-bar-square must be obtained beforehand. Sample variances and proportions of variance explained can come from prior studies of similar populations and treatments. The estimate of the population effect can come from prior studies, from the cost of the treatment, or from some notion of what size effect would be policy-relevant.

When the relative size rather than the absolute size of the population effect is to be specified, a slight modification of (3) can be used. If the relative size is defined as ν,

$$(4)\quad \delta_0 = \nu\bar{y},$$

and

$$(5)\quad n \geq \{4\left\{\frac{z(1-\alpha)+z(1-\beta)}{\nu}\right\}^2\left\{\frac{\sqrt{Var(y)}}{\bar{y}}\right\}^2 \frac{(1-R^2)}{1-R^2_{SZ}}\frac{1}{4c(1-c)}\}+1.$$

It may be easier to choose a relative effect size for (5) than to decide on an absolute effect size for (3), and it may be easier to specify a value for the coefficient of variation in (5) than for the absolute amount of variance in (3).

When, as in block designs, a fixed total sample size must be allocated among design cells, the formulas just presented should be reinterpreted to apply to each cell separately. There is a separate population effect, outcome variance, R-bar-square, proportion of controls, and significance level for each cell. The formula may be solved for statistical power in each cell, and sample may be allocated among cells to equalize power within each cell or minimize some function of weighted power in each cell.[10]

Effects per participant. So far, the effect of assignment to treatment has been the experiment's presumed goal for estimation. When, as frequently occurs in social experimentation, a sizable proportion of those assigned to treatment become nonparticipants who do not receive it, it is necessary to distinguish between the effect of assignment and the effect of actual participation. The effect of assignment is the difference in average outcomes between all those assigned to treatment and all those assigned to control. The effect of participation, however, is the difference in average outcomes between those assigned to treatment who participated and those assigned to control who would have participated given the opportunity.

The simplest way to deal with this issue is to reduce the detectable assignment effect to reflect anticipated nonparticipation. It may be thought that nonparticipants and other groups "water down" the size of the assignment effect that must be found. For example, if only half of the target population would participate in a program of treatment that would raise participants' earnings by $1000, then the detectable effect used in (3) should be $500.

Two alternative approaches, discussed in Cave (1987a), are to estimate switching regression models for potential participants and potential nonparticipants, or to inflate the impact of assignment by dividing it by the sample fraction of participants.

Summary. Expressions (1), (3), and (5) all have the same general form. Careful inspection of these expressions is sufficient to verify each of the assertions made at the beginning of this paper. Halving the size of the effect that must be found quadruples the required sample size. The sample size is proportional to the amount of variance in the outcome, and to the fraction of variation in the outcome that cannot be explained with regression. A smaller rejection region or higher statistical power[11] makes more observations necessary. Moving away from an even split between treatment and control groups increases the required sample size, multiplying it by 1/(4c(1-c)).

This paper also mentioned several other factors which must be considered when determining the sample size for a social experiment. These factors are survey design effects, attrition, randomization design effects, and nonparticipation.

Appendix

The basic model in an analysis of covariance for a one-factor two-level experiment is

$$(A1)\quad y_i = S_i\delta_0 + Z_i'\delta_{1:k} + \delta_{k+1} + \epsilon,$$

where y_i is a post-treatment outcome, S_i is a dummy variable for treatment assignment status, Z_i is a k-vector of covariates observed before treatment assignment, and the error is normally distributed about zero with variance σ^2.

Least squares parameter estimates from a sample of n observations are given by

$$(A2)\quad \hat{\delta} = (X'X)^{-1}X'y,$$

where

$$(A3)\quad X = \begin{pmatrix} S_1 & z_{11} & z_{12} & \cdot & z_{1k} & 1 \\ S_2 & z_{21} & z_{22} & \cdot & z_{2k} & 1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ S_n & z_{n1} & z_{n2} & \cdot & z_{nk} & 1 \end{pmatrix}.$$

Thus

$$(A4)\quad X'X = \begin{pmatrix} \Sigma S_i^2 & \Sigma z_{i1}S_i & \Sigma z_{i2}S_i & \cdot & \Sigma z_{ik}S_i & \Sigma S_i \\ \Sigma z_{i1}S_i & \Sigma z_{i1}^2 & \Sigma z_{i1}z_{i2} & \cdot & \Sigma z_{i1}z_{ik} & \Sigma z_{i1} \\ \Sigma z_{i2}S_i & \Sigma z_{i2}z_{i1} & \Sigma z_{i2}^2 & \cdot & \Sigma z_{i2}z_{ik} & \Sigma z_{i2} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \Sigma z_{ik}S_i & \Sigma z_{ik}z_{i1} & \Sigma z_{ik}z_{i2} & \cdot & \Sigma z_{ik}^2 & \Sigma z_{ik} \\ \Sigma S_i & \Sigma z_{i1} & \Sigma z_{i2} & \cdot & \Sigma z_{ik} & n \end{pmatrix}.$$

Partitioning this expression yields

$$(A5)\quad X'X = \begin{pmatrix} \Sigma S_i^2 & R \\ R' & Q \end{pmatrix},$$

where the first element of the matrix in (A5) is the upper left corner element of the matrix in (A4).

As explained in detail by, for example, Theil (1970, p. 18), the matrix equation

$$(A6)\quad (X'X)^{-1}\begin{pmatrix} \Sigma S_i^2 & R \\ R' & Q \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & I_{k+1} \end{pmatrix}$$

may be expanded into a system of three scalar equations and one matrix equation and solved to give as the inverse of the partitioned matrix

$$(A7)\quad (X'X)^{-1} = \begin{pmatrix} \left(\Sigma S_i^2 - RQ^{-1}R'\right)^{-1} & -\left(\Sigma S_i^2 - RQ^{-1}R'\right)^{-1}RQ^{-1} \\ -Q^{-1}R'\left(\Sigma S_i^2 - RQ^{-1}R'\right)^{-1} & Q^{-1} + Q^{-1}R'\left(\Sigma S_i^2 - RQ^{-1}R'\right)^{-1}RQ^{-1} \end{pmatrix}.$$

Now since

$$(A8)\quad X'y = \begin{pmatrix} \Sigma S_i y_i \\ \Sigma z_{i1} y_i \\ \Sigma z_{i2} y_i \\ \cdot \\ \Sigma z_{ik} y_i \\ \Sigma y_i \end{pmatrix}$$

may also be partitioned into its first element and a k+1-vector called, say, T, the parameter estimates become

$$(A9)\quad \hat{\delta} = \begin{pmatrix} \delta_0 \\ \delta_{1:k+1} \end{pmatrix} = \begin{pmatrix} \left(\Sigma S_i^2 - RQ^{-1}R'\right)^{-1}\Sigma S y_i - \left(\Sigma S_i^2 - RQ^{-1}R'\right)^{-1}RQ^{-1}T \\ -Q^{-1}R'\left(\Sigma S_i^2 - RQ^{-1}R'\right)^{-1}\Sigma S y_i + Q^{-1} + Q^{-1}R'\left(\Sigma S_i^2 - RQ^{-1}R'\right)^{-1}RQ^{-1}T \end{pmatrix}.$$

Consider next the model for the assignment status dummy

$$(A10)\quad S_i = Z_i'\theta_{1:k} + \theta_{k+1} + \xi,$$

Its least-squares estimate from a sample of n observations is

$$(A11)\quad \hat{\theta} = (W'W)^{-1}W'S$$

where

$$(A12)\quad W = \begin{pmatrix} z_{11} & z_{12} & \cdot & z_{1k} & 1 \\ z_{21} & z_{22} & \cdot & z_{2k} & 1 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ z_{n1} & z_{n2} & \cdot & z_{nk} & 1 \end{pmatrix}.$$

As one might expect for the regression of one right-hand-side variable on the others, its least-squares solution is closely related to the solution (given by (A9)) of the model for y, since

$$(A13)\quad W'W = \begin{pmatrix} \Sigma z_{i1}^2 & \Sigma z_{i1}z_{i2} & \cdot & \Sigma z_{i1}z_{ik} & \Sigma z_{i1} \\ \Sigma z_{i2}z_{i1} & \Sigma z_{i2}^2 & \cdot & \Sigma z_{i2}z_{ik} & \Sigma z_{i2} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \Sigma z_{ik}z_{i1} & \Sigma z_{ik}z_{i2} & \cdot & \Sigma z_{ik}^2 & \Sigma z_{ik} \\ \Sigma z_{i1} & \Sigma z_{i2} & \cdot & \Sigma z_{ik} & n \end{pmatrix}$$

is identical to submatrix Q of (A4), and since

$$(A14)\quad W'S = \begin{pmatrix} \Sigma z_{i1}S_i \\ \Sigma z_{i2}S_i \\ \cdot \\ \Sigma z_{ik}S_i \\ \Sigma S_i \end{pmatrix} = R'.$$

In fact, the first element of the matrix in (A7) is the reciprocal of the amount of sample variation in S not explained by (A10):

$$(A15)\quad \left(\Sigma S_i^2 - RQ^{-1}R'\right)^{-1} = \left(S'S - S'W(W'W)^{-1}W'S\right)^{-1}$$
$$= \left(S'S - S'W(W'W)^{-1}(W'W)(W'W)^{-1}W'S\right)^{-1}$$
$$= (S'S - \hat{\theta}'W'W\hat{\theta})^{-1}$$
$$= \{(W\hat{\theta} + \hat{\xi})'(W\hat{\theta} + \hat{\xi}) - \hat{\theta}'W'W\hat{\theta}\}^{-1}$$
$$= (\hat{\xi}'\hat{\xi})^{-1}.$$

Thus

$$(A16)\quad Var(\hat{\delta}_0) = \hat{\sigma}^2(X'X)_{11}^{-1}$$
$$= \hat{\sigma}^2(\hat{\xi}'\hat{\xi})^{-1}$$
$$= \frac{(1-\bar{R}^2)Var(y)}{(n-1)Var(S)(1-R_{Sz}^2)}.$$

This is Pitcher's (1979, p. 70) equation I-1.

If the optimal sample size is large, the sample estimator of the effect is normally distributed, and the critical value for rejecting the hypothesis that there is no effect is

$$(A17)\quad \hat{\delta}_0^c = z(1-\alpha)\sqrt{Var(\hat{\delta}_0)}.$$

For a probability of size B that the null will be accepted even though the true size of the effect is δ_0,

$$(A18)\quad z(B) = -z(1-B) = \frac{\hat{\delta}_0^c - \delta_0}{\sqrt{Var(\hat{\delta}_0)}}$$

is the required z-value. Substituting (A16) and (A17) into the last member of (A18) and rearranging yields the sample size formula given as expression (3) in the text.

Finally, the impact coefficient itself may be expressed in terms of the variables and parameters of the auxiliary regression (A10). Substituting (A11) through (A14) into (A9) yields

$$(A19)\quad \delta_0 = \frac{S'y - \hat{\theta}'W'y}{S'S - \hat{\theta}'W'W\hat{\theta}}.$$

References

Cave, George. "Assignment, Participation, and Attenuated Impacts." Unpublished manuscript, August, 1987.

Cave, George. "Sample Sizes for Impact Differences by Subgroup." Unpublished manuscript, July, 1987.

Cohen, Jacob. Statistical Power Analysis for the Behavioral Sciences. New York: Academic Press, revised ed., 1977.

Conlisk, John. "Choice of Sample Size in Evaluating Manpower Programs: Comments on Pitcher and Stafford." Research in Labor Economics Supplement 1 (1979): 79-96.

Greenberg, David, and Philip Robins. "The Changing Role of Social Experiments in Policy Analysis." Journal of Policy Analysis and Management 5, 2 (Winter 1986): 340-362.

Greene, William. "Estimation of Limited Dependent Variable Models by Ordinary Least Squares and the Method of Moments." Journal of Econometrics 21, 2 (February 1983): 195-212.

Kastenbaum, Marvin A., and David G. Hoel. "Sample Size Requirements: One-Way Analysis of Variance." Biometrika 57, 2 (1970): 421-430.

Lansing, John B., and James N. Morgan. Economic Survey Methods. Ann Arbor: University of Michigan Institute for Survey Research, 1971.

Mace, Arthur E. Sample Size Determination. New York: Reinhold, 1964.

Pitcher, Hugh M. "A Sensitivity Analysis to Determine Sample Sizes for Performing Impact Evaluation of the CETA Programs." Research in Labor Economics Supplement 1 (1979): 37-78.

Stafford, Frank P. "A Decision Theoretic Approach to the Evaluation of Training Programs." Research in Labor Economics Supplement 1 (1979): 9-35.

Stafford, Frank P. "Optimal Sample Sizes for Evaluating JTPA." Unpublished manuscript, May 1985.

Theil, Henri. Principles of Econometrics. New York: Wiley, 1971.

Winer, B.J. Statistical Principles in Experimental Design. New York: McGraw-Hill, 2d ed., 1971, 1962.

1. George Cave is a Senior Research Associate at Manpower Demonstration Research Corporation. The ideas expressed here do not necessarily represent the views of MDRC.

2. That is, the smaller the rejection region for the hypothesis of no effect.

3. In this particular example, policy interest would focus on selection of the best treatment group, and on comparing the relative effectiveness of the two treatments. If "job search only" were believed to have an employment rate effect of 5 percentage points, and "job search plus" were believed to have an effect of 8 percentage points, then the sample size in those two groups taken together should be sufficient to detect a difference of 3 percentage points, while the sample size in the first group and the control group together should be sufficient to detect a 5 point difference. Thus it may be preferable to use a two-group sample size formula even when there are three groups. In the unlikely case, given practical constraints on sample sizes for social experiments, that there are more than three treatment groups, methods based on noncentral F may be relevant for determining sample sizes. See Kastenbaum and Hoel (1970) for a very practical approach using a standardized range parameter, and Mace (1964) or Winer (1971) for a more traditional approach using a noncentrality parameter.

4. See expressions (1), (3), and (5) below.

5. An outcome distribution which has aspects of both the discrete and the continuous approaches, but which requires slightly more complicated empirical techniques, is the tobit. See Greene (1981, p. 203) for a simple expression for the asymptotic variance of the tobit estimator.

6. See Stafford (1979, 1985).

7. See Lansing and Morgan (1971) and Winer (1971). Stratifying decreases, and clustering increases, by a multiplicative factor, the sample size needed when there is simple random sampling.

8. See Mace (1964), p. 101.

9. To ensure the internal validity of inferences about effects, it is important to test this hypothesis for every sample and subsample of complete data used in the analysis of a social experiment.

10. When there is policy interest in the difference in program impact by block or "subgroup", a different approach must be taken to determine the total sample size required. See Cave (1987b).

11. Note that, since $z(1 - 0.50) = 0$, ignoring the power dimension of the analysis and setting out to control only the size of the rejection region is equivalent to requiring only 50% power.