Behavior Research Methods, Instruments, & Computers
2001, 33 (4), 544-548

Sample size determination for a *t* test given a *t* value from a previous study: A FORTRAN 77 program

RAPHAEL GILLETT
University of Leicester, Leicester, England

When uncertain about the magnitude of an effect, researchers commonly substitute in the standard sample-size-determination formula an estimate of effect size derived from a previous experiment. A problem with this approach is that the traditional sample-size-determination formula was not designed to deal with the uncertainty inherent in an effect-size estimate. Consequently, estimate-substitution in the traditional sample-size-determination formula can lead to a substantial loss of power. A method of sample-size determination designed to handle uncertainty in effect-size estimates is described. The procedure uses the *t* value and sample size from a previous study, which might be a pilot study or a related study in the same area, to establish a distribution of probable effect sizes. The sample size to be employed in the new study is that which supplies an expected power of the desired amount over the distribution of probable effect sizes. A FORTRAN 77 program is presented that permits swift calculation of sample size for a variety of *t* tests, including independent *t* tests, related *t* tests, *t* tests of correlation coefficients, and *t* tests of multiple regression *b* coefficients.

The versatility of the *t* statistic has made it one of the most frequently employed statistics in the behavioral sciences. Among many other uses, it is routinely employed to test hypotheses about means, correlation coefficients, and multiple regression coefficients.

Sample-size determination requires accurate information about the size of an effect in the population (Cohen, 1988; Kraemer & Thiemann, 1987; Murphy & Myors, 1998). However, theory in the behavioral sciences is often not sufficiently well developed to be able to provide precise predictions about the magnitude of an effect. This uncertainty creates an obstacle for researchers who wish to ensure that their experiments have an adequate amount of power. If, as commonly occurs, only the direction of an effect is predicted, sample-size determination cannot proceed.

In an attempt to get around this problem, researchers have made use of effect-size estimates from previous studies in the same area. This is a reasonable and justifiable strategy. Unfortunately, the way in which researchers subsequently employ the estimate is frequently inappropriate. Thus, it is common practice for researchers to substitute in the standard sample-size-determination formula an estimate of effect size derived from a previous experiment. A problem with this approach is that the traditional sample-size-determination formula was not designed to deal with the uncertainty inherent in an effect-size estimate. Effect-size estimates display random variation around the true effect size. Consequently, estimate-substitution in the traditional sample-size-determination formula can lead to a substantial loss of power. Gillett (1995, p. 383) provided an example that demonstrated the extent of the power deficit in a *t* test of the difference between two means.

The reason for the loss of power is the nonlinear relationship between power and effect size. For a given sample size, power increases with effect size in a negatively accelerating manner. Thus, a reduction in sample size by a given amount has a greater (negative) impact on power than does the (positive) impact of an increase in sample size by the same amount. As already noted, effect-size estimates vary randomly around the true effect size. Consequently, the power *loss* of studies using a *higher-than-average* effect-size estimate (and, hence, fewer subjects) is much greater than the power *gain* of studies using a *lower-than-average* effect-size estimate (and, hence, more subjects). The net effect is that the average power of studies across a discipline is considerably lower than the nominal level.

The practice of running a small pilot study for the purpose of obtaining an effect-size estimate is particularly problematic. The small size of a pilot study produces a large amount of variance in the effect-size estimate. If the estimate is substituted in the standard sample-size-determination formula, the expected power loss is large. By using the techniques developed by Gillett (1995), it can be shown that the practice of entering a pilot-study estimate in the traditional sample-size formula leads to an average power deficit of at least 10 percentage points.

A more statistically rigorous method of sample-size determination for *t* tests that is able to incorporate effect-size

This article was written during a period of study leave granted by the University of Leicester. Correspondence concerning this article should be addressed to R. Gillett, School of Psychology, University of Leicester, Leicester LE1 7RH, U.K. (e-mail: rtg@le.ac.uk).

estimates from a previous study without loss of power has been developed by Gillett (1995). The only information required by the method is the *t* value and sample size from a previous study.

In this approach, the test statistic and sample size from the earlier study are used to establish a distribution of probable effect sizes. The sample size to be employed in the new study is that which provides an expected power of the desired amount over the distribution of probable effect sizes. The adjective *expected* has a special statistical meaning which corresponds roughly to *average*. In other words, the aim of the expected-power approach is to find the sample size that supplies an average power of the required level across the range of likely effect sizes. By contrast, the objective of the estimate-substitution approach is to choose the sample size that has the desired power to detect a single effect size given by the estimate from the earlier study.

A simplified example may help to clarify the difference between the two approaches. Suppose that a power of 80% is required and that an effect-size estimate of 0.5 from a previous study indicated that the true effect size could be one of three equally likely values: 0.4, 0.5, 0.6. The estimate-substitution approach considers only the original estimate of 0.5 and calculates the sample size that delivers 80% power for that value. On the other hand, the expected-power method takes into account the fact that any of the three values could be the true effect size. It chooses the sample size that provides 80% power on average across the three effect sizes. Thus, a sample size that supplied respective powers of, say, 64%, 83%, and 93% for the three effect sizes, yielding an average of 80%, would be chosen.

The expected-power method may be used to provide the sample size for a main experiment on the basis of the result of a pilot study, or to determine the sample size for a replication attempt. An advantage of the approach is that it guarantees that, over a discipline as a whole, the average power of all studies is equal to the desired nominal level. A proof of this result is given in Gillett (1994).

THE EXPECTED-POWER METHOD

The expected-power approach uses Bayesian methods to establish the distribution of probable effect sizes. Bayesian techniques have often proved helpful in solving design problems in classical statistical procedures. Bayesian insights may be exploited without abandoning the classical inference model. However, from the point of view of a user of the computer program, the only Bayesian construct required is the *prior distribution.*

The prior distribution represents the frequency with which different effect sizes occur in the experiments carried out in a discipline. It is termed a prior distribution because it represents the state of knowledge about an effect size δ before the first experiment was conceived. The expected-power method requires a researcher to specify a prior distribution for an effect size δ. The variance of the prior distribution is termed the *prior variance.* The prior variance represents the spread of effect sizes that are found across a discipline.

The prior distribution provides a very rough, ballpark estimate of the range of values within which an effect size is likely to lie. When information about an effect size is available (e.g., in the form of a *t* value), it is combined with the information from the prior distribution to form a distribution of probable effect sizes, termed the *posterior distribution.* The expected power supplied by a sample size is calculated across the posterior distribution of probable effect sizes. The prior distribution enables an effect-size estimate to be used more effectively, by influencing the location and spread of the posterior distribution. The less reliable an estimate from a previous study is (e.g., the smaller its sample size), the greater the influence of the prior on the posterior will be.

What might be a reasonable choice of prior for the behavioral sciences? The prior distribution represents the state of knowledge about effect size *before the first experiment was conducted.* Since our concern is with studies whose effect size is unknown before the original experiment, and since it is arbitrary whether the first mean is subtracted from the second mean, or vice versa, it follows that the prior distribution must be symmetric about the origin. That is, the prior mean is zero.

A natural choice of prior distribution is the normal distribution. Empirical evidence indicates that the a priori likelihood of an effect tends to be inversely related to its size (Haase, Waechter, & Solomon, 1982). Effect sizes typically encountered in psychology are very roughly normally distributed with zero mean and variance of the order $\sigma_\delta^2 \leq 1$ (Gillett, 1986). Hence, a normal distribution is a reasonable choice for the prior distribution. The sample-size-determination program described in this article allows a researcher to choose a normal prior with zero mean and a value for the prior variance that is appropriate in the light of current knowledge of the distribution of effect sizes in a discipline.

A value of $\sigma_\delta^2 = 1$ is a useful preliminary estimate for the prior variance. This figure is consistent with the frequency distribution of 11,044 effect sizes compiled by Haase et al. (1982). It is well known, however, that such surveys tend to overestimate the variance of δ because they focus largely on published studies. Selection for publication frequently requires the attainment of statistical significance. Censoring of nonsignificant studies leads to an inflated estimate of effect size (Hedges,1984; Lane & Dunlap, 1978). Hence, the value $\sigma_\delta^2 = 1$ is almost certainly a high estimate for the prior variance.

Another possible choice for the role of prior distribution is the uniform distribution. The uniform distribution may be viewed as a special case of the normal distribution in which the variance is very large ($\sigma_\delta^2 = \infty$). Insofar as the uniform distribution is the limiting form of the normal distribution, it can be argued that a uniform prior represents the most liberal assumption that is compatible with the available data on effect sizes. A uniform prior yields a smaller sam-

Table 1
TEPSAM Input Dialog Sequence

1. Enter the *t* value from the previous study
2. Enter *N*, the total number of participants used in the *t* test from the previous study
3. The degrees of freedom for a *t* test are $N-d$, where, for example,
$d = 1$ for a related (or paired) *t* test, or a one-sample *t* test
$d = 2$ for an independent *t* test, or a two-sample *t* test
$d = 2$ for a *t* test of a correlation or simple regression coefficient
$d = k+1$ for a *t* test of a *b* coefficient in multiple regression
where *k* is the number of predictor variables
Enter the *d* value
4. Enter the required power (e.g., .80)
5. Enter alpha, the required significance level (e.g., .05)
6. For normal prior, enter variance (e.g., 1.0)
For uniform prior, enter 99

ple size than a normal prior (Gillett, 1986). Hence, the sample size supplied under a uniform prior represents a *lower limit,* below which the required expected power cannot be obtained on any scientifically reasonable assumption. Therefore, the sample size for a study should not be allowed to fall below the value yielded by a uniform prior.

Computer Program

A computer program written in FORTRAN 77 is available to perform sample-size calculations by using the expected-power method. The program, which is called TEPSAM, runs in a DOS window on a PC or on a Power Macintosh with Windows emulation software. It has been tested on PCs running DOS 5.0, DOS 6.0, Windows 95/98, Windows 2000, and on a Power Macintosh running SoftWindows 95 Version 3.0. A file containing an executable copy of the program, TEPSAM.EXE, may be obtained by sending an e-mail request to the author at rtg@le.ac.uk.

To run the program, first copy the file TEPSAM.EXE to a folder, or directory, of your choice. Open a DOS window and go to that folder. To start the program, simply type TEPSAM and press the enter/return key on the keyboard. (If you prefer, you may use lower case when typing the name of the program, since DOS is not case sensitive.) A series of prompts follows, requesting information about both the previous study and the planned study. For example, the first prompt asks the user to *Enter the t value from the previous study*. Details of the TEPSAM input dialog sequence are provided in Table 1.

Once all the information that the program requires has been entered, computation begins. In most cases, the result will appear almost immediately. Sometimes, when a large sample size is required, there will be a delay. In this event, a *Please Wait* sign is displayed and the output pauses until execution has completed, whereupon the result will appear on the screen.

Occasionally, the algorithm might encounter difficulties. Where there is a risk of numerical overflow or where convergence would take an unacceptably long time, an approximation based on the normal distribution is used and the result is accompanied by statement to that effect. The approximation is described in Gillett (1991). Because the power functions of the normal distribution and the *t* distribution are very similar (Wahlsten, 1991), the approximation is sufficiently accurate for the purpose of sample-size determination in the situations where it is likely to be required.

The output from TEPSAM contains a summary of information about the previous study and a summary of the parameters of the new study that is being planned, followed by a value for *N* indicating the *total* number of participants required. For example, in an independent *t* test, where the means of two unrelated groups are compared, the number of participants per group would be *N*/2. A sample of output from TEPSAM is given in Table 2.

Practical Examples

Related (paired) *t* test. A researcher wishes to compare the performance of the same group on two separate occasions. A similar study in the literature, with 25 participants, obtained a *t* value of 4.2. The researcher considers that a normal prior with a variance of $\sigma_\delta^2 = 0.5$ provides a reasonable representation of the distribution of effect sizes in psychology. How many participants should be recruited for the new study to ensure power of .80 at a significance level of $\alpha = .05$? The values entered in response to TEPSAM's six prompts are 4.2, 25, 1, 0.80, 0.05, and 0.5. The resulting output from the program states that the total number of participants required in the new study is $N = 19$.

Independent *t* test. In preparation for a larger experiment, a pilot study was run in which 20 participants were randomly assigned to two groups, 10 in each. A *t* test of the difference between the means returned a value of $t = 2.4$. The researcher considers that a normal prior with a variance of $\sigma_\delta^2 = 1.0$ is representative of the distribution of effect sizes in psychology. How many participants should be recruited for the main experiment to ensure power of .90 at a significance level of $\alpha = .05$? The values entered

Table 2
Sample Output from TEPSAM

Previous Study	$t = 2.00$ $N = 54$ $df = 52$
New Study	Power = .80 Significance level = .05
	Prior distribution: Normal with variance = 1.00
	Computation in progress – Please Wait
	Computation completed
	Total sample size required is $N = 227$

in response to TEPSAM's six prompts are 2.4, 20, 2, 0.90, 0.05, and 1.0. The output from the program states that the total number of participants required in the main experiment is $N = 166$. That is, 83 in each group.

Correlation. A researcher has come across an experiment in the literature that investigated the relationship between perceived attitude similarity and degree of liking. A total of 35 participants rated their degree of liking for an unknown person on the basis of that person's attitude profile, which the experimenter had arranged to overlap to a different degree, from 0% to 100%, with each participant's own profile. A *t* test of the hypothesis that the correlation between similarity and liking was greater than zero yielded a value of $t = 2.1$. Assuming a normal prior with a variance $\sigma_\delta^2 = 0.5$, how many participants should be recruited for a new experiment to ensure power of .80 at a significance level of $\alpha = .05$? The values entered in response to TEPSAM's six prompts are 2.1, 35, 2, 0.80, 0.05, and 0.5. The output from the program states that the total number of participants required in the new experiment is $N = 184$.

A *b* coefficient in multiple regression. In an investigation of the influence of the study habits of 20 students on subsequent exam performance, a researcher regressed exam performance on average daily time spent studying and number of academic books purchased. The *t* value for the *b* coefficient of number of books purchased was $t = 1.7$, which was *not* significant at $\alpha = 0.05$. Suspecting that low power, owing to the small sample size, prevented the *t* test from reaching significance, the researcher plans a new study. Assuming that a normal prior with a variance of $\sigma_\delta^2 = 1.0$ is roughly representative of the distribution of effect sizes in psychology, how many students should be recruited for the new study to ensure power of .80 at a significance level of $\alpha = .05$? The values entered in response to TEPSAM's six prompts are 1.7, 20, 3, 0.80, 0.05, and 1.0. The output from the program states that the total number of student participants required in the new study is $N = 188$.

The sample size figure produced by the program assumes that the same predictor variables will be used in the new study. Note that a *t* test of a *b* coefficient tells us whether the part of the variance of the dependent variable that a predictor is able to explain contains a unique area that does not overlap with areas explained by other predictor variables. Clearly, if another predictor were to correlate moderately with the predictor variable of interest, and if that predictor were less important to the aims of the investigation, then power could be increased simply by removing the other predictor from the multiple regression. Removal of an overlapping predictor increases the size of the part of dependent variable variance that is uniquely explained by the predictor of interest.

Multiple Regression Assumptions

The dependent variable and the predictor variables are assumed to have a joint multivariate normal distribution. The conditional model of multiple regression is assumed, in which the values of the predictor variables are characterized as fixed and variability resides solely in the dependent variable. Strictly, the findings of the conditional analysis apply only to that section of a population with the same observed scores on the predictor variables. By contrast, the unconditional model treats all variables, both independent and dependent, as truly variable. The conclusions of the unconditional model apply to the whole population from which the participants are sampled.

Although the unconditional model will usually be the more appropriate approach in studies in the behavioral sciences, the conditional model provides a reasonable approximation and is more tractable. Cohen's (1988) power tables assume the conditional model and also employ an additional noncentral χ^2 approximation to the noncentral *F* distribution. In a comparison of exact sample sizes derived from the unconditional model with sample-size values obtained using Cohen's tables, Gatsonis and Sampson (1989) found that "the approximations are quite accurate in many situations of practical interest" (p. 524).

The present approach employs the exact noncentral *F* distribution instead of the noncentral χ^2 approximation used by Cohen (1988). Hence, sample-size values should provide an even better approximation to those of the unconditional model.

It is important to note that Cohen's (1988) tables are based on the traditional sample-size formula and, hence, differ from sample sizes obtained by the expected-power method employed in the present program. Thus, using an effect-size estimate from a previous study in Cohen's tables would produce a power deficit, for the reasons mentioned above.

Algorithm and Functions

The main algorithm employed in the program is described in Gillett (1995, Section 2). Functions used in the program include the root-finding function RTFLSP (Press, Teukolsky, Vetterling, & Flannery, 1992, p. 349) and algorithms AS63, AS109, AS111, and AS245 from the Applied Statistics section of the StatLib (1999) software collection at http://lib. stat.cmu.edu/apstats. AS63 calculates the incomplete beta function ratio. AS109 computes the inverse of the incomplete beta function. AS111 yields the normal deviate corresponding to a lower tail area of the normal distribution. AS245 produces the natural logarithm of the gamma function.

Availability

A file containing an executable copy of the program, TEPSAM.EXE, may be obtained by sending an e-mail request to the author at rtg@le.ac.uk.

CONCLUSION

The expected-power technique, a method of sample-size determination designed to handle uncertainty in effect-size estimates, was described. The procedure uses the *t* value and sample size from a previous study to es-

tablish a distribution of probable effect sizes. The sample size to be employed in the new study is that which supplies an expected power of the desired amount over the distribution of probable effect sizes. An advantage of the expected-power approach is that it guarantees that, over a discipline as a whole, the average power across all studies using the technique is equal to the desired nominal level. A FORTRAN 77 program was presented that permits rapid calculation of sample size for a variety of *t* tests, including independent *t* tests, related *t* tests, *t* tests of correlation coefficients, and *t* tests of multiple regression *b* coefficients.

REFERENCES

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Gatsonis, C., & Sampson, A. R. (1989). Multiple correlation: Exact power and sample size calculations. *Psychological Bulletin*, **106**, 516-524.

Gillett, R. (1986). Sample size determination in replication attempts: The standard normal *z* test. *British Journal of Mathematical & Statistical Psychology*, **39**, 190-207.

Gillett, R. (1991). A FORTRAN 77 program for sample-size determination in replication attempts when effect size is uncertain. *Behavior Research Methods, Instruments, & Computers*, **23**, 442-446.

Gillett, R. (1994). An average power criterion for sample size estimation. *The Statistician*, **43**, 389-394.

Gillett, R. (1995). The expected power of *F* and *t* tests conditional on information from an earlier study. *British Journal of Mathematical & Statistical Psychology*, **48**, 371-384.

Haase, R. F., Waechter, D. M., & Solomon, G. S. (1982). How significant is a significant difference? Average effect size of research in counseling psychology. *Journal of Counseling Psychology*, **29**, 58-65.

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effect of censoring studies yielding insignificant mean differences. *Journal of Educational Statistics*, **9**, 61-85.

Kraemer, H. C., & Thiemann, S. (1987). *How Many Subjects? Statistical Power Analysis in Research*. Newbury Park, CA: Sage.

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical & Statistical Psychology*, **31**, 107-112.

Murphy, K. R., & Myors, B. (1998). *Statistical power analysis*. Mahwah, NJ: Erlbaum.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in FORTRAN: The art of scientific computing* (2nd ed). Cambridge: Cambridge University Press.

StatLib Applied Statistics Algorithms (1999) [On-line]. Available: http://lib.stat.cmu.edu/apstat.

Wahlsten, D. (1991). Sample size to detect a planned contrast and a one degree-of-freedom interaction effect. *Psychological Bulletin*, **110**, 587-595.

(Manuscript received October 20, 2000;
revision accepted for publication August 4, 2001.)