Behavior Research Methods, Instruments, & Computers
1991, 23 (3), 442-446

A FORTRAN 77 program for sample-size determination in replication attempts when effect size is uncertain

RAPHAEL GILLETT
University of Leicester, Leicester, Great Britain

A method of sample-size determination for use in attempts to replicate experiments is described. It is appropriate in situations where there is uncertainty about the magnitude of the effect under investigation. The procedure uses information supplied by the original experiment to establish a distribution of probable effect sizes. The sample size to be used in a replication study is that which provides an expected power of the desired amount over the distribution of probable effect sizes. A FORTRAN 77 program is presented that permits rapid calculation of sample size in replication attempts employing comparisons of means, correlation coefficients, or proportions.

A replication attempt is a study conducted to establish whether or not a previous finding, which may or may not have reached statistical significance, represents a genuine effect. Replication studies play a crucial role in scientific research by providing an independent test of theoretical predictions and by helping to build a body of trustworthy results (Lykken, 1968; Rosenthal, 1979).

The power of a study to detect an experimental effect is a direct function of the sample size chosen by the researcher, other things being equal (Cohen, 1988). Unfortunately, it has been shown that "most psychologists have an exaggerated belief in the likelihood of successfully replicating an obtained finding" (Tversky & Kahneman, 1971, p. 105). Consequently, researchers often seriously underestimate the sample size required to achieve an acceptable probability of uncovering the presumed effect.

The problem is compounded by the fact that in the behavioral sciences there is often a lack of precise, well-formulated theories that are capable of specifying unequivocally the expected magnitude of an experimental effect. Uncertainty on the part of a researcher concerning the size of the effect that is under investigation increases the likelihood of an unsuitable choice of sample size.

To remedy this situation, Gillett (1986) has developed a method of sample-size determination for replication attempts, which is appropriate when there is uncertainty about the magnitude of the predicted effect. It is assumed that the same population effect size underlies the original experiment and the replication attempt. Information from the original study is used to determine a distribution of probable effect sizes. The sample size to be employed in the replication attempt is that which supplies an expected power of the desired amount over the distribution of probable effect sizes. The posterior distribution of effect size is established using Bayes's theorem. Hence, the technique requires the researcher to specify a prior distribution for effect size.

Correspondence should be addressed to Raphael Gillett, Department of Psychology, University of Leicester, Leicester LE1 7RH, Great Britain.

Choosing a Prior Distribution

The prior distribution represents the state of knowledge about effect size before the original 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.

Even where there is some previous information about effect size, it is still advisable to choose a prior density that is symmetric about the origin. A replication attempt should be independent of the experimenter's preconceptions if it is to fulfill its scientific function of providing an unbiased test of the original result.

It is important to realize that the role of the prior distribution in sample-size determination is quite different from its more familiar role in hypothesis testing. Consider the consequences that would follow if a prior with a nonzero mean were chosen. In both sample-size determination and hypothesis testing, a prior with a nonzero mean produces a larger posterior effect size than a prior with a mean of zero, assuming that both the prior mean and the result of the experiment have the same sign. In hypothesis testing, this leads to an *increase* in the probability of detecting an effect. In sample-size determination, however, it causes a *reduction* in the likelihood of detecting an effect because the larger the posterior effect size appears to be, the smaller the sample size required to detect it.

An experimenter who *fails* to replicate an earlier finding is thus placed in an awkward position when a prior with a nonzero mean is used. Such a researcher never knows whether the failure to replicate indicates that the original finding was spurious or whether the prior mean was simply set too high (thereby producing a small sample size with insufficient power to uncover the effect). This ambiguity can only be removed by locating the prior mean at zero.

A natural candidate for the role 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. Effect sizes typically encountered in psychology are very roughly normally distributed with zero mean and unit variance (Gillett, 1986). Hence, a normal distribution is a reasonable choice for the prior distribution. The sample-size program listed in the Appendix allows a researcher to choose a normal prior with zero mean and any value for the variance that seems appropriate (e.g., $\sigma^2 = 1$).

A second candidate for the role of prior distribution is the uniform distribution. It is well known that classical statistical inference concerning the mean yields the same result as the corresponding Bayesian technique in which a uniform prior is assumed for the mean. Thus, from a Bayesian perspective, classical hypothesis testing operates with the uniform distribution as an implicit, undeclared prior distribution.

The uniform distribution may be viewed as a special case of the normal distribution in which the variance is very large. 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. Moreover, a uniform prior yields a smaller sample size than does 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 replication attempt should not be allowed to fall below the value yielded by a uniform prior.

Sample-Size Determination

The sample-size determination procedure may be used in replication attempts involving the comparison of means, the comparison of correlation coefficients, and the comparison of proportions. To illustrate the approach, we will consider the situation where it is desired to replicate a previously obtained difference between the means of two independent samples from populations sharing the same known variance.

Let z denote the standard normal score obtained in an earlier experiment that had compared the means of two groups each containing n subjects. Let m represent the number of subjects per group in a study attempting to replicate the earlier result. Under a normal prior with zero mean and variance σ^2, the expected power $P_N(m; z, n)$ of a replication attempt employing m subjects per group, given the values z and n of the original study, is shown by Gillett (1986) to be

$$P_N(m; z, n) = Q\left(\frac{c - zw\sqrt{\frac{m}{n}}}{\sqrt{\frac{mw}{n} + 1}}\right) + Q\left(\frac{c + zw\sqrt{\frac{m}{n}}}{\sqrt{\frac{mw}{n} + 1}}\right),$$

where $Q(y)$ is the probability that a standard normal variate exceeds the value y, where $w = n\sigma^2/(n\sigma^2+2)$, and where c is the standard normal critical value associated with the desired level of significance. To find the sample size that supplies a desired amount of power, successive trial values of m are entered into the formula until the required level of power is reached.

Since this process can be time-consuming, it would be helpful to use a computer program to perform the calculations on a microcomputer. Accordingly, a FORTRAN 77 program titled ZREPSAM is provided in the Appendix to compute the sample size for a replication study. ZREPSAM may be used for one- and two-sample designs, for z and t tests, and for comparisons of means, comparisons of correlation coefficients, and comparisons of proportions. A practical example illustrates ZREPSAM's mode of operation.

Example

Suppose that an experiment comparing the means of two groups, each containing $n = 28$ subjects, achieved the result $t = 3.6$. In an attempted replication of this experiment, how many subjects should be enlisted per group in order to achieve an expected power of 0.80 under a uniform prior, if a two-tailed test at the $\alpha = 0.05$ significance level is desired?

Run the program by typing its name, ZREPSAM, at the DOS prompt. The program then prompts the user for the following data: (1) the z or t value obtained in the previous study; (2) the sample size n used in the earlier study; (3) the power P required in the replication attempt; (4) the critical value c of the z (not t) statistic associated with the desired significance level; and (5) the variance σ^2 of the prior distribution.

In the present example, these quantities are $t = 3.6$, $n = 28$, $P = 0.80$, $c = 1.96$, and $\sigma^2 = 99$. (A value of $\sigma^2 = 99$ denotes the choice of a uniform prior.)

The program displays the result $m = 23$. This means that a sample size of 23 subjects per group should be used to yield an expected power of 0.80 over the distribution of probable effect sizes.

Further examples of the application of the technique to comparisons of correlation coefficients and comparisons of proportions are provided in Gillett (1986).

Accuracy

The value of m must be an integer. Hence, m is calculated so that $P_N(m-1; z, n) < P_N(m; z, n)$. In the case of the ordinary z test, m is accurate to at least five significant figures. In other words, sample sizes of $m < 100{,}000$ are completely accurate. In the approximation to the t test, m is accurate to within 1 subject, for $n > 10$ and $t > 1.6$, and all errors are conservative (i.e., yield a level of power that is greater than the nominal value). The accuracy of m in approximate tests of correlation coefficients and in approximate tests of proportions is a function of the goodness of fit of the transformed values to the normal distribution. In these tests, the margin of error in m is commonly less than the larger of either 3 subjects or 3%, provided that $n > 20$.

Minimum Effect Size

The following consideration should be borne in mind when determining the sample size for a replication attempt. When the combined evidence of z and n suggests that the population-effect size is small, the present procedure will necessarily indicate that a large sample size is required to detect the effect. Since researchers may wish to avoid dissipating resources in pursuit of very small effects, they might consider specifying in advance the mini-

mum effect size that they would deem to be nontrivial. They could then consult the tables in Cohen (1988) to determine the sample size required to detect this effect. In this way, an uppper limit for m may be obtained. Should the value for m recommended by the present method exceed this upper limit, as it might do if the population-effect size is very small, then researchers could use the upper limit instead. Such a strategy would ensure prudent use of resources.

For example, consider a comparison of means using a t test, for which a power of $P = 0.8$ at $\alpha = 0.05$ (two-tailed) is desired. Cohen (1988) defines an effect size of $\delta = 0.8$ as *large*, an effect size of $\delta = 0.5$ as *medium*, and an effect size of $\delta = 0.2$ as *small*. Suppose that the minimum effect size that a researcher deems to be nontrivial is $\delta = 0.1$. Then Table 2.4.1 in Cohen (1988) indicates that a sample size of 1,571 per group is required to detect an effect as small as $\delta = 0.1$. The value 1,571 could thus serve as an upper limit to m, only to be used in the replication attempt if the sample size indicated by the present technique exceeds it.

Availability

A disk copy of the program in the Appendix can be obtained without charge from Raphael Gillett, Department of Psychology, University of Leicester, Leicester LE1 7RH, Great Britain (e-mail: JANET:rtg@leicester.ac.uk). Please send a formatted (MS-DOS 2.0 or later) 5.25-in. disk, a self-addressed disk mailer, and return postage (e.g., International Reply Coupons, obtainable at the post office).

REFERENCES

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

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

LYKKEN, D. (1968). Statistical significance in psychological research. *Psychological Bulletin*, **70**, 151-159.

ROSENTHAL, R. (1979). Replications and their relative utilities. *Replications in Social Psychology*, **1**, 15-23.

TVERSKY, A., & KAHNEMAN, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, **76**, 105-110.

APPENDIX
Listing of Program ZREPSAM

```
      PROGRAM ZREPSAM
C
C     GIVEN THE VALUES Z (OR T) AND N FROM A PREVIOUS EXPERIMENT, THE
C     PROGRAM CALCULATES THE SAMPLE SIZE M THAT ENSURES THAT A REPLICATION
C     ATTEMPT WILL HAVE AN EXPECTED POWER P OF DETECTING AN EFFECT AT A
C     REQUIRED LEVEL OF SIGNIFICANCE.
C
C     NOTE THAT P REPRESENTS THE EXPECTED, OR 'AVERAGE', POWER OVER THE
C     DISTRIBUTION OF PROBABLE EFFECT SIZES, RATHER THAN THE POWER OF THE Z
C     (OR T) TEST AGAINST A SPECIFIC EFFECT. THE PROGRAM MAY THEREFORE BE
C     USED WHENEVER THERE IS UNCERTAINTY ABOUT THE MAGNITUDE OF THE EFFECT
C     UNDER INVESTIGATION, AS COMMONLY OCCURS WHEN THEORY IN AN AREA IS
C     INSUFFICIENTLY ADVANCED TO PROVIDE PRECISE PREDICTIONS OR WHEN THE
C     ORIGINAL EXPERIMENT WAS SIMPLY AN EXPLORATORY ONE.
C
C     A CHOICE OF UNIFORM OR NORMAL PRIOR DISTRIBUTION FOR EFFECT SIZE IS
C     OFFERED. IF A NORMAL PRIOR IS SELECTED, ITS VARIANCE IS REQUIRED.
C     (A VALUE OF VARIANCE = 1 IS CONSISTENT WITH THE RANGE OF EFFECT SIZES
C     TYPICALLY ENCOUNTERED IN PSYCHOLOGY.) IF A UNIFORM PRIOR IS SELECTED
C     IT SHOULD BE BORNE IN MIND THAT, SINCE A UNIFORM PRIOR REPRESENTS THE
C     MOST LIBERAL ASSUMPTION THAT IS SCIENTIFICALLY ACCEPTABLE, THE VALUE OF
C     M OBTAINED ON THIS ASSUMPTION REPRESENTS A LOWER LIMIT BELOW WHICH
C     SAMPLE SIZE SHOULD NOT BE PERMITTED TO FALL.
C
C     REFERENCE: GILLETT, R. (1986)  SAMPLE SIZE DETERMINATION IN REPLICATION
C                ATTEMPTS: THE STANDARD NORMAL Z TEST.  BRITISH JOURNAL OF
C                MATHEMATICAL AND STATISTICAL PSYCHOLOGY, 39, 190-207.
C
      WRITE(*,10)
10    FORMAT(/' ENTER Z (OR T) VALUE FROM PREVIOUS STUDY')
      READ(*,*)Z
      WRITE(*,20)
20    FORMAT(/' ENTER N FROM PREVIOUS STUDY')
      READ(*,*)N
      WRITE(*,30)
```

```
30   FORMAT(/' ENTER REQUIRED POWER (E.G., 0.80)')
     READ(*,*)P
     WRITE(*,40)
40   FORMAT(/' ENTER CRITICAL Z VALUE (N.B. NOT T) ASSOCIATED WITH'/
    $' REQUIRED SIGNIFICANCE LEVEL (E.G., 1.96)')
     READ(*,*)C
     WRITE(*,50)
50   FORMAT(/' FOR NORMAL PRIOR, ENTER VARIANCE < 99 (E.G., 1)'
    $/' FOR UNIFORM PRIOR, ENTER 99')
     READ(*,*)V
     WRITE(*,60)Z,N,P,C
60   FORMAT(////' PREVIOUS STUDY:   Z = ',F5.2,7X'N =',I7,
    $///' REPLICATION ATTEMPT:'//' POWER = ',F4.2,7X'CRITICAL Z',
    $' VALUE = ',F5.2/)
     IF (V.LT.99) THEN
     WRITE(*,70)V
70   FORMAT(' PRIOR DISTRIBUTION:   NORMAL WITH VARIANCE = ',F4.1)
     ELSE
     WRITE(*,80)
80   FORMAT(' PRIOR DISTRIBUTION:   UNIFORM')
     ENDIF
     CALL SAMPSIZE(Z,N,P,C,V,M,MR)
     WRITE(*,90)M,M+2,MR,M
90   FORMAT(/' STATISTICAL TEST (ONE OR TWO SAMPLE)'17X'SAMPLE SIZE'
    $//5X'Z TEST FOR COMPARISON OF MEANS                    ',I8,//
    $5X'T TEST FOR COMPARISON OF MEANS (*)                  ',I8,//
    $5X'Z TEST FOR COMPARISON OF CORRELATIONS (**)          ',I8,/
    $5X'(USING FISHER"S TRANSFORMATION)'//
    $5X'Z TEST FOR COMPARISON OF PROPORTIONS (**)           ',I8,/
    $5X'(USING ARCSIN TRANSFORMATION)'//
    $5X'(*),(**) APPROXIMATE TESTS:   REQUIRE PREVIOUS N > 10,20'/)
     END

     SUBROUTINE SAMPSIZE(Z,N,P,C,V,M,MR)
C
C
     CALCULATES RK, THE SQUARE ROOT OF M/N, FROM WHICH THE SAMPLE SIZE
C    M FOR A REPLICATION ATTEMPT MAY BE OBTAINED AS M = RK*RK*N
C
     AN=N
     W=1.0
     IF (V.LT.99) W=V*AN/(V*AN+2.0)
C
C    CALCULATE FIRST APPROXIMATION TO RK
C
     X=1024
     DO 50 I=1,20
     G=X
     X=X/2.0
     H=X
     CALL EP(Z,P,C,W,X,D,DD)
     IF (D.GT.0) GOTO 80
50   CONTINUE
     M=1
     MR=3
     RETURN
C
C    USE NEWTON'S METHOD TO CONVERGE ON RK
C
60   X=RK
70   CALL EP(Z,P,C,W,X,D,DD)
```

```
80    RK=X-D/DD
C
C     IF CONVERGENCE FAILS, CHOOSE BETTER FIRST APPROXIMATION
C
      IF (RK.LT.0.0) THEN
      H=H/2.0
      X=G-H
      GOTO 70
      END IF
      IF (ABS(D).GT.0.000001) GOTO 60
      M=INT(RK*RK*AN+1.0)
      MR=INT(RK*RK*(AN-3.0)+4.0)
      RETURN
      END

      SUBROUTINE EP(Z,P,C,W,X,D,DD)
C
C     GIVEN RK, THE SUBROUTINE CALCULATES D, THE DIFFERENCE BETWEEN THE
C     ACTUAL AND REQUIRED EXPECTED POWER, AND ALSO ITS DERIVATIVE DD
C
      SK=SQRT(X*X*W+1)
      S1=(C-X*Z*W)/SK
      S2=(C+X*Z*W)/SK
      D=P-UN(S1)-UN(S2)
      D=-EXP(-0.5*S1*S1)*(Z+C*X)+EXP(-0.5*S2*S2)*(Z-C*X)
      DD=DD*W*0.39894228/SK**3
      RETURN
      END

      FUNCTION UN(Y)
C
C     PROBABILITY THAT A STANDARD NORMAL DEVIATE EXCEEDS Y
C
      A=ABS(Y)
      IF (A.GT.6.0) A=6.0
      B=0
      S=A*0.4714045208
      DO 10 I=1,13
      R=REAL(I)-0.5
      G=EXP(-R*R/9)*SIN(R*S)/R
      B=B+G
10    CONTINUE
      UN=0.5-0.3183098862*B
      IF (Y.LT.0.0) UN=1.0-UN
      RETURN
      END
```

(Manuscript received November 6, 1989;
revision accepted for publication April 4, 1991.)