Confidence intervals and standard errors for ratios of normal variables

WILLIAM P. DUNLAP and N. CLAYTON SILVER
Tulane University, New Orleans, Louisiana

Occasionally, one has need for confidence intervals or tests of significance for ratios of normal variables. For instance, tests of significance on the *therapeutic safety ratio* are needed in pharmacology. This ratio is the dose of a compound that is "lethal" to 50% of subjects tested (LD_{50}) divided by the dose that is "effective" for 50% of treated subjects (ED_{50}). The bigger this ratio is, the larger the lethal dose is compared with the effective dose and, therefore, the safer the drug. Although it is probably safe to assume that estimates such as ED_{50} and LD_{50} have approximately normal distributions, the distribution of the ratio of two normal variables is quite complex. This distribution has been investigated first by Geary (1930), then more extensively by Fieller (1932), and more recently by a number of researchers (see, e.g., Hinkley, 1969; Marsaglia, 1965; Paulson, 1942; Shanmugalingam, 1982). Because of Fieller's extensive early work, the general theory of the distribution of the ratio of two normal variables is called "Fieller's Theorem" by some authors (see, e.g., Davies, 1961).

Fieller's Theorem. A good approximation to the confidence interval for a ratio of normal variables can be solved in the following manner. For normal variables y and x, call their ratio v, where

$$v = y/x. \qquad (1)$$

Now consider the expression

$$d = y - vx, \qquad (2)$$

which, because it is a linear composite of normal variables y and x, must itself be normal. Furthermore, d must have a mean equal to zero. Using expected values, the variance of d can be shown to equal

$$s_d^2 = s_y^2 - 2v\text{Cov}(xy) + v^2 s_x^2. \qquad (3)$$

The quantities x and y might be single observations, in which case s_x^2 and s_y^2 are population variances. More frequently, however, x and y are sample means, in which case s_x^2 and s_y^2 are squared standard errors. Although Equation 3 theoretically should be expressed in terms of population values, in any practical application the population values will have to be estimated from samples; therefore, the symbol s^2 is used for variance in the expressions. The derivations that follow assume that x and

y are normally distributed, which is certainly tenable when x and y are means; the effects of nonnormality on the distribution of a ratio has not to our knowledge been investigated.

Because d in Equation 2 is normally distributed, Equation 2 divided by the square root of Equation 3 is distributed as Student's t (See Kendall & Stuart, 1979, pp. 137-138, for a derivation). Squaring both sides we can write

$$t^2 = (y - vx)^2 / (s_y^2 - 2v\text{Cov}(xy) + v^2 s_x^2). \qquad (4)$$

A confidence interval for the ratio v can be obtained by finding values of v that satisfy the above expression. By rearranging Equation 4, one can write a quadratic equation in terms of v, which can be solved for upper and lower values of the confidence interval using the quadratic formula.

Letting

$$Q = 1 - t^2 s_x^2 / x^2, \qquad (5)$$

the resulting solution is a confidence interval having a center at

$$C = (y/x - t^2 \text{Cov}(xy)/x^2)/Q, \qquad (6)$$

and a standard error,

$$SE = \{s_y^2 - 2y/x\text{Cov}(xy) + y^2/x^2 s_x^2 - t^2 s_x^2 / x^2[s_y^2 - \text{Cov}(xy)^2/s_x^2]\}^{1/2}/(xQ). \qquad (7)$$

Therefore, the confidence interval is

$$C \pm tSE. \qquad (8)$$

The above expressions will work well only when x is large relative to s_x. If this were not the case, Q in Equation 5 could even become negative, and one would be faced with the awkwardness of a negative estimated standard error.

Notice also that the above expressions permit x and y to be correlated; that is, $\text{Cov}(xy)$ does not equal zero. In many applied instances, x and y can be assumed to be independent; thus, terms containing $\text{Cov}(xy)$ will equal zero, and the expressions simplify considerably. Furthermore, if x is very large relative to both s_x and s_y, Q approaches one, and the final term involving t^2 in the standard error, Equation 7, can be dropped, yielding the following expressions, which are simple enough to use by hand:

$$C = y/x, \qquad (9)$$

and

$$SE = (s_y^2 + y^2/x^2 s_x^2)^{1/2}/x. \qquad (10)$$

The Program. The program (see the Appendix) uses the more complex and more accurate expressions (Equations 5 through 8) to estimate the standard error of a ra-

tio of two normal variables and a confidence interval for that ratio. Input required is x, s_x, y, s_y, and the correlation, r_{xy}. Also required is the value of Student's t at the particular level of confidence desired. To look up t, degrees of freedom must be used, which would be $N-1$, if N raw data values were used for both x and y estimates. For the case in which different numbers of data points are involved in numerator and denominator estimates, we have usually set N equal to the smaller number as a conservative measure. If very large numbers of data points are involved in both x and y, normal curve values, such as 1.96 with $\alpha = 0.05$, would suffice.

Accuracy. Because the accuracy of these computations depends on several factors, Monte Carlo simulations were performed to determine conditions under which the estimates computed by the program would be sufficiently accurate for applied work. A program was written to generate 10 thousand samples of ratios of random normal numbers having predetermined means, standard deviations, and numerator–denominator correlations. Whenever the numerator or denominator became negative, the case was discarded, because it was felt that this would most correctly represent real data. From these data, empirical standard errors and confidence intervals were determined under a variety of conditions.

Standard Error of a Ratio. In the first set of Monte Carlo simulations, we compared the "approximate" (Equations 9 and 10) to the "complete" (Equations 5 to 7) solution, with the correlation between numerator and denominator set equal to zero. It became readily apparent that the primary determinant of accuracy in both cases was the ratio of the denominator to its standard deviation, x/s_x. Any definition of "sufficiently accurate" is necessarily arbitrary, but we decided that estimated standard errors that deviated less than 10% from empirical standard errors were acceptable. Using this criterion, the denominator had to be six or more times its standard deviation for sufficient accuracy with the approximate solution. With the complete solution, x/s_x could be as low as four, and still provide reasonable precision. The ratio of the numerator to its standard deviation, y/s_y, although of much less consequence, appeared to work opposite to x/s_x; thus, somewhat poorer accuracy was seen for high values of y/s_y, and low values resulted in more accurate estimates of the standard error, which even became conservative for very low values of y/s_y.

Confidence Intervals. The approximate equations produced inaccurate confidence intervals for all combinations of conditions studied. Because the distribution of a ratio is positively skewed, the approximate solution resulted in too few scores in the lower tail, and too many scores in the upper tail. The complete solution produced confidence intervals that were quite accurate when x and y were large relative to s_x and s_y, respectively, and which became somewhat conservative under less ideal condi-

tions; that is, the actual frequencies of scores in either tail were less than 2.5% for a 95% confidence interval.

Correlation. When the numerator and denominator are correlated, only the complete solution applies. Simulation showed that estimated standard errors were quite good for correlations up to 0.6, but tended to underestimate empirical standard errors for larger correlations. Even with a correlation between y and x of 0.9, however, the underestimation was only 6%. Confidence intervals were quite accurate regardless of the correlation.

Conclusions. The following conclusions were reached: (1) The approximate equations are not recommended unless the ratio of the denominator to its standard deviation is quite large, and approximate confidence intervals will be asymmetric. (2) The complete solution provided by the present program provides useful estimates of the standard error of a ratio whenever the denominator is at least four times its standard deviation, and it also provides reasonably accurate confidence intervals that, at worst, become conservative in adverse cases. (3) The researcher is well advised to choose the variable with the largest value relative to its standard deviation as the denominator, and the variable with the smaller value relative to its standard deviation as the numerator when forming a ratio.

Requirements. The program is written in single-precision FORTRAN IV and runs on a DEC-2060, IBM 3050, and an IBM-PC, although specification of input and output units had to be modified for the different computers. Given such minor modifications, the program should run on any machine with a FORTRAN IV or later compiler.

Availability. A listing of the program may be obtained at no charge from William P. Dunlap. To receive the IBM-PC version, please send a double-sided, double-density, soft-sectored, 5¼-in. floppy disk along with your request.

REFERENCES

DAVIES, O. L. (1961). *Statistical methods in research and production*. London: Oliver & Boyd.

FIELLER, E. C. (1932). The distribution of the index in a normal bivariate population. *Biometrika*, **24**, 428-440.

GEARY, R. C. (1930). The frequency distribution of the quotient of two normal variates. *Journal of the Royal Statistical Society*, **93**, 442-446.

HINKLEY, D. V. (1969). On the ratio of two correlated normal random variables. *Biometrika*, **56**, 635-639.

KENDALL, M., & STUART, A. (1979). *The advanced theory of statistics* (Vol. 2, 4th ed.). New York: Macmillan.

MARSAGLIA, G. (1965). Ratios of normal variables and ratios of sums of uniform variables. *Journal of the American Statistical Association*, **60**, 193-204.

PAULSON, E. (1942). A note on the estimation of some mean values for a bivariate distribution. *Annals of Mathematical Statistics*, **8**, 440-445.

SHANMUGALINGAM, S. (1982). On the analysis of the ratio of two correlated normal variables. *The Statistician*, **31**, 251-258.

APPENDIX

```
C  THIS PROGRAM COMPUTES STANDARD ERRORS AND CONFIDENCE INTERVALS
C     FOR RATIOS OF NORMAL VARIABLES VIA FIELLER'S THEOREM
C
       OPEN (UNIT=9,DEVICE='TTY')
  1    WRITE(9,2)
  2    FORMAT (/' ENTER NUMERATOR, SD OF NUMERATOR (0,0 TO STOP)')
       READ(9,*)Y,SDY
       IF(Y.EQ.0.0)STOP
       WRITE(9,3)
  3    FORMAT(' ENTER DENOMINATOR, SD OF DENOMINATOR')
       READ(9,*)X,SDX
       Y2=Y*Y
       X2=X*X
       V=Y/X
       WRITE(9,4)
  4    FORMAT(' ENTER CORRELATION BETWEEN NUM. AND DEN.')
       READ(9,*)R
       VY=SDY*SDY
       VX=SDX*SDX
       COV=R*SDY*SDX
       WRITE(9,5)
  5    FORMAT(' ENTER YOUR T VALUE')
       READ(9,*)T
       T2=T*T
       Q=1.-T2*VX/X2
       C=(V-T2*COV/X2)/Q
       SE=SQRT(VY-2.*Y/X*COV+Y2/X2*VX-T2*VX/X2*(VY-COV**2/VX))/X/Q
       CI1=C-T*SE
       CI2=C+T*SE
       WRITE(9,6)V,C,SE,CI1,CI2
  6    FORMAT(/' RATIO =', F12.4/' ADJUSTED CENTER =', F12.4/
      *' STANDARD ERROR =' ,F12.4/' CONFIDENCE INTERVAL IS FROM',
      *  F12.4, ' TO' ,F12.4)
       GOTO 1
       END
```