Hypergeometric tests for 2 x k contingency tables

WILLIAM P. DUNLAP

Tulane University, New Orleans, Louisiana

The underlying probability distribution appropriate for a 2×2 contingency table depends upon how the data are collected, and whether the various marginal frequencies are fixed or are free to take on various possible values when the data are collected. Whether the row and/or column marginal frequencies are fixed or are free to vary determines classification into one of three possible experimental designs, called Cases I, II, and III. With Case I, the correlational model, a population is sampled randomly and two binomial aspects, such as sex (male or female) and cigarette smoking (smoker or non-smoker), are measured. Because the subjects are sampled randomly, one cannot know prior to data collection how many males and females or how many smokers and nonsmokers will be found; therefore, no marginal frequencies are fixed. With Case II, the experimental model, a certain number of subjects from each of two groups are tested, (e.g., 20 males and 20 females), and a dichotomous variable is measured (e.g., smoking or non-smoking). For Case II, the independent variable (sex in this example) is fixed, and the dependent variable (smoking), which is free to take on a number of outcomes, is not fixed. With Case III, the randomization model, both marginal frequencies are fixed; therefore, prior to data collection, the user knows the marginal frequencies. Examples of such data are considerably rarer than Case I and Case II data, but there are some examples. A psychology department, for instance, decides to move to a new building. There are 20 faculty members: 13 full professors and 7 of lower rank. Of the 20 available offices, 12 have phones, but 8 do not. We wish to know whether or not office assignment is independent of rank. Prior to the specific allocation of offices, we know precisely what the marginal frequencies will be.

For Case I and Case II data, the correct distributions are the multinomial and a product of two binomials respectively; however, these distributions cannot be used directly because that would require the specification of certain population parameters that are not known. Instead, Pearson's chi-square test is used to approximate the probabilities in both cases, and when the sample sizes are large, the approximation is fairly accurate. For Case III data, the correct underlying distribution is the hypergeometric, and the application of this distribution to 2×2 tables was described independently by Fisher (1934) and Irwin (1935); the test is commonly called the Fisher Exact Test (FET). Yates (1934) offered a correction for continuity

for the Pearson chi-square that made the latter test accurately approximate the FET.

The problem not understood by these early researchers was that the FET was not appropriate for Case I and Case II data, but only provides a quite conservative test for data of those types. This controversy is reviewed by Camilli and Hopkins (1978) who conclude that neither the FET nor the Yates correction of continuity should be used with Case I or II data. However, for Case III data, the FET does provide an exact test and should be used. With hand calculation, the test may be complex and time-consuming, but a computer can handle the problem with ease.

Marascuilo and McSweeney (1977) pointed out that hypergeometric solutions can be extended to more complex contingency tables. The primary reason that these solutions have not been used for more complex tables is that the computational complexity increases dramatically. The program described here solves the hypergeometric probability of $2 \times k$ tables, where k can represent 2, 3, 4, or more columns, but there are only two rows. The program generates all possible tables that could occur with the fixed marginal frequencies, calculates the hypergeometric probability of each table, and sums the probabilities of all those tables whose probability is less than or equal to that of the original table entered. When the value of k is 2, of course, the program computes the FET.

As an example of a 2×3 table, there are 25 applicants for job assignment: 16 are male and 9 are female. There are four supervisors, 8 white collar, and 13 blue collar positions. Therefore, we know in advance of job assignment what all the marginal frequencies will be. (See Table 1). The task is to determine whether significant statistical evidence of sex discrimination exists. The hypergeometric probability of this outcome or a more extreme outcome occurring by chance alone is 0.00454; therefore, the conclusion is that the pattern of job assignment is not independent of the applicant's sex. This example can be used to check whether the program is working accurately if adapted to another computer. Other examples of the FET can be taken from a variety of non-parametric statistics text books; however, the reader is warned that the present program computes two-tailed probabilities, and if the example is for a one-tailed test, the solutions will not agree.

The Program. The user is asked interactively to enter the number of columns (only two rows are assumed).

Table 1
Position

Sex	Supervisor	White Collar	Blue Collar	
Male	1	3	12	16
Female	3	5	1	9
Total	4	8	13	25

The author's mailing address is: Department of Psychology, Tulane University, New Orleans, LA 70118.

Next, the data for Row 1 are entered, separated by commas; then similarly the data for Row 2. The program then asks if the user wishes to see the tables as they are generated. Because there are often quite a few tables, the user should reflect before responding "yes;" however, this option is useful for instructional purposes. Finally, the program responds with the exact probability of an outcome that is as unlikely or more unlikely than the given table to occur by chance alone, given fixed marginal frequencies. With large sample sizes and many columns, there are many tables to generate, and, although the algorithm for table generation is as efficient as I could write it, each table takes some computer time. Because the total run time is directly proportional to the number of tables that must be generated, the program uses a crude algorithm to approximate the number of tables required. If the estimated number of tables is less than 5,000, the program automatically proceeds; however, if the number exceeds 5,000, the program informs the user of the estimate, then asks the user whether to proceed. On the DEC-20 computer, about 2,000 tables can be generated per cpu sec; therefore, a 2×4 problem with a total sample of 160 which required the generation of 45,961 tables took about 27 cpu sec. Because asymptotically the hypergeometric approach and the chi-square approximation are the same, the user may wish to switch to the latter when significant computer time would be required by the exact test.

Requirements. The program is written in single-precision FORTRAN IV and runs on a DEC-20 computer. The program does not require much data storage, so it does not require extensive memory. Because the program is interactive, the input and output are through the computer terminal. Several nonstandard FORTRAN statements are used to make the program run efficiently on the DEC system. Users with other systems may have to modify the assignment of logical units, READ statements with an "*" in place of the usual format statement number (this causes free format on the DEC), and format statements that end with the symbol "\$," which causes the cursor to remain at the end of the line on the DEC.

Program Availability. A listing of the program may be obtained at no charge from the author.

REFERENCES

- CAMILLI, G., & HOPKINS, K. D. (1978). Applicability of chi-square to 2×2 contingency tables with small expected cell frequencies. *Psychological Bulletin*, 85, 163-167.
- FISHER, R. A. (1934). *Statistical methods for research workers* (5th ed.). Edinburgh: Oliver and Boyd.
- IRWIN, J. O. (1935). Tests of significance for differences between percentages based on small numbers. *Metron*, 12, 83-94.
- MARASCUILLO, L. A., & MCSWEENEY, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Monterey: Brooks/Cole.
- YATES, F. (1934). Contingency tables involving small numbers and the chi-square test. *Journal of the Royal Statistical Society (Series B)*, 1, 217-235.

APPENDIX

```

      DIMENSION T(2,10),C(10),R(2),TL(17)
C --- HYPERGEOMETRIC TEST FOR 2 X NC TABLES.  WHEN NC EQUALS TWO
C --- THIS IS THE FISHER EXACT TEST
      COMMON TL
      OPEN (UNIT=5,DEVICE='TTY')
      OPEN (UNIT=6,DEVICE='TTY')
C --- CREATE TABLE OF LOG FACTORIALS
      FL=0.0
      DO 1 I=2,17
        A=I
        FL=FL+A*LOG(A)
      1 TL(I)=FL
C --- ENTER PROBLEM PARAMETERS
      2 WRITE(6,3)
      3 FORMAT(/' ENTER NO. OF COLUMNS OR 0 TO STOP (2 ROWS ASSUMED) ',%)
      READ(5,*)NC
      IF(NC.LE.1)STOP
      IF(NC.EQ.2)WRITE(6,4)
      4 FORMAT(/' WITH TWO COLUMNS THIS IS FISHERS EXACT TEST')
      PROB=0.
      DO 5 I=1,NC
        C(I)=0.
        R(1)=0.
        R(2)=0.
        NTAB=0
        DO 7 J=1,2
          WRITE(6,6)I
        6 FORMAT(/' ENTER DATA FOR ROW ',I,', SEPARATED BY COMMAS ',%)
          READ(5,*)(T(I,J),J=1,NC)
          DO 7 J=1,NC
            C(J)=C(J)+T(I,J)
          7 R(I)=R(I)+T(I,J)
          G=R(1)+R(2)
          FNC=NC
          ANT=(G/FNC)**(NC-1)
          IF(ANT.LT.5000.)GO TO 10
          WRITE(6,8)ANT
        8 FORMAT(/' THE APPROXIMATE NUMBER OF TABLES REQUIRED IS ',F10.0/
          1 ' DO YOU WISH TO PROCEED WITH THE COMPUTATION (Y OR N)? ',%)
          READ(5,9)ANS
          9 FORMAT(A1)
          IF(ANS.EQ.'N'.OR.ANS.EQ.'n')GOTO 2
        10 CON=FACL(R(1))+FACL(R(2))-FACL(G)
          DO 11 J=1,NC
            CON=CON+FACL(C(J))
          PGT=PRB(T,NC,CON)
          WRITE(6,12)
          12 FORMAT(/' DO YOU WANT TO SEE THE TABLES AS THEY ARE GENERATED
            1 (Y OR N)? ',%)
          READ(5,9)ANS
C --- GENERATE ALL POSSIBLE TABLES
          RL=R(1)
          DO 13 J=1,NC
            K=NC-J+1
            T(1,K)=RL
            IF(C(K).LT.RL)T(1,K)=C(K)
            RL=RL-T(1,K)
          13 T(2,K)=C(K)-T(1,K)
          GOTO 19
          14 KK=2
          IF(T(1,2).EQ.0.0)GOTO 15
          IF(T(1,1)+1.0.GT.C(1))GOTO 15
          T(1,2)=T(1,2)+1.
          T(1,1)=T(1,1)-1.
          T(2,2)=C(2)-T(1,2)
          T(2,1)=C(1)-T(1,1)
          GOTO 19
          15 KK=KK+1
          IF(KK.GT.NC)GOTO 24
          IF(T(1,KK).EQ.0.0)GOTO 15
          T(1,KK)=T(1,KK)+1.0
          RL=R(1)
          DO 16 I=KK,NC
            RL=RL-T(1,I)
          16 KKMO=KK-1
          DO 17 I=1,KKMO
            K=KKMO-I+1
            T(1,K)=RL
            IF(C(K).LT.RL)T(1,K)=C(K)
          17 RL=RL-T(1,K)
          IF(RL.GT.0)GOTO 15
          DO 18 I=1,KK
            T(2,I)=C(I)-T(1,I)
          18 P=PRB(T,NC,CON)
          19 P=PRB(T,NC,CON)
          IF(ANS.NE.'Y'.AND.ANS.NE.'y')GOTO 23
          DO 20 I=1,2
            20 WRITE(6,21)(T(I,J),J=1,NC)
          21 FORMAT(1X,10F7.0)
          WRITE(6,22)P
          22 FORMAT(' PROBABILITY OF THIS TABLE IS ',F12.7)
          23 CONTINUE
          NTAB=NTAB+1
          IF(ABS((P-PGT)/PGT).LE.1E-6.OR.P.LE.PGT)PROB=PROB+P
          GOTO 14
          24 WRITE(6,25)PROB,NTAB
          25 FORMAT(/' CUMULATIVE PROBABILITY OF ALL TABLES AS LIKELY OR/'
            1 ' LESS LIKELY THAN THE GIVEN TABLE IS ',F12.7/' TOTAL
            2 NUMBER OF TABLES GENERATED WAS ',I9)
          GOTO 2
        END

```

```

C      FUNCTION PRB(T,NC,CON)
C — PRB RETURNS THE HYPERGEOMETRIC PROBABILITY OF TABLE T, WITH
C — TWO ROWS AND NC COLUMNS
C — CON IS A CONSTANT = LOG(PROD(ROW TOT!)/PROD(COL TOT!)/GRAND TOT!
C — SML = SMALLEST ARGUMENT ACCEPTABLE TO FUNCTION EXP
      DIMENSION T(2,10)
      SML=-89.4
      P=CON
      DO 1 I=1,2
      DO 1 J=1,NC
1  P=P-FACL(T(I,J))
      PRB=0.
      IF(P.GT.SML)PRB=EXP(P)
      RETURN
      END
C
C      FUNCTION FACL(A)
C — RETURNS WITH THE NATURAL LOG OF THE FACTORIAL OF A

```

```

C — TABLE LOOK UP BELOW 18 - STIRLINGS APPROXIMATION ABOVE
      DIMENSION TL(17)
      COMMON TL
      FACL=0.
      IF(A.LE.1.0)RETURN
      IF(A.GT.17.0)GOTO 1
      I=A
      FACL=TL(I)
      RETURN
1  FACL=.918938533+A*LOG(A)*(A+.5)-A+1.0/(12.0*A)
      RETURN
      END

```

(Revision accepted for publication July 11, 1985.)