Journal of Modern Applied Statistical Methods

Volume 3 | Issue 1 Article 27

5-1-2004

JMASM11: Comparing Two Small Binomial Proportions

James F. Reed III
St. Luke's Hospital and Health Network, ReedJ@slhn.org

Journal of Modern Applied Statistical Methods
May, 2004, Vol. 3, No. 1, 250-258

1538 – 9472/04/$95.00

JMASM11: Comparing Two Small Binomial Proportions

James F. Reed III
Research Institute
St. Luke's Hospital and Health Network

A large volume of research has focused on comparing the difference between two small binomial proportions. Statisticians recognize that Fisher's Exact test and Yates chi-square test are excessively conservative. Likewise, many statisticians feel that Pearson's Chi-square or the likelihood statistic may be inappropriate for small samples. Viable alternatives exist.

Key words: Fisher's Exact Test, Fisher's Mid-p, Scaled Chi-square test

Introduction

A large volume of research spanning nearly half a century has focused on comparing the difference between two small binomial proportions. The validity of various testing procedures remains clouded by controversy (Hirji, 1991). Fisher's exact test (FET) (Fisher, 1958) and Yates continuity corrected chi-square test (Yates, 1934) (X^2_y) have numerous criticisms based on theoretical and empirical considerations. Critiques of FET and X^2_y conclude that they are excessively conservative when used with small to moderate sample sizes leading to an implied loss of power which diminishes their utility (Berkson, 1978; Dupont, 1986; D'Agostino, 1988; Haviland, 1990).

D'Agostino showed that even with small sample sizes Pearson's chi-square test (X^2) and the Student t-test based on binary data generally provide observed significance levels not far from the postulated levels (D'Agostino, 1988). FET is also extremely sensitive to minor variations in data, even when the minimum expected cell size is fairly large (Dupont, 1986).

James F. Reed III, Ph.D., is the Director in the Research Institute at St. Luke's Hospital & Health Network in Bethlehem, PA., 18015. Email: ReedJ@slhn.org.

Upton (Upton, 1982) and Overall (Overall, 1987) evaluated a wide variety of test procedures for comparing the difference between two small binomial proportions. They concluded that in both ease of computation and the average or median actual significance level, one should use one of three tests - the X^2, Student's t-test for binomial data (BST), or the scaled chi-square test (X^2_s) for almost all sample sizes. Others have advocated the use of Fisher's Mid-p (MP) based procedure in connection with FET (Miettinen, 1974; Plackett, 1984). Barnard (Barnard, 1989; Barnard, 1990) recommends reporting both the traditional p-value and the MP p-value when using FET.

Among applied statisticians a casual attitude towards using these tests has emerged. Practice appears to be guided by what has been described as "conventional wisdom" (D'Agostino, 1988). When the two sample sizes are large, applied statisticians generally use the X^2 or the likelihood statistic (G^2) and compare with the χ^2_{k-1} distribution. With small to moderate sample sizes, either FET or X^2_y are favored. This strategy is reinforced and seldom questioned as evidenced by the content of statistics texts used in colleges and universities.

Many statisticians recognize that FET and the X^2_y tests for comparing two independent binomial proportions are excessively conservative but continue their use. Likewise, many feel that X^2 and G^2 are inappropriate for small sample sizes. Viable alternatives to X^2

include X^2_s, Fisher's MP (14), or BST. None of these have made an appearance in statistics texts. The objective was to provide the practicing statistician with an executable code that produces these alternatives.

Methodology

The following notation was used to describe the comparison of two independent binomial proportions. Let A and B represent the number of successes in independent samples from two binomial populations (n_1, π_1) and (n_2 π_2). Let $n = n_1 + n_2$. Then the joint probability of a particular outcome is:

$$\Pr(A = a, B = b) = [n_1!/a!\,(n_1 - a)!]\,[n_2!/b!\,(n_2 - a)!]\pi_1^{a}\,(1 - \pi_1)^{c}\,\pi 2^{b}\,(1 - \pi 2)^{d},$$

$$\text{for } a = 0, 1, \ldots, n_1 \text{ and } b = 0, 1, \ldots, n_2,$$
$$c = n_1 - a \text{ and } d = n_2 - b.$$

Pearson and Scaled chi-square tests (X^2 and X^2_S)

For an observed pattern of (a, b), Pearson's chi-square statistic is

$$X^2 = (ad - bc)^2 n/\{n_1 n_2 (a+b)(c+d)\}$$

The scaled chi-square statistic is derived from the mean and variance of the conditional hypergeometric distribution and is defined as

$$X^2_S = X^2(n - 1)/n$$

X^2 and X^2_S are compared with a χ^2 statistic with 1 degree of freedom. X^2 and X^2_S tests are approximate tests because the distributions of X^2 and X^2_S approach χ^2 with 1 degree of freedom only when the sample sizes are large.

Student t-test (BST)

BST uses the means and variances of the two binomial distributions to compute the usual two independent sample t-statistic on a pooled estimate of the variances. BST is then compared with to the Student t distribution with $n_1 + n_2 - 2$ degrees of freedom.

Fisher's Exact test (FET)

FET is constructed using the conditional distribution of A given A+B. FET is defined:

$$f(a, s, \phi) = \Pr(A = a \mid A+B = s; \phi) \text{ and } S(a, s, \phi) = \Pr(A > a \mid A + B = s; \phi), \text{ where } \phi = \pi_1(1 - \pi_2)/(\pi_1 (1 - \pi_1))$$

For a one-sided α-level test, reject Ho if $f(a, s, \phi) + S(a, s, 1) \le \alpha$.

Mid-P (MP)

MP tests the mean of two probabilities obtained by inclusion and exclusion of the observed point in a discrete distribution. This is equivalent to inclusion of half the probability of the observed point in each tail. MP is found by modifying the above one-sided procedure by the following:

$$\text{For a one-sided } \alpha\text{-level test, reject if } 0.5\, f(a, s, 1) + S(a, s, 1) \le \alpha.$$

Conclusion

A summary of the literature based on intuitive and theoretical grounds argues in favor of the use of a MP test (Barnard, 1989; Lancaster 1961). The computational effort required for the MP test is no more than that needed for FET. Further, the basis for MP is a natural adjustment for discreteness; and the test easily generates to r x c contingency tables and other discrete data problems (Hirji, 1991). It is strongly agreed upon that the Yates-corrected chi-square statistic in analyses of 2 x 2 contingency tables are overly conservative and that the Pearson chi-square generally provides adequate control over type I error probabilities (Haviland, 1990).

The two-tailed FET p-value is highly sensitive to small variations in 2 x 2 contingency tables. This sensitivity raises doubts about the utility of the FET as a measure of the relative strength of evidence provided by different tables (Dupont, 1986). Pearson's chi-square statistic generally provides adequate control over type I error probabilities without the severe conservative bias produced by Yates' correction for continuity.

When the analytic problem of comparing two independent binomial proportions the classical FET and the X^2_y chi-square tests are too conservative for practical use. A recommend analytical algorithm is: (1) When the two samples are nearly equal, and when the underlying true binomial value is near 0.5, use one of three statistics: {X^2, X^2_S, BST, MP} for all sample sizes, and (2) in case of unequal sample sizes, or when the common binomial parameter is near 0 or 1, use MP statistic.

An executable Fortran program that produces the statistics outlined in the previous section and sample data is provided in the appendix. A literature search did not produce any references related to public domain software that produces these statistics. The program as written may not be optimal. Any suggestion for refinements to the program would be gratefully accepted.

References

Barnard, G. A. (1989). On alleged gains in power from lower p-values. *Statistics in Medicin*e, *8*, 1469-1477.

Barnard, G. A. (1990). Comment. *Statistics in Medicine*, *9*, 373-375.

Lancaster, H. O. (1961). Significance tests in discrete distributions. *Journal of the American Statistical Association*, *56*, 223-234.

Berkson, J. (1978). In dispraise of the exact test. *Journal of Statistical Inference and Planning*, 2, 27-42.

D'Agostino, R. B., Chase, W., & Belanger, A. (1988). The appropriateness of some common procedures for testing the equality of two independent binomial proportions. *American Statistician*, *42*, 198-202.

Dupont, W. D. (1986). Sensitivity of Fisher's exact test to minor perturbations in 2 x 2 contingency tables. *Statistics in Medicine*, *5*, 629-635.

Fisher, R. A. (1958). *Statistical Methods for Research Workers*. (13th ed.). New York: Hafner.

Haviland, M G. (1990). Yate's correction for continuity and the analysis of 2 x 2 contingency tables. *Statistics in Medicine*, 9, 363-367.

Hirji, K. F., Tan, S. J., & Elashoff, R. M. (1991). A quasi-exact test for comparing two binomial proportions. *Statistics in Medicine*, *10*, 1137-1153.

Miettinen, O. S. (1974). Discussion of Conover's "Some reasons for not using the Yates continuity correction on 2 x 2 contingency tables". *Journal of the American Statistical Association*, *69*, 374-382.

Overall, J. E., Rhoades, H. M., & Starbuck, R. R. (1987). Small-sample tests for homogeneity of response probabilities in 2 x 2 contingency tables. *Psychological Bulletin*, *102*, 307-314.

Plackett, R. L. (1984). Discussion of Yate's "Tests of significance for 2 x 2 contingency tables". *Journal of the Royal Statistical Society*, S*eries* A, *147*, 426-463.

Upton, G. J. G. (1982). A comparison of alternative tests for the 2 x 2 comparative trial. *Journal of the Royal Statistical Society, Series A*, *145*, 86-105.

Yates F. (1934). Contingency tables involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society* (Supplement), *1*, 217-235.

Appendix

An executable Fortran program that produces the statistics outlined in this article follows:

```
      INTEGER O(2,2),ROWS(2),COLS(2)
      INTEGER ROW,COL
      COMMON A,B,C,D

C
      OPEN(8,FILE='PKIN')
      READ(8,*) ROW,COL
      READ(8,*) ((O(I,J),J=1,COL),I=1,ROW)
      CLOSE(8)
      A=O(1,1)
      B=O(1,2)
      C=O(2,1)
      D=O(2,2)
      ROW1=A+B
      ROW2=C+D
      COL1=A+C
      COL2=B+D
      TOT =A+B+C+D

C
      CLOSE(8)
      OPEN(8,FILE='PKOUT')

      CALL CHISQ
      CALL BST
      CALL FISH  (O)

      STOP
      END
C=====================================================================
SUBROUTINE CHISQ
      COMMON A,B,C,D

      ROW1=A+B
      ROW2=C+D
      COL1=A+C
      COL2=B+D
      TOT=A+B+C+D

      PI1=A/(A+C)
      PI2=B/(B+D)

      CHI=((A*D-B*C)**2*TOT)/(COL1*COL2*ROW1*ROW2)

      WRITE(8,101)
      WRITE(8,102) A, B
      WRITE(8,103) C,D
      WRITE(8,104) PI1, PI2

      IDF = 1
      CALL CHIP (CHI,IDF,XPROB)
```

```
      IF (CHI .EQ. 0) XPROB = 1.0D0
      WRITE(8,105) CHI,XPROB
C
      SCS = CHI*(TOT-1)/(TOT)
      CALL CHIP (SCS,IDF,YPROB)
      IF (SCS .EQ. 0) YPROB = 1.0D0
      WRITE(8,106) SCS,YPROB
C
  101 FORMAT(14X,'TRT A',15X,'TRT B',/)
  102 FORMAT(10X,'   a = ',F6.1,7X,'   b = ',F6.1)
  103 FORMAT(10X,'   c = ',F6.1,7x,'   d = ',F6.1,/)
  104 FORMAT(10X,'ã(1) = ',F6.5,7X,'ã(2) = ',F6.3,/)
  105 FORMAT(/,10X,'Xý   = ',F6.2,' , p-value = ',F7.4)
  106 FORMAT(/,10X,'Xý(S)= ',F6.2,' , p-value = ',F7.4)
C
      RETURN
      END
C=====================================================================
SUBROUTINE BST
      COMMON A,B,C,D
      INTEGER DF

      ROW1=A+B
      ROW2=C+D
      COL1=A+C
      COL2=B+D
      TOT =A+B+C+D

      P1 = A/COL1
      P2 = B/COL2
      TN = ABS(P1-P2)
      PI = (A+B)/TOT
      D1 = PI*(1-PI)/COL1
      D2 = PI*(1-PI)/COL2
      TD = SQRT(D1+D2)
      T  = TN/TD
      FT =T*T

      DF = TOT-2
      CALL FAPPROX (FT,1,DF,QX)
      CALL NPROB (QX,TPROB)

      WRITE(8,101) T,TPROB
  101 FORMAT(/,10X,'BST  = ',F6.2,' , p-value = ',F7.4)
      RETURN
      END

C=====================================================================
SUBROUTINE GTEST (O)
      INTEGER O(2,2),ROWS(2),COLS(2)

      COMMON A,B,C,D

      ROWS(1)=A+B
      ROWS(2)=C+D
      COLS(1)=A+C
      COLS(2)=B+D
```

```
      TOT=A+B+C+D
C
      DO 10 I = 1,2
      DO 10 J = 1,2
         IF (O(I,J).NE.0) G = G + O(I,J)*LOG(REAL(O(I,J)))
  10  CONTINUE
      DO 20 I = 1,2
         IF (ROWS(I).NE.0.0) G = G - ROWS(I)*LOG(REAL(ROWS(I)))
  20  CONTINUE
      DO 30 J = 1,2
         IF (COLS(J).NE.0.0) G = G - COLS(J)*LOG(REAL(COLS(J)))
  30  CONTINUE

      G = G + TOT*LOG(TOT)
C
      IF (G.LT.0.0) G = 0.0
      IF (G.GE.0.0) G = 2.0 * G
      IDF = 1
      CALL CHIP(G,IDF,PROB)
      IF (G.EQ.0.0D0) PROB = 1.0
C
      WRITE(8,100)G,PROB
  100 FORMAT(/,10X,'G    = ',F6.2,' , p-value = ',F7.4)
      RETURN
      END
C=====================================================================C
FISHER'S EXACT TEST
C
C     1)  One-tail computations
C     2)  Two-tail computations
C
C---------------------------------------------------------------------
      SUBROUTINE FISH (O)
      INTEGER O(2,2)
      INTEGER A,B,C,D,S
      REAL    P(9),T(20),MFPROB1,MFPROB2
C
      K = 0
      A=O(1,1)
      B=O(1,2)
      C=O(2,1)
      D=O(2,2)

    1 K = K+1
      XL = 0.0D0
      CALL EPROB (A,PR)
      P(2)=PR
      CALL EPROB (B,PR)
      P(3) = PR
      CALL EPROB (C,PR)
      P(4) = PR
      CALL EPROB (D,PR)
      P(5) = PR
      S = A+B
      CALL EPROB (S,PR)
      P(6) = PR
      S = C+D
```

```
      CALL EPROB (S,PR)
      P(7) = PR
      S = B+D
      CALL EPROB (S,PR)
      P(8) = PR
      S = A+C
      CALL EPROB (S,PR)
      P(9) = PR
      S = A+B+C+D
      CALL EPROB (S,PR)
      P(1) = PR
      DO 20 J = 6,9
   20    XL = XL + P(J)
      DO 30 J = 1,5
   30    XL = XL - P(J)
      T(K) = EXP(XL)
C
      IF((A.EQ.0).OR.(B.EQ.0).OR.(C.EQ.0).OR.(D.EQ.0))THEN
C
C     CONDITIONAL PROBABILITY
C
         FPROB1  = 0.0D0
         DO 40 I = 1,K
   40      FPROB1  = FPROB1  + T(I)
C
C        FISHERS EXACT TEST PROBABILITY (Two-Tail)
         PROB2  = 1 - FPROB1 + T(1)
         FPROB2 = 2*MIN(FPROB1,PROB2)
         IF (FPROB2 .GE. 1) FPROB2=1
C
C        FISHER MID-P
         T(1)    = 0.5D0*T(1)
         MFPROB1 = 0.0D0
         DO 50 I = 1,K
   50      MFPROB1 = MFPROB1 + T(I)
         MFPROB2 = 2*MFPROB1
         IF(MFPROB2 .GE. 1) MFPROB2=1
         IF (A.LT.C) MFPROB = 1.0D0 - MFPROB

         WRITE(8,100) FPROB1
C        WRITE(8,101) FPROB2
         WRITE(8,102) MFPROB1
C        WRITE(8,103) MFPROB2
  100    FORMAT(/,10X,'FET (One-tail), p-value = ',F7.4)
C 101    FORMAT(/,10X,'FET (Two-tail), p-value = ',F7.4)
  102    FORMAT(/,10X,'MP  (ONE-tail), p-value = ',F7.4)
C 103    FORMAT(/,10X,'MP  (TWO-tail), p-value = ',F7.4)

         RETURN
      ENDIF
C
      IF (A*D - B*C .LT. 0) THEN
         A = A-1
         D = D-1
         B = B+1
         C = C+1
      ELSE
```

```
         A = A+1
         D = D+1
         B = B-1
         C = C-1
      ENDIF
C
      GO TO 1

      RETURN
      END
C=====================================================================
SUBROUTINE EPROB (S,PR)
      INTEGER S
      REAL PR
      PR = 0
      DO 10 I = 1,S
  10     PR = PR + LOG(REAL(I))
      RETURN
      END
C=====================================================================
SUBROUTINE FAPPROX (F,N1,N2,QX)
      REAL F,V1,V2,XNUM,XDEN,QX
      V1 = REAL(N1)
      V2 = REAL(N2)
      XNUM = F**(1.0/3.0)*(1.0-2.0/(9.0*V2))-(1.0-2.0/(9.0*V1))
      XDEN = 2.0/(9.0*V1)+F**(2.0/3.0)*(2.0/(9.0*V2))
      QX = XNUM/SQRT(XDEN)
      RETURN
      END
C=====================================================================
SUBROUTINE NPROB (X,PROB)
      REAL D,PROB,X
      DATA D1,D2,D3/0.0498673470,0.0211410061,0.0032776263/
      DATA D4,D5,D6/0.0000380036,0.0000488906,0.0000053830/
      PROB = 1.0/(2.0*(1.0+D1*X+D2*X*X+D3*X*X*X
     &                  +D4*X*X*X*X+D5*X*X*X*X*X+D6*X*X*X*X*X*X)**16)
      IF (PROB .GE. 1.0) PROB = 1.0
      RETURN
      END
C=====================================================================C
ABRAMOWITZ & STEGUN
C                              ******************
C                                   Q(X2|DF)
C                               PG 941, EQ 26.4.14
C                               PG 941, EQ 26.4.15
C                               PG 932, EQ 26.2.19
C---------------------------------------------------------------------
      SUBROUTINE CHIP (STAT,IDF,P)
C
      DATA D1/0.0498673470/,D2/0.0211410061/,D3/0.0032776263/
     $     D4/0.0000380036/,D5/0.0000488906/,D6/0.0000053830/
C
C   CUBE ROOT APPROXIMATION
C
      DF = REAL(IDF)
      X  = ((STAT/DF)**(1.0D0/3.0D0)-(1.0D0-(2.0D0/(9.0D0*DF))))/
     $     (SQRT(2.0D0/(9.0D0*DF)))
```

```
C
C   IMPROVED CUBE ROOT APPROXIMATION
C
      IF ((X.GE.-3.5).AND.(X.LE.-3.0))
     $   X=X+0.6060606*(-0.0067+0.0102*(-3.0-X))
      IF ((X.GT.-3.0).AND.(X.LE.-2.5))
     $   X=X+0.6060606*(-0.0033+0.0068*(-2.5-X))
      IF ((X.GT.-2.5).AND.(X.LE.-2.0))
     $   X=X+0.6060606*(-0.0010+0.046*(-2.0-X))
      IF ((X.GT.-2.0).AND.(X.LE.-1.5))
     $   X=X+0.6060606*(0.0001+0.0022*(-1.5-X))
      IF ((X.GT.-1.5).AND.(X.LE.-1.0))
     $   X=X+0.6060606*(0.0006+0.0005*(-1.0-X))
      IF ((X.GT.-1.0).AND.(X.LE.-0.5))
     $   X=X+0.6060606*(-0.0006)
      IF ((X.GT.-0.5).AND.(X.LE.0.0))
     $   X=X+0.6060606*(-0.0002+0.0008*X)
      IF ((X.GT.0.0).AND.(X.LE.0.5))
     $   X=X+0.6060606*(-0.0003-0.001*(0.5-X))
      IF ((X.GT.0.5).AND.(X.LE.1.0))
     $   X=X+0.6060606*(-0.0006-0.006*(1.0-X))
      IF ((X.GT.1.0).AND.(X.LE.1.5))
     $   X=X+0.6060606*(-0.0005+0.0002*(1.5-X))
      IF ((X.GT.1.5).AND.(X.LE.2.0))
     $   X=X+0.6060606*(0.0002+0.0014*(2.0-X))
      IF ((X.GT.2.0).AND.(X.LE.2.5))
     $   X=X+0.6060606*(0.0017+0.003*(2.5-X))
      IF ((X.GT.2.5).AND.(X.LE.3.0))
     $   X=X+0.6060606*(0.0043+0.0052*(3.0-X))
      IF ((X.GT.3.0).AND.(X.LE.3.5))
     $   X=X+0.6060606*(0.0082+0.0078*(3.5-X))
C
      X2 = X*X
      X3 = X2*X
      X4 = X3*X
      X5 = X4*X
      X6 = X5*X
      P  = 0.5*(1.0+D1*X+D2*X2+D3*X3+D4*X4+D5*X5+D6*X6)**(-16.0)
C
C   ERROR CHECKS
C
      IF (P.GT.1.0) P=0.999
      IF (P.LT.0.0) P=0.001
      RETURN
      END
```