8-1971

An Extension of the Cochran Q test

Soo-Young Cho

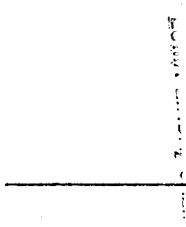
LOMA LINDA UNIVERSITY

Graduate School

AN EXTENSION OF THE COCHRAN Q TEST

by

Soo-Young Cho



A Thesis in Partial Fulfillment

of the Requirements for the Degree

Master of Science in the Field of Biostatistics

August 1971

Each person whose signature appears below certifies that he has read this thesis and that in his opinion it is adequate, in scope and quality, as a thesis for the degree of Master of Science.

Paul Y. Yahiku, Chairman
Paul Y. Yahiku, Assistant Professor of Biostatistics

Jan W. Kuzma
Jan W. Kuzma, Associate Professor of Biostatistics

Grenith J. Zimmerman
Grenith J. Zimmerman, Assistant Professor of Biostatistics

C. Duane Zimmerman
C. Duane Zimmerman, Assistant Professor of Biomathematics

ACKNOWLEDGMENT

The author wishes to express his appreciation to the members of his committee: Drs. Paul Yahiku, Jan Kuzma, Grenith Zimmerman, Duane Zimmerman, and all who helped make this study possible.

Special thanks are due to Dr. Jan Kuzma, Chairman of the Department of Biostatistics, School of Health, for providing me the opportunity to work in the Department, thus receiving needed financial support.

As a foreign student, I also extend my sincere gratitude to Dr. Paul Yahiku, the chairman of my Thesis Advisory Committee, for his continual service, inspiration, help in correcting my English, and guidance in making this thesis possible.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

To compare percentage distributions of categorical data for two or more independent samples the χ^2 test is ordinarily used. A more accurate comparison of the percentages is sometimes obtained if the samples consist of matched individuals. The basis for the matching may be relevant characteristics of the different subjects which may be variables such as age, sex, weight, ethnic group, and the like, or the fact that in each of the samples the same subjects are being used.

When the data consists of dichotomized ordinal information, methods of analysis are available. The McNemar test is generally used when comparing only two related samples. A method for testing whether three or more matched sets of frequencies or proportions differ significantly among themselves is provided by the Cochran Q test.

A wide variety of situations may be suggested for which the data might be analysed by the Cochran Q test. Siegel (1956) uses the following example. On an examination consisting of c questions of a pass-fail nature and administered to N individuals one could test whether these items differ in difficulty. If each person answers all c questions there will be c groups of observations corresponding to these questions. These groups are considered "matched" since they involve the same individuals.

We may wish to compare the responses of N subjects to a given stimulus under c different conditions. Here again we borrow an example from Siegel (1956). We shall interview a group of voters at c speci-

fied times during an election campaign, asking each member to indicate which of two candidates he favors. We can interview them prior to the campaign, at the peaks of the campaigns of each of the two candidates, immediately before the election, and immediately after the results are known. We could then determine whether the responses of the subjects to the stimulus are significantly different under the c conditions.

As indicated above, the Cochran Q test applies only to the case of dichotomized responses. Clearly there is a need for a method of analyzing the more general case where the responses are multinomial. In the present paper an extension of the Cochran Q test is developed for handling such cases.

CHAPTER II

MATHEMATICAL FORMULATION

Throughout this paper the number of groups being compared will be denoted by c, and the number of sets will be denoted by N. Each set will consist of c matched individuals. The response in the ith set and the jth group will be denoted by X_{ij}. These responses can be displayed as in Table 1. It is convenient to introduce some additional notations:

L_{xi} = Total number of times x occurs in the ith set, where x = 0, 1, 2, ..., n.

$T_i = \sum_{j=1}^{c} X_{ij}$, total of ith set.

$S_i = \sum_{j=1}^{c} X_{ij}^2$, sum of squares of scores in the ith set.

$G_j = \sum_{i=1}^{N} X_{ij}$, total of jth group.

These are also displayed in Table 1.

Note that the observations in any given set are independent of observations in any other set since they involve different individuals. The null hypothesis states that the c groups are identical with respect to the characteristic being investigated. It follows that under the null hypothesis the response in the jth group, X_{ij}, and the response in the j'th group, $X_{ij'}$, are identically distributed for all j and j', and that the conditional joint probability function of $X_{i1}, X_{i2}, \ldots, X_{ic}$ given $L_{0i}, L_{1i}, \ldots, L_{ni}$ is

$$f(x_{i1}, x_{i2}, \ldots, x_{ic}) = \frac{1}{\binom{c}{L_{0i}, L_{1i}, \ldots, L_{ni}}}$$

provided the values of $x_{i1}, x_{i2}, \ldots, x_{ic}$ are consistent with $L_{0i}, L_{1i}, \ldots, L_{ni}$. It is zero otherwise.

Cochran's test criterion for the special case (n=1) is $\sum_j (G_j - \overline{G})^2$. He has shown for this case that the statistic

$$Q = \frac{(c-1) \sum_j (G_j - \overline{G})^2}{\sum_i T_i \left(1 - \frac{T_i}{c}\right)}$$

approaches asymptotically a χ^2 distribution with $(c - 1)$ degrees of freedom as N gets large.

In the general case being considered here we again use $\sum_j (G_j - \overline{G})^2$ as the test criterion.

Table 1

A Display of Responses, X_{ij}, Group Totals, G_j, Frequencies, L_{xi}, Set Totals, T_i, and Set Sum of Squared Responses, S_i

Gp. j / i Set	1	2	3	. . .	c	L_{0i}	L_{1i}	L_{2i}	. . .	L_{ni}	T_i	S_i
1	X_{11}	X_{12}	X_{13}	. . .	X_{1c}	L_{01}	L_{11}	L_{21}	. . .	L_{n1}	T_1	S_1
2	X_{21}	X_{22}	X_{23}	. . .	X_{2c}	L_{02}	L_{12}	L_{22}	. . .	L_{n2}	T_2	S_2
3	X_{31}	X_{32}	X_{33}	. . .	X_{3c}	L_{03}	L_{13}	L_{23}	. . .	L_{n3}	T_3	S_3
.	.	.	.	. . .	.	.	.	.	. . .	.	.	.
.	.	.	.	. . .	.	.	.	.	. . .	.	.	.
.	.	.	.	. . .	.	.	.	.	. . .	.	.	.
N	X_{N1}	X_{N2}	X_{N3}	. . .	X_{Nc}	L_{0N}	L_{1N}	L_{2N}	. . .	L_{nN}	T_N	S_N
Tot.	G_1	G_2	G_3	. . .	G_c							

CHAPTER III

DISTRIBUTION OF Q UNDER THE NULL HYPOTHESIS

It should be mentioned that the following derivations very closely parallel those of Cochran's. Whereas Cochran's test is conditional on the T_i's ($i = 1, 2, \ldots, N$) the present test is conditional on L_{0i}, $L_{1i}, \ldots, L_{ni}$ ($i = 1, 2, \ldots, N$). It should be understood that all the following results are conditional on the values of these L's.

Ultimately we shall obtain the means, variances, and the covariances of the G_j's and based on these obtain the limiting distribution of the generalized Q statistic

$$Q = \frac{\sum_j (G_j - \overline{G})^2}{\frac{1}{c}\sum_i S_i - \sum_i W_i} . \tag{1}$$

For the ith row we have

$$E(X_{ij}) = \frac{\sum_{x=0}^{n} L_{xi} \cdot x}{c} = \frac{\sum_{j=1}^{c} X_{ij}}{c} = \frac{T_i}{c} \tag{2}$$

and

$$E(X_{ij}^2) = \frac{\sum_{x=0}^{n} L_{xi} \cdot x^2}{c} = \frac{\sum_{j=1}^{c} X_{ij}^2}{c} = \frac{S_i}{c} . \tag{3}$$

From (2) and (3) we obtain

$$Var(X_{ij}) = E(X_{ij}^2) - E^2(X_{ij}) = \frac{cS_i - T_i^2}{c^2} . \tag{4}$$

The expectation of $X_{ij} \cdot X_{ij'}$ is

$$E(X_{ij} \cdot X_{ij'}) = \sum_{x=0}^{n} \sum_{x'=0}^{n} x \cdot x' \cdot P(X_{ij}=x,\ X_{ij'}=x')$$

$$= \sum_{x=0}^{n} \sum_{x'=0}^{n} x \cdot x' \cdot P(X_{ij}=x) \cdot P(X_{ij'}=x' \mid X_{ij}=x) \tag{5}$$

where

$$P(X_{ij}=x) = \frac{L_{xi}}{c} \tag{6}$$

and

$$P(X_{ij'}=x' \mid X_{ij}=x) = \begin{cases} \dfrac{L_{x'i}}{c-1} & \text{if } x \neq x' \\ \dfrac{L_{xi}-1}{c-1} & \text{if } x = x' \end{cases} . \tag{7}$$

For convenience we shall let $W_i = E(X_{ij} \cdot X_{ij'})$.

The covariance of X_{ij} and $X_{ij'}$ can then be written as

$$\begin{aligned} \mathrm{Cov}(X_{ij}, X_{ij'}) &= E(X_{ij} \cdot X_{ij'}) - E(X_{ij}) \cdot E(X_{ij'}) \\ &= W_i - \frac{T_i^2}{c^2} . \end{aligned} \tag{8}$$

Since X's in any given row are independent of X's in any other row, the means, variances, and the covariances of the column totals, G_j, can be obtained by summing the corresponding expressions in (2), (4), and (8) over i. Thus

$$E(G_j) = \sum_{i=1}^{N} \frac{T_i}{c}, \qquad j = 1, 2, \ldots, c, \tag{9}$$

$$\mathrm{Var}(G_j) = \sigma^2 = \sum_{i=1}^{N} \left(\frac{cS_i - T_i^2}{c^2}\right), \quad j = 1, 2, \ldots, c, \tag{10}$$

and

$$\mathrm{Cov}(G_j, G_{j'}) = \rho\sigma^2 = \sum_{i=1}^{N} \left(W_i - \frac{T_i^2}{c^2}\right), \quad j \neq j' . \tag{11}$$

As in Cochran's case we can expect the joint distribution of the variates $G_1, G_2, \ldots, G_c$ to approach a multivariate normal distribution with the above variances and covariances. As shown by Walsh (1947)

the quantity

$$Q = \frac{\sum_j (G_j - \bar{G})^2}{\sigma^2(1 - \rho)} \tag{12}$$

has as its asymptotic distribution a χ^2 distribution with $(c - 1)$ degrees of freedom. By subtracting (11) from (10) we obtain for the denominator in (12)

$$\sigma^2(1 - \rho) = \sum_i \left(\frac{cS_i - T_i^2}{c^2}\right) - \sum_i \left(W_i - \frac{T_i^2}{c^2}\right)$$

$$= \sum_i \frac{S_i}{c} - \sum_i W_i = \frac{1}{c} \sum_i S_i - \sum_i W_i \ . \tag{13}$$

Finally, substituting this expression into (12), we obtain the generalized formula for Q in (1).

CHAPTER IV

APPLICATION TO AN EXAMPLE

We shall consider an example in which the effectiveness of three drugs are to be compared. Patients are to be assigned to the three drugs in the following manner: twelve sets of patients, each consisting of three matched individuals, are to be formed. In each set the three members are then assigned at random to the three drugs. We would then have three matched groups (c=3) and twelve sets (N=12). At the end of a certain time each patient will be asked to indicate whether he feels worse, no difference, or better. Scores of 0, 1, and 2, respectively, will be assigned to these responses. The results corresponding to Table 1 are given in Table 2.

Table 2

Worse (0), No Difference (1), and Better (2) Responses

By Patients Under 3 Types of Drugs

Set	Drug 1	Drug 2	Drug 3	L_{0i}	L_{1i}	L_{2i}	T_i	S_i
1	1	2	2	0	1	2	5	9
2	0	1	1	1	2	0	2	2
3	1	2	2	0	1	2	5	9
4	2	1	1	0	2	1	4	6
5	0	1	1	1	2	0	2	2
6	0	1	1	1	2	0	2	2
7	1	2	2	0	1	2	5	9
8	0	0	2	2	0	1	2	4
9	0	0	1	2	1	0	1	1
10	2	1	2	0	1	2	5	9
11	1	0	2	1	1	1	3	5
12	0	1	1	1	2	0	2	2
Total	G_1= 8	G_2=12	G_3=18	-	-	-	$\sum_{i=1}^{12} T_i$=38	$\sum_{i=1}^{12} S_i$=60

In order to compute Q as given (1) we need the quantities: c, $\sum_j G_j^2$, $(\sum_j G_j)^2$, $\sum_i S_i$, and $\sum_i W_i$. Except for W_i, these quantities are readily obtained. We have

$$c = 3.$$

$$\sum_j G_j^2 = 8^2 + 12^2 + 18^2 = 64 + 144 + 324 = 532.$$

$$(\sum_j G_j)^2 = (8 + 12 + 18)^2 = 38^2 = 1444.$$

$$\sum_j S_i = 60.$$

The formula for W_i is given in (5). We shall display the computations for W_1:

$$\begin{aligned} W_1 = \sum_{x=0}^{n} \sum_{x'=0}^{n} x \cdot x' \cdot P_1(x,x') &= 0 \cdot 0 \cdot P_1(0,0) \\ &+ 0 \cdot 1 \cdot P_1(0,1) \\ &+ 0 \cdot 2 \cdot P_1(0,2) \\ &+ 1 \cdot 0 \cdot P_1(1,0) \\ &+ 1 \cdot 1 \cdot P_1(1,1) \\ &+ 1 \cdot 2 \cdot P_1(1,2) \\ &+ 2 \cdot 0 \cdot P_1(2,0) \\ &+ 2 \cdot 1 \cdot P_1(2,1) \\ &+ 2 \cdot 2 \cdot P_1(2,2) \end{aligned} \tag{14}$$

where

$$P_1(x,x') = P(X_{1j}=x,\ X_{1j'}=x').$$

Since some of the terms in (14) are zero, this expression simplifies to

$$W_1 = 1 \cdot P_1(1,1)$$

$$+ 2 \cdot P_1(1,2)$$

$$+ 2 \cdot P_1(2,1)$$

$$+ 4 \cdot P_1(2,2).$$

From (6) and (7) it follows that:

$$W_1 = 1 \left(\frac{1}{3}\right)\left(\frac{1-1}{3-1}\right)$$

$$+ 2 \left(\frac{1}{3}\right)\left(\frac{2}{3-1}\right)$$

$$+ 2 \left(\frac{2}{3}\right)\left(\frac{1}{3-1}\right)$$

$$+ 4 \left(\frac{2}{3}\right)\left(\frac{2-1}{3-1}\right)$$

$$= 1 \left(\frac{1}{3}\right)\left(\frac{0}{2}\right) + 2 \left(\frac{1}{3}\right)\left(\frac{2}{2}\right)$$

$$+ 2 \left(\frac{2}{3}\right)\left(\frac{1}{2}\right) + 4 \left(\frac{2}{3}\right)\left(\frac{1}{2}\right)$$

$$= 0 + \frac{2}{3} + \frac{2}{3} + \frac{4}{3} = \frac{8}{3} .$$

Likewise we can compute the remaining W_i's. As can be appreciated, the computations required to obtain these W_i's can be quite tedious. It turns out for this example that $\sum_i W_i = 14.33$. Substituting the above values into formula (1) we obtain $Q = 8.94$. The number of degrees of freedom for this problem is $c - 1 = 3 - 1 = 2$. From χ^2 table we observe that $\chi^2_{.95}$ for 2 degrees of freedom is 5.99. Thus we conclude that the three drugs differ significantly in effectiveness.

A FORTRAN program has been written to perform this extended Cochran Q test. This program is contained in the appendix together with an application to the above example.

BIBLIOGRAPHY

Cochran, W. G. (1950). "The comparison of percentages in matched samples." Biometrika, 37, 256-266.

Maxwell, A. E. (1961). "Cochran's Q-Test for Correlated Quantal Results." Analyzing Qualitative Data, pp. 131, 132. Methuen & Co., LTD.: London.

Siegel, S. (1956). "The Cochran Q Test." Nonparametric Statistics For The Behavioral Sciences, pp. 161-166. McGraw-Hill Book Co., Inc.: New York.

Walsh, J. E. (1947). "Concerning the effect of intraclass correlation on certain significance tests." Ann. Math. Statist., 18:88.

APPENDIX

A FORTRAN PROGRAM

FOR

THE GENERALIZED COCHRAN Q TEST

1. GENERAL DESCRIPTION

   a. This program performs the generalized Cochran Q test.

   b. Output from this program includes:

      (1) Maximum score, n

      (2) Number of samples (or groups), c

      (3) Sum of each column, G_j

      (4) Calculated value of Q

      (5) Degrees of freedom

   c. Limitations per program:

      (1) n, maximum score ($0 \leq n \leq 9$)

      (2) c, maximum number of samples (or groups)

      ($1 \leq c \leq 75$)

      (3) N, maximum number of cases (or sets)

      ($1 \leq N \leq 9999$)

2. ORDER OF CARDS IN DECK

   a. Package program deck

   b. Information card

   c. Data cards

3. CARD PREPARATION

   a. Information card (one information card for each problem)

      Column 1-12 Alphanumeric job code (optional)

13-16 n, maximum score

17-20 c, number of samples (or groups)

21-24 N, number of cases (or sets)

b. Data cards (one data card for each case)

Column 1-4 Identification (optional)

6-80 Scores (one column per score)

4. FORTRAN PROGRAM

```
    DIMENSION S(75), G(75), R(9), P(9,9), IG(75)
    READ 10, MS, LC, NC
 10 FORMAT(12X,3I4)
    SG  = 0.
    SGS = 0.
    SS  = 0.
    SW  = 0.
    M   = NC
    C   = LC
    DO 20 I=1,LC
 20 G(I) = 0.
 30 READ 40, S
 40 FORMAT(5X,75F1.0)
    DO 50 I=1,MS
 50 R(I) = 0.
    DO 80 I=1,LC
    IF( S(I) ) 80, 80, 60
 60 SS = SS + S(I)*S(I)
    G(I) = G(I) + S(I)
    JS = S(I)
    DO 80 J=1,MS
    IF(JS - J) 80, 70, 80
 70 R(J) = R(J) + 1.
 80 CONTINUE
    DO 130 J=1,MS
    DO 130 K=1,MS
    IF(J - K) 110, 100, 110
100 P(J,K) = R(J)/C * (R(K)-1.)/(C-1.)
    GO TO 120
110 P(J,K) = R(J)/C * R(K)/(C-1.)
120 U = J*K
130 SW = SW + U * P(J,K)
    M = M - 1
    IF( M ) 140, 140, 30
140 DO 150 I=1,LC
    SG = SG + G(I)
150 SGS = SGS + G(I)*G(I)
    SSG = SG*SG
    Q = (C*SGS - SSG)/(SS - C*SW)
    LDF = LC - 1
```

```
    PRINT 160
160 FORMAT(1H1)
    DO 170 I=1,4
170 PRINT 180
180 FORMAT(1H0)
    PRINT 190
190 FORMAT(20X,54H**************************************************
   C**)
    PRINT 200
200 FORMAT(20X,1H*,52X,1H*)
    PRINT 220
220 FORMAT(20X,1H*,23X,6HOUTPUT,23X,1H*)
    PRINT 200
    PRINT 230, MS
230 FORMAT(20X,1H*,2X,43HMAXIMUM SCORE  . . . . . ,I
   C5,2X,1H*)
    PRINT 200
    PRINT 240, LC
240 FORMAT(20X,1H*,2X,43HNUMBER OF SAMPLES  . . . . . ,I
   C5,2X,1H*)
    PRINT 200
    PRINT 250, NC
250 FORMAT(20X,1H*,2X,43HNUMBER OF CASES  . . . . . ,I
   C5,2X,1H*)
    DO 260 I=1,LC
    IG(I) = G(I)
    PRINT 200
260 PRINT 270, I, IG(I)
270 FORMAT(20X,1H*,2X,13HSUM OF COLUMN,I3,27H . . . . . .
   C. ,I5,2X,1H*)
    PRINT 200
    PRINT 280, Q
280 FORMAT(20X,1H*,2X,38HCALCULATED VALUE OF Q . . . . . .,F10.5,
   C2X,1H*)
    PRINT 200
    PRINT 290, LDF
290 FORMAT(20X,1H*,2X,43HDEGREES OF FREEDOM . . . . . ,I
   C5,2X,1H*)
    PRINT 200
    PRINT 190
    END
    EOF
```

5. EXAMPLE

```
DRUG TEST       2   3   12

   1 122
   2 011
   3 122
   4 211
   5 011
```

```
 6 011
 7 122
 8 002
 9 001
10 212
11 102
12 011
```

```
*********************************************************
*                                                       *
*                        OUTPUT                         *
*                                                       *
*  MAXIMUM SCORE . . . . . . . . . . . . . .     2     *
*                                                       *
*  NUMBER OF SAMPLES . . . . . . . . . . . .     3     *
*                                                       *
*  NUMBER OF CASES . . . . . . . . . . . . .    12     *
*                                                       *
*  SUM OF COLUMN  1 . . . . . . . . . . . . .     8     *
*                                                       *
*  SUM OF COLUMN  2 . . . . . . . . . . . . .    12     *
*                                                       *
*  SUM OF COLUMN  3 . . . . . . . . . . . . .    18     *
*                                                       *
*  CALCULATED VALUE OF Q . . . . . . . . .  8.94114     *
*                                                       *
*  DEGREES OF FREEDOM . . . . . . . . . . . .     2     *
*                                                       *
*********************************************************
```

LOMA LINDA UNIVERSITY

Graduate School

AN EXTENSION OF THE COCHRAN Q TEST

by

Soo-Young Cho

An Abstract of a Thesis

in Partial Fulfillment of the Requirements

for the Degree Master of Science

in the Field of Biostatistics

August 1971

ABSTRACT

The familiar χ^2 test is ordinarily used to compare percentage distributions of categorical data for two or more independent samples. A more accurate comparison of the percentages is sometimes obtained if the samples consist of matched individuals. The NcNemar test is generally used when comparing only two related samples. Where there are three or more samples the Cochran Q test is used. Both of these tests apply only to the case of dichotomized responses.

This thesis extends the Cochran Q test to the case where the responses are multinomial. A generalized statistic corresponding to Cochran's result is obtained. As in Cochran's case this statistic has asymptotically a χ^2 distribution with degrees of freedom equal to one less than the number of samples.