TESTING EQUALITY OF PROPORTIONS WITH INCOMPLETE CORRELATED DATA

by

Gregory Campbell
Purdue University

Department of Statistics
Division of Mathematical Sciences
Mimeo Series #80-18

July 1980

TESTING EQUALITY OF PROPORTIONS WITH INCOMPLETE CORRELATED DATA

by

Gregory Campbell
Purdue University

Let (ψ_i, ϕ_i) be independent, identically distributed pairs of zero-one random variables with (possible) dependence of ψ_i and ϕ_i within the pair. For n pairs, both variables are observed, but for m_1 additional pairs only ψ_i is observed and for m_2 others ϕ_i is observed. If $p_1=P\{\psi_i=1\}$ and $p_2=P\{\phi_i=1\}$, the problem is to test $p_1=p_2$. Maximum likelihood estimates of p_1 and p_2 are obtained via the EM algorithm. A test statistic is developed whose null distribution is asymptotically chi-square with one degree of freedom (as n and either m_1 or m_2 tend to infinity). If $m_1=m_2=0$ the statistic reduces to that of McNemar's test; if n=0, it is equivalent to the statistic for testing equality of two independent proportions. An example is presented.

TESTING EQUALITY OF PROPORTIONS WITH INCOMPLETE CORRELATED DATA

by

Gregory Campbell
Purdue University

1. Introduction. The large sample test procedures for equality of two proportions are well known. If the two samples utilized to estimate the proportions are independent of each other, then the quite familiar large sample statistic for the difference of two independent proportions is used to test the null hypothesis of equality of proportions. If each individual in a sample undergoes both the trial (with success or failure as an outcome) for the first proportion as well as for the second, the sample proportions are no longer necessarily independent, and hence the test of McNemar (1947) for correlated proportions can be employed. The situation treated in this paper is that in which some of the individuals in the samples undergo both Bernoulli trials whereas others undergo just one of the pair of trials; this will be called the case of incomplete correlated data. An important example of such a situation occurs in a rotating sample for which one wants to assess whether a change in proportion of some response has occurred over a time interval during which the sample has been partially rotated.

The notation for this problem is established in Section 2, along with an introduction of the well-known test statistics for the separate cases of independent samples and of correlated pairs in the sample. The likelihood equation is written in Section 3 for this situation with incomplete correlated data, and the maximum likelihood solution given. This then provides a test for the equality of proportions. In the final section, an example is presented and the procedure discussed.

2. The problem and special cases. Let $\{(\psi_i, \varphi_i)\}$ $(i = 1,2,\ldots)$ denote independent, identically distributed pairs of Bernoulli random variables such that it is not known that ψ_i and φ_i are independent within the pair. It is assumed that for n pairs, both zero-one variables are observed. For m_1 additional pairs only ψ_i in the pair is observed and for another m_2 pairs only φ_i is observed. Let $p_1 = P\{\psi_i = 1\}$ and $p_2 = P\{\varphi_i = 1\}$. The problem is to test the null hypothesis of equality $(p_1=p_2)$ versus the general alternative $(p_1 \neq p_2)$.

Some notation is introduced. Let A be the number of the n pairs for for which $(\psi_i, \varphi_i) = (0,0)$. Let B, C, and D denote the number of n pairs for which (ψ_i, φ_i) is equal to (0,1), (1,0), and (1,1), respectively. Let E(F) denote the number of m_1 for which $\psi_i = 0(1)$ and let G(H) denote the number of m_2 for which $\varphi_i = 0(1)$. The data can be presented in the following table which is supplemented by additional marginals:

ψ \ φ	0	1		
0	A	B		E
1	C	D		F
			n	m_1

G	H	m_2

Assume that n=0 and m_1 and m_2 are both non-zero. Then the test of H_0: $p_1=p_2$ can be treated for this case of independent samples using the test statistic

$$z = \frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{m_1}+\frac{1}{m_2})}} \qquad (1)$$

where $\hat{p}_1 = F/m_1$, $\hat{p}_2 = H/m_2$, and $\hat{p} = (F+H)/(m_1+m_2)$. Under H_0, z is asymptotically standard normal, so that for large m_1 and m_2, z^2 is approximately chi-square with one degree of freedom.

If $m_1=m_2=0$ and n is non-zero, the procedure of McNemar (1947) can be used to test H_0. The test is conditional on (B+C). Under H_0, B has a binomial distribution with parameters $\frac{1}{2}$ and (B+C), so McNemar's test is then simply a binomial test for the parameter $\frac{1}{2}$ using the test statistic B. If B+C is large, then the large sample approximation that under H_0 $X_1^2 = (B-C)^2/(B+C)$ is asymptotically chi-square with one degree of freedom can be used.

Let $\pi_{ij} = P\{\psi = i, \varphi = j\}$ for i,j = 0,1. Of course the sum of the π's is in the following table with row and column totals:

		φ = 0	φ = 1	
ψ	0	π_{00}	π_{01}	$1-p_1$
	1	π_{10}	π_{11}	p_1
		$1-p_2$	p_2	1

Note that since $p_1 = \pi_{10} + \pi_{11}$ and $p_2 = \pi_{01} + \pi_{11}$, testing $p_1 = p_2$ is equivalent to testing $\pi_{10} = \pi_{01}$.

The test for independence ($\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$) has been treated in this case of supplementary margins by Chen and Fienberg (1974) using a log-linear approach. Here the hypothesis of interest is that of symmetry of the π's ($\pi_{10} = \pi_{01}$) or, alternatively, homogeneous marginals. It is possible to develop this problem in a log-linear framework also (see Bishop, Fienberg, and Holland for the case $m_1 = m_2 = 0$.)

3. Maximum likelihood estimation and two test statistics. Consider the multinomial model in which n observations are categorized into the cells of the 2x2 table and m_1 are categorized into row 1 or 2 and m_2 into columns 1 or 2. The sample sizes n, m_1, and m_2 are fixed. The likelihood expression is given by:

$$L = \binom{n}{A,B,C,D} \pi_{00}^{A} \pi_{01}^{B} \pi_{10}^{C} \pi_{11}^{D} \binom{m_1}{E} (\pi_{00} + \pi_{01})^{E} \cdot (\pi_{10} + \pi_{11})^{F} \binom{m_2}{G} (\pi_{00} + \pi_{10})^{G} (\pi_{01} + \pi_{11})^{H}. \qquad (2)$$

The likelihood can be maximized using the EM algorithm of Dempster, Laird, and Rubin (1977). To find the unrestricted maximum likelihood estimates of the π's, proceed as follows:

1. Initialize

$$\hat{E}_{ij}^{(0)} = 0 \qquad i,j = 0,1;$$

$$\hat{\pi}_{00}^{(0)} = A/n; \qquad \hat{\pi}_{01}^{(0)} = B/n;$$

$$\hat{\pi}_{10}^{(0)} = C/n; \qquad \hat{\pi}_{11}^{(0)} = D/n.$$

2. At stage (k+1), (for k = 0,1,2,...), let

$$\hat{E}_{00}^{(k+1)} = A + \frac{\hat{\pi}_{00}^{(k)}}{\hat{\pi}_{00}^{(k)} + \hat{\pi}_{01}^{(k)}} E + \frac{\hat{\pi}_{00}^{(k)}}{\hat{\pi}_{00}^{(k)} + \hat{\pi}_{10}^{(k)}} G$$

$$\hat{E}_{01}^{(k+1)} = B + \frac{\hat{\pi}_{01}^{(k)}}{\hat{\pi}_{00}^{(k)} + \hat{\pi}_{01}^{(k)}} E + \frac{\hat{\pi}_{01}^{(k)}}{\hat{\pi}_{01}^{(k)} + \hat{\pi}_{11}^{(k)}} H$$

$$\hat{E}_{10}^{(k)} = C + \frac{\hat{\pi}_{10}^{(k)}}{\hat{\pi}_{10}^{(k)} + \hat{\pi}_{11}^{(k)}} F + \frac{\hat{\pi}_{10}^{(k)}}{\hat{\pi}_{00}^{(k)} + \hat{\pi}_{10}^{(k)}} G$$

$$\hat{E}_{11}^{(k)} = D + \frac{\hat{\pi}_{11}^{(k)}}{\hat{\pi}_{10}^{(k)} + \hat{\pi}_{11}^{(k)}} F + \frac{\hat{\pi}_{11}^{(k)}}{\hat{\pi}_{01}^{(k)} + \hat{\pi}_{11}^{(k)}} H$$

The computation of the E's is the expectation (E-) step of the EM algorithm. This is followed by the maximization (M-) step which is the calculation of $\hat{\pi}_{ij}^{(k+1)}$:

$$\hat{\pi}_{ij}^{(k+1)} = E_{ij}^{(k+1)}/(n + m_1 + m_2) \quad i,j = 0,1.$$

This iterative procedure converges to the unrestricted maximum likelihood estimates $\hat{\pi}_{ij}$.

It is possible to apply this same technique to find the maximum likelihood estimators under the restriction $\pi_{10} = \pi_{01}$. The procedure is:

1. Initialize

$$\tilde{E}_{ij}^{(0)} = 0 \qquad i,j = 0,1;$$

$$\tilde{\pi}_{00}^{(0)} = A/n; \qquad \tilde{\pi}_{11}^{(0)} = D/n;$$

$$\tilde{\pi}_{10}^{(0)} = \tilde{\pi}_{01}^{(0)} = (B+C)/2n.$$

2. The E-step is exactly as in Step 2 of the unrestricted case except $\tilde{E}$ replaces $\hat{E}$ and $\tilde{\pi}$ replaces $\hat{\pi}$. The M-step differs:

$$\tilde{\pi}_{00}^{(k+1)} = \tilde{E}_{00}^{(k+1)}/(n+m_1+m_2); \; \tilde{\pi}_{11}^{(k+1)} = \tilde{E}_{11}^{(k+1)}/(n+m_1+m_2)$$

$$\tilde{\pi}_{10}^{(k+1)} = \tilde{\pi}_{01}^{(k+1)} = (\tilde{E}_{10}^{(k+1)} + \tilde{E}_{01}^{(k+1)})/2(n+m_1+m_2).$$

This iterative procedure converges to the restricted maximum likelihood estimates $\tilde{\pi}_{ij}$, with $\tilde{\pi}_{10} = \tilde{\pi}_{01}$.

A test of H_0: $\pi_{10} = \pi_{01}$ versus the alternative $\pi_{10} \neq \pi_{01}$ can be obtained from Pearson's X^2:

$$X^2 = \sum_{i=0}^{1} \sum_{j=0}^{1} \frac{(\hat{E}_{ij} - \tilde{E}_{ij})^2}{\tilde{E}_{ij}}, \tag{3}$$

where $\hat{E}_{ij} = (n+m_1+m_2)\,\hat{\pi}_{ij}$ and $\tilde{E}_{ij} = (n+m_1+m_2)\,\tilde{\pi}_{ij}$. Under H_0, X^2 has asymptotically a chi-square distribution with one degree of freedom. The single degree of freedom follows from the fact that under H_0 there are 2 linearly independent parameters in the model, whereas under the alternative there are 3. This is the proposed procedure for testing equality of proportions.

There is another test statistic in the event n, m_1, and m_2 are all large; namely, take z from equation (1) and the statistic X_1^2 of McNemar and calculate $X_2^2 = z^2 + X_1^2$. Since z is independent of X_1^2, the resultant statistic X_2^2 is approximately chi-square with two degrees of freedom.

4. Example and discussion. Consider the following example. One hundred and fifty individuals are asked if they favor or oppose a particular course of action. Six months later, 50 individuals have been randomly chosen to be rotated out of the sample and are replaced by 50 new

individuals. These 150 are then asked the same question. The research question of interest is to decide if the proportion in favor has changed. The data are:

		Currently Oppose	Currently Favor	
Six Months Ago	Oppose	30	20	26
	Favor	30	20	24
		33	17	

The unrestricted maximum likelihood estimates of the π's converge in five or six iterations to:

$$\hat{\pi}_{00} = .3143 \qquad \hat{\pi}_{01} = .1924$$

$$\hat{\pi}_{10} = .3057 \qquad \hat{\pi}_{11} = .1876$$

Subject to the restriction $\pi_{10} = \pi_{01}$ the maximum likelihood estimates are:

$$\tilde{\pi}_{00} = .3150 \qquad \tilde{\pi}_{01} = .2483$$

$$\tilde{\pi}_{10} = .2483 \qquad \tilde{\pi}_{11} = .1884$$

Thus,

$$X^2 = \frac{(62.86-63.00)^2}{63.00} + \frac{(38.48-49.66)^2}{49.66} + \frac{(61.14-49.66)^2}{49.66} + \frac{(37.52-37.68)^2}{37.68}$$

$$= .000 + 2.517 + 2.654 + .001 = 5.172.$$

The P-value of this is about .025.

Consider the second analysis using X_2^2. The McNemar statistic is $X_1^2 = (B-C)^2/(B+C) = 10^2/50 = 2.000$. The statistic z^2 is:

$$z^2 = \frac{(.61-.52)^2}{[(.59)(.41)(\frac{1}{50} + \frac{1}{50})]} = 2.026$$

Thus, $X_2^2 = 4.026$ which corresponds to a P-value using 2 degrees of freedom of approximately .14. Further, note that P-values of X_1^2 and z^2 (using 1 degree of freedom) are approximately .16 each.

At first glance the procedure using the statistic X^2 might be thought to be superior in that the degrees of freedom is smaller than that for X_2^2. However, the sample size for X^2 is $(n+m_1+m_2)$ whereas for X_1^2 and z^2 the sample sizes are n and (m_1+m_2), respectively, so this fact confounds the reduction in degrees of freedom.

However, there is an intuitive reason for preferring X^2, the statistic based on the EM algorithm. It is possible for z^2 to be large due to $\hat{p}_1 < \hat{p}_2$ and yet X_1^2 large in that the estimated proportions are reversed $(\frac{C+D}{n} > \frac{B+D}{n})$. This would result in a large X_2^2, whereas the effects in different directions would to some extent cancel in the statistic X_2^2.

Acknowledgement. The computer programming of this procedure which was used in Section 4 was done by Tom Broene.

REFERENCES

Bishop, Yvonne M.M., Fienberg, Stephen E., and Holland, Paul W. (1975). Discrete Multivariate Analysis. MIT Press, Cambridge, MA.

Chen, Tar and Fienberg, Stephen E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. Biometrics 30 629-642.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc., Ser. B 39 1-22.

McNemar, Quinn. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12 153-157.