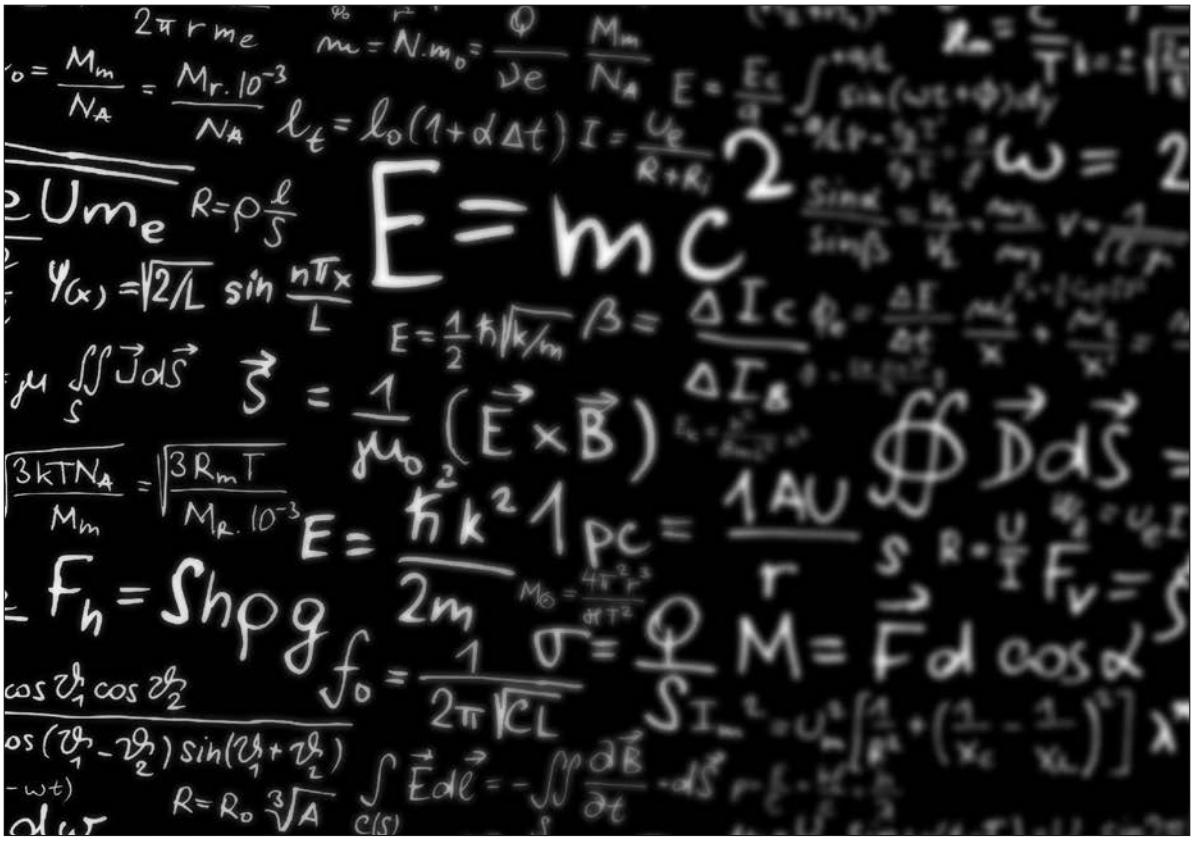


# Asian Journal of Mathematics & Statistics

ISSN 1994-5418

science alert

ANSInet
an open access publisher
http://ansinet.com

Asian Journal of Mathematics and Statistics 2 (1): 15-19, 2009
ISSN 1994-5418


# An Approximate Likelihood Ratio Method for Testing Equality of Two Dependent Proportions

Tadewos Koroto
Department of Statistics and Demography,
The National University of Lesotho, Kingdom of Lesotho

**Abstract:** This study considers an approximate likelihood ratio test for equality of two dependent proportions. A bivariate probability distribution of a specified form is assumed and the likelihood ratio statistic is approximated from this distribution. The distribution accounts for the correlation of the underlying two binomial random variables. The application of the procedure on data resulting from treatment of TB patients shows that the proposed test can be used as an alternative test for data involving non-response.

**Key words:** Dependent proportions, approximate likelihood-ratio test, bivariate binomial distribution, non-response

## INTRODUCTION

Dependent proportions are common in biomedical studies, such as studies that focus on changes in subjects' responses over time, observations on severity of pain at pairs of body locations and retrospective case-control studies. Agresti (2002) describes several inference methods for such data. In a simple study involving a binary response, the data for dependent observations are displayed in a 2×2 contingency table, where, $n_{11}$, $n_{12}$, $n_{21}$ and $n_{22}$ denote, respectively, the number of pairs that are successes for both observations, successes for the first observation but failures for the second observation, failures for the first observation but successes for the second observation and failures for both observations. Commonly the four cell counts of the 2×2 contingency table are assumed to follow a multinomial distribution. Inferences about the parameters of the underlying distribution are based on the cell counts (Newcombe, 1998; Liu *et al*., 2002; Agresti and Min, 2005).

In the presence of non-response the individual cell counts are often difficult to obtain. A Bayesian approach described by Ghosh *et al*. (2000) could be used for dealing with problems of non-response if there is some auxiliary information. Suppose that we could not observe the $n_{ij}$ due to the problem of non-response but the marginal totals are known. Assuming binomial distributions for the first row total and the first column total, the corresponding probabilities are:

$$P(Y_j = y_j) = \binom{n}{y_j} p_j^{y_j} (1-p_j)^{n-y_j}, \quad j = 1, 2$$

Comparison of $p_1$ and $p_2$ is not straight forward because the subjects comprising the row and column counts are not independent. That is to say, if $p_1$ is the proportion of patients who develop a specified type of complication and is the proportion of patients who develop a second type of complication, then the two proportions are not independent as some patients could exhibit both types of complications. Here, a bivariate probability distribution is proposed that accounts for the dependence of $Y_1$ and $Y_2$. Using the joint distribution and the realized values $y_1$ and $y_2$, a likelihood ratio test is suggested for testing the equality of the two proportions.

## LIKELIHOOD RATIO TEST

### Bivariate Binomial Distribution

Suppose different discrete events, which are naturally related, are observed simultaneously. There are a number of multivariate distributions that could be used to model such events. The problem is that a bivariate distribution which is of a binomial type and which allows for dependence is not readily available. If the two random variables satisfy the Poisson assumptions, then one could use the bivariate Poisson distribution introduced by Kocherlakota and Kocherlakota (1992). It reads as:

$$P(Y_1 = y_1, Y_2 = y_2) = e^{-\lambda-\mu-\alpha}\sum_{l=0}^{\min(y_1,y_2)}\frac{\alpha^l\lambda^{y_1-l}\mu^{y_2-l}}{l!(y_1-l)!(y_2-l)!} \quad (1)$$

The resulting marginal distributions of $Y_1$ and $Y_2$ are Poisson with parameters $\lambda$ and $\mu$, respectively. The correlation of $Y_1$ and $Y_2$ is assumed to be positive. The parameter $\alpha$ (>0) represents the correlation of $Y_1$ and $Y_2$. The higher $\alpha$, the stronger the correlation. If $\alpha$ is high, the probability that each of the variables takes on a large value $y_j$ is higher than the probability that one of them takes on a small value and the other one a large value. That is to say, the two variables are highly concordant.

Suppose that two positively correlated binomial random variables $Y_1$ and $Y_2$ are assumed to follow a bivariate distribution where the probability $P(Y_1 = y_1, Y_2 = y_2)$ increases by some factor as the correlation increases. The following distribution could be assumed in this case:

$$P(Y_1 = y_1, Y_2 = y_2) = K\binom{n}{y_1}p_1^{y_1}(1-p_1)^{n-y_1}\binom{n}{y_2}p_2^{y_2}(1-p_2)^{n-y_2}\alpha^{\min(y_1,y_2)} \quad (2)$$

where, $y_1$ and $y_2$ taking values 0, 1, . . . , n.

The factor K does not depend on $y_1$ and $y_2$ as it is used to normalize the distribution, that is, to make the summation over all $y_1$ and $y_2$ equal 1. When the two random variables are independent, or $\alpha = 1$, the joint distribution (Eq. 2) reduces to a product of two binomial distributions. The main problem is that it is not clear whether or not we can get a binomial marginal distribution from the above bivariate distribution. However, since the objective is to test the equality of the two proportions using a likelihood ratio test, it is reasonable to expect the resulting likelihood ratio statistic to provide a valid comparison unless the actual values of $\alpha$ in the numerator and denominator of the likelihood ratio are much larger than 1.

### Approximation of the Likelihood Ratio Statistic

A test for comparing the two probabilities of success is describes here using the bivariate binomial distribution introduced above.

Given n and the realized values $y_1$ and $y_2$ , the objective is to test:

$$H_0: p_1 = p_2 \text{ versus } H_1: p_1 \neq p_2$$

In an ordinary likelihood ratio test one would take the logarithm of the joint probability (Eq. 2) and maximize it with respect to the parameters $p_1$, $p_2$ and under $H_0$ and $H_1$ separately. Direct maximization is not possible because the factor K in Eq. 2 also depends on the parameters. Instead one could try to identify an optimal point $(p_1^*, p_2^*)$ in the neighborhood of $(y_1/n, y_2/n)$ for specified values of $\alpha$. The likely range for $\alpha$ is expected to be small (i.e., 1 to 2, or 1 to 3). As $\alpha$ increases the value of K decreases and as a result the likelihood function starts to decline. Therefore, it is sufficient to try

values of $\alpha$ such as 1.1, 1.2, …, 3 and then refine the search once the likely range is identified. The resulting optimal values , $p_1^*$, $p_2^*$ and $\alpha^*$ are taken as the likelihood estimates under $H_1$.

A similar search procedure is used to identify $p_0 = p_1 = p_2$ and $\alpha^0$ that maximize the likelihood function under $H_0$. If the optimal value of is saved for each pair $p_1$ and $p_2$ during the search for $p_1^*$, $p_2^*$ and $\alpha^*$ then one could simply look for the optimal among points where $p_1 = p_2$.

Finally, the estimates are substituted for $p_1$, $p_2$ and $\alpha$ in the likelihood function to get $l^*$ and $l_0$, the maximized log-likelihood under $H_1$ and $H_0$, respectively. The resulting approximate likelihood-ratio statistic is defined as:

$$G^2 = -2(l^* - l^*) \tag{3}$$

where, $H_0$ is true $G^2$ is assumed to follow a chi-square distribution with 1 degree of freedom for large n.

The test is, therefore, to reject the null hypothesis when $G^2$ is higher than $\chi_1^2$ at specified level of significance. The basic S-plus commands for finding the optimal values of $p_1$, $p_2$, $\alpha$ and the corresponding value of the likelihood are shown in the Appendix.

**Appendix**

S-plus commands for obtaining the estimates:

```
n<- 30
y1<-14
y2<-5
p1min<-0.11
p1max<-0.65
pint<-0.02
p2min<-0.05
p2max<-0.39
amin<-1.00
amax<-2.5
aint<-0.02
n<-n+1
h1<-rep(.0,n)
g<-h1
t1<-h1
v<-h1
v2<-h1
nh<-ceiling((p1max-p1min)/pint)
nl<-ceiling((p2max-p2min)/pint)
nr<-ceiling((amax-amin)/aint)
dimx<-nh*nl
xval<-matrix(.0, nrow=dimx , ncol=4)
lk<-1
t <-min(y1,y2)
p1<-p1min
for(l1 in 1:nh)
      {
     p2<-p2min
     for(nk in 1:nl)
        {
        for(I in 1:n)
             {
              v[i] <-choose(n-1,i-1)*p1^(I-1)* (1-p1)^(n-I)
              g[i] <-choose(n-1,i-1)*p2^(I-1)* (1-p2)^(n-I)
             }
        vml<-matrix(.0, nrow=nr, ncol=2)
        tvml<-vml
        a<-amin
        for(j1 in 1:nr)
```

```
        {
        pnum<-choose(n-1,y1)*p1^y1* (1-p1)^(n-1-y1)*
                  choose(n-1,y2)*p2^y2* (1-p2)^(n-1-y2)*a^t
         v2<-rep(.0,n)
         for(I in 1:n)
             {
              t1[i] <-a^(I-1)
              for(j in 1:i) v2[j] <-a^(j-1)*(1- a^(I-j))
              h1[i] <-sum(v2*v)
              }
          pr<-sum(g*(h1+t1))
          prq<-pnum/pr
          vml[j1,] <-c(a,prq)
           a<-a + aint
         }
      tvml<-vml[order(vlm[,2]),1:2]
      cp1 <-tvml[nr,1]
      cp2 <-tvml[nr,2]
      xval[lk, ] <-c(p1,p2,cp1,cp2)
      lk<-lk+1
      p2<-p2+ pint
      }
   p1 <-p1+ pint
   }  # xval is a matrix of the estimates, each row contains p1, p2, the corresponding optimal and the value of the likelihood function when the parameters take on these values.
```

## EXAMPLE

From the records of TB patients treated at a tuberculosis treatment center in Lesotho, Southern Africa, a sample of 30 TB patients was selected. Fourteen of the patients had improper follow-up (not being assigned an observer, non-compliance, controlled tests not done as prescribed, treatment taking longer than 6 months, etc.). Five of the patients either died or failed to respond to the treatment. The objective is to test whether there is a significant difference between the proportion of patients who had improper follow-up and the proportion of patients who died or failed to respond to treatment.

Suppose $Y_1$ and $Y_2$ denote the number of patients in a sample of size 30 that exhibit the respective outcomes. Therefore, in the notations of the preceding sections, $n = 30$, $y_1 = 14$ and $y_2 = 5$. Following the procedure described earlier it is assumed that the distribution of $Y_1$ and $Y_2$ is bivariate binomial introduced earlier. To speed up the identification of optimal points, the normalizing factor is found as follows:

$$\sum_{i=0}^{n}\sum_{j=0}^{n}P(Y_1 = i, Y_2 = j) = K\sum_{i=0}^{n}\sum_{j=0}^{n}\binom{n}{i}p_1^{\,i}(1-p_1)^{n-i}\binom{n}{j}p_2^{\,j}(1-p_2)^{n-j}\alpha^{\min(i,j)} = 1$$

That is,

$$\begin{aligned}\frac{1}{K} &= \sum_{j=0}^{n}\binom{n}{j}p_2^{\,j}(1-p_2)^{n-j}\sum_{i=0}^{j}\binom{n}{i}p_1^{\,i}(1-p_1)^{n-i}\alpha^{i}\\ &+ \sum_{j=0}^{n}\binom{n}{j}p_2^{\,j}(1-p_2)^{n-j}\sum_{i=j+1}^{n}\binom{n}{i}p_1^{\,i}(1-p_1)^{n-i}\alpha^{j}\\ &= \sum_{j=0}^{n}\binom{n}{j}p_2^{\,j}(1-p_2)^{n-j}\sum_{i=0}^{j}\binom{n}{i}p_1^{\,i}(1-p_1)^{n-i}\alpha^{i}(1-\alpha^{j-i})\\ &+ \sum_{j=0}^{n}\binom{n}{j}p_2^{\,j}(1-p_2)^{n-j}\alpha^{j}\end{aligned}$$

This value is substituted for K in the likelihood when the algorithm is used to identify the optimal point.

The test of interest is

$$H_0: p_1 = p_2 \text{ versus } H_1: p_1 \neq p_2$$

Since, $y_1/n = 14/30 = 0.47$, $y_2/n = 5/30 = 0.17$, it is sufficient to search for $p_1$ in 0.11 to 0.65 and for $p_2$ in 0.05 to 0.39. Then choose evenly spaced points within the two ranges of values with 0.02 as spacing value. Initially the value of $\alpha$ was made to vary in [1, 4] using 0.1 as spacing value. After observing that, the likelihood started to decline for all possible points $(p_1,p_2)$ when $\alpha$ exceeded 2.5, the refined search is restricted to this range with spacing value of 0.02. The optimal values under the null hypothesis were $p_1 = p_2 = 0.35$ and $\alpha = 1.08$ and the resulting log-likelihood was -6.50229. Under the alternative hypothesis $p_1 = 0.43$, $p_2 = 0.19$ and $\alpha = 1.14$ with log-likelihood of -3.5791. This resulted in an approximate likelihood-ratio statistic $G^2 = -2(-3.5791-(-6.50229)) = 5.846$, which has a p-value of 0.01 assuming a chi-square distribution with 1 degree of freedom. Since, the p-value is small, there is a strong evidence to conclude that the proportion of patients who received improper follow-up and the proportion of patients who died or failed to respond to the treatment are not equivalent.

## CONCLUSIONS

In the study, an approximate likelihood-ratio test is proposed for comparing two dependent proportions. Unlike the standard analysis, the procedure uses only the marginal totals of the contingency table, which makes it useful for dealing with non-response. The algorithms presented in the study can also be applied for comparing pairs of such proportions in terms of the strength of their correlation.

The main advantage of the proposed test is that it does not require sophisticated computation adopted by other procedures for making inference when the data have problems of non-response. The procedures used in the test are simple and can be applied without requiring extensive re-sampling methods.

Although, no attempt was made in the study to see the comparative performance of the proposed test procedure, the mathematical arguments and computation of the test statistic are simple and straightforward. Regarding the power of the test, it is clear that no theoretical significance is placed on the distribution adopted for the test statistic. However, as the test statistic resembles a likelihood-ratio statistic, the power of the test is expected to be comparable to that of techniques that employ re-sampling or Monte Carlo estimation to handle non-response problems.

## REFERENCES

Agresti, A., 2002. Categorical Data Analysis. 2nd Edn. John Wiley and Sons, Inc., Publication, New Jersey, ISBN: 9780471360933.

Agresti, A. and Y. Min, 2005. Simple improved confidence intervals for comparing matched proportions. Statist. Med., 24: 729-740.

Ghosh, M., M. Chen, A. Ghosh and A. Agresti, 2000. Hierarchical Bayesian analysis of binary matched pairs data. Statist. Sinica, 10: 647-657.

Kocherlakota, S. and K. Kocherlakota, 1992. Bivariate Discrete Distributions. 1st Edn., Marcel Dekker, New York, ISBN: 9780824787028.

Liu, J., H. Hsueh, E. Hsieh and J.J. Chen, 2002. Test for equivalence or non-inferiority for paired binary data. Statist. Med., 21: 231-245.

Newcombe, R., 1998. Improved confidence intervals for the difference between binomial proportions based on paired data. Statist. Med., 17: 2635-2650.