Conference on Applied Statistics in Agriculture | 2003 - 15th Annual Conference Proceedings

HOTELLING'S T^2 APPROXIMATION FOR BIVARIATE DICHOTOMOUS DATA

Pradeep Singh

Imad Khamis

James Higgins

Recommended Citation

Singh, Pradeep; Khamis, Imad; and Higgins, James (2003). "HOTELLING'S T^2 APPROXIMATION FOR BIVARIATE DICHOTOMOUS DATA," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1187

HOTELLING'S T^2 APPROXIMATION FOR BIVARIATE DICHOTOMOUS DATA

Pradeep Singh
Department of Mathematics
SE Missouri State University
Cape Girardeau. MO 63701
Email: psingh@semo.edu

Imad Khamis
Department of Mathematics
SE Missouri State University
Cape Girardeau, MO 63701
Email: ikhamis@semo.edu

James Higgins
Department of Statistics
Kansas State University
Manhattan, KS 66506
Email: higgins@stat.ksu.edu

ABSTRACT

The comparison of the means of two treatments or populations when more than one variable is measured may be done using Hotelling's T^2 statistic. In many real world situations the data obtained are dichotomous, and the assumption of multivariate normality upon which Hotelling's T^2 is based is no longer valid. In this paper, an approximate Hotelling T^2 test is proposed for bivariate dichotomous data and empirically evaluated in terms of Type I error rate. It is shown that the approximation does a good job of controlling the Type I error rate for a range of bivariate parameters even for relatively small sample sizes.

Key Words and Phrases: bivariate data, dichotomous response, Hotelling T^2, multivariate analysis

1. Introduction

It is very common to have multivariate data in which the individual variates are dichotomous, i.e. take one of just two possible values, 0 or 1. Multivariate models with binary response have found extensive application in reliability and biostatistics. In meat sciences this type of data may arise in comparing the contamination of beef carcasses under two methods of decontamination where bivariate responses are presence or absence of two types of bacteria on the carcasses.

If observations are selected randomly from multivariate normal populations, a common multivariate statistic for comparing two populations is Hotelling T^2 [Anderson (1984)]. A permutation test that is based on the computation of the t-statistic for each of the response variables is also appropriate for multivariate data. Blair et. Al. [1994] showed that one sided multivariate tests can enjoy substantial power advantages over Hotelling T^2 test under certain conditions.

In a two group experiment with binary responses, the central problem is to describe the joint distribution of a set of binary variables. The oldest approach to multivariate binary data is to define indices of association following essentially Yule. Goodman and Kruskal [1954, 1959,

1963] have reviewed and extended this work. Recently, Bilder [2000] has used a Pearson like chi-square statistic to analyze multi-response contingency tables.

In this paper, an adaptation of Hotelling T^2 is proposed for comparison of two populations having bivariate dichotomous responses. An empirical study is done to examine the type I error rate.

2. Bivariate Dichotomous Data

We consider the problem of comparing two treatments in experiments in which bivariate dichotomous response variables are measured on each experimental unit. For example, suppose an entomologist wishes to compare the effectiveness of a broad spectrum insecticide in controlling two pests affecting a particular plant. Suppose one set of experimental plants in a completely random design is treated with the insecticide and the other set acts as control.

For group I, the response would be X_1 and X_2

$$X_1 = \begin{cases} 1 \text{ when pest1 is present on a plant} \\ 0 \text{ otherwise} \end{cases}$$

$$X_2 = \begin{cases} 1 \text{ when pest2 is present on a plant} \\ 0 \text{ otherwise} \end{cases}$$

For group II, the response would be Y_1 and Y_2

$$Y_1 = \begin{cases} 1 \text{ when pest1 is present on a plant} \\ 0 \text{ otherwise} \end{cases}$$

$$Y_2 = \begin{cases} 1 \text{ when pest2 is present on a plant} \\ 0 \text{ otherwise} \end{cases}$$

In matrix notation we can represent the data as

$$X = \begin{matrix} X_1 & X_2 \end{matrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 1 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} \text{ and } Y = \begin{matrix} Y_1 & Y_2 \end{matrix} \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 0 & 1 \end{pmatrix} \quad (1)$$

The hypotheses to be tested are H_0: $P_X = P_Y$, H_a: $P_X \neq P_Y$ where $P_X = (p_{x_1} \quad p_{x_2})$ and $P_Y = (p_{y_1} \quad p_{y_2})$ are vectors of expected proportions or population proportions of pest 1 and pest 2 present on plants.

3. Permutation Test

A two sample permutation test is carried out by randomly assigning experimental units or subjects to one of two treatments. All possible two-sample data sets are obtained by permuting $m + n$ observations among two groups. There are $\binom{m+n}{m}$ such data sets.

The permutation principle states that the permutation distribution is an appropriate reference distribution for determining the p-value of a test and deciding whether or not a test is statistically significant. One may extend permutation tests to the multivariate setting. Here one permutes observed vectors among the groups, keeping the vectors intact in doing the permutations. See Higgins [2003] Chapter 6 for more details.

A multivariate permutation test for this problem may be carried out using PROC MULTTEST in SAS®. The permutation test is based on the computation of a t-statistic for each of the response variables. Let t_j denote the two sample t-statistic for testing the difference between the means of treatments 1 and 2 on response variable j, $j = 1,2,....k$. The statistic computed in MULTTEST is maximum of the absolute values of the t-statistics

$$T_{max\ abs} = max(|t_1|, |t_2|, .., |t_k|).$$

The permutation p-value for the jth variate is the proportion of the permutation distribution of $T_{max\ abs}$ greater than or equal to the observed value of $|t_j|$. Because the permutation distribution is used as the reference distribution, the statistic may be applied to dichotomous data as well as continuous data without concern about the violation of the normality assumption associated with the parametric test. One may also use bootstrap sampling, or sampling with replacement from the set of multivariate vectors, instead of permutation sampling. Bilder [2000] considered bootstrap sampling for this problem. One may also carry out a one-sided multivariate permutation test although this is not implemented in SAS®.

4. Hotelling T^2 Approximation

Suppose we have n observations from population 1 and m observations from population 2. There are k response variables for each population. The response matrices are represented by

$$X = \begin{bmatrix} x_{11} & \dots & x_{1k} \\ x_{21} & \dots & x_{2k} \\ . & & . \\ . & & . \\ . & & . \\ x_{n1} & \dots & x_{nk} \end{bmatrix} \qquad Y = \begin{bmatrix} y_{11} & \dots & y_{1k} \\ y_{21} & \dots & y_{2k} \\ . & & . \\ . & & . \\ . & & . \\ y_{m1} & \dots & y_{mk} \end{bmatrix} .$$

We assume that responses are distributed as multivariate normal with mean and covariance as shown below.

$X = [X_1 X_2 ..... X_k] \sim MVN(\mu_X, \Sigma_X)$, where $X_i = [x_{i1} ... .x_{in}]^{/}$, $i = 1, ..., n.$

$Y = [Y_1 Y_2 ..... Y_k] \sim MVN(\mu_Y, \Sigma_Y)$, where $Y_i = [y_{i1} ... .y_{im}]^{/}$, $i = 1, ..., m.$

Hotelling's T^2 statistic, which assumes that $\Sigma_X = \Sigma_Y = \Sigma$, is given by

$$T^2 = \frac{mn}{n+m}\left(\bar{X} - \bar{Y}\right)^{'} \left(S_{pooled}\right)^{-1} \left(\bar{X} - \bar{Y}\right), \qquad (2)$$

where $S_{pooled} = \dfrac{(n-1)S_X + (m-1)S_Y}{n+m-2}$ and S_X, S_Y are sample variance–covariance matrices of X and Y respectively. Under the null hypothesis $F = \dfrac{m+n-k-1}{(n+m-2)k} T^2$ has an F-distribution with degrees of freedom k and $m+n-k-1$.

Now suppose the data are dichotomous. Because of the Central Limit Theorem, the analysis of univariate dichotomous data may be done with normal approximations for large samples. The approximations are generally good even for moderate sample sizes if the population proportions are not too close to 0 or 1. The question of interest here is the possible use of multivariate normal methods to analyze bivariate dichotomous data.

The suggested approach is to apply (2) directly to the dichotomous data just as if 0's and 1's are quantitative observations. The expected values are given by

$$E(X) = P_x = \left[p_{x1}\ p_{x2}\right] \text{ and } E(Y) = P_y = \left[p_{Y1}\ p_{Y2}\right],$$

where $p_{x1} = P(X_1 = 1)$, $p_{x2} = P(X_2 = 1)$, $p_{y1} = P(Y_1 = 1)$, $p_{y2} = P(Y_2 = 1)$. The covariance between X_1 and X_2 is given by

$$Cov(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2)$$
$$= p_{x12} - p_{x1}\, p_{x2}.$$

Similarly the covariance between Y_1 and Y_2 is given by

$$Cov(Y_1, Y_2) = E(Y_1Y_2) - E(Y_1)E(Y_2)$$
$$= p_{y12} - p_{y1}\, p_{y2}.$$

Note that p_{x12} is the probability that both X_1 and X_2 are 1, and similarly p_{y12} is the probability that both Y_1 and Y_2 are 1. The variance-covariance matrix for X can be written as

$$V(X) = \Sigma_x = \begin{bmatrix} p_{x1}(1-p_{x1}) & p_{x12} - p_{x1}p_{x2} \\ p_{x12} - p_{x1}p_{x2} & p_{x2}(1-p_{x2}) \end{bmatrix}. \qquad (3)$$

The variance-covariance matrix for Y can be written as

$$V(Y) = \Sigma_y = \begin{bmatrix} p_{y1}(1-p_{y1}) & p_{y12} - p_{y1}p_{y2} \\ p_{y12} - p_{y1}p_{y2} & p_{y2}(1-p_{y2}) \end{bmatrix}. \quad (4)$$

The sample statistics are given by

$$\hat{p}_{x1} = \frac{\sum_{i=1}^{n} x_{i1}}{n}, \ \hat{p}_{x2} = \frac{\sum_{i=1}^{n} x_{i2}}{n} \text{ and } \hat{p}_{y1} = \frac{\sum_{i=1}^{m} y_{i1}}{m}, \ \hat{p}_{y2} = \frac{\sum_{i=1}^{m} y_{i2}}{m},$$

$$\hat{\Sigma}_x = \begin{bmatrix} \hat{p}_{x1}(1-\hat{p}_{x1}) & \hat{p}_{x12} - \hat{p}_{x1}\hat{p}_{x2} \\ \hat{p}_{x12} - \hat{p}_{x1}\hat{p}_{x2} & \hat{p}_{x2}(1-\hat{p}_{x2}) \end{bmatrix}, \ \hat{\Sigma}_y = \begin{bmatrix} \hat{p}_{y1}(1-\hat{p}_{y1}) & \hat{p}_{y12} - \hat{p}_{y1}\hat{p}_{y2} \\ \hat{p}_{y12} - \hat{p}_{y1}\hat{p}_{y2} & \hat{p}_{y2}(1-\hat{p}_{y2}) \end{bmatrix} \quad (5)$$

where $\hat{p}_{x12} = \frac{\sum_{i=1}^{n} x_{i1}x_{i2}}{n}$ and $\hat{p}_{y12} = \frac{\sum_{i=1}^{m} y_{i1}y_{i2}}{m}$.

The unbiased estimates of the variance-covariance matrices for the two groups are $S_X = \frac{n}{n-1}\hat{\Sigma}_x$ and $S_Y = \frac{m}{m-1}\hat{\Sigma}_y$. Under the assumption that $\Sigma_x = \Sigma_y = \Sigma$, we use the usual pooled estimate S_{pooled} to estimate Σ and then apply the formula for T^2 defined in (2) to the dichotomous data.

A modification of this procedure is to use the estimated variance-covariance matrix defined by

$$\hat{\Sigma} = \frac{(n-1)\hat{\Sigma}_x + (m-1)\hat{\Sigma}_y}{n+m-2}. \quad (6)$$

The test statistic for testing the hypothesis H_0: $P_x = P_y$ is given by

$$\hat{T}^2 = \frac{mn}{m+n}\left(\hat{P}_x - \hat{P}_y\right)' \hat{\Sigma}^{-1}\left(\hat{P}_x - \hat{P}_y\right), \quad (7)$$

where $\hat{P}_x = \left[\hat{p}_{x1} \ \hat{p}_{x2}\right]$ and $\hat{P}_y = \left[\hat{p}_{Y1} \ \hat{p}_{Y2}\right]$. As with the use of T^2 we assume that the distribution of $F = \frac{m+n-2-1}{(n+m-2)2}\hat{T}^2$ approximately follows an F-distribution with numerator degrees of freedom 2 and denominator degrees of freedom $n + m - 3$. This test statistic and the test statistic in (2) applied directly to dichotomous data are evaluated with respect to probability of type 1 error rate.

5. Simulation

In order to study the performance of these tests, random samples were generated from bivariate Bernoulli distributions. The two dichotomous variables have to be correlated for bivariate analysis to be relevant. We used SAS® for our simulation study. We generated two Bernoulli distributions for the two variables. The first variable X_1 was generated from Bernoulli(p_1). The second variable X_2 was generated conditional on X_1 from a Bernoulli(p_2) where p_2 is obtained as follows:

$$P(X_2=1 \mid X_1=0) = p_{10}\,,\ P(X_2=1 \mid X_1=1) = p_{11}$$

$$\begin{aligned} P(X_2=1) &= P(X_2=1 \mid X_1=0)P(X_1=0) + P(X_2=1 \mid X_1=1)\,P(X_1=1) \\ &= p_{10}\,(1-p_1) + p_{11}p_1 \\ &= p_{2.} \end{aligned}$$

The conditional probabilities p_{10} and p_{11} were chosen so that X_1 and X_2 have a specified correlation ρ. The same procedure was repeated for the other population. The variance-covariance matrix was the same for both cases under H_0.

Random samples of size n and m were drawn from population1 and population2 respectively for each correlation value. For this study n and m were equal. The different values of n and m were 10, 20, 30, and 40. The two statistics given in (2) and (7) were computed. Each combination of p_1, p_2, n, m, and ρ was repeated 5000 times and the test statistics were computed at each repetition. The type 1 error rate is taken to be the relative frequency with which the test statistics given by (2) and (7) exceeded the critical value in 5000 replications. The critical value is computed at 5% significance level.

6. Conclusions

Probability of Type1 errors is given in Table 1. The correlation coefficient ρ ranges from .25 to .90. The sample size varies from 10 to 40. There are two values of type1 error rates- one using the test statistic given in (2) and the other by using the test statistic given in (7).

The statistic T^2 defined in (2) does a good job of controlling the Type 1 error rate for the cases considered. The elements of the variance-covariance matrix S_{pooled} in (2) are slightly larger than those of $\hat{\Sigma}$ for (7). Because the statistics involve the inverse of the variance-covariance matrix, it follows that T^2 is smaller than $\hat{T}^2$, and so the Type 1 error rate of $\hat{T}^2$ will be higher as is evident from the table. However, the difference is not particularly large for samples of larger size.

We thought that the correlation might be a determining factor in terms of Type 1 error rate being closer to its nominal value. The simulation results indicate that it is not so for the cases considered. Table1 shows that the change in correlation coefficient ρ from .25 to .90 does not affect the probability of Type 1 error for either of the test statistics. The marginal probabilities p_1 and p_2 range from .4 to .9. Thus, in terms of controlling Type I error, the use of Hotelling's T^2

appears to be an acceptable methodology for analyzing bivariate, and by logical extension, multivariate dichotomous data.

7. Future Investigation

The proposed test for bivariate dichotomous data was studied in terms of controlling the Type I error rate. The power of this test deserves further investigation. We are also studying the power of this test when the variance-covariance matrices are not pooled.
With respect to the confidence interval, we can make a confidence ellipsoid as in the usual multivariate case. Confidence region for difference of proportions would be an ellipsoid centered at the observed proportion difference, whose axes are determined by the eigenvalues and eigenvectors of S_{pooled}. However, further investigation of its properties is needed.

Table 1

p_1	p_{11}	p_{10}	p_2	ρ	n	m	T^2	$\hat{T}^2$
.60	.95	.05	.59	.90	10	10	.050	.071
					20	20	.054	.062
					30	30	.056	.061
					40	40	.056	.058
.50	.90	.05	.475	.85	10	10	.048	.067
					20	20	.050	.058
					30	30	.047	.051
					40	40	.052	.054
.60	.90	.15	.60	.70	10	10	.044	.065
					20	20	.048	.054
					30	30	.056	.060
					40	40	.052	.056
.90	.90	.45	.90	.50	10	10	.044	.066
					20	20	.056	.063
					30	30	.057	.057
					40	40	.048	.050
.40	.75	.50	.60	.25	10	10	.044	.067
					20	20	.048	.058
					30	30	.057	.063
					40	40	.056	.059

Summary

In this paper, we consider the problem of comparing two treatments in experiments in which bivariate dichotomous response variables are measured on each experimental unit. A

multivariate permutation test for this problem may be carried out using PROC MULTTEST in SAS®. The question of interest here is the possible use of multivariate normal methods to analyze bivariate dichotomous data. An approximate Hotelling T^2 test is proposed for bivariate dichotomous data and empirically evaluated in terms of Type I error rate. It is shown that the approximation does a good job of controlling the Type I error rate for a range of bivariate parameters even for relatively small sample sizes. Thus, in terms of controlling Type I error, the use of Hotelling's T^2 appears to be an acceptable methodology for analyzing dichotomous data.

Acknowledgements

The authors wish to thank the referee(s) for their valuable comments and suggestions in improving the manuscript.

References

Agresti, A. and Liu, I.-M. (1999) *Modeling a Categorical Variable Allowing Arbitrary Many Category Choices*. Biometrics, 55, 935-943.

Anderson, T.W. (1984) *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, New York.

Bilder, C.R., Loughin, T.M., and Nettleton, D. (2000) *Multiple Marginal Independence Testing for pick any/c Variables*. Communications in Statistics: Simulation and Computation, 29(4), 1285-1316.

Blair, R.C., Higgins, J.J., Karniski, W., and Kromery, J.D. (1994) *A Study of Multivariate Permutation Tests which may repalce Hotelling's T-square Test in Prescribed Circumstances*. Multivariate Behavioral Research, 29, 141-163.

Cox, D.R. (1972) *The Analysis of Multivariate Binary Data*. Applied Statistics, 21, 113-120.

Goodman, L.A. and Kruskal, W.H. (1954) *Measure of Association for Cross Classifications*. Journal of the American Statistical Association, 49, 732-764.

Goodman, L.A. and Kruskal, W.H. (1959) *Measure of Association for Cross Classifications*. Journal of the American Statistical Association, 54, 123-163.

Goodman, L.A. and Kruskal, W.H. (1963) *Measure of Association for Cross Classifications*. Journal of the American Statistical Association, 58, 310-364.

Higgins, J.J. (2004) An *Introduction to Modern Nonparametric Statistics*. Thompson Brooks/Cole, Duxbury Advanced Series, Pacific Grove, CA.

Knocke, J.D. (1976) *Multiple Comparisons with Dichotomous Data*. Journal of the American Statistical Association, 71, 849-853.