

11-1-2016

# A Comprehensive Review of the Two-Sample Independent or Paired Binary Data, with or without Stratum Effects

Dewi Rahardja

*U.S. Department of Defense, rahardja@gmail.com*


Ying Yang

*U.S. Food and Drug Administration*

Zhiwei Zhang

*U.S. Food and Drug Administration*

Follow this and additional works at: <http://digitalcommons.wayne.edu/jmasm>

 Part of the [Applied Statistics Commons](#), [Social and Behavioral Sciences Commons](#), and the [Statistical Theory Commons](#)

## Recommended Citation

Rahardja, Dewi; Yang, Ying; and Zhang, Zhiwei (2016) "A Comprehensive Review of the Two-Sample Independent or Paired Binary Data, with or without Stratum Effects," *Journal of Modern Applied Statistical Methods*: Vol. 15 : Iss. 2 , Article 16.

DOI: 10.22237/jmasm/1478002440

Available at: <http://digitalcommons.wayne.edu/jmasm/vol15/iss2/16>

This Regular Article is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized editor of DigitalCommons@WayneState.

---

# A Comprehensive Review of the Two-Sample Independent or Paired Binary Data, with or without Stratum Effects

## **Cover Page Footnote**

Disclaimer Statement: This research represents the authors own work and opinion. It does not reflect any policy nor represent the official position of the U.S. Department of Defense nor any other U.S. Federal agency.

# A Comprehensive Review of the Two-Sample Independent or Paired Binary Data, with or without Stratum Effects

**Dewi Rahardja**

U.S. Department of Defense  
Fort Meade, Maryland

**Ying Yang**

U. S. Food and Drug Administration  
Silver Spring, Maryland

**Zhiwei Zhang**

U. S. Food and Drug Administration  
Silver Spring, Maryland

---

Various statistical hypotheses testing for discrete or categorical or binary data have been extensively discussed in the literature. A comprehensive review is given for the two-sample binary or categorical data testing methods on data with or without Stratum Effects. The review includes traditional methods such as Fisher's Exact, Pearson's Chi-Square, McNemar, Bowker, Stuart-Maxwell, Breslow-Day and, Cochran-Mantel-Haenszel, as well as newly developed ones. We also provide the roadmap, in a figure or diagram format to which methods are available in the literature. In addition, the implementation of these methods in popular statistical software packages such as SAS and/or R is also presented. This will be helpful for researchers to determine which (categorical-data) testing method is available to use in various fields of study such as clinical trials, epidemiology, etc., both for the design phase of a study in prospective study, cross-sectional or retrospective study analysis.

*Keywords:* Cochran-Mantel-Haenszel (CMH) test, common odds ratio (OR), common risk difference (CRD), homogeneous stratum effect (HSE), McNemar's test, paired binary data, stratified data, Bowker's test, marginal homogeneity, stratified test, Stuart-Maxwell's test, symmetry, Fisher's exact test, chi-squared test

---

## Introduction

Many real-world data, such as data in clinical trials, financial data, epidemiology, sociology, etc. often use outcome variables that are categorical or binary in nature, that is, for example, in binary case, there are two possible outcomes for each subject. Without loss of generality (WLOG), these two outcomes are mutually exclusive and are categorized as success or failure. A frequent task in many fields of study, such as medical statistics (or any other field) is to compare two (independent or paired) binomial proportions. It can occur both in randomized

---

*Dr. Rahardja is a Statistician. Email her at: rahardja@gmail.com,  
Dewi.G.Rahardja.civ@mail.mil.*

controlled trials and in observational studies. The outcomes of two groups can be summarized in a single  $2 \times 2$  contingency table. The number of subjects in each group ( $n_{1+}$  and  $n_{2+}$ ) is assumed to be fixed by the design. Assume that the subjects in group 1 have probability of success equal to  $p_1$ , and that the subjects in group 2 have probability of success equal to  $p_2$ . The following Table 1 illustrates a single  $2 \times 2$  contingency table.

**Table 1.** Comparing 2 groups of binomial data in a *single*  $2 \times 2$  contingency-table format.

	Success	Failure	
Group 1	$n_{11}$	$n_{12}$	$n_{1+}$
Group 2	$n_{21}$	$n_{22}$	$n_{2+}$

Let  $n = \{n_{11}, n_{12}, n_{21}, n_{22}\}$  be the observed values as in Table 1. The number of successes in group 1 is binomially distributed with parameters  $n_{1+}$  and  $p_1$ . In a similar manner, the number of successes in group 2 is binomially distributed with parameters  $n_{2+}$  and  $p_2$ . The parameters  $p_1$  and  $p_2$  are estimated by the sample proportions

$$\hat{p}_1 = \frac{n_{11}}{n_{1+}} \quad \text{and} \quad \hat{p}_2 = \frac{n_{21}}{n_{2+}},$$

which are the maximum likelihood estimates.

The followings are the three most common measures to compare between two groups in a study. They are, the proportion difference, proportion (risk) ratio, and the odds ratio:

Parameter:	Notation
Difference:	$\hat{p}_1 - \hat{p}_2$
Risk ratio:	$OR = \hat{p}_1 / \hat{p}_2$
Odds Ratio:	$\frac{\hat{p}_1 / (1 - \hat{p}_1)}{\hat{p}_2 / (1 - \hat{p}_2)}$

The two groups being considered can be classified either as independent or matched pairs. Independent groups mean that the two samples taken are

independent, that is, sample values selected from one population are not related in any way to sample values selected from the other population. Matched pairs consist of two samples that are dependent or paired outcomes. The two variables may be two responses on a single individual or two responses from a matched pair (as in matched case-control studies). Table 2 summarizes the outcomes of matched pair two groups in a  $2 \times 2$  contingency-table format.

**Table 2.** Matched pair two groups in a  $2 \times 2$  contingency-table format.

Control	Case		Total
	Success	Failure	
Success	$n_{11} (p_{11})$	$n_{12} (p_{12})$	$n_{1+} (p_{1+})$
Failure	$n_{21} (p_{21})$	$n_{22} (p_{22})$	$n_{2+} (p_{2+})$
Total	$n_{+1} (p_{+1})$	$n_{+2} (p_{+2})$	

where  $p_{+1}$  and  $p_{1+}$  are the marginal probabilities of a success response for the case and control subjects, respectively.

In a stratified design (or multiple  $2 \times 2$  contingency tables), the subjects are selected from two or more strata which are formed from important covariates such as gender, income level, marital status, etc. The number of subjects in each of the two groups in each stratum is set (fixed) by the design. A separate  $2 \times 2$  table is formed for each stratum. Hence, there are multiple  $2 \times 2$  contingency tables. The data can be represented as a set of  $K$   $2 \times 2$  tables as the following Table 3.

**Table 3.** Comparing 2 groups of binomial data in a *multiple*  $2 \times 2$  contingency-table format.

	Success	Failure	
Group 1	$n_{11k}$	$n_{12k}$	$n_{1+k}$
Group 2	$n_{21k}$	$n_{22k}$	$n_{2+k}$

where  $k = 1, \dots, K$  stratum.

Thus, the purpose of this review of is to consider the existing testing methods in the literatures on the two independent or matched pair samples with binary data with or without stratum effects.

## Hypothesis Testing

Consider two independent groups without stratum effect (i.e., a single  $2 \times 2$  contingency table). The hypotheses for two independent proportions can be written as  $H_0: p_1 = p_2$  and  $H_1: p_1 \neq p_2$ . A Chi-square test is often used test the hypotheses.

In SAS PROC FREQ, the CHISQ option is used in the TABLES statement to obtain the test statistic and its associated  $p$ -value. By the famous rule of thumb, the Cochran's rule, the Chi-square test assumes that the expected value for each cell is five or higher. However, if this assumption is not met, the Fisher's exact test can be used regardless of how small the expected frequency is. The Fisher's exact test can be used with the FISHER option on the TABLES statement. However, Fisher's exact test is computationally explosive for large sample size and hence the Chi-square test is needed for large sample size approximation.

When subjects from two groups are independently sampled from two or more strata (i.e., with stratum effect; or a multiple  $2 \times 2$  contingency table), the null hypothesis of the interest can be to test whether odds ratios are the same across strata, that is,  $H_0: OR_1 = OR_2 = \dots = OR_k$  (or, homogeneity across strata). The Breslow-Day (BD) test (1980) for homogeneous odds ratios across strata can be used to test for the stratum effect. If the BD test is rejected, then the treatment comparison should be performed by strata; otherwise, the Cochran Mantel Haenzel (CMH) test (Cochran, 1954; Mantel & Haenszel, 1959) can be used to test whether the common odds ratios across strata is equal to 1, i.e., if all the  $OR_i = 1$ , for  $i = 1, \dots, k$ . In SAS PROC FREQ, the CMH option can be used for testing whether the common odds ratios are equal to one. The CMH option also provides logit estimates of the common odds ratio and the common relative risks.

Next, consider binary data collected on matched pairs. The sampling unit is not one individual but a pair of related individuals, which could be two parts of or two occasions for the same individual. For example, the binary response is a voter's choice from two presidential candidates and the two occasions could be two different time points before the presidential election.

For unstratified paired binary data, McNemar's test (1947) is commonly used to test whether the risk difference is zero. Such a null hypothesis is more commonly known as marginal homogeneity or symmetry of the  $2 \times 2$  contingency table. This null hypothesis of homogeneity can be written as  $H_0: p_{1.} = p_{.1}$ , where  $p_{1.} = p_{11} + p_{12}$  and  $p_{.1} = p_{11} + p_{21}$ , or equivalently, the null hypothesis of symmetry,  $H_0: p_{12} = p_{21}$ . McNemar's test can be calculated using AGREE option in PROC FREQ. Developed from asymptotic theory, McNemar's test requires a large

number of observations (say 5, by Cochran's rule) in each cell of discordance. For small samples, an exact binomial test can be used to test the null hypothesis of symmetry.

When the paired categorical random variables take  $K$  ( $K > 2$ ) values, Bowker's test (1948) can be used to test the symmetry  $H_0: p_{ij} = p_{ji}$ , for all  $i \neq j$ , where  $i, j \in \{1, 2, \dots, K\}$ .

If the test of marginal homogeneity is of the interest, the generalization of the McNemar's test, commonly referred to as generalized McNemar's or Stuart-Maxwell test (1955) can be used,  $H_0: p_{i.} = p_{.i}$ , where  $i = 1, 2, \dots, K$ . Note that for  $K > 2$ , the null hypothesis of symmetry is not equivalent to the null hypothesis of homogeneity. In fact, rejection of marginal homogeneity implies rejection of symmetry, but not vice versa. Therefore, practitioners need to decide which hypothesis to test for a particular application. A SAS macro by Sun and Yang (2008) has been developed for Stuart-Maxwell statistic.

Zhao, Rahardja, Wang and Shen (2014) considered a series of independent paired binary data in which the series is defined by a stratification factor, the null hypothesis of interest is to test the homogeneous stratum effects. In analogy, this is similar to the Breslow-Day (BD) test (1980) for homogeneous odds ratios across a series of stratified  $2 \times 2$  contingency tables in which the binary data are unpaired. The null hypothesis can be written as  $H_0: p_{1.1} - p_{.11} = \dots = p_{1.K} - p_{.1K}$ , or equivalently,  $H_0: p_{121} - p_{211} = \dots = p_{12K} - p_{21K}$ . The R-code of testing HSE in stratified paired binary data is available in the Zhao et al (2014) manuscript. If the homogeneous stratum effect (HSE) test is rejected, then the data should be analyzed by strata; otherwise the common risk difference (CRD) test for paired binary data in Zhao-Rahardja (2013) manuscript can be used to estimate the CRD. The test for CRD is analogous to the CMH test when the binary data are unpaired. Table 4 summarizes the above discussion.

## Roadmap

WLOG, the figure/diagram below (see Figure 1) provides the roadmap for practitioners to choose a suitable testing method for their categorical data analysis. In the Figure 1, the roadmap is provided by whether or not stratification table or multiple contingency tables is necessary.

COMPREHENSIVE REVIEW OF CATEGORICAL-DATA TESTING METHODS

**Table 4.** Listing of Sample Type with the appropriate testing, test statistics, and SAS command or R code.

Sample Type	Null Hypothesis ( $H_0$ )	Test statistics	SAS command or other option
Independent samples:			
Single $2 \times 2$ or Single $K \times K$ table	$H_0: p_1 = p_2$	Fisher's exact test (small sample)	<b>PROC FREQ</b> using <b>/Fisher</b> option
	$H_0: p_1 = p_2$	Chi-square test (large sample)	<b>PROC FREQ</b> using <b>/Chisq</b> option
Stratified independent samples:			
Multiple $2 \times 2$ tables	$H_0: OR_1 = OR_2 = \dots = OR_k$	Breslow-Day test for testing common odds ratio (OR) across strata  Cochran-Mantel- Haenszel (CMH) for estimating <i>Common</i> OR	<b>PROC FREQ</b> using <b>/CMH</b> option
Dependent/matched pairs:			
Single $2 \times 2$ table	$H_0: p_{+1} = p_{1+}$	McNemar's test	<b>PROC FREQ</b> using <b>/Agree</b> option
Single $K \times K$ table	$H_0: p_{ij} = p_{ji}$	Bowker test for symmetry	SAS macro by Sun and Yang (2008)
	$H_0: p_{i.} = p_{.i}$	Stuart-Maxwell test for marginal homogeneity	
Stratified dependent/matched pairs:			
Multiple $2 \times 2$ tables	$H_0: p_{1.1} - p_{.11} = \dots = p_{1.K} - p_{.1K}$	Homogenous stratum effect (HSE) test for homogeneity	<b>R-code</b> in Zhao <i>et</i> <i>al.</i> (2014)
	$H_0: p_{1.1} - p_{.11} = \dots = p_{1.K} - p_{.1K} = 0$	Common risk difference(CRD) test for estimating common risk difference	



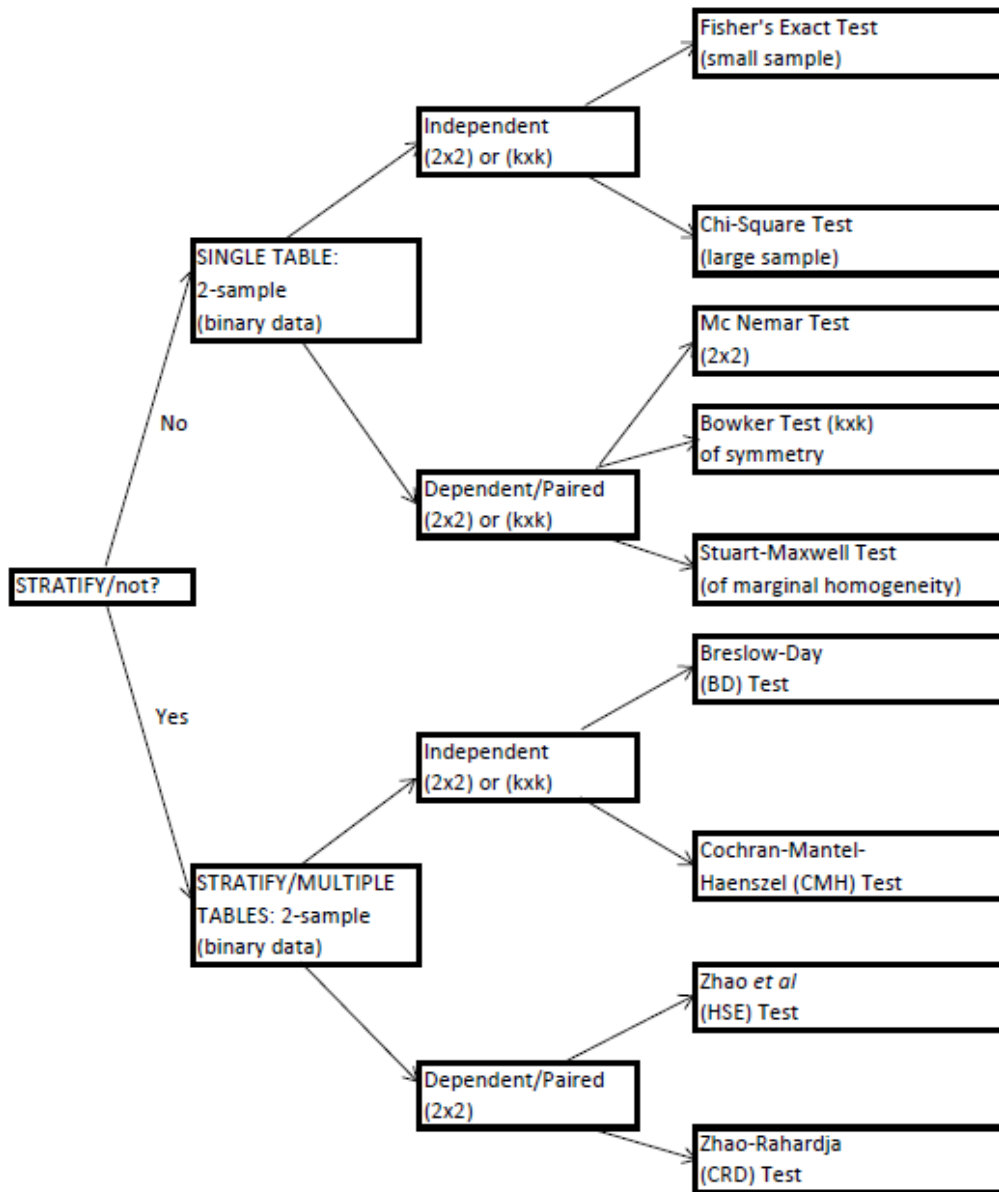


Figure 1. Categorical-Data Roadmap by Stratify or Not

## Summary

Categorical Data or most commonly binary or dichotomous outcome (i.e., success vs. failure, dead vs. alive, 1 vs. 0) is very common in real-data applications such as clinical trials, financial data, epidemiology, sociology, etc. The analysis of such categorical outcomes has a long history, beginning with the single  $2 \times 2$  table, multiple/stratified  $2 \times 2$  tables, matched/paired  $2 \times 2$  tables, to big table such as  $K \times K$  tables. In this paper, we provide a comprehensive review of the hypothesis testing procedures that are available in the literature for various types of categorical data. In summary, this review will be helpful for the practitioners in various fields of study (such as clinical trials, financial data, epidemiology, sociology, etc.) to determine the appropriate method according to the provided roadmap in Figure 1.

## Disclaimer

This research represents the authors own work and opinion. It does not reflect any policy nor represent the official position of the U.S. Department of Defense nor any other U.S. Federal agency.

## References

- Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association*, 43(244), 572-574. doi: 10.2307/2280710
- Breslow, N. E., & Day, N. E. (1980). *Statistical methods in cancer research: The analysis of case-control studies*. Lyon, France: International Agency for Research on Cancer.
- Cochran, W. G. (1954). Some methods for strengthening the common  $\chi^2$  tests. *Biometrics*, 10(4), 417-451. doi: 10.2307/3001616
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153-157. doi: 10.1007/BF02295996
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4), 719-748.

Maxwell, A. E. (1970). Comparing the classification of subjects by two independent judges. *British Journal of Psychiatry*, 116(535), 651-655.

doi: 10.1192/bjp.116.535.651

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, 42(3/4), 412-416. doi: 10.2307/2333387

Sun, X., & Yang, Z. (2008). *Generalized McNemar's Test for Homogeneity of the Marginal Distributions* (Paper 382-2008). SAS Global Forum 2008, Statistics and Data Analysis, San Antonio, Texas.

Zhao, Y. D., & Rahardja, D. (2013). Estimation of the common risk difference in stratified paired binary data with homogeneous stratum effect. *Journal of Biopharmaceutical Statistics*, 23(4), 848-855.

doi: 10.1080/10543406.2013.789888

Zhao, Y. D., Rahardja, D., Wang, D.-H., & Shen, H. (2014). Testing homogeneity of stratum effects in stratified paired binary data. *Journal of Biopharmaceutical Statistics*, 24(3), 600-607.

doi: 10.1080/10543406.2014.888440