Journal of Modern Applied Statistical Methods

Volume 16 | Issue 1 Article 2

5-1-2017

A Reinterpretation and Extension of McNemar's Test

Chauncey M. Dayton
University of Maryland, College Park and BDS Data Analytics, LLC, cdayton@umd.edu

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 20-33.
doi: 10.22237/jmasm/1493596860

ISSN 1538 − 9472

A Reinterpretation and Extension of McNemar's Test

Chauncey M. Dayton
University of Maryland, College Park
College Park, MD

The McNemar test is extended to multiple groups based on a latent class model incorporating classes representing consistent responders and a single latent error rate. The method is illustrated with data from a CDC survey of immunizations for flu and pneumonia for which a part-heterogeneous model is selected for interpretation.

Keywords: McNemar test, latent class analysis, marginal homogeneity, response error

Introduction

The McNemar chi-square test is the procedure of choice in studies assessing marginal homogeneity for repeated dichotomous classifications. Typical applications involve two independent raters or assays providing dichotomous judgments for the same set of stimuli, or a panel of independent judges responding on two occasions to the same dichotomous variable. The research question is whether or not it is reasonable to describe the two marginal classification rates for, say, a positive classification as equivalent (i.e., homogeneous). The chi-square significance test for this case is attributed to McNemar (1947) and the generalization to square tables larger than 2×2 is often referred to as the Stuart-Maxwell test (Stuart, 1955; Maxwell, 1970). Although alternatives to the McNemar test have been proposed, the original procedure performs well in comparative simulations as shown by Fagerland, Lydersen, and Laake (2013). Also, methods for performing multiple comparisons involving several sets of 2×2 tables have been presented by Westfall, Troendle and Pennello (2010).

Dr. Dayton is Professor Emeritus of Measurement, Statistics and Evaluation, as well as principal researcher with BDS Data Analytics. Email him at: cdayton@umd.edu.

For dichotomous variables, *A* and *B*, let π_{ij} represent the theoretic proportion for level *i* of variable *A* and level *j* of variable *B* (Table 1). Marginal homogeneity implies that $\pi_{1.} = \pi_{.1}$ or

Table 1. Theoretic Proportions for 2 × 2 Table

	B+	B−	Row
A+	π_{11}	π_{12}	$\pi_{1.}$
A−	π_{21}	π_{22}	$\pi_{2.}$
Column	$\pi_{.1}$	$\pi_{.2}$	

equivalently, that $\pi_{2.} = \pi_{.2}$. Assuming a sample of *N* cases and observed frequencies, n_{ij}, this implies symmetry because $\pi_{1.} = N(\pi_{11} + \pi_{12})$ and $\pi_{.1} = N(\pi_{11} + \pi_{21})$ so that π_{12} must be equal to π_{21}. Note, however, that marginal homogeneity does not imply symmetry for tables larger than 3 × 3.

The test for symmetry and, per force, the test for marginal homogeneity, reduces to a two-celled goodness-of-fit test based on the observed frequencies n_{12} and n_{21} with the null hypothesis $\pi_{12} = \pi_{21}$, or equivalently, $\pi_{12} = \pi_{21} = .5$. Note that the expected frequencies are both equal to $(n_{12} + n_{21})/2$. In terms of observed frequencies, the McNemar statistic in the form of a Pearson chi-square, with one degree of freedom, can be written as: $\chi^2 = \dfrac{(n_{12} - n_{21})^2}{(n_{12} + n_{21})}$.

An asymptotically equivalent test statistic can be based on a likelihood-ratio chi-square of the form $L^2 = -2\left(n_{12} Ln \dfrac{2n_{12}}{n_{12} + n_{21}} + n_{21} Ln \dfrac{2n_{21}}{n_{12} + n_{21}} \right)$. Often a correction for continuity is applied to the Pearson chi-square statistic to improve accuracy (Fleiss, 1981) and there are recent modifications such as mid-p computations (Fagerland, Lydersen & Laake, 2013). Agresti and Klingenburg (2005), and Klingenberg and Agresti (2006), have presented multivariate extensions of the McNemar test. Also, Durkalski, Palesch, Lipsitz, and Rust (2003) have introduced adaptations to account for clustering of observations.

The focus in the current study is on the issue of stratified homogeneity. Stratified homogeneity implies that marginal homogeneity for variables *A* and *B*, say, holds across the levels of a third variable (e.g., time, strata or groups). Feuer and Kessler (1989) considered a two-sample case, but the approach considered here is more general and based on latent variable modeling. Although stratified procedures can be conceptualized in log-linear terms (Bishop, Fienberg, &

Holland, 1975), the present approach exploits a result from Dayton and Macready (1983) who showed that the model underlying the McNemar test is equivalent to a restricted two-class latent class model for a 2×2 contingency table.

Latent Class Analysis

The mathematical model for latent class analysis (LCA) can be conceptualized as follows. Let $Y_s = \{y_{sj}\}$ be the vector-valued response for observed variables $j = 1, \ldots, J$, for the s^{th} respondent. Let the response options for the variables be defined over a set of distinct, mutually-exclusive values $r = 1, \ldots, R_j$ for the j^{th} variable (e.g., for dichotomous responses these values would be $r = (1,2)$). Then, for C distinct latent classes, an unrestricted latent class model is defined as:

$$P(Y_s) = \sum_{c=1}^{C} \theta_c \prod_{j=1}^{J} \prod_{r=1}^{R_j} \alpha_{cjr}^{\delta_{sjr}} \ .$$

The latent class (mixing) proportions are θ_c, $c = 1, \ldots, C$, with the restriction that these non-negative proportions sum to one. The latent class proportions represent the sizes of the unobserved latent classes. The α_{cjr} are conditional probabilities associated with the observed variables. That is, they represent the probability of response r to variable j given membership in the c^{th} latent class. Thus, for each variable, there is a vector of R_j conditional probabilities and these conditional probabilities sum to one for each variable within each latent class.

The δ_{sjr} terms are introduced in the manner of Kronecker deltas to include the appropriate conditional probabilities in the model based on the observed responses for the s^{th} respondent. Thus, $\delta_{sjr} = 1$ if $y_{sj} = r$ but $\delta_{sjr} = 0$ otherwise. In effect, the latent class model is based on the assumption that, conditional on latent class membership, the responses to the variables are independent. To make the model explicit, consider three dichotomously-scored variables and two latent classes. Within latent class 1, the probabilities for a 1 response (e.g., positive, yes or agree) are α_{111}, α_{121}, and α_{131} and within latent class 2 these probabilities are α_{211}, α_{221}, and α_{231}. The observed response {1,2,1}, for example, has conditional probability $\alpha_{111}\,(1 - \alpha_{121})\,\alpha_{131}$ within latent class 1 and conditional probability $\alpha_{211}\,(1 - \alpha_{221})\,\alpha_{231}$ within latent class 2, so that the unconditional probability for this response is $\theta_1\alpha_{111}\,(1 - \alpha_{121})\,\alpha_{131} + (1 - \theta_1)\,\alpha_{211}\,(1 - \alpha_{221})\,\alpha_{231}$. From a psychological measurement perspective, each conditional probability can be viewed as an item difficulty (or easiness) that may vary across the unobserved latent classes.

The log-likelihood for a latent class model with observations, $Y_s = \{y_{sj}\}$, is $\Im = \sum_s LnP(Y_s)$. To generate maximum-likelihood estimates (MLEs) for the parameters in the model, a set of normal equations must be solved simultaneously: $\frac{d\Im}{d\theta_c} = 0$ for each latent class proportion and $\frac{d\Im}{d\alpha_{cjr}} = 0$ for each conditional probability. However, a specific model will involve restrictions that must be introduced into the solution for the estimates. For example, the latent class proportions must sum to 1 across the classes and the conditional probabilities may be constrained in various ways including, at least, summing to 1 across the response options. Unfortunately, the presence of additive terms within the logarithmic operator means that the model is non-linear in the parameters and, except for special cases, cannot be solved by algebraic approaches.

However, given suitable restrictions, maximum-likelihood estimation is usually possible using iterative procedures such as Newton-Raphson algorithms as in Haberman's program LAT (1979) or by estimation-maximization (EM) algorithms as in Vermunt's program LEM (1997). These procedures are *regula falsi* methods that are subject to various computing complications including local maxima, boundary conditions, etc. (Dayton, 1999). Based on the MLE's, model fit can be assessed by Pearson or likelihood-ratio chi-square statistics computed from the cross-tabulation of the observed responses (e.g., the 2^J table for J dichotomous variables). In general, the degrees of freedom for these tests are $\#Cells - 1 - \#Pars$ where $\#Pars$ is the number of independent parameters estimated by MLE. However, it is possible that the parameters in a latent class model are not identified even though there are positive degrees of freedom. Programs such as LEM (Vermunt, 1997) provide some useful information on model identification although this can be a complex issue. These methods, as well as related descriptive approaches to assessing model fit, are summarized in Dayton (1999).

Two Repeated Dichotomous Classifications

The McNemar test is based on a 2×2 table with observed cell frequencies n_{ij} and cell proportions $p_{ij} = n_{ij} / N$ where N is the total sample size. Assuming an unrestricted two-class latent class model, the expected cell proportions are:

$$E(p_{11}) = \theta_1\alpha_{111}\alpha_{121} + \theta_2\alpha_{211}\alpha_{221}$$
$$E(p_{12}) = \theta_1\alpha_{111}\alpha_{122} + \theta_2\alpha_{211}\alpha_{222}$$
$$E(p_{21}) = \theta_1\alpha_{112}\alpha_{121} + \theta_2\alpha_{212}\alpha_{221}$$
$$E(p_{22}) = \theta_1\alpha_{112}\alpha_{122} + \theta_2\alpha_{212}\alpha_{222}$$

Given the usual restrictions on probabilities, there are five independent parameters, θ_1, α_{111}, α_{121}, α_{211}, and α_{221}, but only three independent observed proportions, p_{11}, p_{12}, and p_{21}. Therefore, the model cannot be identified unless at least two more restrictions are imposed. Imposing two restrictions would not yield positive degrees of freedom for assessing fit, so, in order to assess fit of the model, a total of three additional restrictions is required. The first two restrictions can be: $\alpha_{111} = \alpha_{121} \equiv \alpha_{11}$ and $\alpha_{211} = \alpha_{221} \equiv \alpha_{21}$; i.e., equating conditional probabilities across the two variables. If we interpret the first class as favoring a "1" response and the second class as favoring a "2" response, then a third restriction of the form $1 - \alpha_{11} = \alpha_{21} \equiv \alpha_e$ allows a single conditional probability, α_e, to be viewed as a response error. It should be noted that Proctor (1970) suggested the use of a restricted latent class model that involved response errors for the analysis of Guttman scales and that his approach was expanded by Dayton and Macready (1976). Given these restrictions, the equations above reduce to:

$$E(p_{11}) = \theta_1(1-\alpha_e)^2 + (1-\theta_1)\alpha_e^2$$
$$E(p_{12}) = \theta_1(1-\alpha_e)\alpha_e + (1-\theta_1)\alpha_e(1-\alpha_e)$$
$$E(p_{21}) = \theta_1\alpha_e(1-\alpha_e) + (1-\theta_1)(1-\alpha_e)\alpha_e$$
$$E(p_{22}) = \theta_1\alpha_e^2 + (1-\theta_1)(1-\alpha_e)^2$$

The two latent classes can be interpreted as comprised of respondents who consistently use the response category 1 or, alternately, consistently use the response category 2. Inconsistent responses such as {1,2} or {2,1} are assumed to occur as a result of response errors that represent lack of consistency. Note responses such as {1,1} and {2,2} require that respondents either do not make a response error or that they make two response errors (e.g., a respondent in the latent class associated with a {1,1} response makes two response errors and responds {2,2}).

For this relatively simple model, the log-likelihood and normal equations can be set up and solved algebraically as shown in Dayton and Macready (1983).

However, an alternative approach is based on the realization that the expected and observed frequencies are equal for responses {1,1} and {2,2}; i.e., $p_{11} = E(p_{11}) = \theta_1(1-\alpha_e)^2 + (1-\theta_1)\alpha_e^2$ and $p_{22} = E(p_{22}) = \theta_1\alpha_e^2 + (1-\theta_1)(1-\alpha_e)^2$. Thus, algebraically solving these two equations for values of the parameters yields, per force, the maximum likelihood estimators:

$$\hat{\theta}_1 = \frac{p_{11} - \hat{\alpha}_e^2}{1 - 2\hat{\alpha}_e} \text{ and } \hat{\alpha}_e = .5 - \sqrt{.25 - (p_{12} + p_{21})/2}\,.$$

Note that $\hat{\alpha}_e$ is undefined for $p_{12} + p_{21} > .5$ so that it is necessary to reverse the coding for one of the variables if this occurs in practice. The restricted latent class model yields expected frequencies that are consistent with the McNemar test in the sense that $\hat{p}_{11} = p_{11}$, $\hat{p}_{22} = p_{22}$, and $\hat{p}_{12} = p_{21} = (p_{12} + p_{21})/2$. Also, the resulting chi-square value for model fit is exactly the same as the uncorrected McNemar chi-square statistic with one degree of freedom. Thus, the McNemar may be viewed as testing the null the hypothesis $\alpha_{11} = 1 - \alpha_{21}$ versus the alternative $\alpha_{11} \neq 1 - \alpha_{21}$.

This conceptualization of the McNemar test focuses on response consistency rather than marginal homogeneity although the implications for observed responses are the same. However, estimates for the latent class parameters provide a measure of the agreement between classifications that is not available in a conventional McNemar analysis. For example, consider the exemplary before/after treatment results in Table 2. Positive responses occur at a rate of 40.3% before treatment and at a rate of 47.6% after treatment. The 6.3% difference is significant based on an uncorrected McNemar chi-square value of 4.55 (p = .033). Our latent class model yields estimated parametric values of .423 for the latent class proportion, θ_1, and .074 for the error rate, α_e. The value .423, or 42.3%, is an estimate for the proportion of respondents who have positive responses at both the before and after occasions of observation. Note that the conventional McNamar procedure does not provide a comparable statistic. Also, the value .074, or 7.4%, is an estimated error rate that applies to both the positive/positive and negative/negative latent response groups. Once again, this a value that has no direct analog in a McNemar analysis (although roughly similar to the before/after relative change in this example).

Table 2. Exemplary Pre/Post Data

		After		
		Positive	***Negative***	***Total***
Before	***Positive***	59	6	65
	Negative	16	80	96
	Total	75	86	161

Stratified McNemar Test

Consider cross-tabulations similar to those in Table 1 for two or more strata within a population (or for the same population at different points in time or for samples from several populations). Letting the strata be represented by $y = 1, \ldots, Y$, the expected cell proportions for a given stratum can be written as:

$$
\begin{aligned}
E\left(p_{11y}\right) &= \theta_{1y}\left(1-\alpha_{ey}\right)^2 + \left(1-\theta_{1y}\right)\alpha_{ey}^2 \\
E\left(p_{12y}\right) &= \theta_{1y}\left(1-\alpha_{ey}\right)\alpha_{ey} + \left(1-\theta_{1y}\right)\alpha_{ey}\left(1-\alpha_{ey}\right) \\
E\left(p_{21y}\right) &= \theta_{1y}\alpha_{ey}\left(1-\alpha_{ey}\right) + \left(1-\theta_{1y}\right)\left(1-\alpha_{ey}\right)\alpha_{ey} \\
E\left(p_{22y}\right) &= \theta_{1y}\alpha_{ey}^2 + \left(1-\theta_{1y}\right)\left(1-\alpha_{ey}\right)^2
\end{aligned}
$$

Maximum likelihood estimation for the stratified model follows the same approach as for any latent class model in general but requires that suitable restrictions be imposed on the estimated parameters. In addition, issues related to identification of the model must be considered (Dayton, 1999). Because the strata are independent, it is apparent that jointly estimating the parameters in the heterogeneous form of the stratified model is the same as fitting the model separately to each stratum but does provide an overall measure of fit in the form of a chi-square statistic with Y degrees of freedom. However, the major advantage of conceptualizing the model in this form is that it allows for imposing across-strata restrictions on the error rates. The most highly restricted case results in a homogeneous model with $2Y - 1$ degrees of freedom that is based on restrictions of the form $\alpha_{ey} = \alpha_e \forall y$. However, a variety of part-heterogeneous models may be suggested by theory (or, the data) and tested accordingly. Closed-form estimates are not, in general, available for the stratified model. Fortunately, as illustrated below, available programs for latent class analysis allow for these restrictions and associated MLEs.

A similar conceptualization, known as the Hui-Walter model (Hui & Walter, 1980), has been presented in the context of repeated assays for the purpose of estimating false-positive and false-negative rates. This model is saturated so that fit to data cannot be assessed by ordinary procedures and is based on a different set of restrictions. Biemer (2011) presents an extended discussion with examples of the Hui-Walter model.

Application for Two Immunization Survey Items

The CDC Behavioral Risk Factor Surveillance System (BRFSS) is a large-scale telephone survey that tracks health risks in the United States. The CDC web-enabled analysis tool for BRFSS (http://nccd.cdc.gov/s_broker/WEATSQL.exe/weat/index.hsql) was used to produce cross-tabulations of responses to two items, referred to as Flu and Pneumonia, for adults aged 65 and older:

Flu:	Had a flu shoot within past 12 months.
Pneumonia:	Ever had a pneumonia vaccination.

The item responses were Yes/No and, for the year 2011, there were responses available for a total of 143,002 people across the United States. A large variety of demographic variables is included in the data system and, using CDC labeling, we chose to compare race/ethnicity groups divided into the strata: (1) White, Non-Hispanic; (2) Black, Non-Hispanic; (3) Hispanic; and (4) Other which comprised multiracial and other races. Cross-tabulated frequency data for the four race/ethnicity groups are presented in Table 3.

Table 3. Cross-Tabulation of Two Immunization Variables for Four Race/Ethnic Groups

Flu:	**Yes**	**Yes**	**No**	**No**		**McNemar**	
Pneumonia:	**Yes**	**No**	**Yes**	**No**	**Total**	**G^2**	**Prob.**
White, Non-Hispanic	64,446	12,729	23,792	21,279	122,246	3404.05	0.000
Black, Non-Hispanic	3,367	1,107	1,728	2,575	8,777	137.14	0.000
Hispanic	2,050	1,005	1,123	2,251	6,429	6.55	0.011
Other	2,641	679	1,105	1,125	5,550	102.71	0.000
Total	*72,504*	*15,520*	*27,748*	*27,230*	*143,002*	*3503.30*	*0.000*

Our focus was on the relative rates of flu and pneumonia immunizations across the race/ethnic groups. As shown in Table 4, the marginal immunization

rates are moderately different for three of the four race/ethnic groups but very similar for Hispanics (i.e., .48 and .49 for flu and pneumonia, respectively).

Table 4. Marginal Rates

Race/Ethnic Group	Flu	Pneumonia
White, Non-Hispanic	0.63	0.72
Black, Non-Hispanic	0.51	0.58
Hispanic	0.48	0.49
Other	0.60	0.67
Total	*0.62*	*0.70*

In Table 3, the column labeled McNemar G^2 presents McNemar likelihood-ratio chi-square fit statistics for each race/ethnic group as well as for the total sample. These tests are consistent with our observation concerning the marginal rates with only the Hispanic group failing to be significant beyond the .01 level.

Homogeneous, heterogeneous and part-heterogeneous stratified McNemar models were fit to the cross-tabulations of the two immunization items for the four race/ethnic groups. The homogeneous model posits a single response error rate, α_e, for the four strata whereas the heterogeneous model posits unique error rates, α_{e1}, α_{e2}, α_{e3}, and α_{e4}, for the four strata. In both cases, the size of the latent class, θ_1, corresponding to a Yes response to both items, {1, 1}, is allowed to vary by group in order to fix the marginal distributions for the race/ethnic groups. The part-heterogeneous model, which equated error rates for all groups except White, Non-Hispanic, was suggested by the fact that the error rates for these three strata were quite similar for the heterogeneous model (i.e., .206, .209 and .201, respectively). MLE parameter estimation and model fit were conducted using the latent variable program, LEM (Vermunt, 1997). Although lacking a modern computer interface, LEM has the dual advantages of being (a) available free for download for Microsoft operating systems and (b) extremely flexible in terms of the latent class models that can be estimated. Sample LEM program set-ups for the homogeneous and heterogeneous models are included in the Appendix. Model fit statistics and parameter estimates are presented in Table 5. Given the large sample size, it was not unexpected that all three models result in rejection of the hypothesis of equal error rates across the four race/ethnic groups.

Table 5. Stratified McNemar Models Fit to Vaccination Variables

Model	DF	Chi-Sq (G^2)	AIC	Homogenous Groups	Error Rates	Class Size
Homogeneous	7	3709.95*	513, 360.7	[1234]	.186	.78, .57, .47, .72
Part-Heterogeneous	6	3652.38*	513, 305.1	[1],[234]	.183, .204	.78, .58, .47, .73
Heterogeneous	4	3650.85*	513, 307.6	[1],[2],[3],[4]	.183, .206,.209, .201	.77, .58, .47, .73
Collapsed	1	3503.30*	N/A	[1]	.186	.75

Note: *All *p*-values are less than .001

Using the Akaike (1973) information measure as suggested by Dayton (1999) for comparing latent class models, a min(AIC) criterion indicates that the part-heterogeneous model is best among the models being compared. Because the three models are nested, it is appropriate to test differences among them using likelihood-ratio chi-square (G^2) statistics. These comparisons are:

Homogeneous vs. Part-Heterogeneous: $\Delta(G^2) = 57.57$, DF = 1, $p < .01$;
Homogeneous vs. Heterogeneous: $\Delta(G^2) = 59.10$, DF = 3, $p < .01$;
Part-Heterogeneous v.s Heterogeneous: $\Delta(G^2) = 1.53$, DF = 2, $p < .05$.

The Part-Heterogeneous model fits the data no worse than the Heterogeneous model, whereas both of these models provide better fit than the Homogeneous model.

As noted above, in order to fix the marginal distributions at observed values for the four race/ethic groups, it was necessary to posit separate latent class proportions for the strata. These proportions are quite consistent across the models that were evaluated with White, Non-Hispanic and Hispanic showing considerably larger latent class proportions than the other two groups. If race/ethnicity is ignored and a non-stratified latent class model is fitted to the (marginal) 2×2 table of immunization rates, a latent class proportion of .75 is estimated. An error rate of .186 was estimated for the homogeneous model which is essentially identical to that from the marginal 2×2 model although this is driven by the fact that about 85% of the total sample is comprised of White, Non-Hispanic respondents,

In order to allow for the observed lack of agreement in immunizations rates for flu and pneumonia vaccinations, the latent class models suggest a rate of inconsistencies (errors) of approximately 18% - 20%. That is, about one in five individuals in a latent class that represents consistently Yes (or consistently No) respondents would, in fact, fail to respond consistently. From Table 2 it is notable

that inconsistencies tend to be in the direction of failing to obtain a flu vaccination, which may suggest some educational strategy in this regard for the 65 and older age group.

Capitalizing on the fact that the McNemar test can be conceptualized as a restricted latent class model, we have defined homogeneous, heterogeneous and part-heterogeneous models with parameter estimates that have interpretations that could be of interest in applied research settings such as immunization patterns for the 65-and-over population. Furthermore, estimation and significance testing are available using widely available latent-class programs.

References

Agresti, A., & Klingenberg, B. (2005). Multivariate tests comparing binomial probabilities, with application to safety studies for drugs. *Applied Statistics, 54*(4), 691-706. doi: 10.1111/j.1467-9876.2005.05437.x

Akaike, H. (1973). Information theory and an extension of the maximum-likelihood principle. In B. N. Petrov and F. Csake (Eds.), *Second international symposium on information theory*. Akademiai Kiado: Budapest, 267-281.

Biemer, P. P. (2011). *Latent class analysis of survey error*. New Jersey: Wiley.

Bishop, Y. M. M., Fienberg, S. E. & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.

Dayton, C. M. (1999). *Latent class scaling analysis*. New York: Sage Publications. doi: 10.4135/9781412984720

Dayton, C. M. & Macready, G. B. (1976). A probabilistic model for the validation of behavioral hierarchies. *Psychometrika, 41*(2), 189-204. doi: 10.1007/bf02291838

Dayton, C. M. & Macready, G. B. (1983). Latent structure analysis of repeated classifications with dichotomous data. *British Journal of Mathematical & Statistical Psychology, 36*(2), 189-201. doi: 10.1111/j.2044-8317.1983.tb01124.x

Durkalski, V. L., Palesch, Y. Y., Lipsitz, S. R. & Rust, P.F. (2003). Analysis of clustered matched-pair data. *Statistics in Medicine, 22*(15), 2417-2428. doi: 10.1002/sim.1438

Fagerland, M. W., Lydersen, S. & Laake, P. (2013). The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology, 13*(1), 91. doi: 10.1186/1471-2288-13-91

Feuer, E. J. & Kessler, L. J. (1989). Test statistic and sample size for a two-sample McNemar test. *Biometrics, 45*(2), 629–636. doi: 10.2307/2531505

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. New York: Wiley.

Haberman, S. J. (1979). *Analysis of qualitative data, volume 2: New developments*. New York: Academic Press.

Hui, S. L. & Walter, S. D. (1980). Estimating the error rates of diagnostic tests. *Biometrics, 36*(1), 167-171. doi: 10.2307/2530508

Klingenberg, B. & Agresti, A. (2006). Multivariate extensions of McNemar's test. *Biometrics, 62*(3), 921-928. doi: 10.1111/j.1541-0420.2006.00525.x

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

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

Proctor, C. H. (1970). A probabilistic formulation and statistical analysis of Guttman scaling. *Psychometrika, 35*(1), 73-78. doi: 10.1007/bf02290594

Stuart, A. A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika, 42*(3-4), 412-416. doi: 10.1093/biomet/42.3-4.412

Vermunt, J. K. (1997). *LEM: A general program for the analysis of categorical data*. Department of Methodology & Statistics, Tilburg University.

Westfall, P. H., Troendle, J. F. & Pennello, G. (2010). Multiple McNemar Tests. *Biometrics, 66*(4), 1185-1191. doi: 10.1111/j.1541-0420.2010.01408.x

Appendix

LEM input file for Homogeneous model

```
* CDC Behavioral Risk Factor Surveillance System
* Elderly flu shot last 12 months
* Elderly pneumonia vaccination ever
* Four ethnic groups - white, black, Hispanic, other
* Stratified McNemar test
* Homogenous Model [1234]
lat 1
man 3
dim 2 4 2 2
lab X Y F P * X = latent variable; Y = Ethnic;
              F = Flu, P = Pneumonia
mod Y
X|Y
F|XY eq2
P|XY eq2
des [ 0 2 0 2 0 2 0 2  2 0 2 0 2 0 2 0
      0 2 0 2 0 2 0 2   2 0 2 0 2 0 2 0 ]
dat [64446 12729 23792 21279     3367 1107 1728 2575
      2050  1005  1123  2251     2641  679 1105 1125]
```

LEM input file for Heterogeneous model

```
* CDC Behavioral Risk Factor Surveillance System
* Elderly flu shot last 12 months
* Elderly pneumonia vaccination ever
* Four ethnic groups - white, black, Hispanic, other
* Stratified McNemar test
* Heterogeneous Model [1],[2],[3],[4]
lat 1
man 3
dim 2 4 2 2
lab X Y F P * X = latent variable; Y = Ethnic;
              F = Flu, P = Pneumonia
mod Y
      X|Y
```

```
       F|XY eq2
       P|XY eq2
des [ 0 2 0 4 0 6 0 8   2 0 4 0 6 0 8 0
      0 2 0 4 0 6 0 8   2 0 4 0 6 0 8 0 ]
dat [64446 12729 23792 21279     3367 1107 1728 2575
      2050  1005  1123  2251     2641  679 1105 1125]
```