TWO-SAMPLE MCNEMAR TESTS FOR COMPLEX SURVEY DATA

Jenny Thompson and Robin Fisher
Jenny Thompson, Bureau of the Census,
Washington, DC 20233

KEY WORDS: Current Population Survey, Parallel Survey, Nonparametric Statistics.

Introduction

The McNemar test (1947) has been generalized to a two-sample situation where the hypothesis of interest is that the marginal changes in each of two independent samples' 2 x 2 tables are equal. (Feuer and Kessler, 1989). The application presented was for a two-sample cohort analysis and assumed simple random sampling.

Further modifications of the test statistic are necessary for a complex survey data application of the McNemar test. First, because the data are not obtained through a simple random sample, a different estimate of the variance is required. Second, unless the survey has a longitudinal design, a separate link of individuals in two consecutive months' of data must be performed. In general, such a link will use a set of demographic variables and will include some false matches. This induces another variance component to the model, the error due to false matches.

We show two refinements of this test for complex survey data, which require different estimates of variance. We first provide some general background about the McNemar tests. We then describe our modifications, including some remarks on applications to several months' data. Finally, we present our applications of these tests to the Current Population Survey's Parallel Survey split panel study and to the Current Population Survey's CATI Phase-in Project.

General Test

A sample is randomly split into two independent representative samples (split panels). After a baseline measurement is taken in both panels, a new technique is administered in one panel, the treatment panel. The other panel serves as a control.

The responses are matched longitudinally after the second measurement is taken. A response can be +, - , or * (missing). Since this is matched data, the "**" cell will be empty.

This scenario is represented pictorially as

Treatment Panel
Month 2
Treatment

Month 1 / No Treatment	+	-	*	
+	x_{++}	x_{+-}	x_{+*}	$x_{+.}$
-	x_{-+}	x_{--}	x_{-*}	$x_{-.}$
*	x_{*+}	x_{*-}		$x_{*.}$
	$x_{.+}$	$x_{.-}$		n

Control Panel
Month 2
No Treatment

Month 1 / No Treatment	+	-	*	
+	x_{++}'	x_{+-}'	x_{+*}'	$x_{+.}'$
-	x_{-+}'	x_{--}'	x_{-*}'	$x_{-.}'$
*	x_{*+}'	x_{*-}'		$x_{*.}'$
	$x_{.+}'$	$x_{.-}'$		n′

where n is not necessarily equal to n′.

For each panel, define

$M_{(12)}$ as the set of cases which have month 1 and month 2 responses (matched cases). This set contains $n_{(12)} = (x_{++} + x_{+-} + x_{-+} + x_{--})$ elements;
$M_{(10)}$ as the set of cases which have month 1 responses, but no month 2 response. This set contains $n_{(10)} = (x_{+*} + x_{-*})$ elements;
$M_{(02)}$ as the set of cases which have month 2 responses, but no month 1 response. This set contains $n_{(02)} = (x_{*+} + x_{*-})$ elements.

First, consider the one-sample case. Traditionally, the one-sample McNemar test statistic is constructed from the $n_{(12)}$ and $n_{(12)}'$ matched responses. In the one-sample scenario, we test the hypothesis

H_0: $\text{Prob}(x_{+-}) = \text{Prob}(x_{-+})$
H_1: Not H_0

i.e., the hypothesis that the movement from one state to the other (+ to -, or - to +) is zero. We also refer to this movement as the flux.

The one-sample test can be a useful diagnostic in the two-sample situation. We examine the Control panel estimates to see if there is zero movement. Any significant movement in the Treatment panel can be measured as a deviation from zero flux or as a change in the probability of a "+."

The two-sample hypothesis is

H_0: $Prob(x_{+\cdot}) - Prob(x_{\cdot+}) = Prob(x'_{+\cdot}) - Prob(x'_{\cdot+})$
H_1: Not H_0

In other words, the difference in the probabilities of switching in the two directions is the same, regardless of the treatment, or equivalently, the difference in panel fluxes is zero.

Complex Survey Modifications

Modification One: Longitudinally Linked Data

This method is a straightforward application of the two-sample McNemar test, using longitudinally linked data from a complex survey. The domain for both months of data is given by $M_{(12)}$.

Note, in one panel,

$$\begin{aligned} Prob(x_{+\cdot}) - Prob(x_{\cdot+}) &= [(Prob(X_{++}) + Prob(X_{+-})) - \\ &\quad (Prob(X_{++}) + (Prob(X_{-+}))] \\ &= [Prob(x_{+\cdot})^{\circ} - (Prob(x_{\cdot+})^{\circ}] \\ &= p_2^{\circ} - p_1^{\circ} \end{aligned}$$

where p_2° is the marginal probability of a + response in month 2, given that the respondent responded both months;
and where p_1° is the marginal probability of a + response in month 1, given that the respondent responded both months.

The one-sample test statistic is

$$Z_1^{\circ} = \frac{p_2^{\circ} - p_1^{\circ}}{\sqrt{Var(p_2^{\circ} - p_1^{\circ})}},$$

$$\text{where } p_1^{\circ} = \frac{x_{++} + x_{+-}}{n_{(12)}},\ p_2^{\circ} = \frac{x_{++} + x_{-+}}{n_{(12)}}$$

Given two independent panels, the two-sample test statistic is
If the survey is designed to collect longitudinal data, then this modification is a natural extension of the method described by Feuer and Kessler. The extension is the use of weighted estimates and

$$Z^{\circ} = \frac{(p_2^{\circ} - p_1^{\circ}) - (p_2'^{\circ} - p_1'^{\circ})}{\sqrt{Var(p_2 - p_1) + Var(p_2'^{\circ} - p_1'^{\circ})}},\ \text{where} \qquad (1a)$$

$$p_1'^{\circ} = \frac{x'_{++} + x'_{+-}}{n'_{(12)}},\ p_2'^{\circ} = \frac{x'_{++} + x'_{-+}}{n'_{(12)}}$$

complex survey variances and covariances in place of simple random sample variances. For this type of survey design, an effective mechanism to link individuals from month to month is presumably in place. Often, however, this is not the case, and one data set must be physically linked to another. Consequently, the $n_{(12)}$ elements in the domain will contain some false matches, and some actual matches may be inadvertently excluded.

Modification Two: Unlinked Data

This method omits the longitudinal linkage step altogether, noting that the construction of the test statistic relies on estimates of marginal probabilities. Assume that under the null hypothesis, the expected value of $(Prob(x_{*\cdot}) - Prob(x_{\cdot*}))$ is zero. This is described for a simple random sampling application in Marascuilo et al (1988).

The domain for the first month of data is given by $M_{(12)} \cup M_{(10)}$ which contains $n_{(12)} + n_{(10)} = n_1$ elements. The domain for the second month of data is given by $M_{(12)} \cup M_{(02)}$ which contains $n_{(12)} + n_{(02)} = n_2$ elements.

The one-sample test statistic constructed from the unlinked data is given by

$$Z_1 = \frac{p_2 - p_1}{\sqrt{Var(p_2 - p_1)}},$$

$$\text{where } p_1 = \frac{x_{+\cdot}}{n_1},\ p_2 = \frac{x_{\cdot+}}{n_2}$$

Given two independent panels, the two-sample test statistic is

$$Z = \frac{(p_2 - p_1) - (p_2' - p_1')}{\sqrt{Var(p_2 - p_1) + Var(p_2' - p_1')}}, \qquad (1b)$$

$$\text{where } p_1' = \frac{x'_{+\cdot}}{n_1'},\ p_2' = \frac{x'_{\cdot+}}{n_2'}$$

As with the application described above, all estimates are weighted estimates, and variances are complex survey variances.

Linear Combinations

We can use our estimated covariance matrix to test linear combinations of $\underline{\lambda}_1$, $\underline{\lambda}_c$, or $\underline{\delta}$ over time, where $\underline{\lambda}_T = \underline{p}_2 - \underline{p}_1$, $\underline{\lambda}_C = \underline{p}'_2 - \underline{p}'_1$, and $\underline{\delta} = \underline{\lambda}_T - \underline{\lambda}_C$, and $\underline{p}_1$, $\underline{p}_2$, $\underline{p}'_1$, and $\underline{p}'_2$ are vectors containing the marginal probabilities for the time period.

Perhaps the most interesting (to our applications) of these tests is of the hypothesis H_0: $\underline{1}'\underline{\mu} = 0$,where $\underline{\mu}$ is the expected value of one of the vectors described above. Other general linear hypotheses of this form could be equally interesting.

Another interesting test is the "omnibus hypothesis," where we test H_0: $\underline{\mu} = \underline{0}$. The test statistics for this hypothesis are $\underline{\lambda}'_T \Sigma_{\lambda(T)}^{-1} \underline{\lambda}_T$, $\underline{\lambda}'_C \Sigma_{\lambda(C)}^{-1} \underline{\lambda}_C$, and $\underline{\delta}' \Sigma_{\delta}^{-1} \underline{\delta}$ each of which has an approximate chi-squared distribution with r degrees of freedom, where r is the dimension of the vector of interest.

Applications

1. Background

The official monthly civilian labor force estimates from January 1994 onward are based on data from a comprehensively redesigned Current Population Survey (CPS). The redesign included implementation of a new, fully computerized questionnaire and an increase in centralized computer-assisted telephone interviewing (CATI). To gauge the effect of the CPS redesign on published estimates, a Parallel Survey (PS) was conducted using the new questionnaire and data collection procedures from July 1992 through December 1993. Special studies were embedded in both the PS and the CPS during the same time period to provide data for testing hypotheses about the effects of the new methodological differences on labor force estimates: the PS split panel study and the CPS CATI Phase-in Project (a continuation of the study presented in Shoemaker, 1993).
Findings described in Shoemaker (1993) had shown that including centralized telephone interviews yielded a larger unemployment rate. The two-sample McNemar test appeared to be a good vehicle for examining this phenomenon. For both surveys, the initial (first and fifth interviews) are conducted by a personal visit, and the subsequent interviews are conducted by telephone whenever possible. Thus the initial interviews provide a baseline measurement of labor force status; the second and sixth interviews provide a "post-treatment" measurement of labor force status.

To create the panels for both studies, sample within selected sample areas was randomly divided into two representative panels. The treatment panel was designated as CATI eligible. Sample households in the panel were eligible for CATI interviewing after the initial interviews. To be interviewed by CATI, a respondent must have a telephone, speak English or Spanish, and agree to future telephone interviews. Not all households in this panel were interviewed by CATI. The other panel served as a control.

The monthly unemployment rate is the primary statistic published from CPS data. Our goal was to understand how including CATI interviews influenced the probability of changing labor force status, in this case from unemployed to not unemployed (or vice versa).

2. Estimates

Each month/panel estimate is an unbiased estimate: each weight is the product of the baseweight, the weighting control factor, and an adjustment factor for the probability of inclusion in a split panel.

Variances of level were computed with generalized variance functions (GVFs). For more details, see Fisher et al (1993). Robert Fay used his VPLX software to calculate replicate estimates of correlation between rotation groups for unemployed and for civilian labor force using 10/92 through 12/93 CPS data. We used these correlations for both sets of test statistics based on unlinked data, assuming that they would not differ by survey or by geography. We derived an expression for the within-panel correlation for civilian population by relating previously calculated autocorrelations (Fisher and McGuinness, 1993) and variance estimates to the individual rotation group estimates.

We used the unlinked data correlations as a poor approximation for the linked data. Regrettably, we did not have direct replicate estimates of linked data correlation, which we would intuitively expect to be higher than the unlinked. We were also unable to determine the same sort of unique relationship between our autocorrelations and our monthly estimates of variance for obtaining linked data correlations. The consequence of using unlinked data correlations to approximate linked data was artificially similar results for the two modifications' applications. Moreover, there were other unresolved problems with the variance estimator for the linked data. This led us to omit our tests based on linked data.

3. Results

Parallel Survey Split Panel Study

This section presents the formal results from the one and two sample McNemar tests using unlinked Parallel Survey (PS) split panel data. Although this data was collected monthly, small expected cell sizes in the control panel led us to omit several sets of adjacent months from this analysis. Table One provides summary statistics for the one-sample "monthly" tests for each panel which were based on unlinked data from the PS's split panels. Table Two provides summary statistics for the two-sample tests based on unlinked data.

The reported values of p_1, p_2, p_1', and p_2' are percentages of estimated unemployed to estimated total population for the panel. Here, p_1 and p_1' are the panel ratio of estimated unemployed from the first and fifth interviews to the estimated panel population from the first and fifth interviews; p_2 and p_2' are the panel ratio of estimated unemployed from the second and sixth interviews to the estimated panel population from the second and sixth interviews.

The one-sample McNemar tests in Table One test the probability that the proportion unemployed does not change between the initial and the subsequent interview within the same panel. We use the Control panel to examine the unemployment flux from one month to the next in the absence of CATI. Note that the two significant point estimates are in the opposite direction.

The omnibus hypothesis test was significant (p-value=0.00), so we tested the mean of these points. Because we were unable to reject this test (p-value=0.24), we did not test any further linear combinations.

Note the negative unemployment flux in the Treatment panel. This observation is substantiated by the significant result from the formal test of the omnibus test (p-value=0.00), and the significant result for the hypothesis $\underline{1}'\underline{\mu}=0$ (p-value=0.00).

Consider the two-sample McNemar test results in Table Two. Individually, the monthly results do not demonstrate a clear difference in the unemployment flux between the two panels. On the other hand, the omnibus test is significant (p-value=0.00). The mean unemployment flux seems to be lower in the treatment panel as evidenced by the significant test results of the hypothesis $\underline{1}'\underline{\mu} = 0$, where $\underline{\mu}$ is the vector of (p_2-p_1)-$(p_2'-p_1')$'s, with each element corresponding to a month's estimate (p-value=0.01).

The two-sample t-tests presented in Thompson (1994) failed to detect a difference by panel in mean unemployment rate using the PS split panel data. This contrasts with the CPS CATI Phase-in results: over two years, the CATI (Treatment) panel had consistently significantly higher unemployment rates than the non-CATI (Control) panel. See Shoemaker (1993). In this analysis of PS split panel data, we have evidence that unemployment is lower in the presence of CATI. There are, however, some problems with the data. First, there is some confounding in the Treatment (CATI) panel, since not all respondents in this panel have their second interview conducted from a centralized telephone facility. Second, the expected sample size in the pertinent Control panel cells was near ten, which could be small enough to make the distribution behave unpredictably. This latter problem is not an issue with the analysis presented below.

CPS CATI Phase-in Project Results

The CPS CATI Phase-in Project was a continuation of the study presented in Shoemaker (1993). CATI interviewers used an automated version of the old CPS paper questionnaire, with a modified version of the lead-in labor force question. More details are provided in Thompson (1994). The data considered in this paper are from the same time period as the PS split panel data: 10/92 - 12/93, omitting the 2/93 - 3/93 time frame. Expected cell sizes in both the panels were over one hundred, and so all other sets of data are included.

The one-sample McNemar test results for both panels are presented in Table Three. Test statistics are constructed with unlinked data. The Control panel estimates the unemployment flux from one month to the next in the absence of CATI. The monthly tests for the Control panel do not appear to exhibit any particular movement. Furthermore, the omnibus hypothesis test was not significant (p-value=0.29), so we did not test any further linear combinations.

On the average, although quite variable, the estimates of p_1' are about 4 percent larger than the estimates of p_2'. The Treatment (CATI) panel estimates of p_2 are larger on the average than the estimates of p_1. Given the Control panel's estimates behavior, this phenomenon provides some evidence of a CATI effect.

Note the movement in the Treatment panel from **not** unemployed to unemployed. This observation is substantiated by the significant result from the formal test of the omnibus test (p-value=0.00), and the significant result for the hypothesis $\underline{1}'\underline{\mu}=0$ (p-value=0.00). In contrast to the PS results, this data provides some evidence that unemployment rate is higher in the presence of CATI.

This evidence is further substantiated by the two-sample McNemar test results provided Table Four. These individual monthly results provide some evidence of difference in the unemployment flux. Moreover, the omnibus test is significant (p-value=0.00). The mean unemployment flux in the Treatment panel seems to be higher as evidenced by the significant test results of the hypothesis $\underline{1}'\underline{\mu} = 0$.

The two-sample t-tests presented in Thompson (1994) detected a **positive** difference by panel in mean unemployment rate using the CPS split panel data. These results were consistent with the CPS CATI Phase-in results presented in Shoemaker (1993). This analysis of CPS split panel data reinforces that conclusion. Again, it is impossible to attribute the positive net migration from not unemployed to unemployed entirely to the effect of CATI.

Discussion

Our results appear to yield opposite conclusions about the effect of CATI on unemployment flux. The CATI effect is not, however, the same in both tests.

Perhaps the key difference is the questionnaire. The PS data was collected using the redesigned CPS questionnaire. The new questionnaire was designed as an automated instrument. In contrast, the old CPS questionnaire used for the CPS CATI Phase-in Project was designed as a paper instrument. Field interviewers were required to memorize complicated skip patterns. To minimize respondent burden, CPS interviews generally last about twenty minutes. Using an automated questionnaire, an interviewer can collect more (and more detailed) information in the same amount of tir , since she no longer has to determine the path of the interview. The wording of the labor force questions also differs between the two questionnaires.

PS interviews were conducted using the same questionnaire both in the field interviews (using a laptop computer) or from the CATI facility. In contrast, the CPS CATI Phase-in interviews used two different versions of the old questionnaire: a paper version for the field interviews; and an automated version, with a slightly modified lead-in labor force question for the CATI interviews.

Given these questionnaire differences, and the caveats about the PS split panel data, it would be unwise to draw any clear conclusion about the effect of CATI alone from these two studies. Instead, we recommend continuing to examine this effect by using two-sample McNemar techniques on the new CPS split panel data, which uses the old CATI Phase-in design and the redesigned, fully automated questionnaire.

Conclusion

We have presented two modifications of the two-sample McNemar test using complex survey data, with applications from the unlinked data modification. If the survey does not have a longitudinal design, then the application using the linked data will have an unknown covariance structure and will include a variance component due to matching error. In this case, using the unlinked data makes sense with respect to the model's interpretation, although the statistic based on the (unlinked) estimates of marginal probabilities may be inferior to a well-developed linked model.

Acknowledgements

We thank James Hartman, Alfredo Navarro, James Roebuck, and Lynn Weidman for their useful comments. We would also like to thank Sue Chandler for her typing of this paper.

References

Feuer, Eric J. and Kessler, Larry G. (1989), Test Statistic and Sample Size for a Two-Sample McNemar Test, *Biometrics* ***45***, 629-636.

Fisher, Robin; Robison, Edwin; Thompson, Jenny; Welch, Michael (1993), Variance Estimation in the CPS Overlap Test, *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Fisher, Robin and McGuinness, Richard (1993), "Correlations and Adjustment Factors for CPS," internal memorandum, Demographic Statistical

Methods Division, Bureau of the Census, Washington, D.C.

Marascuilo, Leonard A.; Omelich, Carol L.; Gokhale, D.V. (1988), Planned and Post Hoc Methods for Multiple-Sample McNemar (1947) Tests With Missing Data. *Psychological Bulletin 103*, 238-245.

Shoemaker, Harland H. (1993), Results from the Current Population Survey CATI Phase-in Project, *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Thompson, Jenny (1994), Mode Effects Analysis of Labor Force Estimates, *CPS Overlap Analysis Team Technical Report 3*, Bureau of the Census, Washington, D.C.

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

Table One

	Treatment			Control		
Time Frame	p_2-p_1	se(p_2-p_1)	P-Value	p_2'-p_1'	se(p_2'-p_1')	P-Value
10/92 - 11/92	-0.62	0.29	0.03	2.44	0.81	0.00
11/92 - 12/92	-0.47	0.28	0.09	0.11	0.83	0.89
04/93 - 05/93	-0.76	0.27	0.00	0.20	0.72	0.78
06/93 - 07/93	-0.04	0.27	0.88	0.97	0.71	0.17
08/93 - 09/93	-0.66	0.27	0.02	-1.73	0.68	0.01

One-Sample McNemar Tests for PS Split Panel Study (Unlinked Data)

Table Two

Time Frame	$(p_2$-$p_1)$-$(p_2'$-$p_1')$	se$[(p_2$-$p_1)$-$(p_2'$-$p_1')]$	P-Value
10/92 - 11/92	-3.06	0.86	0.00
11/92 - 12/92	-0.58	0.88	0.51
04/93 - 05/93	-0.95	0.77	0.22
06/93 - 07/93	-1.02	0.76	0.18
08/93 - 09/93	1.08	0.74	0.14

Two-Sample McNemar Tests for PS Split Panel Study (Unlinked Data)

Table Three

	Treatment			Control		
Time Frame	p_2-p_1	se(p_2-p_1)	P-Value	p_2'-p_1'	se(p_2'-p_1')	P-Value
10/92 - 11/92	1.13	0.16	0.00	0.05	0.47	0.92
11/92 - 12/92	0.07	0.17	0.66	-0.14	0.47	0.76
12/92 - 01/93	0.43	0.13	0.00	0.72	0.43	0.09
01/93 - 02/93	0.00	0.14	0.97	-0.91	0.43	0.03
03/93 - 04/93	-0.25	0.14	0.07	-0.16	0.39	0.69
04/93 - 05/93	0.63	0.13	0.00	-0.18	0.43	0.67
05/93 - 06/93	0.88	0.13	0.00	0.47	0.38	0.22
06/93 - 07/93	0.84	0.13	0.00	-0.32	0.46	0.49
07/93 - 08/93	-0.07	0.14	0.61	-0.52	0.39	0.19
08/93 - 09/93	0.42	0.13	0.00	-0.54	0.44	0.23
09/93 - 10/93	0.06	0.12	0.60	-0.08	0.37	0.83
10/93 - 11/93	1.05	0.12	0.00	-0.63	0.42	0.13
11/93 - 12/93	0.18	0.14	0.20	-0.09	0.37	0.82

One-Sample McNemar Tests for CPS CATI Phase-in Project (Unlinked Data)

Table Four

Time Frame	$(p_2$-$p_1)$-$(p_2'$-$p_1')$	se$[(p_2$-$p_1)$-$(p_2'$-$p_1')]$	P-Value
10/92 - 11/92	1.18	0.50	0.02
11/92 - 12/92	0.22	0.50	0.67
12/92 - 01/93	-0.29	0.45	0.52
01/93 - 02/93	0.92	0.45	0.04
03/93 - 04/93	-0.10	0.42	0.81
04/93 - 05/93	0.81	0.45	0.07
05/93 - 06/93	0.41	0.41	0.31
06/93 - 07/93	1.16	0.48	0.02
07/93 - 08/93	0.45	0.42	0.28
08/93 - 09/93	0.95	0.46	0.04
09/93 - 10/93	0.14	0.39	0.71
10/93 - 11/93	1.69	0.44	0.00
11/93 - 12/93	0.26	0.40	0.51

Two-Sample McNemar Tests for CPS CATI Phase-in Project (Unlinked Data)