ADAPTATION OF A CHI-SQUARE TEST TO COMPLEX NHDS SAMPLES

Iris M. Shimizu and Jai Choi, National Center for Health Statistics
Iris M. Shimizu, NCHS, 6525 Belcrest Road, Room 915, Hyattsville, MD 20782

KEY WORDS: Complex samples, Correlation, Hospitals

1. Introduction

A test of independence for aggregate annual estimates between years was desired for a specific class of hospital discharges for a study involving data from the National Hospital Discharge Survey (NHDS). Because the NHDS uses a stratified cluster sample with hospitals being selected at the first or second stage and discharges being sampled within hospitals, the classic tests, such as the χ^2-tests, are inappropriate because the classic tests require simple random samples. Furthermore, the same hospitals are in the NHDS sample for several years.

Choi and McHugh (1989) recently developed a corrected χ^2-test which can be used with complex samples, but it requires information on correlation in the sample. The adaptation of that test to NHDS was enabled by correlation coefficients estimated for NHDS by Bean and Hoffman (1992).

This paper presents the Choi-McHugh test statistic for a four stage sample and its adaption to the NHDS sample. The NHDS survey design is outlined in the next section while Section 3 presents the corrected χ^2-test statistic for a four stage sample. Sections 4-6 discuss and illustrate the corrected test as adapted to the NHDS.

2. NHDS sample design

The NHDS universe consists of non-institutional, non-Federal hospitals in the 50 States and the District of Columbia which have an average length of stay for all patients of less than 30 days and have six or more beds staffed for inpatient use. After 1987, the universe also included all general hospitals.

For the period 1965-87, hospitals with 1,000 or more beds were selected with certainty. The remaining sample was a two-stage, stratified sample. At the first stage, hospitals were sampled from strata defined by the four Census regions and ownership. At the second stage, discharges were selected within systematic random sampling from discharge lists maintained at the sample hospitals. See Simmons and Schnack (1970) for details of the 1965 design.

For the period 1988 to the present, hospitals with 1,000 or more beds or 40,000 discharges annually were selected with certainty. The remaining sample was a three-stage, stratified sample. Selected at the first stage are primary sampling units (PSUs) which are counties or groups of counties or county equivalents or towns and townships (for some PSUs in New England and Hawaii). At the second stage, hospitals were selected with probability proportional to their annual volume of discharges. Within the sampled hospitals, discharges were selected with systematic random sampling from discharge lists. PSUs were stratified by region and population size while hospitals in sample PSUs were stratified by PSU group and their status as subscribers of commercial abstracting services for their medical records. See Shimizu (1990) for details of the 1988 NHDS sampling design.

The sample hospitals and PSUs are retained in the NHDS sample until the sample is redesigned. That is, hospitals selected in 1965 were retained until 1988 when the NHDS was redesigned. PSUs and hospitals selected in 1988 will be retained until the sample is again redesigned. However, the sample is used primarily to produce annual statistics. Hence, in reality, the selection of data years to produce statistics adds another sampling stage to the NHDS. For the subsequent discussion on χ^2 tests, the sample of data years is treated as though it were selected from within sample hospitals because the PSU and hospital samples are fixed without regard to the data years used. That is, the sequence of sampling stages is assumed to be: PSU, hospital, year, and discharge.

3. Corrected χ^2 tests for Complex Samples

The classic χ^2-test statistic must be corrected before it is applied to data from complex samples. The test statistic developed by Choi and McHugh (1989) is:

$$\frac{\chi^2}{c} \qquad (1)$$

where χ^2 is the usual chi-square statistic that may be formulated as:

$$\chi^2 = \sum_i \frac{[Y_i - E(Y_i)]^2}{E(Y_i)} \qquad (2)$$

with Y_i as an aggregate or a percent statistic.

The correction factor **c** in the denominator of (1) is a linear function of the correlations that exist between units at each stage of sampling. For a four stage sample, **c** is formulated as:

$$c = W[1 + \theta_1 \bar{n}_{..} b \bar{m} (\bar{h} - 1) + \theta_2 \bar{n}_{..} b (\bar{m} - 1) + \theta_3 \bar{n}_{..} (b - 1) + \theta_4 (\bar{n}_{..} - 1)] \quad (3)$$

where W is an average of sample weights, the θs are the correlations between the units at each sampling stage, and the coefficients of the θs are the products of average numbers of sample units selected at the different sampling stages.

Under certain assumptions, equation (3) simplifies. For an extreme case, suppose that all of the correlations are equal to one, that is, $\theta_1 = \theta_2 = \theta_3 = \theta_4 = 1$, then

$$c = W[\bar{n}\, b\, \bar{m}\, h] .$$

On the other hand, if there is no correlation, that is, if $\theta_1 = \theta_2 = \theta_3 = \theta_4 = 0$, then:

$$c = W ,$$

and there a weighting effects, but no design effects, on the test.

4. Application to NHDS

For the NHDS, the W average of sample weights in the correction factor **c** [equation (3)] may be formulated as:

$$W = \frac{\hat{N}_.}{b \sum_{k=1}^{b} \frac{\hat{N}_k}{\hat{N}_.} n_k} \quad (4)$$

where

b = number of years of data used in the study. (Groups of years may be used in place of individual years.)

k denotes year (group of years).

$\hat{N}_k$ = estimated total number of discharges for the kth year.

$$\hat{N}_. = \sum_{k=1}^{b} \hat{N}_k \quad (5)$$

is the estimated total discharges across years included in the study.

n_k = unweighted total number of sample discharges in the kth year.

The remaining parameters for NHDS in equation (3) for **c** may be defined as:

θ_1 = correlation between PSUs within PSU strata.

θ_2 = correlation between hospitals within PSU.

θ_3 = correlation between years within hospital.

θ_4 = correlation between discharges within year.

$\bar{h}$ = average number of sample PSUs with NHDS data within PSU stratum.

$\bar{m} = \frac{\sum_k m_k}{b g \bar{h}}$ is the average number of respondent hospitals per PSU per year.

m_k = number of hospitals in kth year.

g = number of PSU strata.

$\bar{n}_{..} = \frac{\sum_k n_k}{\sum_k m_k}$ is the average number of sample discharges per year per hospital.

Knowledge about the correlations between sampling units at each stage is required, as well as sample sizes. Starting at the PSU sampling stage, the term involving correlation between PSUs drops out for both the 1965–87 and the present sample (after 1987). The 1965–87 sample did not have geographic PSUs. The present sample only has one PSU per stratum, so that the coefficient $(\bar{h} - 1)$ of the PSU correlation coefficient is zero.

For the remaining terms involving correlation in the correction factor **c**, the following assumptions are made with regard to the characteristic of interest in the study:

- There is no correlation between hospitals. For example, a high volume for the characteristic at one hospital does not suggest that a high volume for that characteristic occurs at another hospital. Correlation, if any exists, is likely to be negative within small areas for diagnostic and procedural variables, especially if hospitals avoid duplicating specialty services provided at other hospitals in their areas. Negative correlations should be set equal to zero in the correction factor to assure a conservative test.
- Within an individual hospital, there is correla–

tion between years. That is, if a hospital has a high volume for a variable relative to other hospitals in a year, then that hospital will likely have a relatively high volume for that variable in other years.

- Discharges are not correlated to each other (or the correlation is negligible). It is likely that multiple discharges to a few individuals are selected to the sample from some hospitals and those may be correlated. However, the numbers of such multiple discharges in the sample are small relative to the size of the annual sample of discharges. There may also be correlation between discharges if a common disaster sends people to hospitals in a community (for example, a major passenger train crash), but these disasters are rare and, hence, ignored in this paper.

With this set of information and assumptions, the correction factor to the χ^2-test statistic c simplifies to:

$$c = W[1 + \theta_3 \bar{n}_{..}(b - 1)] \ . \qquad (6)$$

That is, the correction factor is a linear function of only one correlation, the correlation over the years included in the study.

5. Estimating correlation between NHDS years

The correlation coefficients between years in NHDS are not in general readily available. However, Bean and Hoffman (1992) recently calculated the correlation coefficients for hospitals between years for two pairs of years (1982-83 and 1983-84) for selected variables. These are presented in Table 1. If the table includes the variable of interest, but not for the set of years included in one's study, the analyst may estimate the correlation required for his test statistic by taking the average of the two coefficients given in the table for that variable. The averages probably overstate the true correlation coefficients for periods of more than two years, and as such, will result in conservative tests.

For variables not included in the table, analysts must obtain their own correlation coefficient estimates or assume that the correlation for their variable is similar to that for one of the variables included in the tables.

6. Illustration of the adapted test

The question that originally lead to the development of a χ^2-test for use in the NHDS was whether independence existed across years in the volume of hospital discharges having a particular characteristic. More frequently, though, analysts opt to use multiple years of NHDS data only when their study involves rare characteristics for which a single year of NHDS sample is inadequate to yield reliable estimates. Hence, both situations are included in the following illustrations.

The following example attempts to demonstrate use of Table 1 to determine the correlation required in the adapted test. The first example takes a figure directly from the table. The second example requires an extension of the table results to sets of years not covered in the table.

The examples require sample sizes and estimated universe size for each year of data in the analysis. These are given in Table A for both examples.

Example 1. Independence of variables in 1982-83

For simplicity, suppose one wants to test the independence of two variables for hospital discharges in 1982-83. There are b =2 years of data in the analysis. From Table A, the other items needed to compute the factor c in the corrected χ^2-test are:

$$\bar{n}_{..} = \frac{n_{1982} + n_{1983}}{m_{1982} + m_{1983}} = 497.6 \ ,$$

$$\hat{N}_{.} = \hat{N}_{1982} + \hat{N}_{1983} = 77{,}373{,}000 \ ,$$

and for the W in equation (4),

$$W = \frac{\hat{N}_{.}}{b\left(\frac{\hat{N}_{1982}}{\hat{N}_{.}} n_{1982} + \frac{\hat{N}_{1983}}{\hat{N}_{.}} n_{1983}\right)} = 184.2 \ .$$

Hence, for the 1982-83 NHDS, equation (6) for the correction factor becomes:

$$c = 184.2(1 + \theta_3 \times 497.6) \qquad (7)$$

TABLE A: Respondent sample units and estimated total number of discharges by year: NHDS

	Respondent sample		Estimated total
	Hospitals	Discharges	discharges
Year	m_k	n_k	$\hat{N}_k$
1982	426	214,000	38,594,000
1983	418	206,000	38,784,000
1984	407	192,000	37,162,000
1985	414	195,000	35,057,000

which is now a function only of the common correlation θ_3 between years at hospitals for the variable(s) being analyzed in the study.

Suppose for the sake of illustration that one wants to test the independence of sex and age in 1982–83. It can be seen in Table 1 that the median of the six correlation coefficients for age and sex in 1982–83 NHDS is 0.91. Hence, we assume for our test that $\theta = 0.91$. (To be most conservative, one could instead assume that $\theta = 0.96$, the maximum of the six correlations.) Substituting for $\theta = 0.91$ in equation (7) yields $c = 83{,}617.0$.

To compute the numerator of the test statistic, the aggregate estimates for discharges by sex and age in 1982–83 are provided in Table B.

If i and j denote sex and age groups, respectively, then the expected values $E(Y_{ij})$ given in B are derived by :

$$E(Y_{ij}) = \frac{\sum_i Y_{ij} \sum_j Y_{ij}}{\sum_j \sum_i Y_{ij}} .$$

The classic χ^2 statistic for this test is:

$$\chi^2 = \sum_i^I \sum_j^J \frac{[Y_{ij} - E(Y_{ij})]^2}{E(Y_{ij})} = 2{,}710{,}683.6 .$$

This statistic has $(I-1)(J-2)=(2-1)(3-1)=2$ degrees of freedom (for two sex and three age groups).

Hence, from equation (1), the corrected test statistic is:

$$\frac{\chi^2}{c} = \frac{2{,}710{,}683.6}{83{,}613.0} = 32.4 ,$$

which is significant for 2 degrees of freedom. Thus, based on this test, one is justified in stating that age and sex are independent variables.

TABLE B. Aggregate estimates for discharges (in thousands) by age and sex from the NHDS: United States, 1982-83.

Age	Sex				
	Both	Male		Female	
		Y_{Mj}	$E(Y_{Mj})$	Y_{Fj}	$E(Y_{Fj})$
All years	77,378	31,043		46,335	
<15years	7,308	4,182	2,932	3,126	4,376
15–44 yrs.	30,823	9,139	12,366	21,684	18,457
>44years	39,247	17,722	15,745	21,525	23,502

Example 2. Independence across 1984–85

Suppose again for the sake of illustration that one wants to test independence across years 1982–85 for discharges to males. Here $b = 4$ and from equation (4) and Table A, it can be shown that,

$$\bar{n}_{..} = 484.7 \text{ and } W = 185.1$$

and

$$c = 185.1[1 + \theta \times 484.7 \times (4 - 1)]$$

$$= 185.1(1 + \theta \times 1{,}454.1) .$$

Table 1 does not include any correlation coefficients across four years which are needed for the present test. The needed correlation for c may be approximated by the average of coefficients that are present for that variable. It can be seen in Table 1 that the coefficients for discharges to males is 0.91 and 0.95 for 1982–83 and 1983–84, respectively. The coefficient for our test is approximated by the average:

$$\theta_3 = (0.91 + 0.95)/2 = 0.93 .$$

The correction factor is then:

$$c = 250{,}541 .$$

The aggregate estimates for discharges by males in the years 1982–85 are provided in Table C. Here,

$$E(Y_i) = \frac{\sum_i Y_i}{b}$$

is the simple average of discharges to males over the four years. From Table C, it can be shown that the classic χ^2 statistic is:

$$\chi^2 = \sum_k^4 \frac{[Y_i - E(Y_i)]^2}{E(Y_i)} = 83{,}884.3 ,$$

with $4 - 1 = 3$ degrees of freedom. Adding the correction factor c, the corrected test statistic from equation (1) becomes:

TABLE C. Aggregate estimates for discharges (in thousands) to males by year from the NHDS: United States, 1982-85.

Year	Discharges to Males (in thousands)	
	Y_i	$E(Y_i)$
Total	60,104	
1982	15,470	15,026
1983	15,573	15,026
1984	14,900	15,026
1985	14,161	15,026

$$\chi^2 = \frac{83,884}{250,541} = 0.335 ,$$

which is not significant at the 5 percent level.

7. Summary

The classic χ^2-test statistic must be corrected before it is applied to data from complex samples such as the stratified, cluster samples used in the National Hospital Discharge Survey. A corrected χ^2-test developed in recent literature for such samples is adapted for use with NHDS data. When multiple years of data are used, the correction factor requires knowledge about the common correlation between years for hospitals. With the aid of coefficient estimates that are available, the corrected test is illustrated with multi-year NHDS data.

REFERENCE

Bean, JA and Hoffman, KL (1992). Covariances for Estimated Totals When Comparing Between Years. National Center for Health Statistics, Vital and Health Statistics 2 (No. 114).

Choi, JW, and McHugh, RB (1989). "A Reduction Factor in Goodness-of-Fit and Independence Tests for Clustered and Weighted Observations." Biometrics 45. pp. 979-996.

Shimizu, IM (1990). "The New Statistical Design of the National Hospital Discharge Survey." American Statistical Association 1990 Proceedings of the Section on Survey Research Methods. pp. 702-707.

Simmons, WR, and Schnack, GA (1970). Development of the Design of the NCHS Hospital Discharge Survey. National Center for Health Statistics, Vital and Health Statistics 2 (No. 39).

Table 1. Correlation coefficients for the number of discharges, days of care, and surgical procedures for patients discharged from short-stay hospitals, by selected characteristics: United States, 1982-84

Characteristic	No. of discharges		Days of care		Characteristic	Procedures	
	1982-3	1983-4	1982-3	1983-4		1982-3	1993-4
	Coefficients					Coefficients	
					All procedures:		
Under 15 years . . .	0.96	0.98	0.94	0.96	Total	0.89	0.95
15-44 years	0.92	0.96	0.81	0.92	Under 15 years . .	0.93	0.94
45-64 years	0.88	0.95	0.84	0.89	15-44 years	0.91	0.95
65 years and over .	0.91	0.95	0.87	0.91	45-65 years	0.87	0.93
					65 years and over .	0.86	0.93
Female	0.91	0.96	0.85	0.93			
Male	0.91	0.95	.088	0.91	Female	0.90	0.96
					Male	0.86	0.92
Malignant neoplasm							
of lung	0.49	0.61	0.44	0.46	Extraction of lens,		
of lung, male . . .	0.45	0.41	0.34	0.29	65 years and over .	0.86	0.79
of breast, female .	0.52	0.49	0.47	0.32			
					Myringotomy,		
Mental disorders . .	0.95	0.94	0.92	0.93	under 15 years . . .	0.82	0.83
15-44 years	0.96	0.93	0.92	0.93			
					Tonsillectomy,		
Cataract	0.89	0.77	0.85	0.74	under 15 years . . .	0.74	0.75
65 years and over	1.88	0.76	0.82	0.70			
					Cardiac catheterization	0.90	0.92
Acute myocardial							
infraction	0.72	0.75	0.53	0.63	Direct heart		
					revasculatization .	0.84	0.85
Atherosclerotic							
heart disease	0.79	0.83	0.69	0.79	Prostatectomy	0.68	0.71
					65 years and over .	0.64	0.67
Inguinal hernia, male	0.58	0.57	0.40	0.32			
					Circumcision,		
Congenital anomaly,					under 15 years . . .	0.85	0.91
under 15 years . .	0.95	0.96	0.93	0.89			
					Bilateral occlusion of		
Fracture of neck					fallopian tube,		
of femur					15-44 years	0.93	0.93
65 years and over	0.67	0.60	0.53	0.41			
					Hysterectomy	0.88	0.92
Females					45-64 years	0.63	0.65
with deliveries . .	0.95	0.96	0.90	0.95	65 years and over .	0.43	0.20
					Cesarean section:		
					All ages	0.91	0.94
					Arthroplasty and		
					replacement of hip:		
					65 years and over .	0.53	0.47
					Female	0.55	0.43

SOURCE: National Hospital Discharge Survey, 1982-84 and Bean and Hoffman (1992)