
Lecture 25: Models for Matched Pairs

Dipankar Bandyopadhyay, Ph.D.

BMTRY 711: Analysis of Categorical Data Spring 2011

Division of Biostatistics and Epidemiology

Medical University of South Carolina

Matched Pairs

- We are going to conclude the semester examining models for repeated categorical data
- Repeated observations are common in longitudinal studies
- Since the data are matched on subject, they are often statistically dependent
- This dependence needs to be incorporated into our statistical models
- Our first example is when we have two repeated observations

Example

- 1600 British citizens were surveyed on the Prime Minister's job performance
- Each citizen rated the Prime Minister as “approve” or “disapprove”
- Then, after 6 months, each citizen re-rates the Prime Minister
- A (2×2) table on the matched ratings is provided below

First Survey	Second Survey		
	Approve	Disapprove	
Approve	794	150	944
Disapprove	86	570	656
Total	880	720	1600

Summary Measures

- One way we can begin to examine these data is to calculate the Odds Ratio

$$\begin{aligned}OR &= \frac{794 \cdot 570}{86 \cdot 150} \\ &= 35.1\end{aligned}$$

- An OR of 35 indicates a strong association among opinions
- That is, the odds of approving after 6 months are over 35 times higher if a citizen approved initially
- Also, 1364 of the 1600 did not change opinion after 6 months
- 236 changed opinion
- We are interested in seeing whether or not the general opinion is changing over time

Dependent Proportion

- Let π_{ab} be the probability of outcome a for the first observation and b for the second
- For example, π_{11} is the probability of approving at baseline and after 6 months still approving
- For a (2×2) table, the following table summarizes the true probabilities

First Survey	Second Survey		
	Approve	Disapprove	
Approve	π_{11}	π_{12}	π_{1+}
Disapprove	π_{21}	π_{22}	π_{2+}
Total	π_{+1}	π_{+1}	$\pi_{++} = 1$

Marginal Homogeneity

- **Marginal homogeneity** occurs when

$$\pi_{1+} = \pi_{+1}$$

- Note, since

$$\pi_{1+} - \pi_{+1} = 0$$

- Then

$$\pi_{1+} - \pi_{+1} = (\pi_{11} + \pi_{12}) - (\pi_{11} + \pi_{21}) = \pi_{12} - \pi_{21} = 0$$

- Then,

$$\begin{aligned}\pi_{21} - \pi_{12} &= (\pi_{22} + \pi_{21}) - (\pi_{22} + \pi_{12}) \\ &= \pi_{2+} - \pi_{+2}\end{aligned}$$

- Thus, the **marginal homogeneity** model has the marginal column probabilities equaling the marginal row probabilities
- Or, the off diagonal probabilities are equal
- To develop a statistical test for this model, we will now consider the observed sample information

- let n_{ab} be the count of subjects moving from a to be b
- Then, in terms of the (2×2) , the observed counts are

First Survey	Second Survey		
	Approve	Disapprove	
Approve	n_{11}	n_{12}	n_{1+}
Disapprove	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+1}	n_{++}

Observed Proportions

- In terms of observed proportions

First Survey	Second Survey		
	Approve	Disapprove	
Approve	p_{11}	p_{12}	p_{1+}
Disapprove	p_{21}	p_{22}	p_{2+}
Total	p_{+1}	p_{+2}	$p_{++} = 1$

- where $p_{ab} = n_{ab}/n_{++}$
- Let $d = p_{+1} - p_{1+} = p_{2+} - p_{+2}$
- Note, d is the difference of two **dependent** proportions

- If p_{+1} and p_{1+} were independent, then

$$\begin{aligned} \text{var}(d) &= \text{var}(p_{+1} - p_{1+}) \\ &= \text{var}(p_{+1}) + \text{var}(p_{1+}) \\ &= \frac{p_{+1}(1-p_{+1})}{n_{+1}} + \frac{p_{1+}(1-p_{1+})}{n_{1+}} \end{aligned}$$

- However, since p_{+1} and p_{1+} are dependent, then we need to subtract out the covariance
- Thus

$$\begin{aligned} \text{var}(d) &= \text{var}(p_{+1} - p_{1+}) - 2\text{cov}(p_{+1}, p_{1+}) \\ &\quad \text{where } \text{cov}(p_{+1}, p_{1+}) \text{ equals} \\ &\quad (p_{11}p_{22} - p_{12}p_{21})/n \end{aligned}$$

- Therefore

$$\text{var}(\sqrt{nd}) = p_{+1}(1 - p_{+1}) + p_{1+}(1 - p_{1+}) - 2(p_{11}p_{22} - p_{12}p_{21})$$

- After some algebra,

$$\widehat{var}(d) = [(p_{12} + p_{21}) - (p_{12} - p_{21})^2] / n$$

- This is the large sample (i.e., Wald or non-null) variance
- Thus, the test of marginal homogeneity is the same as testing

$$H_0 : d = 0$$

- Using

$$z = \frac{d}{\sqrt{\widehat{var}(d)}}$$

McNemar's Test

- Just as we have had all semester long, we could consider the variance under the null
- If $\pi_{12} - \pi_{21}$, then an alternative variance under the null is

$$\text{var}_0(d) = \frac{p_{12} + p_{21}}{n} = \frac{1/n(n_{12} + n_{21})}{n} = \frac{n_{12} + n_{21}}{n^2}$$

- Then testing using the null standard error would yield

$$\begin{aligned} z_0 &= \frac{p_{+1} - p_{1+}}{\sqrt{\text{var}_0(d)}} \\ &= \frac{(n_{11} + n_{21} - n_{11} - n_{12})/n}{\sqrt{\text{var}_0(d)}} \\ &= \frac{n_{21} - n_{12}}{n\sqrt{\text{var}_0(d)}} \\ &= \frac{n_{21} - n_{12}}{(n_{12} + n_{21})^{1/2}} \end{aligned}$$

- McNemar's test is the square of z_0 and is distributed as chi-square with 1 *df*.

Revisiting Example

- We have observed a strong association (OR > 35), but does the marginal homogeneity model hold?

First Survey	Second Survey		
	Approve	Disapprove	
Approve	794	150	944
Disapprove	86	570	656
Total	880	720	1600

- For this data

$$p_{1+} = 944/1600 = 0.59$$

and

$$p_{+1} = 880/1600 = 0.55$$

Difference in Dependent Proportions

- A large sample comparison of these proportions is

$$d = 0.55 - 0.59 = -0.04$$

- with standard error of

$$\widehat{var}(d) = [(150 + 86) - (150 - 86)^2] / 1600 = 0.000091188$$

- Thus, a large sample confidence interval is

$$-0.04 \pm 1.96\sqrt{0.000091188} = (-0.06, -0.02)$$

- Thus, the rating of the Prime Minister appears to have dropped between 2% and 6%

McNemar's

- McNemar's Test is

$$z_0^2 = \left[\frac{86 - 150}{(86 + 150)^{1/2}} \right]^2 = (-4.17)^2 = 17.3559$$

- Which is less than -1.96, so we have strong evidence to support a drop in rating

Using SAS

```
data one;
  input pre post count;
  cards;
  1 1 794
  1 2 150
  2 1 86
  2 2 570
  ;
run;
proc freq;
  tables pre*post /agree;
  weight count;
run;
```

Statistics for Table of pre by post

McNemar's Test

Statistic (S) 17.3559
DF 1
Pr > S <.0001

Small Sample p -value

- The null hypothesis of marginal homogeneity is

$$H_0 : \pi_{12} = \pi_{21}$$

- Which is the same as

$$\frac{\pi_{21}}{\pi_{21} + \pi_{12}} = 0.5$$

since

$$\pi_{21} = 0.5\pi_{21} + 0.5\pi_{12}$$

-
- Let $n^* = n_{21} + n_{12}$ (that is n^* is the number of discordant (off diagonal pairs))
 - Under H_0 , $n_{21} \sim b(n^*, 0.5)$
 - With $E(n_{21}) = 0.5n^*$
 - That is, the discordant data should be equally distributed if marginal homogenous
 - The exact p -value is the binomial tail probability.
 - Recall, exact binomial inference is based on the “as extreme or more extreme” principle

Example

- Consider the following made up example

	Rating 2		
Rating 1	1	2	
1	35	12	47
2	3	40	43
	38	52	90

- 90 subjects appears to be a large sample, but McNemar's is only based on discordant pairs
- For this example, we have 15 (=12+3) pairs
- We should consider using an exact method based on the binomial

Binomial PDF macro

Recall our Binomial PDF macro

```
%macro mybinomialpdf(p,n);  
dm "output" clear; dm "log" clear;  
options nodate nocenter nonumber;  
data myexample;  
  do i = 0 to &n;  
    prob = PDF('BINOMIAL',i,&p,&n) ;  
    cdf = CDF('BINOMIAL',i,&p,&n) ;  
    mlcdfprob = 1-cdf+prob;  
  
    output;  
  end;  
  label i = "Number of *Successes";  
  label prob = "P(Y=y) ";  
  label cdf = "P(Y<=y)";  
  label mlcdfprob="P(Y>=y)";  
run;  
  
title "Binomial PDF for N=&n and P=&p";  
proc print noobs label split="*";  
run;  
  
%mend mybinomialpdf;  
  
%mybinomialpdf(0.5,15); <-- null prob = 0.5, n* = 15
```

Results

Binomial PDF for N=15 and P=0.5

Number of
Successes

P(Y=y)

P(Y≤y)

P(Y≥y)

0	0.00003*	0.00003	1.00000
1	0.00046*	0.00049	0.99997
2	0.00320*	0.00369	0.99951
3	0.01389*	0.01758	0.99631
4	0.04166	0.05923	0.98242
5	0.09164	0.15088	0.94077
6	0.15274	0.30362	0.84912
7	0.19638	0.50000	0.69638
8	0.19638	0.69638	0.50000
9	0.15274	0.84912	0.30362
10	0.09164	0.94077	0.15088
11	0.04166	0.98242	0.05923
12	0.01389*	0.99631	0.01758 ** Observed case
13	0.00320*	0.99951	0.00369
14	0.00046*	0.99997	0.00049
15	0.00003*	1.00000	0.00003

* indicate as extreme or more extreme values

Exact P-value equals $P(Y \geq 12) + P(Y \leq 3) = 0.01758 + 0.01758 = 0.03516$

Using SAS the easy way

- Using a “new” (version 9+?) version of SAS, you can get the exact p -value directly from proc freq

```
data one;
  input rate1 rate2 count @@;
  cards;
  1 1 35    1 2 12    2 1 3    2 2 40
  ;
run;
proc freq data=one;
  weight count;
  tables rate1*rate2 / agree;
  exact mcnem;
run;
```

- Note the “exact” statement in proc freq

Selected Results

Statistics for Table of rate1 by rate2

McNemar's Test

```
-----  
Statistic (S)          5.4000  
DF                      1  
Asymptotic Pr > S    0.0201  
Exact      Pr >= S    0.0352
```

- The computed exact p -value matches our binomial pdf result
- There is some disagreement among the exact and chi-square based p -value, but the statistical conclusions are the same
- Just remember “large samples” for McNemar’s test implies a large number of discordant pairs (and that the concordant pairs do not contribute to the test statistic)

Marginal Homogeneity for > 2 levels

- McNemar's Test is defined for a simple 2×2 table
- What if you have more than two levels in the classification?
- Consider a study in which 539 individuals were asked to what extent they agree with the notation that either their fellow church members or other friends encouraged them to exercise more
- The research question is to understand how the two different social support mechanisms (church v. other friends) may be contributing to increasing exercise in overweight individuals

Data

	Outside Friend				
Church Member	SA	A	D	SD	Total
SA	47	20	2	2	71
A	39	139	32	3	213
D	24	87	106	7	224
SD	5	5	11	10	31
Total	115	251	151	22	539

- Like McNemar's, we are interested in testing

$$\pi_{k+} = \pi_{+k} \quad \forall k$$

- Or, in terms of cell probabilities

$$\pi_{ij} = \pi_{ji} \quad \forall i, j$$

- Hence, we are testing for symmetry of the probabilities above and below the concordant diagonal.
- which again translates into testing that the two marginal distributions are similar

Using SAS

```
data two;
  input ch_rate fr_rate count @@;
  cards;
1 1 47 1 2 20 1 3 2 1 4 2
2 1 39 2 2 139 2 3 32 2 4 3
3 1 24 3 2 87 3 3 106 3 4 7
4 1 5 4 2 5 4 3 11 4 4 10
  ;
run;
proc freq data=two;
  weight count;
  tables ch_rate* fr_rate / agree;
run;
```

Selected Results

- The test results are:

```
      Test of Symmetry
-----
Statistic (S)      52.8288
DF                  6
Pr > S              <.0001
```

- However, interpreting the results, in this case, is easier if we collapse the categories
- This works since our data is ordinal and “agree” and “disagree” are easily understood (groupings of nominal levels may not make sense)

```
data twob;
  set two;
  if ch_rate > 2 then ch_rate2 = 2; else ch_rate2 = 1;
  if fr_rate > 2 then fr_rate2 = 2; else fr_rate2 = 1;
run;
proc freq data=twob;
  weight count;
  tables ch_rate2* fr_rate2 / agree;
run;
```

- Here, we are back to the regular McNemar’s test

Selected results

Statistics for Table of ch_rate2 by fr_rate2

McNemar's Test

```
-----  
Statistic (S)      42.0250  
DF                 1  
Pr > S            <.0001
```

- 39 observations indicated that they agree church members are encouraging and disagreed that other friends are encouraging
- 121 observations indicated that they agree friends are encouraging and that church members are not
- Our results suggests that a respondent's belief is that friends are more likely to encourage increased exercise (p -value < 0.0001)

Measures of Agreement

- Suppose we have a double dichotomy formed by each of n biostatisticians being rated by two raters on a dichotomous scale:

		RATER 2		
		(1)	(2)	
		Psychot	Neurotic	Total
R	A	Psychotic		
T	(1)	50	10	60
E				
R	1	Neurotic		
	(2)	7	40	47
	Total	57	50	107

Data

		Variable (W)		
		1		2
Variable (X)	1	Y_{11}	Y_{12}	Y_{1+}
	2	Y_{21}	Y_{22}	Y_{2+}
		Y_{+1}	Y_{+2}	n

- In general,

$$X = (\text{rating from rater 1}) = 1 \text{ or } 2$$

$$W = (\text{rating from rater 2}) = 1 \text{ or } 2$$

- X and W can take on the same possible values, say,

$$1 = \text{'positive rating'};$$

$$2 = \text{'negative rating'}.$$

- Neither margin is fixed by design, although the total number of subjects, n , usually is.

-
- Two raters 'agree' if they give a subject the same rating, i.e., the raters agree if they both rate subject i as a '1' or they both rate subject i as a '2'.
 - The simplest choice for a measure of agreement is the overall probability of agreement:

$$\eta = p_{11} + p_{22}$$

DISTINCTION BETWEEN AGREEMENT AND ASSO- CIATION

- For two raters, X and W , to agree, they must fall in a diagonal of the (2×2) table.
- However, in general, for X and W to be associated, you should be able to predict one from the other.

- Consider the following (2×2) table

		Rater 2 (W)	
		1	2
Rater 1 (X)	1	0	p_{12}
	2	p_{21}	0

- For this table, the $OR = 0$, ($\log(OR) = -\infty$), which is perfect negative association: Given the rating from Rater 1, you can perfectly predict the rating of Rater 2, i.e., if $X = 1$, then $W = 2$. and if $X = 2$, then $W = 1$. Thus, the observations are highly associated.
- However, the raters completely disagree, and agreement is low. Using the above measure of agreement, it is 0, i.e.,

$$\eta = p_{11} + p_{22} = 0$$

Another Measure

- Look at the previous measure of agreement,

$$\eta = p_{11} + p_{22}$$

We will see that this is not a great measure of agreement, because it tends to be too high, just by chance.

- If two raters independently rate the same subject, you would expect your measure of agreement to be 0.
- By chance, if two raters independently rate the same subject,

$$p_{jk} = p_{j+}p_{+k},$$

and the measure of agreement η , is

$$\eta = p_{1+}p_{+1} + p_{2+}p_{+2} > 0 ,$$

- For example, suppose the probability of giving a positive rating (rating=1) is high:

$$p_{1+} = p_{+1} = .9 \text{ and thus } p_{2+} = p_{+2} = .1,$$

If the two raters are completely independent, still, by chance alone, the measure of agreement is high:

$$\eta = .9^2 + .1^2 = .82,$$

- Thus, we need to scale this measure somehow; further, if the raters were independent, we would want a measure of agreement to be 0.

The Kappa Coefficient

- Cohen (1960) suggested the following measure, called 'Kappa',

$$\begin{aligned}\kappa &= \frac{\eta - (p_{1+} + p_{+1} + p_{2+} + p_{+2})}{1 - (p_{1+} + p_{+1} + p_{2+} + p_{+2})}, \\ &= \frac{(p_{11} + p_{22}) - (p_{1+} + p_{+1} + p_{2+} + p_{+2})}{1 - (p_{1+} + p_{+1} + p_{2+} + p_{+2})}\end{aligned}$$

- Here, we have subtracted off the value η under independence, so that, if the raters are independent, $\kappa = 0$.
- Further, when agreement is perfect,

$$\eta = p_{11} + p_{22} = 1,$$

which leads to the maximum value of the numerator,

$$\begin{aligned}\max\{(p_{11} + p_{22}) - (p_{1+} + p_{+1} + p_{2+} + p_{+2})\} &= \\ 1 - (p_{1+} + p_{+1} + p_{2+} + p_{+2})\end{aligned}$$

- The denominator of κ is the maximum value of the numerator
- As such, the maximum value κ can take is 1:

$$\kappa = \frac{\max\{\eta - (p_{1+p+1} + p_{2+p+2})\}}{1 - (p_{1+p+1} + p_{2+p+2})} = 1$$

- However, κ can also be negative when agreement is poor. Consider the following (2×2) table:

		Rater 2	
		1	2
	1	0	.5
Rater 1			
	2	.5	0

For this table,

$$\begin{aligned} \kappa &= \frac{(p_{11} + p_{22}) - (p_{1+}p_{+1} + p_{2+}p_{+2})}{1 - (p_{1+}p_{+1} + p_{2+}p_{+2})} \\ &= \frac{0 - (.5^2 + .5^2)}{1 - (.5^2 + .5^2)} = -\frac{1}{3} \end{aligned}$$

Example

- Suppose we have $n = 107$ biostatisticians rated by two raters on a dichotomous scale:

		RATER 2		
		Psychot	Neurotic	Total
R A T E R 1	Psychotic	50	10	60
	Neurotic	7	40	47
	Total	57	50	107

- Just looking at the table, there appears to be high agreement, but is it greater than would be expected by chance?

Using SAS

```
data one;
input rater1 rater2 count;
cards;
1 1 50
1 2 10
2 1 7
2 2 40
;

proc freq;
  table rater1*rater2 / agree;
weight count;
run;
```

```
/* SELECTED OUTPUT */
```

```
Simple Kappa Coefficient
```

```
-----
Kappa          0.6797 <--- Okay, what does this
ASE            0.0711      value mean?
95% Lower Conf Limit  0.5403
95% Upper Conf Limit  0.8190
```

```
Sample Size = 107
```

Interpretation of Kappa

- Interpretation of Kappa is typically based on the guidelines by Landis and Koch (1977)
- The break the values into three categories
- $\kappa > 0.75$ represents excellent agreement beyond chance
- $0.40 \leq \kappa \leq 0.75$ represents fair to good agreement beyond chance
- $\kappa < 0.40$ represents poor agreement beyond chance

Relationship of Kappa with McNemar's Test

- McNemar's (and the generalized test for symmetry) test the null hypothesis of marginal homogeneity
- If McNemar's test is statistically significant, the marginal probabilities differ
- Marginal probabilities differing suggests that the rows and columns are different
- If you want to test if the rows and columns are similar, you should consider Kappa
- In general, if Kappa is high, McNemar's is low and vice versa

Agreement with more than 2 categories

- Diagnoses of $n = 400$ biostatisticians by 2 raters:

		Rater 2			
		Psychot	Neurot	Organic	Total
R a t e r 1	Psychot	75	1	4	80
	Neurot	5	4	1	10
	Organic	0	0	10	10
	Total	80	5	15	400

Weighted v. Unweighted Kappa

- If The ratings are ordinal, allowing for some tolerance in disagreement is feasible
- For example, If one rater strongly agrees and the other just agrees, it seems like we should account for the fact that both raters gave approximately the same rating
- One solution is to use the weighted κ (Fliess, 1980) for a $K \times K$ table.
- Suppose that agreement weights, say w_{jk} , are assigned on rational grounds to the cells of the $(K \times K)$ table:
 - 1. $w_{jj} = 1$
 - 2. $0 \leq w_{jk} < 1$ for $j \neq k$.
 - 3. $w_{jk} = w_{kj}$, (the two raters are considered symmetrically).

- Then, the ‘weighted’ proportion agreement is

$$\eta_w = \sum_{j=1}^K \sum_{k=1}^K w_{jk} p_{jk}$$

and

- The, the ‘chance-expected’ weighted proportion agreement is

$$\eta_c = \sum_{j=1}^K \sum_{k=1}^K w_{jk} p_{j+k}$$

- Then, the WEIGHTED KAPPA is

$$\kappa_w = \frac{\eta_w - \eta_c}{1 - \eta_c}.$$

-
- One choice for the weights is the ‘simple weights’:
 - 1. $w_{jj} = 1$
 - 2. $w_{jk} = 0$ for $j \neq k$.
so that we are measuring agreement along the diagonal. When $K = 2$, this reduces to the κ coefficient for a (2×2) table.
 - Simple weights are appropriate if you are studying nominal ratings

- Another possibility is

$$w_{jk} = 1 - \frac{|j - k|}{K - 1},$$

which gives observations further apart in table less weight; for the above example, with $K = 3$,

$$w_{12} = w_{21} = .5$$

$$w_{23} = w_{32} = .5$$

$$w_{13} = w_{31} = 0$$

and

$$w_{11} = w_{22} = w_{33} = 1.$$

- When $K = 2$, this also reduces to the κ coefficient for a (2×2) table.
- This style of weights is appropriate for ordinal ratings

Example

Diagnoses in $n = 400$ biostatisticians by 2 raters:

		Rater 2			
		Psychot	Neurot	Organic	Total
R a t e r 1	Psychot	75	1	4	80
	Neurot	5	4	1	10
	Organic	0	0	10	10
	Total	80	5	15	400

- Just looking at the table, there appears to be high agreement, but is it greater than would be expected by chance?

Using SAS

```
data one;
input rater1 rater2 count;
cards;
1 1 75
1 2 1
1 3 4
2 1 5
2 2 4
2 3 1
3 1 0
3 2 0
3 3 10
;

proc freq;
  table rater1*rater2 / agree ;
weight count;
run;
```

```
/* SELECTED OUTPUT */
```

Kappa Statistics

Statistic	Value	ASE	95% Confidence Limits	
Simple Kappa	0.6765	0.0877	0.5046	0.8484
Weighted Kappa	0.7222	0.0843	0.5570	0.8874

Sample Size = 100

- Which Kappa is to be reported?
- I would report the simple kappa since the classifications are nominal

Revisiting Exercise Encouragement Example

- Recall our exercise example:

	Outside Friend				
Church Member	SA	A	D	SD	Total
SA	47	20	2	2	71
A	39	139	32	3	213
D	24	87	106	7	224
SD	5	5	11	10	31
Total	115	251	151	22	539

- Here, a weighted Kappa is feasible
- Although, this example isn't the typical kappa example (here, two ratings are given by the subject; usually, two independent raters give ratings about a single subject)

Selected Results

- From the same PROC FREQ statement specified previously, we also get the following output

Kappa Statistics

Statistic	Value	ASE	95% Confidence Limits
Simple Kappa	0.3428	0.0313	0.2815 0.4041
Weighted Kappa	0.3942	0.0314	0.3326 0.4559

Sample Size = 539

- Here, the weighted Kappa indicates “poor” agreement, but allowing for some disagreement (weighting) does improve the overall agreement
- Poor agreement should not be a surprise, since the test for symmetry suggested levels of support for exercise differed by type of person (church v. regular friend)

SAS Notes

- SAS uses the scores assigned to each variable in the calculation of the sampling weights
- Linear transformations of scores yield the same weighted kappa
- Non linear transformations yield different weighted kappa scores
- The same slides pertaining to the selection of scores for ordinal measures of association apply to the selection of scores for weighted kappa

Kappa Extensions

- If you have a stratification variable, you can calculate a stratum-specific kappa (κ_w)
- There is a “pooled” estimate of kappa that is calculated as

$$\widehat{\kappa}_{pooled} = \frac{\sum_{w=1} \frac{\widehat{\kappa}_w}{V_m(\widehat{\kappa}_w)}}{\sum_{w=1} \frac{1}{V_m(\widehat{\kappa}_w)}}$$

- and

$$V_m(\widehat{\kappa}_w) = \frac{A + B - C}{n(1 - p_e)^2}$$

- where

$$A = \sum_{i=1}^k \pi_{ij} [1 - (\pi_{i+} + \pi_{+i})(1 - \widehat{\kappa})]^2$$

$$B = (1 - \widehat{\kappa})^2 \sum_{i \neq j} \pi_{ij} (\pi_{+i} + \pi_{j+})^2$$

$$C = [\widehat{\kappa} - p_e(1 - \widehat{\kappa})]^2$$

$$p_e = \sum_{i=1}^k \pi_{i+} \pi_{+i}$$

General Notes

- We have been concerned with “square tables”
- That is, the rows and columns have the same levels (descriptions and counts)
- Agreement, as measured by Kappa, is similar to a correlation and represents the amount of density on the main diagonal
- McNemar’s measures marginal homogeneity...that is, does the proportion of a specific response change between responses.
- Since the two matched responses tend to be positively correlated, we gain precision (decrease the standard errors)
- That is, matched data design is more efficient and would require fewer subjects