Using **Proc Freq** in the SAS System to Perform Categorical Data Analysis

By Fairouz Makhlouf

Statistical Support Staff

Center for Information Technology

National Institute of Health

Contents:

- Scales of measurements
- Defining categorical data
- Using **PROC FREQ**
- Measures of Association
- Relative Risk & Odds Ratio
- Tests of No Association
- Stratified Analysis

- Nonparametric Analysis and Agreement Statistics
- Exercises

Scales of Measurement

- The Nominal Scale
- The Ordinal Scale
- Ratio Scale
- The Interval

Nominal Scale

- In a ***Nominal Scale*** measurement, observations are labeled and categorized into a mutually exclusive categories.
- Examples:
  - Race (African American, White, Hispanic, etc..)
  - Sex (Male, Female)

Ordinal Scale

- In an ***Ordinal Scale*** measurement, observations are ranked in terms of size or magnitude
- Examples:
  - Grades (A,B,C,D)
  - Rank (1,2,3,4)

Ratio Scale

- In a ***Ratio Scale***, ratios of numbers do reflects ratios of magnitudes. Such a scale has an absolute Zero point.
- Example:
  - The height for a group of people recorded as (60,62,59,65 and so on).
    - A value of zero represents no height
    - It is possible to use these measurements to form ratios. A person who is 72 inches is twice as tall as a 36 inch-tall person.

Interval Scale

- In an ***Interval Scale*** ,equal differences between numbers on the scale reflect equal differences in magnitude. Such a scale does not have an absolute Zero point.
- Example:
  - The height for a group of people recorded as the difference between the person height and the average height. So, +1 means that the person is 1 inch greater than the average height and -2 means that the person is 2 inches less than the average.

- In this case a zero does not indicate a zero height but that the person has an average height.
- Also the difference between +9 an +3 is 6 which means that the first person is 6 inches taller than the second one. And it does not mean that the first person is 3 times as tall as the second one (ratio comparison are not possible).

What are categorical data?

- They are data that takes on relatively distinct values and often presented as the frequency of subjects that fall into various categories.

Example 1:

- You ask 20 students in which school they are enrolled. (Notice here School is Nominal)

School	Frequency
Art and Science	6
Bussiness	5
Education	9

Example 2:

- You ask 20 students in their grades. (Notice here Grade is Ordinal)

Grade	Frequency
A	2
B	12
C	5
D	1

Example 3:

- You record the age of 20 students. (Notice here age lies in the Ratio Scale)

Age	Frequency
20	3
21	4
22	8
23	3

Use the FREQ procedure to specify two-way contingency tables

Example:

Suppose you have the following data:

student	grade	agegrp
1	a	1
2	c	2
3	b	2
4	b	2
5	d	1
6	c	2
7	b	2

Enter the data set into SAS?

```
data class;
  input id grade $ agegrp $;
cards;
1 a 1
2 c 2
3 b 2
4 b 2
5 d 1
6 c 2
7 b 2
;

proc freq data=class;
 tables agegrp*grade;
run;
```

```
TABLE OF AGEGRP BY GRADE

AGEGRP     GRADE

Frequency,
Percent  ,
Row Pct  ,
Col Pct  ,a       ,b       ,c       ,d       ,  Total
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
1        ,      1 ,      0 ,      0 ,      1 ,      2
         ,  14.29 ,   0.00 ,   0.00 ,  14.29 ,  28.57
         ,  50.00 ,   0.00 ,   0.00 ,  50.00 ,
         , 100.00 ,   0.00 ,   0.00 , 100.00 ,
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
2        ,      0 ,      3 ,      2 ,      0 ,      5
         ,   0.00 ,  42.86 ,  28.57 ,   0.00 ,  71.43
         ,   0.00 ,  60.00 ,  40.00 ,   0.00 ,
         ,   0.00 , 100.00 , 100.00 ,   0.00 ,
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
Total           1        3        2        1        7
            14.29    42.86    28.57    14.29   100.00
```

Example:

Suppose you have the following table

	Party Identification		
Race	Democrat	Independent	Republican
White	341	105	405
Black	103	15	11

Enter the data set into SAS?

```
PROC FORMAT;
  VALUE RACE 1='WHITE'
             2='BLACK';
  VALUE P_IDENT 1='DEMOCRAT'
                2='INDEPENDENT'
                3='REPUBLICAN';
RUN;

DATA PARTY;
  INPUT RACE PARTY_ID COUNT;
  FORMAT RACE RACE. PARTY_ID P_IDENT.;
CARDS;
1 1 341
1 2 105
1 3 405
2 1 103
2 2  15
2 3  11
;
PROC PRINT DATA=PARTY;
RUN;
```

The SAS System

Obs	RACE	PARTY_ID	COUNT
1	WHITE	DEMOCRAT	341
2	WHITE	INDEPENDENT	105
3	WHITE	REPUBLICAN	405
4	BLACK	DEMOCRAT	103
5	BLACK	INDEPENDENT	15
6	BLACK	REPUBLICAN	11

Using Proc Freq create a Two-Way contingency table with Race to represent the rows and Party_Id to represent the Columns?

```
PROC FREQ DATA=PARTY;
   TABLES RACE*PARTY_ID;
   WEIGHT COUNT;
RUN;
```

```
RACE      PARTY_ID

Frequency,
Percent  ,
Row Pct  ,
Col Pct  ,DEMOCRAT,INDEPEND,REPUBLIC,  Total
         ,        ,ENT     ,AN      ,
ffffffffˆffffffffˆffffffffˆffffffffˆ
WHITE    ,    341 ,    105 ,    405 ,    851
         ,  34.80 ,  10.71 ,  41.33 ,  86.84
         ,  40.07 ,  12.34 ,  47.59 ,
         ,  76.80 ,  87.50 ,  97.36 ,
ffffffffˆffffffffˆffffffffˆffffffffˆ
BLACK    ,    103 ,     15 ,     11 ,    129
         ,  10.51 ,   1.53 ,   1.12 ,  13.16
         ,  79.84 ,  11.63 ,   8.53 ,
         ,  23.20 ,  12.50 ,   2.64 ,
ffffffffˆffffffffˆffffffffˆffffffffˆ
Total         444      120      416      980
            45.31    12.24    42.45   100.00
```

Measures of Association

- They are Statistics that indicates the strength of association between two variables

Example:

Use **Proc Freq** to obtain measures of association between Race and Party_Id?

To do so we need to specify the ***measures*** option.

```
PROC FREQ DATA=PARTY / noprint measures;
   TABLES RACE*PARTY_ID;
   WEIGHT COUNT;
RUN;
```

The SAS System
The FREQ Procedure

Statistics for Table of RACE by PARTY_ID

Statistic	Value	ASE
Gamma	-0.7080	0.0527
Kendall's Tau-b	-0.2733	0.0237
Stuart's Tau-c	-0.2023	0.0213
Somers' D C|R	-0.4426	0.0350
Somers' D R|C	-0.1688	0.0179
Pearson Correlation	-0.2847	0.0245
Spearman Correlation	-0.2847	0.0246
Lambda Asymmetric C|R	0.1194	0.0478
Lambda Asymmetric R|C	0.0000	0.0000
Lambda Symmetric	0.0962	0.0391
Uncertainty Coefficient C|R	0.0470	0.0088
Uncertainty Coefficient R|C	0.1183	0.0203
Uncertainty Coefficient Symmetric	0.0673	0.0122

Sample Size = 980

```
Gamma
Kendall's Tau-b
Stuart's Tau-c

Somers' D C|R
Somers' D R|C

Pearson Correlation
Spearman Correlation
```

These measures require an ordinal measurements scale for both variables

```
Lambda Asymmetric C|R
Lambda Asymmetric R|C
Lambda Symmetric

Uncertainty Coefficient C|R
Uncertainty Coefficient R|C
Uncertainty Coefficient Symmetric
```

If at least one of the variables is nominally scaled

How does Proc Freq order the levels?

- **Numeric variables**: The levels are in numeric order
- **Character variables**: By default the levels are alphabetically ordered.

If a variable is from an ordinal scale and the levels in the contingency table are not in the correct order then the first seven of the measurements of association are wrong.

- To order them differently
  - Specify the levels in the correct order in the DATA step and then Use the **ORDER=DATA** option in **Proc Freq**
    - Note: Sometimes this dangerous because **PROC Freq** ignores observations with a zero frequency and so it might be difficult to get the data sorted in the correct order.
  - Create a numeric variable and assign a format to the levels of the numeric variable. In this case **Proc Freq** prints the formatted value in the table using the default **ORDER=INTERNAL**

Example:

- For the variables Softness (Soft,Medium and Hard) and Temp (Low, Medium and High), use Proc Freq to Create a 2-way contingency table.
  - Do not show the overall and the column percentages. (I.e. use nocol and nopersent options)
  - Request measures of association

```
title 'DETERGENT PREFERENCE STUDY';
data deterg;
  input softness $ temp $ count @@;

cards;
Soft Low 29 Soft Med 22 Soft High 11
Med  Low 14 Med  Med 35 Med  High 19
Hard Low  9 Hard Med 22 Hard High 31
;

proc freq data=deterg;
tables softness*temp/ nocol nopercent measures;
weight count;
title 'Variable Levels Ordered Incorrectly';
run;
```

```
            Variable Levels Ordered Incorrectly
TABLE OF SOFTNESS BY TEMP

SOFTNESS     TEMP

Frequency,
Row Pct  ,High    ,Low     ,Med     ,  Total
ffffffffˆffffffffˆffffffffˆffffffffˆ
Hard     ,     31 ,      9 ,     22 ,     62
         ,  50.00 ,  14.52 ,  35.48 ,
ffffffffˆffffffffˆffffffffˆffffffffˆ
Med      ,     19 ,     14 ,     35 ,     68
         ,  27.94 ,  20.59 ,  51.47 ,
ffffffffˆffffffffˆffffffffˆffffffffˆ
Soft     ,     11 ,     29 ,     22 ,     62
         ,  17.74 ,  46.77 ,  35.48 ,
ffffffffˆffffffffˆffffffffˆffffffffˆ
Total          61       52       79    192
```

STATISTICS FOR TABLE OF SOFTNESS BY TEMP

Statistic	Value	ASE
Gamma	0.175	0.091
Kendall's Tau-b	0.120	0.063
Stuart's Tau-c	0.119	0.063
Somers' D C|R	0.119	0.063
Somers' D R|C	0.121	0.063
Pearson Correlation	0.153	0.069
Spearman Correlation	0.141	0.072
Lambda Asymmetric C|R	0.142	0.084
Lambda Asymmetric R|C	0.218	0.069
Lambda Symmetric	0.181	0.069
Uncertainty Coefficient C|R	0.062	0.024
Uncertainty Coefficient R|C	0.061	0.024
Uncertainty Coefficient Symmetric	0.062	0.024

Sample Size = 192

- We need to order the variables logically . To do this we need
  - Create temporary character variables
  - Transform the values to numeric using user-created informats
  - Handle any case sensitive issues
  - Format the numeric variables with user defined format

```
proc format;
  invalue softns 'SOFT'=1 'MED'=2 'HARD'=3;
  value softness 1='Soft' 2='Medium' 3='Hard';
  invalue tempr 'LOW'=1 'MED'=2 'HIGH'=3 ;
  value temp 1='Low' 2='Medium' 3='High';
run;

data deterg2 (drop=s t);
  input s $ t $ count @@;
  softness=input(upcase(s),softns.);
  temp=input(upcase(t),tempr.);
  format softness softness. temp temp.;

cards;
Soft Low 29 Soft Med 22 Soft High 11
Med  Low 14 Med  Med 35 Med  High 19
Hard Low  9 Hard Med 22 Hard High 31
;

proc freq data=deterg2;
   tables softness*temp/ nocol nopercent measures;
   weight count;
   title 'Variable Levels Ordered correctly';
run;
```

```
Variable Levels Ordered correctly

TABLE OF SOFTNESS BY TEMP

SOFTNESS     TEMP

Frequency,
Row Pct  ,Low     ,Medium  ,High    ,  Total
ffffffffˆffffffffˆffffffffˆffffffffˆ
Soft     ,     29 ,     22 ,     11 ,     62
         ,  46.77 ,  35.48 ,  17.74 ,
ffffffffˆffffffffˆffffffffˆffffffffˆ
Medium   ,     14 ,     35 ,     19 ,     68
         ,  20.59 ,  51.47 ,  27.94 ,
ffffffffˆffffffffˆffffffffˆffffffffˆ
Hard     ,      9 ,     22 ,     31 ,     62
         ,  14.52 ,  35.48 ,  50.00 ,
ffffffffˆffffffffˆffffffffˆffffffffˆ
Total          52       79       61      192
```

STATISTICS FOR TABLE OF SOFTNESS BY TEMP

Statistic	Value	ASE
Gamma	0.450	0.085
Kendall's Tau-b	0.307	0.062
Stuart's Tau-c	0.304	0.062
Somers' D C|R	0.305	0.062
Somers' D R|C	0.309	0.062
Pearson Correlation	0.339	0.068
Spearman Correlation	0.338	0.068
Lambda Asymmetric C|R	0.142	0.084
Lambda Asymmetric R|C	0.218	0.069
Lambda Symmetric	0.181	0.069
Uncertainty Coefficient C|R	0.062	0.024
Uncertainty Coefficient R|C	0.061	0.024
Uncertainty Coefficient Symmetric	0.062	0.024

Sample Size = 192

Note:

- The first seven Statistics depend on the correct ordering or the levels of the ordinal variables. So the value of those statistic change.
- The last six statistics remain the same since they do not take the order into account.
- A table that summarize both variables are given in the next slide.

Statistic	Correct Ordering Value	Incorrect Ordering Value
Gamma	0.450	0.228
Kendall's Tau-b	0.307	0.155
Stuart's Tau-c	0.304	0.154
Somers' D C|R	0.305	0.154
Somers' D R|C	0.309	0.157
Pearson Correlation	0.339	0.188
Spearman Correlation	0.338	0.181
Lambda Asymmetric C|R	0.142	0.142
Lambda Asymmetric R|C	0.218	0.218
Lambda Symmetric	0.181	0.181
Uncertainty Coefficient C|R	0.062	0.062
Uncertainty Coefficient R|C	0.061	0.061
Uncertainty Coefficient Symmetric	0.062	0.062

Using the "order=data" option

```
data deterg;
  input softness $ temp $ count @@;

cards;
Soft Low 29 Soft Med 22 Soft High 11
Med  Low 14 Med  Med 35 Med  High 19
Hard Low  9 Hard Med 22 Hard High 31
;

proc freq data=deterg order=data;
  tables softness*temp/ nocol nopercent measures;
  weight count;
  title 'Variable Levels Ordered Correctly using order=data';
run;
```

```
Variable Levels Ordered Correctly using order=data

TABLE OF SOFTNESS BY TEMP

SOFTNESS     TEMP

Frequency,
Row Pct  ,Low     ,Med     ,High    ,  Total
ffffffffˆffffffffˆffffffffˆffffffffˆ
Soft     ,     29 ,     22 ,     11 ,     62
         ,  46.77 ,  35.48 ,  17.74 ,
ffffffffˆffffffffˆffffffffˆffffffffˆ
Med      ,     14 ,     35 ,     19 ,     68
         ,  20.59 ,  51.47 ,  27.94 ,
ffffffffˆffffffffˆffffffffˆffffffffˆ
Hard     ,      9 ,     22 ,     31 ,     62
         ,  14.52 ,  35.48 ,  50.00 ,
ffffffffˆffffffffˆffffffffˆffffffffˆ
Total          52       79       61      192
```

What would go wrong when we use the following data with “order=data” option?

```
data deterg3;
  input softness $ temp $ count @@;
cards;
Soft Low  0 Soft Med 22 Soft High 11
Med  Low 14 Med  Med 35 Med  High 19
Hard Low  9 Hard Med 22 Hard High 31
;

proc freq data=deterg3 order=data;
   tables softness*temp/ nocol nopercent measures;
   weight count;
   title 'Variable Levels NOT Ordered Correctly';
   title2 'using order=data';
run;
```

```
Variable Levels NOT Ordered Correctly
using order=data

TABLE OF SOFTNESS BY TEMP

SOFTNESS     TEMP

Frequency,
Row Pct   ,Med      ,High     ,Low      ,  Total
ffffffffˆffffffffˆffffffffˆffffffffˆ
Soft      ,      22 ,      11 ,       0 ,     33
          ,   66.67 ,   33.33 ,    0.00 ,
ffffffffˆffffffffˆffffffffˆffffffffˆ
Med       ,      35 ,      19 ,      14 ,     68
          ,   51.47 ,   27.94 ,   20.59 ,
ffffffffˆffffffffˆffffffffˆffffffffˆ
Hard      ,      22 ,      31 ,       9 ,     62
          ,   35.48 ,   50.00 ,   14.52 ,
ffffffffˆffffffffˆffffffffˆffffffffˆ
Total            79        61        23    163
```

Wrong Order!

Values of Measures Of Association

Example: Perfect Positive Association

```
data per_pos;
  input var1 var2 count;
cards;
1 1 10
2 2 13
3 3 20
;

proc freq data=per_pos;
   tables var1*var2 / nocol nopercent measures;
   weight count;
run;
```

```
 Perfect Positive Association
TABLE OF VAR1 BY VAR2

VAR1      VAR2

Frequency,
Row Pct  ,       1,       2,       3,  Total
ffffffffˆffffffffˆffffffffˆffffffffˆ
       1 ,      10 ,       0 ,       0 ,      10
         , 100.00 ,    0.00 ,    0.00 ,
ffffffffˆffffffffˆffffffffˆffffffffˆ
       2 ,       0 ,      13 ,       0 ,      13
         ,    0.00 , 100.00 ,    0.00 ,
ffffffffˆffffffffˆffffffffˆffffffffˆ
       3 ,       0 ,       0 ,      20 ,      20
         ,    0.00 ,    0.00 , 100.00 ,
ffffffffˆffffffffˆffffffffˆffffffffˆ
Total           10       13       20       43
```

STATISTICS FOR TABLE OF VAR1 BY VAR2

Statistic	Value	ASE
Gamma	1.000	0.000
Kendall's Tau-b	1.000	0.000
Stuart's Tau-c	0.957	0.046
Somers' D C|R	1.000	0.000
Somers' D R|C	1.000	0.000
Pearson Correlation	1.000	0.000
Spearman Correlation	1.000	0.000
Lambda Asymmetric C|R	1.000	0.000
Lambda Asymmetric R|C	1.000	0.000
Lambda Symmetric	1.000	0.000
Uncertainty Coefficient C|R	1.000	0.000
Uncertainty Coefficient R|C	1.000	0.000
Uncertainty Coefficient Symmetric	1.000	0.000

Sample Size = 43

- All measures except Stuart’s tau-c are 1.
- Stuart’s Tau-c takes into account the size of the table.

Example: Perfect negative Association

```
data per_neg;
  input var1 var2 count;
cards;
1 3 10
2 2 13
3 1 20
;
proc freq data=per_neg;
   tables var1*var2 / nocol nopercent measures;
   weight count;
   title ' Perfect Negative Association';
run;
```

Perfect Negative Association

TABLE OF VAR1 BY VAR2

```
VAR1      VAR2

Frequency,
Row Pct  ,       1,       2,       3,  Total
ffffffffˆffffffffˆffffffffˆffffffffˆ
       1 ,      0 ,      0 ,     10 ,     10
         ,   0.00 ,   0.00 , 100.00 ,
ffffffffˆffffffffˆffffffffˆffffffffˆ
       2 ,      0 ,     13 ,      0 ,     13
         ,   0.00 , 100.00 ,   0.00 ,
ffffffffˆffffffffˆffffffffˆffffffffˆ
       3 ,     20 ,      0 ,      0 ,     20
         , 100.00 ,   0.00 ,   0.00 ,
ffffffffˆffffffffˆffffffffˆffffffffˆ
Total          20       13       10       43
```

STATISTICS FOR TABLE OF VAR1 BY VAR2

Statistic	Value	ASE
Gamma	-1.000	0.000
Kendall's Tau-b	-1.000	0.000
Stuart's Tau-c	-0.957	0.046
Somers' D C|R	-1.000	0.000
Somers' D R|C	-1.000	0.000
Pearson Correlation	-1.000	0.000
Spearman Correlation	-1.000	0.000
Lambda Asymmetric C|R	1.000	0.000
Lambda Asymmetric R|C	1.000	0.000
Lambda Symmetric	1.000	0.000
Uncertainty Coefficient C|R	1.000	0.000
Uncertainty Coefficient R|C	1.000	0.000
Uncertainty Coefficient Symmetric	1.000	0.000

Sample Size = 43

- All of the first seven measurements has a value -1 except Stuart's Tau-c .
- The last six have a value of 1 since they do not take into account the ordinal nature of the data.
- Stuart's Tau-c takes into account the size of the table. (check the formula)

Example: No Association

```
data no_assoc;
  input var1 var2 count@@;
cards;
1 1 20 2 1 12 3 1 6
1 2 10 2 2  6 3 2 3
;
proc freq data=no_assoc;
   tables var1*var2 / nocol nopercent measures;
   weight count;
   title 'No Association';
run;
```

```
No Association

TABLE OF VAR1 BY VAR2

VAR1      VAR2

Frequency,
Row Pct  ,       1,       2,  Total
ffffffffˆffffffffˆffffffffˆ
       1 ,      20 ,      10 ,      30
         ,   66.67 ,   33.33 ,
ffffffffˆffffffffˆffffffffˆ
       2 ,      12 ,       6 ,      18
         ,   66.67 ,   33.33 ,
ffffffffˆffffffffˆffffffffˆ
       3 ,       6 ,       3 ,       9
         ,   66.67 ,   33.33 ,
ffffffffˆffffffffˆffffffffˆ
Total          38       19       57
```

Notice that the proportion in each row is the same

```
STATISTICS FOR TABLE OF VAR1 BY VAR2

Statistic                                  Value        ASE
ffffffffffffffffffffffffffffffffffffffffffffffffffffffff
Gamma                                      0.000       0.245
Kendall's Tau-b                            0.000       0.127
Stuart's Tau-c                             0.000       0.130

Somers' D C|R                              0.000       0.109
Somers' D R|C                              0.000       0.147

Pearson Correlation                        0.000       0.132
Spearman Correlation                       0.000       0.132

Lambda Asymmetric C|R                      0.000       0.000
Lambda Asymmetric R|C                      0.000       0.000
Lambda Symmetric                           0.000       0.000

Uncertainty Coefficient C|R                0.000       0.000
Uncertainty Coefficient R|C                0.000       0.000
Uncertainty Coefficient Symmetric          0.000       0.000

Sample Size = 57
```

All correlation are 0.

Ordinal Data Measures

- If both variables are from an ordinal scale
  - measures of association are valid for all sample sizes
  - measures of association are between -1 and 1 where
    - +1 means perfect positive association
    - -1 means perfect negative association
    - 0 means no association

- ASE: Asymptotic Standard error requires a large sample size to be valid
  - at lease 80% of the cells have an expected frequency of 5 or greater
  - a sample size of 25 for each degree of freedom in the Pearson Chi-square statistics
- If the sample size is large then the measure of association is approximately normal. And a 95% confident interval could be calculated be
  - measure of association ± (1.96 * ASE)
- If you specify the CL option in the table statement SAS will calculate the 95% confidence interval for you.

Example:

```
proc freq data=deterg2;
   tables softness*temp
            /nocol nopercent measures CL;
   weight count;
   title 'CL option';
run;
```

STATISTICS FOR TABLE OF SOFTNESS BY TEMP

Statistic	Value	ASE	95% Confidence Bounds	
Gamma	0.450	0.085	0.282	0.617
Kendall's Tau-b	0.307	0.062	0.185	0.428
Stuart's Tau-c	0.304	0.062	0.183	0.426
Somers' D C|R	0.305	0.062	0.183	0.426
Somers' D R|C	0.309	0.062	0.187	0.431
Pearson Correlation	0.339	0.068	0.206	0.471
Spearman Correlation	0.338	0.068	0.205	0.471
Lambda Asymmetric C|R	0.142	0.084	-0.022	0.305
Lambda Asymmetric R|C	0.218	0.069	0.083	0.353
Lambda Symmetric	0.181	0.069	0.045	0.318
Uncertainty Coefficient C|R	0.062	0.024	0.015	0.110
Uncertainty Coefficient R|C	0.061	0.024	0.014	0.108
Uncertainty Coefficient Symmetric	0.062	0.024	0.015	0.109

Sample Size = 192

Pearson Vs. Other Ordinal measures

- Pearson correlation
  - Measures the linear association
  - uses the actual variable values unless in the SCORES= option you specify SCORES=RANK. In this case the Pearson correlation is identical to the Spearman correlation
- Spearman uses rank scores for the row and the columns scores

- Gamma, Tau-b, Tau-c, and Somers'D
  - use the relative ordering of the levels to measure positive and negative association.
  - They produce measures that are similar so it is hard to recommend one over the other
  - they are computed by using a ratio where the numerator is the same for every measure but they differ in the denominator where adjustments for ties, sample size or table size takes place.

Concordance & Discordance

- The pair is called concordant if the subject ranking higher on the variable X also ranks higher on the variable Y
- The pair is called discordant if the subject ranking higher on the variable X ranks lower on the variable Y
- The pair is called tied if the subject has the same classification on X and/or Y

A and D are concordant

B and C are discordant

A and B are tied

A and C are tied

B and D are tied

C and D are tied

		Y	
		Low	High
X	Low	A	B
	High	C	D

For the following table let us find out the number of concordant, discordant and tied pairs.

		Education	
		Low	High
Income	Low	40	5
	High	10	35

Consider a pair of subjects in the cell (Low Income ,Low Education) and the other in (High Income, High Education). This pair is concordance because the second subject is ranked higher than the first on both income and education.

- Concordant: 40*35=1400
- Discordant: 10*5 = 50
- Tied: (40*5 + 40*10 + 5*35 + 10*35) = 1125
- The number of Concordant pairs is greater than the number of Discordant pairs which suggest that there is a tendency for low income to occur with low education and high income with high education.

Gamma Statistics

- The Gamma Statistics is calculated by

$$\frac{P-Q}{P+Q}$$

  - P is twice the number of concordance pairs
  - Q is twice the number of discordance pairs

– In our example it is 0.931

- So, the estimator of gamma
  - is based on the number of concordant and discordant pairs of observations
  - It ignores tied pairs
  - If the two variables are independent then the gamma estimator tends to equal to 0.
  - Its ranges from -1 to 1.
  - It is appropriate when both variables lie on an ordinal scale.

Kendall's Tau-b Statistics

- The Kendall's Tau-b Statistics is calculated by

$$\frac{P-Q}{\sqrt{w_r w_c}}$$

where

$$w_r = n^2 - \sum_i n_{i.}^2$$

$$w_c = n^2 - \sum_j n_{.j}^2$$

  - $n_{.j}$ Is the jth column total
  - $n_{i.}$ Is the ith row total

– So in our example it is 0.671

- So, the estimator of Kendall's tau-b
  - is based on the number of concordant and discordant pairs of observations taking into consideration the tied pairs.
  - Its range is from -1 to 1.
  - It is appropriate when both variables lie on an ordinal scale.

Stuart's Tau-c Statistics

- The Stuart's Tau-c Statistics is calculated by

$$\frac{P-Q}{n^2(m-1)/m}$$

  - P is twice the number of concordance pairs
  - Q is twice the number of discordance pairs
  - n overall total
  - m = min(# of rows, # of columns)
  - So in our example it is 0.667

- So, the estimator of Stuart's tau-c
  - is based on the number of concordant and discordant pairs of observations taking into consideration the tied pairs and the table size.
  - Its range is from -1 to 1.
  - It is appropriate when both variables lie on an ordinal scale.

Somers' D(C|R) Statistics

- The Somers' D(C|R) Statistics is calculated by

$$\frac{P-Q}{w_r}$$

  - So in our example it is 0.667
  - C|R denotes
    - that the row variable is regarded as independent variable
    - that the column variable is regarded as dependent variable

- So, the estimator of Somers' D(C|R)
  - is based on the number of concordant and discordant pairs of observations taking into consideration the tied pairs on the independent variable.
  - It ranges from -1 to 1.
  - It is appropriate when both variables lie on an ordinal scale.

Somers' D(R|C) Statistics

- The Somers' D(R|C) Statistics is calculated by

$$\frac{P - Q}{w_c}$$

  – So in our example it is 0.675

Example:

```
data Con_Dis;
  input income educ count;
cards;
1 1 40
1 2 5
2 1 10
2 2 35
;
proc freq data=Con_Dis;
 tables income* educ/measures;
 weight count;
 title 'Concordance and Discordance';
run;
```

Concordance and Discordance

STATISTICS FOR TABLE OF INCOME BY EDUC

Statistic	Value	ASE
Gamma	0.931	0.040
Kendall's Tau-b	0.671	0.077
Stuart's Tau-c	0.667	0.078
Somers' D C|R	0.667	0.078
Somers' D R|C	0.675	0.077
Pearson Correlation	0.671	0.077
Spearman Correlation	0.671	0.077
Lambda Asymmetric C|R	0.625	0.103
Lambda Asymmetric R|C	0.667	0.081
Lambda Symmetric	0.647	0.090
Uncertainty Coefficient C|R	0.361	0.091
Uncertainty Coefficient R|C	0.357	0.091
Uncertainty Coefficient Symmetric	0.359	0.091

- The most popular statistics are Spearman and Kendall's Tau-b
  - Use Spearman if you are looking at ranks
  - Use Kendall's Tau-b if you are at concordance Vs discordance
- If you have Dependant and Independent variables use Somers' D C|R or Somers' D R|C

Nominal Data Measures

- If at least one of the variables in the two-way table is nominally scaled we use
  - the lambda coefficients
  - the uncertainty coefficients
- Each of these measures has three forms
  - C|R column variable is treated as dependent and the row variable is fixed
  - R|C row variable is treated as the dependent and the column is fixed
  - Symmetric no dependent and independent variables
- They range from 0 to 1.

Lambda Coefficient

- It is defined as
  - The proportional reduction of the expected number of predictive errors gained by the knowledge of the independent variable. **Or,**
  - The probable improvement in predicting the column variable given that the one has the knowledge of the row variable.
- The symmetric lambda is computed by averaging the two asymmetric lambdas

Uncertainty Coefficient

- It is the proportion of the entropy (uncertainty) in the dependent variable that is explained by the independent variable

Relative Risk & Odds Ratio

- They are mesures of association that are used in special cases of a **2-by-2** contingency table
  - **Relative risk** is used in ***cohort*** study where the researcher enroll study subjects and then follow overtime to observe the results
  - **Odds Ratio** is used in ***retrospective*** study where the researcher looks at the outcome and then looks back to the patient's chart to determine which actions were related to the event. *(This is used when you have rare disease)*

Example *: Relative Risk

To investigate the association between exercise and cardiovascular disease, the researcher considered one group of 500 patients who do not exercise regularly and another group of 500 patients who do. The researcher will follow the patients over a 5 year period and observe how many develop cardiovascular disease. The table is given

		Cardiovascular disease?	
		Yes	No
Exercise Regularly	Yes	35	465
	No	50	450

- What is the risk of developing cardiovascular disease for someone who exercise regularly?
  - Pr(Event|Exercise) = 35/500 = 0.07
- What is the risk of developing cardiovascular disease for someone who does not exercise regularly?
  - Pr(Exercise) = 50/500 = 0.1
- The relative risk is the ratio of the two risks
  - RR = 0.07/0.1 = 0.7
- So, exercising regularly can protect you from cardiovascular disease.

		Cardiovascular disease?		
		Yes	No	Total
Exercise Regularly	Yes	35	465	500
	No	50	450	500
	Total	85	915	1000

Values of Relative Risk

- Values of relative risk ranges between 0 and infinity.
  - A value of 1 indicates that the risks are the same, so there is no association between the study factor and the outcome.
  - A value between 0 and 1 indicates that the experimental group is at smaller risk of the event.
  - A value greater than 1 indicates that the study factor places the individual at an increased risk for the outcome

Using Proc FREQ to calculate the Relative Risk

```
data relrisk;
 input Exer $ cardio $ count;
cards;
1-Yes 1-Yes 35
1-Yes 2-No  465
2-No  1-Yes 50
2-No  2-No  450
;

Proc Freq data= relrisk;
tables Exer*cardio/nocol norow nopercent measures;
weight count;
run;
```

```
TABLE OF CARDIO BY EXER

CARDIO     EXER

Frequency,1-Yes   ,2-No    ,  Total
ffffffffˆffffffffˆffffffffˆ
1-Yes    ,     35 ,    465 ,    500
ffffffffˆffffffffˆffffffffˆ
2-No     ,     50 ,    450 ,    500
ffffffffˆffffffffˆffffffffˆ
Total          85      915     1000
```

```
 Estimates of the Relative Risk (Row1/Row2)

                                                  95%
Type of Study                Value       Confidence Bounds
ffffffffffffffffffffffffffffffffffffffffffffffffffffffff
Case-Control                 0.677        0.432      1.063
Cohort (Col1 Risk)           0.700        0.463      1.059
Cohort (Col2 Risk)           1.033        0.995      1.073


Sample Size = 1000
```

As you can see a 95% confidence Interval is computed.

To compute another confidence interval you can use the ALPHA= option in the TABLES statement.

The sample size must be large for the confidence interval to be accurate.

Important Note

- Proc FREQ expects the variables to be ordered so that the cell corresponding to (Yes,Yes) is in cell (1,1) so either,
  - Use the format/informat statement in SAS
  - Or do as I did in this example

Example *: Odds Ratio

To investigate the association between Heavy Drinking and Bladder Cancer, the researcher started with 60 patients who were recently diagnosed with bladder cancer and control group of 60 patients who were free of bladder cancer. Then the researcher asks each patients about their past Alcohol consumption and then classify them as heavy drinker or not. The table is given

		Bladder Cancer?		
		Yes	No	Total
Heavy Drinker ?	Yes	34	24	58
	No	26	36	62
	Total	60	60	120

- The odds are defined as
  - Pr(Event happen)/Pr(Event does not happen)
- For our example
  - Pr(heavy drinker for cancer group) = 34/60
  - Pr(not heavy drinker for cancer group) = 26/60
  - So, the odds of being heavy drinker for the Cancer group = 34/26 = 1.31
  - Pr(heavy drinker not for cancer group) = 24/60
  - Pr(not heavy drinker not for cancer group) = 36/60
  - So, the odds of being heavy drinker not for the Cancer group = 24/36 = 0.67

- Hence, the odds ratio is
  - OR = 1.31/0.67 = 1.96
  - So, the odds are 1.96 times higher that someone with bladder cancer was heavy drinker compared to someone without bladder cancer.

- **Values properties for odds ratio are similar to the relative risk values**

Using Proc FREQ to calculate the Odds Ratio

```
data oddratio;
  input drink $ cancer $ count;
cards;
1-Yes 1-Yes 34
1-Yes 2-No  24
2-No  1-Yes 26
2-No  2-No  36
;

Proc Freq data= oddratio;
tables drink*cancer/nocol norow nopercent measures;
weight count;
run;
```

```
TABLE OF DRINK BY CANCER

DRINK     CANCER

Frequency,1-Yes   ,2-No    ,  Total
ffffffffˆffffffffˆffffffffˆ
1-Yes    ,     34 ,     24 ,     58
ffffffffˆffffffffˆffffffffˆ
2-No     ,     26 ,     36 ,     62
ffffffffˆffffffffˆffffffffˆ
Total          60       60     120
```

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Bounds	
Case-Control	1.962	0.949	4.055
Cohort (Col1 Risk)	1.398	0.971	2.012
Cohort (Col2 Risk)	0.713	0.491	1.034

Sample Size = 120

Tests of No Association

- Chi-square test
- Fisher's exact test

Chi-Square test

- The chi-squared test, tests the following hypothesis
  - H_0: Row and column variables are independent(or no association between the row variable and the column variable)
- It ***measures*** the strength of the ***evidence*** that an association exist
- ***Do not measure*** the ***strength*** of an association
- Depends on the sample size
- It compares the observed to the expected

Using Proc Freq to get the Chi-square statistic

```
PROC FREQ DATA=PARTY;
   TABLES RACE*PARTY_ID/nocol norow nopercent chisq;
   WEIGHT COUNT;
RUN;
```

```
TABLE OF RACE BY PARTY_ID

RACE      PARTY_ID

Frequency,DEMOCRAT,INDEPEND,REPUBLIC,  Total
         ,        ,ENT     ,AN      ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
WHITE    ,    341 ,    105 ,    405 ,    851
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
BLACK    ,    103 ,     15 ,     11 ,    129
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
Total         444      120      416      980


STATISTICS FOR TABLE OF RACE BY PARTY_ID

Statistic                     DF     Value        Prob
ƒƒƒƒƒƒƒƒƒ
```

- The Likelihood-ratio chi-square and the Pearson chi-square are asymptotically significant
- The mantel-Haenszel chi-square requires both variables to be ordinal.
- The Chi-square option produces three measures of association based on the chi-square
  - phi coefficient
  - contingency coefficient
  - Cramer's V

Phi coefficient	is in the range (-1,1) for a 2-by-2 tables and in the range of (0,sqrt(min(r-1,c-1))) for larger tables, where r: the number of rows c: the number of column
Contingency Coefficient	is in the range (0,1)
Cramer's V	is in the range (-1,1) for a 2-by-2 table and (0,1) for larger tables

Values further away from 0 indicate the presence of relatively strong association

Using Proc Freq to get the Expected frequency and the contribution of each cell to the total Chi-square statistics

```
PROC FREQ DATA=PARTY;
   TABLES RACE*PARTY_ID/nocol norow nopercent chisq
                        expected cellchi2;
   WEIGHT COUNT;
RUN;
```

$$\text{Expected} = \frac{\text{Row Total} * \text{Column Total}}{\text{Total}}$$

$$\text{chisqof cell} = \frac{(\text{Observed - Expected})^2}{\text{Expected}}$$

```
TABLE OF RACE BY PARTY_ID

RACE            PARTY_ID

Frequency      ,
Expected       ,
Cell Chi-Square,DEMOCRAT,INDEPEND,REPUBLIC,  Total
               ,        ,ENT     ,AN      ,
ffffffffffffffˆffffffffˆffffffffˆffffffffˆ
WHITE          ,    341 ,    105 ,    405 ,    851
               , 385.56 ,  104.2 , 361.24 ,
               , 5.1488 , 0.0061 , 5.3008 ,
ffffffffffffffˆffffffffˆffffffffˆffffffffˆ
BLACK          ,    103 ,     15 ,     11 ,    129
               , 58.445 , 15.796 , 54.759 ,
               , 33.966 , 0.0401 , 34.969 ,
ffffffffffffffˆffffffffˆffffffffˆffffffffˆ
Total               444      120      416      980
```

Using Proc Freq to Output the Expected frequency and the contribution of each cell to the total Chi-square statistics to a Dataset

```
PROC FREQ DATA=PARTY;
   TABLES RACE*PARTY_ID/ out=freqout noprint outexpect
                         outpct sparse;
   WEIGHT COUNT;
RUN;
```

OBS	RACE	PARTY_ID	COUNT	EXPECTED	PERCENT	PCT_ROW	PCT_COL
1	WHITE	DEMOCRAT	341	385.555	34.7959	40.0705	76.8018
2	WHITE	INDEPENDENT	105	104.204	10.7143	12.3384	87.5000
3	WHITE	REPUBLICAN	405	361.241	41.3265	47.5911	97.3558
4	BLACK	DEMOCRAT	103	58.445	10.5102	79.8450	23.1982
5	BLACK	INDEPENDENT	15	15.796	1.5306	11.6279	12.5000
6	BLACK	REPUBLICAN	11	54.759	1.1224	8.5271	2.6442

Statistics	Scale of Measurement	Alternative Hypothesis	DF
Pearson Chi-square	No restriction	General Association	(r-1)*(c-1)
Likelihood ratio chi-square	No restriction	General Association	(r-1)*(c-1)
Mantel-Haenzel chi-square	Ordinal	Linear Association	1

- The Mantel-Haenszel chi-square is computed like the Pearson chi-square but it requires that the levels of the ordinally scaled variables to be in the correct order in the contingency table.
- The Mantel-Haenszel chi-square is sensitive only to linear association
  - all of the MH statistics' power is concentrated toward the objective of detecting linear association and is contained in 1 df. (I.e MH is more powerful in detecting linear association)
  - The power of the general association statistics is dispersed over a greater number of alternative and degrees of freedom.

Example: General Association

```
data general;
  input var1 var2 count;
cards;
1 3 15
2 2 15
2 4 15
3 1 15
3 5 15
;
proc freq data=general;
   tables var1*var2 / nocol nopercent norow chisq;
   weight count;
   title 'Not linear Association';
run;
```

Not linear Association

TABLE OF VAR1 BY VAR2

VAR1 VAR2

```
Frequency,       1,       2,       3,       4,       5,  Total
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
       1 ,      0 ,      0 ,     15 ,      0 ,      0 ,     15
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
       2 ,      0 ,     15 ,      0 ,     15 ,      0 ,     30
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
       3 ,     15 ,      0 ,      0 ,      0 ,     15 ,     30
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
Total          15       15       15       15       15       75
```

```
STATISTICS FOR TABLE OF VAR1 BY VAR2

Statistic                     DF      Value        Prob
ffffffffffffffffffffffffffffffffffffffffffffffffffffff
Chi-Square                     8    150.000       0.001
Likelihood Ratio Chi-Square    8    158.238       0.001
Mantel-Haenszel Chi-Square     1      0.000       1.000
Phi Coefficient                       1.414
Contingency Coefficient               0.816
Cramer's V                            1.000

Sample Size = 75
WARNING: 33% of the cells have expected counts less than 5.
Chi-Square may not be a valid test.
```

Significant general association but not linear.

Example: Linear Association

```
data lin_asso;
  input var1 var2 count@@;
cards;
1 1 3 2 1 7 3 1 10
1 2 4 2 2 7 3 2 5
1 4 6 2 4 7 3 4 4
1 5 10 2 5 7 3 5 3
;
proc freq data=lin_asso;
   tables var1*var2 / nocol norow nopercent chisq;
   weight count;
   title 'Detecting Linear Association with MH';
run;
```

```
Detecting Linear Association with MH

TABLE OF VAR1 BY VAR2

VAR1      VAR2

Frequency,       1,       2,       4,       5,  Total
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
       1 ,      3 ,      4 ,      6 ,     10 ,     23
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
       2 ,      7 ,      7 ,      7 ,      7 ,     28
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
       3 ,     10 ,      5 ,      4 ,      3 ,     22
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
Total          20       16       17       20       73
```

```
STATISTICS FOR TABLE OF VAR1 BY VAR2

Statistic                          DF      Value         Prob
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
Chi-Square                          6      8.859        0.182
Likelihood Ratio Chi-Square         6      8.888        0.180
Mantel-Haenszel Chi-Square          1      7.963        0.005
Phi Coefficient                            0.348
Contingency Coefficient                    0.329
Cramer's V                                 0.246

Sample Size = 73
```

Significant linear association but not general.

Problem: Small Sample Size

```
data per_neg;
  input var1 var2 count;
cards;
1 3 10
2 2 13
3 1 20
;
proc freq data=per_neg;
   tables var1*var2 / nocol nopercent norow chisq;
   weight count;
   title 'Problem';
run;
```

```
Problem

TABLE OF VAR1 BY VAR2

VAR1      VAR2

Frequency,       1,       2,       3,  Total
ffffffffˆffffffffˆffffffffˆffffffffˆ
       1 ,      0 ,      0 ,     10 ,     10
ffffffffˆffffffffˆffffffffˆffffffffˆ
       2 ,      0 ,     13 ,      0 ,     13
ffffffffˆffffffffˆffffffffˆffffffffˆ
       3 ,     20 ,      0 ,      0 ,     20
ffffffffˆffffffffˆffffffffˆffffffffˆ
Total          20       13       10       43
```

```
STATISTICS FOR TABLE OF VAR1 BY VAR2

Statistic                       DF     Value        Prob
ffffffffffffffffffffffffffffffffffffffffffffffffffffff
Chi-Square                       4    86.000       0.001
Likelihood Ratio Chi-Square      4    90.894       0.001
Mantel-Haenszel Chi-Square       1    42.000       0.001
Phi Coefficient                        1.414
Contingency Coefficient                0.816
Cramer's V                             1.000

Sample Size = 43

WARNING: 67% of the cells have expected counts less than 5.
 Chi-Square may not be a valid test.
```

Chi-square test is not appropriate for small samples because of the Normality assumption and the large sample theory.

Exact Test for 2-by-2 Tables

- For a 2-by-two table
  - Fisher's Exact Test tests the null hypothesis of no association between the row variable and the column variable
  - The alternative Hypothesis for Fisher's test that there is some kind of association. In this case general association is the same as linear association
  - The test is valid for any sample size

- Fisher's Exact test produces a p-value that represent the probability of observing a table at least as extreme as the one actually observed given that there is no association between the variables
- Assumption for Fisher's Exact test is that the row and the column marginal totals are fixed instead of being random variables. If the assumption is not satisfied by the sampling process, it is usually satisfied by conditioning on the observed marginal

Using Proc Freq to produce Fisher's Exact test

```
Data fisher;
  input married $ car $ count;
cards;
Y Y 4
Y N 1
N Y 1
N N 4
;

proc Freq data=fisher order=data;
   tables married*Car/ nocol nopercent norow chisq;
   weight count;
   title 'Fisher"s Exact Test';
run;
```

A random Sample of ten people is classified according to their marital status and whether or not they drive an American car

```
Fisher"s Exact Test

TABLE OF MARRIED BY CAR

MARRIED     CAR

Frequency,Y        ,N        ,  Total
ffffffffˆffffffffˆffffffffˆ
Y         ,      4 ,      1 ,      5
ffffffffˆffffffffˆffffffffˆ
N         ,      1 ,      4 ,      5
ffffffffˆffffffffˆffffffffˆ
Total            5        5       10
```

```
STATISTICS FOR TABLE OF MARRIED BY CAR

Statistic                          DF      Value         Prob
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
Chi-Square                          1      3.600        0.058
Likelihood Ratio Chi-Square         1      3.855        0.050
Continuity Adj. Chi-Square          1      1.600        0.206
Mantel-Haenszel Chi-Square          1      3.240        0.072
Fisher's Exact Test (Left)                               0.996
                    (Right)                              0.103
                    (2-Tail)                             0.206
Phi Coefficient                            0.600
Contingency Coefficient                    0.514
Cramer's V                                 0.600

Sample Size = 10
WARNING: 100% of the cells have expected counts less than 5.
Chi-Square may not be a valid test.
```

- As you can see Proc Freq produces
  - the left-tail test which is used when testing

    H_0: p(American|married) = p(American|Not married)

    vs.

    H_a: p(American|married) < p(American|Not married)
  - The right tail test

    H_0: p(American|married) = p(American|Not married)

    vs.

    H_a: p(American|married) > p(American|Not married)
  - the two tail test

    H_0: p(American|married) = p(American|Not married)

    vs.

    H_a: p(American|married) ≠ p(American|Not married)

Exact Test for r-by-c Tables

- For a r-by-c table
  - Fisher's Exact Test tests the null hypothesis of no association between the row variable and the column variable
  - The alternative Hypothesis for Fisher's test that there is some kind of association.
  - This is only two test hypothesis. (since the alternative hypothesis is not linear association)

Using Proc Freq to produce Fisher's Exact test for r-by-c tables

```
data exact;
  input var1 var2 count@@;
cards;
1 1 3 2 1 0 3 1 2
1 0 4 2 2 4 3 2 5
1 4 6 2 4 2 3 4 4
1 5 10 2 5 3 3 5 3
;
proc freq data=exact;
   tables var1*var2 / nocol norow nopercent exact;
   weight count;
   title 'Exact Test for r-by-c table';
run;
```

```
Exact Test for r-by-c table

TABLE OF VAR1 BY VAR2

VAR1      VAR2

Frequency,       0,       1,       2,       4,       5,  Total
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
       1 ,      4 ,      3 ,      0 ,      6 ,     10 ,     23
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
       2 ,      0 ,      0 ,      4 ,      2 ,      3 ,      9
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
       3 ,      0 ,      2 ,      5 ,      4 ,      3 ,     14
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
Total           4        5        9       12       16       46
```

```
STATISTICS FOR TABLE OF VAR1 BY VAR2

Statistic                     DF      Value        Prob
ffffffffffffffffffffffffffffffffffffffffffffffffffffff
Chi-Square                     8     15.750       0.046
Likelihood Ratio Chi-Square    8     21.702       0.005
Mantel-Haenszel Chi-Square     1      0.185       0.667
Fisher's Exact Test (2-Tail)                       0.021
Phi Coefficient                       0.585
Contingency Coefficient               0.505
Cramer's V                            0.414

Sample Size = 46
WARNING: 87% of the cells have expected counts less than 5.
Chi-Square may not be a valid test.
```

Stratified Analysis

Consider the following study:

Odds Ratio = 1.74

	Head Colds	No Head Colds
Poor Temp Control	70	150
Good Temp Control	55	205

So, the poor temp control is about 1.74 likely to cause head colds as the good temp control.

Now, divide the data into two groups, smokers and nonsmokers we get the following tables

Smokers

	Head Colds	No Head Colds
Poor Temp Control	30	50
Good Temp Control	20	55

Odds Ratio 1.65

None-Smokers

	Head Colds	No Head Colds
Poor Temp Control	40	100
Good Temp Control	35	150

Odds Ratio 1.714

The previous process of separating the objects into separate groups is called ***stratification***. And the process of analyzing the contingency tables from two or more strata is called ***stratified analysis***.

Stratification Using Proc FREQ

```
Data study;
  input T_Contrl $ Group: $10. Cold $ count;
cards;
1-POOR SMOKERS 1-YES 30
1-POOR SMOKERS 2-NO  50
1-POOR NONSMOKERS 1-YES 40
1-POOR NONSMOKERS 2-NO  100
2-GOOD SMOKERS 1-YES    20
2-GOOD SMOKERS 2-NO     55
2-GOOD NONSMOKERS 1-YES 35
2-GOOD NONSMOKERS 2-NO  150
;
proc freq data=study;
  tables group * t_contrl * cold/nocol norow nopercent
                                  chisq measures;
  weight count;
  title ' Stratification';
run;
```

```
 Stratification

TABLE 1 OF T_CONTRL BY COLD
CONTROLLING FOR GROUP=NONSMOKERS

T_CONTRL     COLD

Frequency,1-YES   ,2-NO    ,  Total
ffffffffˆffffffffˆffffffffˆ
1-POOR   ,     40 ,    100 ,    140
ffffffffˆffffffffˆffffffffˆ
2-GOOD   ,     35 ,    150 ,    185
ffffffffˆffffffffˆffffffffˆ
Total          75      250      325

STATISTICS FOR TABLE 1 OF T_CONTRL BY COLD
CONTROLLING FOR GROUP=NONSMOKERS

Statistic                     DF     Value        Prob
ffffffffffffffffffffffffffffffffffffffffffffffffffff
Chi-Square                     1     4.183       0.041
Likelihood Ratio Chi-Square    1     4.151       0.042
Continuity Adj. Chi-Square     1     3.657       0.056
Mantel-Haenszel Chi-Square     1     4.170       0.041
Fisher's Exact Test (Left)                       0.985
                    (Right)                      0.028
                    (2-Tail)                     0.046
Phi Coefficient                      0.113
Contingency Coefficient              0.113
Cramer's V                           0.113
```

Statistic	Value	ASE
Gamma	0.263	0.123
Kendall's Tau-b	0.113	0.056
Stuart's Tau-c	0.095	0.047
Somers' D C|R	0.097	0.048
Somers' D R|C	0.133	0.065
Pearson Correlation	0.113	0.056
Spearman Correlation	0.113	0.056
Lambda Asymmetric C|R	0.000	0.000
Lambda Asymmetric R|C	0.036	0.061
Lambda Symmetric	0.023	0.040
Uncertainty Coefficient C|R	0.012	0.012
Uncertainty Coefficient R|C	0.009	0.009
Uncertainty Coefficient Symmetric	0.010	0.010

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Bounds	
Case-Control	1.714	1.020	2.882
Cohort (Col1 Risk)	1.510	1.015	2.246
Cohort (Col2 Risk)	0.881	0.777	0.999

Sample Size = 325

```
 Stratification

TABLE 2 OF T_CONTRL BY COLD
CONTROLLING FOR GROUP=SMOKERS

T_CONTRL     COLD

Frequency,1-YES   ,2-NO    ,  Total
ffffff
```

Statistic	Value	ASE
Gamma	0.245	0.164
Kendall's Tau-b	0.116	0.079
Stuart's Tau-c	0.108	0.074
Somers' D C|R	0.108	0.074
Somers' D R|C	0.124	0.085
Pearson Correlation	0.116	0.079
Spearman Correlation	0.116	0.079
Lambda Asymmetric C|R	0.000	0.000
Lambda Asymmetric R|C	0.067	0.132
Lambda Symmetric	0.040	0.080
Uncertainty Coefficient C|R	0.011	0.015
Uncertainty Coefficient R|C	0.010	0.013
Uncertainty Coefficient Symmetric	0.010	0.014

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Bounds	
Case-Control	1.650	0.833	3.267
Cohort (Col1 Risk)	1.406	0.879	2.250
Cohort (Col2 Risk)	0.852	0.685	1.060

Sample Size = 155

Cochran-Mantel-Haenszel Statistics

- It test the following hypothesis

  H_0: no association between the row variable and the column variable in any of the strata

  vs.

  H_a:there is association between the row variable and the column variable in at least one strata

- Under the null Hypothesis the distribution of CMH statistics is approximately chi-square if the sample size is sufficiently large.

- CMH measures components from each stratum, add the components across the strata, and then computes the chi-square statistics based on these sums.
- CMH requires
  - a large overall sample size
  - it does not require a large sample size in each stratum
  - it requires fixed row and column marginal totals in each stratum which can be obtained by conditional distribution arguments

- CMH is not appropriate if the association changes dramatically from stratum to stratum. I.e. it does not assume stratum*row*column interaction
- The reason that CMH may fail to reject the null hypothesis
  - There is no association anywhere
  - Positive patterns of association in some strata are being offset by negative pattern by other stratum
  - The sample size is not large enough

- CMH statistic does not change when the row (or column) scores are changed in the following way
  - The same number is added to each score
  - Each score is multiplied by the same numbers

Three Alternative Hypothesis

- PROC FREQ computes three different types of CMH statistics depending on the hypothesis it is testing
- The three types are
  - Nonzero correlation
  - row mean scores differ
  - general association

Type 1: Nonzero Correlation

- It is called **Mantel-Haenszel Statistics**
- This statistics is sensitive to linear association.
- It has DF = 1
- It is only appropriate when both the row and the column variables are ordinal.
- For one table this is the same as Mantel-Haenszel chi-square obtained by the CHISQ option

Type 2: Row mean score differ

- It is called the ***ANOVA*** statistics or the ***mean score statistics***.
- This statistics requires the column variable to be ordinally scaled
- It has DF = r -1 where,
  - r is number of rows
- For one table this is the same as analysis of variance and if rank score is used it is identical to Kruskal-Wallis test.

Type 3: General Association

- It is a generalization of the Pearson chi-square.
- DF = (r - 1) * (c - 1)
  - r is number of rows
  - c number of columns
- It does not require an ordinal scale for either variable.
- For one table this is the same as Pearson chi-square

Option In PROC FREQ to calculate CMH statistics

- CMH all three CMH statistics
- CMH1 correlation statistics
- CMH2 correlation and Anova

Example: *

- Four surgical operation for a duodenal ulcer were compared in a clinical trial
  a. Drainage and Vagotomy
  b. 25% Resection and Vagotomy
  c. 50% Resection and Vagotomy
  d. 75% Resection
- The response variable is the severity (None, Slight, and Moderate) of an undesirable complication called dumping syndrome
- Is there an association between the type of operation and the severity of the response?

```
data syndrome;
  input hospital $ trtment $ severity $ count @@;
cards;
1 a None 23  1 a Slight  7 1 a Moderate 2
1 b None 23  1 b Slight 10 1 b Moderate 5
1 c None 20  1 c Slight 13 1 c Moderate 5
1 d None 24  1 d Slight 10 1 d Moderate 6
2 a None 18  2 a Slight  6 2 a Moderate 1
2 b None 18  2 b Slight  6 2 b Moderate 2
2 c None 13  2 c Slight 13 2 c Moderate 2
2 d None  9  2 d Slight 15 2 d Moderate 2
3 a None  8  3 a Slight  6 3 a Moderate 3
3 b None 12  3 b Slight  4 3 b Moderate 4
3 c None 11  3 c Slight  6 3 c Moderate 2
3 d None  7  3 d Slight  7 3 d Moderate 4
4 a None 12  4 a Slight  9 4 a Moderate 1
4 b None 15  4 b Slight  3 4 b Moderate 2
4 c None 14  4 c Slight  8 4 c Moderate 3
4 d None 13  4 d Slight  6 4 d Moderate 4
;
proc freq data=syndrome order=data;
  tables hospital*trtment*severity/all;
  weight count;
run;
```

The All Option is the same as using CMH, CHISQ MEASURES

Mantal-Haenzel Test

TABLE 1 OF TRTMENT BY SEVERITY
CONTROLLING FOR HOSPITAL=1

TRTMENT SEVERITY

```
Frequency,
Percent  ,
Row Pct  ,
Col Pct  ,None    ,Slight  ,Moderate,  Total
ffffffffˆffffffffˆffffffffˆffffffffˆ
a        ,     23 ,      7 ,      2 ,     32
         ,  15.54 ,   4.73 ,   1.35 ,  21.62
         ,  71.88 ,  21.88 ,   6.25 ,
         ,  25.56 ,  17.50 ,  11.11 ,
ffffffffˆffffffffˆffffffffˆffffffffˆ
b        ,     23 ,     10 ,      5 ,     38
         ,  15.54 ,   6.76 ,   3.38 ,  25.68
         ,  60.53 ,  26.32 ,  13.16 ,
         ,  25.56 ,  25.00 ,  27.78 ,
ffffffffˆffffffffˆffffffffˆffffffffˆ
c        ,     20 ,     13 ,      5 ,     38
         ,  13.51 ,   8.78 ,   3.38 ,  25.68
         ,  52.63 ,  34.21 ,  13.16 ,
         ,  22.22 ,  32.50 ,  27.78 ,
ffffffffˆffffffffˆffffffffˆffffffffˆ
d        ,     24 ,     10 ,      6 ,     40
         ,  16.22 ,   6.76 ,   4.05 ,  27.03
         ,  60.00 ,  25.00 ,  15.00 ,
         ,  26.67 ,  25.00 ,  33.33 ,
ffffffffˆffffffffˆffffffffˆffffffffˆ
Total          90       40       18      148
            60.81    27.03    12.16   100.00
```

STATISTICS FOR TABLE 1 OF TRTMENT BY SEVERITY
CONTROLLING FOR HOSPITAL=1

Statistic	DF	Value	Prob
Chi-Square	6	3.421	0.754
Likelihood Ratio Chi-Square	6	3.553	0.737
Mantel-Haenszel Chi-Square	1	1.568	0.210
Phi Coefficient		0.152	
Contingency Coefficient		0.150	
Cramer's V		0.108	

Statistic	Value	ASE
Gamma	0.135	0.110
Kendall's Tau-b	0.086	0.071
Stuart's Tau-c	0.083	0.068
Somers' D C|R	0.074	0.060
Somers' D R|C	0.102	0.083
Pearson Correlation	0.103	0.079
Spearman Correlation	0.098	0.080
Lambda Asymmetric C|R	0.000	0.000
Lambda Asymmetric R|C	0.028	0.044
Lambda Symmetric	0.018	0.029
Uncertainty Coefficient C|R	0.013	0.014
Uncertainty Coefficient R|C	0.009	0.009
Uncertainty Coefficient Symmetric	0.010	0.011

Sample Size = 148

Mantal-Haenzel Test

TABLE 2 OF TRTMENT BY SEVERITY
CONTROLLING FOR HOSPITAL=2

TRTMENT SEVERITY

```
Frequency,
Percent  ,
Row Pct  ,
Col Pct  ,None    ,Slight  ,Moderate,  Total
ƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
a        ,     18 ,      6 ,      1 ,     25
         ,  17.14 ,   5.71 ,   0.95 ,  23.81
         ,  72.00 ,  24.00 ,   4.00 ,
         ,  31.03 ,  15.00 ,  14.29 ,
ƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
b        ,     18 ,      6 ,      2 ,     26
         ,  17.14 ,   5.71 ,   1.90 ,  24.76
         ,  69.23 ,  23.08 ,   7.69 ,
         ,  31.03 ,  15.00 ,  28.57 ,
ƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
c        ,     13 ,     13 ,      2 ,     28
         ,  12.38 ,  12.38 ,   1.90 ,  26.67
         ,  46.43 ,  46.43 ,   7.14 ,
         ,  22.41 ,  32.50 ,  28.57 ,
ƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
d        ,      9 ,     15 ,      2 ,     26
         ,   8.57 ,  14.29 ,   1.90 ,  24.76
         ,  34.62 ,  57.69 ,   7.69 ,
         ,  15.52 ,  37.50 ,  28.57 ,
ƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
Total          58       40        7      105
            55.24    38.10     6.67   100.00
```

STATISTICS FOR TABLE 2 OF TRTMENT BY SEVERITY
CONTROLLING FOR HOSPITAL=2

Statistic	DF	Value	Prob
Chi-Square	6	10.921	0.091
Likelihood Ratio Chi-Square	6	11.174	0.083
Mantel-Haenszel Chi-Square	1	7.063	0.008
Phi Coefficient		0.323	
Contingency Coefficient		0.307	
Cramer's V		0.228	

Statistic	Value	ASE
Gamma	0.385	0.120
Kendall's Tau-b	0.253	0.081
Stuart's Tau-c	0.243	0.078
Somers' D C|R	0.216	0.069
Somers' D R|C	0.297	0.096
Pearson Correlation	0.261	0.092
Spearman Correlation	0.286	0.091
Lambda Asymmetric C|R	0.128	0.097
Lambda Asymmetric R|C	0.091	0.095
Lambda Symmetric	0.105	0.082
Uncertainty Coefficient C|R	0.061	0.035
Uncertainty Coefficient R|C	0.038	0.022
Uncertainty Coefficient Symmetric	0.047	0.027

Sample Size = 105

Mantal-Haenzel Test

TABLE 3 OF TRTMENT BY SEVERITY
CONTROLLING FOR HOSPITAL=3

TRTMENT SEVERITY

```
Frequency,
Percent  ,
Row Pct  ,
Col Pct  ,None    ,Slight  ,Moderate,  Total
ƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
a        ,      8 ,      6 ,      3 ,     17
         ,  10.81 ,   8.11 ,   4.05 ,  22.97
         ,  47.06 ,  35.29 ,  17.65 ,
         ,  21.05 ,  26.09 ,  23.08 ,
ƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
b        ,     12 ,      4 ,      4 ,     20
         ,  16.22 ,   5.41 ,   5.41 ,  27.03
         ,  60.00 ,  20.00 ,  20.00 ,
         ,  31.58 ,  17.39 ,  30.77 ,
ƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
c        ,     11 ,      6 ,      2 ,     19
         ,  14.86 ,   8.11 ,   2.70 ,  25.68
         ,  57.89 ,  31.58 ,  10.53 ,
         ,  28.95 ,  26.09 ,  15.38 ,
ƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
d        ,      7 ,      7 ,      4 ,     18
         ,   9.46 ,   9.46 ,   5.41 ,  24.32
         ,  38.89 ,  38.89 ,  22.22 ,
         ,  18.42 ,  30.43 ,  30.77 ,
ƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
Total          38       23       13       74
            51.35    31.08    17.57   100.00
```

STATISTICS FOR TABLE 3 OF TRTMENT BY SEVERITY
CONTROLLING FOR HOSPITAL=3

Statistic	DF	Value	Prob
Chi-Square	6	3.123	0.793
Likelihood Ratio Chi-Square	6	3.309	0.769
Mantel-Haenszel Chi-Square	1	0.156	0.693
Phi Coefficient		0.205	
Contingency Coefficient		0.201	
Cramer's V		0.145	

Statistic	Value	ASE
Gamma	0.068	0.149
Kendall's Tau-b	0.047	0.101
Stuart's Tau-c	0.047	0.103
Somers' D C|R	0.042	0.091
Somers' D R|C	0.052	0.112
Pearson Correlation	0.046	0.117
Spearman Correlation	0.052	0.117
Lambda Asymmetric C|R	0.000	0.104
Lambda Asymmetric R|C	0.056	0.060
Lambda Symmetric	0.033	0.069
Uncertainty Coefficient C|R	0.022	0.023
Uncertainty Coefficient R|C	0.016	0.017
Uncertainty Coefficient Symmetric	0.019	0.020

Sample Size = 74

TABLE 4 OF TRTMENT BY SEVERITY
CONTROLLING FOR HOSPITAL=4

TRTMENT SEVERITY

Frequency, Percent , Row Pct , Col Pct	None	Slight	Moderate	Total
a	12	9	1	22
	13.33	10.00	1.11	24.44
	54.55	40.91	4.55	
	22.22	34.62	10.00	
b	15	3	2	20
	16.67	3.33	2.22	22.22
	75.00	15.00	10.00	
	27.78	11.54	20.00	
c	14	8	3	25
	15.56	8.89	3.33	27.78
	56.00	32.00	12.00	
	25.93	30.77	30.00	
d	13	6	4	23
	14.44	6.67	4.44	25.56
	56.52	26.09	17.39	
	24.07	23.08	40.00	
Total	54	26	10	90
	60.00	28.89	11.11	100.00

STATISTICS FOR TABLE 4 OF TRTMENT BY SEVERITY
CONTROLLING FOR HOSPITAL=4

Statistic	DF	Value	Prob
Chi-Square	6	5.264	0.510
Likelihood Ratio Chi-Square	6	5.501	0.481
Mantel-Haenszel Chi-Square	1	0.661	0.416
Phi Coefficient		0.242	
Contingency Coefficient		0.235	
Cramer's V		0.171	

Statistic	Value	ASE
Gamma	0.084	0.144
Kendall's Tau-b	0.054	0.093
Stuart's Tau-c	0.052	0.089
Somers' D C|R	0.046	0.079
Somers' D R|C	0.064	0.108
Pearson Correlation	0.086	0.103
Spearman Correlation	0.063	0.106
Lambda Asymmetric C|R	0.000	0.000
Lambda Asymmetric R|C	0.046	0.109
Lambda Symmetric	0.030	0.071
Uncertainty Coefficient C|R	0.034	0.027
Uncertainty Coefficient R|C	0.022	0.018
Uncertainty Coefficient Symmetric	0.027	0.022

Sample Size = 90

Mantal-Haenzel Test

SUMMARY STATISTICS FOR TRTMENT BY SEVERITY
CONTROLLING FOR HOSPITAL

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	6.340	0.012
2	Row Mean Scores Differ	3	6.590	0.086
3	General Association	6	10.598	0.102

Total Sample Size = 417

Results:

- **Correlation statistics** = 6.34 with p-value=0.012 so, there is a significant linear association for at least one of the hospitals.
- The **General association** statistics = 10.598 with p-value=0.102 so it is not significant.

Correlation Analysis by Strata

Hospital	Sample Size	Pearson Correlation	Mantel-Haenszel Chi-square	DF	Prob
1	148	0.10	1.57	1	0.21
2	105	0.26	7.06	1	0.01
3	74	0.05	0.16	1	0.69
4	90	0.09	0.66	1	0.42

So, the significant nonzero correlation is mainly from Hospital 2.

Measures of Partial Association

- Adjusted relative risk
- Adjusted odds ratio
- Breslow-Day Statistics

Adjusted Relative Risk

- It measures the average extent of association across the strata.
- They are appropriate for any sample size but the confidence interval requires overall large sample size.
- Proc Freq computes two types of adjusted relative risk
  - The Mantel-Haenszel (MH) estimator which handles zero frequencies with no difficulty.
  - The Logit estimator which does not handle zero frequencies. So 0.5 is added to each cell that contains zero.

Adjusted Odds Ratio

- They are estimates of the adjusted relative risk for case-control studies. Provided that the outcome of interest is rare.
- They are useful measures of association independent of relative risk and regardless of how the data were collected.

Homogeneity Of Odds Ratio

- The odds ratio are said to be homogeneous across the strata, if all the strata shows the same strength of association.
- If not then there is an interaction among the row variable, the column variable and the stratification variable.
- Breslow-Day Statistics is used to test the homogeneity of the odds ratio

Breslow-Day Statistics

- It has a chi-square statistics when the sample size is large and the null hypothesis is true.
- The df=q-1 where q is the number of contingency tables included in the calculations. (Tables with zero row or columns are not included in the calculations)
- requires a large sample size in each stratum.

Example:**

The following data are based on a study on coronary disease (Kock,Imrey, et al. 1985). Investigators were interested in whether electrocardiogram (ECG) measurement for segment depression was associated with disease status. Gender was thought to be associated with disease status, so the investigators poststratified the data into male and female group. In addition, there was an interest in examining the odds ratio.

Sex	ECG	Disease	No Disease	Total
Female	< 0.1 ST	4	11	15
Female	≥ 0.1 ST	8	10	18
Male	< 0.1 ST	9	9	18
Male	≥ 0.1 ST	21	6	27
Total		32	36	78

```
data ca;
  input gender $ ECG $ disease $ count;
  cards;
female <0.1  yes  4
female <0.1  no  11
female >=0.1 yes  8
female >=0.1 no  10
male   <0.1  yes  9
male   <0.1  no   9
male   >=0.1 yes 21
male   >=0.1 no   6
;
proc freq data=ca;
  tables gender*disease / nocol nopercent
                          norow chisq;
  tables gender*ECG*disease / nocol nopercent
                              norow all;
  weight count;
run;
```

```
TABLE OF GENDER BY DISEASE

GENDER     DISEASE

Frequency,no      ,yes     ,  Total
ffffffffˆffffffffˆffffffffˆ
female   ,     21 ,     12 ,     33
ffffffffˆffffffffˆffffffffˆ
male     ,     15 ,     30 ,     45
ffffffffˆffffffffˆffffffffˆ
Total          36       42       78


STATISTICS FOR TABLE OF GENDER BY DISEASE

Statistic                     DF     Value        Prob
fffffffffffffffffffffffffffffffffffffffffffffffffffff
Chi-Square                     1     7.035       0.008
Likelihood Ratio Chi-Square    1     7.121       0.008
Continuity Adj. Chi-Square     1     5.868       0.015
Mantel-Haenszel Chi-Square     1     6.944       0.008
Fisher's Exact Test (Left)                        0.998
                    (Right)                    7.51E-03
                    (2-Tail)                      0.011
Phi Coefficient                      0.300
Contingency Coefficient              0.288
Cramer's V                           0.300

Sample Size = 78
```

- Looking at the table gender*disease, the chi-square test statistic is 7.035 with p-value=0.008. So there is strong association between gender and disease status. Males are much more likely to have symptoms of coronary artery disease than females.

```
TABLE 1 OF ECG BY DISEASE
CONTROLLING FOR GENDER=female

ECG        DISEASE

Frequency,no       ,yes      ,  Total
ffffffffˆffffffffˆffffffffˆ
<0.1      ,     11 ,      4 ,     15
ffffffffˆffffffffˆffffffffˆ
>=0.1     ,     10 ,      8 ,     18
ffffffffˆffffffffˆffffffffˆ
Total          21        12       33
```

```
STATISTICS FOR TABLE 1 OF ECG BY DISEASE
CONTROLLING FOR GENDER=female

Statistic                        DF     Value        Prob
ffffffffffffffffffffffffffffffffffffffffffffffffffffff
Chi-Square                        1     1.117       0.290
Likelihood Ratio Chi-Square       1     1.134       0.287
Continuity Adj. Chi-Square        1     0.481       0.488
Mantel-Haenszel Chi-Square        1     1.084       0.298
Fisher's Exact Test (Left)                          0.923
                    (Right)                         0.245
                    (2-Tail)                        0.469
Phi Coefficient                         0.184
Contingency Coefficient                 0.181
Cramer's V                              0.184

 Estimates of the Relative Risk (Row1/Row2)

                                            95%
Type of Study            Value      Confidence Bounds
ffffffffffffffffffffffffffffffffffffffffffffffffffffff
Case-Control             2.200       0.504      9.611
Cohort (Col1 Risk)       1.320       0.790      2.206
Cohort (Col2 Risk)       0.600       0.224      1.607


Sample Size = 33
```

For females MH statistic = 1.084, with p-value = 0.298. The odds ratio for the females is 2.2, with 95% confidence interval that includes 1. Those females with higher ST segment depression levels had 2.2 times the odds of CA disease than females with lower levels.

```
TABLE 2 OF ECG BY DISEASE
CONTROLLING FOR GENDER=male

ECG        DISEASE

Frequency,no       ,yes      ,  Total
ffffffffˆffffffffˆffffffffˆ
<0.1     ,       9 ,       9 ,     18
ffffffffˆffffffffˆffffffffˆ
>=0.1    ,       6 ,      21 ,     27
ffffffffˆffffffffˆffffffffˆ
Total            15        30       45
```

```
STATISTICS FOR TABLE 2 OF ECG BY DISEASE
CONTROLLING FOR GENDER=male

Statistic                      DF     Value        Prob
ffffffffffffffffffffffffffffffffffffffffffffffffffffff
Chi-Square                      1     3.750       0.053
Likelihood Ratio Chi-Square     1     3.729       0.053
Continuity Adj. Chi-Square      1     2.604       0.107
Mantel-Haenszel Chi-Square      1     3.667       0.056
Fisher's Exact Test (Left)                        0.988
                    (Right)                       0.054
                    (2-Tail)                      0.105
Phi Coefficient                       0.289
Contingency Coefficient               0.277
Cramer's V                            0.289

 Estimates of the Relative Risk (Row1/Row2)

                                            95%
Type of Study           Value      Confidence Bounds
ffffffffffffffffffffffffffffffffffffffffffffffffffffff
Case-Control            3.500       0.959     12.778
Cohort (Col1 Risk)      2.250       0.968      5.230
Cohort (Col2 Risk)      0.643       0.388      1.064


Sample Size = 45
```

For males MH statistic = 3.667, with p-value = 0.056. The odds ratio for the males is 3.5, with 95% confidence interval that barley includes 1. Those males with higher ST segment depression levels had 3.5 times the odds of CA disease than males with lower levels.

```
SUMMARY STATISTICS FOR ECG BY DISEASE
CONTROLLING FOR GENDER


  Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic   Alternative Hypothesis    DF       Value      Prob
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
   1        Nonzero Correlation        1       4.503     0.034
   2        Row Mean Scores Differ     1       4.503     0.034
   3        General Association        1       4.503     0.034
```

The CMH statistic = 4.503 with p-value=0.034. So, by combining the genders, the power has been increased so that the association detected is significant at 0.05.

```
 Estimates of the Common Relative Risk (Row1/Row2)
                                                          95%
Type of Study    Method                 Value   Confidence Bounds
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
Case-Control     Mantel-Haenszel        2.847      1.083    7.482
  (Odds Ratio)   Logit                  2.859      1.081    7.565

Cohort           Mantel-Haenszel        1.641      1.039    2.594
  (Col1 Risk)    Logit                  1.525      0.983    2.365

Cohort           Mantel-Haenszel        0.630      0.411    0.965
  (Col2 Risk)    Logit                  0.634      0.405    0.993

The confidence bounds for the M-H estimates are test-based.



Breslow-Day Test for Homogeneity of the Odds Ratios

Chi-Square =   0.215           DF =   1           Prob = 0.643


Total Sample Size = 78
```

- The estimate of the common MH odds ratio is 2.847 and the logit is 2.859. The confidence interval for the logit estimator does not include 1.
- Breslow-Day test are
  - Chi-Square = 0.215 DF=1 p-value=0.643
  - It indicates no significant difference between the odds ratio for males and females.

Nonparametric Analysis and Agreement Statistics

- Test of Association
  - Wilcoxon-Mann-Whitney
  - Kruskal-Wallis

Wilcoxon-Mann-Whitney Test

- It is a non-parammetric test that tests

  H_0: the distribution of the response variable is the same for the two groups

  vs.

  H_a: That there is a location difference between the two groups

- It is used when there are two independent groups
- Ordinal scaled response variables

- For small sample size (<= 10) the distribution of the test is tabled exactly
- For large samples, an asymptotic distribution is used to find the p-value

Example:

- A credit card organization wants to compare two free-schedule methods with respect to their effect on recruiting new customers
  - Method 1: No annual fee for the card, but there is a service charge of $1.5 per month and interest charges on each cash advances from date of receipt. There are no interest payments on the purchases if the entire balance is paid in full each month.
  - Method 2: $18 annual fee for the card, but no interest payments on cash advances or purchases if the entire balance is paid in full each month
- The credit card organization randomly assigns 24 member banks to the two methods and then counts the number of new customers recruited in the first week. Is there an association between the fee schedule method and the number of new customers recruited?

Assuming normality of the data in the two groups we can use ttest as follows:

```
data credit;
  input method recruits @@;
  cards;
1 21 1 34 1 26 1 29 1 18 1 28
1 22 1 30 1 21 1 28 1 20 1 18
2 12 2 21 2 24 2 17 2 18 2 26
2 21 2 18 2 15 2 15 2 15 2 12
;
proc ttest data=credit;
  class method;
  title 'Two-Sample T-test';
run;
proc freq data=credit;
  tables method*recruits/noprint chisq cmh scores=rank;
  title 'Wilcoxon-Mann-Whitney statistics';
run;
```

```
Two-Sample T-test

TTEST PROCEDURE

Variable: RECRUITS

METHOD       N         Mean      Std Dev    Std Error
-----------------------------------------------------
     1      12  24.58333333   5.24765460   1.51486740
     2      12  17.83333333   4.44835686   1.28413002

Variances        T       DF    Prob>|T|
----------------------------------------
Unequal     3.3990     21.4      0.0027
Equal       3.3990     22.0      0.0026


For H0: Variances are equal, F' = 1.39    DF = (11,11)    Prob>F' = 0.5929
```

So,

The mean number of customers recruited for method 1 was 24.58.

The mean number of customers recruited for method 2 was 18.25.

Wilcoxon-Mann-Whitney statistics

STATISTICS FOR TABLE OF METHOD BY RECRUITS

Statistic	DF	Value	Prob
Chi-Square	12	14.000	0.301
Likelihood Ratio Chi-Square	12	19.408	0.079
MH Chi-Square (Rank Scores)	1	8.265	0.004
Phi Coefficient		0.764	
Contingency Coefficient		0.607	
Cramer's V		0.764	

Sample Size = 24
WARNING: 100% of the cells have expected counts less than 5.
Chi-Square may not be a valid test.

```
SUMMARY STATISTICS FOR METHOD BY RECRUITS


  Cochran-Mantel-Haenszel Statistics (Based on Rank Scores)

Statistic   Alternative Hypothesis    DF       Value      Prob
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
   1        Nonzero Correlation        1       8.265     0.004
   2        Row Mean Scores Differ     1       8.265     0.004
   3        General Association       12      13.417     0.339

Total Sample Size = 24
```

The MH Chi Square option in the table and statistics 1 and 2 in the second table indicates that the Wilcoxon-Mann-Whitney statistics 8.265 with p-value of 0.004.

So, there are significantly more customers are recruited with Method1 than Method 2

Kruskal-Wallis Test

- It is a non-parammetric test that tests

  H_0: the distribution of the response variable is the same for all of the groups

  vs.

  H_a: That there is a location difference among the groups

- It is used when there are three or more independent groups
- Ordinal scaled response variables

- For small sample size (<= 10 subjects and <= 5 groups) the distribution of the test is tabled exactly
- For large samples, an asymptotic distribution is used to find the p-value

Example: *

```
data kruskal;
  input group response @@;
  cards;
1 21 1 34 1 26 1 29 1 18 1 28
1 22 1 30 1 21 1 28 1 20 1 18
2 12 2 21 2 24 2 17 2 18 2 26
2 21 2 18 2 15 2 15 2 15 2 12
3 21 3 22 3 18 3 24 3 24 3 28
3 18 3 18
;

proc freq data=kruskal;
  tables group*response/noprint cmh scores=rank;
  title 'Kruskal-Wallis Test statistics';
run;
```

```
SUMMARY STATISTICS FOR GROUP BY RESPONSE


  Cochran-Mantel-Haenszel Statistics (Based on Rank Scores)

Statistic   Alternative Hypothesis    DF       Value      Prob
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
   1        Nonzero Correlation        1       2.259     0.133
   2        Row Mean Scores Differ     2       9.544     0.008
   3        General Association       24      24.253     0.447

Total Sample Size = 32
```

The 2nd CMH statistics indicates that the Kruskal-Wallis statistic is 9.544 with a p-value 0.008. So we conclude that there is a significant association between GROUP and RESPONSE.

Stratified Analysis

- Friedman's chi-square: Nonparametric nway Anova.
- McNemar's Test: Matched pairs
- Simple Kappa
- Weighted Kappa
- Symmetry Test

Friedman's chi-square

- It is used when
  - there are two or more groups
  - the response variable is ordinally scaled
  - to test no partial association between the grouping variable and the response variable after adjusting for one or more blocking factors.
  - Subjects are randomly assigned to the groups
  - Each block contains one observation from each group

- The distribution of the test statistics is exact when the sample size is small (<=3 groups or <=9 blocks). Otherwise the distribution is asymptotically distributed as chi-square.

Example: *

- Four different probes on a hardness machine are compared with respect to the readings that they give for four different specimens of a metal. For each piece of metal, the order in which the probes are used is randomly determined. Is there an association between the type of probe and the hardness reading?

Note: In this problem we can't use GLM because of normality assumption.

```
Data probe;
   input specimen probe hardness @@;
   cards;
1 1  9.3 1 2 9.4 1 3 9.2 1 4  9.7
2 1  9.4 2 2 9.3 2 3 9.4 2 4  9.6
3 1  9.6 3 2 9.8 3 3 9.5 3 4 10.0
4 1 10.0 4 2 9.9 4 3 9.7 4 4 10.2
;
proc freq data=probe;
 table specimen*probe*hardness /cmh2 scores=rank;
 title 'Friedman"s Chi-Square';
run;
```

```
Friedman"s Chi-Square

SUMMARY STATISTICS FOR PROBE BY HARDNESS
CONTROLLING FOR SPECIMEN


  Cochran-Mantel-Haenszel Statistics (Based on Rank Scores)

Statistic   Alternative Hypothesis    DF       Value      Prob
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
   1        Nonzero Correlation        1       1.969     0.161
   2        Row Mean Scores Differ     3       8.846     0.031

Total Sample Size = 16
```

So, there is a significant shift in hardness readings among the four probe types. (Friedman's chi-square is 8.846 with p-value = 0.031)

If you did not adjust for the metal Specimen we would use Kruskal-Wallis test.

```
proc freq data=probe;
 table probe*hardness / noprint cmh2
                        scores=rank;
 title 'Kruskal-Wallis test';
run;
```

```
Kruskal-Wallis test

SUMMARY STATISTICS FOR PROBE BY HARDNESS


  Cochran-Mantel-Haenszel Statistics (Based on Rank Scores)

Statistic   Alternative Hypothesis    DF       Value      Prob
ffffff
```

McNemar's Test

- It is used when 2-by-2 table contains information from matched pairs in which the response variable has two levels
- Examples of matched pairs
  - left eye and right eye measurements
  - before and after measurements

Example:

- To test the effect of anti-cigarette advertisement on people's attitudes toward smoking (either positive or negative), a 100 people were asked for their attitude toward smoking. Then they were let to see shown an anticigarette advertisement and then asked their smoking attitude. The results were as follows

After

Before

	Negative	Positive
Negative	30	10
Positive	45	15

```
Data cigar;
  input before $ after $ count;
  cards;
negative negative 30
negative positive 10
positive negative 45
positive positive 15
;

proc freq data=cigar;
 table before*after /nocol nopercent norow
                     agree;
 weight count;
 title 'McNemar"s Test for paired data';
run;
```

```
McNemar"s Test for paired data

TABLE OF BEFORE BY AFTER

BEFORE     AFTER

Frequency,negative,positive,  Total
ffffffffˆffffffffˆffffffffˆ
negative ,     30 ,     10 ,     40
ffffffffˆffffffffˆffffffffˆ
positive ,     45 ,     15 ,     60
ffffffffˆffffffffˆffffffffˆ
Total          75       25      100


STATISTICS FOR TABLE OF BEFORE BY AFTER

McNemar's Test
--------------
Statistic = 22.273       DF = 1        Prob = 0.001
```

So, there is good evidence that the anticigarette advertisement was effective to change people's attitude toward smoking.

Simple Kappa, Weighted Kappa, and Symmetry Tests

- Simple Kappa
  - It's a measure of the agreement. Its value is
    - 0 when the agreement equals that expected by chance
    - 1 when there is a perfect agreement
    - negative values occur when agreement is weaker than expected by chance
    - The stronger the agreement the higher the value
  - It is used for Nominal Data

- Weighted Kappa
  - It’s a measure of the agreement. Its value is
    - values of 0.4 and above are considered to indicate moderate agreement
    - values of 0,8 and above indicates excellent agreement
  - It is used when the categories are ordered and the seriousness of the agreement depends on the difference between the ratings.

- Symmetry Test
  - It test $p_{ij}=p_{ji}$
  - It implies marginal homogeneity
  - In a 2-by-2 table symmetry and marginal homogeneity are equivalent so McNemar's test is the same as Bowker's test.

Example: **

The following data come from a study concerning the diagnostic classification of multiple sclerosis patients. Patients from Winnipeg and News Orleans were classified into one of four diagnostic classes by Winnipeg and News Orleans neurologist

Winnipeg Neurologist

New Orleans Neurologist	**1**	**2**	**3**	**4**
1	38	5	0	1
2	33	11	3	0
3	10	14	5	6
4	3	7	3	10

```
data classify;
  input NewOrl Winnipeg count @@;
cards;
1 1 38 1 2  5 1 3 0 1 4 1
2 1 33 2 2 11 2 3 3 2 4 0
3 1 10 3 2 14 3 3 5 3 4 6
4 1  3 4 2  7 4 3 3 4 4 10
;
proc freq data=classify;
 table NewOrl*Winnipeg /nocol nopercent
                        norow  agree;
 weight count;
run;
```

```
TABLE OF NEWORL BY WINNIPEG

NEWORL     WINNIPEG

Frequency,       1,       2,       3,       4,  Total
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
       1 ,     38 ,      5 ,      0 ,      1 ,     44
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
       2 ,     33 ,     11 ,      3 ,      0 ,     47
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
       3 ,     10 ,     14 ,      5 ,      6 ,     35
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
       4 ,      3 ,      7 ,      3 ,     10 ,     23
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
Total          84       37       11       17    149
```

```
STATISTICS FOR TABLE OF NEWORL BY WINNIPEG

Test of Symmetry
----------------
Statistic = 46.749        DF = 6        Prob = 0.001


Kappa Coefficients

Statistic          Value      ASE    95% Confidence Bounds
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
Simple Kappa       0.208    0.050        0.109     0.307
Weighted Kappa     0.380    0.052        0.278     0.481


Sample Size = 149
```

- Kappa has a value of 0.208. This indicates a slight agreement. Since the confidence bounds do not contain the value 0, we can reject the hypothesis that kappa is equal to zero (no agreement) at 0.05 level of significance.
- The weighted kappa has the value 0.38. Which means that if you consider disagreement close to the diagonal less heavily than disagreement further away from the diagonals, you get higher agreement.

Summary Table

Row Variable	Column Variable	Test	Option
Nominal	Nominal	Chi-square	CHISQ
Nominal	Ordinal	Mean Score	CMH2
Ordinal	Ordinal	MH chi-square	CHISQ

For Stratified Analysis

Row Variable	Column Variable	Test	Option
Nominal	Nominal	CMH Type III	CMH
Nominal	Ordinal	CMH Type III	CMH2
Ordinal	Ordinal	CMH Type I	CMH1

Exercises

Q1. The table below contains results of a study by Mendenhall et al. (1984) to compare radiation therapy with surgery in treating cancer of the larynx.

Does Radiation therapy Control Cancer significantly better than Surgery?

	Cancer Controlled	Cancer Not Controlled
Surgery	21	2
Radiation Therapy	15	3

Q2. The table below contains results of a study by Srole et al. (1978, p.289) to describe the relationship between mental impairment and parents' socioeconomic status for a sample of residents of Manhattan.

A. Is there an association between Parents socioeconomic status and the Mental Health status.

B. Measure the extent of the association between parents' socioeconomic status and mental health.

Mental Health Status

Parents' Socioeconomic Status	Well	Mild Symptom Formation	Moderate Symptom Formation	Impaired
A (high)	64	94	58	46
B	57	94	54	40
C	57	105	65	60
D	72	141	77	94
E	36	97	54	78
F (low)	21	71	54	71

Q3. These data are from a study concerned with the evaluation of patients receiving one of two treatments in a multiple clinic trial.

	Drug A		Drug B	
Clinic	Yes	No	Yes	No
1	22	4	13	17
2	23	12	24	12
3	7	3	4	7
4	16	3	15	8

Evaluate the significance of the difference between the two drugs. Discuss the association in in the individual clinics, and evaluate the overall association.

Solution: Q1

```
proc format;
  value trtment 1 = 'Surgery'
                2 = 'Radiation';
  value result  1 = 'Controlled'
                2 = 'Not Controlled';
run;
data cancer;
  input trtment result count;
  format trtment trtment. result result.;
cards;
1 1 21
1 2  2
2 1 15
2 2  3
;
proc freq data=cancer;
  tables trtment*result / nopercent nocol
                          norow chisq;
  weight count;
run;
```

```
TABLE OF TRTMENT BY RESULT

TRTMENT     RESULT

Frequency ,Controll,Not Cont,  Total
          ,ed      ,rolled  ,
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
Surgery   ,     21 ,      2 ,     23
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
Radiation ,     15 ,      3 ,     18
ƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒˆ
Total           36        5       41


STATISTICS FOR TABLE OF TRTMENT BY RESULT

Statistic                     DF     Value        Prob
ƒƒƒƒƒƒƒƒƒ
```

Fisher's exact test is the appropriate test for determining whether there is an association in this table. The 2-tail probability is 0.638, which indicates **NON-Significance Association**. The new radiation therapy does not control larynx cancer better than surgery.

Solution: Q2

```
proc format;
  value Status   1 = 'Well'
                 2 = 'Mild'
                 3 = 'Moderate'
                 4 = 'Impaired';
run;
data mental;
  input p_status $ m_status  count @@;

  label p_status = 'Parents Socio Status'
        m_status = 'Mental Health Status';

  format m_status Status.;
cards;
A 1 64 B 1 57 C 1  57  D 1  72 E 1 36 F 1 21
A 2 94 B 2 94 C 2 105  D 2 141 E 2 97 F 2 71
A 3 58 B 3 54 C 3  65  D 3  77 E 3 54 F 3 54
A 4 46 B 4 40 C 4  60  D 4  94 E 4 78 F 4 71
;
proc freq data=mental;
   tables p_status* m_status/ nopercent nocol
                              norow chisq measures;
   weight count;
run;
```

```
TABLE OF P_STATUS BY M_STATUS

P_STATUS(Parents Socio Status)     M_STATUS(Mental Health Status)

Frequency,Well    ,Mild    ,Moderate,Impaired,  Total
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
A        ,     64 ,     94 ,     58 ,     46 ,    262
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
B        ,     57 ,     94 ,     54 ,     40 ,    245
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
C        ,     57 ,    105 ,     65 ,     60 ,    287
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
D        ,     72 ,    141 ,     77 ,     94 ,    384
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
E        ,     36 ,     97 ,     54 ,     78 ,    265
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
F        ,     21 ,     71 ,     54 ,     71 ,    217
ffffffffˆffffffffˆffffffffˆffffffffˆffffffffˆ
Total         307      602      362      389     1660
```

A. The Pearson Chi-Square Statistics is 45.985 with df=15 and a p-value of 0.001. So there is a strong evidence of an association between parents; socioeconomic and mental health status.

B. Since both variables of interest are ordinally scaled, we should use one of the measures of association that take ordinality into account. (If we assume that Mental Health Status are equally spaced, then we can also look at the measures of correlation)
All of the measures are small.

Statistic	Value	ASE
Gamma	0.154	0.025
Kendall's Tau-b	0.120	0.019
Stuart's Tau-c	0.125	0.020
Somers' D C|R	0.113	0.018
Somers' D R|C	0.128	0.021
Pearson Correlation	0.150	0.024
Spearman Correlation	0.149	0.024

Solution: Q3

```
data clinical;
  input clinic drug $ result $ count @@;
cards;
1 A Cured 22 2 A Cured 23
1 A Not    4 2 A Not   12
1 B Cured 13 2 B Cured 24
1 B Not   17 2 B Not   12
3 A Cured  7 4 A Cured 16
3 A Not    3 4 A Not    3
3 B Cured  4 4 B Cured 15
3 B Not    7 4 B Not    8
;

proc freq data=clinical;
  tables clinic*drug*result/nocol norow
                            nopercent all;
weight count;
run;
```

A stratified analysis is appropriate for this data, where interest lies in investigating the association between treatment and cure rate, adjusting for any effect of the clinics. Therefore, the analysis consist of looking at four 2-by-2 tables, one table for each clinic. The CMH statistic for partial association has the value Q(CMH)= 8.053 with 1 degrees of freedom which is significantly significant (p-value = 0.005). Thus, we can conclude that there is significant association between treatment and cure rate.

The relative risk and the odds ratios for the individual tables indicate that clinic 1 displays most of the association and is the major contributor to the summary statistics. There appears to be a little or no measurable association in clinic 2, and moderate association in clinic 3 and 4, although insignificant if considered by themselves. A clinical trial can be considered a cohort study , and thus the relative risk is an appropriate measure to examine.

References:

- *SAS Institute, Modeling Techniques for Categorical Response Data, Course Notes.
- **Maura E. Stokes, Charles S. Davis, Gary G. Koch, Categorical Data Analysis Using the SAS System, SAS Institute Inc.
- *** Agresti, A. (1990), Categorical Data Analysis, New York: John Wiley & Sons, Inc.
- **** Agresti, A. (1996), Introduction to Categorical Data Analysis, New York: John Wiley & Sons, Inc.