SPSS Program Notes
Biostatistics: A Guide to Design, Analysis, and Discovery Second Edition
by Ronald N. Forthofer, Eun Sul Lee, Mike Hernandez
Chapter 10: Analysis of Categorical Data

(Note: these program notes were developed under an older version of SPSS. The current version of IBM® SPSS® Statistics is version 22)

Program Notes Outline

Chapter 10 Formulas

Test	Test Statistic	Distribution of Test Statistic
Chi-square test using Yate's correction	$X^2_{YC} = \sum_i \sum_j \frac{\left(\left\| n_{ij} - m_{ij} \right\| - 0.5\right)^2}{m_{ij}}$	for a 2 by 2 table, chi-square with 1 degree of freedom; for an r by c table, chi-square with (r-1)(c-1) degrees of freedom
Simplified version of chi-square test using Yate's correction for a 2 by 2 contingency table	$X^2_{YC} = \frac{n\left(\left\| n_{11}n_{22} - n_{12}n_{21} \right\| - n/2\right)^2}{n_{1\bullet}\, n_{2\bullet}\, n_{\bullet 1}\, n_{\bullet 2}}$	chi-square with 1 degree of freedom
McNemar's test	$X^2_M = \frac{\left(\left\| d_1 - d_2 \right\| - 1\right)^2}{d_1 + d_2}$	for a 2 by 2 table, chi-square with 1 degree of freedom
Trend test	$X^2 = \frac{\left(\sum_{j=1}^{c} n_{\bullet j}\left(p_j - \bar{p}\right)\left(S_j - \bar{S}\right)\right)}{\bar{p}\bar{q}\sum_{j=1}^{c} n_{\bullet j}\left(S_j - \bar{S}\right)^2}$	See pages 284 to 286 in text

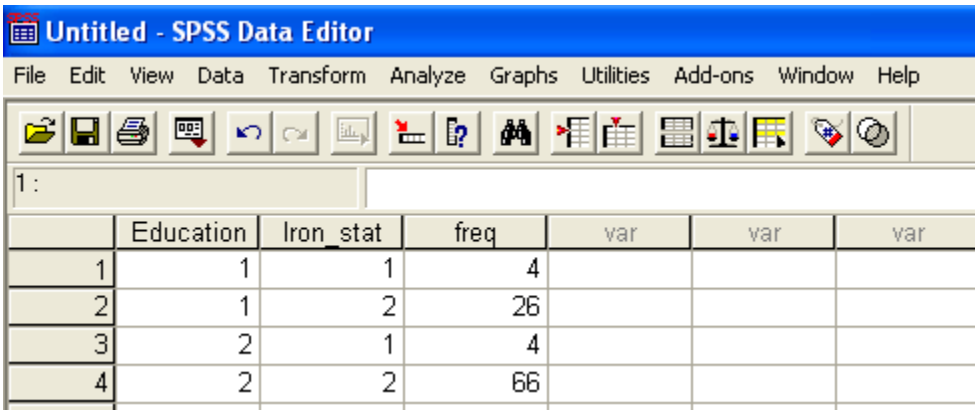
Cochran-Mantel-Haenszel chi-square test and odds ratio	See page 289 to 291 in text for formulas dealing with Cochran-Mantel-Haenszel chi-square test and odds ratio	See page 289 to 291 in text

Odds ratio estimate	$OR = \frac{n_{11}n_{22}}{n_{21}n_{12}}$
Estimate of standard error for the log of the odds ratio	$\hat{\sigma}_{\ln(OR)} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$
(1 - α/2)100 percent confidence interval for the log of the odds ratio	$\ln(OR) \pm z_{1-\alpha/2}\hat{\sigma}_{\ln(OR)}$

Program Note 10.1 – Chi-square Test for a 2 by 2 Contingency Table

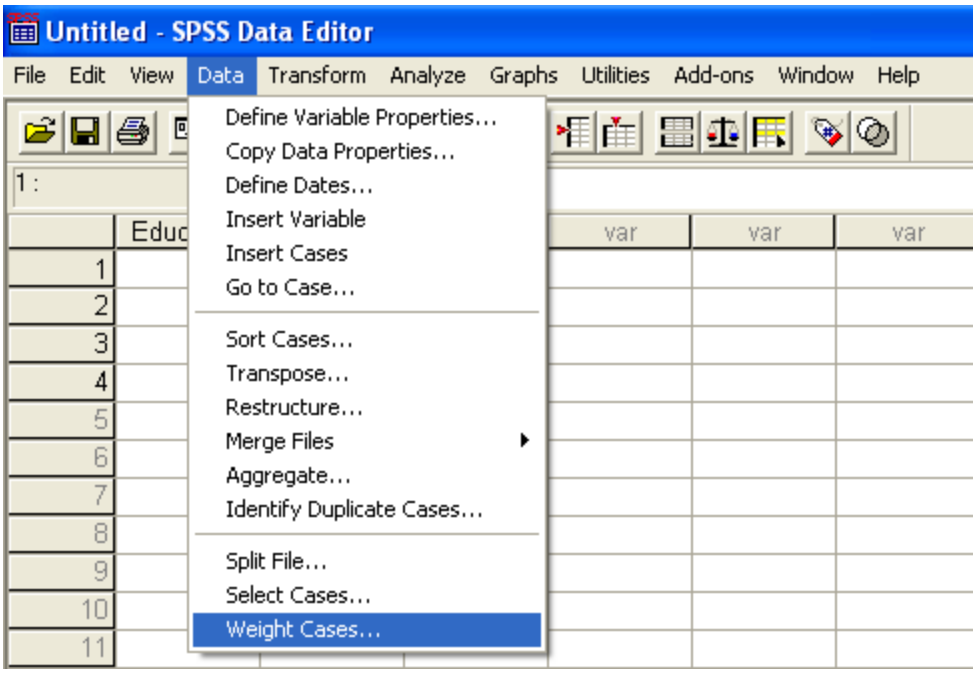
1. Uncorrected chi-square test, Yate's corrected chi-square test, and the odds ratio

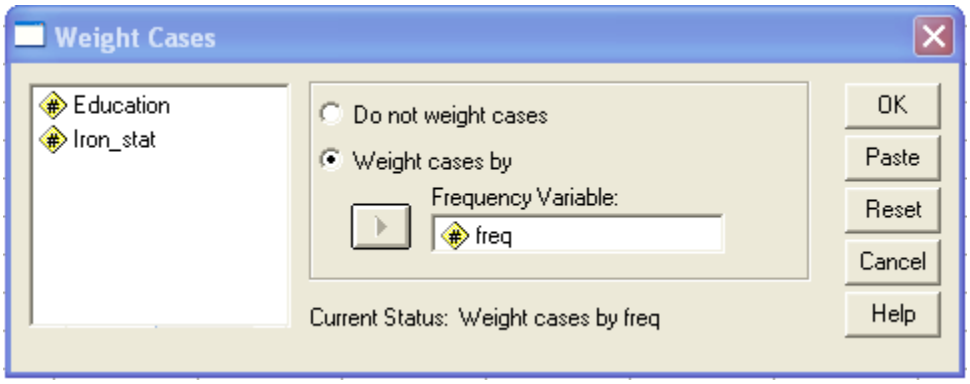
In Example 10.3, we wish to analyze the education and iron status data presented in Table 10.6 to determine if there is a statistically significant relationship between these two variables. We show the calculation for the Yate's corrected chi-square statistic. SPSS provides both the test statistic for the uncorrected Pearson chi-square statistic and Yate's corrected chi-square statistic which is referred to as **Continuity Correction**. First, we present the data from the 2 by 2 table shown in the **SPSS Data Editor** below. The variable Education has two values: '1' representing '< 12 years' and '2' representing '>= 12 years'. The variable Iron_stat also has two values: '1' representing 'Deficient' and '2' representing 'Acceptable'. Finally, the last variable freq represents the number of observations in each cell of the 2 by 2 table.



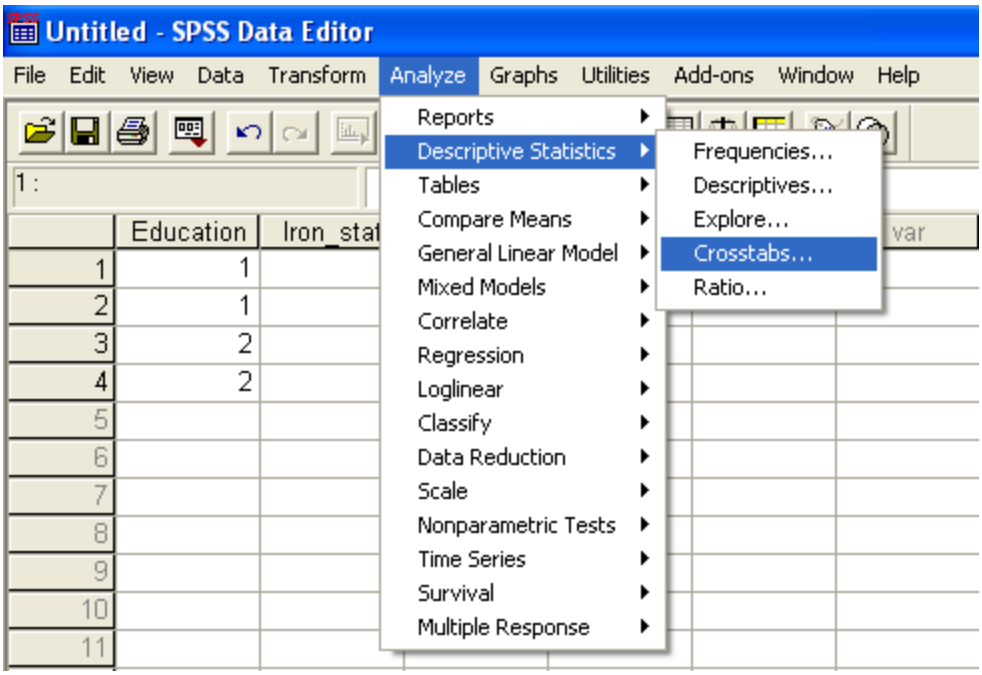
	Education	Iron_stat	freq	var	var	var
1	1	1	4			
2	1	2	26			
3	2	1	4			
4	2	2	66			

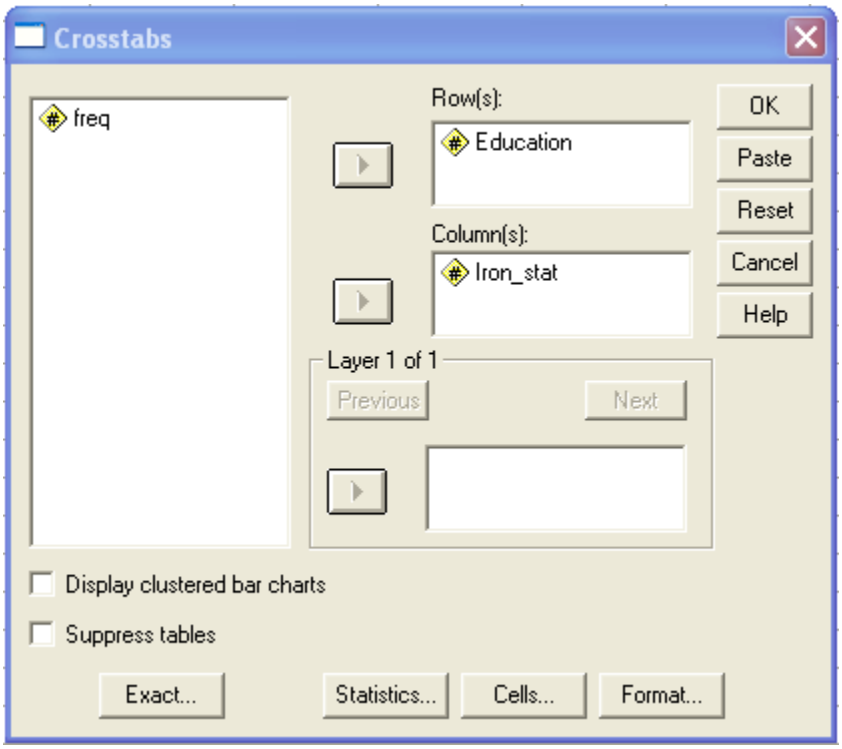
Because the data is summarized, we need to use the SPSS procedure **Data** -> **Weight Cases…** before going to the **Analyze** command. Note that weighting cases is not necessary when the data is presented in an expanded format.



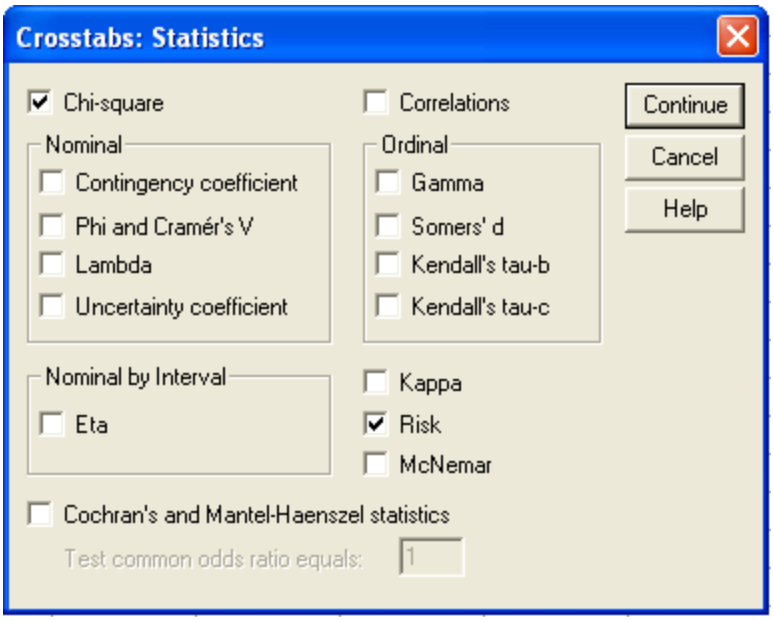


Once we weight cases using the variable freq, we can use the SPSS procedure **Analyze** -> **Descriptive Statistics** -> **Crosstabs…** to conduct the analysis and obtain the chi-square statistic of interest.





In the **Crosstabs: Statistics** window, we select both **Chi-square** and **Risk**. The **Risk** option provides the odds ratio estimate along with values of the lower and upper 95% confidence interval.



Part of the SPSS output is shown below. In the table of **Chi-Square Tests**, the first value of 1.656 corresponds to the uncorrected Pearson's chi-square statistic. The second value of 0.783 corresponds to Yate's corrected chi-square statistic. In the **Risk Estimate** table, the first value of 2.538 is the odds ratio estimate.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.656[b]	1	.198		
Continuity Correction[a]	.783	1	.376		
Likelihood Ratio	1.529	1	.216		
Fisher's Exact Test				.236	.185
Linear-by-Linear Association	1.640	1	.200		
N of Valid Cases	100				

a. Computed only for a 2x2 table

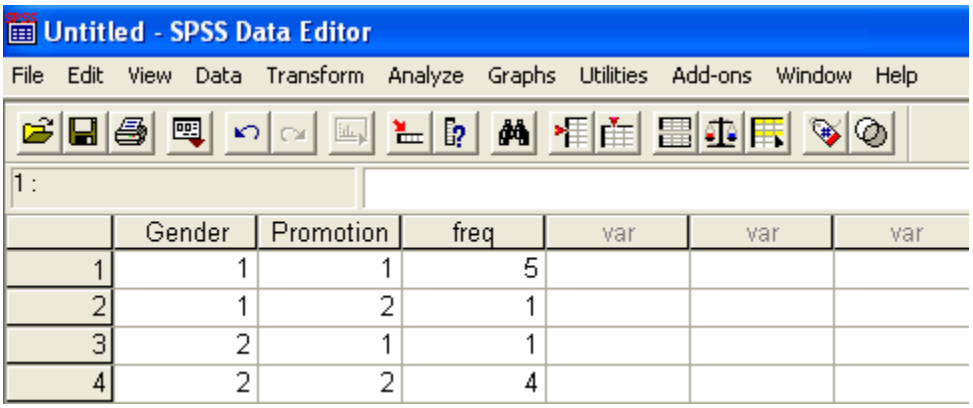
b. 1 cells (25.0%) have expected count less than 5. The minimum expected count is 2.40.

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Education (1 / 2)	2.538	.591	10.912
For cohort Iron_stat = 1	2.333	.624	8.719
For cohort Iron_stat = 2	.919	.790	1.070
N of Valid Cases	100		

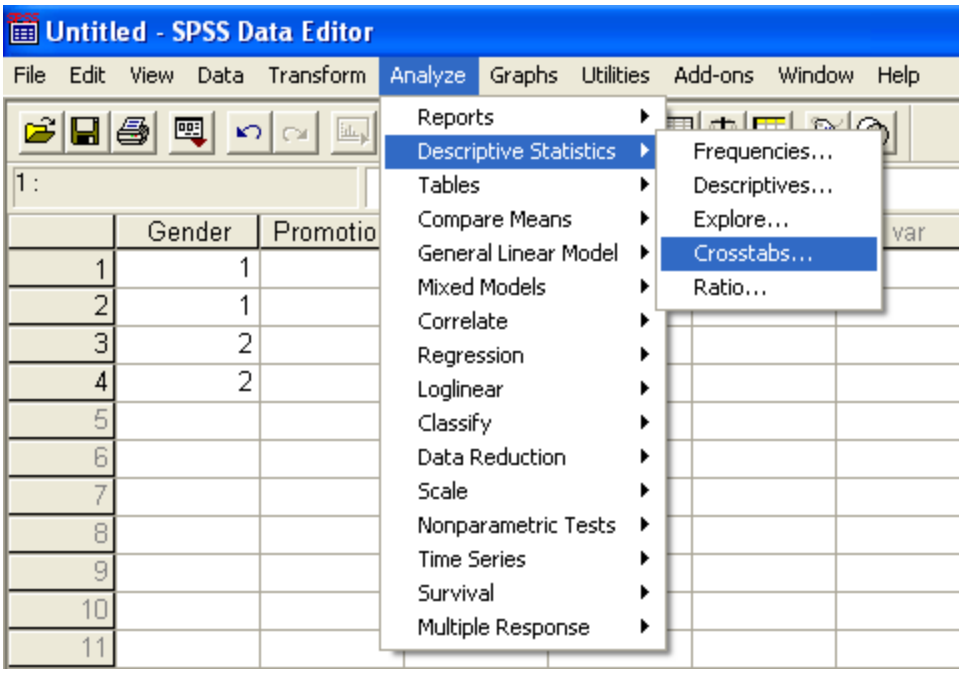
2. Fisher's Exact Test

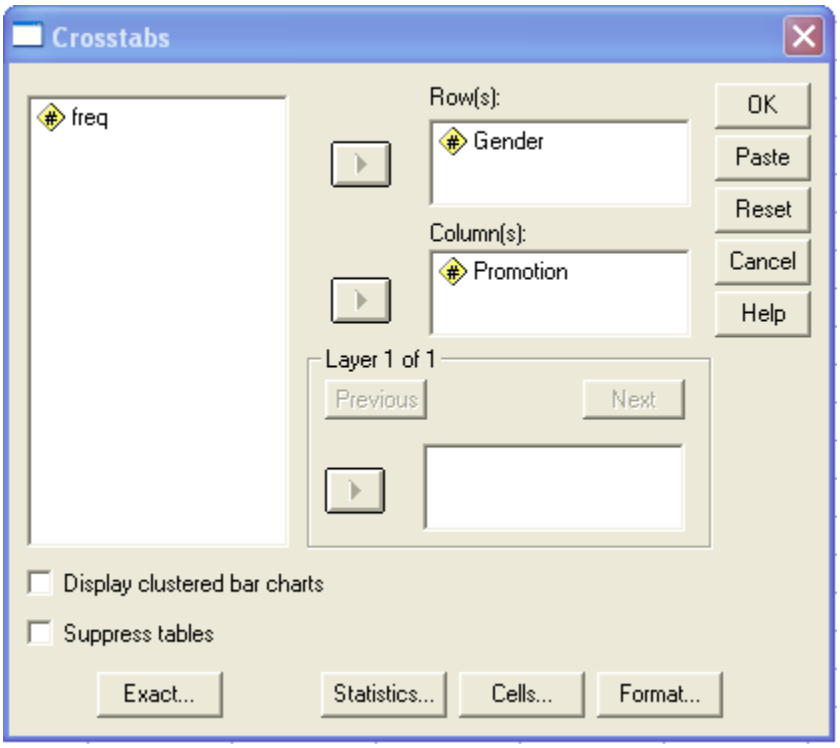
Here we present data from Example 10.5 that examines records of promotion by gender. For the variable Gender, '1' represents Male and '2' represents Female, and for the variable Promotion, '1' represents Yes and '2' represents No. The variable freq specifies the number of observations in each cell of the 2 by 2 contingency table.



	Gender	Promotion	freq	var	var	var
1	1	1	5			
2	1	2	1			
3	2	1	1			
4	2	2	4			

Although we do not show it here, you must use the SPSS procedure **Data** -> **Weight Cases…** before going to the **Analyze** command. Next we use the SPSS procedure **Analyze** -> **Descriptive Statistics** -> **Crosstabs…** to calculate the p-value for a Fisher's Exact test.





Part of the SPSS output is shown below. In the text we present the p-value 0.067.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	4.412[b]	1	.036	.080	.067	
Continuity Correction[a]	2.228	1	.136			
Likelihood Ratio	4.747	1	.029	.080	.067	
Fisher's Exact Test				.080	.067	
Linear-by-Linear Association	4.011[c]	1	.045	.080	.067	.065
N of Valid Cases	11					

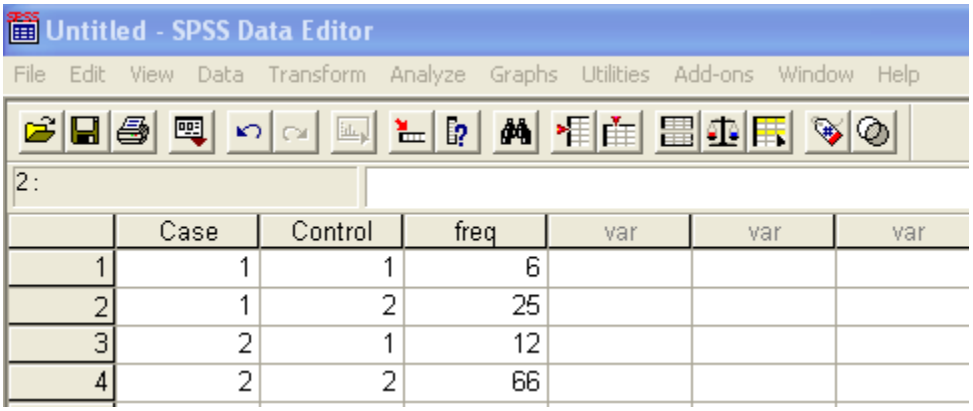
a. Computed only for a 2x2 table

b. 4 cells (100.0%) have expected count less than 5. The minimum expected count is 2.27.

c. The standardized statistic is 2.003.

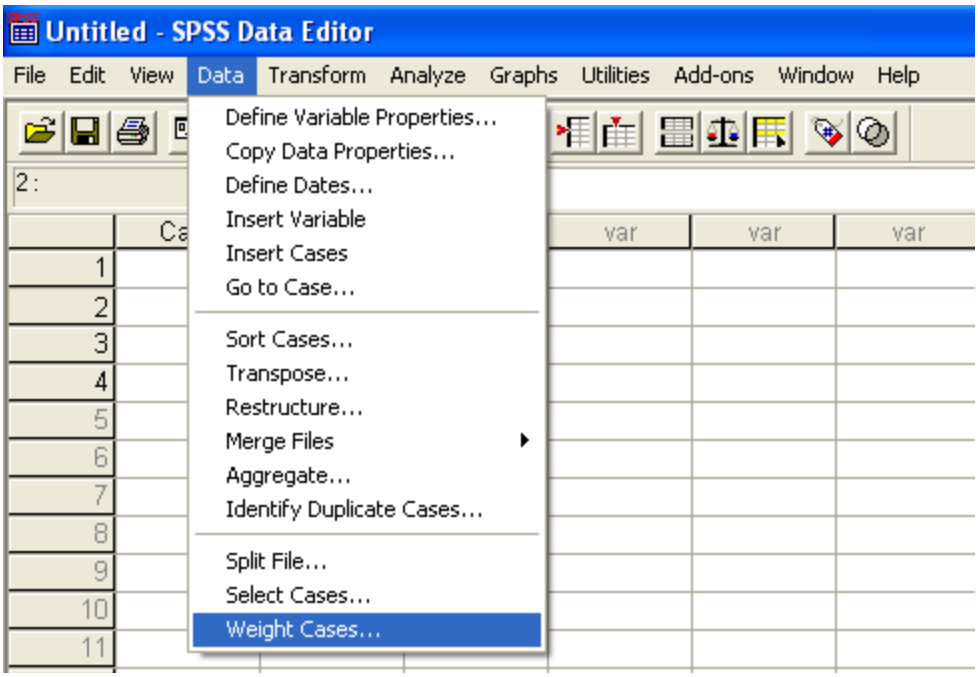
3. McNemar's Test

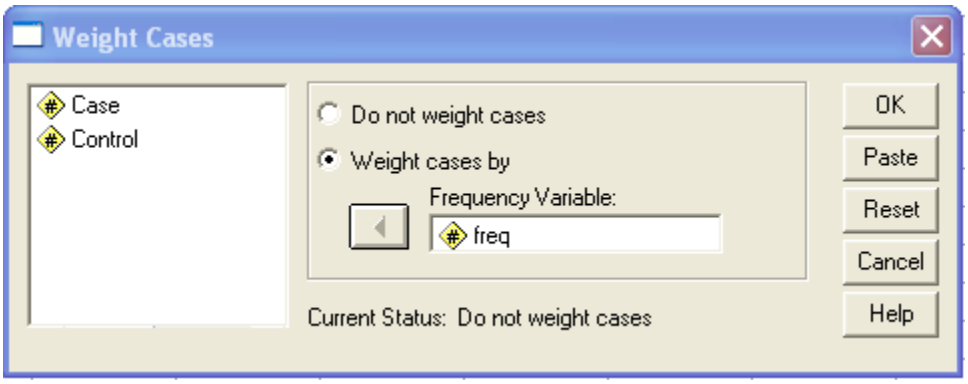
Here we present data from Example 10.6 where we conduct an analysis of matched pairs data. In Example 10.6, we examine family history of dementia among cases and controls. For both the Case and Control variables, '1' represents Present or that family history of dementia is present and '2' represents absent. The variable freq specifies the number of observations in each cell of the 2 by 2 contingency table.



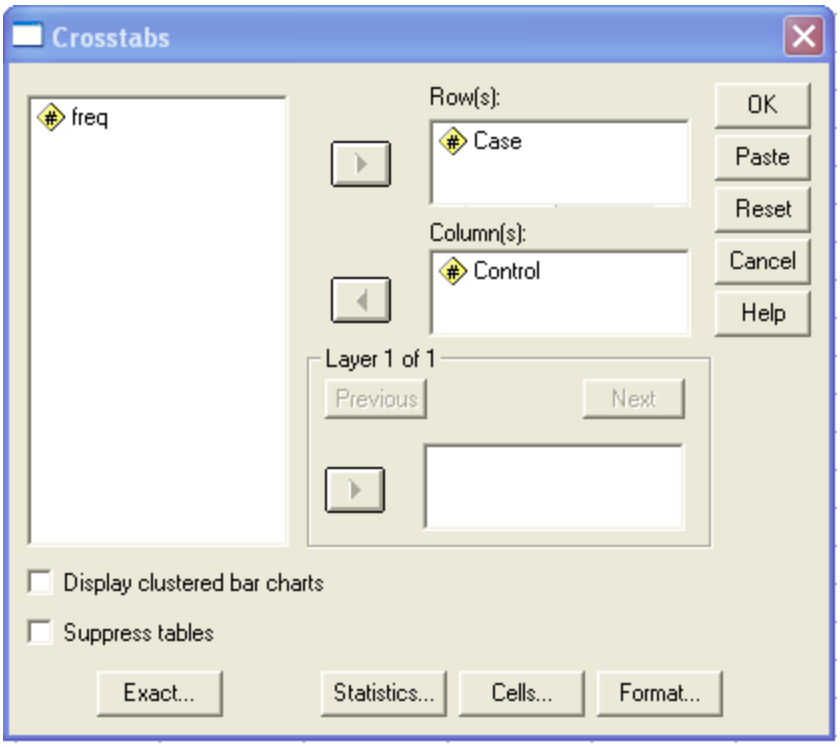
	Case	Control	freq	var	var	var
1	1	1	6			
2	1	2	25			
3	2	1	12			
4	2	2	66			

Once again, you must use the SPSS procedure **Data** -> **Weight Cases…** before going to the **Analyze** command.

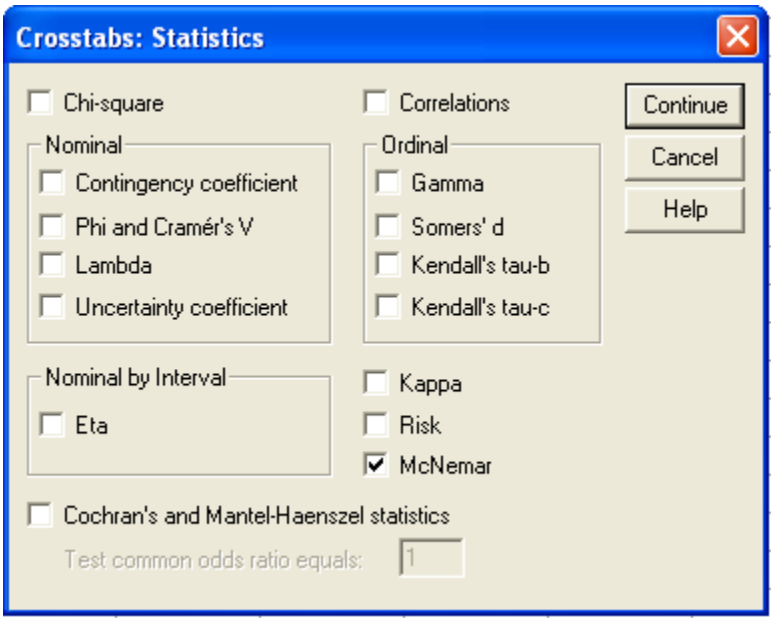




Next we use the SPSS procedure **Analyze** -> **Descriptive Statistics** -> **Crosstabs…** to obtain the **Crosstabs** window below.



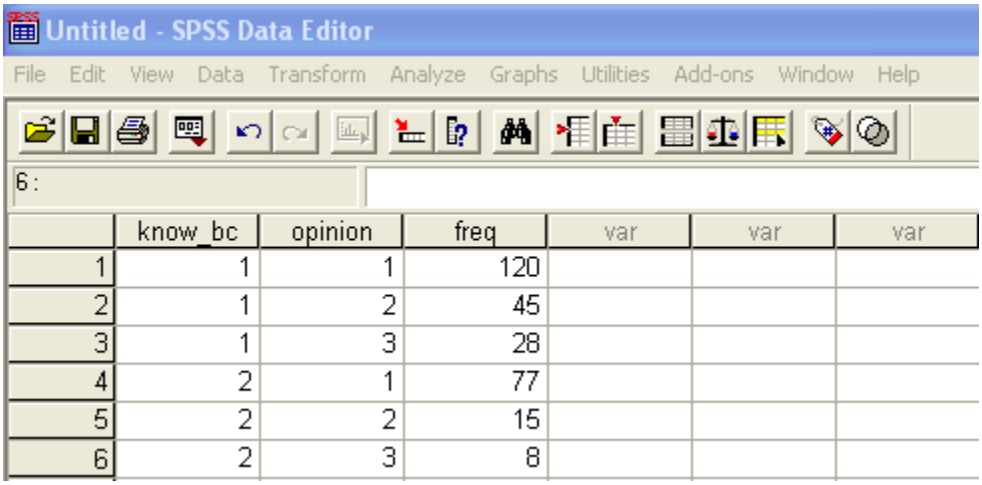
In the **Crosstabs: Statistics** window, select **McNemar**.



Click on **Continue** to produce the test statistic for McNemar's test.

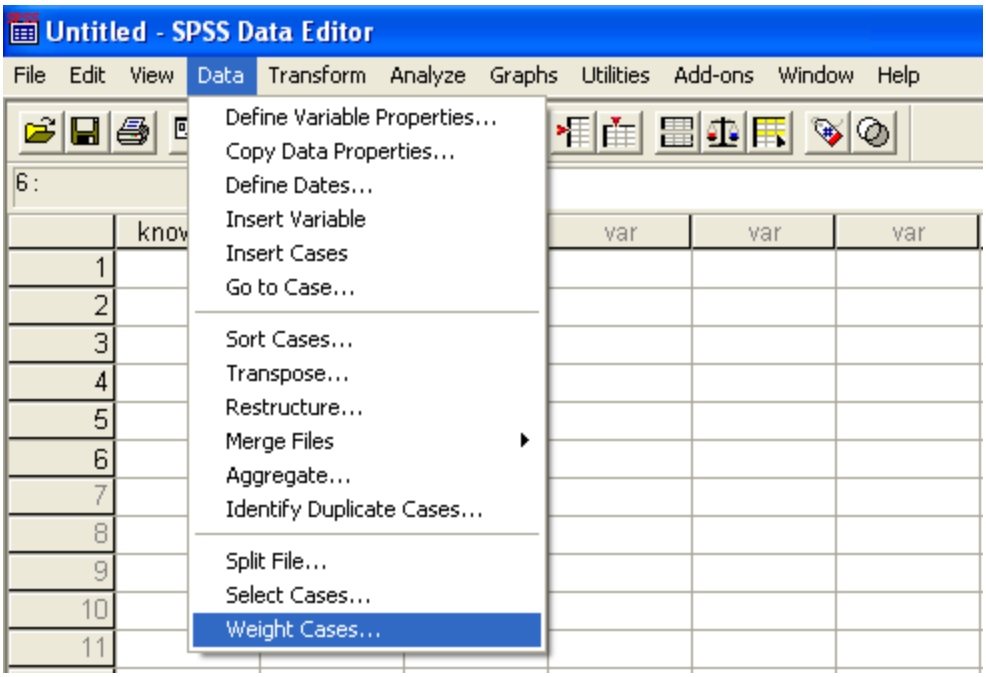
Program Note 10.2 – Chi-square Test for an r by c Contingency Table

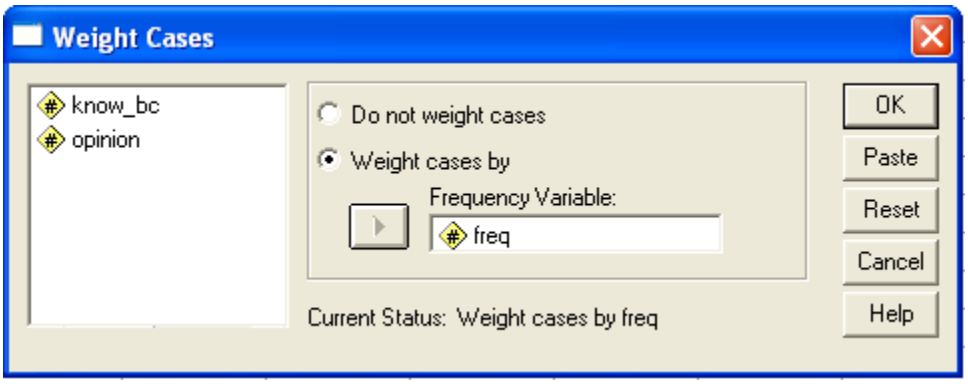
Here we present data from Example 10.7 on the women's opinion of mamography. For the variable know_bc, '1' represents Yes or knowing someone with breast cancer and '2' represents No or not knowing someone with breast cancer. For the variable opinion, '1' represents Positive or having a positive opinion about mammography, '2' represents Neutral, and '3' represents Negative. The variable freq specifies the number of observations in each cell of the 2 by 3 contingency table.

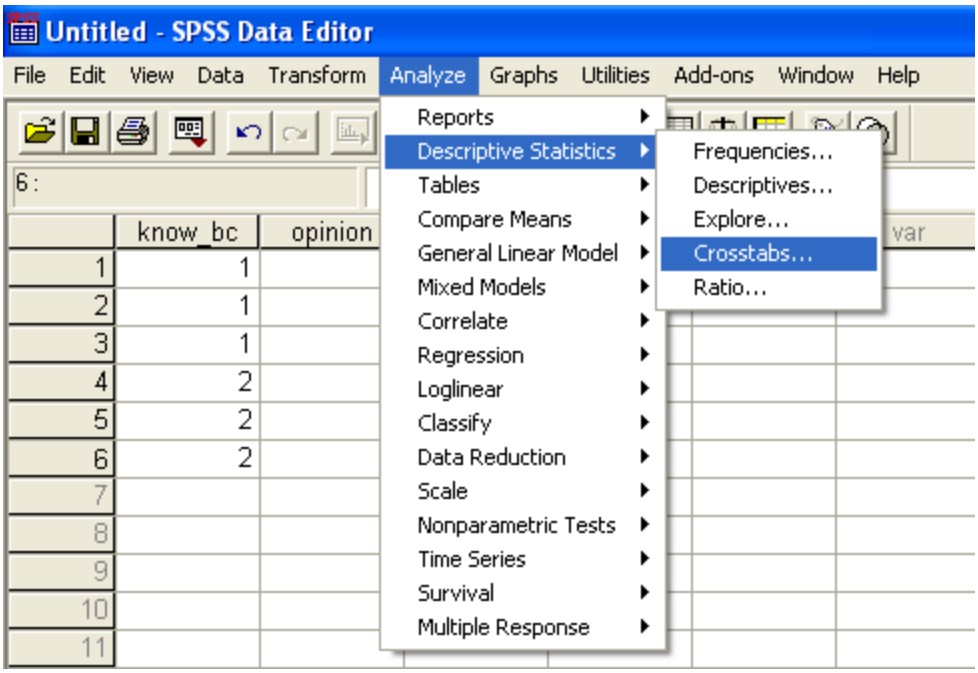


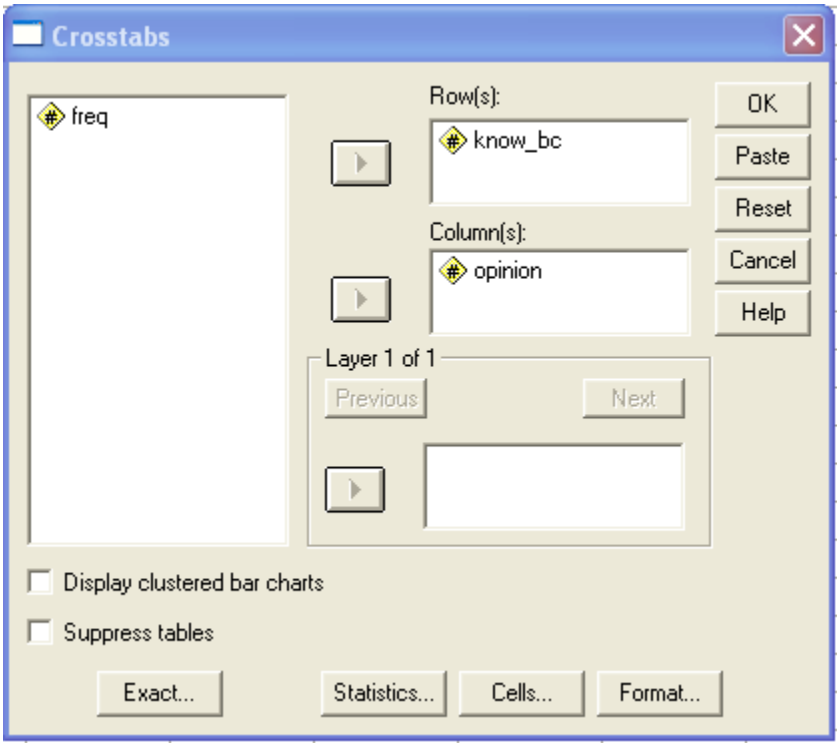
	know_bc	opinion	freq	var	var	var
1	1	1	120			
2	1	2	45			
3	1	3	28			
4	2	1	77			
5	2	2	15			
6	2	3	8			

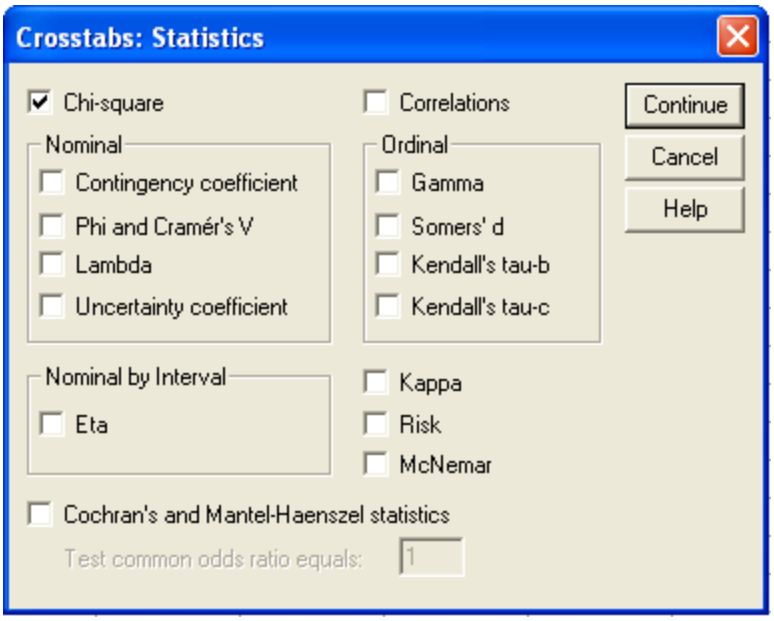
Once again, you must use the SPSS procedure **Data** -> **Weight Cases…** before going to the **Analyze** command.


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Define Variable Properties...
Copy Data Properties...
Define Dates...
Insert Variable
Insert Cases
Go to Case...
Sort Cases...
Transpose...
Restructure...
Merge Files
Aggregate...
Identify Duplicate Cases...
Split File...
Select Cases...
Weight Cases...
6 :
know
var
var
var


Weight Cases
know_bc
opinion
Do not weight cases
Weight cases by
Frequency Variable:
freq
Current Status: Weight cases by freq
OK
Paste
Reset
Cancel
Help


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Frequencies...
Descriptives...
Explore...
Crosstabs...
Ratio...
6 :
know_bc
opinion
var
1 1
2 1
3 1
4 2
5 2
6 2
7
8
9
10
11


Crosstabs
freq
Row(s):
know_bc
Column(s):
opinion
OK
Paste
Reset
Cancel
Help
Layer 1 of 1
Previous
Next
Display clustered bar charts
Suppress tables
Exact...
Statistics...
Cells...
Format...



The SPSS output is shown below.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
know_bc * opinion	293	100.0%	0	.0%	293	100.0%

know_bc * opinion Crosstabulation

Count

		opinion			
		Positive	Neutral	Negative	Total
know_bc	Yes	120	45	28	193
	No	77	15	8	100
Total		197	60	36	293

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6.648[a]	2	.036
Likelihood Ratio	6.891	2	.032
Linear-by-Linear Association	6.056	1	.014
N of Valid Cases	293		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 12.29.

Program Note 10.3 – Trend Test

To illustrate the trend test, we use the data from Example 10.7 and the same SPSS procedure as the one presented above. However, in the SPSS output below, we will focus on the **Linear-by-Linear Association** value from the **Chi-Square Tests** table. The value of 6.056 corresponds to the trend test statistic.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
know_bc * opinion	293	100.0%	0	.0%	293	100.0%

know_bc * opinion Crosstabulation

Count

		opinion			
		Positive	Neutral	Negative	Total
know_bc	Yes	120	45	28	193
	No	77	15	8	100
Total		197	60	36	293

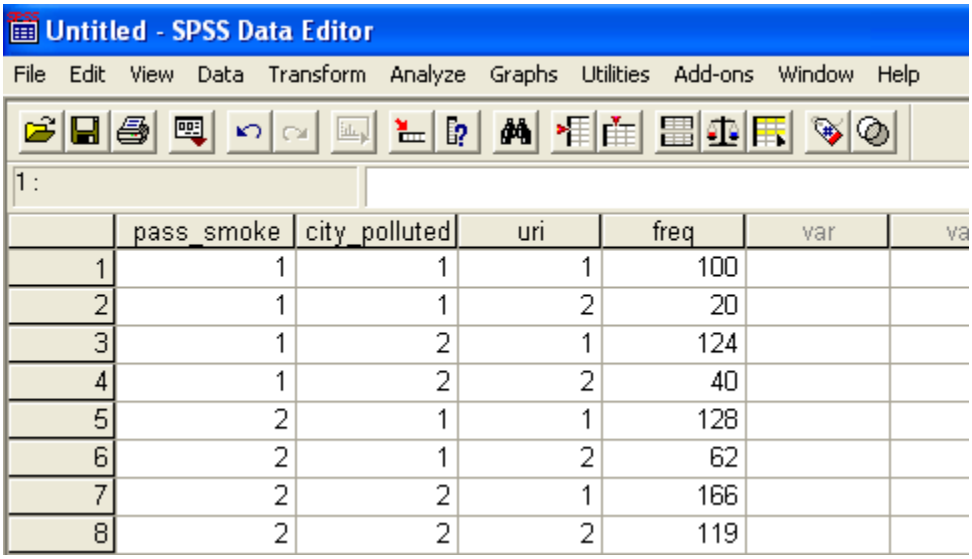
Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6.648[a]	2	.036
Likelihood Ratio	6.891	2	.032
Linear-by-Linear Association	6.056	1	.014
N of Valid Cases	293		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 12.29.

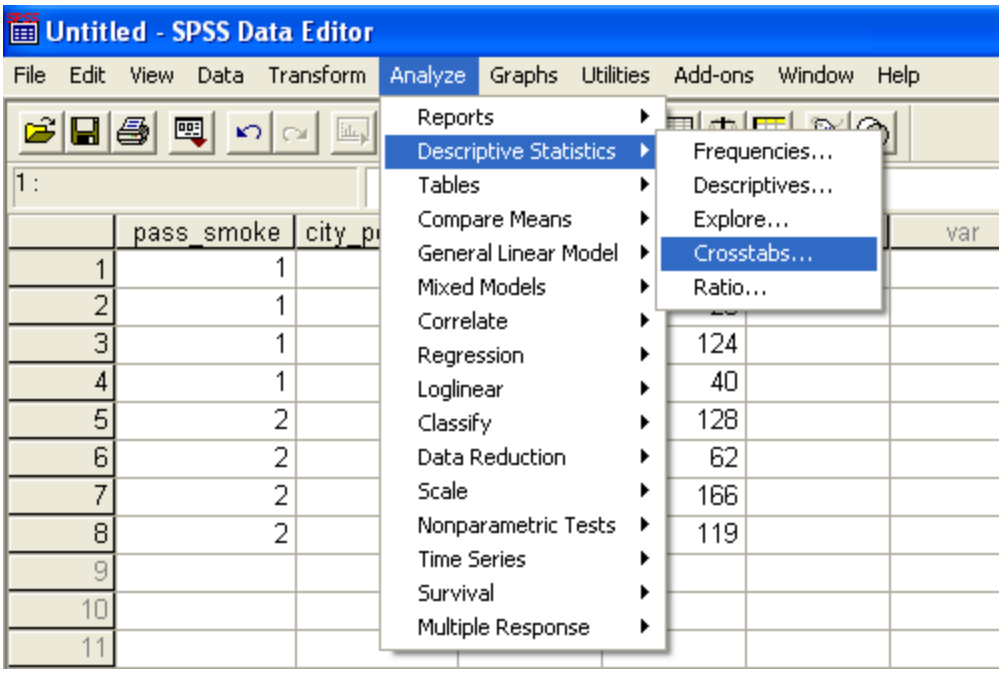
Program Note 10.4 – The Mantel-Haenszel Odds Ratio

To illustrate the Cochran-Mantel-Haenszel test, we use the data from Table 10.8 referred to in Example 10.9. The variable pass_smoke refers to passive smoke in the house where '1' represents Yes and '2' represents No. The variable city_polluted refers to level of city pollution where '1' represents high pollution and '2' represents low. The variable uri refers to upper respiratory infection during the previous 12 months where '1' represents Some and '2' represents None. The variable freq specifies the number of observations in each cell of the contingency table.

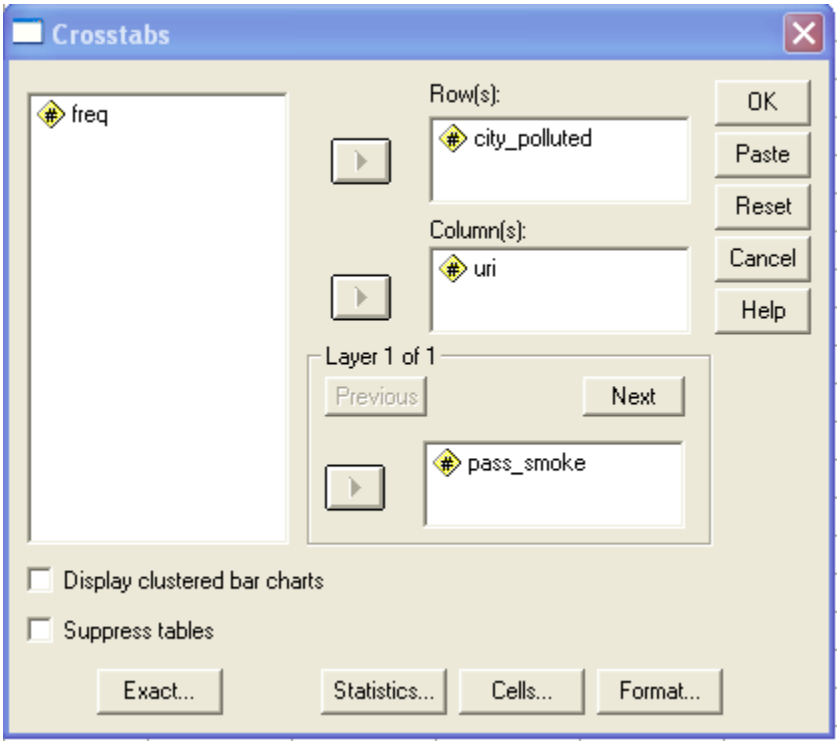


	pass_smoke	city_polluted	uri	freq	var	va
1	1	1	1	100		
2	1	1	2	20		
3	1	2	1	124		
4	1	2	2	40		
5	2	1	1	128		
6	2	1	2	62		
7	2	2	1	166		
8	2	2	2	119		

Although we do not show it here, you must use the SPSS procedure **Data** -> **Weight Cases…** before going to the **Analyze** command.



In the Crosstabs window, we make the following selection. Notice that the variable pass_smoke goes in the **Layer 1 of 1** window.



The partial SPSS output is shown below.

city_polluted * uri * pass_smoke Crosstabulation

Count

pass_smoke			uri		Total
			Some	None	
Yes	city_polluted	High	100	20	120
		Low	124	40	164
	Total		224	60	284
No	city_polluted	High	128	62	190
		Low	166	119	285
	Total		294	181	475

Chi-Square Tests

pass_smoke		Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Yes	Pearson Chi-Square	2.481[b]	1	.115		
	Continuity Correction[a]	2.039	1	.153		
	Likelihood Ratio	2.528	1	.112		
	Fisher's Exact Test				.141	.076
	Linear-by-Linear Association	2.472	1	.116		
	N of Valid Cases	284				
No	Pearson Chi-Square	4.023[c]	1	.045		
	Continuity Correction[a]	3.645	1	.056		
	Likelihood Ratio	4.056	1	.044		
	Fisher's Exact Test				.054	.028
	Linear-by-Linear Association	4.014	1	.045		
	N of Valid Cases	475				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 25.35.

c. 0 cells (.0%) have expected count less than 5. The minimum expected count is 72.40.

Tests of Homogeneity of the Odds Ratio

	Chi-Squared	df	Asymp. Sig. (2-sided)
Breslow-Day	.056	1	.812
Tarone's	.056	1	.812

Tests of Conditional Independence

	Chi-Squared	df	Asymp. Sig. (2-sided)
Cochran's	6.456	1	.011
Mantel-Haenszel	6.037	1	.014

Under the conditional independence assumption, Cochran's statistic is asymptotically distributed as a 1 df chi-squared distribution, only if the number of strata is fixed, while the Mantel-Haenszel statistic is always asymptotically distributed as a 1 df chi-squared distribution. Note that the continuity correction is removed from the Mantel-Haenszel statistic when the sum of the differences between the observed and the expected is 0.

Mantel-Haenszel Common Odds Ratio Estimate

Estimate			1.518
ln(Estimate)			.418
Std. Error of ln(Estimate)			.165
Asymp. Sig. (2-sided)			.011
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	1.099
		Upper Bound	2.097
	ln(Common Odds Ratio)	Lower Bound	.095
		Upper Bound	.740

The Mantel-Haenszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1.000 assumption. So is the natural log of the estimate.