

SAS Program Notes

Biostatistics: A Guide to Design, Analysis, and Discovery

Chapter 10: Analysis of Categorical Data

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

In Example 10.3, we use a contingency table, Table 10.6, to display data on education and iron status. We would like to determine if these variables are independent. The SAS procedure **PROC FREQ** can be used to produce contingency tables along with statistics for the chi-square test of independence as well as Fisher's Exact test. In the SAS command lines below, we use the variable **EDUCAT** to indicate level of education and the variable **IRON** to indicate if the individual is iron deficient or not. Because the levels of the variable **EDUCAT** fall along the rows and the levels of the variable **IRON** fall along the columns, the variables are joined by an asterisk '*' with the row variable **EDUCAT** first followed by the column variable **IRON** after the **TABLES** statement. Notice that we use a forward slash '/' follow by **CHISQ** to request statistics for the 2 by 2 table. In Example 10.3, we use Yate's correction and that test statistic is found next to the **Continuity Adj. Chi-Square**. The value of the test statistic is **0.7829** and its corresponding p-value is **0.3763**.

SAS commands:

```
DATA EDU_IRON;
INPUT EDUCAT $ 1-13 IRON $ 15 N 17-18;
DATALINES;
UNDER 12YRS D 4
UNDER 12YRS A 26
12YRS OR MORE D 4
12YRS OR MORE A 66
;
PROC FREQ ORDER=DATA;
  TABLES EDUCAT*IRON/CHISQ;
  WEIGHT N;
RUN;
```

SAS output:

The SAS System
The FREQ Procedure

Table of EDUCAT by IRON

EDUCAT	IRON		
Frequency			
Percent			
Row Pct			
Col Pct	D	A	Total

UNDER 12YRS	4	26	30
	4.00	26.00	30.00
	13.33	86.67	
	50.00	28.26	
12YRS OR MORE	4	66	70
	4.00	66.00	70.00
	5.71	94.29	
	50.00	71.74	
Total	8	92	100
	8.00	92.00	100.00

Statistics for Table of EDUCAT by IRON

Statistic	DF	Value	Prob
Chi-Square	1	1.6563	0.1981
Likelihood Ratio Chi-Square	1	1.5289	0.2163
Continuity Adj. Chi-Square	1	0.7829	0.3763
Mantel-Haenszel Chi-Square	1	1.6398	0.2004
Phi Coefficient		0.1287	
Contingency Coefficient		0.1276	
Cramer's V		0.1287	

WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Fisher's Exact Test

Cell (1,1) Frequency (F)	4
Left-sided Pr <= F	0.9496
Right-sided Pr >= F	0.1855
Table Probability (P)	0.1350
Two-sided Pr <= P	0.2362

In the SAS commands below, we use a forward slash '/' follow by **REL RISK** in the **TABLES** statement to request odds ratio estimates for the 2 by 2 table. Notice that the SAS output produces an odds ratio estimate followed by values for a 95% confidence interval.

SAS commands:

```
PROC FREQ DATA=EDU_IRON ORDER=DATA;
  TABLES EDUCAT*IRON/REL RISK;
  WEIGHT N;
RUN;
```

SAS commands:

The SAS System
The FREQ Procedure

Statistics for Table of EDUCAT by IRON

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	2.5385	0.5905	10.9121
Cohort (Col1 Risk)	2.3333	0.6244	8.7194
Cohort (Col2 Risk)	0.9192	0.7898	1.0698

Sample Size = 100

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

In Example 10.7, the 2 by 3 contingency table results from the cross-tabulation of whether a woman knew someone with breast cancer with her opinion about mammography. In the SAS commands below, we use a forward slash '/' follow by **CHISQ** and **EXPECTED** in the **TABLES** statement to request statistics and to produce the expected values in the 2 by 3 contingency table.

SAS commands:

```
DATA BREASTCANCER;  
INPUT KNOW $ OPINION $ N;  
DATALINES;  
YES POSITIVE 120  
YES NEUTRAL 45  
YES NEGATIVE 28  
NO POSITIVE 77  
NO NEUTRAL 15  
NO NEGATIVE 8  
;  
PROC FREQ ORDER=DATA;  
TABLE KNOW*OPINION/CHISQ EXPECTED;  
WEIGHT N;  
RUN;
```

SAS output:

The SAS System
The FREQ Procedure

Table of KNOW by OPINION

KNOW	OPINION			Total
	POSITIVE	NEUTRAL	NEGATIVE	
Frequency				
Expected				
Percent				
Row Pct				
Col Pct				
YES	120	45	28	193
	129.76	39.522	23.713	
	40.96	15.36	9.56	65.87
	62.18	23.32	14.51	
	60.91	75.00	77.78	
NO	77	15	8	100
	67.235	20.478	12.287	
	26.28	5.12	2.73	34.13
	77.00	15.00	8.00	
	39.09	25.00	22.22	
Total	197	60	36	293
	67.24	20.48	12.29	100.00

Statistics for Table of KNOW by OPINION

Statistic	DF	Value	Prob
Chi-Square	2	6.6479	0.0360
Likelihood Ratio Chi-Square	2	6.8914	0.0319
Mantel-Haenszel Chi-Square	1	6.0556	0.0139
Phi Coefficient		0.1506	
Contingency Coefficient		0.1489	
Cramer's V		0.1506	

Sample Size = 293

Note 10.3 – Trend Test

In Example 10.8, we want to test the null hypothesis of no linear trend in opinion about mammography. In the SAS commands below, we use a forward slash '/' follow by **TREND** in the **TABLES** statement to request statistics and to produce the test statistic for the **Cochran-Armitage Trend Test**. Notice in the textbook that we compute a test statistic that follows a chi-square distribution, in particular, it is a chi-square distribution with one degree of freedom. Therefore, to obtain the result provided in the textbook, we should just square the test statistic (2.465 squared is approximately equal to 6.10) since a standard normal value squared is equivalent to a chi-square value with one degree of freedom.

SAS commands:

```
PROC FREQ DATA=BREASTCANCER ORDER=DATA;  
  TABLE KNOW*OPINION/TREND;  
  WEIGHT N;  
RUN;
```

SAS output:

The SAS System
The FREQ Procedure

Table of KNOW by OPINION

KNOW	OPINION			Total
	POSITIVE	NEUTRAL	NEGATIVE	
YES	Frequency			
	Percent			
	Row Pct			
	Col Pct			
	120	45	28	193
	40.96	15.36	9.56	65.87
	62.18	23.32	14.51	
	60.91	75.00	77.78	
NO	Frequency			
	Percent			
	Row Pct			
	Col Pct			
	77	15	8	100
	26.28	5.12	2.73	34.13
	77.00	15.00	8.00	
	39.09	25.00	22.22	
Total	197	60	36	293
	67.24	20.48	12.29	100.00

Statistics for Table of KNOW by OPINION

Cochran-Armitage Trend Test

Statistic (Z)	2.4650
One-sided Pr > Z	0.0069
Two-sided Pr > Z	0.0137

Sample Size = 293

In Example 10.10, we use the Mantel-Hanzel estimate of the common odds ratio to adjust for the presence of passive smoke. In the SAS commands below, we use a forward slash '/' follow by **RELRISK** and **CMH** in the **TABLES** statement to request the Cochran-Mantel-Hanzel statistics. Notice that by using the **RELRISK** option, odds ratios are provided by **PASSIVE_SMOKE** status.

SAS commands:

```

DATA RESP_INFECTION;
INPUT PASSIVE_SMOKE $ POLLUTION $ INFECTION $ N;
DATALINES;
YES HIGH SOME 100
YES HIGH NONE 20
YES LOW SOME 124
YES LOW NONE 40
NO HIGH SOME 128
NO HIGH NONE 62
NO LOW SOME 166
NO LOW NONE 119
;
PROC FREQ ORDER=DATA;
    TABLE PASSIVE_SMOKE*POLLUTION*INFECTION/RELRISK CMH;
    WEIGHT N;
RUN;

```

SAS output:

The SAS System
The FREQ Procedure

Table 1 of POLLUTION by INFECTION
Controlling for PASSIVE_SMOKE=YES

POLLUTION	INFECTION		Total
	SOME	NONE	
HIGH	Frequency		
	Percent		
	Row Pct		
	Col Pct		
	100	20	120
	35.21	7.04	42.25
	83.33	16.67	
	44.64	33.33	
LOW	Frequency		
	Percent		
	Row Pct		
	Col Pct		
	124	40	164
	43.66	14.08	57.75
	75.61	24.39	
	55.36	66.67	
Total	224	60	284

78.87 21.13 100.00

Statistics for Table 1 of POLLUTION by INFECTION
Controlling for PASSIVE_SMOKE=YES

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	1.6129	0.8870	2.9328
Cohort (Col1 Risk)	1.1022	0.9793	1.2404
Cohort (Col2 Risk)	0.6833	0.4218	1.1069

Sample Size = 284

Table 2 of POLLUTION by INFECTION
Controlling for PASSIVE_SMOKE=NO

POLLUTION	INFECTION		Total
	SOME	NONE	
Frequency			
Percent			
Row Pct			
Col Pct			
HIGH	128	62	190
	26.95	13.05	40.00
	67.37	32.63	
	43.54	34.25	
LOW	166	119	285
	34.95	25.05	60.00
	58.25	41.75	
	56.46	65.75	
Total	294	181	475
	61.89	38.11	100.00

Statistics for Table 2 of POLLUTION by INFECTION
Controlling for PASSIVE_SMOKE=NO

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	1.4800	1.0081	2.1727
Cohort (Col1 Risk)	1.1566	1.0060	1.3298
Cohort (Col2 Risk)	0.7815	0.6110	0.9995

Sample Size = 475

Summary Statistics for POLLUTION by INFECTION

Controlling for PASSIVE_SMOKE

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

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	6.4396	0.0112
2	Row Mean Scores Differ	1	6.4396	0.0112
3	General Association	1	6.4396	0.0112

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	Mantel-Haenszel	1.5182	1.0992	2.0968
	Logit	1.5176	1.0987	2.0963
Cohort (Col1 Risk)	Mantel-Haenszel	1.1326	1.0309	1.2443
	Logit	1.1246	1.0276	1.2307
Cohort (Col2 Risk)	Mantel-Haenszel	0.7558	0.6063	0.9421
	Logit	0.7601	0.6105	0.9464

Breslow-Day Test for
Homogeneity of the Odds Ratios

Chi-Square	0.0563
DF	1
Pr > ChiSq	0.8125

Total Sample Size = 759