Categorical Data Analysis
Lab material #2

Table 1 summarizes the information from a randomized clinical trial that compared two treatments (test, placebo) for a respiratory disorder.

Table 1: Respiratory Outcomes

Treatment	Favorable	Unfavorable	Total
Placebo	16	48	64
Test	40	20	60

The question of interest is whether the rates of favorable response for test (67 percent) and placebo (25 percent) are the same. You can address this question by investigating whether there is a statistical association between treatment and outcome.

The following PROC FREQ statements produce a frequency table and the chi-square statistics for the data in Table 1. The data are supplied in frequency, or count, form. An observation is supplied for each configuration of the values of the variables TREAT and OUTCOME. The variable COUNT holds the total number of observations that have that particular configuration. The WEIGHT statement tells the FREQ procedure that the data are in frequency form and names the variable that contains the frequencies.

The CHISQ option in the TABLES statement produces chi-square statistics.

```
* example2_1.sas;
data respire;
  input treat $ outcome $ count;
  cards;
  placebo f 16
  placebo u 48
  test f 40
  test u 20
  ;
proc freq;
  weight count;
  tables treat*outcome / chisq;
run;
```

Output 1 displays the data in a 2×2 table. With an overall sample size of 124, and all expected cell counts greater than 10, the sampling assumptions for the chi-square statistics are met. Output 1 also contains the table with the chi-square statistic. The Pearson chi-square X^2 is labeled "Chi-Square" and the Likelihood ratio statistic G^2 is labeled "Likelihood Ratio Chi-Square." X^2 has a value of 21.7087 and p-value is less than 0.0001; G^2 has a value of 22.377 and p-value is less than 0.0001. Both of these statistics are clearly significant. There is a strong association between treatment and outcome such that the test treatment results in a more favorable response outcome than the placebo.

Output 1 Frequency Table and Table Statistics

Table of treat by outcome

treat	outcome		
Frequency| Percent | Row Pct | Col Pct	f	u	Total
placebo	16	48	64
	12.90	38.71	51.61
	25.00	75.00	
	28.57	70.59	
test	40	20	60
	32.26	16.13	48.39
	66.67	33.33	
	71.43	29.41	
Total	56	68	124
	45.16	54.84	100.00

Statistics for Table of treat by outcome

Statistic	DF	Value	Prob
Chi-Square	1	21.7087	<.0001
Likelihood Ratio Chi-Square	1	22.3768	<.0001
Continuity Adj. Chi-Square	1	20.0589	<.0001
Mantel-Haenszel Chi-Square	1	21.5336	<.0001
Phi Coefficient		-0.4184	
Contingency Coefficient		0.3860	
Cramer's V		-0.4184	

Fisher's Exact Test

Cell (1,1) Frequency (F)	16
Left-sided Pr <= F	2.838E-06
Right-sided Pr >= F	1.0000
Table Probability (P)	2.397E-06
Two-sided Pr <= P	4.754E-06

Sample Size = 124

Sometimes your data include small and zero cell counts. For example, consider the following data from a study on treatments for healing severe infections. A test treatment and a control are compared to determine whether the rates of favorable response are the same.

Table 2: Severe Infection Treatment Outcomes

Treatment	Favorable	Unfavorable	Total
Test	10	2	12
Control	2	4	6
Total	12	6	18

Obviously, the sample sizes requirements for the chi-square tests are not met by these data. However, if you can consider the margins (12, 6, 12, 6) to be fixed, then you can conduct Fisher's exact test. The following SAS code produces the 2×2 frequency table for Table 2. Specifying the CHISQ option also produces the Fisher's exact test for a 2×2 table. In addition, the ORDER=DATA option specifies that PROC FREQ orders the levels of the rows (columns) in the same order in which the values are encountered in the data set.

```
* example2_2.sas;
data severe;
  input treat $ outcome $ count;
  cards;
  Test f 10
  Test u 2
  Control f 2
  Control u 4
  ;
proc freq order=data;
  tables treat*outcome / chisq nocol;
  weight count;
run;
```

The NOCOL option suppresses the column percentages, as seen in Output 2. Output 2 contains the chi-square statistics, including the exact test. Note that the sample size assumptions are not met for the chi-square tests: the warning beneath the table asserts that this is the case.

Output 2 Frequency Table and Table Statistics

```
     Table of treat by outcome

treat     outcome

Frequency|
Percent  |
Row Pct  |f       |u       |  Total
---------+--------+--------+
Test     |     10 |      2 |     12
         |  55.56 |  11.11 |  66.67
         |  83.33 |  16.67 |
---------+--------+--------+
Control  |      2 |      4 |      6
         |  11.11 |  22.22 |  33.33
         |  33.33 |  66.67 |
---------+--------+--------+
Total          12        6       18
            66.67    33.33   100.00
```

Statistics for Table of treat by outcome

Statistic	DF	Value	Prob
Chi-Square	1	4.5000	0.0339
Likelihood Ratio Chi-Square	1	4.4629	0.0346
Continuity Adj. Chi-Square	1	2.5313	0.1116
Mantel-Haenszel Chi-Square	1	4.2500	0.0393
Phi Coefficient		0.5000	
Contingency Coefficient		0.4472	
Cramer's V		0.5000	

WARNING: 75% 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)	10
Left-sided Pr <= F	0.9961
Right-sided Pr >= F	0.0573
Table Probability (P)	0.0533
Two-sided Pr <= P	0.1070

Sample Size = 18

Note that the SAS System produces both a left-tail and right-tail p-value for Fisher's exact test. The left-tail probability is the probability of all tables such that the (1,1) cell value is less than or equal to the one observed. The right-tail probability is the probability of all tables such that the (1,1) cell value is greater than or equal to the one observed. Thus, the one-sided p-value is the same as the right-tailed p-value in this case, since large values for the (1,1) cell correspond to better outcomes for Test treatment.

Both the two-sided p-value of 0.1070 and the one-sided p-value of 0.0573 are larger than the p-values associated with X^2 ($p = 0.0339$) and G^2 ($p = 0.0346$). Depending on you significance criterion, you may reach different conclusions with these three test statistics. The sample size requirements for the chi-square distribution are not met with these data; hence the test statistics using this approximation are questionable. This example illustrates the usefulness of Fisher's exact test when the sample size requirements for the usual chi-square tests are not met. Note that Fisher's exact test is always appropriate, even when the sample size is large.

Table 3 contains data from a study on how general daily stress affects one's opinion on a proposed new health policy. Since information on stress level and opinion were collected at the same time, the data are cross-sectional.

Table 3: Opinion on New-Health Policy

Stress	Favorable	Unfavorable	Total
Low	48	12	60
High	96	94	190

To produce the odds ratio and other measures of association from PROC FREQ, you specify the MEASURES option in the TABLES statement. The ORDER=DATA option is used in the PROC FREQ statement to produce a table that looks the same as that displayed in Table 3. Without this option, the row corresponding to high stress would come first and the row corresponding to low stress would come last.

```
* example2_3.sas;
data stress;
  input stress $ outcome $ count;
  cards;
  low f 48
  low u 12
  high f 96
  high u 94
  ;
proc freq order=data;
  tables stress*outcome / chisq measures nocol nopct;
  weight count;
run;
```

Output 3 contains the resulting frequency table. Since the NOCOL and NOPCT options are specified, only the row percentages are printed. Eighty percent of the low stress group were favorable, while the high stress group was nearly evenly split between favorable and unfavorable. Output 3 displays the chi-square statistics, measures of association, and a table labeled "Estimates of the Relative Risk (Row1/Row2)." The statistics X^2 and G^2 indicate a strong association, with values of 16.2198 and 17.3520, respectively. Note how close the values for these statistics are for a sample size of 250.

Measures of association such as Kendall's tau-b, the Pearson Correlation, Spearman Correlation, and uncertainty coefficients are listed. See Lecture note 4 for more information about some of these measures.

The odds ratio value is listed beside "Case-Control" in the section labeled "Estimates of the Relative Risk (Row1/Row2)." The estimated OR is 3.9167, which means that the odds of a favorable response are 3.9167 times higher for those with low stress than those with high stress. The confidence intervals are labeled "Confidence Bounds" and are 95 percent confidence intervals by default. To change them, use the ALPHA=option in the TABLES statement.

The values listed for "Cohort (Col1 Risk)" and "Cohort (Col2 Risk)" are the estimates of relative risk for a cohort (prospective) study. Since these data are cross-sectional, you cannot estimate relative risk (conceptually). However, the value 1.5833 is the ratio of the prevalence of favorable opinions for the low stress group compared to the high stress group. (The value 0.4043 is the prevalence ratio of the unfavorable opinions of the low stress group compared to the high stress group.)

Output 3 Frequency Table and Table Statistics

Table of stress by outcome

stress	outcome		
Frequency| Row Pct	f	u	Total
low	48	12	60
	80.00	20.00	
high	96	94	190
	50.53	49.47	
Total	144	106	250

Statistics for Table of stress by outcome

Statistic	DF	Value	Prob
Chi-Square	1	16.2198	<.0001
Likelihood Ratio Chi-Square	1	17.3520	<.0001
Continuity Adj. Chi-Square	1	15.0354	0.0001
Mantel-Haenszel Chi-Square	1	16.1549	<.0001
Phi Coefficient		0.2547	
Contingency Coefficient		0.2468	
Cramer's V		0.2547	

Fisher's Exact Test

Cell (1,1) Frequency (F)	48
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	3.247E-05
Table Probability (P)	2.472E-05
Two-sided Pr <= P	4.546E-05

Statistics for Table of stress by outcome

Statistic	Value	ASE
Gamma	0.5932	0.1147
Kendall's Tau-b	0.2547	0.0551
Stuart's Tau-c	0.2150	0.0489
Somers' D C|R	0.2947	0.0631
Somers' D R|C	0.2201	0.0499
Pearson Correlation	0.2547	0.0551
Spearman Correlation	0.2547	0.0551
Lambda Asymmetric C|R	0.0000	0.0000
Lambda Asymmetric R|C	0.0000	0.0000
Lambda Symmetric	0.0000	0.0000
Uncertainty Coefficient C|R	0.0509	0.0231
Uncertainty Coefficient R|C	0.0630	0.0282
Uncertainty Coefficient Symmetric	0.0563	0.0253

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	3.9167	1.9575	7.8366
Cohort (Col1 Risk)	1.5833	1.3104	1.9131
Cohort (Col2 Risk)	0.4043	0.2389	0.6841

Sample Size = 250