BIOS 6244 Analysis of Categorical Data
October 19, 2005
Computer Lab 2

As we saw in Computer Lab 1, PROC FREQ can be used to perform many of the statistical procedures for categorical data that we will cover in this course.

In this Computer Lab, we will cover the SAS code required to produce the analyses that we covered in the lectures on October 5, 10, 12, and 17. The examples will be presented in the order in which they were covered in class. The page numbers indicate the pages in the lecture notes.

The Odds Ratio

Reading Assignment: Stokes et al., Section 2.5, pp. 32-39
(available on the course website)

Physician's Aspirin Study Example (pp. 19-21 in lecture notes)

The following SAS code (available on the course website) creates a 2x2 table for the data in Table 2.3 from the Physician's Aspirin Study that examined the association between taking aspirin and 5-year risk of MI.

```
data aspirin;
input aspirin mi count @@;
cards;
1 1 189  1 2 10845
2 1 104  2 2 10933
;
```

To calculate the OR and find an approximate 95% CI(OR), use the following code:

```
proc freq order=data; weight count;
   tables aspirin * mi / nocol nopct relrisk;
   title 'Physicians Aspirin Study';
   title2 'Aspirin vs. MI';
   title3 'Odds Ratio';
run;
```

As pointed out by Dr. Laviolette, the ORDER = DATA option appearing in the PROC FREQ statement orders the data values according to their order in the input data set. We include this option so that SAS will not reorder them in some unexpected way.

The NOCOL and NOPCT options in the TABLES statement suppresses the printing of column percentages and cell percentages, respectively. We really need only the row percentages to be able to interpret the table.

The relevant output produced by this code is as follows:

```
                          Statistics for Table of aspirin by mi

                       Estimates of the Relative Risk (Row1/Row2)
Type of Study               Value       95% Confidence Limits
----------------------------------------------------------------
Case-Control (Odds Ratio)   1.8321        1.4400        2.3308      [This is the point estimate & an
Cohort (Col1 Risk)          1.8178        1.4330        2.3059      approximate 95% CI(OR).]
Cohort (Col2 Risk)          0.9922        0.9892        0.9953
```

Notice that the highlighted values agree with the values presented in our text and in the lecture notes (pp. 20-21).

Chi-Square and Likelihood Ratio Tests

Reading Assignment: Stokes et al., Section 2.2, pp. 20-23
(available on the course website)

Mechanism of Traumatic Injury Example (pp. 28-30 in lecture notes)

The following SAS code (available on the course website) creates a 2x3 table for the data from the Toxicology in Trauma study that examined the association between a positive screen for cocaine and mechanism of injury. It then calculates the chi-square and likelihood ratio test statistics and the approximate p-values based on the chi-square distribution.

```
data cocaine;
input cocaine mechanism count @@;
cards;
1 1 626   1 2 21  1 3 73
2 1 77    2 2 4   2 3 22
;

proc freq order = data; weight count;
   tables cocaine * mechanism / nocol nopct chisq expected;
   title 'Toxicology in Trauma Example';
   title2 'Cocaine Screen vs. Mechanism of Injury';
   title3 'Chi-Square Test';
run;
```

The EXPECTED option in the TABLES statement produces the expected cell frequencies.

The relevant output produced by this SAS code is as follows:

Table of cocaine by mechanism

cocaine mechanism

Frequency Expected Row Pct	1	2	3	Total
1	626	21	73	720
	615.02	21.871	83.111	
	86.94	2.92	10.14	
2	77	4	22	103
	87.982	3.1288	11.889	
	74.76	3.88	21.36	
Total	703	25	95	823

[These are the expected cell frequencies.]

(SAS output continued on following page)

Statistics for Table of cocaine by mechanism

Statistic	DF	Value	Prob
Chi-Square	2	11.6719	0.0029
Likelihood Ratio Chi-Square	2	10.0237	0.0067
Mantel-Haenszel Chi-Square	1	11.6341	0.0006
Phi Coefficient		0.1191	
Contingency Coefficient		0.1183	
Cramer's V		0.1191	

[These are the test statistics, df's and approximate p-values.]

Unfortunately, SAS cannot produce the adjusted residuals presented in the table on p. 28 of the lecture notes. However, SPSS can perform these calculations and the SPSS syntax for this example is provided on the course website.

Tests for Trend

Alcohol and Infant Malformation Example (pp. 32-34 in lecture notes)

The following SAS code (available on the course website) creates a 5x2 table for the data in Table 2.7 in our textbook that examines the association between maternal alcohol consumption and congenital sex organ malformation in newborn infants. The row scores are the midpoints of the categories of alcohol consumption. The SAS code also calculates the test statistic for the test for trend, along with the approximate and exact p-values for this test.

```
data infants;
input alcohol malform count @@;
cards;
0 0   17066  0.5 0 14464  1.5 0   788  4.0 0   126  7.0 0    37
0 1      48  0.5 1    38  1.5 1     5  4.0 1     1  7.0 1     1
;

proc freq  order = data; weight count;
   tables alcohol * malform / trend;
   exact trend;
   title 'Maternal Alcohol vs. Infant Malformation Example';
   title2 'Test for Trend';
   title3 'Grouped Data Midpoint Scores';
   run;
```

The relevant output produced by this SAS code is as follows:

```
Statistics for Table of alcohol by malform

  Cochran-Armitage Trend Test
  ---------------------------
  Statistic (Z)        -2.5632          [This is the test statistic for the test for linear trend.]

  Asymptotic Test
  One-sided Pr <  Z     0.0052          [Approximate p-value based on standard normal.]
  Two-sided Pr > |Z|    0.0104

  Exact Test
  One-sided Pr <=  Z    0.0168          [Exact p-value for test for linear trend.]
  Two-sided Pr >= |Z|   0.0172
```

To specify a different set of row scores, once must change the DATA step accordingly. For example, to use linear scores (e.g., 0, 1, 2, 3, 4) , the following code would be used:

```
data infants;
input alcohol malform count @@;
cards;
0 0   17066  1 0 14464  2 0   788  3 0   126  4 0  37
0 1      48  1 1    38  2 1     5  3 1     1  4 1   1
;
proc freq  order = data; weight count;
   tables alcohol * malform / trend;
   exact trend;
   title 'Maternal Alcohol vs. Infant Malformation Example';
   title2 'Test for Trend';
   title3 'Linear Scores';
   run;
```

and the following output would obtain:

(SAS output is given on following page)

```
Statistics for Table of alcohol by malform

 Cochran-Armitage Trend Test
----------------------------
Statistic (Z)        -1.3520

Asymptotic Test
One-sided Pr <  Z     0.0882
Two-sided Pr > |Z|    0.1764

Exact Test
One-sided Pr <=  Z    0.1046
Two-sided Pr >= |Z|   0.1790
```

Notice how the choice of different row scores changes the conclusion. Also note that use of the exact distribution for the test statistic has very little effect on the p-value, despite the extreme lack of balance in the contingency table.

The Cochran-Armitage test applies only if there are 2 columns. For more than 2 ordinal columns, the following SAS code should be used:

```
data infants;
input alcohol malform count @@;
cards;
0 0   17066  0.5 0 14464  1.5 0   788  4.0 0   126  7.0 0    37
0 1      48  0.5 1    38  1.5 1     5  4.0 1     1  7.0 1     1
;

proc freq  order = data; weight count;
   tables alcohol * malform / measures;
   exact pcorr;
   title 'Maternal Alcohol vs. Infant Malformation Example';
   title2 'Test for Trend';
   title3 'Grouped Data Midpoint Scores';
   run;
```

The relevant output produced by this SAS code is as follows:

```
Statistics for Table of alcohol by malform

Pearson Correlation Coefficient
--------------------------------
Correlation (r)           0.0142
ASE                       0.0106
95% Lower Conf Limit     -0.0066
95% Upper Conf Limit      0.0350

Test of H0: Correlation = 0

ASE under H0              0.0107
Z                         1.3226
One-sided Pr >  Z         0.0930
Two-sided Pr > |Z|        0.1860

Exact Test
One-sided Pr >=  r        0.0168          [This is exact p-value for the general test of linear
Two-sided Pr >= |r|       0.0172          association.]
```

Note that the exact p-value for the general test is identical to that for the Cochran-Armitage test when there are only 2 columns.

Analyses for 2xJ Tables When Columns Are Ordinal

Positive Cocaine Tox Screen vs. Alcohol Disorder Example (pp. 36-37 in lecture notes)

The following SAS code (available on the course website) inputs the data from the Toxicology in Trauma study that examined the association between a positive screen for cocaine and presence of an alcohol disorder. It then performs the Mann-Whitney-Wilcoxon test and produces approximate and exact p-values.

(SAS code given on following page.)

```
data cocaine;
input cocaine disorder count @@;
cards;
1 1 62   1 2 35  1 3 7   1 4 6
2 1 578  2 2 152 2 3 24  2 4 23
;

proc npar1way wilcoxon; FREQ count;
   class cocaine;  var disorder;
   exact wilcoxon;
   title 'Cocaine Screen vs. Alcohol Disorder';
run;
```

The relevant output produced by this SAS code is as follows:

```
Cocaine Screen vs. Alcohol Disorder

The NPAR1WAY Procedure

Wilcoxon Two-Sample Test

Statistic (S)               56700.0000

Normal Approximation
Z                              3.9853
One-Sided Pr >  Z              <.0001
Two-Sided Pr > |Z|             <.0001

t Approximation
One-Sided Pr >  Z              <.0001
Two-Sided Pr > |Z|             <.0001

Exact Test
One-Sided Pr >=  S           6.472E-05
Two-Sided Pr >= |S - Mean|   7.364E-05      [This is the exact p-value for the Mann-Whitney-Wilcoxon test.
                                            We did not specify an alternative hypothesis, so use the 2-tailed p-value.]
```

Note that it doesn't matter in this example what method you use to calculate the p-value. They all yield highly significant results, and the p-value should be reported as "$p < .001$."

Exact Tests for 2x2 Tables

Reading Assignment: Stokes et al., Section 2.5, pp. 23-29 (available on the course website)

Tea-Tasting Example (pp. 40-45 in lecture notes)

The following SAS code (available on the course website) creates the 2x2 table for Fisher's tea tasting experiment. It then performs the χ^2, LR, and Fisher's exact tests and produces approximate and exact p-values for each one. It also finds an approximate and an exact 95% CI(OR).

```
data tea;
input poured guess count @@;
cards;
1 1 3  1 2 1
2 1 1  2 2 3
;

proc freq order=data; weight count;
   tables poured*guess / chisq relrisk;
   exact  CHISQ OR;
   title 'Tea Tasting Example';
   title2 'Fishers Exact Test';
run;
```

The relevant output produced by this SAS code is as follows:

```
       Tea Tasting Example

 The FREQ Procedure

Statistics for Table of poured by guess

Pearson Chi-Square Test
----------------------------------
Chi-Square                   2.0000
DF                                1
Asymptotic Pr >  ChiSq       0.1573          [This is the p-value for the usual chi-square test.]
Exact      Pr >= ChiSq       0.4857          [This is the p-value for the exact version of the chi-square test.]
```

(SAS output is continued on next page.)

```
 Likelihood Ratio Chi-Square Test
 ──────────────────────────────────
 Chi-Square                 2.0930
 DF                              1
 Asymptotic Pr >  ChiSq     0.1480          [This is the p-value for the usual likelihood ratio test.]
 Exact      Pr >= ChiSq     0.4857          [This is the p-value for the exact version of the likelihood ratio test.]

Fisher's Exact Test
 ──────────────────────────────────
Cell (1,1) Frequency (F)         3
Left-sided Pr <= F          0.9857
Right-sided Pr >= F         0.2429          [This is the exact 1-tailed p-value for Fisher's exact test.]

Table Probability (P)       0.2286          [This is the point probability that would be used to calculate the mid p-value.]
Two-sided Pr <= P           0.4857          [This is the exact 2-tailed p-value for Fisher's exact test.]


Estimates of the Relative Risk (Row1/Row2)

Odds Ratio (Case-Control Study)
──────────────────────────────────
Odds Ratio                  9.0000

Asymptotic Conf Limits
95% Lower Conf Limit        0.3666          [This is usual approximate 95% CI(OR).]
95% Upper Conf Limit      220.9270

Exact Conf Limits
95% Lower Conf Limit        0.2117          [This is exact 95% CI(OR).]
95% Upper Conf Limit      626.2435
```

Exact Tests for IxJ Tables

Oral Lesions Example (pp. 45-46 in lecture notes)

The following SAS code (available on the course website) creates the 9x3 table for the data on oral lesions in India. It then performs the χ^2, LR, and Fisher-Freeman-Halton tests and produces approximate and exact p-values for each one.

(SAS Code is given on following page.)

```
data lesions;
input site region count @@;
cards;
1 1 0  1 2 1  1 3 0
2 1 8  2 2 1  2 3 8
3 1 0  3 2 1  3 3 0
4 1 0  4 2 1  4 3 0
5 1 0  5 2 1  5 3 0
6 1 0  6 2 1  6 3 0
7 1 0  7 2 1  7 3 0
8 1 1  8 2 0  8 3 1
9 1 1  9 2 0  9 3 1
;

proc freq order=data; weight count;
   tables site * region / chisq;
   exact chisq Fisher;
   title 'Oral Lesions Example';
   title2 'Fisher-Freeman-Halton Test';
run;
```

The relevant output produced by this SAS code is as follows:

```
Oral Lesions Example

Pearson Chi-Square Test
----------------------------------
Chi-Square                 22.0992
DF                              16
Asymptotic Pr >  ChiSq      0.1400          [These are the approximate and exact p-values for chi-square test.]
Exact      Pr >= ChiSq      0.0269

Likelihood Ratio Chi-Square Test
----------------------------------
Chi-Square                 23.2967
DF                              16
Asymptotic Pr >  ChiSq      0.1060          [These are the approximate and exact p-values for likelihood ratio test.]
Exact      Pr >= ChiSq      0.0356
```

(SAS output continued on next page.)

Oral Lesions Example

The FREQ Procedure

Statistics for Table of site by region

Fisher's Exact Test

Table Probability (P)	5.334E-06
Pr <= P	0.0101

[This is the exact p-value for the Fisher-Freeman-Halton test.]