BIOS 6244 Analysis of Categorical Data
Assignment 3 Solutions

1. Consider Exercise 2.22, p. 50. Combine the data for males and females so that you have a single 2 x 4 table. Write the SAS code to carry out the appropriate analysis of these data, including the DATA step. (Hint: Assume that "Response to Chemotherapy" is ordinal.)

   Solution

   Combining the data as indicated above yields the following 2x4 table:

	Response to Chemo			
Therapy	**Progressive Disease**	**No Change**	**Partial Remission**	**Complete Remission**
Sequential	32	57	34	28
Alternating	53	51	23	21

```
data exercise2_22;
input therapy response count @@;
cards;
1 1 32   1 2 57  1 3 34  1 4 28
2 1 53   2 2 51  2 3 23  2 4 21
;

proc npar1way wilcoxon; FREQ count;
   class therapy;
   var response;
   exact wilcoxon;
   title 'Exercise 2.22, p. 50';
   title2 'Mann-Whitney-Wilcoxon Test';
   run;
```

2. Consider Exercise 2.29, p. 51.

   i. Write the SAS code, including the DATA step, to carry out the analyses requested in parts (a) and (b) of this Exercise.

(Solution continued on next page.)

Exercise (2i), cont.

Solution

```
data exercise2_29;
input x y count @@;
cards;
1 1 4  1 2 2  1 3 0
2 1 2  2 2 2  2 3 2
3 1 0  3 2 2  3 3 4
;

proc freq order=data; weight count;
   tables x * y / measures chisq;
   exact pcorr chisq;
   title 'Exercise 2.29';
   run;
```

ii. Part a. of Exercise 2.29

Solution

The following SAS output provided with the assignment is used to answer this question:

```
Pearson Chi-Square Test
----------------------------------
Chi-Square                 8.0000
DF                              4
Exact      Pr >= ChiSq     0.1642
```

Thus, $X^2 = 8$, df = 4, p = .164. Since .164 > .05, we conclude there is insufficient evidence to conclude that X & Y are not independent. Therefore, based on these data, the assumption of independence appears to be reasonable.

Part b. of Exericse 2.29

Solution

The following SAS output provided with the assignment is used to answer this question:

```
Pearson Correlation Coefficient
----------------------------------
Correlation (r)           0.6667

Exact Test
Two-sided Pr >= |r|       0.0069
```

Thus, the test statistic is r = .67, p = .007. Since .007 < .05, we conclude that X & Y are not independent and that there is a significant linear trend between the row & column variables.

The results in part b. above differ so much from those of part a. because, in the analysis in part b., we are taking advantage of the fact that X & Y are both ordinal. The analysis in part a. above ignores the ordinality of X & Y.

3. Consider Exercise 2.14, p. 48. Table 2.4 is reproduced below.

Table 2.4 Cross Classification of Smoking Status and Myocardial Infarction (MI)

Ever Smoker	Myocardial Infarction	Controls
Yes	172	173
No	90	346

Source: A. Gramenzi et al., *J. Epidemiol. and Commun. Health, 43*: 214–217 (1989).
Reprinted with permission of BMJ Publishing Group.

i. Calculate the usual χ^2 test statistic and the approximate p-value using the appropriate χ^2 distribution.

Solution

For Table 2.4, the row and column totals are as given below:

		MI	Controls	
Ever Smoked	Yes	172	173	345
	No	90	346	436
		262	519	781

For a 2x2 table, the χ^2 test statistic simplifies to

$$X^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1+}n_{2+}n_{+1}n_{+2}} = \frac{(781)[(172)(346) - (90)(173)]^2}{(345)(436)(262)(519)} = 73.73, \text{ df} = (2\text{-}1)(2\text{-}1) = 1.$$

Thus, the approximate p-value calculated using a $\chi^2(1)$ distribution is given by:

$$\Pr(X^2 \geq 73.37 \mid \text{df} = 1) < .0001 \text{ by SimCalc.}$$

This p-value provides extremely strong evidence against the null hypothesis of independence; therefore, we conclude that X & Y are not independent and that there is a significant positive association between MI and smoking among women < 69 years of age.

ii. Calculate the odds ratio and find an approximate 95% CI(OR).

Solution

$$\widehat{OR} = \frac{(172)(346)}{(90)(173)} = 3.82$$

EXTRA CREDIT

An approximate 95% CI[log(OR)] is given by :

$$\log(\widehat{OR}) \pm 1.96\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}} = \log(3.8222) \pm 1.96\sqrt{\frac{1}{172}+\frac{1}{90}+\frac{1}{173}+\frac{1}{346}}$$

$$= 1.3408 \pm 1.96(.1600) = 1.3408 \pm .3136$$

$$= (1.0272, 1.6544)$$

Back-transforming, we obtain an approximate 95% CI(OR): $(e^{1.0272}, e^{1.6544}) = (2.79, 5.23)$.

Thus, there is evidence of a strong positive association between MI and smoking. We are 95% sure that the true OR is as low as 2.79 or as high as 5.23.

4. Work Exercise 2.23, p. 50.

Solution

The following 2x2 table obtains from the information given in the problem:

		Outcome	
		Normalization	Controls
Group	Prednisolone	7	8
	Control	0	15

Recalling that the marginal totals are assumed to be fixed when applying Fisher's exact test, we see that the only table in the reference set that is favorable to the alternative hypothesis that the results were better for treatment than for control is the one that was obtained. Therefore, the upper-tailed exact p-value for Fisher's exact test is given by the hypergeometric probability calculated using Equation (5) on p. 39 of the lecture notes:

$$\text{p-value} = P(n_{11}) = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{+1}-n_{11}}}{\binom{n}{n_{+1}}},$$

where $n_{11} = 7$, $n_{1+} = 15$, $n_{2+} = 15$, $n = 30$, $n_{+1} = 7$:

$$\text{p-value} = P(7) = \frac{\binom{15}{7}\binom{15}{0}}{\binom{30}{7}} = \frac{\left(\frac{15!}{7!8!}\right)(1)}{\left(\frac{30!}{7!23!}\right)} = \frac{143}{45240} = .003.$$

Since $.003 < .05$, we reject H_0 and conclude that prednisolone is significantly better than control in normalizing calcium levels in women with metastatic breast CA.