BIOS 625 Fall 2015 Homework Set 3 Solutions

1. Agresti 2.20

Table 2.12 is from an early study on the death penalty in Florida. Analyze these data and show that Simpson's Paradox occurs.

		Death Penalty	
Victim's Race	**Defendant's Race**	**Yes**	**No**
White	**White**	19	132
	Black	11	52
Black	**White**	0	9
	Black	6	97

Let X, Y, Z denote defendant's race, death penalty, and victim's race, respectively.

For white victims, the odds ratio of death penalty for white defendant to black is:

$$\theta_{XY(Z=W)} = \frac{19 \times 52}{11 \times 132} \approx 0.68 < 1$$

For black victims, the corresponding odds ratio is $\theta_{XY(Z=B)} = 0 < 1$.

However, the marginal odds ratio is $\theta_{XY} = \frac{19 \times 149}{17 \times 141} \approx 1.18 > 1$.

The result shows that Simpson's paradox occurs since the two conditional odds ratios are both less than 1 while the marginal odds ratio is greater than 1. In other words, we have the reversal in the association after controlling for the victim's race.

NOTE: Can also show this by calculating probabilities and then stratified probabilities.

2. Agresti 2.24

Table 2.14 cross classifies job satisfaction by race. Determine whether the groups are stochastically ordered, and estimate the difference between the probability that job satisfaction is higher for blacks than whites and the probability that job satisfactions is higher for whites than blacks.

	Job Satisfaction			
Race	**Dissatisfied**	**Neutral**	**Fairly Satisfied**	**Very or Completely Satisfied**
Black	19	13	42	59
White	47	40	215	430

Sample conditional distributions are:

Black: (19/133, 13/133, 42/133, 59/133)

White: (47/732, 40/732, 215/732, 420/732)

i.e., Black: (0.142, 0.098, 0.316, 0.443)

White: (0.064, 0.055, 0.294, 0.587)

The groups are stochastically ordered, with the black group tending to have more dissatisfaction with their work.

$$\Delta = \frac{\left[59(47+40+215)+42(47+40)+13(47)\right]-\left[19(40+215+430)+13(215+430)+42(430)\right]}{133\times 732}$$

$\Delta = -0.1784$

Δ estimates the difference between the probability that the satisfaction level is higher for black than white groups and the probability that the satisfaction level is higher for white than black groups.

3. Agresti 3.11

Refer to Table 3.11, GSS data on party ID and race.

	Party Identification		
Race	**Democrat**	**Independent**	**Republican**
Black	192	75	8
White	459	586	471

a. Using X^2 and G^2, test the hypothesis of independence between party identification and race. Report the *p*-values and interpret.

$X^2 = 177.3, df = 2, p < 0.001$

$G^2 = 197.4, df = 2, p < 0.001$

Strong evidence of association.

b. Use standardized residuals to describe the evidence of association.

Table of race by party

race	party	Frequency	Std Residual
Black	**Democrat**	192	12.5421
	Independent	75	-3.5986
	Republican	8	-9.7063
	Total	275	
White	**Democrat**	459	-12.5421
	Independent	586	3.5986
	Republican	471	9.7063

Table of race by party

race	party	Frequency	Std Residual
	Total	1516	

Party identification leans towards democrat for blacks and republican for whites.

c. **Partition chi-squared into components regarding the choice between Democrat and Independent and between these two combined and Republican.**

Between Democrat and Independent: $X^2 = 66.63, df = 1, p < 0.001$

Between Democrat/Independent and Republican: $X^2 = 94.1, df = 2, p < 0.001$

4. Agresti 3.12

Using the 2008 GSS, we cross-classified party ID with gender. Table 3.12 shows some results. Explain how to interpret all the results on this printout.

The values $X^2 = 8.29(df = 2, p = 0.0158)$ and $G^2 = 8.31(df = 2, p = 0.0157)$ show considerable evidence against the hypothesis of independence. The standardized Pearson residuals show Female Democrats and Male Independents are much greater than expected under independence, and the number of Female Independents and Male Democrats is significantly less than expected under independence.

5. Agresti 3.16

A study on educational aspirations of high school students measured aspirations with the scale (some high school, high school graduate, some college, college graduate). The student counts in these categories were (9, 44, 13, 10) when family income was low, (11, 52, 23, 22) when family income was middle, and (9, 41, 12, 27) when family income was high.

a. **Test the independence of educational aspirations and family income using X^2 or G^2. Explain the deficiency of this test for these data.**

$X^2 = 8.8709, df = 6, p = 0.181$

$G^2 = 8.9165, df = 6, p = 0.1783$

Both tests indicate not rejecting the null hypothesis. There is no significant association between family income level and student educational aspirations.

Deficiency of the tests for these data: When Pearson or Likelihood ratio chi-square tests are used to test independence between X and Y, the methods themselves ignore ordinal information, so it is possible that the tests gives us a large *p*-value even when there is truly a trend between X and Y.

b. **Find the standardized residuals. Do they suggest any association pattern?**

Table of income by education

income	education	Frequency	Std Residual
low	<HS	9	0.4061
	HS	44	1.5828
	<College	13	-0.1286
	College	10	-2.1078
	Total	76	
middle	<HS	11	-0.1898
	HS	52	-0.5441
	<College	23	1.3042
	College	22	-0.4032
	Total	108	
high	<HS	9	-0.1903
	HS	41	-0.9459
	<College	12	-1.2374
	College	27	2.4360
	Total	89	

The standardized residuals show higher income could be associated with higher educational aspirations.

c. Conduct an alternative test that may be more powerful. Interpret.

Considering the ordinal feature for these data, M^2 would be appropriate to test for independence. Thus, we have $M^2 = (n-1)r^2 = (273-1)(0.1321)^2 = 4.75$ with df=1. This shows strong evidence of association (*p*=0.029). Also you can report the Mantel-Haenszel Chi-square test from SAS. The test statistic is 4.7489 with df=1 and *p*=0.0293, leading to a very similar result.

6. Agresti 3.19

A study in the Department of Wildlife Ecology at the University of Florida sampled wild common carp fish from a wetland in central Chile. One analysis investigated whether the fish muscle had lead pollutant and whether there was evident malformation in the fish. Of 25 fish without lead, 7 had malformation. Of 14 with lead, 7 had malformation. Report and interpret the *p*-value for Fisher's exact test for a one-sided alternative of a greater chance of malformation when there is lead pollutant.

Sample SAS code looks like:

```
data table;
input lead$ malformation$ count @@;
datalines;
no yes 7 no no 18
```

```
yes yes 7 yes no 7
;
proc freq order=data; weight count;
tables lead*malformation;
exact fisher;
run;
```

Try to figure out the correct *p*-value.

```
                 Fisher's Exact Test
        Cell (1,1) Frequency (F)        7
        Left-sided Pr <= F         0.1526
        Right-sided Pr >= F        0.9568

        Table Probability (P)      0.1094
        Two-sided Pr <= P          0.2966
```

7. Agresti 3.26

Using the delta method as in Section 3.1.6, show that the Wald confidence interval for the logit of a binomial parameter π is

$$\log\left[\hat{\pi}/(1-\hat{\pi})\right] \pm z_{\alpha/2}/\sqrt{n\hat{\pi}(1-\hat{\pi})}.$$

Explain how to use this interval to obtain one for π itself.

For the binomial parameter π, we have its MLE $\hat{\pi}$ and estimated variance $\operatorname{var}(\hat{\pi}) = \hat{\pi}(1-\hat{\pi})/n$

Now based on the delta method, the estimated variance for $g(\hat{\pi}) = \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)$ is

$$\begin{aligned}
\operatorname{var}\left[g(\hat{\pi})\right] = \left[g'(\hat{\pi})\right]^2 \operatorname{var}(\hat{\pi}) &= \left[\left(\log\frac{\hat{\pi}}{1-\hat{\pi}}\right)'\right]\operatorname{var}(\hat{\pi}) \\
&= \left\{\frac{1-\hat{\pi}}{\hat{\pi}}\left[\frac{1}{1-\hat{\pi}} + \frac{\hat{\pi}}{(1-\hat{\pi})^2}\right]\right\}^2 \frac{\hat{\pi}(1-\hat{\pi})}{n} \\
&= \frac{1}{n\hat{\pi}(1-\hat{\pi})}
\end{aligned}$$

Thus the 95% CI for the logit of a binomial parameter is: $\log\left[\hat{\pi}/(1-\hat{\pi})\right] \pm z_{\alpha/2}/\sqrt{n\hat{\pi}(1-\hat{\pi})}$.

Based on the 95% CI for the logit of π, we could derive the 95% CI for just π as follows:

$$y = \text{logit}(\pi) = \log\frac{\pi}{1-\pi} \qquad \Rightarrow \qquad \pi = \text{logit}^{-1}(y) = \frac{\exp(y)}{1+\exp(y)}$$

Note: the inverse-logit function is also called the expit function.

Thus the 95% CI for π is:

$$\text{logit}^{-1}\left[\log\left[\frac{\hat{\pi}}{(1-\hat{\pi})}\right] \pm z_{\alpha/2}\Big/\sqrt{n\hat{\pi}(1-\hat{\pi})}\right] \Rightarrow \frac{\exp\left[\log\left[\frac{\hat{\pi}}{(1-\hat{\pi})}\right] \pm z_{\alpha/2}\Big/\sqrt{n\hat{\pi}(1-\hat{\pi})}\right]}{1+\exp\left[\log\left[\frac{\hat{\pi}}{(1-\hat{\pi})}\right] \pm z_{\alpha/2}\Big/\sqrt{n\hat{\pi}(1-\hat{\pi})}\right]}$$

After some algebra, we could see that the Wald (1-α) CI for π is:

$$\left(\frac{\hat{\pi}}{\hat{\pi}+(1-\hat{\pi})\exp\left(\frac{z_{\alpha/2}}{\sqrt{n\hat{\pi}(1-\hat{\pi})}}\right)}, \frac{\hat{\pi}\exp\left(\frac{z_{\alpha/2}}{\sqrt{n\hat{\pi}(1-\hat{\pi})}}\right)}{(1-\hat{\pi})+\hat{\pi}\exp\left(\frac{z_{\alpha/2}}{\sqrt{n\hat{\pi}(1-\hat{\pi})}}\right)}\right)$$

8. Agresti 3.34

For a 2 x 2 table, show that:

a. The four Pearson residuals may take different values.

Suppose we have a 2x2 table, where the column total is labeled as $n_{+j}\,(j=1,2)$ and row total is $n_{i+}\,(i=1,2)$. The overall total is $n_{++}=N$.

The Pearson residuals for any of these 4 cells would be calculated as follows, indicating that it's likely that the Pearson residuals may take different values:

$$e_{ij} = \frac{n_{ij}-\hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}} = \frac{n_{ij}-\frac{n_{i+}n_{+j}}{N}}{\sqrt{\frac{n_{i+}n_{+j}}{N}}} = n_{ij}\sqrt{\frac{N}{n_{i+}n_{+j}}} - \sqrt{\frac{n_{i+}n_{+j}}{N}}$$

b. All four standardized residuals have the same absolute value.

The absolute value of the standardized residual would be calculated like this:

$$\left|r_{ij}\right| = \frac{\left|n_{ij} - \hat{\mu}_{ij}\right|}{\sqrt{\hat{\mu}_{ij}\left(1-p_{i+}\right)\left(1-p_{+j}\right)}} = \frac{\left|n_{ij} - \frac{n_{i+}n_{+j}}{N}\right|}{\sqrt{\frac{n_{i+}n_{+j}}{N}\left(1-\frac{n_{i+}}{N}\right)\left(1-\frac{n_{+j}}{N}\right)}} = \frac{\frac{1}{N}\left|n_{ij}N - n_{i+}n_{+j}\right|}{\frac{1}{N}\sqrt{n_{i+}n_{+j}n_{i'+}n_{+j'}\frac{1}{N}}} = \frac{\left|n_{ij}N - n_{i+}n_{+j}\right|}{\sqrt{n_{i+}n_{+j}n_{i'+}n_{+j'}\frac{1}{N}}}$$

From the formula above, for every four $\left|r_{ij}\right|\left(i=1,2;\, j=1,2\right)$, the denominator includes all four marginal totals (two row totals and two column totals) and the overall totals for this 2x2 table. Therefore, the denominator would be the same for four $\left|r_{ij}\right|\left(i=1,2;\, j=1,2\right)$, no matter which cell is chosen. Regarding the numerator, we can have the following:

$$\begin{aligned}
\left|n_{ij}N - n_{i+}n_{+j}\right| &= \left|n_{ij}\left(n_{i+}+n_{i'+}\right) - n_{i+}\left(n_{ij}+n_{i'j}\right)\right| \\
&= \left|n_{ij}n_{i+} + n_{ij}n_{i'+} - n_{i+}n_{ij} - n_{i+}n_{i'j}\right| \\
&= \left|n_{ij}n_{i'+} - n_{i+}n_{i'j}\right| \\
&= \left|n_{ij}\left(n_{i'j}+n_{i'j'}\right) - \left(n_{ij}+n_{ij'}\right)n_{i'j}\right| \\
&= \left|n_{ij}n_{i'j'} - n_{ij'}n_{i'j}\right|
\end{aligned}$$

The numerator is just the absolute value of the cross product for this 2x2 table no matter which cell you choose. In other words, the numerator is the same when you calculate any of four absolute standardized residuals. Therefore, all four standardized residuals would have the same absolute values.

c. The square of each standardized residual equals X^2.

For the 2x2 table we have the formula:

$$X^2 = \frac{n\left(n_{11}n_{22} - n_{12}n_{21}\right)^2}{n_{1+}n_{2+}n_{+1}n_{+2}}$$

Also the square of each standardized residual is:

$$r_{ij}^2 = \frac{\left(n_{ij}n_{i'j'} - n_{ij'}n_{i'j}\right)^2}{n_{i+}n_{+j}n_{i'+}n_{+j'}\frac{1}{N}} = \frac{N\left(n_{11}n_{22} - n_{12}n_{21}\right)^2}{n_{1+}n_{2+}n_{+1}n_{+2}}$$

Therefore, we can see that the square of each standardized residual equals X^2.