

PubH 7407, Spring 2007: Midterm

There are five problems, each with multiple parts. Please start a new page for each problem. Put your name at the top of each page.

1. **Brief answer.** Answer the following questions with one or two sentences.

- (a) What does $\gamma = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d}$ measure? For what type of data is this measure appropriate?
- (b) Consider the following table of probabilities $\pi_{ij} = P(X = i, Y = j)$:

| | | |
|---------|---------|---------|
| | $Y = 1$ | $Y = 2$ |
| $X = 1$ | 0.10 | 0.30 |
| $X = 2$ | 0.05 | 0.15 |
| $X = 3$ | 0.10 | 0.30 |

Are X and Y independent? That is, is $X \perp Y$?

- (c) For the previous table, what is $P(X = Y)$?
- (d) $n_{++} = 200$ people are randomly sampled and cross-classified according to their gender and political affiliation in the following table:

| | | | | |
|--------|----------|------------|----------|----------------|
| | Democrat | Republican | Other | |
| Male | n_{11} | n_{12} | n_{13} | n_{1+} |
| Female | n_{21} | n_{22} | n_{23} | n_{2+} |
| | n_{+1} | n_{+2} | n_{+3} | $n_{++} = 200$ |

Is this an example of *multinomial* or *product multinomial* sampling? Why?

- (e) $n_{1+} = 100$ men and $n_{2+} = 100$ women are randomly sampled within their gender and classified according to their political affiliation in the following table:

| | | | | |
|--------|----------|------------|----------|----------------|
| | Democrat | Republican | Other | |
| Male | n_{11} | n_{12} | n_{13} | $n_{1+} = 100$ |
| Female | n_{21} | n_{22} | n_{23} | $n_{2+} = 100$ |
| | n_{+1} | n_{+2} | n_{+3} | $n_{++} = 200$ |

Is this an example of *multinomial* or *product multinomial* sampling?

2. **True/false.** Write **true** or **false** for each statement.

- (a) In a 2×2 table, the odds ratio $\theta = 1$ is equivalent to $X \perp Y$.
- (b) The odds ratio, relative risk, and difference in proportions are all valid measures for summarizing a 2×2 tables in a case-control study.
- (c) For testing independence with in an $I \times J$ contingency table from a random sample, Pearson's X^2 and the LRT statistic G^2 both have $\chi^2_{(I-1)(J-1)}$ distributions for any sample size.
- (d) In parts 1(d) and 1(e) it does not matter how the data are sampled when determining if X is related to Y using the X^2 and G^2 test of association. That is, the p -values are the same.

- (e) Say X and Y are binary and Z has K categories. One can test $X \perp Y|Z$ using a Wald or likelihood ratio test of $H_0 : \beta = 0$ in the logistic regression model:

$$\text{logit } P(Y = 1) = \alpha + \beta X + \beta_1 z_1 + \cdots + \beta_{K-1} z_{K-1},$$

where $z_k = 1$ for observations in category k of Z and $z_k = 0$ otherwise.

3. Consider the horseshoe crab data from your text. Recall that color is 1, 2, 3, 4 for light-medium, medium, dark-medium, and dark. We wish to investigate how color and weight affect the probability of having one or more satellites. The following SAS program was fit:

```
data crabs1;
input color spine width satell weight;
  weight=weight/1000; color=color-1;
  y=0; n=1; if satell>0 then y=1;
datalines;
3 3 28.3 8 3050
4 3 22.5 0 1550
2 1 26.0 9 2300
...
2 1 28.0 0 2625
5 3 27.0 0 2625
3 2 24.5 0 2000
;
proc logistic data=crabs1 descending;
  class color / param=ref; model y = color weight color*weight / lackfit;
run;
```

Yielding the annotated output:

Class Level Information

| Class | Value | Design | Variables |
|-------|-------|--------|-----------|
| color | 1 | 1 | 0 0 |
| | 2 | 0 | 1 0 |
| | 3 | 0 | 0 1 |
| | 4 | 0 | 0 0 |

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|----------------|----|----------|----------------|-----------------|------------|
| Intercept | 1 | -1.1868 | 2.2999 | 0.2663 | 0.6058 |
| color | 1 | -0.4335 | 5.4046 | 0.0064 | 0.9361 |
| color | 2 | -1.2654 | 2.5847 | 0.2397 | 0.6244 |
| color | 3 | -6.7289 | 3.4351 | 3.8371 | 0.0501 |
| weight | 1 | 0.1947 | 1.0303 | 0.0357 | 0.8501 |
| weight*color 1 | 1 | 0.8536 | 2.1551 | 0.1569 | 0.6920 |
| weight*color 2 | 1 | 1.2149 | 1.1419 | 1.1319 | 0.2874 |
| weight*color 3 | 1 | 3.5596 | 1.5633 | 5.1846 | 0.0228 |

Hosmer and Lemeshow Goodness-of-Fit Test

| Chi-Square | DF | Pr > ChiSq |
|------------|----|------------|
| 3.6059 | 8 | 0.8908 |

- (a) Describe the association of color and weight on the estimated odds of having one or more satellites. In particular, describe how the estimated odds of having one or more satellites changes from dark-medium to dark, and when increasing the weight by one unit (here, kg).

- (b) Is there any indication of lack-of-fit? Why or why not?
- (c) Write the two estimated probability functions $\hat{\pi}(w, c)$ where (w, c) is (weight, color), for arbitrary w and $c = 3$ and $c = 4$, the dark-medium and dark crabs.
- (d) For crabs that weigh 3.0 kg, what is the estimated *relative risk* of having a satellite for dark versus medium-dark crabs?
4. The following SAS program was used to assess the marginal relationship of color to the probability of having one or more satellites, note the `link=identity` here:

```
proc genmod data=crabs1 descending;
class color; model y=color / dist=bin link=identity;
```

The output is:

| Analysis Of Parameter Estimates | | | | | | | | |
|---------------------------------|----|----------|----------------|--------|-----------------------|------------|------|--------|
| Parameter | DF | Estimate | Standard Error | Wald | 95% Confidence Limits | Chi-Square | Pr > | ChiSq |
| Intercept | 1 | 0.3182 | 0.0993 | 0.1236 | 0.5128 | 10.27 | | 0.0014 |
| color | 1 | 0.4318 | 0.1596 | 0.1189 | 0.7447 | 7.32 | | 0.0068 |
| color | 2 | 0.4081 | 0.1093 | 0.1938 | 0.6224 | 13.94 | | 0.0002 |
| color | 3 | 0.2727 | 0.1239 | 0.0299 | 0.5156 | 4.84 | | 0.0277 |
| color | 4 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | | . |

- (a) What is the estimated relative risk of having one or more satellites for medium-light versus dark crabs?
- (b) What is the estimated difference in the probabilities of having one or more satellites for light-medium versus dark crabs? Test that this difference is significant at the 5% level.
5. A study on the educational aspirations of high school students measured aspiration $X = 1, 2, 3, 4$ for levels (some high school, high school graduate, some college, college graduate). Also recorded was $Y = 1, 2, 3$ the family income level (low, middle, and high). The data are

| | Low | Middle | High |
|----------------------|-----|--------|------|
| Some high school | 9 | 11 | 9 |
| High school graduate | 44 | 52 | 41 |
| Some college | 13 | 23 | 12 |
| College graduate | 10 | 22 | 27 |

The following code was used to analyze these data:

```
data table;
input Aspiration$ Income$ count @@;
datalines;
1 1 9 2 1 44 3 1 13 4 1 10 1 2 11 2 2 52 3 2 23 4 2 22 1 3 9 2 3 41 3 3 12 4 3 27
;
proc freq order=data; weight count;
tables Aspiration*Income / expected chisq plcorr;
proc genmod order=data; class Aspiration Income;
model count = Aspiration Income / dist=poi link=log residuals;
```

With the following output:

The FREQ Procedure

Table of Aspiration by Income

| Aspiration | Income | | | |
|------------|--------|--------|--------|-------|
| Frequency | | | | |
| Expected | 1 | 2 | 3 | Total |
| 1 | 9 | 11 | 9 | 29 |
| | 8.0733 | 11.473 | 9.4542 | |
| 2 | 44 | 52 | 41 | 137 |
| | 38.139 | 54.198 | 44.663 | |
| 3 | 13 | 23 | 12 | 48 |
| | 13.363 | 18.989 | 15.648 | |
| 4 | 10 | 22 | 27 | 59 |
| | 16.425 | 23.341 | 19.234 | |
| Total | 76 | 108 | 89 | 273 |

Statistics for Table of Aspiration by Income

| Statistic | DF | Value | Prob |
|-----------------------------|----|--------|--------|
| Chi-Square | 6 | 8.8709 | 0.1810 |
| Likelihood Ratio Chi-Square | 6 | 8.9165 | 0.1783 |

Statistics for Table of Aspiration by Income

| Statistic | Value | ASE |
|------------------------|--------|--------|
| Gamma | 0.1625 | 0.0795 |
| Polychoric Correlation | 0.1491 | 0.0722 |

The GENMOD Procedure

| Observation | Resraw | Reschi | Resdev | StResdev | StReschi | Reslik |
|-------------|-----------|-----------|-----------|-----------|-----------|-----------|
| 1 | 0.9267389 | 0.3261617 | 0.3202024 | 0.3987119 | 0.4061323 | 0.4013622 |
| 2 | 5.8608012 | 0.9490109 | 0.926141 | 1.5446752 | 1.582819 | 1.5692137 |
| 3 | -0.362639 | -0.099204 | -0.099658 | -0.129226 | -0.128637 | -0.128988 |
| 4 | -6.424924 | -1.585318 | -1.710424 | -2.274189 | -2.107847 | -2.203483 |
| 5 | -0.472527 | -0.139507 | -0.140482 | -0.191137 | -0.189812 | -0.190529 |
| 6 | -2.1978 | -0.298536 | -0.300589 | -0.547803 | -0.544062 | -0.545191 |
| 7 | 4.0109894 | 0.9204503 | 0.8906041 | 1.2618685 | 1.3041566 | 1.2832659 |
| 8 | -1.340677 | -0.277503 | -0.280225 | -0.407119 | -0.403164 | -0.405042 |
| 9 | -0.454212 | -0.147722 | -0.14893 | -0.191884 | -0.190329 | -0.191268 |
| 10 | -3.663004 | -0.548105 | -0.555865 | -0.959298 | -0.945905 | -0.950423 |
| 11 | -3.648352 | -0.922279 | -0.962127 | -1.290907 | -1.237442 | -1.26742 |
| 12 | 7.7655521 | 1.770649 | 1.6679729 | 2.2947526 | 2.4360117 | 2.3624328 |

- Test $H_0 : X \perp Y$ using X^2 or G^2 ; what do you conclude? Are these tests approximately valid here?
- Are these data nominal or ordinal? If ordinal, are there any other tests of association you might consider? Describe the association with an estimate and 95% CI. Note that $z_{0.025} = 1.96$. What do you conclude?
- Create a table of “+” and “-” for the signs of the standardized Pearson residuals. Do you see any patterns? if so, describe.