BIOS625, Fall 2015: Midterm Samples

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 part (c), 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.

3. 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

(a) Test $H_0 : X \perp Y$ using X^2 or G^2; what do you conclude? Are these tests approximately valid here?

(b) 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?

(c) Create a table of "+" and "−" for the signs of the standardized Pearson residuals. Do you see any patterns? if so, describe.

4. Consider data relating political affiliation (Democrat, Republican, or Independent) to the college of enrollment of U.S. university students (Letters – essentially literature, Engineering, Agriculture, or Education). SAS's PROC FREQ and PROC GENMOD produce the following table of observed and expected counts, likelihood ratio and Pearson tests for independence, as well as the standardized Pearson residuals (the `r` below).

Table of College by Affiliation

College (Frequency / Expected)	Republican	Democrat	Independent	Total
Letters	34	61	16	111
	38.313	50.845	21.842	
Engineering	31	19	17	67
	23.126	30.69	13.184	
Agriculture	19	23	16	58
	20.019	26.568	11.413	
Education	23	39	12	74
	25.542	33.897	14.561	
Total	107	142	61	310

Statistics for Table of College by Affiliation

Statistic	DF	Value	Prob
Chi-Square	6	16.1613	0.0129
Likelihood Ratio Chi-Square	6	16.3901	0.0118

Obs	College	Affiliation	count	r
1	Letters	Republican	34	-1.07469
2	Letters	Democrat	61	2.41451
3	Letters	Independent	16	-1.74079
4	Engineering	Republican	31	2.28541
5	Engineering	Democrat	19	-3.23767
6	Engineering	Independent	17	1.32451
7	Agriculture	Republican	19	-0.31226
8	Agriculture	Democrat	23	-1.04285
9	Agriculture	Independent	16	1.68036
10	Education	Republican	23	-0.71235
11	Education	Democrat	39	1.36463
12	Education	Independent	12	-0.85835

(a) Do you accept or reject that the college of enrollment is independent of political affiliation? Why or why not? Comment on the validity of the test's p-value in terms of the expected cell counts.

(b) Are any cells particularly ill-fit by the model of independence? If so, for which college(s) does this occur? Are any pairs of colleges particularly "unlike" each other in terms of political affiliation?

Combining Letters, Agriculture, and Education into one category called Other:

Table of College by Affiliation

College (Frequency / Expected)	Affiliation: Republican	Democrat	Independent	Total
Engineering	31	19	17	67
	23.126	30.69	13.184	
Other	76	123	44	243
	83.874	111.31	47.816	
Total	107	142	61	310

Statistics for Table of College by Affiliation

Statistic	DF	Value	Prob
Chi-Square	2	10.5103	0.0052
Likelihood Ratio Chi-Square	2	10.8539	0.0044

Omitting Engineering from the table:

Table of College by Affiliation

College (Frequency / Expected)	Affiliation: Republican	Democrat	Independent	Total
Letters	34	61	16	111
	34.716	56.185	20.099	
Agriculture	19	23	16	58
	18.14	29.358	10.502	
Education	23	39	12	74
	23.144	37.457	13.399	
Total	76	123	44	243

Statistics for Table of College by Affiliation

Statistic	DF	Value	Prob
Chi-Square	4	5.7698	0.2170
Likelihood Ratio Chi-Square	4	5.5361	0.2366

(c) Verify that $G_1^2 + G_2^2$ for the collapsed and reduced tables above add up to G^2 for the full table on the previous page. Verify that $df_1 + df_2 = df$ as well.

(d) Partitioning the chi-squared G^2 attempts to locate *why* the original test of $H_0 : X \perp Y$ is rejected. Carefully interpret the followup tests for independence in the collapsed and partial tables. What do you conclude about political affiliation and college of enrollment among U.S. university students?