BIOS 625: Categorical Data Analysis & Generalized Linear Models
Spring 2018
Midterm I SOLUTIONS [100 points]

NAME:

INSTRUCTIONS:

The following questions comprise Exam 1. Each question should be completed on the blank white paper supplied. You may combine answers on the same page, or you may choose to use one page per question. In either case, ensure that you have placed you name on each page. Once the exam has been completed, staple all pages in the order of the questions.

Note: If you do not have a calculator, you may leave the calculations in the form of a fraction. Also, there will be part marking. So, try to go as far as possible for each problem.

1. [10 points] For testing $H_0 = \pi_j = \pi_{j0}, j = 1, \ldots, c$ using sample multinomial proportions $\hat{\pi}_j$, show that the likelihood-ratio statistic $G^2 = -2n \sum_j \hat{\pi}_j \log(\pi_{j0}/\hat{\pi}_j) \geq 0$, with equality if and only if $\hat{\pi}_j = \pi_{j0}$ for all j.

   Answer: See Homework.

2. [10 points] For counts n_i, the *power divergence statistic* for testing goodness-of-fit is

$$\frac{2}{\lambda(\lambda+1)} \sum n_i[(n_i/\hat{\mu}_i)^\lambda - 1], -\infty < \lambda < \infty$$

   (a) For $\lambda = 1$, show that this equals χ^2.

   Answer: See Homework.

3. [10 points] An article in the New York Times (Feb. 17, 1999) about the PSA blood test for detecting prostate cancer stated "The test fails to detect prostate cancer in 1 in 4 men who have the disease (false-negative results), and as many as two-thirds of the men tested receive false positive results." Let $D+$ be the event that a randomly drawn man has the disease and $T+$ be the event that the PSA blood test comes up positive; denote $D-$ and $T-$ as the compliments of these events.

   (a) What is the sensitivity and specificity of the PSA blood test?

   (b) Say that in a certain subpopulation, the disease prevalence is $P(D+) = 0.13$. What is the predictive value positive in this subpopulation $P(D+|T+)$?

   (c) **Bonus, 5 points**: A diagnostic test has specificity 0.98 and sensitivity 0.92. Find the odds ratio between the true disease status and the diagnostic test results.

   Answer:

   a. The sensitivity is $P(T+|D+) = 1 - P(T-|D+) = 1 - 1/4 = 3/4$ and the specificity is $P(T-|D-) = 1 - P(T+|D-) = 1 - 2/3 = 1/3$.

   b.

$$\begin{aligned} P(D+|T+) &= \frac{P(T+|D+)P(D+)}{P(T+|D+)P(D+) + P(T+|D-)P(D-)} \\ &= \frac{(3/4)0.13}{(3/4)0.13 + (2/3)0.87} \\ &= 0.144 \end{aligned}$$

c. (Bonus) Sensitivity is $P(T+|D+)$ and specificity is $P(T-|D-)$, so

$$\frac{\frac{P(D+|T+)}{P(D-|T+)}}{\frac{P(D+|T-)}{P(D-|T-)}} = \frac{\frac{P(T+|D+)}{P(T-|D+)}}{\frac{P(T+|D-)}{P(T-|D-)}} = \frac{\frac{0.92}{0.08}}{\frac{0.02}{0.98}} = 563.5.$$

4. [20 points] The following data results from a Case Control study designed to estimate the effects of alcohol consumption with esophageal cancer. Denote
p_1 as the $P(\text{CASE}|80+\text{ mg/day})$,
p_2 as the $P(\text{CASE}|0\text{ - }79\text{ mg/day})$,
π_1 as the $P(80+\text{ mg/day}|\text{CASE})$ and
π_2 as the $P(80+\text{ mg/day}|\text{CONTROL})$.
When appropriate, estimate each of these quantities and provide a valid summary measure of association. For the presented measure of association, provide a valid interpretation.

		Esophageal Cancer		
		Case	Control	
Alcohol	80+ (mg/day)	96	109	205
	0-79 (mg/day)	104	666	770
	Total	200	775	975

Answer: Please see Lecture notes # 6 [page 27 onwards], where it is stated that the odds ratio $OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$ can be estimated as: $OR = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$, since the OR can be equivalently defined in terms of the p's or the π's. From the Table, we have:

$$\pi_1 = \frac{96}{200}$$

and

$$\pi_2 = \frac{109}{775}.$$

Thus, the (estimated) odds ratio is:

$$\widehat{OR} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = 5.640$$

Interpretation: Estimates odds of esophageal cancer among High alcohol group (80+ mg/day) is 5.64 times the estimated odds for the Low alcohol group. Below, you will also see that the 95% CI of this OR *excludes* 1, which means that it is significant.
Now, the 95% CI for the odds ratio is:

$$\exp\{\log(\widehat{OR}) \pm 1.96\sqrt{\widehat{Var}[\log(\widehat{OR})]}\}$$

Using formulas [see notes],

$$\widehat{Var}[\log(\widehat{OR}) = 1/96 + 1/109 + 1/104 + 1/666 = 0.0307$$

Hence, the 95% CI is:

$$(\exp(\log(5.640) - 1.96 * \text{sqrt}(0.0307)), \exp(\log(5.640) + 1.96 * \text{sqrt}(0.0307))),$$

which is

$$(4.001, 7.951)$$

Caveat: You can easily check the results obtained here by looking at Page 29 in Lecture 6.

5. [50 points] The following `SAS` code reads in data that relates education to attitude toward legalized abortion from a General Social Survey.

```
data abort; input school$ attitude$ count @@;
datalines;
1 1 209 1 2 101 1 3 237
2 1 151 2 2 126 2 3 426
3 1  16 3 2  21 3 3 138
;
proc format;
 value $sc ’1’=’< high school’ ’2’=’high school’ ’3’=’> high school’;
 value $ac ’1’=’generally disapprove’ ’2’=’middle position’ ’3’=’generally approve’;
proc freq; weight count;
 format school $sc. attitude $ac.;
 table school*attitude/ expected chisq plcorr nopercent nocol norow;
```

The output is:

```
  school         attitude

  Frequency     |
  Expected      |generall|middle p|generall|  Total
                |y disapp|osition |y approv|
                |rove    |        |e       |
  --------------+--------+--------+--------+
  < high school |    209 |    101 |    237 |    547
                | 144.33 | 95.197 | 307.47 |
  --------------+--------+--------+--------+
  high school   |    151 |    126 |    426 |    703
                | 185.49 | 122.35 | 395.16 |
  --------------+--------+--------+--------+
  > high school |     16 |     21 |    138 |    175
                | 46.175 | 30.456 | 98.368 |
  --------------+--------+--------+--------+
  Total              376      248      801     1425

      Statistics for Table of school by attitude

Statistic                     DF       Value     Prob
------------------------------------------------------
Chi-Square                     4     93.0338    <.0001
Likelihood Ratio Chi-Square    4     96.5267    <.0001

                                       Value  Std. Error
------------------------------------------------------
Gamma                                 0.3873    0.0366
Pearson Correlation                   0.2530    0.0240
Polychoric Correlation                0.3432    0.0325


                  Sample Size = 1425
```

a. [10 points] Is ‘education level’ ordinal or nominal? Is ‘attitude toward legalized abortion’ ordinal, nominal, interval, methodological, or continuous?
Answer: Both variables are ordinal.

b. [10 points] Formally test that ‘education level’ and ‘attitude toward legalized abortion’ are independent using X^2 and/or G^2. Are these tests valid here? Why or why not?
Answer: The Pearson $X^2 = 93.0$ with p-value < 0.0001; the likelihood ratio test $G^2 = 96.5$ with p-value < 0.0001. Either test leads us to strongly reject $H_0 : X \perp Y$. Yes these tests are valid. All expected cell counts are at least one; in fact they’re all over ten.

c. [10 points] Create a 3×3 table of '+' and '−' for each cell based on the sign of the Pearson (or raw) residual. Describe *and interpret* any pattern that you see.

Answer:

+	+	−
−	+	+
−	−	+

We see, *qualitatively*, larger counts along the diagonal (and just above the diagonal) than what we'd expect under independence. This is in line with "concordant" outcomes of more approval with higher education. Information on whether the residuals are 'significantly' larger than what we'd expect is not provided.

d. [10 points] The gamma, Pearson correlation, and polychoric correlation statistics are also reported. Obtain a 95% CI for each statistic.

Answer: For gamma, the CI is $(0.316, 0.459)$, for Pearson based on default scores $(0.206, 0.300)$, and for polychoric $(0.280, 0.407)$. These are all obtained as the MLEs ± 1.96 times their standard error.

e. [10 points] Briefly (i.e. in one sentence each, not formulas) describe what each statistic in part (d) measures and formally test that these are zero.

Answer: All statistics are between -1 and 1 measure the strength of a particular kind of trend, pattern, or association in the data. The gamma statistic estimates the probability of concordance versus the probability of discordance. The Pearson correlation is literally the Pearson correlation based on replacing variable levels with scores, here $\{1, 2, 3\}$ for both X and Y. The polychoric correlation is the maximum likelihood estimate of the correlation of underlying continuous *latent* variables (Z_1, Z_2) that have a bivariate normal distribution.

The Pearson statistic measures 'linear trend' in the ordinal variables, the gamma statistic is said to measure 'monotone association.'

All measures show a positive, weak to moderate but significant association between education and attitude. We would reject than any of the three measures are zero at the $\alpha = 5\%$ significance level based on the CIs reported in part (d).