Exam I (2021)

Introduction to Categorical Data Analysis

1. (30 points) For each of the following statements, answer true (T) or false (F):

   (a) Eduation (<high school, high school, college, graduate school) is a ordinal variable.

   (b) Cohort study belongs to prospective design.

   (c) In contingency tables used to evaluate the results from a logistic regression analysis, the sum of the sensitivity and specificity will always equal one.

   (d) The likelihood-ratio goodness-of-fit test statistic for a logit model assesses the adequacy of fit of the model by comparing how well it conforms to the data as opposed to the saturated logit model.

   (e) For an $I \times J$ table, $\binom{I}{2}\binom{J}{2}$ odds ratios may be constructed; these odds ratios are all functionally independent.

   (f) It is possible to construct a 2×2 table of joint probabilities π_{ij} in which $\pi_{ij} = \pi_{i+}\pi_{+j}$ (for $i = 1, 2$; $j = 1, 2$), and yet the odds ratio θ is unequal to 1.

   (g) For a sample of retired subjects in Florida, a contingency table is used to relate X =chelesterol (8 ordered levels) to Y =whether the subject has symptoms of heart disease (yes= 1, no= 0). For the linear logit model $\text{logit}P(Y = 1) = \alpha + \beta x$ fitted to the 8 binomials in the 8×2 contingency table by assigning scores to the 8 cholesterol levels, the deviance statistic equals 6.0. Thus, this model provides a poor fit to the data.

   (h) In the example just mentioned in (g), at the lowest cholesterol level, the observed number of heart disease cases equals 31. The standardized residual equals 1.35. This mean that the model predicted 29.65 cases (i.e., 1.35=31-29.65).

   (i) The difference of proportions, relative risk, and odds ratio are valid measures for summarizing 2×2 tables for either prospective or retrospective (e.g., case-control) studies.

   (j) A British study reported in the *New York Times*: (Dec. 3, 1998) stated that of smokers who get lung cancer, "women were 1.7 times more vulnerable than men to get small-cell lung cancer." The number 1.7 is a sample odds ratio.

   (k) For testing independence with random samples, Pearson's χ^2 statistic and the likelihood-ratio G^2 statistic both have chi-squared distributions for any sample size, as long as the sample was randomly selected.

   (l) Fisher's exact test is a test of the null hypothesis of independence for 2×2 contingency tables that fixes the row and column totals and uses a hypergeometric distribution for the count in the first cell. For a one-sided alternative of a positive association (i.e., odds ratio > 1), the p-value is the sum of the probabilities of all those tables that have count in the first cell at least as large as observed, for the given marginal totals.

(m) Probit models are generalized linear models in which the random component is normal and the link is the inverse of a normal cdf.

(n) For generalized linear models, the Wald test of $\beta = 0$ is equivalent to the test based on the likelihood-ratio statistic, in the sense that they give identical p-values.

(o) If X and Y are marginally independent, X and Z are marginally independent, and Y and Z are marginally independent, then X, Y, and Z are necessarily mutually independent.

2. (35 points) A common perception among sports analysts is that a team playing on its home field is more likely to win than a team which is visiting. The 2×2 contingency table below cross classifies all 162 games played in the 2000 season by the St. Louis Cardinals according to (i) whether the game was played at home or away (X), and (ii) whether the game was won or lost (Y). The accompanying SAS output provides an analysis of the table. Use the output to address the questions which follow.

(a) The output provides p-values for three separate tests for the hypothesis of independence of X and Y. Name each test and provide the corresponding p-value. Summarize the conclusions in terms of the context of the application.

(b) What is the estimate of the relative risk $\pi_{1|1}/\pi_{1|2}$? Provide a brief interpretation of this statistic in terms of the context of the application.

(c) SAS provides an approximate 95% confidence interval for the odds ratio θ. Explain how such a confidence interval may be used for testing the hypothesis of independence of X and Y. In the present application, do the confidence interval and the tests in part (a) convey the same conclusion? Briefly explain.

```
          The FREQ Procedure
       Table of field by win_loss
Frequency|
Percent  |
Row Pct  |
Col Pct  |Win     |loss    | Total
----------------------------
Home     |     50 |     31 |     81
         |  30.86 |  19.14 |  50.00
         |  61.73 |  38.27 |
         |  52.63 |  46.27 |
----------------------------
Away     |    45 |    36 |    81
         | 27.78 | 22.22 | 50.00
         | 55.56 | 44.44 |
         | 47.37 | 53.73 |
----------------------------
Total         95      67    162
           58.64    41.36   100.00

      Statistics for Table of field by win_loss
Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1      0.6363    0.4251
Likelihood Ratio Chi-Square    1      0.6368    0.4249
Continuity Adj. Chi-Square     1      0.4072    0.5234
Mantel-Haenszel Chi-Square     1      0.6324    0.4265
Phi Coefficient                       0.0627
Contingency Coefficient               0.0625
Cramer's V                            0.0627

       Fisher's Exact Test
----------------------------------
Cell (1,1) Frequency (F)        50
Left-sided Pr <= F          0.8308
Right-sided Pr >= F         0.2618

Table Probability (P)       0.0925
Two-sided Pr <= P           0.5235

           Statistics for Table of field by win_loss
           Estimates of the Relative Risk (Row1/Row2)
Type of Study                   Value       95% Confidence Limits
-------------------------------------------------------------------
Case-Control (Odds Ratio)      1.2903        0.6894        2.4149
Cohort (Col1 Risk)             1.1111        0.8571        1.4403
Cohort (Col2 Risk)             0.8611        0.5957        1.2448

                     Sample Size = 162
```

3. (35 points) The following SAS output features the results of a logistic regression analysis. The objective of the analysis is to determine whether a student's final exam grade in STAT200 can be effectively used to predict whether the student will earn an A in the course.

   (a) Perform a hypothesis test based on the likelihood-ratio test statistic to determine the goodness-of-fit of the linear logit model.

(b) Consider a student who earns a 180 on the final exam. According to the fitted model, what are the odds the student will earn an A for his/her term grade? What is the probability the student will earn an A for his/her term grade?

(c) Assume that student Z's final exam grade is 1 unit higher than that of student V, what is the odds ratio of earning an A for student Z vs. student V. Find a 95% confidence interval for the odds ratio and interpret.

```
                                The GENMOD Procedure

                           Ordered                 Total
                             Value   choice    Frequency

                                 1        1           18
                                 2        0           31


                   Deviance and Pearson Goodness-of-Fit Statistics

Criterion                  Value           DF        Value/DF        Pr>ChiSq

Deviance                 20.0285           38          0.5271          0.9927
Pearson Chi-Square       22.9104           38          0.6029          0.9747

                            Number of Observations Read          40
                            Number of Observations Used          40

                      Analysis of Parameter Estimates

                                                Standard
Parameter        DF            Estimate            Error   Chi-Square  Pr > ChiSq
Intercept         1            -25.5250           7.5672      11.3779      0.0007
FIN_EXAM          1              0.1496           0.0442      11.4414      0.0007
```