**STA 4504/5503** **Name**: _____

**Sample test questions for Exam 1**

1. For the following statements, answer true(T) or false(F).

   a. _____ In 2×2 tables, statistical independence is equivalent to a population odds ratio value of $\theta = 1.0$.

   b. _____ 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.

   c. _____ Using data from the Harvard Physician's Health Study, we find a 95% confidence interval for the relative risk relating having a heart attack to drug (placebo, aspirin) to be (1.4, 2.3). If we had formed the table with aspirin in the first row (instead of placebo), then the 95% confidence interval would have been (1/2.3, 1/1.4) = (.4, .7).

   d. _____ Pearson's chi-squared test of independence treats both the rows and the columns of the contingency table as nominal scale; thus, if either or both variables are ordinal, the test ignores that information.

   e. _____ For testing independence with random samples, Pearson's $X^2$ 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.

   f. _____ 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.

   g. _____ 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.

   h. _____ In a 5×2 contingency table that compares 5 groups on a binary response variable, the $G^2$ chi-squared statistic with $df = 4$ for testing independence can be exactly partitioned into 4 separate independent chi-squared statistics that each have $df = 1$ by comparing row 1 to row 5, row 2 to row 5, row 3 to row 5, and row 4 to row 5.

   i. _____ The person who first got the formula correct for df for the chi-squared test of independence was Karl Pearson.

j. _____ An ordinary regression model that treats the response $Y$ as having a normal distribution is a special case of a generalized linear model, with normal random component and identity link function. As a result, one can do ordinary regression and ANOVA using software (such as PROC GENMOD in SAS) that is designed to fit generalized linear models.

k. _____ One question in a recent General Social Survey asked subjects how many times they had had sexual intercourse in the previous month. The sample means were 5.9 for the male respondents and 4.3 for the female respondents; the sample variances were 54.8 and 34.4. The modal response for each gender was 0. Since the response variable is a count, the best way to model this count with gender as an indicator explanatory variable would be to use a Poisson generalized linear model.

2. Each of 100 multiple-choice questions on an exam has five possible answers but one correct response. For each question, a student randomly selects one response as the answer.

   a. Specify the probability distribution of the student's number of correct answers on the exam, identifying the parameter(s) for that distribution.

   b. Would it be surprising if the student made at least 50 correct responses? Explain your reasoning.

3. Consider the following data from a women's health study:

| | | Myocardial Infarction | |
|---|---|---|---|
| | | Yes | no |
| Oral Contraceptives | Used | 23 | 34 |
| | Never Used | 35 | 132 |

   a. Construct a 95% confidence interval for the population odds ratio.

   b. Based on (a), does it seem plausible that the variables are independent? Explain.

4. For adults who sailed on the Titanic on its fateful voyage, the odds ratio between gender (female, male) and survival (yes, no) was 11.4.

   a. What is wrong with the interpretation, "The probability of survival for females was 11.4 times that for males."

   b. When would the quoted interpretation be approximately correct? Why?

   c. The odds of survival for females equaled 2.9. For each gender, find the proportion who survived.

5. In a General Social Survey, gender was cross-classified with party identification. Table 1 attached shows some results.

Table 1: Table for Problem 5

```
-------------------------------------------------------------------
             Frequency|
             Expected |dem     |indep   |repub   |
             ---------+--------+--------+--------+
             female   |    279 |     73 |    225 |
                      | 261.42 | 70.653 | 244.93 |
             ---------+--------+--------+--------+
             male     |    165 |     47 |    191 |
                      | 182.58 | 49.347 | 171.07 |
             ---------+--------+--------+--------+

       Statistic                       DF       Value      Prob
       ------------------------------------------------------------
       Chi-Square                       2      7.0095    0.0301
       Likelihood Ratio Chi-Square      2      7.0026    0.0302

 Observ  Resraw  Reschi  StReschi   Observ  Resraw  Reschi  StReschi
   1     17.584   1.088   2.293       4    -17.584  -1.301  -2.293
   2      2.347   0.279   0.465       5     -2.347  -0.334  -0.464
   3    -19.931  -1.274  -2.618       6     19.931   1.524   2.618
-------------------------------------------------------------------
```

a. Explain what the second set of numbers are in the contingency table. Show how to obtain 261.42.

b. Explain how to interpret the Prob value listed for the Chi-Square statistic.

c. Explain carefully how to interpret the values listed under StReschi. In which cases were there significantly more people than one would expect if party identification were independent of gender?

6. Explain what is meant by *overdispersion*, and explain how it can occur for Poisson generalized linear models for count data.

7. Explain two ways in which the generalized linear model extends the ordinary regression model that is commonly used for quantitative response variables.

8. For the 23 space shuttle flights that occurred before the Challenger mission in 1986, Table 2 shows the temperature ($^\circ F$) at the time of the flight and whether at least one of the six primary O-rings suffered thermal distress (1 = yes, 0 = no). The first attached SAS printout shows the use of various models for analyzing these data.

(a) For the logistic regression model using temperature as a predictor for the probability of thermal distress, calculate the estimated probability of thermal distress at 31°, the temperature at the time of the Challenger flight.

(b) At the temperature at which the estimated probability equals 0.5, give a linear approximation for the change in the estimated probability per degree increase in temperature.

(c) Interpret the estimated effect of temperature on the odds of thermal distress.

(d) Test the hypothesis that temperature has no effect, using the likelihood-ratio test. Interpret results.

(e) Suppose we treat the (0, 1) response as if it has a normal distribution, and fit a linear model for the probability. Report the prediction equation, and find the estimated probability of thermal distress at 31°. Comment on the suitability of this model.

(f) Suppose you also wanted to include in the model the month during which the launch occurred (January, February, etc.). Show how you could add indicator variables to the model to allow this. Explain how to interpret the coefficients of the indicator variables.

(g) Refer to the previous part. Explain how you could further generalize the model to allow interaction between temperature and month of the launch, and explain how you could conduct a test to investigate whether you need the interaction terms.

Table 2. Space shuttle data

| Ft | Temp | TD | Ft | Temp | TD | Ft | Temp | TD | Ft | Temp | TD |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 66 | 0 | 2 | 70 | 1 | 3 | 69 | 0 | 4 | 68 | 0 |
| 5 | 67 | 0 | 6 | 72 | 0 | 7 | 73 | 0 | 8 | 70 | 0 |
| 9 | 57 | 1 | 10 | 63 | 1 | 11 | 70 | 1 | 12 | 78 | 0 |
| 13 | 67 | 0 | 14 | 53 | 1 | 15 | 67 | 0 | 16 | 75 | 0 |
| 17 | 70 | 0 | 18 | 81 | 0 | 19 | 76 | 0 | 20 | 79 | 0 |
| 21 | 75 | 1 | 22 | 76 | 0 | 23 | 58 | 1 | | | |

Note: Ft = flight no., Temp = temperature, TD = thermal distress (1 = yes, 0 = no). Data based on Table 1 in *J. Amer. Statist. Assoc*, 84: 945-957, (1989), by S. R. Dalal, E. B. Fowlkes, and B. Hoadley.

```
Results of model fitting for space shuttle data of Problem 8:

-------------------------------------------------------------
 Model 1

                Criteria For Assessing Goodness Of Fit
          Criterion             DF         Value      Value/DF
          Deviance              21       20.3152        0.9674
          Pearson Chi-Square    21       23.1691        1.1033
          Log Likelihood         .      -10.1576             .

      Parameter    DF    Estimate     Std Err   ChiSquare  Pr>Chi
      INTERCEPT     1     15.0429      7.3786      4.1563  0.0415
      TEMP          1     -0.2322      0.1082      4.6008  0.0320

-------------------------------------------------------------
Model 2

                Criteria For Assessing Goodness Of Fit
          Criterion             DF         Value      Value/DF
          Deviance              22       28.2672        1.2849
          Pearson Chi-Square    22       23.0000        1.0455
          Log Likelihood         .      -14.1336             .

      Parameter    DF    Estimate     Std Err   ChiSquare  Pr>Chi
      INTERCEPT     1     -0.8267      0.4532      3.3278  0.0681

-------------------------------------------------------------
Model 3
```

Criteria For Assessing Goodness Of Fit

| Criterion | DF | Value | Value/DF |
|---|---|---|---|
| Deviance | 21 | 3.3386 | 0.1590 |
| Pearson Chi-Square | 21 | 3.3386 | 0.1590 |
| Log Likelihood | . | -10.4411 | . |

| Parameter | DF | Estimate | Std Err | ChiSquare | Pr>Chi |
|---|---|---|---|---|---|
| INTERCEPT | 1 | 2.9048 | 0.8046 | 13.0323 | 0.0003 |
| TEMP | 1 | -0.0374 | 0.0115 | 10.5473 | 0.0012 |

**Formulas**

Binomial $\quad P(y) = \dfrac{N!}{y!(N-y)!}\pi^y(1-\pi)^{N-y}\ ,\ y = 0, 1, 2, ..., N, \quad \mu = N\pi, \sigma = \sqrt{N\pi(1-\pi)}$

$$\text{Hypergeometric} \quad P(n_{11}) = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{+1}-n_{11}}}{\binom{n}{n_{+1}}}, \quad \binom{a}{b} = a!/[b!(a-b)!]$$

$$\text{odds} = \pi/(1-\pi), \quad \pi = \text{odds}/(1+\text{odds}), \quad \text{relative risk} = \pi_1/\pi_2$$

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}, \qquad \hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

$$\text{SE}(\log\hat{\theta}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

$$X^2 = \Sigma\frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}, \quad \hat{\mu}_{ij} = (n_{i+}n_{+j})/n, \quad df = (I-1)(J-1)$$

$$G^2 = 2\Sigma n_{ij}\ \log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right), \quad df = (I-1)(J-1)$$

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

For $H_0 : \beta = 0$, LR statistic $= -2(L_0 - L_1)$, Wald statistic $z = \hat{\beta}/\text{SE}$

Logistic regression $\pi(x) = \alpha + \beta x \qquad \pi = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$

$\log(\frac{\pi}{1-\pi}) = \alpha + \beta x$

$\hat{\pi} = .5$ at $x = -\hat{\alpha}/\hat{\beta}$, incremented rate of change $= \hat{\beta}\hat{\pi}(1-\hat{\pi})$, $e^{\hat{\beta}} =$ odds ratio

$\text{logit}(\pi) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k \qquad \pi = \frac{\exp(\alpha+\beta_1 x_1+\cdots+\beta_k x_k)}{1+\exp(\alpha+\beta_1 x_1+\cdots+\beta_k x_k)}$