QUALIFYING EXAM SOLUTIONS
Statistical Methods
Fall, 2008

1. Write the whole paragraph by yourself. The main points are below:

   - The odds ratio is
   $$\hat{\theta} = \frac{17 \times 16}{5 \times 18} = 3.02$$
   and
   $$s_{\log \hat{\theta}} = \sqrt{1/17 + 1/5 + 1/18 + 1/16} = 0.6139.$$
   Then, the 95% confidence interval is
   $$3.02e^{\pm 1.96 \times 0.6139} = [1.1309, 8.0647].$$
   Thus, it is significant. Therefore, the reduction is significant.
   - The decreases in Delaware difference is significant, but in Monroe is not.
   - The difference between the two counties are not signigicant before, but it is significant after.

2. The key points of the problems are below:

   - The odds ratio and the confidence interval between "Too small" and "Other". The 95% confidence interval is
   $$1.4213e^{\pm 1.960.4879} = [0.5462, 3.6982]$$
   which is not significant.
   - The odds ratio and the confidence interval between "Large enoug" and "Other". The 95% confidence interval is
   $$2.5625e^{\pm 1.96 \times 0.3334} = [1.3331, 4.9256],$$
   which is significant.
   - The odds ratio from first to the second, and from the second to the thid. From "too small" to "intermediate": the 95% confidence interval is
   $$0.5966e^{\pm 1.96 \times 0.5966} = [0.1852, 1.9208],$$
   which is not significant; the 95% confidence interval from "intermediate" to "large enough" is
   $$3.20e^{\pm 1.96 \times 0.4097} = [1.3629, 6.7414],$$
   which is significant.

   Try to interpret the odds ratios by yourself. The key is: the increase from "too small" to "intermediate" is not significant, but from "intermediate" to "large" is significant.

3. (a) The survival function is

$$\hat{S}(t) = \begin{cases} 0.9167 & \text{when } t = 49 \\ 0.825 & \text{when } t = 56 \\ 0.6875 & \text{when } s = 69 \\ 0.55 & \text{when } t = 70 \\ 0.275 & \text{when } t = 74 \\ 0.1375 & \text{when } t = 75 \\ 0 & \text{when } t = 81 \end{cases}$$

(b) Let $S_M(t)$ be the survival function for male, and $S_F(t)$ be for female. We test

$$H_0 : S_M(t) = S_F(t) \leftrightarrow H_1 : S_M(t) \neq S_F(t).$$

The value of the log-rank test statistic is 0.299. Comparing to $\chi^2_{0.05,1} = 3.84$, we accept H_0 and conclude $S_M(t) = S_F(t)$.

(c) The survivial function is $S(t) = e^{-\lambda t}$, where

$$\hat{\lambda} = e^{-4.772} = 0.008463.$$

Then, $\hat{S}(75) = e^{-75\hat{\lambda}} = 0.5301$.

(d) The survival function has the relationship

$$S_M(t) = S_M(t)^{1.30}.$$

4. Let c_i be the number of parasites in the i-th jar for $i = 1, \cdots, 120$. Let X_{i1} be the type of parasites, and X_{i2} be the type of bettle. Then, the data are given by (c_i, X_{i1}, X_{i2}). We can fit a regression model as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, i = 1, 2, j = 1, 2, 3, k = 1, \cdots, 20.$$

We can analyze the main effects α_i for parasites, and β_j for beetles, and their interaction effects $(\alpha\beta)_{ij}$. We will use the ANOVA method to analyze the data. The response $Y_i = c_i$ may be either the original count, or count using a logistic transformation as

$$\log \frac{c_i}{30 - c_i}.$$

In addition, we can also use a logistic regression to analyze the data. The model is

$$\log \frac{\pi_i}{1 - \pi_i} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}.$$

The aim of the first method is to model the count of parasites in a bettle. The second and the third methods is to model the probability for a prarsit to be contained in a bettle.

5. In this problem, one cannot measure exactly whether the disease can be present. Therefore, it can be only known with 85% or 95%. Then, we can assign a prior distribution with 95% (or 85%) if it is a case or not. The likelihood function then can be derive. A logistic regression method then can be derived.

6. (a) This method can reduce the multicollinearity.

   (b) The model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \beta_6 X_2 X_3 + \beta_7 X_1 X_2 X_3 + \epsilon.$$

   The test of the F-statistic is

$$H_0 : \beta_1 = \cdots = \beta_7 = 0 \leftrightarrow H_1 : \text{not all of } \beta_i \text{ is zero.}$$

   Based on the p-value, we reject the null hypothesis.

   (c) For male, the model is

$$\hat{Y} = -4.88282 - 3.93895 X_1 + 4.62358 X_3 + 0.058 X_1 X_3$$

   and for female the model is

$$\hat{Y} = -4.9363 - 3.88095 X_1 + 3.93307 X_3 - 8.6826 X_1 X_3.$$

   Then, try to interpret it by yourself.

7. (a) Let X_1 be HEALTH and X_2 be GRP. The model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 I_{X_2=2} + \beta_3 X_1 I_{X_2=2} + \epsilon, \epsilon \sim N(0, \sigma^2).$$

   (b) We may consider the log-transformation because it transforms the value of the response from $(0, \infty)$ to $(-\infty, \infty)$ which can better match the linear function of independent variables.

   (c) For group 1, we have mean after the transformaion is $2.2443+(0.007499+0.00324)75.62-0.6777 = 2.378$ and before the transformaion is 10.79. For group 2, they are $2.2443 + (0.007499)74.59 = 2.804$ and 16.51 respectively.

   (d) We do not have significant concerns from the plot because residual plot does not present any heteoregeity of variance, and the histogram plot shows the residual is almost normal.

   (e) Poisson method is to model the count directly but regression treats count as a continuous vbariable.

   (f) SAS results show the interaction is not significant (from SS3 values). In group 1, the predicted value is $e^{1.7255+(0.0050+0.0035)\times 75.62-0.5888} = 5.9267$, for group 2, the predicted value is $e^{1.7255+0.0050\times 74.59} = 8.1535$. They are lower than the predicted value in (c).

   (g) We recommend the ANCOVA model instead of the Poisson model because HEALTH is a continuous variable.