STAT 557 **Assignment 3** **Name** _____
FALL 2017

Reading Assignment: Agresti, Chapters 4 and 5, Section 7.1

Written Assignment: Due Thursday, October 26, in class.

1. For a particular plant, the leaf shape (cut or potato) and plant size (tall or dwarf) are controlled by independent genes *a* and *b*. Mendelian genetics tells us that if two characteristics are distributed independently in a dihybrid cross then four phenotypes ab, Ab, aB, AB, are expected to occur in proportions 9/16, 3/16, 3/16, 1/16, respectively. The data in the following table are counts of the four possible phenotypes for samples of plants collected in three different regions. (Data from Lloyd, *Statistical Analysis of Categorical Data*).

	Tall/Cut	Tall/Potato	Dwarf/Cut	Dwarf/Potato
Region 1	926	288	293	104
Region 2	467	151	150	47
Region 3	693	234	219	70

Consider each row of the table to have an independent multinomial distribution.

a) Fit the model dictated by the Mendelian theory to the data in this table. Calculate the value of the deviance statistic $G^2 = 2\sum_{i=1}^{3}\sum_{j=1}^{4} Y_{ij}\log(\hat{m}_{A,ij}/\hat{m}_{0,ij})$, where $\hat{m}_{A,ij}$ is the expected count under the alternative hypothesis and $\hat{m}_{o,ij}$ is the expected count under the null hypothesis. Here the alternative model is that each row of the table has a different multinomial distribution, the general alternative. Report the value of G^2, degrees of freedom, and a p-value. State your conclusion.

b) If the populations in these regions are not in equilibrium then different proportions phenotypes may be present, i.e., $p_a p_b$, $(1-p_a)p_b$, $p_a(1-p_b)$, $(1-p_a)(1-p_b)$. Estimate p_a and p_b separately for each region. Report the value of the deviance statistic for this model (with respect to the general alternative), its degrees of freedom, and a p-value. State your conclusion.

c) Fit the model in part (b) with the constraints that p_a has the same value for all three regions and p_b has the same value for all three regions. Report the value of the deviance statistic for this model (with respect to the general alternative), its degrees of freedom, and a p-value. State your conclusion.

d) The models in Parts (a), (b), and (c) are a set of nested models. Report an analysis of deviance table for this set of models.

e) Do any of these models provide an adequate description of the data? If so, identify the simplest model that provides an adequate description. Justify your answer.

2. Let $Y_1, Y_2, Y_3, \ldots, Y_n$ denote a set of n independent observations from a Poisson distribution with mean μ. Write out the formula for the log-likelihood function. Show that the Fisher scoring algorithm converges after one step, regardless of the initial value, $\hat{\mu}^{(0)} > 0$, used to start the process. What happens with the Newton-Raphson algorithm?

3. (a) For a fixed value of the dispersion parameter, $k > 0$, show that the negative binomial distribution is a member of the exponential family of distributions. Use the parameterization of the negative binomial distribution presented by Agresti on page 127, i.e.,

$$f(y;k,\mu) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}\left(\frac{k}{\mu+k}\right)^k\left(1-\frac{k}{\mu+k}\right)^y, \quad y=0, 1, 2, \ldots..$$

For this parameterization, $E(Y) = \mu$ and $Var(Y) = \mu + \frac{\mu^2}{k}$.

(b) Is the negative binomial distribution a member of the exponential family when k is treated as an unknown parameter? Justify your answer.

4. The sex of a certain species of turtle is thought to be largely determined by the incubation temperature of the eggs. To examine this relationship turtle eggs were collected in Illinois and incubated at various temperatures. A number of eggs were placed in a box that was incubated at a certain temperature. Different boxes were incubated at different temperatures and different boxes contained different number of eggs. The numbers of male and female turtles that hatched from the eggs were recorded for each box. The data are posted in the file **teggll.dat**. The columns of this file correspond to

Box	box identification number
Temperature	incubation temperature (°C) for the box
Females	number of females emerging from the eggs in the box
Males	number of males emerging from the eggs in the box
Total	number of eggs in the box

(a) Fit a logistic regression model of the form

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \text{ (temperature)}$$

where π_i represents the conditional probability that a female turtle hatches from an egg incubated at the specified temperature. Report mle's for the parameters β_0 and β_1, and Report the standard errors of the estimates.

(b) Test the null hypothesis $H_0 : \beta_1 = 0$ against the alternative $H_a : \beta_1 \neq 0$. Report a p-value and state your conclusion.

(c) Estimate the ratio of the odds that a female hatches from an egg incubated at x+1 degrees relative to the odds that a female hatches from an egg incubated at x degrees. Construct a 95% confidence interval for the odds ratio. State your conclusion.

(d) Estimate the probability that a female turtle hatches from an egg incubated at 28 degrees. Report the standard error of the estimated probability

(e) Use the logistic regression model to estimate the temperature at which 50% of the eggs produce female turtles and 50% of the eggs produce male turtles. Report the value of the temperature estimate and a 95% confidence interval.

(f) A second sample of eggs was obtained from the same turtle species in New Mexico. These data are posted in the file **teggnm.dat**. Fit the logistic regression model described in part (a) to these data. Report values for the parameter estimates and their standard errors.

(g) Use the model in part (f) to estimate the incubation temperature at which 50% of the eggs produce females. Report a 95% confidence interval.

(h) Test the null hypothesis that the incubation temperatures at which 50% of the eggs produce females are the same for the Illinois and New Mexico populations of turtles. Report your conclusion.

(i) Fit complimentary log-log models of the form

$$\log(-\log(1-\pi_i)) = \alpha_0 + \alpha_1 \text{ (temperature)}$$

to the turtle egg data from Illinois and the turtle egg data from New Mexico. Report estimates of the parameters and their standard errors for each model. Does this model provide a better description of the data than the logistic regression models? If so, re-estimate the incubation temperature at which 50% of the eggs produce females for each population of turtles and compare the estimates. Report your conclusions.

(j) Can you find a model that fits the data for Illinois and New Mexico better than either the logit model or complimentary log-log model considered in previous parts of this question. If so, provide some evidence that your model is better and estimate the temperatures in Illinois and New Mexico at which 50% of the eggs produce females. Report confidence intervals for your estimates.

5. Some studies produce count data with a higher frequency of zero counts than can be can be accommodated by Poisson or negative binomial models. This may occur because a sample is taken from a population that is a mixture of two populations; members of one population have positive probabilities of providing positive counts while members of the other population always provide a

zero count. The members of the first population, for example, may be susceptible to a particular disease and members of the population may experience one or more infections during the study follow-up period, while members of the second population are immune to the disease and do not experience any infections during the follow-up period. The numbers of infection events for member of the first population may follow a Poisson distribution, but selecting members of the second population will inflate the number of subjects reporting zeros counts. Distributions that place greater probability on zero are often called zero-inflated distribution.

In this exercise we will illustrate zero-inflated Poisson and zero-inflated negative binomial distributions with applications to data on milk consumption (glasses of milk) reported by a random sample of 1900 subjects in the U.S. The data are summarized in the following table.

Milk Consumption (in glasses)	Observed Counts
0	767
1	557
2	333
3	142
4	62
5	23
6	16
Total	**1900**

Some of the subjects who reported drinking no milk may never drink milk, while others who reporting drinking no milk may sometimes drink milk but did not drink any milk on the day of the study. One type of zero-inflated Poisson (ZIP) model that is often considered in such situations has a probability function of the form

$$\Pr(X = x) = \omega I_{(x=0)} + (1-\omega)\frac{e^{-\mu}\mu^{x}}{\Gamma(x+1)} \quad \text{for } x=0, 1, 2, \ldots.$$

where ω is the probability that the respondent was selected from the population of people who never drink milk and $1-\omega$ is the probability of selecting a person from the population of people who never drink milk. In this notation, $I_{(x=0)}$ is an indicator function that is one when the response is zero, and it is zero when the response is a positive count. We will refer to this as the ZIP (zero-inflated Poisson) model.

You may find the SAS code posted on blackboard as **milkzip.sas** helpful for answering the following questions. This program uses the genmod procedure in SAS to fit find maximum likelihood estimates for the ZIP model. Prior to that, this program has code for a SAS macro, called gof_test that computes the Pearson goodness-of-fit test statistic. The first argument in the

input to this macro is a file that has just two columns. The first column, called y, contains the observed frequencies for each of the categories and the second column contains the maximum likelihood estimates for the category probabilities for the model under consideration. It is assumed that the categories are ordered as 0 glasses, 1 glass, 2 glasses, …. No combining of categories is done to make sure that most expected counts are at least 5. If this condition is not satisfied, you would need to combine categories and change the values in the two columns of the input data file by hand, or you need to include code from a program used earlier in the course to check this condition and combine categories as needed. It would be easy to modify this code to add the deviance statistic, or to convert this code into a function in R.

(a) Interpret the parameters ω and μ in the context of the milk consumption study.

(b) Derive formulas for the mean and variance of the ZIP distribution.

(c) Find the maximum likelihood estimates for the parameters in the ZIP model for the data from the milk consumption study. Also, report the large sample estimates of their standard errors.

(d) Construct an approximate 95% confidence interval for the mean number of glasses of milk consumed per day.

(e) What is the probability that a person randomly selected from the mixture of these two populations gives a zero response? Construct an approximate 95% confidence interval for this probability.

(f) Construct an approximate 95% confidence interval for the proportion of the population that never drinks milk.

(g) Compute the value of a Pearson goodness-of-fit statistic and the corresponding p-value for fitting a Poisson distribution to the milk consumption data. Compute the value of the Pearson goodness-of-fit statistics and the corresponding p-value for fitting the ZIP model. State your conclusions.

6. Another family of models for dealing with excessive zeros is the Hurdle family of models. In this family the entire probability of observing a zero count is modelled with a parameter, say π, and the conditional distribution of the positive counts is modelled with a truncated-at-zero distribution. The probability distribution for the truncated-at-zero Poisson distribution, for example, is

$$\Pr(Y = x) = \frac{e^{-\mu}(\mu)^x}{(1-e^{-\mu})\Gamma(x+1)} \quad \text{for } x=1, 2, 3, 4,....$$

and the probability function for the Poisson Hurdle model is

$$\Pr(X = x) = \begin{cases} \pi & \text{if } x=0 \\ (1-\pi)\dfrac{e^{-\mu}(\mu)^x}{(1-e^{-\mu})\Gamma(x+1)} & \text{if } x=1, 2, 3, 4,.... \end{cases}$$

Note that the Poisson Hurdle distribution allows for the possibility of a deficiency of zero counts, which cannot be accommodated by the ZIP model.

The code posted on Blackboard as **milkhurdle.sas** uses NLMIXED, the SAS procedure for fitting non-linear mixed models, to find maximum likelihood estimates for the parameters in Poisson Hurdle model for the milk consumption data. It also includes the macro for computing the value of the Pearson goodness-of-fit test.

(a) Interpret the parameters π and μ in the context of the milk consumption study.

(b) Derive formulas for the mean and variance of the Poisson Hurdle distribution.

(c) Find the maximum likelihood estimates for the parameters in the Poisson Hurdle model for the data from the milk consumption study. Also, report the large sample estimates of their standard errors.

(d) What is the probability that a person randomly selected from the mixture of these two populations gives a zero response? Construct an approximate 95% confidence interval for this probability.

(e) Construct an approximate 95% confidence interval for the proportion of the population that never drinks milk.

(f) Compute the value of a Pearson goodness-of-fit statistic and the corresponding p-value for fitting a Poisson Hurdle distribution to the milk consumption data. State your conclusions.

7. Repeat problem 5 for the zero inflated negative binomial (ZINB) model. Use the following parameterization of the ZINB probability function.

$$\Pr(X = x) = \omega I_{(x=0)} + (1-\omega)\frac{\Gamma\left(x+\frac{1}{\kappa}\right)}{\Gamma\left(\frac{1}{\kappa}\right)\Gamma(x+1)}\left(\frac{1}{1+\kappa\mu}\right)^{1/\kappa}\left(\frac{\kappa\mu}{1+\kappa\mu}\right)^x \quad \text{for } x=0, 1, 2, ...$$

The mean and variance of this distribution are

$$E(X) = (1-\omega)\mu \quad \text{and} \quad Var(X)=(1-\omega)[\mu(1+\mu\kappa)+\omega\mu^2].$$

Note that the variance is larger than the mean and it is also larger than the variance for the ZIP distribution in problem 5. This model can account for more variability while retaining the same

zero probability provided by the ZIP distribution. (With this information, you should be able to do part (b) correctly.)

8. Repeat problem 6 for the negative binomial Hurdle model. Use the following parameterization of the ZINB probability function

$$\Pr(X = x) = \begin{cases} \pi & \text{if } x=0 \\ (1-\pi)\dfrac{\Gamma\left(x+\frac{1}{\kappa}\right)}{\left(1-\left(\frac{1}{1+\kappa\mu}\right)^{1/\kappa}\right)\Gamma\left(\frac{1}{\kappa}\right)\Gamma(x+1)}\left(\frac{1}{1+\kappa\mu}\right)^{1/\kappa}\left(\frac{\kappa\mu}{1+\kappa\mu}\right)^{x} & \text{if } x=1, 2, 3, 4,.... \end{cases}$$

The mean and variance are

$$E(X) = \left(\frac{1-\pi}{1-\theta}\right)\mu \quad \text{and} \quad Var(X)=\left(\frac{1-\pi}{1-\theta}\right)\mu(1 + \mu + \mu\kappa) - \left(\frac{1-\pi}{1-\theta}\right)^2 \mu^2,$$

where

$$\theta = \left(\frac{1}{1+\kappa\mu}\right)^{1/\kappa}.$$