Due: Friday, 29 Sept, 5pm, to my office or mailbox (in 1121 Snedecor)

Reminders:

1. Please answer 3 out of the 4 questions.
2. Please see the Homework guidelines page of the class web site for homework policies.
3. In particular, problem 4 is a “major data analysis” problem. Please look at the guidelines for how to write up your answer.
4. Remember, you are encouraged to work together. However, you must write up your own answers.
5. This is not a programming class. Ask for help if R or MARK is being troublesome.

Problems:

1. These data come from a study of a deer mice population in sagebush steppe in Utah. Individuals were trapped on 5 nights. Summary statistics are:

<table>
<thead>
<tr>
<th>Trap night</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_i$</td>
<td>14</td>
<td>9</td>
<td>12</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>$u_i$</td>
<td>14</td>
<td>5</td>
<td>11</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>$m_i$</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>$M_i$</td>
<td>0</td>
<td>14</td>
<td>19</td>
<td>30</td>
<td>37</td>
</tr>
</tbody>
</table>

   $M_{t+1} = 42$

   Log likelihoods for models $M_0$ and $M_t$ are given in the Chao and Huggins reading. The log-likelihood used in the exercise is the log likelihood for model $M_b$, with $c$ in the usual notation called $p_2$ in the exercise. R functions to calculate all three log likelihoods are in lnlmodels.r, although note that I omit the optional constants from the log likelihood calculation. The sufficient statistics used in my functions (and in the Chao and Huggins equations) are:

   $t$ Number of trapping occasions
   $n$ number caught on each occasion (vector of length $t$)
   $n_i$ ndot: total number of captures, $n = \sum_{i=1}^{t} n_i$
   $M$ number of tags in population just prior to occasion $i$, $M_i = \sum_{j=1}^{i-1} u_j$ except $M_1 = 0$.
   $M_t$ Mdot = $\sum_{i=1}^{t} M_i$, note, does not include $M_{t+1}$.
   $M_{t+1}$ Mt1: Total number of unique individuals seen
   $u$ number of unmarked individuals caught on each occasion (vector of length $t$)
   $m$ number of marked individuals caught on each occasion (vector of length $t$)
   $m_i$ mdot: total captures of marked individuals, $m_i = \sum_{i=1}^{t} m_i$
   note: $n_i = u_i + m_i$ for all $i$. 

1
(a) Estimate the population size, \( N \), and its standard error under model \( M_0 \).

(b) Estimate the population size, \( N \), and its standard error under model \( M_t \).

(c) Estimate the population size, \( N \), and its standard error under model \( M_b \).

(d) Which model, \( M_0 \), \( M_t \), or \( M_b \) is the best for these data? Briefly explain your choice.

(e) If X was your answer to question 1d, is it appropriate to say something like “model X is a good fit to the data.”? Briefly explain why or why not.

(f) The log likelihoods given in Chao and Huggins use \( N \) as the parameter. You will sometimes see \( f_0 \), defined as the number of “never seen” individuals, used instead of \( N \). Note that \( N = M_{t+1} + f_0 \). What is the log-likelihood of model \( M_0 \) when \( f_0 \) is the parameter instead of \( N \), i.e. what is \( \ln L(f_0, p|n, M_{t+1}) \)? If you fit both the model using \( N \) and the model using \( f_0 \), would you get the same log likelihood value? Briefly explain why or why not.

2. This problem continues the analysis of deer mice population size from problem 1. The parts of this question are based on the fits of models \( M_0 \), \( M_b \), and \( M_t \) from problem 1. If you could not do problem 1 and want to do this problem, let me know and I will provide needed results from problem 1.

(a) Using model \( M_b \), calculate a Wald 95% confidence interval for \( N \).

(b) Using model \( M_b \), use log transformed \( N \) to calculate a Wald 95% confidence interval for \( N \).

(c) Using model \( M_b \), calculate a 95% profile confidence interval for \( N \).

Note: For model \( M_0 \), the conditional mle of \( p \) given \( N \) is \( \frac{n}{N} \). For model \( M_t \), the conditional mle of \( p_i \) given \( N \) is \( \frac{n_i}{N} \). For model \( M_b \), the conditional mle’s of \( p \) and \( c \) given \( N \) are \( \hat{p} = \frac{M_{t+1}}{1\hat{N} - M} \) and \( \hat{c} = \frac{m}{M} \).

(d) Look at the plot of profile likelihood vs. \( N \). Which is more appropriate: a quadratic approximation in terms of \( N \) or a quadratic approximation in terms of log \( N \)? Briefly explain your choice and include your plot (or plots) of the profile likelihood.

(e) Wald intervals are much more commonly used than profile intervals (at least now). Based on your results from questions 2c and 2d, would you recommend using Wald intervals based on \( N \) or those based on log \( N \)? Briefly explain your answer.

(f) Using AIC weights, compute the model averaged estimate of \( N \) and its standard error, using both the Buckland se formula and the Revised formula.

(g) Imagine you had three different studies of the same population. These are conducted by different people at slightly different times, but we will assume that the population remains closed across the three studies (so the true \( N \) is the same for all three studies). The study results are:

<table>
<thead>
<tr>
<th>Study</th>
<th>( \hat{N} )</th>
<th>( \text{Var} \hat{N} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>145</td>
<td>30</td>
</tr>
<tr>
<td>B</td>
<td>147</td>
<td>30</td>
</tr>
<tr>
<td>C</td>
<td>143</td>
<td>30</td>
</tr>
</tbody>
</table>
The three studies can be assumed to be independent estimates of \( N \). If you average the results from the three studies, what is the pooled estimate, \( \hat{N} \), and its standard error.

Note: because the within-study variances are the same, the equally-weighted and inverse-variance-weighted estimates are identical.

(h) Now imagine the same results were obtained from fitting three models to results from Study A. Those results, now including model weights, are:

<table>
<thead>
<tr>
<th>Model</th>
<th>( \hat{N} )</th>
<th>( \text{Var}\hat{N} )</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>145</td>
<td>30</td>
<td>1/3</td>
</tr>
<tr>
<td>B</td>
<td>147</td>
<td>30</td>
<td>1/3</td>
</tr>
<tr>
<td>C</td>
<td>143</td>
<td>30</td>
<td>1/3</td>
</tr>
</tbody>
</table>

Compute the model averaged estimate of \( N \) and its standard error (Buckland formula is sufficient).

(i) In statistical terms, why are the results from question 2h not identical to those from question 2g?

Note: At least some values should be different.

3. This problem is a simplified version of something in regular use. Wind turbines kill birds and bats. The Fish and Wildlife Service cares (a lot) when those are endangered or threatened bats (because of the Endangered Species Act) or eagles (because of the Bald and Golden Eagle Protection Act). Hence, owners of wind turbine farms are required to estimate the number of killed bats and eagles. Bats are the bigger problem. Eagle kill is mostly in the winter and an eagle carcass on snow is pretty easy to spot. Bat kill is mostly late summer and early fall. Dead bats are hard to spot; they are small, dark brown critters that fall on bare ground, grass/prairie, or in the middle of a crop field. The data to estimate bat kill are collected by trained technicians who search the ground around each turbine and identify each carcass they find. They search every week and sometimes more frequently. The number killed is larger than the number seen for three different reasons:

- only a portion of the area is searched
- scavengers, e.g. coyotes or raccoons, remove a carcass before the technician looks for it
- the searcher doesn’t see a carcass that is in the search area and not scavenged.

The portion of area searched, \( \pi_a \), is a known constant. The searcher efficiency and scavenger removal rate are estimated by placing a known number of marked dead bats on the ground and seeing how many are found by the technician. In actual use, the searcher efficiency and scavenger removal rate are estimated by two different studies. I am simplifying this problem by combining the two rates into a single probability: \( \pi_s \) is the probability that a carcass falling in the searched area is found by a searcher. \( \pi_S \) can be estimated by placing a known number of marked dead bats in the search area and recording how many are found by a searcher.

Note: there is a temporal aspect to carcass removal (how many days has it been on the ground) that I am ignoring here to keep the problem from getting to complicated.

The data for this problem are made up based on typical numbers for a non-endangered bat species at an Iowa wind farm with a very, very intensive monitoring program. The
data are:

\( Y = 74 \): the number of unmarked bats (so killed by the turbines).

In the carcass detection trial, 100 marked bats were placed in the search area. The technicians found:

\( Z = 27 \) marked bats (so some of the 100 that were deliberately placed in the search area). The search area proportion is known to be \( \pi_a = 0.2 \).

A reasonable model for the searcher trial is \( Z \sim \text{Bin}(100, \pi_s) \). A reasonable model for the turbine mortality is \( Y \sim \text{Bin}(N, \pi_a \pi_s) \). At the end of the problem, we will consider a different model for turbine mortality: \( Y \sim \text{Pois}(N \pi_a \pi_s) \)

(a) What are the expected values of \( Y \) and \( Z \)?
(b) What is the method-of-moments estimate of \( N \)?
(c) What is the log likelihood function for the parameters \((N, \pi_s)\) given the data, \((Y, Z)\) and known constant, \( \pi_a \).

Use the Binomial model for \( Y \).
(d) Find the mle’s of \( N \) and \( \pi_s \) and their standard errors.
(e) Find the mle’s of \( N \) and \( \pi_s \) and their standard errors under the Poisson model for \( Y \).

Note: The searcher trial model remains a Binomial because the number of placed carcasses is not random.

(f) Bonus point: Use what you know (or can look up) about the relationship between a Binomial and a Poisson distribution to explain why the two models give similar mle’s, slightly different se’s, and why the se from the Poisson distribution is larger than that from the Binomial model.

Note: The detection probabilities here are very large and can be obtained only by a very expensive monitoring program. This problem gets a lot more interesting and a lot more difficult when the detection probabilities are much smaller, e.g. \( \pi_a = 0.1 \) and \( \pi_s = 0.05 \) and no dead bats are found (\( Y = 0 \)). That situation is an active research area.

4. This is a “major data question”. Please organize your answer as described in the homework guidelines. You do need to describe how you arrived at your answer. Provide supporting information in your answer. Do not give me estimates of \( N \) from a large number of models.

The data in newt.txt are from a study to estimate the number of adult great crested newts resident in an area soon to be developed. Adults can be individually identified by the pattern of spots on their belly. Individuals were caught on 12 nights, identified and sexed. The population can be assumed closed over the 12 capture occasions. The data in newtB.txt are the capture histories and sex (1 or 2) for each individual.

Your goal is to provide an estimate of the number of newts in the population. At a minimum, you need to provide an estimate and standard error. A confidence interval would be nice.

Notes: I suggest you avoid model Mtb in its most general forms (different \( p \) for each time; a single \( c \) or different \( c \) for each time) because the population size is not identifiable in that model. A model with a linear effect of time may be reasonable.