Homework #4

(due Monday, March 18)

#1. You want to estimate the relative frequency of the rare white color morph in a populations of *Cirsium palustre* in France; most of the flowers are purple in this species. So you want to estimate some parameter p, which is the probability that a randomly selected individual will be a white flower. After obtaining a generous travel award, you collect individuals in two sampling events. The data for a total of 461 plants are:

Event	sample size	# plants with white flowers	# plants with purple flowers
Sampling event 1	$n_1 = 199$	$x_1 = 4$	195
Sampling event 2	$n_2 = 262$	$x_2 = 8$	254

(A) Write a likelihood equation p for a sample generically, then given the equation with the numbers for the first sample.

$$L(p \mid x) = \mathbb{P}(x \mid n, p) = \binom{n}{x} p^x (1-p)^{n-x} \quad (1)$$

$$L(p \mid x) = \mathbb{P}(x = 4 \mid n = 199, p) = \binom{199}{4} p^4 (1-p)^{195} \quad (2)$$

(B) Consult the https://en.wikipedia.org/wiki/Conjugate_prior#Table_of_conjugate_distributions and then choose an appropriate prior distribution. You have to pretend you didn't see the data, but just thought that the white morph was rare. There is not one specific answer I'm looking for here – it is *your* prior.

The Beta distribution is conjugate to the binomial likelihood. If we think that the frequency of the white morph is rare, and choose (somewhat arbitrarily) that we'd like the mean of our prior on p to be around 0.1, then we could use a $p \sim \text{Beta}(\alpha = 1, \beta = 9)$ as our prior. I'll use this prior for the rest of the answer...

(C) If you analyzed your data after the first sample, what would your posterior distribution on p have been?

using the notation on the wikipedia page, for the first trial: $n = 1$ (because n is counting the number of trials), and $N_1 = 199$ (the sample size), so the posterior is:

$$\text{Beta}(\alpha = 1 + 4, \beta = 9 + 195) = \text{Beta}(\alpha = 5, \beta = 204)$$

(D) What is you 95% credible interval for p after seeing the first sample?

`qbeta(0.025, 5, 204)` and `qbeta(0.975, 5, 204)` in `R` reveal that the 95%credible interval is $(0.00785, 0.0485)$

(E) What is the *maximum a posteriori* (MAP) estimate of p after seeing the first sample?

Using the formuala for the mode of a Beta:

$$MAP(p) = \hat{p} = \frac{\alpha - 1}{\alpha + \beta - 2} = \frac{4}{207} = 0.0193$$

(F) Imagine that you take your prior from the posterior sampling event and use it as your prior for the second sampling event. Give the posterior, MAP, and 95% C.I. after considering the second sample after the first sample.

We end up with a

$$\text{Beta}(\alpha = 5 + 8, \beta = 204 + 254) = \text{Beta}(\alpha = 13, \beta = 458)$$

the 95%credible interval is $(0.0148, 0.0442)$. The MAP is $MAP(p) = \hat{p} = 0.0256$

(G) Now consider what would you would have gotten (posterior, MAP, and 95% C.I.) if you had started with your naive prior and analyzed all of the data at one step. Do you get a different posterior distribution?

No, you end up with the same posterior working with all the data or in a two-step analysis.

(Question #2 is on second page)

(2) You are either a botanist or someone who enjoys pretending to be a botanist. You want to determine if there is evidence of crowding effects in reproductive success in your favorite annual angiosperm. You set up an experiment in which you select a set of experimental plots which do not have any seeds of the plant. Then you transplant 1, 2, or 3 plants into each plot, and you allow them to grow for the year. Each plot has a barrier to prevent seeds from entering from the outside. The next year, you painstakingly count the number of small plants of your favorite species that emerge.

Your data:

Plot	# plants grown last year in the plot	# of tiny plants this year
1	1	19
2	2	25
3	3	39
4	1	17
5	2	28
6	3	35

For the sake of this question, let's just assume that the number of plants should follow a Poisson model with the expectation set by either a null model or an alternative.

You would like to test the null hypothesis (H_0) of no evidence of a competitive interaction. In other words, if the null is correct the expected number of plants in this year is simply a linear function of the number of plants in the plot last year. Furthermore, it is a linear function with a y-intercept of 0 (because we are assuming there was no seed bank from prior years, so you'd get no plant this year if you hadn't put any into the ground the previous year). So the null model has some slope that is unknown but represents the number of plants you expect to get in year $t+1$ from each plant that you put in the ground in year t.

Test that null against 2 alternatives.

The first alternative hypothesis, H_1, states that: (a) there is some expected reproductive success for the first plant in a plot, but (b) if a plot has > 1 progenitor planted, then the expected number of plants in the next year is a linear function of the number of "extra" parental plants (the number of plants over the first plant).

The third alternative (H_2) is simply that you have different expected numbers of progeny plants based on how many plants you put in the ground in the first year. So with H_2 have 3 expected numbers with no constrained relationship between them. (This may be the easiest hypothesis to start working on).

Perform a maximum likelihood hypothesis test comparing these 3 models. Report the parameter estimates, and log-likelihoods, and the test statistics used to decide which model is preferred.

There are 6 trials, but only 3 starting conditions. Given that we'd have the same expected number of seedlings given the same input, we have three distinct expected numbers of offspring:

We can use j to index the trial within a given starting number, and we use i to index the # of plants put into the ground. We'll ignore the denominator (factorials) in the Poisson likelihood because they depend on the data and not the parameters (so they are the same for all models).

We'll use x_{i*} for $\sum_{j=1}^{2} x_{i,j}$. So this is the total number of seedlindgs for each number put in. In our data: $x_{1*} = 36$ and $x_{2*} = 53$ and $x_{3*} = 74$

The most general model is H_2:

$$\ln L(\lambda_1, \lambda_2, \lambda_3 \mid X) = \sum_{i=1}^{3}\sum_{j=1}^{2}(x_{i,j}\ln\lambda_i - \lambda_i) \quad (3)$$

$$= \sum_{i=1}^{3} x_{i*}\ln(_i) - 2\lambda_i \quad (4)$$

Working through the derivatives, we find that the 3 sampling set ups don't affect each other. So the MLEs are just the means for each condition:

$$\widehat{\lambda_i} = \frac{x_{i*}}{2} \quad (5)$$

$$\widehat{\lambda_1} = \frac{36}{2} = 18 \quad (6)$$

$$\widehat{\lambda_2} = \frac{53}{2} = 26.5 \quad (7)$$

$$\widehat{\lambda_3} = \frac{74}{2} = 37 \quad (8)$$

$$\ln L(\lambda_1 = 18, \lambda_2 = 26.5, \lambda_3 = 37 \mid X) = 381.95 \quad (9)$$

In model H_0 we have a constraint:

$$\lambda_1 = m \quad (10)$$

$$\lambda_2 = 2m \quad (11)$$

$$\lambda_3 = 3m \quad (12)$$

where m (the number of seedlings per plant in a density-independent model) is unknown. This let's us rewrite the likelihood in terms of our unknown:

$$\ln L(m \mid X, H_0) = \sum_{i=1}^{3} x_{i*}\ln(im) - 2im \quad (13)$$

$$= x_{1*}\ln(m) - 2m + x_{2*}\ln(2m) - 4m + x_{3*}\ln(3m) - 6m \quad (14)$$

$$= (x_{1*} + x_{2*} + x_{3*})\ln(m) - 12m + x_{2*}\ln(2) + x_{3*}\ln(3) \quad (15)$$

$$\frac{\partial \ln L}{\partial m} = \frac{x_{1*} + x_{2*} + x_{3*}}{m} - 12 \quad (16)$$

$$\hat{m} = \frac{x_{1*} + x_{2*} + x_{3*}}{12} = 13.58 \quad (17)$$

$$\lambda_1 = \hat{m} = 13.58 \quad (18)$$

$$\lambda_2 = 2\hat{m} = 27.16 \quad (19)$$

$$\lambda_3 = 3\hat{m} = 40.75 \quad (20)$$

$$\ln L(m = 13.58 \mid X) = 380.38 \quad (21)$$

H_1 is a middle ground with a different set of constraints: In model H_0 we have a constraint:

$$\lambda_1 = b \quad (22)$$

$$\lambda_2 = b + c \quad (23)$$

$$\lambda_3 = b + 2c \quad (24)$$

so:

$$\ln L(b, c \mid X, H_1) = \sum_{i=1}^{3} x_{i*} \ln(b + (i-1)c) - 2(b + (i-1)c)m \tag{25}$$

$$= x_{1*} \ln(b) + x_{2*} \ln(b + c) + x_{3*} \ln(b + 2c) - 6b - 6c \tag{26}$$

$$\frac{\partial \ln L}{\partial c} = \frac{x_{2*}}{b + c} + \frac{2x_{3*}}{b + 2c} - 6 \tag{27}$$

$$\frac{\partial \ln L}{\partial b} = \frac{x_{1*}}{b} + \frac{x_{2*}}{b + c} + \frac{x_{3*}}{b + 2c} - 6 \tag{28}$$

$$\tag{29}$$

So, finding the joint MLEs, we get 2 equations:

$$\frac{x_{2*}}{\hat{b} + \hat{c}} + \frac{2x_{3*}}{\hat{b} + 2\hat{c}} = 6 \tag{30}$$

$$\frac{x_{1*}}{\hat{b}} + \frac{x_{2*}}{\hat{b} + \hat{c}} + \frac{x_{3*}}{\hat{b} + 2\hat{c}} = 6 \tag{31}$$

Rearranging the second then substituting into the first:

$$\frac{x_{2*}}{\hat{b} + \hat{c}} + \frac{x_{3*}}{\hat{b} + 2\hat{c}} = 6 - \frac{x_{1*}}{\hat{b}} \tag{32}$$

$$\frac{x_{3*}}{\hat{b} + 2\hat{c}} + 6 - \frac{x_{1*}}{\hat{b}} = 6 \tag{33}$$

$$\frac{x_{3*}}{\hat{b} + 2\hat{c}} = \frac{x_{1*}}{\hat{b}} \tag{34}$$

$$\hat{b} x_{3*} = \hat{b} x_{1*} + 2\hat{c} x_{1*} \tag{35}$$

$$\hat{c} = \hat{b} \frac{x_{3*} - x_{1*}}{2x_{1*}} \tag{36}$$

$$\tag{37}$$

Substituting in for $\hat{b}$ into Equation 30 we get:

$$\frac{x_{2*}}{\hat{b} + \hat{b}\frac{x_{3*} - x_{1*}}{2x_{1*}}} + \frac{2x_{3*}}{\hat{b} + 2\hat{b}\frac{x_{3*} - x_{1*}}{2x_{1*}}} = 6 \tag{38}$$

$$\frac{x_{2*}}{1 + \frac{x_{3*} - x_{1*}}{2x_{1*}}} + \frac{2x_{3*}}{1 + \frac{2x_{3*} - 2x_{1*}}{2x_{1*}}} = 6\hat{b} \tag{39}$$

$$\frac{2x_{1*}x_{2*}}{2x_{1*} + x_{3*} - x_{1*}} + \frac{4x_{1*}x_{3*}}{2x_{1*} + 2x_{3*} - 2x_{1*}} = 6\hat{b} \tag{40}$$

$$\frac{2x_{1*}x_{2*}}{x_{1*} + x_{3*}} + 2x_{1*} = 6\hat{b} \tag{41}$$

$$\hat{b} \approx 17.782 \tag{42}$$

$$\hat{c} \approx 9.385 \tag{43}$$

$$\ln L(b = 17.782, c = 9.385 \mid X, H_1) = 381.926 \tag{44}$$

So the H_0 cannot be rejected by the 1.92 lnL difference rule.