EXERCISE 6: SINGLE SEASON OCCUPANCY MIXTURE MODELS

Please cite this work as: Donovan, T. M. and J. Hines. 2007. Exercises in occupancy modeling and estimation.
<http://www.uvm.edu/envnr/vtcfwru/spreadsheets/occupancy.htm>

TABLE OF CONTENTS

SINGLE-SPECIES, SINGLE-SEASON MIXTURE MODELS SPREADSHEET EXERCISE

OBJECTIVES:

- To learn and understand the single-season two-group mixture model, and how it fits into a multinomial maximum likelihood analysis.
- To use Solver to find the maximum likelihood estimates for the probability of group membership, and the probability of detection and site occupancy for each group.
- To assess the $-2Log_eL$ of the saturated model.
- To introduce concepts of model fit.
- To learn how to simulate single-season occupancy mixture data.

BASIC INFORMATION

Now that you have a solid handle on single-season occupancy modeling with both site and survey level covariates, you're ready to learn a few interesting spin-offs of the basic model. In this exercise, we describe mixture models, which are described in Chapter 5 of the book, "Occupancy Estimation and Modeling" (section 5.1). Click on the worksheet labeled "Mixtures" and we'll get started.

BACKGROUND

The idea behind mixture models, also called heterogeneity models, is that sites in the study area are unique in some way, such that there is heterogeneity among sites in terms of detection and occupancy probability. In previous exercises, we explored how to use covariates to handle individual differences among sites. For example, we modeled occupancy as a function of habitat type and patch size, and

we modeled p as a function of precipitation. In these cases, site to site differences are handled by including covariates in the modeling process. Mixture models, in contrast, deal with unobservable differences among sites.

Before describing the concept of unobservable heterogeneity among sites, let's first reinforce the concept of observable heterogeneity. Observable heterogeneity refers to situations when the factors causing the differences can be identified.

> Example 1: Let's assume we are surveying a collection of sites for the occupancy of a small passerine bird. Each of the study sites has a different detection probability that is related to patch size: the larger the patch size, the lower the detection probability. The difference in detection probability may occur because males might be less likely to be paired in small patches if the small patches are lower in quality, and unmated males may sing more often to attract a mate, hence increasing the detection probability of the species at the site. If we can determine the patch size of each site, then we can treat this situation by analyzing patch as a covariate for p in the Design Matrix in MARK or PRESENCE. The data from just 5 study sites might look like this:

Site	Patch Size	p
1	0.1625369	0.583574
2	0.8273964	0.418874
3	-1.940898	0.919893
4	-0.7114	0.770547
5	-1.821647	0.910654

You run models where detection probabilities are constrained to be a function of a site's patch size, and then compare the results to models where detection probabilities are not constrained. Using AICc methods, you can decide if a model that includes the covariate patch size is better supported by the data. So, with this example, the covariate patch size allows a unique estimate of detection probability for each site, and this covariate might account for potential differences in detection probabilities.

Example 2: Detection of sites is a function of habitat because habitats vary in the density of a species of interest (some habitats have low densities, while others have high densities). If the probability of detecting a species of interest increases as population density increases (a reasonable assumption in most cases), then the covariate "habitat type" could be included in the p side of an occupancy model.

Both of these examples describe observable heterogeneity, and you've had some practice with covariate modeling in previous exercises.

Unobservable heterogeneity, in contrast, refers to situations when the factors causing differences in either occupancy probability or detection probability cannot be readily identified. This could simply mean we have absolutely no clue what might cause differences, but are willing to accept that there might be differences that we cannot measure. For instance, if food resources are a critical predictor of occupancy but cannot be measured readily across sites, it might impose heterogeneity among the study sites, where some sites are rich in food resources and others are poor, even though food was not measured directly. This

unobservable heterogeneity is the focus of this exercise. A key concept in heterogeneity models is that there may be a number of possible values for the detection probability at each site, and the likely value for a particular site is NOT know *a priori*. Heterogeneity models are well described for estimating animal abundance with closed capture methods (see Williams et al. 2001 for a description of the various estimators); MacKenzie et al. extend these ideas into the arena of occupancy modeling (section 5.1)

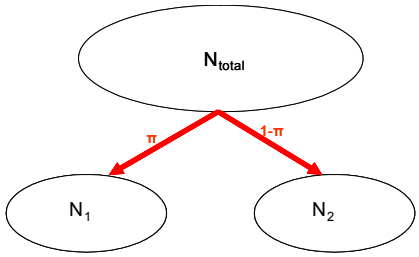
So, how does one model unobservable heterogeneity? Well, the basic idea is that the study sites can be divided into multiple groups, and each group (not each site) has unique detection probabilities and a unique probability of occupancy. The number of groups can be either a discrete number (e.g., 2 groups, 3 groups, etc.) or an infinite number, in which the probability of being in a particular group can be thought of as a random effect drawn from some statistical distribution. In this exercise, we will focus on a heterogeneity model in which the group number is discrete (n = 2), and heterogeneity is modeled for detection probability only.

Let's assume that we are conducting a survey of 250 sites for a songbird of interest, and that most of our observations come from the detection of singing males. Each site is surveyed 4 times under the assumption that the population is closed to changes in occupancy status, and with no errors in species identification. Let's assume that the sites differ in quality, so that fewer or no birds occur in the low quality sites than on the high quality sites. On high-quality sites where density is high, birds are readily detected because many birds sing during the 10 minute survey period. On low-quality sites where the population density is low, the species is harder to detect because very few animals occur on the site. Let's further

assume that the surveys are "rapid assessments" – meaning that sites are surveyed for some minimal amount of time so that all 250 sites can be studied. Thus, observers simply survey the sites quickly and do not have time to collect more detailed data, such as metrics that might describe habitat quality (a common design for large-scale surveys…..we'll cover study designs in a later chapter). In this example, we have two kinds of sites in the population: high quality sites that have many animals and high detection probability, and low quality sites that have few individuals and low detection probability. At any given site, the investigator does not know the group membership.

Field biologists conduct 4 surveys on all 250 sites, and no covariates are measured. Thus, there are $2^4 = 16$ possible encounter histories for the study: 1111, 1110, 1101, 1100, 1011, 1010, 1001, 1000, 0111, 0110, 0101, 0100, 0011, 0010, 0001, and 0000, and the frequency of each history is recorded.

Now, after all that work, we suspect that our sites differ in detection probability and occupancy probability, but we have no covariates to work with in the analysis. Now what? We could model our field data using a two-point mixture model (or two-group heterogeneity model). You can also run three-point mixtures, four-point mixtures, and so on. When analyzing real data, you would want to evaluate different numbers of mixtures (including 1) and see which mixture is best supported by the data. In this exercise, we'll compare two-group mixture models with one-group mixture models.



Let's look at this in more detail. We said above that when using mixture models we divide the study sites into groups. So, for the two-point mixture model we divide the study sites into two groups. The population of study sites (N_{total}) is divided into group 1 and group 2 such that $N_{total} = N_1 + N_2$, where N_1 is the total number of sites in group 1, and N_2 is the total number of sites in group 2. The first trick is to figure out the proportion of sites that belong to group 1 and the proportion of sites that belong to group 2. We let π represent the proportion of sites in group 1, so by definition $N_1 = \pi N_{total}$. Because there are only two groups being considered, if we know N_1 then we can derive N_2 as $N_2 = (1 - \pi) N_{total}$ because a site must be in group 2 if it wasn't in group 1. After we estimate the proportion of the population in each group (π and 1-π), the next step is to estimate the detection and occupancy probabilities for each group separately.

Group 1: proportion of sites in group 1 = πN_{total}, and we estimate detection parameters p_1, p_2, p_3, p_4 and the occupancy parameter ψ specific to group 1.

Group 2: proportion of sites in group 1 = $(1-\pi)N_{total}$, and we estimate detection parameters, p_1, p_2, p_3, p_4 and the occupancy parameter ψ specific to group 2. (Note: as you'll soon see, we really can't estimate a unique ψ for group 2)

MODEL PARAMETERIZATION AND OVERPARAMETERIZATION

The first thing you need to know about mixture models is that they are very data hungry. Because each group can have its own unique parameters, we are estimating up to twice as many parameters than in the 1 group, single season mixture model. Therefore, care must be taken to not over-parameterize a model. What is an

overparameterized model? Well, a model is overparameterized if the number of occupancy parameters to be estimated is greater than or equal to the number of unique encounter histories. A model that is overparameterized may not be identifiable – that is, it might have more than one solution. Clearly, you want to avoid this.

An example might make this more clear. Suppose you conduct a single-season survey with <u>three</u> different survey periods. In this case, there are $2^3 = 8$ possible encounter histories. By our definition of overparameterization –- a model is overparameterized if the number of parameters estimated is greater than or equal to the number of unique histories -- you can run a model that estimates 7 parameters at most. If you run a two-point mixture analysis, and attempt to estimate π, in addition to ψ, p_1, p_2, and p_3 for each group, you will be estimating 9 total parameters. Whoops! This model MIGHT run, and MIGHT even give you parameter estimates, but the estimates would not be valid. In other words, it's up to you to determine if the model is overparameterized or not.

Now, what if we surveyed the sites 4 times instead of 3 times? There would be $2^4 = 16$ types of encounter histories, and thus you can run a model with 15 or fewer occupancy parameters. Now, if you run this model for two groups, you would estimate π, ψ_1, $p_{1,1}$, $p_{2,1}$, $p_{3,1}$, $p_{4,1}$, ψ_2, $p_{1,2}$, $p_{2,2}$, $p_{3,2}$, $p_{4,2}$ – which is a total of 11 parameters. Thus, the model degrees of freedom would be 16 (the number of unique histories from the saturated model) minus 11 (from the occupancy model) = 5, so you haven't overparameterized this model. HOWEVER, and this is important, you must constrain $\psi_1 = \psi_2$ or the model is not identifiable. With that additional constraint, you would be estimating 10 total parameters. The take-home message

is that if you want to estimate time-specific p's in the mixture model, you'll need at least 4 occasions (16 possible histories).

ENCOUNTER HISTORIES

Although there are a lot more parameters to consider, the encounter history probabilities are a simple extension of what we did for the single-season, single-species exercise – the only addition is the group membership identification (π) or ($1-\pi$), and the group-specific detection and occupancy parameters. Because the detection parameters are group specific, we need to add some notation for a two-point mixture. Let detection probabilities be denoted as p_{ij}, where i denotes the occasion and j denotes the group. For example, p_{11} is the probability of detection for the first survey in a site in the first mixture and p_{12} is the probability of detection for the first survey for a site in the second mixture. So the second number that is subscripted identifies the group.

Let's start by writing out the 1111 encounter history probability for a single-season, single-species, 1 group model:
Probability of 1111 = $\psi\ p_1\ p_2\ p_3\ p_4$. Right?

OK, now let's extend this to a two-group mixture:
Probability of 1111 = $\pi\ \psi_1\ p_{1,1}\ p_{2,1}\ p_{3,1}\ p_{4,1} + (1-\pi)\ \psi_2\ p_{1,2}\ p_{2,2}\ p_{3,2}\ p_{4,2}$

In other words, the probability of obtaining a 1111 history is the sum of two probability terms, one for group 1 and the second for group 2. The probability of getting a 1111 history for group 1 is $\pi\ \psi_1\ p_{1,1}\ p_{2,1}\ p_{3,1}\ p_{4,1.}$ The site must first be part of group 1 (π), the site must be occupied (ψ_1), the species must be detected on the

first survey ($p_{1,1}$), the species must be detected on the second survey ($p_{2,1}$), the species must be detected on the third survey ($p_{3,1}$), and the species must be detected on the fourth survey ($p_{4,1}$). All of these terms are multiplied together because all of them must occur to generate a 1111 history for a site in group 1. The probability of getting a 1111 history for group 2 is $(1-\pi)\ \psi_2\ p_{1,2}\ p_{2,2}\ p_{3,2}\ p_{4,2}$. The site must first be part of group 2 ($1-\pi$), the site must be occupied (ψ_2), the species must be detected on the first survey ($p_{1,2}$), the species must be detected on the second survey ($p_{2,2}$), the species must be detected on the third survey ($p_{3,2}$), and the species must be detected on the fourth survey ($p_{4,2}$). All of these terms are multiplied together because all of them must occur to generate a 1111 history for a site in group 2. The two terms are added together because a site can be in either group 1 (in which case the parameters apply to group 1) OR it can be in group 2 (in which case the occupancy parameters apply to group 2).

Let's try another history, 0101. We know that ψ_1 must be equal to ψ_2, so we can drop the subscripts for ψ. In a one-group, single season model, the probability is $\psi\ (1-p_1)\ p_2\ (1-p_3)\ p_4$. We know the site was occupied because the species was detected at least once (ψ), and we know that it was detected on surveys two (p_2) and four (p_4) but not on surveys one ($1-p_1$) or three ($1-p_3$). If we extend this to a two-group mixture model, the probability of a 0101 is:

$\pi\ \psi\ (1-p_{1,1})\ p_{2,1}\ (1-p_{3,1})\ p_{4,1} + (1-\pi)\ \psi\ (1-p_{1,2})\ p_{2,2}\ (1-p_{3,2})\ p_{4,2}$.

Remember that the second subscript of a particular parameter ($p_{1,1}$) defines the group membership.

OK, let's go through just one more, 0000. In a one-group, single season model, the probability is $\psi\ (1-p_1)\ (1-p_2)\ (1-p_3)\ (1-p_4)+(1-\psi)$. The site could have been occupied

but missed on all four surveys, OR it could have been unoccupied. Extending this to two group mixture model, the probability of a 0000 history would have four terms, the first two apply to sites in group 1, while the last two apply to sites in group 2: $\pi \psi (1-p_{1,1}) (1-p_{2,1}) (1-p_{3,1}) (1-p_{4,1}) + \pi (1-\psi) + (1-\pi) \psi (1-p_{1,2}) (1-p_{2,2}) (1-p_{3,2}) (1-p_{4,2}) + (1-\pi) (1-\psi)$. Make sense? Don't forget that you have to distribute the π or $(1-\pi)$ terms for both options (the site is occupied but not detected, or the site is not occupied).

Let's see what happens if you let π equal 1. The probability statement for the 1111 encounter history becomes:

$$p(1111) = (1) \psi p_{1,1} p_{2,1} p_{3,1} p_{4,1} + (1 - 1) \psi_2 p_{1,2} p_{2,2} p_{3,2} p_{4,2}$$

or

$$p(1111) = (1) \psi p_{1,1} p_{2,1} p_{3,1} p_{4,1}$$

which should look very familiar by now. So, when you ran the basic single-species, single season occupancy model with no covariates in exercise 3, you ran a "mixture" model in which there was only one group. Hopefully, the structure of the model is making sense to you. Spend some time working through the other histories before working on the spreadsheet.

TWO-GROUP OCCUPANCY MIXTURE MODEL SPREADSHEET INPUTS

OK, now let's get oriented to the spreadsheet. In this example, the investigator surveys 250 study sites, with each site being surveyed 4 times. The encounter histories are recorded in cells B4:B19, and the frequency of each history is recorded in cells C4:C19. The total number of sites is given in cell C20, and the number of unique histories is given in cell C21 (which you might remember indicates

the number of terms in our multinomial likelihood function). To avoid over-parameterization, you can only run models with 15 or fewer parameters. The naïve estimate for occupancy (occupancy unadjusted for detection probability) is computed in cell C22 as the total number of sites which had one or more detections divided by the total number of sites.

	B	C	D	E	F	G
3	History	Frequency	Parameter	Estimate?	Betas	MLE
4	1111	15	π	1		0.50000
5	1110	17	**Mixture 1**			
6	1101	12	ψ	1		0.50000
7	1100	18	p1	1		0.50000
8	1011	8	p2	1		0.50000
9	1010	13	p3	1		0.50000
10	1001	9	p4	1		0.50000
11	1000	20	**Mixture 2**			
12	0111	5	ψ	0	0.0000	0.50000
13	0110	9	p1	1		0.50000
14	0101	7	p2	1		0.50000
15	0100	15	p3	1		0.50000
16	0011	5	p4	1		0.50000
17	0010	12				
18	0001	9	Click Button to Add Results to Results Table			
19	0000	76				
20	# Sites =	250				
21	# Histories =	16	$\ln(L(p_i \mid n_i, y_i) \propto y_1 \ln(p_1) + y_2 \ln(p_2) + y_3 \ln(p_3) + ... + y_{16} \ln(p_{16})$			
22	Naïve Estimate	0.696				

OK, now let's look at the parameters. Notice the spreadsheet is divided into three sections. In the first section (cells D4:G4), the parameter π is listed. This is the proportion of the total sites that belong to mixture 1. In a two-group mixture model, you MUST estimate π, or the proportion of sites that belong to group 1. The second section of the spreadsheet lists the parameters (ψ, p_1, p_2, p_3, p_4) for mixture 1 (cells D6:G10), and the third portion of the spreadsheet lists the parameters for mixture 2 (cells D12:G16). As with other spreadsheet exercises, you enter a 1 when a parameter is being uniquely estimated, or enter a 0 if the

parameter is being forced to be equal to some other parameter. Don't forget that ψ_1 must equal ψ_2, so we will not be estimating ψ_2 and instead will force it to equal ψ_1 (hence the 0 in cell E12).

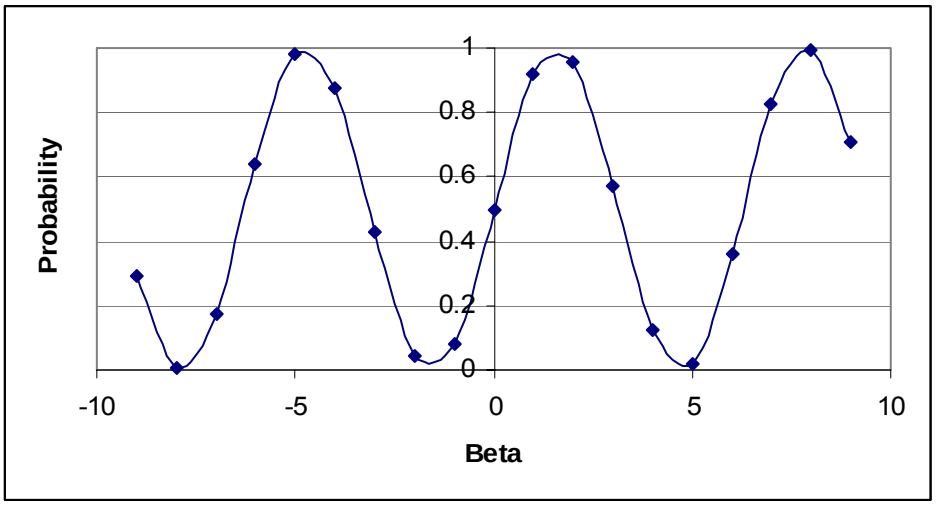
MIXTURE LINKS

The betas for each parameter are listed in cells F4:F16. We don't know what these values should be and will let Solver find them for us. Before running a mixture model in the spreadsheet, the betas should be cleared out (with the exception of cell F12). The parameter estimates that correspond to each beta are computed in cells G4:G16 through a sin link. Click on cell G4 and you'll see the sin link transformation: =(SIN(F4)+1)/2. Remember the sin link constrains the MLE's to be between 0 and 1, which is what we want for π, ψ, and the p_{ij}'s. We can use this link (instead of the now-familiar logit link) because covariates aren't included in mixture models. This is the default link in PRESENCE for two-group mixture models. We introduced the sin link in Exercise 3, but it never hurts to review. In the example below, we entered beta values of -2, -1, 0, 1, and 2 for ψ, p_1, p_2, p_3, and p_4, respectively.

	D	E	F	G
6	ψ	1	-2.0000	0.04535
7	p1	1	-1.0000	0.07926
8	p2	1	0.0000	0.50000
9	p3	1	1.0000	0.92074
10	p4	1	2.0000	0.95465

These betas correspond to the following probabilities: ψ = 0.04535, p_1 = 0.07926, p_2 = 0.5000, p_3 = 0.92074, and p_4 = 0.95465. Note that the betas can take on any

value, and the link function constrains the MLE's to be between 0 and 1, which is necessary because π, p_1, p_2, p_3, p_4,and ψ are probabilities, and probabilities range between 0 and 1. The figure below shows betas that range from -9 to +9, and the corresponding, sin "transformed" probability estimate. Look for the beta values of -2, -1, 0, 1, and 2, and find their corresponding probabilities on the graph:



Now enter some numbers of your choice in the beta column (cells F4:F16) and examine the MLE's. You should see that no matter what beta values you enter, the corresponding parameters are constrained between 0 and 1.

Keep in mind that we don't know what the beta values are…..we are going to let Solver find the betas that maximize the multinomial log likelihood function (see below). If the connection between betas and MLE's is still rusty to you, go back to Exercise 3 and review the material there…things will only get worse if you plow ahead without understanding this very fundamental concept.

MULTINOMIAL LOGIT LINK

OK, time for a little side comment. If you run heterogeneity models with more than two groups, the sin or logit link won't work. Why? Well, in our two-group example, we estimate π as the proportion of study sites in mixture 1. If we know the proportion of sites that belong to group 1, by definition we know the proportion of sites that belong to group 2 (which is 1-π). Now, what if we had three groups? We need to estimate the proportion of sites in group 1, group 2, and group 3. The sum of these proportions must be 1. With two groups this was easily handled by subtraction. But with three or more groups, there might be the potential, say, for the proportion for group 1 to be 0.5 and the proportion for group 2 to be 0.7. This won't do. We need a special link that will force the sum of the proportions to equal 1. This link is called the multinomial link, and is what PRESENCE uses for any heterogeneity model with three or more groups.

The multinomial logit link for a three group problem has the form:
Proportion in Group 1 = exp(B1)/(1 + exp(B1) + exp(B2))
Proportion in Group 2 = exp(B2)/(1 + exp(B1) + exp(B2))
Proportion in Group 3 = 1 – proportion in Group 1 – proportion in Group 2.

SPREADSHEET HISTORY PROBABILITIES

OK! Back to the spreadsheet. Now we are ready to compute the probability of realizing each history. Let's start with the first history listed, 1111, in cell B4.

The probability of realizing a 1111 history is estimated for each group separately. If a site is in group 1, the probability of realizing a 1111 history is $\pi\ \psi_1\ p_{1,1}\ p_{2,1}\ p_{3,1}\ p_{4,1}$. This equation is entered in cell H4: =G4*G6*G7*G8*G9*G10. If a site is in group 2, the probability of realizing a 1111 history is $(1-\pi)\ \psi_2\ p_{1,2}\ p_{2,2}\ p_{3,2}\ p_{4,2}$ This

equation is entered in cell I4: =(1-G4)*G12*G13*G14*G15*G16. Some portion of the 250 study sites will be in group 1, and some portion will be in group 2. Across both groups, the probability of realizing a 1111 history is the sum of the two mixing probabilities, given in cell J4 (=H4+I4). The natural log of the combined history probabilities is computed in cells K4:K19.

Make sense? Spend time now clicking on the formula for each mixture, and be sure to recognize that obtaining the final history probabilities depends on 1) the proportion of sites in group 1 versus group 2, and 2) the estimates of ψ, p_1, p_2, and p_3 that are unique to each group.

Notice that the sum of cells J4:J19 must equal 1 (cell J20): there are 16 possible histories, and each history has a probability of being realized, but the sum of the probabilities must be 1.00.

THE MIXTURE MODEL MULTINOMIAL LOG LIKELIHOOD

The goal of the analysis, as you might have guessed, is to find the combination of betas that maximizes the multinomial log likelihood function. Remember, by changing the betas, we change the parameter estimates linked to each beta, which changes the probability of each encounter history, which changes the Log_eL.

Betas → MLEs → Encounter Histories → Log_eL

All that's left is to compute the log likelihood, given the frequencies of each history and the history's probability. The multinomial log likelihood formula that we've been using is in the blue box below.

$$\ln(L(p_i \mid n_i, y_i) \propto y_1 \ln(p_1) + y_2 \ln(p_2) + y_3 \ln(p_3) + \ldots\ldots + y_{16} \ln(p_{16})$$

There are 16 terms in this function, one for each of the encounter histories. The y_i in the blue box are the frequencies of each kind of history and the p_i in the blue box equation above are the history probabilities. The Log_eL is computed in cell B26 with the equation =SUMPRODUCT(C4:C19,K4:K19), which corresponds to the general formula in the blue box. Now all we have to do is maximize this value to find the MLE's for our dataset.

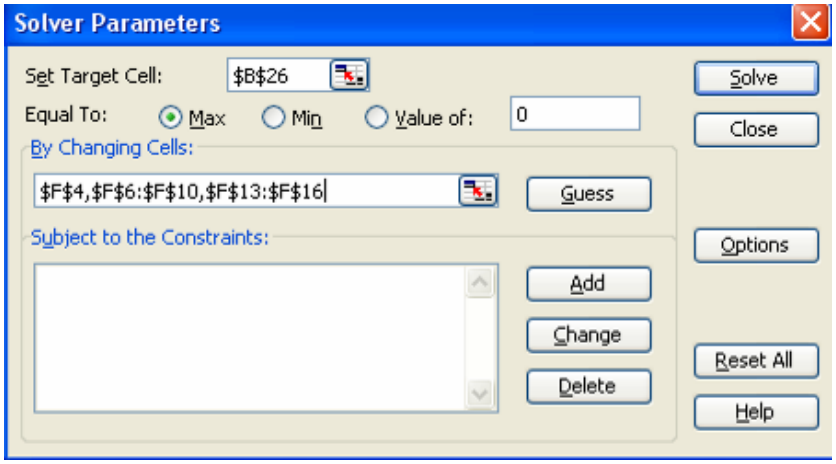
MAXIMIZING THE LOG LIKELIHOOD

Before we run our first model, we need to make sure that $\psi_1 = \psi_2$, so set your spreadsheet up as follows:

	D	E	F	G
3	Parameter	Estimate?	Betas	MLE
4	π	1		=(SIN(F4)+1)/2
5	**Mixture 1**			
6	ψ	1		=(SIN(F6)+1)/2
7	p1	1		=(SIN(F7)+1)/2
8	p2	1		=(SIN(F8)+1)/2
9	p3	1		=(SIN(F9)+1)/2
10	p4	1		=(SIN(F10)+1)/2
11	**Mixture 2**			
12	ψ	0	=F6	=(SIN(F12)+1)/2
13	p1	1		=(SIN(F13)+1)/2
14	p2	1		=(SIN(F14)+1)/2
15	p3	1		=(SIN(F15)+1)/2
16	p4	1		=(SIN(F16)+1)/2

Make sure that the beta cells are cleared out. OK, now we're ready to run this model. We can call this model "ψ, p(t) – two groups" to indicate that we're

estimating π and ψ, plus p_1, p_2, p_3, and p_4 for each group. You know the drill. Open Solver, and set cell B26 to a maximum by changing cells F4, F6:F10, F13:F16.



Press Solve and Solver will work through the various combinations of betas until it finds the maximum.

MIXTURE MODEL OUTPUT

First, let's take a look at the parameter estimates found by Solver:

	D	E	F	G
3	Parameter	Estimate?	Betas	MLE
4	π	1	0.3468	0.66997
5	**Mixture 1**			
6	ψ	1	0.6133	0.78778
7	p1	1	-0.1302	0.43510
8	p2	1	-0.2408	0.38076
9	p3	1	-0.3431	0.33179
10	p4	1	-0.4728	0.27231
11	**Mixture 2**			
12	ψ	0	0.6133	0.78778
13	p1	1	0.7474	0.83988
14	p2	1	0.4888	0.73479
15	p3	1	0.2399	0.61881
16	p4	1	0.0484	0.52417

For the observed field data, Solver maximized the likelihood by placing the study sites into one of two groups. The proportion of sites belonging to group 1 is 0.66997 (cell G4). By subtraction, the proportion of sites belonging to group 2 is (1- 0.66997) = 0.33003. This means that there are two distinct kinds of "sites" in the data, none of which was directly observed by the researcher. For group 1, the probability of occupancy is 0.78778 (cell G6), p_1 = 0.43510 (cell G7), p_2 = 0.38076 (cell G8), p_3 = 0.33179 (cell G12), and p_4 = 0.27231. For group 2, the probability of occupancy is still 0.78778 (cell G12), p_1 = 0.83988, p_2 = 0.73479, p_3 = 0.61881, and p_4 = 0.52417. So the two groups have equal occupancy probabilities (as modeled), but their detection rates are quite different: group 2 has higher detection rates than group 1 in all four surveys.

Now let's look at the remaining output given in cells B25:L26.

	B	C	D	E	F	G	H	I	J	K	L
24	OUTPUTS										
25	Log_eL	$-2Log_eL$	K	AIC	AICc	$-2Log_eL$ Sat	Deviance	Model DF	C-hat	Chi-Square	P value
26	-611.24	1222.4847	10	1242.48	1243.41	1222.4472	0.0376	6	0.006259	0.0182	1.0000

The Log_eL is given in cell B26. Cell C26 is -2 times cell B26, and is the $-2Log_eL$. K is the number of parameters in any given model, and the underlying equation is =SUM(E4,E6:E10,E12:E16). AIC is computed as the $-2Log_eL$ + 2*K. AICc is the second order correction of AIC, and uses the number of study sites in the calculation. Deviance is computed as the difference between the saturated model's $-2Log_eL$ and the current model's $-2Log_eL$; the lower the number the better. Remember that by definition the saturated model is a model in which the data "fit" the model perfectly. The saturated model's $-2Log_eL$ is computed in the usual way (as in previous exercises) in cells N4:O21. The model we just ran had a deviance of 0.0376, which means it is about as good as you can possibly get. The Model Degrees of Freedom is the number of unique histories minus K. In a model without covariates, as long as the Model Degrees of Freedom is positive, you haven't overparameterized your model. C-hat is computed in cells J26 as Deviance divided by DF. The C-hat in this case is close to 0. C-hats larger than 1 might indicate some kind of lack of fit, which we don't need to worry about for this example. The Chi-Square statistic and associated p-value are given in cells K18:L18. The Chi-square computations are provided in the orange cells L4:M19. As you can see, the model results generate expected values that almost perfectly match the observed values, so the chi-square value is about 0 (cell M20) and the p value is 1 (cell L26). All in all, the parameters found by Solver provide a nearly perfect match to the data (perhaps due to the fact that the data were generated with this exact model by expectation!).

Click on the button labeled Model 1 to add your results to the Results Table.

	C	D	E	F	G	H	I	J
29		Model	Log_eL	$-2Log_eL$	K	AIC	AICc	Rank
30	1	ψ p(t)2groups	-611.2489363	1222.497873	10	1242.498	1243.4184	1
31	2	ψ p(.)2groups						#N/A
32	3	ψ p(t)						#N/A
33	4	ψ p(.)						#N/A

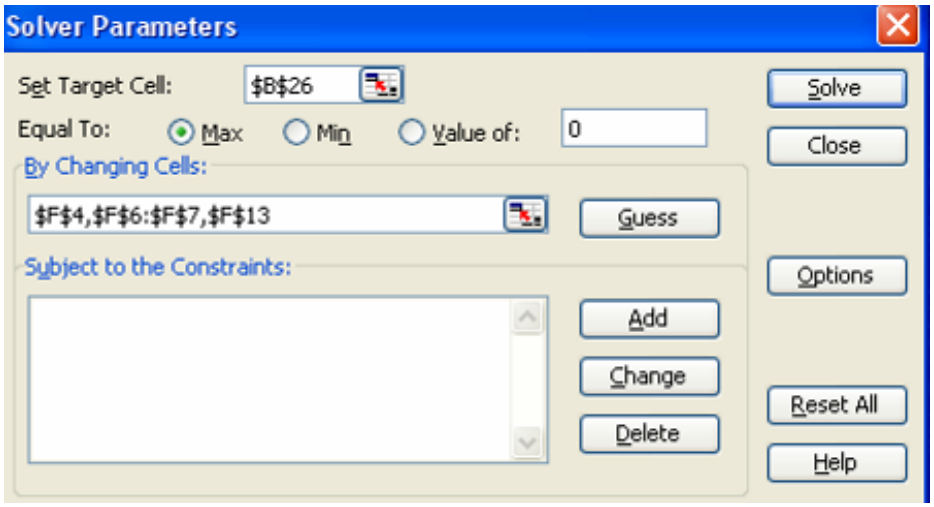
OK, now that you've run one model, we'll run three more: a second mixture model where p is estimated for each group, but is not time-dependent, and then the standard models ψp(t) and ψp(.) models (with one group). Then we'll run the same analyses in PRESENCE to learn some very important concepts.

MODEL P(.)PSI FOR TWO GROUPS

In this model, we will estimate a single p for each group, and will force ψ to be equal for both groups. That is, for group 1, $p_{1,1} = p_{2,1} = p_{3,1}$, and for group 2, $p_{1,2} = p_{2,2} = p_{3,2}$. Think about how you would set this up in the spreadsheet.

	D	E	F	G
3	Parameter	Estimate?	Betas	MLE
4	π	1		=(SIN(F4)+1)/2
5	**Mixture 1**			
6	ψ	1		=(SIN(F6)+1)/2
7	p1	1		=(SIN(F7)+1)/2
8	p2	0	=F7	=(SIN(F8)+1)/2
9	p3	0	=F7	=(SIN(F9)+1)/2
10	p4	0	=F7	=(SIN(F10)+1)/2
11	**Mixture 2**			
12	ψ	0	=F6	=(SIN(F12)+1)/2
13	p1	1		=(SIN(F13)+1)/2
14	p2	0	=F13	=(SIN(F14)+1)/2
15	p3	0	=F13	=(SIN(F15)+1)/2
16	p4	0	=F13	=(SIN(F16)+1)/2

First, we must estimate π, so we enter a 1 in cell E4. Then, for mixture 1, we estimate ψ, and enter a 1 in cell E6. Then we estimate p_1, and enter a 1 in cell E7. Then, we force p_2, p_3, and p_4 in mixture 1 to be equal to p_1 in mixture 1 and enter 0's in cells E8:E10. For mixture 2, we force ψ for mixture 2 to be equal to ψ for mixture 1. We will estimate p_1, so we enter a 1 in cell E13. Then we force p_2, p_3, and p_4 to be equal to p_1 for mixture 2. So, the total number of parameters to be estimated for this model is 4. Let's run it and see if it is more parsimonious than the previous model. Open Solver, and set cell B26 to a maximum by changing cells F4, F6:7, F13. Then Solve.



Now let's look at the output:

	D	E	F	G
3	Parameter	Estimate?	Betas	MLE
4	π	1	0.0000	0.50000
5	**Mixture 1**			
6	ψ	1	0.5188	0.74793
7	p1	1	-0.0267	0.48668
8	p2	0	-0.0267	0.48668
9	p3	0	-0.0267	0.48668
10	p4	0	-0.0267	0.48668
11	**Mixture 2**			
12	ψ	0	0.5188	0.74793
13	p1	1	-0.0267	0.48668
14	p2	0	-0.0267	0.48668
15	p3	0	-0.0267	0.48668
16	p4	0	-0.0267	0.48668

First, note that the Log_eL is 624.56. Now take at look at the MLE's -- a warning sound should be going off in your head. This model estimated that the two groups have equal membership (p = 0.500), and that ψ is 0.74793. Even though the p_{ij}'s were modeled to be group specific, the model results indicate that both groups have a probability of detection of 0.48668. You might be thinking, what is the chance of getting exactly the same detection probabilities for both groups, when they were modeled separately? Very, very low. So, let's try another experiment. This time, when you set up your model, instead of clearing the betas in cells F4, F6:7, F13, enter a random number, or =rand() in those cells. These are "seed" values and Solver will start with these beta values when it searches for the maximum likelihood.

	D	E	F	G
3	Parameter	Estimate?	Betas	MLE
4	π	1	=RAND()	=(SIN(F4)+1)/2
5	**Mixture 1**			
6	ψ	1	=RAND()	=(SIN(F6)+1)/2
7	p1	1	=RAND()	=(SIN(F7)+1)/2
8	p2	0	=F7	=(SIN(F8)+1)/2
9	p3	0	=F7	=(SIN(F9)+1)/2
10	p4	0	=F7	=(SIN(F10)+1)/2
11	**Mixture 2**			
12	ψ	0	=F6	=(SIN(F12)+1)/2
13	p1	1	=RAND()	=(SIN(F13)+1)/2
14	p2	0	=F13	=(SIN(F14)+1)/2
15	p3	0	=F13	=(SIN(F15)+1)/2
16	p4	0	=F13	=(SIN(F16)+1)/2

Now, run Solver again and see what happens. Here are our new results:

	D	E	F	G
3	Parameter	Estimate?	Betas	MLE
4	π	1	0.2403	0.61899
5	**Mixture 1**			
6	ψ	1	0.6208	0.79085
7	p1	1	-0.3156	0.34480
8	p2	0	-0.3156	0.34480
9	p3	0	-0.3156	0.34480
10	p4	0	-0.3156	0.34480
11	**Mixture 2**			
12	ψ	0	0.6208	0.79085
13	p1	1	0.3002	0.64786
14	p2	0	0.3002	0.64786
15	p3	0	0.3002	0.64786
16	p4	0	0.3002	0.64786

You can quickly see that the MLE's are different than the previous model run.

Each group now has unique detection parameters, as specified by the model. Also,

it is important to note that the Log_eL is 622.60, lower than the first model run. So Solver found a solution in the first run that really wasn't the maximum. The results from PRESENCE match this second run. What happened is that Solver found a local maximum in the first run – a point on a bumpy-shaped likelihood surface that it thought was the peak, when unknown to it there was a slightly higher peak somewhere else on the likelihood surface. On the second run Solver found the true maximum.

Press the Model 2 button (around cell E19) to add your results to the Results Table.

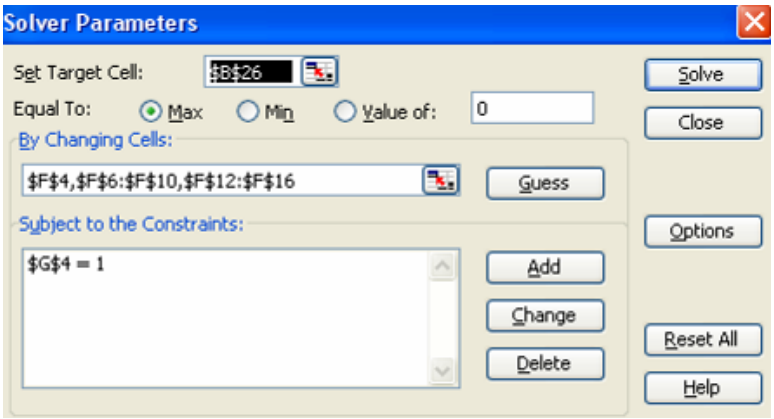
	C	D	E	F	G	H	I	J
29		Model	Log_eL	$-2Log_eL$	K	AIC	AICc	Rank
30	1	ψ p(t)2groups	-611.2489363	1222.497873	10	1242.498	1243.4184	1
31	2	ψ p(.)2groups	-622.5977932	1245.195586	4	1253.196	1253.3589	2
32	3	ψ p(t)						#N/A
33	4	ψ p(.)						#N/A

MODEL P(t)PSI (ONE GROUP)

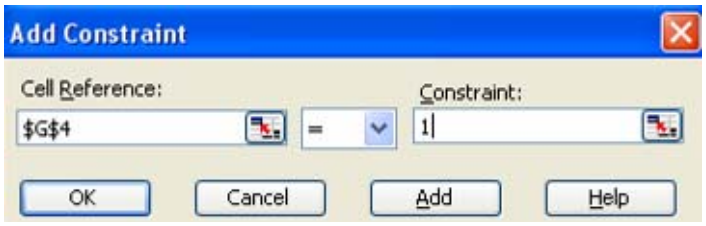
OK, two more models to go, the first of which is the standard model ψ,p(t), just like you ran in Exercise 3. In this model, all the sites belong to a single group, so we will not be estimating π (we know it will be 1), and we will not be estimating any of the parameters associated with group 2. So this model will estimate 5 total parameters:

	D	E	F	G
3	Parameter	Estimate?	Betas	MLE
4	π	0		0.50000
5	**Mixture 1**			
6	ψ	1		0.50000
7	p1	1		0.50000
8	p2	1		0.50000
9	p3	1		0.50000
10	p4	1		0.50000
11	**Mixture 2**			
12	ψ	0		0.50000
13	p1	0		0.50000
14	p2	0		0.50000
15	p3	0		0.50000
16	p4	0		0.50000

OK, now to run this model, we will estimate all the betas uniquely, but will have to add a constraint within Solver itself. Open Solver, and set cell B26 to a maximum by changing cells F4, F6:F10, and F12:F16. This time, however, add the constraint that cell G4 (π) must be equal to 1.



We added this constraint by clicking on the Add button within the Solver dialogue box, typing in the constraints, and pressing the OK button:



Now when you run Solver, you should get the following results:

	D	E	F	G
3	Parameter	Estimate?	Betas	MLE
4	π	0	1.5708	1.00000
5	**Mixture 1**			
6	ψ	1	0.5094	0.74384
7	p1	1	0.2060	0.60228
8	p2	1	0.0540	0.52699
9	p3	1	-0.0967	0.45171
10	p4	1	-0.2497	0.37642
11	**Mixture 2**			
12	ψ	0	0.0000	0.50000
13	p1	0	0.0000	0.50000
14	p2	0	0.0000	0.50000
15	p3	0	0.0000	0.50000
16	p4	0	0.0000	0.50000

The only estimates that we need to be concerned with are the parameters associated with group 1. As indicated, π is 1.000. It doesn't matter what the estimates are for mixture 2. Why? Because each history probability in mixture 2 is multiplied by 1-π, or 0. This is reflected in columns H and I, where only mixture 1 has non-zero history probabilities.

Go ahead and add the results of this model to the Results Table:

	C	D	E	F	G	H	I	J
29		Model	Log_eL	$-2Log_eL$	K	AIC	AICc	Rank
30	1	ψ p(t)2groups	-611.2489363	1222.497873	10	1242.498	1243.4184	2
31	**2**	ψ p(.)2groups	-622.5977932	1245.195586	4	1253.196	1253.3589	3
32	**3**	ψ p(t)	-613.9628668	1227.925734	5	1237.926	1238.1716	1
33	4	ψ p(.)						#N/A

You can see that this model beats the two mixture models (even though you'll see later that the data were simulated with strong mixture effects). We'll explore this more when we run the models in PRESENCE.

MODEL P(.)PSI (ONE GROUP)

OK, the last model is the one-group, constant p model, so you will be estimating just two parameters. Go ahead and set this model up and run it. Don't forget to add the constraint that $\pi = 1$ within Solver. Here are our results:

	D	E	F	G
3	Parameter	Estimate?	Betas	MLE
4	π	0	1.5708	1.00000
5	**Mixture 1**			
6	ψ	1	0.5188	0.74793
7	p1	1	-0.0267	0.48668
8	p2	0	-0.0267	0.48668
9	p3	0	-0.0267	0.48668
10	p4	0	-0.0267	0.48668
11	**Mixture 2**			
12	ψ	0	0.0000	0.50000
13	p1	0	0.0000	0.50000
14	p2	0	0.0000	0.50000
15	p3	0	0.0000	0.50000
16	p4	0	0.0000	0.50000

Click on the Model 4 button to add the results to the Results Table:

	C	D	E	F	G	H	I	J
29		Model	Log_eL	$-2Log_eL$	K	AIC	AICc	Rank
30	1	ψ p(t)2groups	-611.2489363	1222.497873	10	1242.498	1243.4184	2
31	**2**	ψ p(.)2groups	-622.5977932	1245.195586	4	1253.196	1253.3589	4
32	**3**	ψ p(t)	-613.9628668	1227.925734	5	1237.926	1238.1716	1
33	4	ψ p(.)	-624.5631978	1249.126396	2	1253.126	1253.175	3

Our top-ranked model is model ψ p(t), with an AICc score of 1238.2. The second best model is the ψ p(t) two group mixture model, with an AICc score of 1243.4. This is a difference of 5.2 AICc units; there is some support for the mixture model, but it is not overwhelming. (This would be a good time to model average!). Take a look at the Log_eL's of the two models.

	C	D	E	F	G	H	I	J
29		Model	Log_eL	$-2Log_eL$	K	AIC	AICc	Rank
30	1	ψ p(t)2groups	-611.2489363	1222.497873	10	1242.498	1243.4184	2
31	**2**	ψ p(.)2groups	-622.5977932	1245.195586	4	1253.196	1253.3589	4
32	**3**	ψ p(t)	-613.9628668	1227.925734	5	1237.926	1238.1716	1
33	4	ψ p(.)	-624.5631978	1249.126396	2	1253.126	1253.175	3

The two group mixture model (model 1) had a Log_eL of -611.2, whereas the top ranked model (model 3) had a Log_eL of -613.9. So model 1 "fit" the data better than the top-ranked model (remember that it had a deviance of almost 0). But because model 1 estimated 10 parameters compared to model 3 (which estimated 5 parameters), its AICc score was increased. We'll explore the trade-offs between model fit, number of parameters, and precision in PRESENCE.

PRESENCE INPUT FILES

	A
2	Tally
3	0
4	15
5	32
6	44
7	62
8	70
9	83
10	92
11	112
12	117
13	126
14	133
15	148
16	153
17	165
18	174
19	250

The histories and corresponding frequencies given in cells B4:B11 cannot be input directly into PRESENCE (most users of PRESENCE include covariates in the analysis, so the input files are set up on a site-by-site basis). So, we've entered some formulae in columns P:T to convert the summarized data to site-specific data. But before we cover the equations, first look at cells A2:A19, which are shaded grey on the spreadsheet. These cells are a running tally of the total number of sites in the study. Beginning with the first history (1111), the cell A4's formula counts the number of sites that are 1111. The next cell (cell A5) counts the number of 1110 sites + the 1111 sites. The next cell (cell A6) counts the number of 1101, 1110, and 1111 sites, and so on. We will use this running tally to create PRESENCE input files.

Now let's turn our attention to columns Q:V. In column Q, the sites are listed from 1 to 250 down the column. In column R, we assign a history to each site, using the tally in cells A3:A19. Click on cell R4. The equation there is =LOOKUP(Q4-1,$A$3:$A$19,$B$4:$B$19). The function looks up the value in Q4 (the site number) minus 1 in the tally column (A3:A19), and then returns the corresponding history listed in cells B4:B19. Because the lookup vector (the tally) is sorted in ascending order, this equation "works" for our purposes because the LOOKUP function doesn't need to find an exact match. Take a look again at cells B4:C19. Notice that there are 15 sites with a 1111 history, 17 sites with a 1110 history, 12 sites with a 1101 history, and so on. When the lookup function in column R is copied down the column, the result is that the first 15 sites are given a 1111

history, the next 17 sites are given a 1110 history, the next 12 sites are given a 1101 history, and so on. Given the histories in column R, columns S:V simply split each history into survey-specific results (using LEFT, RIGHT, and MID functions). When it's time to create a PRESENCE input file, simply copy cells S4:V253 and paste them into the PRESENCE datasheet.

	Q	R	S	T	U	V
3	Site	History	Survey 1	Survey 2	Survey 3	Survey 4
4	1	1111	1	1	1	1
5	2	1111	1	1	1	1
6	3	1111	1	1	1	1
7	4	1111	1	1	1	1
8	5	1111	1	1	1	1

SIMULATING TWO GROUP MIXTURE DATA

OK, we're almost finished with the spreadsheet exercise. The last section of the spreadsheet is for simulating new data, either by expectation or with stochasticity. Take a look at cells X3:AE6.

	X	Y	Z	AA	AB	AC	AD	AE
3	Parameter	**p1**	**p2**	**p3**	**p4**	**ψ**	**N**	**π**
4	MLE Group 1	0.8	0.7	0.6	0.5	0.8	100	0.4
5	MLE Group 2	0.4	0.35	0.3	0.25	0.8	150	0.6
6						Total Sites:	250	

In this section of the spreadsheet, you enter parameter values (p_1, p_2, p_3, ψ, N, and π) for group 1 and for group 2. The data we've been analyzing in this exercise were obtained from the parameter estimates shown above. Note that you make your entries in the cells shaded in blue – the other cells are computed. For example, in cell AD6, enter the total number of sites. In cell AE4, enter the proportion of sites that are in group 1. Cell AE5 is computed as $1-\pi$. Cells AD4 and AD5 are also computed based on the entries for total sites and π.

Given these entries, we can create data in two ways. First, we can create data based on expectation.

	AB	AC	AD	AE	AF
8	**Summarized Expected Data:**				
9		Mix 1	Mix 2	Total	Rounded
10	1111	13.44	1.26	14.7	15
11	1110	13.44	3.78	17.22	17
12	1101	8.96	2.94	11.9	12
13	1100	8.96	8.82	17.78	18
14	1011	5.76	2.34	8.1	8
15	1010	5.76	7.02	12.78	13
16	1001	3.84	5.46	9.3	9
17	1000	3.84	16.38	20.22	20
18	0111	3.36	1.89	5.25	5
19	0110	3.36	5.67	9.03	9
20	0101	2.24	4.41	6.65	7
21	0100	2.24	13.23	15.47	15
22	0011	1.44	3.51	4.95	5
23	0010	1.44	10.53	11.97	12
24	0001	0.96	8.19	9.15	9
25	0000	20.96	54.57	75.53	76
26		100	150	250	250

Creating data based on expectation is really quite simple. In cells AB10:AB25, we list the possible encounter histories. Then, given the parameter estimates described above, we compute the number of sites in each group that should have a particular history. For example, in cell AC10 we compute the number of sites in mixture 1 that should have a 1111 history. The equation is =AC4*Y4*Z4*AA4*AB4*AD4, which is $N_1 \psi_1 p_{1,1} p_{2,1} p_{3,1} p_{4,1}$. In cell AD10 we compute the number of sites in mixture 2 that should have a 1111 history. The

equation is =AC5*Y5*Z5*AA5*AB5*AD5, which is $N_2 \psi_2 p_{1,2} p_{2,2} p_{3,2} p_{4,2.}$ In cell AE10 we add the group 1 + group 2 sites. In cell AF10, we simply round the total for data analysis purposes. The reason that the some of the models we ran earlier fit so well is that we analyzed data created by expectation. As long as Solver finds the MLE's, the expected frequency of encounter histories should match the observed frequency of encounter histories.

The second method of creating data for analysis is with some stochasticity. This method is demonstrated in cells X29:AG279.

	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
28								History	History	Actual
29	Site	Group	rand() ψ	rand() p1	rand() p2	rand() p3	rand() p4	1	2	History
30	1	2	0.793062	0.606474	0.593843	0.7477917	0.80213	1100	0000	0000
31	2	2	0.750786	0.096962	0.845042	0.9888574	0.62718	1000	1000	1000
32	3	1	0.890587	0.887353	0.975259	0.814502	0.84084	0000	0000	0000
33	4	1	0.27869	0.090697	0.428131	0.153707	0.91658	1110	1010	1110
34	5	2	0.108711	0.748171	0.850801	0.9430909	0.28547	1001	0000	0000

In column X, we list the 250 sites. In column Y, we assign a group membership to each site. Click on cell Y30 and you'll see the formula =IF(RAND()<$AE$4,1,2). If a random number is less than π (given in cell AE4), the site belongs to group 1, otherwise it belongs to group 2. This formula is copied down for the remaining sites until all 250 sites are assigned a group membership. As in previous exercises, the next five columns (columns AA:AD are simply random numbers, with the equation =RAND().

OK, now we're ready to create an encounter history for each site. One easy way to do that is to write out an encounter history for a site if it belongs to group 1 (column AE) and then to write a second encounter history for the same site if it belongs to group 2 (column AF). The encounter histories follow the exact same

logic as we discussed in the introduction of this exercise. Click on cell AE30. The equation is: =

=IF(AND(Z30<$AC$4,AA30<$Y$4),1,0)&IF(AND(Z30<$AC$4,AB30<$Z$4),1,0)&IF(AND(Z30<$AC$4,AC30<$AA$4),1,0)&IF(AND(Z30<$AC$4,AD30<$AB$4),1,0).

This formula should look at least vaguely familiar to you. It is composed of four parts (written above in black, blue, red, and green -- one for each of the four survey periods). Let's look at the first part:

=IF(AND(Z30<$AC$4,AA30<$Y$4),1,0), which is a IF function with a nested AND function within it. IF cell Z30 <$AC$4 (the random ψ for group 1 is less than the specified ψ for group 1) AND if cell AA30 <$Y$4 (the random p_1 is less than the specified p_1 for group 1), then return the number 1 (the species was detected); otherwise return a 0 (the species was not detected). This portion of the formula therefore returns a 1 or 0 for the first survey for site 1. The other portions of the formula follow the same logic…work your way through them now. In column AF, we repeat the exercise, but this time use the parameters associated with group 2.

Finally, in column AG, we assign the actual history for each site with a HLOOKUP function. Click on cell AG30 and you'll see the formula =HLOOKUP(Y30,$AE$29:$AF$279,X30+1). This formula looks up the assigned group membership for site 1 (listed in cell Y30) in a large table in which the first column is either a 1 or a 2 (cells $AE$29:$AF$279). If the site is in group 1, the function returns the history associated with group 1, and if the site is in group 2, the function returns the history associated with group 2. If you haven't used HLOOKUP or LOOKUP functions yet, we encourage you to spend a bit of time learning about them – they're pretty handy functions for many tasks.

The stochastic data are summarized in cells X10:Y26. If you create data this way, copy cells Y10:Y25 and paste them into cells C4:C19 (the histories are ordered in the same way). Then you can copy the data in columns Q:V for inputting into PRESENCE.

	X	Y	Z
8	**Summarized Stochastic Data:**		
9			
10	1111	21	
11	1110	14	
12	1101	13	
13	1100	20	
14	1011	7	
15	1010	13	
16	1001	14	
17	1000	25	
18	0111	3	
19	0110	6	
20	0101	5	
21	0100	18	
22	0011	3	
23	0010	15	
24	0001	10	
25	0000	63	
26		250	

SINGLE SEASON OCCUPANCY MODELS ANALYSIS IN PRESENCE

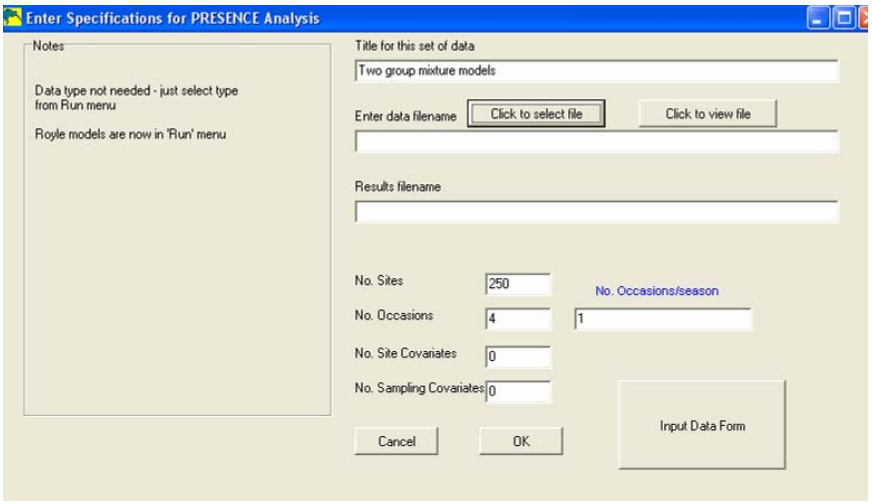
OBJECTIVES

- To become familiar with PROGRAM PRESENCE
- To run the two group, single-season occupancy model in PRESENCE
- To understand the PRESENCE mixture model output.
- To review concepts of model selection.

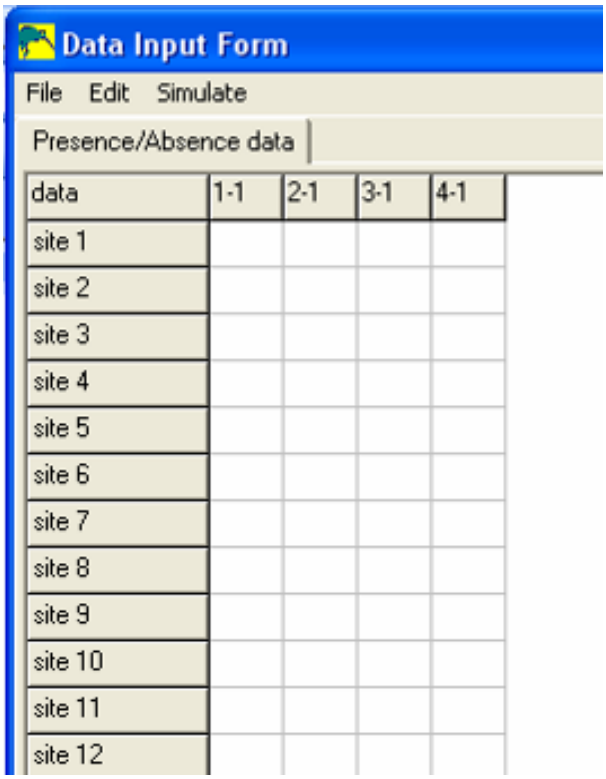
GETTING STARTED

In this exercise, we will be analyzing the data we explored in the previous (spreadsheet) chapter. When you open PRESENCE, you'll see the following screen:

Choose File | New Project, and you'll see the following screen:

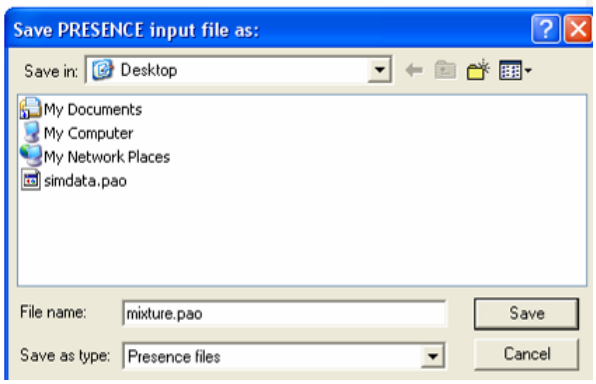


Enter a title for the analysis (e.g., "Two group mixture models"). Then enter 250 for No. Sites and 4 for No. Occasions. Then click on the button labeled "Input Data Form":

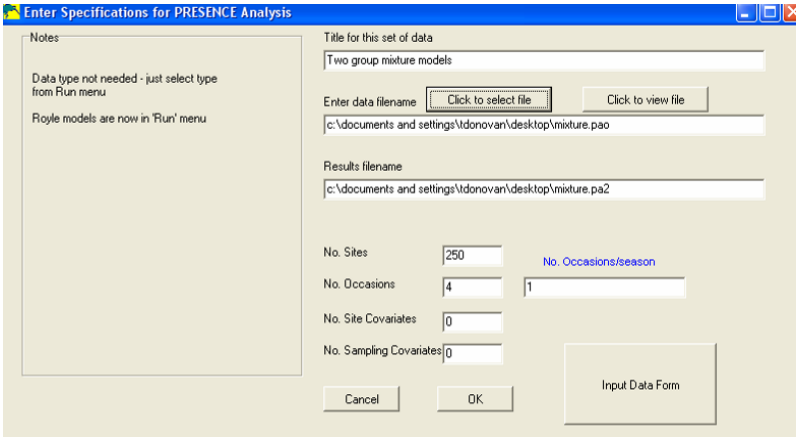


Now you are ready to paste in your spreadsheet data. Return to the spreadsheet and copy cells S4:V253. Then return to the data entry page in PRESENCE, and

select Edit | Paste | Paste Values. You should see data for all 250 sites. Next, SAVE this file. Click on File | Save As, and then type in a file name and location to store your input file:



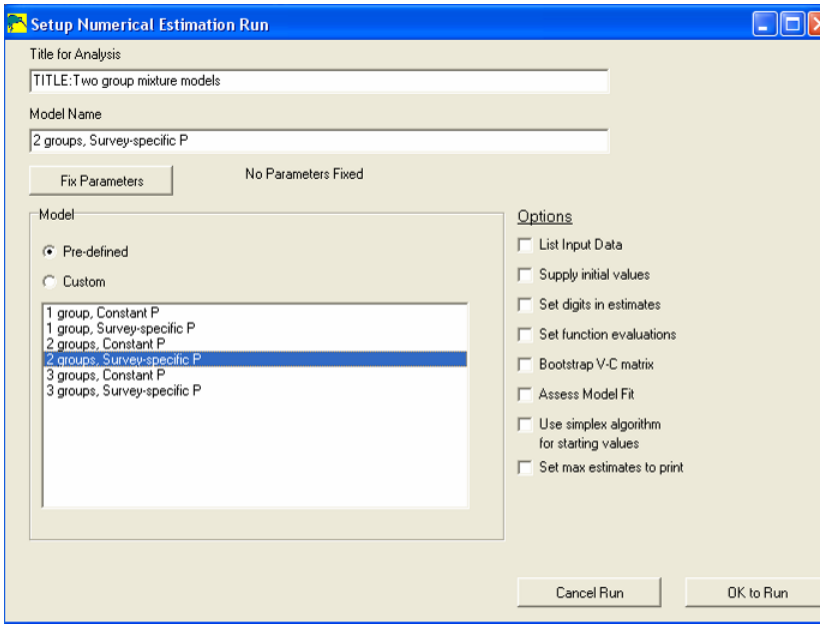
We saved our file as "mixture.pao" and put it on our desktop. Click Save and then close the PRESENCE datasheet. Now, back on the PRESENCE Enter Specifications page, click on the button labeled "Click to Select File" and navigate to your freshly saved input file. Your screen should look something like this:



Click OK, and you're ready to run some models. You'll quickly see that you can run all four models in no time.

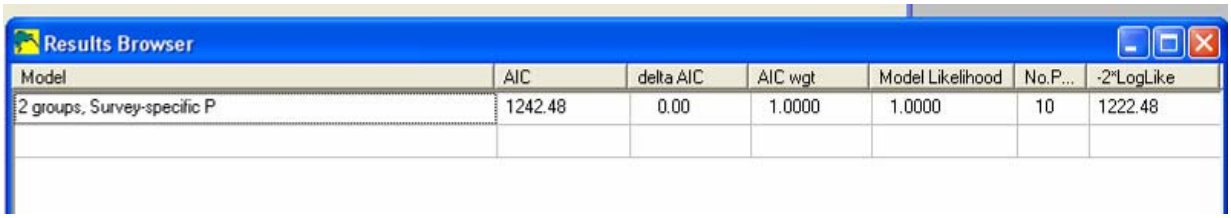
MODEL P(t)PSI FOR TWO GROUPS

Our first model is the two-group mixture model where t is uniquely estimated for each group. To run this model in PRESENCE, choose Run | Analysis from the main toolbar, and then select the pre-defined model named 2 groups, Survey-specific P.



Make sure the "Set digits in estimates" in NOT checked. Click "OK to Run".

Wasn't that easy? Add your results to the Results Browser:



Now, right-click on the model name to pull up the model results:

```
Predefined Model: Detection probabilities are time-specific

Number of groups               = 2
Number of sites                = 250
Number of sampling occasions   = 4
Number of missing observations = 0

Number of parameters           = 10
**** Numerical convergence may not have been reached.
**** Parameter estimates converged to approximately 2.735163
**** significant digits.
-2log(likelihood)              = 1222.484711
AIC                            = 1242.484711
Naive estimate                 = 0.696000

Proportion of sites occupied    (Psi)   = 0.7878 (0.065269)
Probability of group membership (Theta) = 0.3300, 0.6700
Detection probabilities          (p):
  grp   srvy      p           se(p)
  ---   ----   ---------    -----------
   1      1    0.839888     ( 0.277299)
   1      2    0.734797     ( 0.244678)
   1      3    0.618815     ( 0.208769)
   1      4    0.524175     ( 0.192945)

   2      1    0.435101     ( 0.226657)
   2      2    0.380766     ( 0.200851)
   2      3    0.331794     ( 0.165585)
   2      4    0.272310     ( 0.142224)
```

PRESENCE indicates that this is a two-group, predefined model with 250 study sites, 4 sampling occasions, 0 missing observations, and 10 parameters. Note the starred sentence that reads, "Numerical convergence may not have been reached. Parameter estimates converged to approximately 2.735163 significant digits." This is an important warning. It means that the optimization program may not have found the peak of the likelihood surface because the log likelihood value was still changing after the specified number of iterations. You have to decide whether 2 significant digits are enough to satisfy you. (We're fine with two significant digits, but get in the habit of reading this warning and not blindly skipping over it).

Note: If you want to improve the precision of the estimates even more, delete the model you just ran from the Results Browser (right-click on the model name and select "delete"), and re-run the model. But this time, click on the option labeled

"Set function evaluations" on the run page and increase the number of functions that will be evaluated.

OK, now let's compare these results to the spreadsheet:

	D	G
3	Parameter	MLE
4	π	0.66997
5	**Mixture 1**	
6	ψ	0.78778
7	p1	0.43510
8	p2	0.38076
9	p3	0.33179
10	p4	0.27231
11	**Mixture 2**	
12	ψ	0.78778
13	p1	0.83988
14	p2	0.73479
15	p3	0.61881
16	p4	0.52417

```
Predefined Model: Detection probabilities are time-specific

Number of groups               = 2
Number of sites                = 250
Number of sampling occasions   = 4
Number of missing observations = 0

Number of parameters           = 10
**** Numerical convergence may not have been reached.
**** Parameter estimates converged to approximately 2.735163
**** significant digits.
-2log(likelihood)              = 1222.484711
AIC                            = 1242.484711
Naive estimate                 = 0.696000

Proportion of sites occupied    (Psi)   = 0.7878 (0.065269)
Probability of group membership (Theta) = 0.3300, 0.6700
Detection probabilities          (p):
  grp   srvy      p          se(p)
  ---   ----   ---------   -----------
   1      1    0.839888     ( 0.277299)
   1      2    0.734797     ( 0.244678)
   1      3    0.618815     ( 0.208769)
   1      4    0.524175     ( 0.192945)

   2      1    0.435101     ( 0.226657)
   2      2    0.380766     ( 0.200851)
   2      3    0.331794     ( 0.165585)
   2      4    0.272310     ( 0.142224)
```

Although we haven't shown it, the -2Log$_e$L, AIC and naïve estimates match, and you can see above that the MLE's match as well. Note that PRESENCE estimated π (which is labeled theta) as 0.3300 for group 1, and 1-π = 0.6700 for group 2. The spreadsheet estimated π as 0.6700. It doesn't matter which group is which, as long as you match the correct, corresponding MLE's with each group.

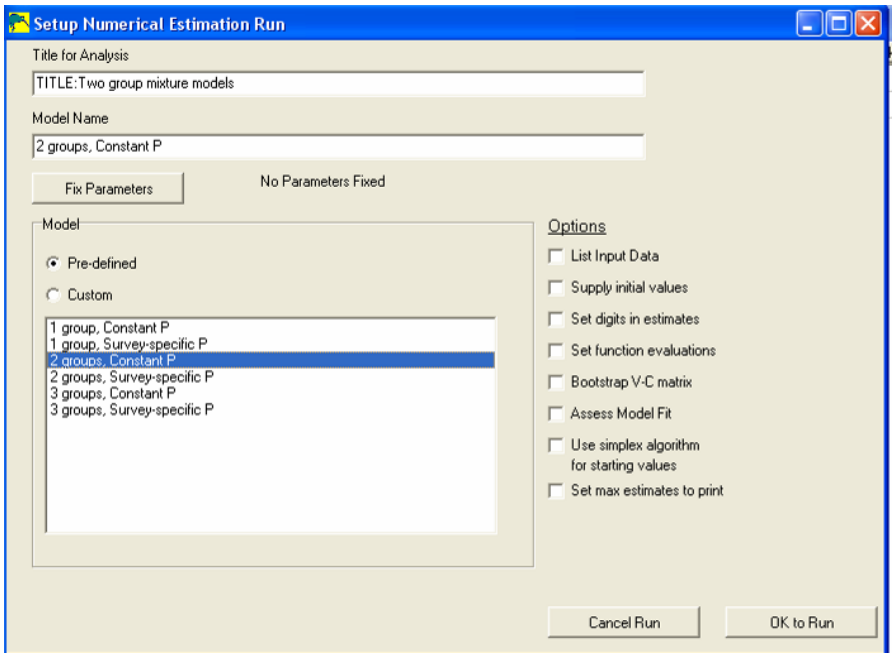
You might also recall from the spreadsheet exercise that this model had a very, very low Deviance score – the model fits the observed field data almost perfectly. But at what price? Take a look now at the standard errors associated with each parameter……they are VERY large for all of the detection parameters. For instance, $p_{1,1}$ (in PRESENCE) is estimated as 0.8399, with a standard error of 0.277. This corresponds to a lower 95% confidence interval of 0.8399 - 1.96 *

0.277= 0.2969 and an upper 95% confidence interval of 0.8399 + 1.96 * 0.277= 1.383. We wouldn't put much confidence in the 0.8399 estimate.

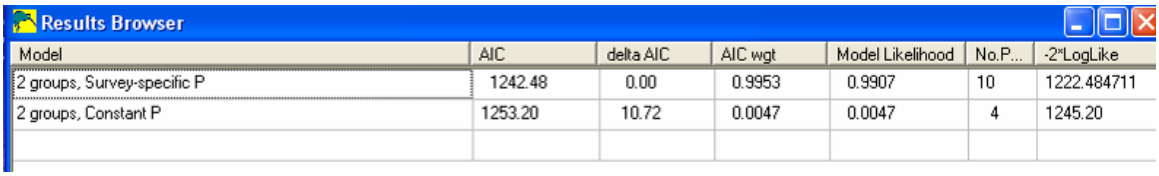
The take-home message is these models really need a lot of sites in order to work reasonably well. Even with the large differences in the p's in this example, the standard errors are pretty big. Would this model be selected as the best model in a model selection exercise? Probably not (as you've seen from the spreadsheet runs), because 10 parameters were estimated to make the model more "realistic". Furthermore, each of the 10 parameters has poor precision (high standard errors). In other words, adding complications to the model to make it more realistic comes at a price - not being able to estimate those parameters as well (i.e. high variances). Let's run the remaining models and then study the Results Browser in detail.

MODEL P(.)PSI FOR TWO GROUPS

OK, our next model is the two-group constant p model, which is another pre-defined model in PRESENCE. (In fact, the mixture models are all pre-defined – you can try to customize the mixture model in the Design Matrix and will quickly see that you can't run anything but the pre-defined models). Go to Run | Analysis – Single-Season and select the 2-group, Constant-P model.



Click OK to Run, and add the results to the Results Browser:



Model	AIC	delta AIC	AIC wgt	Model Likelihood	No.P...	-2*LogLike
2 groups, Survey-specific P	1242.48	0.00	0.9953	0.9907	10	1222.484711
2 groups, Constant P	1253.20	10.72	0.0047	0.0047	4	1245.20

The results of this model match the spreadsheet results, and here are the estimates from PRESENCE (as shown below). Remember, this is the model where we had to "seed" the beta values with random numbers to get Solver to find the appropriate solution. JIM, HOW WOULD ONE KNOW IF THIS IS A PROBLEM IN PRESENCE? ARE THERE ANY OPTIONS IN PRESENCE (LIKE MARK) WHERE YOU CAN RUN AN ALTERNATIVE OPTIMIZATION THAT ENSURES THAT YOU HAVEN'T FOUND A LOCAL MAXIMUM?

pres4203.tmp - Notepad

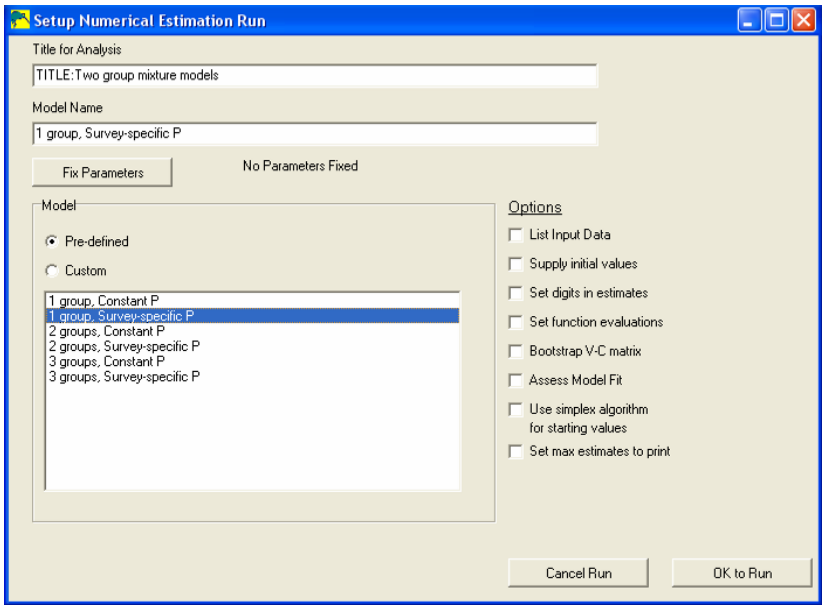
File Edit Format View Help

```
=======(2 groups, Constant P)=======


PRESENCE - Presence/Absence-Site Occupancy data analysis
Sat Dec 16 07:46:17 2006,       Version 2.051114
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
model=12c N,T-->250,4
modtype-->1 Single-Season data Model selected
NSi-->0
NSa-->0
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
 Two group mixture models
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
modtype=1 N=250 T=4 Groups=2 bootstraps=0

==>0


Predefined Model: Detection probabilities are NOT time-specific

Number of groups               = 2
Number of sites                = 250
Number of sampling occasions   = 4
Number of missing observations = 0

Number of parameters           = 4
**** Numerical convergence may not have been reached.
**** Parameter estimates converged to approximately 5.109557
**** significant digits.
-2log(likelihood)              = 1245.195586
AIC                            = 1253.195586
Naive estimate                 = 0.696000

Proportion of sites occupied    (Psi)   = 0.7908 (0.073372)
Probability of group membership (Theta) = 0.3810, 0.6190
Detection probabilities         (p):
  grp   srvy      p            se(p)
  ---   ----   ---------    ----------
   1      1    0.647858     ( 0.226284)

   2      1    0.344800     ( 0.227289)

Variance-Covariance Matrix
   psi  Group 2   p(G1)  p(G2)
 0.0054 -0.0355 -0.0122 -0.0143
-0.0355  0.3884  0.1385  0.1377
-0.0122  0.1385  0.0512  0.0482
-0.0143  0.1377  0.0482  0.0517
-----------------------------------------------------------
CPU time: 0.0 seconds
```

Note that, although this model estimated only 4 parameters, each of the detection probability estimates still have large standard errors. In this case, the large standard errors for the detection parameters are not due to a large K (number of model parameters). Rather, we know that the data were simulated with time-specific p's, so the standard errors reflect the fact there is quite a bit of variance in p within each group.

MODEL P(t)PSI (ONE GROUP)

All right; two down and two to go. The last two models are the standard, ψ p(t) and ψ p(.) models, where the number of groups is one. Running these should be second nature by now:



Run this model and add the results to the Results Browser. As you can see, this model is our top model, and the results match the spreadsheet.

Model	AIC	delta AIC	AIC wgt	Model Likelihood	No.P...	-2*LogLike
1 group, Survey-specific P	1237.93	0.00	0.9064	0.8215	5	1227.93
2 groups, Survey-specific P	1242.48	4.55	0.0932	0.0845	10	1222.484711
2 groups, Constant P	1253.20	15.27	0.0004	0.0004	4	1245.20

Knowing that the data were simulated with two-groups, and time specific p's (the first model you ran), let's now look at estimates from the same model with only 1 group:

```
Proportion of sites occupied    (Psi)   = 0.7878 (0.065269)
Probability of group membership (Theta) = 0.3300, 0.6700
Detection probabilities         (p):
  grp   srvy      p           se(p)
  ---   ----   ---------    ----------
   1      1    0.839888     ( 0.277299)
   1      2    0.734797     ( 0.244678)
   1      3    0.618815     ( 0.208769)
   1      4    0.524175     ( 0.192945)

   2      1    0.435101     ( 0.226657)
   2      2    0.380766     ( 0.200851)
   2      3    0.331794     ( 0.165585)
   2      4    0.272310     ( 0.142224)
```

You can see that we still have the same problem of high standard errors, even though this model estimated only 5 parameters. Why are the standard errors high in this model? Well, the data were simulated with two-groups and survey-specific p's, where p's decline from $p_1 \rightarrow p_4$ and group 2's p's are $\frac{1}{2}$ that of group 1's:

	X	Y	Z	AA	AB	AC	AD	AE
3	Parameter	**p1**	**p2**	**p3**	**p4**	**ψ**	**N**	**π**
4	MLE Group 1	0.8	0.7	0.6	0.5	0.8	100	0.4
5	MLE Group 2	0.4	0.35	0.3	0.25	0.8	150	0.6
6						Total Sites:	250	

The one-group, survey-specific model forces p_1 to be the same for both groups, p_2 to be the same for both groups, etc. Because there is a lot of variation in the data, the standard errors are high, indicating that the precision of the MLE's is low.

MODEL P(.)PSI (ONE GROUP)

Our last model is the standard, one group model ψ p(.). Go ahead and run this model, and add the results to the Results Browser:

Results Browser

Model	AIC	delta AIC	AIC wgt	Model Likelihood	No.P...	-2*LogLike
1 group, Survey-specific P	1237.93	0.00	0.9060	0.8208	5	1227.93
2 groups, Survey-specific P	1242.48	4.55	0.0931	0.0844	10	1222.484711
1 group, Constant P	1253.13	15.20	0.0005	0.0004	2	1249.13
2 groups, Constant P	1253.20	15.27	0.0004	0.0004	4	1245.20

Here are the spreadsheet results for comparison:

	C	D	G	H	I	J
29		Model	K	AIC	AICc	Rank
30	1	ψ p(t)2groups	10	1242.498	1243.4184	2
31	**2**	ψ p(.)2groups	4	1253.196	1253.3589	4
32	**3**	ψ p(t)	5	1237.926	1238.1716	1
33	**4**	ψ p(.)	2	1253.126	1253.175	3

The bottom line is that these mixture models really need a lot of data to perform well in a model-selection exercise. The number of parameters that are estimated can get very large, very quickly, resulting perhaps in a better-fitting model (lower -2Log_eL's) but at a cost: the standard errors can be quite large, with low precision.

That wraps up this exercise. Hopefully you have a good idea of what mixture models are all about and can use them wisely in your own work.