EXERCISE 16: MULTIPLE OCCUPANCY STATES 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-SEASON OCCUPANCY MODELS WITH MULTIPLE OCCUPANCY STATES

OBJECTIVES:

- To learn and understand the single-season occupancy model that includes multiple states of occupancy, and how it fits into a multinomial maximum likelihood analysis.
- To use Solver to find the maximum likelihood estimates for the probability of occupancy in state 1 and state 2, 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 data with multiple states.

BASIC INFORMATION

If you've been completing the exercises in this book in order, you've learned a great deal about the single-season occupancy modeling, and some interesting variations of the basic model. In this exercise, we describe another spin-off of the single-season model, in which "occupancy" can include more than 1 state. This model is described in section 10.1 of the book, Occupancy Estimation and Modeling. Click on the worksheet labeled "Single Season Multi-State" and we'll get started.

BACKGROUND

Hopefully by now you have a solid understanding that the general occupancy model assesses occupancy of a site, such that any encounter history with a "1" in it indicates (e.g., "001") that the site was occupied by the target species. In the single season model, occupancy is dichotomous (the site is occupied or not).

Often, however, an investigator can separate "occupied" into two or more categories, which leads to a multi-state occupancy model. Here's a quick example. In breeding bird atlases (such as the Vermont Breeding Bird Atlas), field observers not only record which species occur in a census block, but further break down the categories into "present," "probable breeder," and "confirmed breeder." In the Occupancy Estimation and Modeling book, the authors describe a scenario where individual ponds (sites) are monitored for evidence of breeding by amphibians, which might include finding egg masses, new metamorphs, or observing animals in amplexus. In this case, the site is classified into one of two states: it contains breeders or it contains non-breeders. In some situations the investigator might classify a site into one of three (or even more) categories. Why might this kind of information be useful? Well, there are many reasons, the simplest of which is that sites with breeding populations are more likely to be of high value than site filled with only non-breeders from a metapopulation perspective.

The multi-state occupancy model is a straight-forward extension of the single-season occupancy model, and also incorporates some elements of the error model that we discussed in previous exercises. In this exercise, we

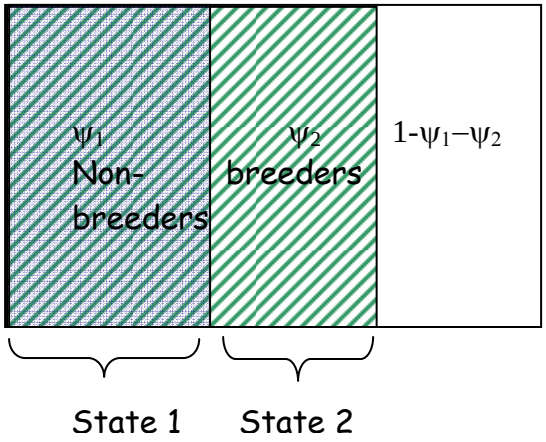
will consider only two states of occupancy: the site is occupied by non-breeders (state 1) versus the site is occupied by breeders (state 2). An important and fundamental concept to these models is that one of the states is observed without error. In our example, we assume that that the breeding state (state 2) is known with certainty (e.g., if you see frogs in amplexus, you know breeding is at least being attempted, if you see egg masses, you know breeding has occurred, and so on.) Thus, it is not possible that observers record the species as being breeders when in fact they are non-breeders. The non-breeding state, in contrast, can include some uncertainty in assignment. For instance, we might hear a singing amphibian in a pond but find no evidence of breeding; it doesn't mean that the species was not breeding; it means that the species was detected but there was not sufficient evidence to indicate breeding status. So errors can be made in assigning the species as non-breeding when it should have been assigned as breeding, but the reverse cannot happen.

MULTI-STATE OCCUPANCY PARAMETERS

So, how can the single-season occupancy model be extended to account for multiple occupancy states? To begin, we'll define our two states again for clarity: state 1 will be the "non-breeding" state in which the species was detected but no evidence of breeding was observed, while state 2 will be the "certain state" in which the species was detected and breeding was confirmed. Now, let's define the occupancy parameters associated with each state:

- ψ_1 is the probability that the site is occupied by individuals in state 1 (non-breeders).
- ψ_2 is the probability that the site is occupied by individuals in state 2 (breeders).
- And of course, $(1-\psi_1 - \psi_2)$ is the probability that the site is not occupied.

A simple diagram will help differentiate between ψ, ψ_2, and $1-\psi_1-\psi_2$. Note the total area of the box must be 1.



It's critical to note that the breeding state is assumed to be constant across all animals at a site throughout the entire study period.

We also need our standard detection probabilities, p, associated with each survey. Note that detection probabilities, p, are state-specific (as you might guess because breeders and non-breeders may have very different probabilities of detection). Thus, with a three survey study, there are 6 p

estimates, denoted by a subscript in which the first number indicates the survey period and the second number indicates the state:

- $P_{1,1}$ = the probability the species is detected in survey 1, given it is in state 1 (non-breeding)
- $P_{2,1}$ = the probability the species is detected in survey 2, given it is in state 1 (non-breeding)
- $P_{3,1}$ = the probability the species is detected in survey 3, given it is in state 1 (non-breeding)
- $P_{1,2}$ = the probability the species is detected in survey 1, given it is in state 2 (breeding)
- $P_{2,2}$ = the probability the species is detected in survey 2, given it is in state 2 (breeding)
- $P_{3,2}$ = the probability the species is detected in survey 3, given it is in state 2 (breeding)

OK. Now, given that the species is detected, we also need to determine whether errors in assignment were made. The figure below will help us sort through the possibilities. In this table, the true state is given in cells H39:I39, while the field observations are given in cells G40:G41. In cell H40, the true state of the site was state 1 (non-breeding), and the field data indicated this, yielding a correct classification. In cell H41, the true state of the site is state 1 (non-breeding), but the field data indicated the state was 2 (breeding). We previously indicated that this error is not possible: if you observe state 2, then state 2 it is. In cell I41, the true state of the site is state 2, and the field data indicate such, yielding a

correct classification. We'll let the parameter delta (δ) be the probability of correct assignment into state 2. Delta is formally defined as the probability that the site is identified as state 2, given it is in state 2 (breeders). And finally, in cell I40, the true state of the site is state 2, but field data indicated it was state 1, yielding an error of (1-δ).

	F	G	H	I
38			**Truth**	
39			*State 1*	*State 2*
40	**Data**	*1*	correct	incorrect: (1-δ)
41		*2*	not possible	correct: δ

So the delta parameters (δ and 1-δ) will come into play when discussing state 2 only. Delta is survey specific, so if you conduct a study in which there are three surveys per site, you'll estimate δ_1, δ_2, and δ_3 (or you can constrain them to be equal). Make sense? Study this table now because it will come into play when we go through the encounter history probabilities in the next section.

MULTI-STATE ENCOUNTER HISTORIES

The easiest way to learn how the multi-state model works is to dive right into the encounter histories. Now, you've probably guessed that instead of entering 0 vs 1 data, you now have three data options: 0 (species was undetected), 1 (the species was detected and classified as non-breeding), or 2 (the species was detected with breeding status confirmed). With three options per survey, this quickly leads to a large number of possible encounter histories. If there are two surveys conducted per site, the number of

possible encounter histories is $3^2 = 9$. If three surveys are conducted per site, there are $3^3 = 27$ possible histories. And if there are 4 surveys per site, the number of possible encounter histories is $3^4 = 81$ possible encounter histories. In our spreadsheet exercise, we'll consider a study in which 3 surveys are done in a single season, resulting in 27 histories (cells D4:D30). The number of sites with each history is given in cells E4:E30, for a total of 250 sites (cell E31). The naïve estimate of occupancy is the sum of all site frequencies except the "000" history, divided by the total number of sites.

	D	E
3	History	Frequency
4	111	10
5	112	1
6	110	10
7	121	1
8	122	5
9	120	3
10	101	10
11	102	3
12	100	11
13	211	1
14	212	5
15	210	3
16	221	5
17	222	22
18	220	12
19	201	3
20	202	12
21	200	6
22	011	10
23	012	3
24	010	11
25	021	3
26	022	12
27	020	6
28	001	11
29	002	6
30	000	63
31	# Sites =	250
32	# Histories =	27
33	Naive Estimate =	0.749

Just so we're all on the same page, a history of 111 indicates that the species was observed in all three sampling periods, but no evidence of breeding was found on any occasion. A history of 121 indicates that the species was observed in all three sampling periods, and on occasion 2 evidence of breeding was detected but on occasions 1 and 3 no evidence of breeding was observed. Because we know that mistakes can't be made for observations of breeding, this history indicates that assignment errors were made in periods 1 and 3. A history of 002 indicates the species was detected in the breeding state on the third survey, but was totally missed in surveys 1 and 2. No

assignment errors were made in surveys 1 or 2 because the species was not detected in those surveys. A history of 000 still means that the species was not detected at the site.

MULTI-STATE ENCOUNTER HISTORY PROBABILITIES

Now, let's go through some examples of how the encounter history probabilities are determined. Let's start with an easy history: 222. In this case, the species was detected in all three surveys, and on each occasion evidence of breeding was recorded. So, we know this site is occupied by breeders, and we also know breeders were detected on every survey. So the history is $\psi_2 * p_{1,2} * \delta_1 * p_{2,2} * \delta_2 * p_{3,2} * \delta_3$. In other words, the site was occupied in state 2 (ψ_2), the species was detected on survey 1 and correctly classified as state 2 ($p_{1,2} * \delta_1$), the species was detected on survey 2 and was correctly classified as state 2 ($p_{2,2} * \delta_2$), and the species was detected on survey 3 and correctly classified as state 2 ($p_{3,2} * \delta_3$).

What about history 102? Well, again, we know that the site was occupied by breeders because they were detected on the third survey. So we will use the parameters associated with state 2 only. The species was detected in survey 1, but a mistake was made in state assignment. Then, the species was not detected at all in the second survey. Finally, it was detected and correctly classified into state 2 in the third survey. So the history probability is:

$\psi_2 * (p_{1,2}) * (1-\delta_1) * (1-p_{2,2}) * p_{3,2} * \delta_3$. In other words, the site was occupied by breeders (ψ_2), the species was detected on survey 1 but was not correctly

classified as breeding $(p_{1,2})*(1-\delta_1)$, the species was not detected on survey 2 $(1-p_{2,2})$, but was detected on survey 3 with probability δ_3 of being correctly identified as breeding $(p_{3,2}*\delta_3)$.

These can be brain-teasers, so let's go through two more, starting with 111. In the multi-state model, a 111 history indicates that an observer went into the field and documented that the species was present in all three surveys, but on no survey was evidence of breeding detected. Because we know that non-breeding observations can be made with error, we have two alternative possibilities to consider.

In the first option, the species was truly in state 1 (non-breeding), which is $\psi 1*p_{1,1}*p_{2,1}*p_{3,1}$.
In the second option, the species could really be in state 2 (breeding), in which three errors in assignment were made (the history really should have been "222"). In this case, the probability is which is $\psi_2*p_{1,2}*(1-\delta_1)*$ $p_{2,2}\,(1-\delta_2)*p_{3,2*}(1-\delta_3)$.
Because both options are possible, we add the two probabilities together so that the probability of observing this history is:
$\psi 1*p_{1,1}*p_{2,1}*p_{3,1} + \psi_2*p_{1,2}*(1-\delta_1)*\ p_{2,2}\,(1-\delta_2)*p_{3,2*}(1-\delta_3)$.

OK, the last history we'll review is 000. In this case, there are three options. First, the site could have been unoccupied, which is $1-\psi_1-\psi_2$. OR, the site could have been occupied by non-breeders, but missed on all three surveys, which is $\psi_1*(1-p_{1,1})*(1-p_{2,1})*(1-p_{3,1})$. OR, the site could have been occupied by breeders, and missed on all three surveys: $\psi_2*(1-p_{1,2})*(1-$

$p_{2,2}$)*(1-$p_{3,2}$). This history should hammer home the point that the delta parameters only come into play when detection occurs; δ_{ij} is the probability that the species was identified as breeders, given the patch is occupied by breeders and animals are detected.

MULTI-STATE MODEL SPREADSHEET INPUTS

OK, with that background, let's get oriented to the spreadsheet. In this example, the investigator surveys 250 study sites, with each site being surveyed 3 times. The encounter histories are recorded in cells D4:D30, and the frequency of each history is recorded in cells E4:E30. The total number of sites is given in cell E31, and the number of unique histories is given in cell E32 (which you might remember indicates the number of terms in our multinomial likelihood function). To avoid over-parameterization, you can only run models with 26 or fewer parameters. We definitely don't need to worry about overparameterization in this exercise. The naïve estimate for occupancy (occupancy unadjusted for detection probability) is computed in cell E33 as the total number of sites which had one or more detections divided by the total number of sites. In this case the estimate is around 75%.

OK, now let's look at the parameters. Notice the spreadsheet is divided into two sections.

	F	G	H	I
3	Parameter	Estimate?	Betas	MLE
4	State 1			
5	ψ_1	1		0.50000
6	p1,1	1		0.50000
7	p2,1	1		0.50000
8	p3,1	1		0.50000
9	State 2			
10	$\psi 2$	1		0.50000
11	p1,2	1		0.50000
12	$\delta 1$	1		0.50000
13	p2,2	1		0.50000
14	$\delta 2$	1		0.50000
15	p3,2	1		0.50000
16	$\delta 3$	1		0.50000

In the first section (cells F5:I8), we list the parameters associated with state 1, consisting of sites that are truly occupied by non-breeders ($\psi 1$, $p_{1,1}$, $p_{2,1}$, $p_{3,1}$). The second section of the spreadsheet (cells F10:I16) lists the parameters for state 2, which consists of sites that are truly occupied by breeders (ψ_2, $p_{1,2}$, δ_1, $p_{2,2}$, δ_2, $p_{3,2}$, δ_3). 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. This set-up is very similar to the mixture models we covered in previous exercises.

MULTI-STATE LINKS

The betas for each parameter are listed in cells H5:H8, H10:H16, and the MLE parameter estimates that correspond to each beta are computed in cells I5:I8, I10:I16 through a logit link. Click on cell I5 and you'll see the logit link transformation: =EXP(H5)/(1+EXP(H5)). Remember the logit link constrains the MLE's to be between 0 and 1, which is what we want for ψ, and the p_i's, and the δ_i's. Because this exercise doesn't include covariates,

we also could have used the sin link, but we stuck with the logit in case we decide to add covariates some day.

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).

SPREADSHEET HISTORY PROBABILITIES

OK! Now we are ready to compute the probability of realizing each history. Let's start with the first history listed, 111, in cell J4.

As we indicated previously, the probability of realizing a 111 history considers two options: the probability of realizing a 111 history if the site is truly in state 1, plus the probability of realizing a 111 history if the site is truly in state 2. If sites are truly in state 1, the probability of realizing a 111 history is $\psi_1 * p_{1,1} * p_{2,1} * p_{3,1}$. If sites are truly in state 2, the probability of realizing a 111 history is $\psi_2 * p_{1,2} *(1-\delta_1) * p_{2,2}*(1-\delta_2) * p_{3,2} * (1-\delta_3)$. Across both states, the probability of realizing a 111 history is the sum of the two probabilities, and is entered in cell H4: =I5*I6*I7*I8+I10*I11*(1-I12)*I13*(1-I14)*I15*(1-I16). The natural log of the combined history probabilities is computed in cell K4. And so it goes for the remaining histories.

Make sense? Spend time now clicking on the formula for each history and for each group. In our experience, if students understand how the encounter histories are calculated, the rest is a piece of cake.

Notice that the sum of cells J4:J30 must equal 1 (cell J31): there are 27 possible histories, and each history has a probability of being realized, but the sum of the probabilities must be 1.00.

THE MULT-STATE 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.

$$\ln(L(p_i \mid n_i, y_i) \propto y_1 \ln(p_1) + y_2 \ln(p_2) + y_3 \ln(p_3) + ... + y_{27} \ln(p_{27})$$

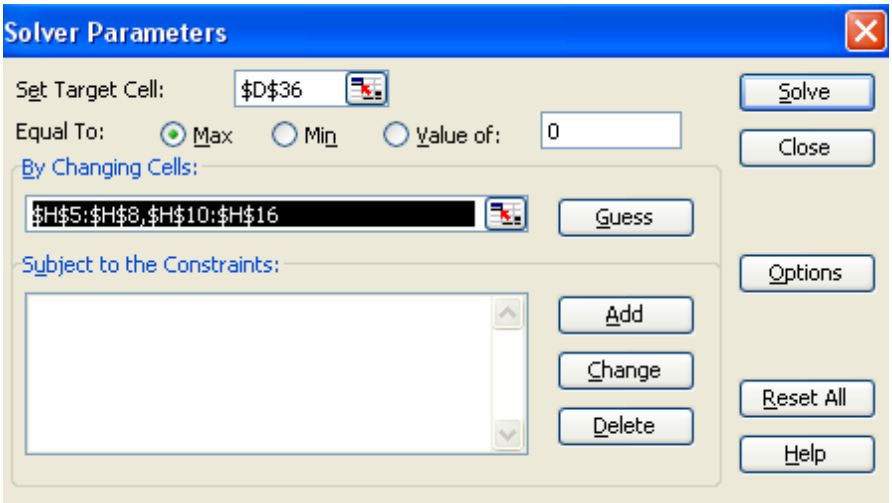
There are 27 terms in this function, one for each of the encounter histories. The y_i in the blue box are the frequencies of each kind of history and the p_i in the blue box equation above are the history probabilities. The Log_eL is computed in cell D36 with the equation =SUMPRODUCT(E4:E30,K4:K30), 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, make sure the betas are cleared out and that your spreadsheet is set up as follows:

	F	G	H	I
3	Parameter	Estimate?	Betas	MLE
4	State 1			
5	ψ_1	1		0.50000
6	p1,1	1		0.50000
7	p2,1	1		0.50000
8	p3,1	1		0.50000
9	State 2			
10	$\psi 2$	1		0.50000
11	p1,2	1		0.50000
12	$\delta 1$	1		0.50000
13	p2,2	1		0.50000
14	$\delta 2$	1		0.50000
15	p3,2	1		0.50000
16	$\delta 3$	1		0.50000

OK, now we're ready to run this model. We can call this model "$\psi_{1,2}$, p(1t,2t)δ(t)" to indicate that we're estimating ψ_1, $p_{1,1}$, $p_{2,1}$, and $p_{3,1}$ for state 1, and ψ_2, $p_{1,2}$, $p_{2,2}$, $p_{3,2}$ and δ_1, δ_2, δ_3 for state 2. You know the drill. Open Solver, and set cell D36 to a maximum by changing cells H5:H8,H10:H16.



Press Solve and Solver will work through the various combinations of betas until it finds the maximum. You probably should also add the constraint that $\psi_1 + \psi_2 <= 1$ (although Solver finds the correct estimates without this constraint).

MULTI-STATE MODEL OUTPUT

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

	F	G	H	I
3	Parameter	Estimate?	Betas	MLE
4	State 1			
5	ψ_1	1	-0.847290928	0.30000
6	p1,1	1	3.50942E-05	0.50001
7	p2,1	1	-2.55984E-05	0.49999
8	p3,1	1	-5.59954E-06	0.50000
9	State 2			
10	$\psi 2$	1	-4.01545E-06	0.50000
11	p1,2	1	0.847295017	0.70000
12	$\delta 1$	1	1.386368846	0.80001
13	p2,2	1	0.84731947	0.70000
14	$\delta 2$	1	1.386232122	0.79999
15	p3,2	1	0.847319494	0.70000
16	$\delta 3$	1	1.38628594	0.80000

The proportion of sites that were truly occupied by non-breeders (state 1) is 0.300 (cell I5). Non-breeders had a 50% probability of being detected, given they were present, across all three survey periods. The proportion of sites that were occupied by breeders (state 2) is 0.5 (cell I10). In all survey periods, breeders had about a 70% probability of being detected. Additionally, in all survey periods the probability of correctly classifying sites to state 2 was 80%. Even though we ran a model where the p's and d's were time specific, Solver found estimates that indicated a simpler model

with fewer parameters (where p's and δ's are constrained to be constant across periods) would be a better fit in terms of AIC.

And, in fact, this will be the case because the data you just analyzed were simulated by expectation with the parameters indicated below:

	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1	Simulate Data											
2	State 1				State 2							N
3	ψ_1	$p_{1,1}$	$p_{2,1}$	$p_{3,1}$	ψ_2	$p_{1,2}$	δ_1	$p_{2,2}$	δ_2	$p_{3,2}$	δ_3	
4	0.3	0.5	0.5	0.5	0.5	0.7	0.8	0.7	0.8	0.7	0.8	250

We'll revisit this section of the spreadsheet in a few minutes.

One word of caution is in order at this point. Remember that $\psi_1 + \psi_2$ cannot be greater than 1. In our example, $\psi_1 = 0.3$ and $\psi_2 = 0.5$, so the proportion of sites that are occupied by either state 1 or state 2 is 0.3 + 0.5 = 0.8. You might want to add the constraint to Solver that the sum of $\psi_1 + \psi_2 <= 1$ within the Solver dialogue box if you want to play it safe.

Now let's look at the remaining output given in cells D35:N36.

	D	E	F	G	H	I	J	K	L	M	N
34	OUTPUTS										
35	Log_eL	$-2Log_eL$	K	AIC	AICc	$-2Log_eL$ Sat	Deviance	Model DF	C-hat	Chi-Square	P value
36	-711.84	1423.681	11	1445.68	1446.79	1423.6808	0.0000	16	9.44222E-09	0.0000	1.0000

The Log_eL is given in cell D36. Cell E36 is -2 times cell D26, and is the $-2Log_eL$. K is the number of parameters in any given model, and the underlying equation is =SUM(G5:G8,G10:G16). AIC is computed as the $-2Log_eL + 2*K$ (cell G36). AICc is the second order correction of AIC and uses the number of study sites in the calculation. Deviance (cell J36) 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$_e$L is computed in the usual way (as in previous exercises) in cells L4:M30. The model we just ran had a deviance of 0 because we analyzed data that were generated by expectation, and Solver found the true estimates. 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 L36 as Deviance divided by DF. C-hats larger than 1 might indicate some kind of lack of fit. The Chi-Square statistic and associated p-value are given in cells M36:N36. The Chi-square computations are provided in the orange cells N4:O30. The spreadsheet shows the Chi-Square test statistic is 0 (because the observed exactly equals the expected values, which won't happen in "real" situations), and the associated p value is 1.

That's really all there is to the multi-state model. We covered the simplest model in which there are two possible states, but you can extend this model to include more states. And, of course, covariates can be included into the multi-state model, which makes this model a very important model option for investigators who can assign occupancy into two or more states.

SIMULATING MULTI-STATE DATA

Before we finish, we want to demonstrate how the data were simulated for this exercise. We already mentioned that the data were simulated where ψ_1 = 0.3, $p_{1,1}$, $p_{2,1}$, $p_{3,1}$ = 0.5 for state 1, and ψ_2 = 0.5, $p_{1,2}$ $p_{2,2}$, $p_{3,2}$ = 0.7, and δ_1, δ_2, δ_3 = 0.8 for state 2. You can simulate any estimates you'd like. But adhere to the cautionary note that $\psi_1 + \psi_2$ must be less than 1. Start by entering

the total number of sites in cell AB4, and then enter the state-specific parameters in cells Q4: AA4.

	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1	**Simulate Data**											
2	State 1				State 2							N
3	ψ_1	$p_{1,1}$	$p_{2,1}$	$p_{3,1}$	ψ_2	$p_{1,2}$	δ_1	$p_{2,2}$	δ_2	$p_{3,2}$	δ_3	
4	0.3	0.5	0.5	0.5	0.5	0.7	0.8	0.7	0.8	0.7	0.8	250
5	Note: $\psi_1 + \psi_2$ must be <= 1.											

As with the other spreadsheet exercises, we will simulate data in two ways: by expectation and with stochasticity. The expected data are simulated in cells W10:Y36. The expected frequency of each history is computed in cells Y10:Y36 as N (cell AB4) times the encounter history probability for each history.

	W	X	Y
8	**Summarized Expected Data:**		
9	History	Probability	Frequency
10	111	0.038872	10
11	112	0.005	1
12	110	0.040	10
13	121	0.005	1
14	122	0.022	5
15	120	0.012	3
16	101	0.040	10
17	102	0.012	3
18	100	0.044	11
19	211	0.005	1
20	212	0.022	5
21	210	0.012	3
22	221	0.022	5
23	222	0.088	22
24	220	0.047	12
25	201	0.012	3
26	202	0.047	12
27	200	0.025	6
28	011	0.040	10
29	012	0.012	3
30	010	0.044	11
31	021	0.012	3
32	022	0.047	12
33	020	0.025	6
34	001	0.044	11
35	002	0.025	6
36	000	0.251	63
37			250

As we've indicated in a previous exercise, analyzing data created by expectation is one good way to evaluate your model, and also is useful for assessing model bias.

The stochastic data are created in a similar way, but we use random numbers to determine each site's encounter history. First, we number sites from 1 to

250 in row Q. Then, we establish whether each site was unoccupied, occupied in state 1, or occupied in state 2. In columns R and S, a random number is entered with the =RAND() function. A key feature here is that the random numbers for ψ_1 are set to equal those for ψ_2. In cell T41, we enter an equation that will return a 0, 1, or 2 to indicate the true state of the site. The formula is =IF(R41<$Q$4,1,IF(S41<$Q$4+$U$4,2,0)), which is a nested IF function. In the first IF function, if the random number is less than the value in cell Q4 (the proportion of sites in state 1), then a "1" is returned. Otherwise Excel moves to the second IF function; if the random number is less than the sum of Q4 + U4 (sites in state 1 or state 2), then a 2 is returned; otherwise a 0 is returned and the site is empty. The sum trick works for state 2 because the two random numbers for ψ_1 and ψ_2 are equal, and because the second IF function will be invoked only if the site is not in state 1. This formula is copied down for the remaining sites.

	Q	R	S	T
39		Random Numbers		
40	**Site**	**ψ1**	**ψ2**	**State**
41	1	0.879933778	0.879933778	0
42	2	0.589737925	0.589737925	2
43	3	0.177855579	0.177855579	1
44	4	0.994831009	0.994831009	0
45	5	0.815681201	0.815681201	0

Next, we establish random numbers for sites truly occupied in state 1 in columns U:X.

	U	V	W	X
39	State 1			
40	ψ_1	p1,1	p2,1	p3,1
41	0.18139682	0.1621778	0.742232	0.089501979
42	0.307345496	0.4651561	0.129779	0.529642555
43	0.336102258	0.1840292	0.573336	0.927838735
44	0.161919659	0.0753212	0.974657	0.241836756
45	0.932762679	0.4975329	0.911415	0.667932069

Similarly, we establish random numbers for sites truly occupied in state 2 in columns Y:AE. Note that more random numbers are associated with state 2 because we need to include the delta parameters:

	Y	Z	AA	AB	AC	AD	AE
39	State 2						
40	$\psi 2$	p1,2	$\delta 1$	p2,2	$\delta 2$	p3,2	$\delta 3$
41	0.57961	0.11587	0.17888	0.88925	0.95147	0.90977	0.71573
42	0.94496	0.00629	0.83228	0.4044	0.3152	0.55266	0.6467
43	0.49827	0.74667	0.71729	0.69595	0.79956	0.91726	0.38331
44	0.65238	0.74485	0.78679	0.61435	0.38046	0.12121	0.44161
45	0.18354	0.63761	0.32357	0.49627	0.64107	0.03555	0.22983

Finally, we enter equations to generate encounter histories for each state (0, 1, 2) separately.

	AF	AG	AH
39	Possible Histories		
40	0	1	2
41	000	101	200
42	000	110	122
43	000	100	020
44	000	101	022
45	000	100	222

If a site is in state 0, its encounter history will be "000" by default (no false positives), so this is entered for all sites in column AF. If a site is in state 1, the encounter history is generated with the formula =IF(V41<$R$4,1,0)&IF(W41<$S$4,1,0)&IF(X41<$T$4,1,0) for site 1. This is

three IF functions whose results are concatenated with the & sign. The first IF function returns a 1 if the random number associated with survey 1 (cell V41) is less than the $p_{1,1}$ given in cell R4. The second IF function returns a 1 if the random number associated with survey 2 (cell W41) is less than the $p_{2,1}$ given in cell S4. The third IF function returns a 1 if the random number associated with survey 3 (cell X41) is less than the $p_{3,1}$ given in cell T4. This formula is copied down for all sites.

For sites in state 2, the equation is trickier, but only a little. The equation is cell AH41 is

=IF(AND(Z41<$V$4,AA41<$W$4),2,IF(AND(Z41<$V$4,AA41>$W$4),1,0))&
IF(AND(AB41<$X$4,AC41<$Y$4),2,IF(AND(AB41<$X$4,AC41>$Y$4),1,0))&
IF(AND(AD41<$Z$4,AE41<$AA$4),2,IF(AND(AD41<$Z$4,AE41>$AA$4),1,
0)).

Yikes! But look carefully and you'll see that once again we generate an outcome for each survey, and then join the outcomes with an & sign. Remember that sites in state 2 can be unobserved (0), incorrectly observed in state 1 (1), or correctly observed in state 2 (2). For the first survey, the outcome is determined by the formula

=IF(AND(Z41<$V$4,AA41<$W$4),2,IF(AND(Z41<$V$4,AA41>$W$4),1,0)),

which is two nested IF functions with embedded AND functions. To begin, IF the random number associated with survey 1 (Z41) is less than $p_{1,2}$ AND the random delta 1 is less than δ_1 in cell W4, a 2 is returned. If either of these conditions is not true, Excel moves to the next IF function. IF the random number associated with survey 1 (Z41) is less than $p_{1,2}$ AND the

random delta 1 is greater than δ_1 in cell W4, a mistake was made and a 1 is returned. If either of these conditions is not true, Excel returns a 0 to indicate that the species was not detected on the survey. Hopefully this makes sense. The outcomes for survey 2 and survey 3 are done in a similar fashion. The formula is copied down for the other sites.

In cell AI41, the equation =HLOOKUP(T41,$AF$40:$AH$290,Q41+1) looks up the site's true state, and returns the appropriate encounter history. The summarized stochastic data are given in cells S10:T36.

	S	T
9	History	Frequency
10	111	2
11	112	4
12	110	0
13	121	3
14	122	9
15	120	9
16	101	1
17	102	8
18	100	3
19	211	4
20	212	14
21	210	3
22	221	10
23	222	49
24	220	21
25	201	6
26	202	24
27	200	14
28	011	1
29	012	6
30	010	0
31	021	3
32	022	22
33	020	13
34	001	3
35	002	9
36	000	9

PRESENCE INPUT FILES

	B
2	Tally
3	0
4	9.718
5	11.09
6	21.2
7	22.572
8	28.06
9	31
10	41.11
11	44.05
12	55
13	56.372
14	61.86
15	64.8
16	70.288
17	92.24
18	104
19	106.94
20	118.7
21	125
22	135.11
23	138.05
24	149
25	151.94
26	163.7
27	170
28	180.95
29	187.25
30	250

The multi-state model is not currently available in PRESENCE, but it will be in the near future. The histories and corresponding frequencies given in cells E4:E30 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 AJ:AN to convert the summarized data to site-specific data. But before we cover the equations, first look at cells B3:B30, 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 (111), the cell B4's formula counts the number of sites that are 111. The next cell (cell B5) counts the number of 111 sites + the 112 sites. The next cell (cell B6) counts the number of 111, 112, and 110 sites, and so on. We will use this running tally to create PRESENCE input files.

Now let's turn our attention to columns AJ:AN. In column AJ, the sites are listed from 1 to 250 down the column. In column AK, we assign a history to each site, using the tally in cells B3:B30. Click on cell AK3. The equation there is =LOOKUP(AJ3-1,$B$3:$B$30,$D$4:$D$30). The function looks up the value in AJ3 (the site number) minus 1 in the tally column ($B$3:$B$30), and then returns the corresponding history listed in cells $D$4:$D$30. 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. We've used this trick in previous exercises too.