Lab 5 - Introduction to RMark

Erik Blomberg (edited by Matt Mensinger & Liam Berigan)

07/27/2023

Contents

1 Estimating Demograrphic Parameters in the R package 'RMark' 2

1.1 Learning Objectives 2

1.2 First steps 3

1.3 Modelling categorical effects 4

1.4 The Cormack-Jolly-Seber (CJS) Model 10

1.5 White-throated dippers and riparian systems 12

2 First Exercise: Getting Started With RMark 13

3 Second Exercise – Running competing models to test hypotheses. 19

3.1 Building competing models 20

3.2 Custom time effects for biological hypotheses 23

3.3 Wrapping things up 26

4 The Three Big Things You Should Have Learned Today 27

5 Lab Assignment – Expanded Abstract 27

6 Quick Reference for Important R Commands in This Lab 29

7 For More Information 31

1 Estimating Demograrphic Parameters in the R package 'RMark'

In this lab we will start our deep-dive into the realm of mark-recapture estimation. As we saw during Lab 1 with the Lincoln-Peterson estimator, and during Lab 4 with binomial probabilities, some mark-recapture and survival estimation is relatively straightforward and can be accomplished by fairly simple calculations (i.e. things we can accomplish in Excel). But as we move into more complex models that rely on **multinomial probabilities**, the math underlying the models become increasingly difficult to complete 'by hand', and eventually it becomes computationally intractable altogether. At that point, we move into the realm of computing and the use of numerical optimization routines to solve the likelihoods for us. This is where specialized software comes in.

We'll be using the R package RMark, which in turn will rely on a called Program MARK for its computational abilities. You should recognize that these softwares, and many that came before them, have almost all been developed by biologists for application by biologists. Most are made available for free, including the ones we will use for the next few weeks. If you wish to complete this or future labs on your own computer, you will need to install program MARK in addition to loading the RMark Package. Getting RMark to work on a Mac is technically possible, but very difficult. If your personal computer is a Mac, I would highly recommend using one of the computers in the Nutting 254/235 computer labs to complete these assignments. Note that computers in these labs can be accessed remotely: see Brightspace for more details.

This lab expands on the concepts related to maximum likelihood and AIC that we covered during the last lab. We will use a maximum likelihood estimation technique to estimate the survival and detection probabilities that are most likely to have produced the mark-recapture data contained in the input file. It will also produce estimates of AICc from the model likelihoods, which we can use to compare among a series of differing model structures.

1.1 Learning Objectives

1) Introduce the concept of 'categorical effects' for modeling mark-recapture data and testing hypotheses about variation in demographic parameters such as survival.

2) Conduct a Cormack-Jolly-Seber live encounter analysis using the R Package ‘RMark’. This will both provide you with an introduction to the RMark package and expose you to deriving demographic estimates using maximum likelihood estimation.

3) Provide the opportunity to evaluate hypotheses using AIC, and interpret biological relevance using the real parameter estimates from the model output.

1.2 First steps

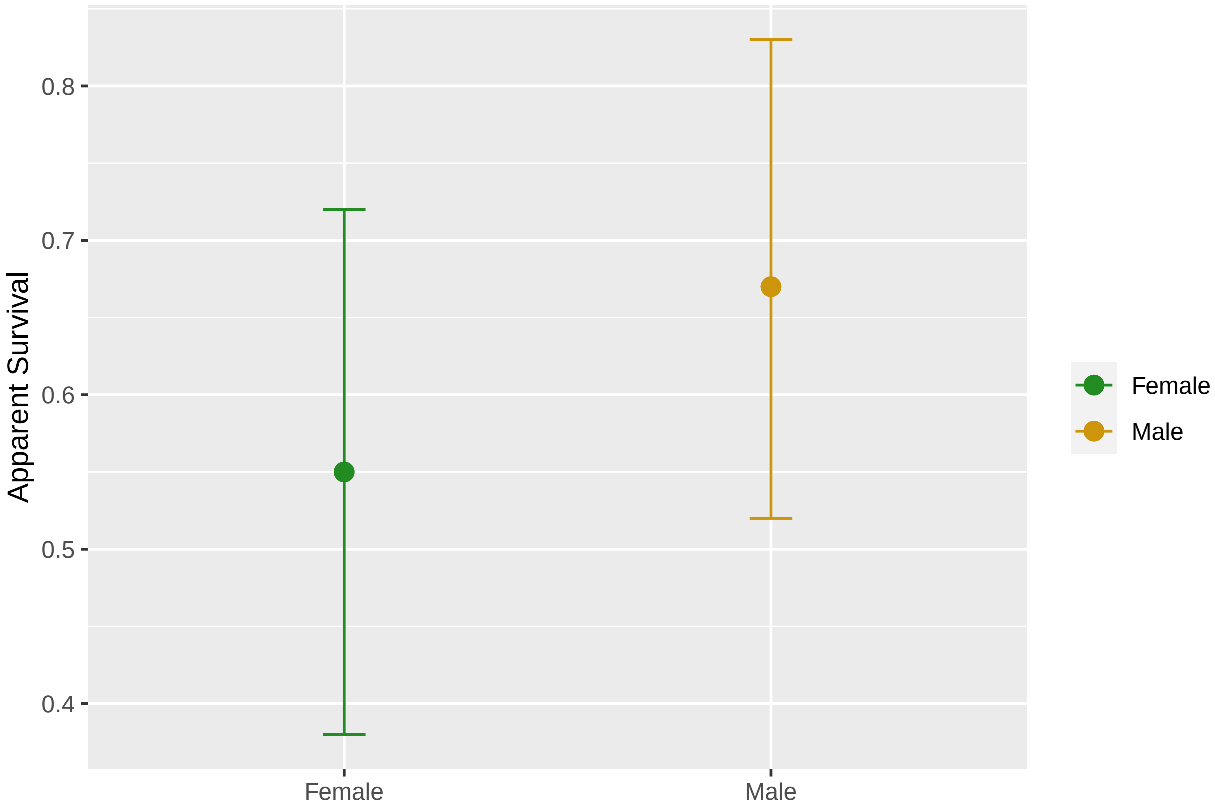
There is a lot of background material in this handout - probably more so than most weeks. But before you jump into reading through that material, I want to check and make sure your computing situation is set up to run RMark successfully. Do that by executing the following code in R. The goal here for now is to simply make sure your computer is set up for R and Mark to talk with each other. Don’t worry too much about what the code is doing, but when you run the Example object line, it should spit out a bunch of numbers in your console that ends with ‘Stop Normal Exit’ and without throwing an error. If you get anything else, or you’re unsure, ask for help.

```r
# install.packages('RMark')

library(RMark)

# this loads the data

data(dipper)

# this allows you to view it

View(dipper)

# this runs a simple model

Example <- mark(dipper, output = FALSE)

# and this deletes it

rm(Example)
```

1.3 Modelling categorical effects

In today's lab we will be testing some relatively basic, but fundamentally important, hypotheses about variation in survival using what we will term **categorical effects**. These effects, as the name implies, accommodate situations where observations can be grouped into discrete categories, and we can ask whether a biological process, such as survival, differs among the categories. Before we get into the meat of this lab where we actually run survival analyses in program R, I want to cover this concept of categorical variation in survival from a more conceptual point of view.

A very basic example would be the categorical effect of animal sex, where modeling differences in survival among sexes allows us to ask the question of whether males and females differ in their survival. A hypothetical example might look like this:



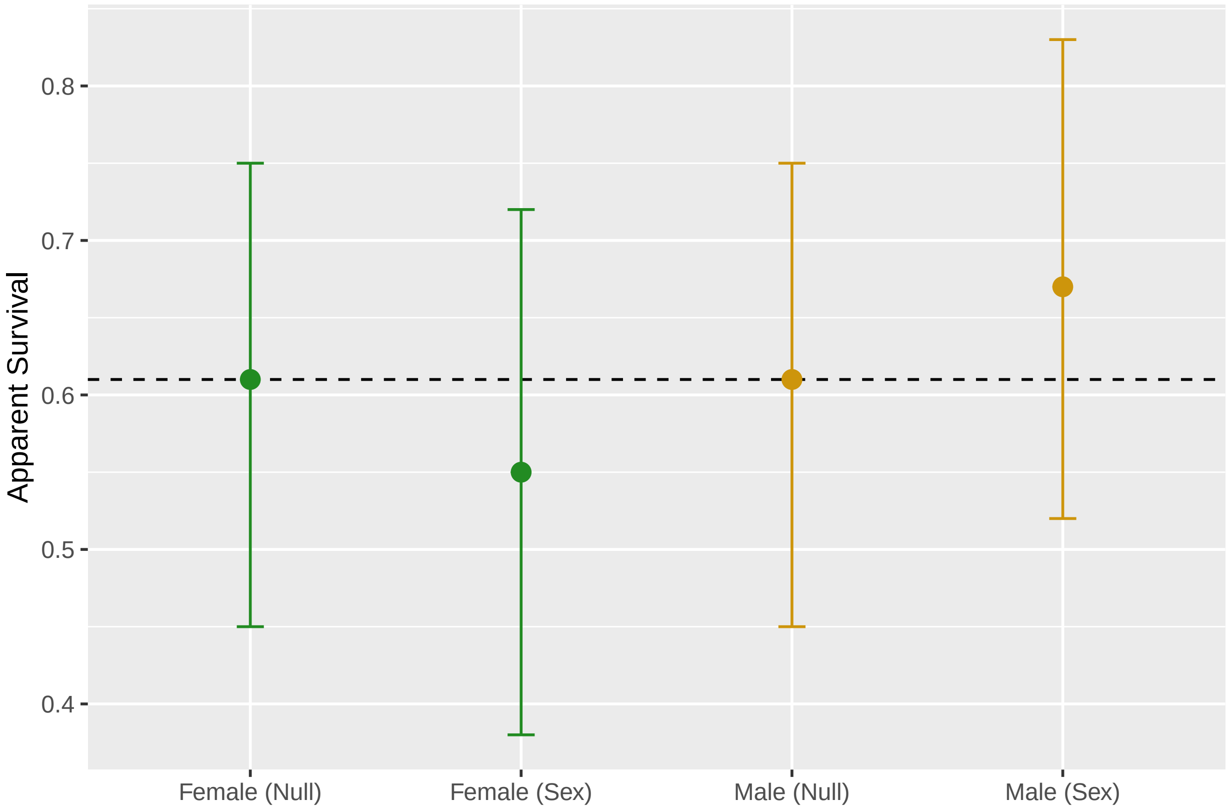
Where in this case, the estimates suggest that males have an ~0.12 greater apparent survival probability than females. We will refer to this as the 'Group' model, and the model notation we will use for it is as follows:

$$\phi(Group) \tag{1}$$

However you will also note that the 95% confidence intervals (error bars in the figure) overlap fairly widely, so we will also want to consider whether these differences can be considered significantly different from each other in a statistical sense. Here, we could run an alternative model where we take away the categorical effect of sex, and assume no differences among males and females in survival. This is generally termed a 'null' model, and the notation involves using a (.) as

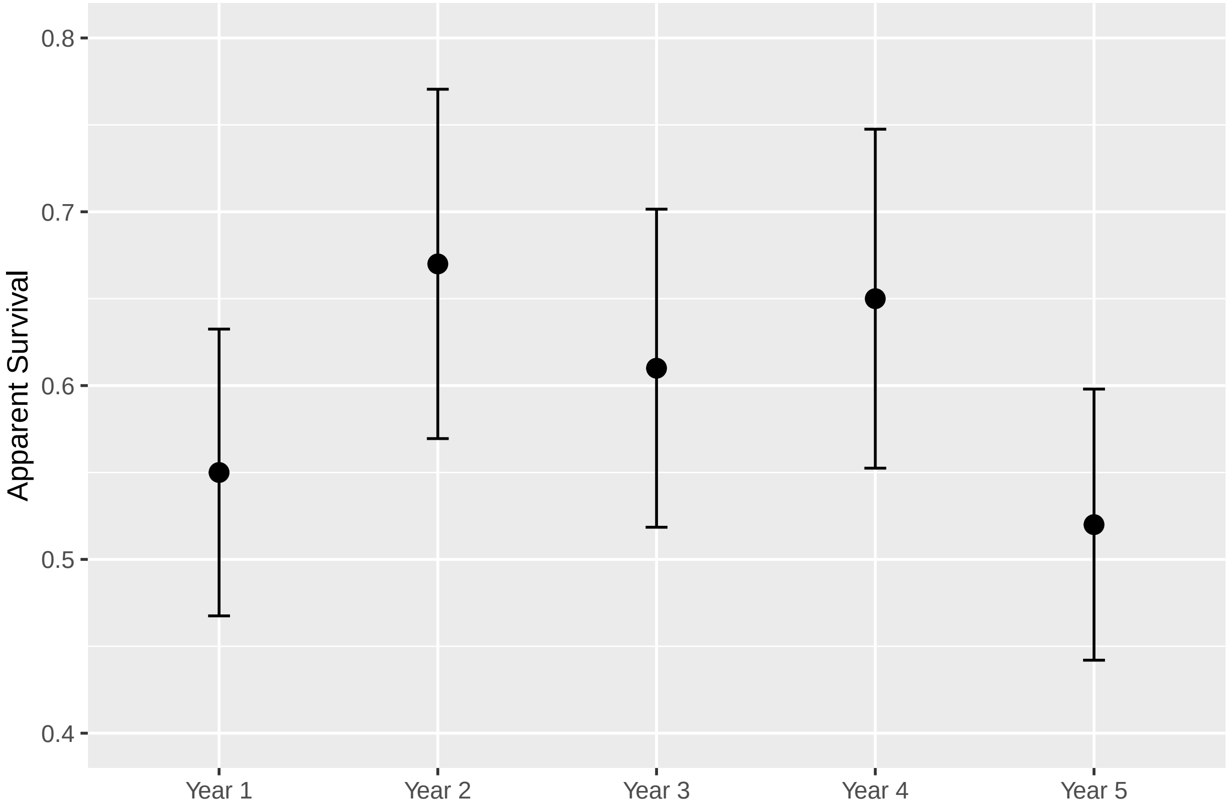
$$\phi(.) \tag{2}$$

Take a look at the following figure, where we add some hypothetical estimates from a null model to those from our previous sex-specific model:



Where the dashed line is showing that the null model predicts the same, constant, survival rate for both males and females. As we work through today's lab we will illustrate, using results from the analysis, how we can construct these two alternative models, use maximum likelihood to derive the most likely estimates for survival under each model and given the data, and use AIC as a tool for testing the statistical hypothesis that one model is better-supported by the data than the other.

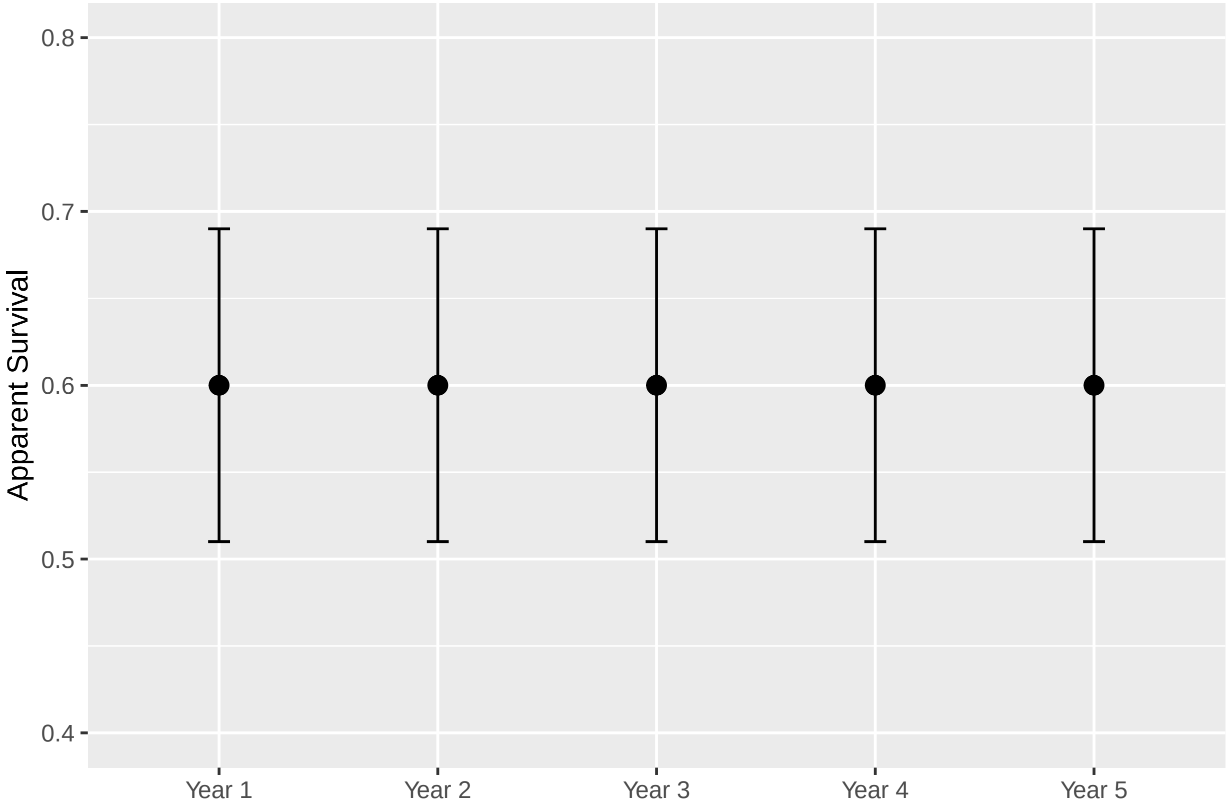
Now, let's extend these principles of categorical variables to another important source of variation in animal survival - variation associated with time. Often we think of temporal variation in the annual survival probability - that is, why do fewer animals survive in one year compared with other years - and that is what we will explore today. However, we will see in future weeks that we can model survival across other time scales such as weeks or months. In any event, from a conceptual standpoint developing a model that allows survival probabilities to differ among years is another form of a categorical effect. Just as with the example for the grouping variable of sex, we can consider each individual year in our analysis as a discrete 'group', where all animals alive and marked during that year held membership in the group. In the model, we allow independent estimates of survival for each year - for example:



Importantly here there is no trend or dependency on the past year's estimates - all years are fully independent. We will refer to this as the 'time' model.

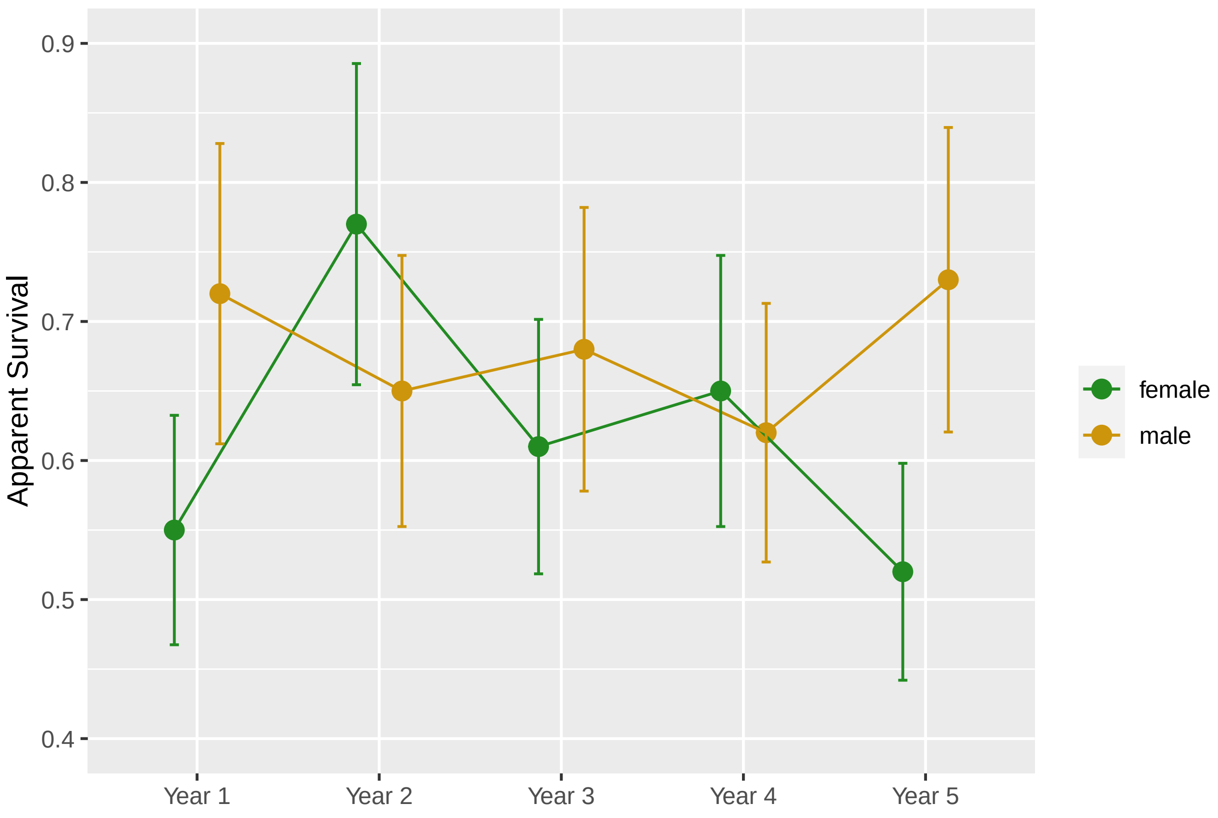
$$\phi(time) \tag{3}$$

As with the sex-level group effect, we could contrast this model with a null hypothesis that there is no annual variation in survival.



Where this is the same model as the $\phi(.)$ model above, but in the graph we are reflecting the fact that survival was identical for all 5 years of the study.

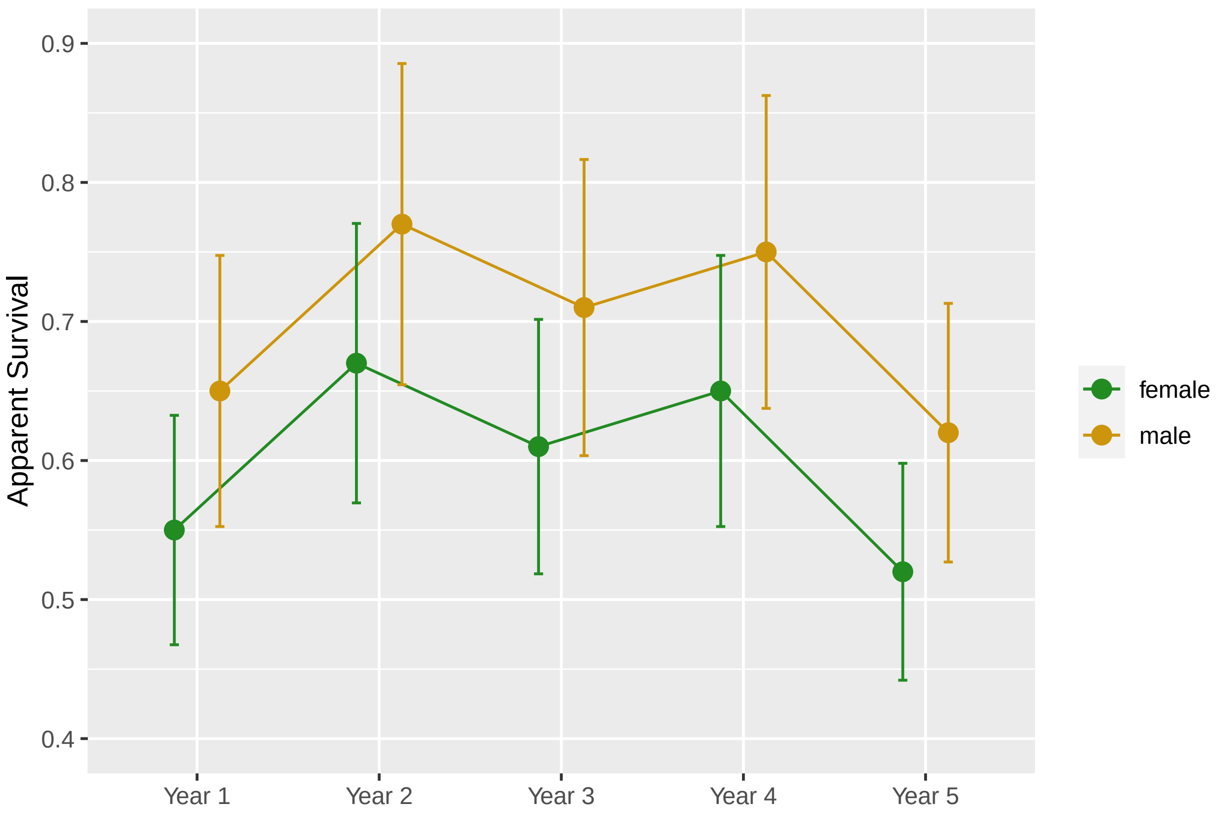
Lastly, I want to combine these concepts of group- and time-varying models into a single framework. Here, we will assume that survival differs among the sexes, that it differs across years, and that the variation among years is independent for the two groups.



The important thing to notice from this graph are that for males and females the apparent survival probability changes each year, but in different ways. That is, the dynamics of survival are independent among the groups. When we develop a model to reflect this hypothesis, we will call it the *group.by.time* model as it represents an interaction between the effects of group (in this case sex) and time (in this case years). The model notation would be

$$\phi(Group \times time) \tag{4}$$

This model differs from an alternative hypothesis, where we allow there to be an effect of group membership, and we allow for an effect of time, but we **constrain** the model such that the effect of time is the same for each group.



The key difference here is that for each sex, the annual estimates of survival change through time, but the change occurs in parallel. We refer to this model as the *group.plus.time* model, or in notation form as:

$$\phi(Group + time) \tag{5}$$

The distinction between the additive ϕ(Group + time) and interactive ϕ(Group x time) forms is important. In practice the big difference you can take away is that with the additive form the lines connecting the years occur in parallel, while for the interactive form, they do not. But this is a concept I find students often struggle with, so let's briefly unpack the distinctions using a biological example.

> ***Sidebar: Additive vs. Interactive Models*** - Imagine we're studying a population of snapping turtles living in a wetland complex, and we expect to see annual variation in survival due to differences in hydroperiod among years that changes how much wetland habitat is available. We might also expect that female turtles have lower survival than males, because they leave the wetlands to nest and have higher rates of road mortality. In this case the factors affecting

> variability in annual survival are largely independent from those affecting differences between the sexes, so a ϕ(Group + time) model might reflect a more plausible hypotheis. However, imagine an alternative scenario where the female turtles are more susceptible to drought conditions because they are larger and have greater energy demands due to nesting. Now, we can't assume that the effect of hydroperiod is totally independent from the effect of sex. In this scenario, a ϕ(Group x time) model would be the better choice. As we will see in the lab exercise today, we don't need to arbitrarily choose one over the other, as we can run both as alternative models and use AIC to test each hypothesis.

Using the R package RMark, we can build models reflecting each of the different categorical effect structures above, and fit them to mark-recapture data. RMark will use principles of maximum likelihood to derive the most likely probability values given the data, or said differently, if the particular hypothesized model structure is correct, what values of survival would be most likely to produce the mark-recapture data that were collected. We will explore these principles using the Cormack-Jolly-Seber modelling framework.

1.4 The Cormack-Jolly-Seber (CJS) Model

In this lab we will be using the RMark package to analyze mark-recapture data. The CJS model contains two different parameters - the apparent survival probability, denoted by the Greek letter ϕ (pronounced phi or fee), and the detection or recapture probability, p. We will fit each of the model structures described above (e.g. Group, time, Null, etc.) on both p and ϕ.

For full understanding you will want to review the conceptual foundation behind these analyses provided by Powell and Gayle Chapter 10, as well as the materials we cover in lecture on survival estimation. Here I'll give just a very brief overview of some key concepts so we are all on the same page going into this lab.

1.4.1 Apparent survival

In the CJS framework we will estimate the **apparent survival probability**, ϕ. This survival probability is termed 'apparent' to distinguish it from the **true survival probability** because we assume that some animals may leave the system due to **permanent emigration**, and as a consequence some marked animals are not available for recapture. This means that mortality and permanent emigration can't be distinguished in the data, so the apparent survival probability is generally thought to be lower than the true survival probability.

Apparent survival probabilities are relevant to the **intervals** between **capture occasions**; they describe the probability that a marked individual survived from capture at time t to recapture at time t+1. Because they are interval-specific, you will always have 1 fewer survival probability estimate than there are capture occasions. It also means that the timing of sampling dictates the time scale to which your survival estimates are relevant; if you catch animals once a year during the month of June, then you will produce an estimate that an animal survives from June one year to June the next year (i.e. an annual apparent survival probability).

These survival probabilities can also be generalized to the population level (assuming adequate sampling, of course) such that if you have an estimate of ϕ=0.5, it implies both that an individual animal has a 0.5 probability of surviving the interval, and that roughly 50% of the population is expected to survive a single interval. Although a statistician would probably argue with me, switching back and forth between thinking of these values as proportions (1/2), probabilities (0.5), or percentages (50%) is fine.

1.4.2 Detection probability

The **detection probability** may also referred to as the **recapture probability**, although it won't always be necessary to physically recapture animals to apply a CJS model. For example, avian ecologists often use color banding to mark songbirds, and resightings of previously-marked birds could be used in place of physical captures. Unlike ϕ, p is considered relevant to the sampling occasions themselves, and it defines the probability that a previously-marked animal will be detected *given that it is alive and available for detection.* This last italicized point is important, because it provides a fundamental assumption of the model that an animal can't be detected if it is not alive or has left the study system.

Like with ϕ, for a study duration of t occasions we will only obtain t-1 estimates of p. Because p gives the probability of recapture, we must first have marked animals available to be recaptured. Since by definition we will initially mark animals during the first occasion of the study, we won't have any available for recapture until the second year, or t=2. Thus when we obtain estimates of p from the data, the first estimate will be the probability of recapture during year 2 of the study.

1.5 White-throated dippers and riparian systems

In this lab we will estimate detection and apparent survival probabilities from mark-recapture data on White-throated Dippers (*Cinclus cinclus*). Dippers are a very unique family (Cinclidae) of passerine birds that have fascinating aquatic feeding habits. They are found along fast-flowing streams and rivers, and they actually dive under the water, where they walk along submerged rocks to forage on aquatic invertebrate larva. They use their wings like flippers, their claws for gripping slippery rocks, and have an oily preen gland like ducks and geese for waterproofing their plumage. Although we are working with data from a Eurasian species of dipper today, their close cousin the American Dipper (*Cinclus mexicanus*) can be found in riparian areas throughout western North America. In general, these are seriously cool birds.

Figure 1: **White-throated Dipper**, source = Wikimedia Commons

The data we'll use today is a classic data set that comes from a field study of dippers in eastern France, where birds were captured and banded along streams for 7 consecutive years. For today's purposes, we will assume captures occurred during the month of June, which provides us with our sampling occasions, and the period between one June and the next represent our intervals. In this analysis, our apparent survival probability reflects annual survival. During intervals 2 and 3 there were major flooding events in the system, whereas the other intervals did not contain flood events. We will use these data in a Cormack-Jolly-Seber (CJS) framework to estimate annual probabilities of ϕ and p to evaluate if annual survival changed among years during this study, and to ask questions about whether survival varied among individuals based on their sex.

2 First Exercise: Getting Started With RMark

As with all R analyses, our first step will be loading the package that we'll need to use. Remember, if you have never used the package before, you should remove the # in front of the install.packages() command and run that function.

```
# install.packages('RMark')
library(RMark)
```

A number of example data files are included in the RMark package, including the mark-recapture data on white-throated dippers, which we can access using the data() command.

```
data(dipper)  # this loads the data

View(dipper)  # this allows you to view it
```

Viewing this input file, you'll see that it includes two columns; one labeled 'ch' which is our **capture history**, and the second which identifies the sex of each individual in the capture history. Each row of data reflects a unique individual (dipper) in the dataset. The string of successive ones and zeros is the encounter history, where each 1 represents an occasion where the individual was observed, and each 0 where it was not. Notice the number of encounter occasions is defined by the total length (number of ones or zeros) in the history, so we can infer that there were 7 capture occasions and 6 survival intervals. Also in this case there was "staggered entry", where new individuals were captured during each occasion. We can tell this because for some individuals there are a number of zeros on the front-end of the history before the first one, which is the first time the dipper was captured.

Using the summary() command, we can also get an idea of the sex-specific sample sizes

```
summary(dipper$sex)
```

The first step in any marked analysis will be converting the input file data into a format that marked will use when running our models. We will use the process.data() and make.design.data() commands to accomplish this.

```
# dipper = name of file, model = the type of
# analysis nocc= number of capture occasions,
# groups = grouping variables


dipper.process <- process.data(dipper, model = "CJS",
    nocc = 7, groups = c("sex"))

# applies command to previous object

dipper.ddl <- make.design.data(dipper.process)
```

Let's run an initial model and look at the output. For this model, we will assume that both ϕ and p are **time-varying**, and for now we will ignore the presence of any group effects. So to come back to the introductory materials from the lab, this model would have the standard notation

$$\phi(time)p(time) \tag{6}$$

and in the R code below we will name the object Phi.time.p.time. The mark() command will execute the model using the dipper.process and dipper.ddl objects we just made, and the model.parameters statement is where we specify the model structures for ϕ and p, indicating this model has a 'time' structure for both parameters.

```
Phi.time.p.time <- mark(data = dipper.process, ddl = dipper.ddl,
    model.parameters = list(Phi = list(formula = ~time),
        p = list(formula = ~time)), brief = TRUE, silent = TRUE)
```

In the short time (probably a second or so) after you clicked run, a bunch of stuff just happened 'under the hood'. In the background R 'called' the Program Mark executable file on your computer and used MARK to execute the FORTRAN code that conducts the likelihood estimation for the model. The particular model structure we defined in the code above was fit to the data contained in the capture history, using maximum likelihood to arrive at the most likely values for ϕ and p given those data and the model structure. We'll dig into this a bit more fully as we explore the model results. First, I want to take a quick look at the model structure to orient you to the concept of parameter indexing using the PIMS() command

```r
PIMS(model = Phi.time.p.time, parameter = "Phi", simplified = TRUE)
```

PIMS stands for Parameter Index Matrix, and it's the mechanism that MARK uses to identify whether one parameter, such as the survival probability in year 1, is different from any other parameter (e.g. the survival probabilities in all other years). In the PIM matrices, each column represents an interval where we will estimate survival (ϕ). Each row represents a different **cohort** of individuals, or animals that are originally marked in a shared occasion. We begin with those marked in occasion 1 (row 1), those marked in occasion 2 (row 2), and so forth. Notice that each row becomes one column shorter as we move down the matrix; animals marked in occasion 1 cannot be part of the occasion 2 cohort, those marked in occasion 2 are not part of the subsequent cohorts, etc.

Notice also there is a different value in each cell for each column, but not for each row. What these tell us is how ϕ is going to be modeled for each cohort in each occasion. These numbers vary across columns because we have specified that for each interval, the survival parameter will be estimated independently of any other interval. For example, the apparent survival rate for males in year 1 will be different than for males in year 2. However, numbers do not vary across rows because we assume that survival is consistent in any given year across cohorts. For example, whether you were originally tagged in year 1 or year 2 does not affect your survival in year 3.

Lastly, notice that we have identical PIMs for the male and female groups, because this is the ϕ(time) model and not a model that includes a group effect [e.g.ϕ(Group + time)]. To highlight the differences here, let's run a second model and look at its PIM structure. Keep in mind, we haven't looked at any results yet - we are simply working to understand some fundamentals of the model characteristics. Run the following code, and then execute the PIMS() command to return the PIMS of the resulting model, which has the notation

$$\phi(Group)p(time) \tag{7}$$

```r
Phi.Group.p.time <- mark(data = dipper.process, ddl = dipper.ddl,
    model.parameters = list(Phi = list(formula = ~sex),
        p = list(formula = ~time)), brief = TRUE, silent = TRUE)
```

The biggest difference in the above code from the last model we ran are 1) the object has a new name and 2) in the Phi=list() statement, it uses a different formula applying the effect of sex, rather than time. When you execute the PIMS() command successfully, you should get this as an output

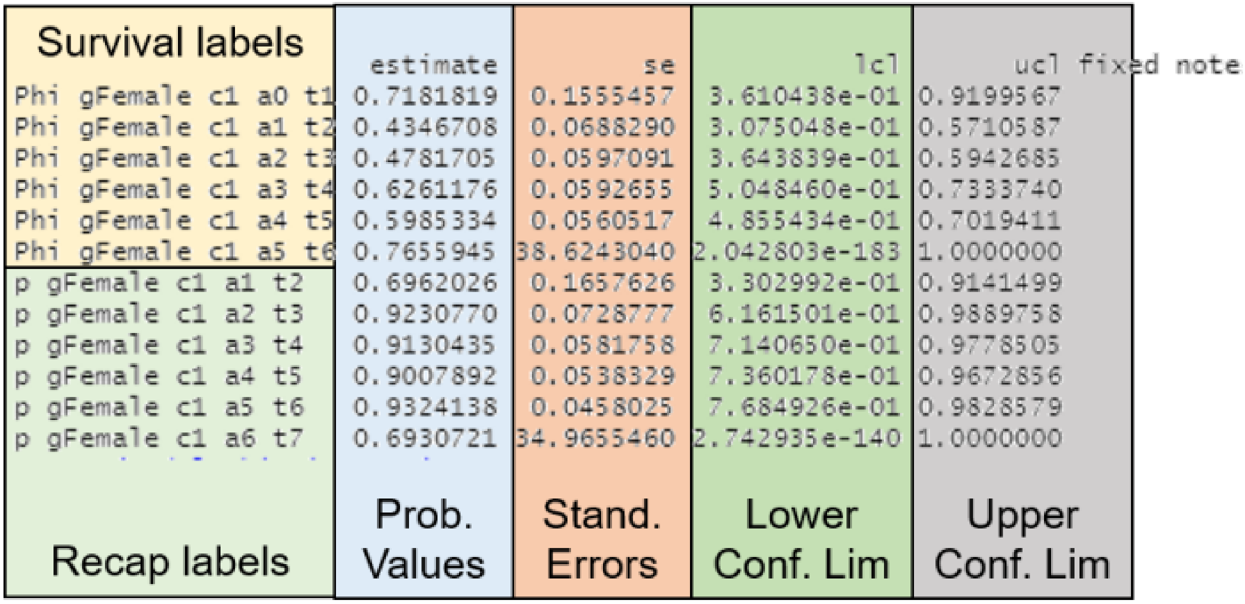
```
## group = sexFemale
##    1  2  3  4  5  6
## 1  1  1  1  1  1  1
## 2     1  1  1  1  1
## 3        1  1  1  1
## 4           1  1  1
## 5              1  1
## 6                 1
## group = sexMale
##    1  2  3  4  5  6
## 1  2  2  2  2  2  2
## 2     2  2  2  2  2
## 3        2  2  2  2
## 4           2  2  2
## 5              2  2
## 6                 2
```

Compare these PIMS to the first set we returned from the ϕ(time) structure, and you should see a clear difference. For this model the index values do not change across columns, so there is no time structure inherent to the model. But, there are different values between the male and female matrices, so the model accommodates different estimates between the sexes. I'll give you some more insight into how the PIMs facilitate the modeling in a second, but first let's take a look at some results from these two models that we have ran and connect them back to these PIM structures. Various results of each model are stored within the model object that was created when we executed the mark() command.

First let's pull the **real parameter estimates** for from the ϕ(time) p(time) model. The 'real' estimates will differ from other model estimates, namely the **beta parameter estimates**, which we will get into during the next lab. Also we can print the PIMs again just to line them up with the model result.

```
Phi.time.p.time$results$real

PIMS(model = Phi.time.p.time, parameter = "Phi", simplified = TRUE)
```

For the model results, notice that we have 6 lines of parameter estimates. The column on the left provides the labels, and then we have the estimates themselves (probability values), the standard errors of those estimates, and the lower and upper limits of the 95% confidence intervals. Your ability to read and understand these results outputs will be really important moving forward, so I've included a color-coded image below just to make sure these components are clear.



```
                             estimate          se          lcl       ucl fixed note
Phi gFemale c1 a0 t1 0.7181819  0.1555457  3.610438e-01 0.9199567
Phi gFemale c1 a1 t2 0.4346708  0.0688290  3.075048e-01 0.5710587
Phi gFemale c1 a2 t3 0.4781705  0.0597091  3.643839e-01 0.5942685
Phi gFemale c1 a3 t4 0.6261176  0.0592655  5.048460e-01 0.7333740
Phi gFemale c1 a4 t5 0.5985334  0.0560517  4.855434e-01 0.7019411
Phi gFemale c1 a5 t6 0.7655945 38.6243040 2.042803e-183 1.0000000
p gFemale c1 a1 t2   0.6962026  0.1657626  3.302992e-01 0.9141499
p gFemale c1 a2 t3   0.9230770  0.0728777  6.161501e-01 0.9889758
p gFemale c1 a3 t4   0.9130435  0.0581758  7.140650e-01 0.9778505
p gFemale c1 a4 t5   0.9007892  0.0538329  7.360178e-01 0.9672856
p gFemale c1 a5 t6   0.9324138  0.0458025  7.684926e-01 0.9828579
p gFemale c1 a6 t7   0.6930721 34.9655460 2.742935e-140 1.0000000
```

Figure 2: **Results Output**

So, to walk through some interpretation here, for the first interval we have a survival estimate of 0.718, with a SE=0.155 and a 95% confidence interval from 0.36 to 0.92. This indicates that roughly 72% of dippers marked during the first occasion survived to the second occasion, with some fairly wide uncertainty. In the second line, we see that ϕ=0.434, or roughly 43% of marked dippers survived the second year of the study.

Jumping down to line 7, we see that the estimated recapture probability in the second occasion (pause for a second to remember why the first recapture estimate is for the second occasion) was p=0.696 with an SE=0.165 and 95% CIs from 0.33 to 0.91, for the third occasion p=0.923, and so forth.

There are a few additional quirks about the results printouts we should cover. In the labels you will see the term 'Female' pop up repeatedly, even though this model doesn't have a group effect. What's the deal with that? The simple answer is that RMark is reporting outputs for us associated with the first group that has parameters 1:6 in the PIM structure for this model, which in our case happens to be the Female group (see your prior printing of the PIMS() command for this model to confirm this for yourself), and it is using that as the parameter estimate label in the results. As we build more complex models, we will see how this plays out to label the group effects differently between females and males.

Also note that some of the estimates, in particular the lower confidence limits, are being expressed as scientific notation. For example, the very first lower confidence limit is 3.610438e-01, which is shorthand for 3.610438 x 10^{-1}. If you remember your basic principles of scientific notation, this means we need to slide the decimal point one space to the left, so this value is actually 0.3610438.

The last thing I want to highlight with this is the final estimate for ϕ and p. You'll notice that their estimates look fairly similar to the other values in the results, but their SEs are either considerably larger or smaller (like orders of magnitude) than the others, and their CIs should either be unreasonably narrow or extend from basically 0 to 1. In general terms when we see the SEs and CIs behaving poorly like this, it indicates that the model **failed to converge**, which is a technical way of saying that the computer failed to settle on the most likely value for the parameter, and as a result couldn't derive the variance estimate. Often this can occur due to data limitations, for example if you failed to catch many animals in a particular group or year it will be difficult to estimate a detection probability for that occasion. In this case, there is something specific to the CJS that helps us to explain this convergence issue.

> ***Sidebar - Confounding between ϕ and* p** - This is one very particular issue with the CJS model that I'll highlight here, but we won't get too far into technical details. For a dataset of N years, we already know that at most we can get N-1 estimates each of ϕ and p. But it's more complicated than that, because when we allow for temporal variation in both parameters [i.e. ϕ(time) p(time)], the very last estimates are said to be 'confounded' with each other, which simply means that they are not reliable and should not be interpreted. For example, if you have a 4-year study, your 3rd survival probability and your detection probability in the last occasion are confounded. You can see page 136 of Powell and Gayle for a bit more insight on this subject if you're curious, but the practical implications are that for a study of N years we should actually only expect to get N-2 reliable estimates of ϕ and p

Alright, that's enough of getting into the weeds for the time being. Now that we've covered the basics of building and running a model, we'll get to the business of addressing some competing hypotheses.

3 Second Exercise – Running competing models to test hypotheses.

Next we'll use the models we just constructed as a starting point for testing hypotheses about dipper survival by running additional alternative models and evaluating the relative statistical support among them using AIC. Every time we run a model, that model has a particular structure (based on how we define it in the coding) which defines whether distinct groupings of animals are allowed to have different survival probabilities, or are **constrained** to be the same. As we change the nature of the constraints among models, we are explicitly representing alternative biological hypotheses about survival, and these hypotheses are defined by the model structure we fit to the survival terms.

In part 1, we generated one hypothesis that survival differed among years but was the same between the sexes. If we want to test alternative hypothesis about how survival varies, we can develop some additional models to reflect those hypotheses. Specifically we're going to ask the following questions of our data:

1) Is there evidence that survival of dippers is variable among years, or is it more-or-less constant through time?

2) Do male and female dippers have the same general mortality risk, or can we detect sex-related difference in survival?

3) If there is an effect of sex, do both sexes share the same sources of annual variation in survival?

4) Did the high flood events during intervals 2 and 3 affect dipper survival?

Before we jump into running models, take a few minutes to consider the questions above and relate them to the conceptual model forms that I illustrated with the figures in section 1.3 of the lab. For questions 1, 2, and 3, you should be able to pick out the model form(s) that match. Question #4 isn't directly represented in a figure we've discussed yet, so take a second to sketch out on a piece of paper what you think that might look like (seriously-sketching things out on paper is a great way to wrap your head around new models). We will compare it to the model results we get later.

3.1 Building competing models

Running additional competitive models with RMark is almost trivially easy. So much so that you might risk just cranking through running them without really stopping to reflect on what the model represents, so I will try to walk you through each of the additional models and link them back to the hypotheses. We will do this by further exploring the PIM structures of the models.

First, we will run the most general model structure, which allows for the interaction between sex and time [i.e. ϕ(group*time)]. This model will allow us to address question #3, and the code is as follows:

```
# pay attention to the Phi=list() component of
# the code.

Phi.Group.by.time.p.time <- mark(data = dipper.process,
    ddl = dipper.ddl, model.parameters = list(Phi = list(formula = ~sex *
        time), p = list(formula = ~time)), brief = TRUE,
    silent = TRUE)
```

For this and all models today, we will simply assume a p(time) structure on the recapture term, and focus our work on the survival term. But you should note the line of code where I've instructed you to pay attention. This chunk of code is identical to earlier code where we ran the ϕ(group) p(time) model, except that I have changed the model object name, and added a ’*time’ component to the ϕ structure. This code ‘tells’ RMark to build a model that allows for an interaction between the categorical group effect (sex) and the categorical effect of time (years). Let's take a quick look at the PIMs for this model.

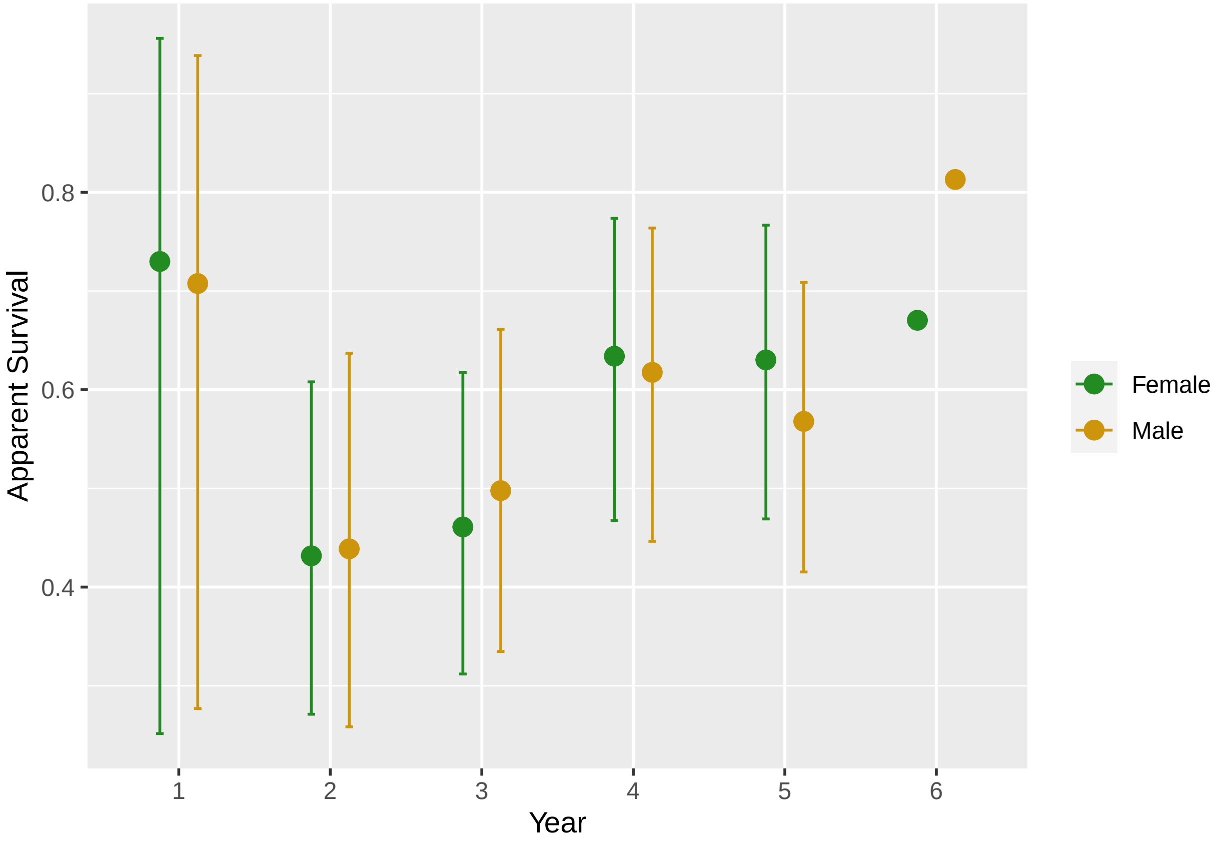
```
PIMS(model = Phi.Group.by.time.p.time, parameter = "Phi",
    simplified = TRUE)
```

```
## group = sexFemale
##    1  2  3  4  5  6
## 1  1  2  3  4  5  6
## 2     2  3  4  5  6
## 3        3  4  5  6
## 4           4  5  6
## 5              5  6
## 6                 6
## group = sexMale
##    1  2  3  4  5  6
## 1  7  8  9 10 11 12
## 2     8  9 10 11 12
## 3        9 10 11 12
## 4          10 11 12
## 5             11 12
## 6                12
```

You should notice here that the PIM values now differ among the male and female groups AND also change across years (columns). The numbers again don't have any bearing on how the survival probabilities are allowed to be different, they just allow them to differ. Let's take a look at the real parameter estimates from this model.

```
Phi.Group.by.time.p.time$results$real
```

Note that we've been given 12 different survival estimates, and they are all at least slightly different from each other. This might be best visualized as a figure (shown below).



Compare the figure to the real parameter estimates you printed in your R console. You'll notice here there is not THAT much difference between the sexes (although it is there), but there is considerable variation among years. We will have to see how this plays out as we run some additional models and use AIC to rank them and identify which are supported.

From here, let's run a few additional models, which will include the model representing an additive effect of age + sex (the ϕ(group + time) model) and a 'null' model, $\phi(.)$

```
# here we just replaced the * with a + in the
# Phi=list() code

Phi.Group.plus.time.p.time <- mark(data = dipper.process,
    ddl = dipper.ddl, model.parameters = list(Phi = list(formula = ~sex +
        time), p = list(formula = ~time)), brief = TRUE,
    silent = TRUE)

# ~1 is the R notation for only fitting an
# intercept on Phi=list()
```

```
Phi.dot.p.time <- mark(data = dipper.process, ddl = dipper.ddl,
    model.parameters = list(Phi = list(formula = ~1),
        p = list(formula = ~time)), brief = TRUE, silent = TRUE)
```

We've now run 5 models total with structures ϕ(group x time), ϕ(group + time), ϕ(group), ϕ(time), and ϕ(.), each with a structure of p(time) on the recapture term. We can use the collect.models() command to build an AIC table from these models.

```
CJS.results <- collect.models()

CJS.results
```

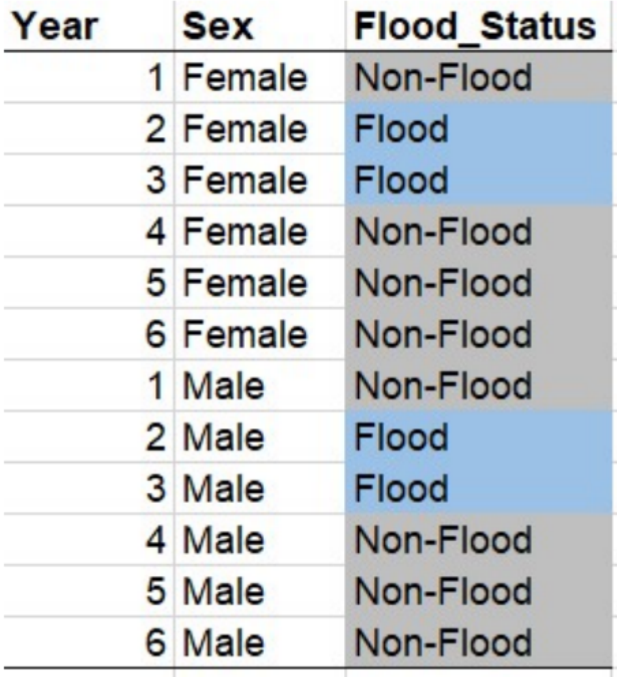
```
##                        model npar     AICc DeltaAICc       weight Deviance
## 1           Phi(~1)p(~time)    7 678.7481  0.000000 0.5889868325 82.00306
## 3         Phi(~sex)p(~time)    8 680.6496  1.901481 0.2276167399 81.82716
## 5        Phi(~time)p(~time)   12 681.7057  2.957545 0.1342402559 74.47310
## 4 Phi(~sex + time)p(~time)   13 683.7328  4.984723 0.0487177065 74.37223
## 2 Phi(~sex * time)p(~time)   18 693.1539 14.405757 0.0004384652 72.99617
```

Using what you've learned about AIC and criteria for model selection, you should be able to infer from this table that the null model is best supported, with one additional model falling within 2.0 ΔAICc of the best model.

3.2 Custom time effects for biological hypotheses

At this point we've only tested 3 of our 4 questions, and we have one more to assess. We want to know if during the 2 flood years (Years 2 and 3), survival differed from all other years. We could possibly infer this from our current results - the fact that none of the models which incorporated 'time' were supported suggests a lack of temporal variation in general - BUT - we haven't yet tested the hypothesis explicitly, which we can do by building a custom model structure.

Let's take a second to think, conceptually, about what this would look like. The key here is recognizing the timing of the flood years (in light blue) vs the non-flood years (light gray) relative to the values indexed by 'time' in our existing models (column 1 in the table).



Year	Sex	Flood_Status
1	Female	Non-Flood
2	Female	Flood
3	Female	Flood
4	Female	Non-Flood
5	Female	Non-Flood
6	Female	Non-Flood
1	Male	Non-Flood
2	Male	Flood
3	Male	Flood
4	Male	Non-Flood
5	Male	Non-Flood
6	Male	Non-Flood

Figure 3: **Flood Hypothetical**

We can replicate this concept in our model structure by modifying the dipper.ddl object, which is what defines the PIM structure for each model. This part is going to get just a little bit 'under the hood' for RMark use - so don't be too concerned if it seems very abstract - but it's also a very powerful way to test specific hypotheses about the mechanisms governing annual variation in demographics, in this case the impact of flooding on Dipper survival.

First, let's take a look at the dipper.dll, which contains the information on the underlying structure of ϕ that RMark uses for fitting models. Enter the following code into your script:

```
dipper.ddl$Phi
```

Pay specific attention to the 'time' column - you will see that it repeats a string of 1 through 6, indicating the 6 annual intervals of the study. Also note that there is an uppercase 'Time' column as well - ignore it for now, we'll come back to it later. Our task is to add a new variable to the design data that tells MARK which of these intervals were flood years vs non-flood years, which we do as follows:

```
# this creates a string of 6 values, each placed
# in the correct sequence of flood vs non-flood
# intervals
```

```
Flood <- data.frame(time = seq(1, 6, 1), Flood = as.factor(c("Non-Flood",
    "Flood", "Flood", "Non-Flood", "Non-Flood", "Non-Flood")))

# here, we append the flood sequence as a new
# variable

dipper.ddl$Phi <- merge_design.covariates(dipper.ddl$Phi,
    Flood, bygroup = FALSE, bytime = TRUE)
```

With this chunk of code, we are telling RMark to add a variable "Flood" to the design data (dipper.ddl) that will create a new grouping variable that is assigned to the individual time periods delineating whether they are a "Flood" or a "Non-Flood" year. In the string above, I've created a vector of 7 terms, with "Flood" in the second and third position. RMark knows to repeat this string and fill in a column throughout dipper.ddl. Let's take a look at the first 18 rows:

```
dipper.ddl$Phi[1:18, ]
```

We've just created a new categorical variable, Flood, that we can use like we have been using the grouping variable Sex. The big difference here is that the grouping is associated with time intervals, rather than individual birds. We can apply it to the code just as we would a group variable.

```
# note Flood term in the Phi=list() statement

Phi.Flood.p.time <- mark(data = dipper.process, ddl = dipper.ddl,
    model.parameters = list(Phi = list(formula = ~Flood),
        p = list(formula = ~time)), brief = TRUE, silent = TRUE)
```

Let's take a look at the real parameter estimates, which give the estimated survival probabilities for Flood and Non-Flood years.

```
Phi.Flood.p.time$results$real
```

Now, there is a really important element here to interpreting your results, and one of the cases where RMark expects that you understand the model structure well enough that you can interpret it without being told explicitly which of the two survival terms are which. In printing the results here, RMark is giving us only two estimates of ϕ, even though we know there are 6 years and two groups, which should = 12 estimates.

When RMark prints the real parameter estimates, it distills them to the simplest version possible given the model. In this case we've created a model with two survival probabilities - one for the Non-Flood years and one for the Flood years. So, it has pulled the first survival estimate from a Non-Flood years, which in the dipper.ddl file is year 1 for the females (hence the labeling in the first line) and the first estimate for a Flood year (hence the labeling in the second line). It's up to us to recognize that, given the structure of our model, all other year/group combinations will have the same estimates as those two because we did not include other effects, such as sex, in the model.

Just to drive home this point home, also take a look at rows 3:8 in the real parameter estimates. These are the recapture probabilities, and you'll see they are only listed for females. Well, this is because our model structure is p(time), and we have assumed no differences in recapture probability between the two sexes.

3.3 Wrapping things up

Alright, we've now run models to test each of our 4 questions, but to get a 'final answer' we need to update our AIC table.

```
CJS.results.final = collect.models()

CJS.results.final
```

You should find that our new Flood hypothesis is best supported by the data as it falls more than 2.0 ΔAICc above the next best model. And based on the real parameter estimates from the model, we know that during flood years survival was reduced by ~0.155 compared with non-flood years. With these final AIC results, you have all you need to complete today's assignment.

4 The Three Big Things You Should Have Learned Today

FIRST The conceptual differences between group*time, group+time, group, time, and null models. These very general model structures will be useful in most if not all mark-recapture analyses and are fundamental for studying the effects of categorical variables.

SECOND How to run individual models using RMark, how to access the real parameter estimates from those models, and how to generate an AIC table of results.

THIRD How to develop a custom model that tests a biological hypothesis about why survival probabilities are similar in some years and different in others.

5 Lab Assignment – Expanded Abstract

For this lab I'm asking you to write an expanded abstract of the work we completed during the lab. You're probably familiar with the concept of an abstract – it's a short summary (usually 300 words or so) that typically accompanies a scientific article, or other forms of science communication such a conference talk. An abstract is designed to give an overview of the work that was completed so that your audience can have a basic frame of reference before they digest the larger body of the work (i.e. the paper). Or, if they aren't planning to read the paper in its entirety, they can at least get a quick appreciation for the background, objectives, methods, results, and interpretation of what you did.

An expanded abstract differs from a typical abstract in that it contains a bit more detail (and may therefore be longer) and also includes some visual aids, such as tables and/or figures. Often you may be asked to write an expanded abstract when, for example, submitting a proposal to give a talk at a conference, or when applying for a grant. Your assignment for this lab is to produce an expanded abstract describing the analysis you performed and its results. The format is much like an abbreviated version of a scientific paper. You'll need to include background information on the study organism, the biological questions you are asking, the data that were used to address them, the mark-recapture methods you used, the results you found, and a short interpretation of the results.

I've posted an example abstract on Brightspace (under the lab folder) that you can read for a very clear picture of what I am asking for, and you can also look to abstracts in recent peer-reviewed journal articles for inspiration, too.

Specifics: Your expanded abstract should include the following information. See formatting guidelines below - these headings are not to be used in the abstract itself but exist to guide you through it.

1. *Title* – Give your abstract an informative title consistent with a scientific investigation – don't just call it "Lab 5 Assignment"
2. *Background* – describe the study organism and the systems they inhabit; 2-3 sentences.
3. *Objectives* – what was the motivation for conducting this work or the question(s) you were hoping to address with it (I realize your true motivation is to complete an assignment, but pretend you are motivated by some desire to better understand dipper ecology or conservation). 2-3 sentences.
4. *Methods* – Typically you would talk about field methods, but in this case you weren't involved in data collection which makes that tough. Glean some information from the Lab 5 materials, and at least mention something general about how the data were collected. Also describe your approach to the survival analysis. 2-4 sentences.
5. *Results* – Describe your results in terms of which hypothesis (or hypotheses) were supported by the data based on the AIC model selection statistics (e.g.ΔAICc and model weights). Also report what your estimates of apparent annual survival and recapture rates were for the best supported model(s). Please include an appropriate measure of precision with any estimates you report – for example, an appropriate way to report an average apparent survival estimate for male dippers would be (ϕMale = 0.55 ± 0.05 SE). 3-5 Sentences.
6. *Conclusions* – Sum up the work with some statement of its relevance to either basic or applied science. You've tested some hypotheses, now tell the reader why the answers to those questions are important. 2-3 Sentences.
7. *Table* – Create a table that presents your AIC Model selection results. You should also refer to this table in the text when discussing support for your competitive models (see my example).
8. *Figure* – produce one figure that visually depicts some aspect of your results, and reference the figure in text.

Formatting: Your abstract should be double spaced with 12 point Times New Roman Font, and should be no longer than 500 words (shorter is ok too) with 1 figure (use Arial font) and 1 table (use Times New Roman). You should also include an appropriate title, your name, affiliation (department and university), and email address, following the formatting in the example I've provided on Brightspace.

Note - there are some tricks in the reference section of this document for outputting your R results to Excel. There is no expectation that you make your tables and figures in R.

Your figures and tables should follow general formatting convention that we've used throughout the semester. Also, you should report 2 decimal places in all tables unless it makes intuitive sense to do otherwise. For example, the number of parameters (K) are whole numbers (1,2,3, etc.) so it doesn't make sense to report decimal places.

The suggested number of sentences and word limit I've listed above are more guidelines than rules; as always, you should use the appropriate amount of space to clearly describe what you did. Being excessively verbose is no better than leaving out important details.

Grading: I am looking for you to demonstrate that you understand how the analysis was conducted and why this type of survival analysis is important to begin with, that you can appropriately interpret the results (both AIC and parameter estimates), and that you can effectively communicate your understanding of all the above. I also expect that you will follow the formatting conventions I've laid out for you. Most importantly, I want to see evidence that you've THOUGHT about the analysis you've just conducted and the implications of its results. Because of this last point, I don't recommend waiting until the last minute to complete this assignment.

Please turn in your completed expanded abstract as a Word document with table and figure embedded, as well as your completed R script, via the Assignments page on Brightspace.

If you've got any questions about the assignment, or need help with the analysis, please ask!

6 Quick Reference for Important R Commands in This Lab

RMark models are structured as follows:

```
Model_Name<- mark(data= ,# process data object
                ddl= ,  #design data object
                model.parameters=list(Phi=list(formula=~Variables),  #Phi structure
                                      p=list(formula=~Variables)),   #p structure
                brief=TRUE, silent=TRUE) # commands to suppress output
```

Code to output the real parameter estimates (survival and recapture probabilities) from a particular model, create a data frame for that table, and export it as a csv file to access in Excel. Note you'll need to change the file path in the write.csv statement:

```
## Throughout: change Model_Name to the name of
## your most supported model

# Return the real parameter estimates of a model
# object

Model_Name$results$real

# create dataframe of estimates

Survival.Estimates <- data.frame(Model_Name$results$real)

# exports the dataframe change the file path to
# be appropriate for your computer but ensure
# that it ends with 'RealEstimates.csv'
write.csv(Survival.Estimates, file = "C:/Your Path/Subdirectory/RealEstimates.csv")
```

Code to access an AIC table for all models you have run, create a dataframe for that table, and export it as a csv file to access in Excel. Note you'll need to change the file path in the write.csv statement:

```
# pull the models together

AIC.Table.Object <- collect.models()

# print the AIC table

AIC.Table.Object
```

```r
## extract the table as a dataframe

AIC.Table <- data.frame(AIC.Table.Object$model.table)

# write the excel file

write.csv(AIC.Table, file = "C:/Your Path/Subdirectory/AICTable.csv")
```

7 For More Information

These labs are meant to be a passing introduction to a vast subject area, and if you find yourself moving forward conducting mark-recapture estimation as part of your graduate studies or in a professional setting, the resources available the phidot web page are a great place to pick up where we leave off. Astute observers will note that the website is named after one of the models we introduced today…

8 GLOSSARY OF KEY TERMS

Apparent Survival Probability In the context of a mark-recapture analysis, survival is said to be 'apparent' when data do not exist to separate the mortality process from permanent emigration. Apparent survival is typically thought to be biased low relative to true survival.

Beta Parameter Estimates Coefficients estimated from the model that require transformation to convert to true probability values. More on these in subsequent labs.

Capture Occasions The time periods during which sampling to detect marked animals occurs. Recapture probabilities are relevant to the capture occasions.

Categorical Effects A set of fixed effects in a model where observations may be grouped into discrete categories. For example, sex, age class, study area, or year may all be categorical in nature.

Capture History The sequence of data that defines the frequency at which animals were observed during a study; in most cases the base input data for a mark-recapture analysis.

Constraint In the case of mark-recapture modeling, creating a model structure where parameter estimates that could in theory be unique are forced to be similar. For example, in this lab the */phi*(Flood) model structure is a constrained form of the */phi*(time) model structure. Both describe temporal variation in survival, but the */phi*(time) model is more general in nature.

Detection Probability The probability that a previously-marked individual will be encountered during a capture occasion, given that it is alive and available for detection. Synonymous with recapture probability.

DNS Daily nest survival, or the probability that a nest will remain active for a 1-day (i.e 24-hour) period

Intervals The periods of time that occur in between capture occasions. The apparent survival probability values are relative to the interval periods.

Model Convergence Indicates whether the variances, SEs, and CIs for all parameters can be estimated through maximum likelihood. Extremely large of very small SEs and CIs indicate lack of convergence.

Multinomial Probability A probability value drawn from a distribution in which there are >2 potential outcomes.

Permanent Emigration When an animal leaves the study area permanently and is not available for recapture or detection.

Real Parameter Estimates The estimated probability values from a given model. In the case of the CJS, these are the apparent survival probabilities and the recapture probabilities.

Recapture Probability The probability that a previously-marked individual will be encountered during a capture occasion, given that it is alive and available for detection. Synonymous with recapture probability.

True Survival Probability The true underlying probability that an animal will remain alive throughout an interval of a given length.