# Survival analysis in an experimental microbial aquatic community

08 September 2021

A single species extinction can cause a chain of secondary extinctions leading to a trophic cascade in a food web (Dunne and Williams 2009). Hence, it is important to improve our predictive understanding of the causes and consequences of biodiversity loss in an ecosystem. Knowing the reason behind species extinction can help us improve conservation efforts.

The order in which species go extinct in a food web influences the consequences of the extinction event. For example, loss of larger species first, followed by smaller species, can cause a more rapid loss of ecosystem functioning in marine sediments than a random extinction order, according to simulation-based predictions (Solan 2004). Therefore, other than knowing which species might suffer extinction, it is also crucial to know when the species will go extinct.

Species in a community are linked by a complex network of interactions such as competition, predation and mutualisms. This can result in the extinction of one species caused by the loss of other species (Paine 1966). Species extinction can also be influenced by abiotic environmental factors (Cardillo 2005).

In this project, we would like to study the effect of environmental factors and interspecific interactions on the extinction rates of species in an experimental microbial aquatic community composed of 17 protist species. In the experiment, four different treatment combinations had been used, varying temperature and nutrient concentration in the microcosms. We would like investigate the combined effect of interspecific interactions and environmental factors on the determinants of extinction events.

We will use survival analysis with a Bayesian approach to estimate the mean extinction times (MTEs) of the species. Possible research questions of this project are:

- What would be the effect of using informative priors derived from empirical studies or elicited from experts for the parameters of the model?
- How can we change the survival model presented here to make them more realistic/interesting than the preliminary study below?
- How extinction times distributions differ among the 17 species of the community?
- What is the effect of the different treatments (eg temperature and nutrient concentration) on the MTE of the different species?
- How can we take into account species traits (eg body size) and food web structure in the analysis?
- What is the effect of speies interactions and/or environmental factors on the collpase of this little ecological community?

## Aquatic food web

The aquatic food web consists of a 17 taxa of aquatic eukaryotic microorganisms, unknown heterotrophic nanoflagellates and an unknown bacterial flora (Fig. @ref(fig:fig1)). The species in the food web (Fig. @ref(fig:fig1)) are: 3 *Blepharisma japonicum*, 4 *Chilomonas paramecium*, 5 *Colpidium striatum*, 6 *Colpoda cucculus*, 7 *Cyclidium glaucoma*, 8 *Didinium nasutum*, 9 *Dileptus anser*, 10 *Entosiphon*, 11 *Euplotes patella*, 12 *Loxocephallus*, 13 *Lepadella*, 14 *Dicanophoridae*, 15 *Paramecium bursaria*, 16 *Paramecium caudatum*, 17 *Telotrochiudium*, 18 *Tetrahymena piriformis* and 19 *Vorticella*.

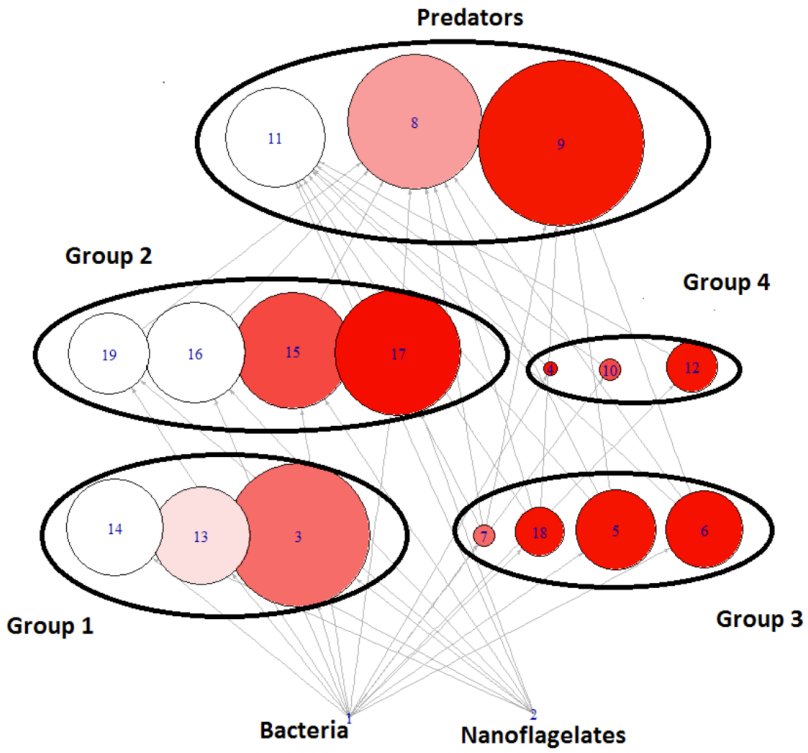



Figure 1: Food web: species are grouped according to their trophic group. The size of each node is proportional to the logarithm of the body size of the species and color indicates the mean extinction time of the species computed by maximizing likelihood across all treatments, with a darker shade of red corresponding to lower extinction times [@palamaraTheoreticalEmpiricalStudies2015].

## Dataset

The experiment for the current dataset was conducted by Nicholas Worsfold at the University of Sheffield (Worsfold 2007). The data is from a highly replicated microcosm experiment where the extinction times of freshwater protists species forming a small food web was recorded. Four different treatment combinations have been used, varying temperature and nutrient concentration in the microcosms.

The raw data from the experiment provides the presence or absence of each species for each of the eight weeks of the experiment, for each of the 200 replicates. We cleaned the data removing Lazarus effects (species recorded as extinct and then found again in a subsequent week), and we present the last week at which the species was observed present in the microcosm.

## Preliminary study using a subset of the dataset

Below, we subset the dataset one species *P. bursaria*. We fit simple survival model using a Bayesian approach, to compute The Mean Time to Extinction for the Species, considering all experimental treatments.

```
## Loading some packages
if ( ! require(runjags  ))   { install.packages("runjags" );   library(runjags ) }
if ( ! require(rjags    ))   { install.packages("rjags"   );   library(rjags   ) }
if ( ! require(bbmle    ))   { install.packages("bbmle"   );   library(bbmle   ) }
if ( ! require(ggpubr   ))   { install.packages("ggpubr"  );   library(ggpubr  ) }
if ( ! require(reshape2 ))   { install.packages("reshape2");   library(reshape2) }
if ( ! require(IDPmisc  ))   { install.packages("IDPmisc" );   library(IDPmisc ) }


## Importing the dataset
## Change the path accordingly
dd <- readRDS("Dataset/extinction.week.Rdata")

## Check the structure of the data set
str(dd)
```

```
## 'data.frame':    3400 obs. of  5 variables:
##  $ jar         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ temp        : int  15 20 20 20 15 20 20 20 15 20 ...
##  $ energy      : num  0.275 0.275 0.275 0.825 0.825 0.825 0.275 0.275 0.275 0.825 ...
##  $ species     : chr  "blepharisma" "blepharisma" "blepharisma" "blepharisma" ...
##  $ week.persist: int  8 8 4 3 6 1 6 6 8 2 ...
```

```
##############################################

## Just for some plotting, not for analysis
## Add 0.5 to get approx. extinction time
dd$week.persist <- dd$week.persist + 0.5

## Set the extant populations to an extinction time of 20 weeks
dd <- transform(dd, dd.fm=ifelse(week.persist==8.5, 20, week.persist))

## For potential survival analysis
## Make binary (0 or 1) variables for temperature and energy treatments
dd <- transform(dd, temp.var=ifelse(temp==15, 0, 1))
dd <- transform(dd, energy.var=ifelse(energy==0.275, 0, 1))
```

```r
## Define day1 (the last week at which the species was observed)
## Define day2 (the first week when the species was not observed)
dd <- transform(dd, day1=week.persist-0.5)
dd <- transform(dd, day2=week.persist+0.5)

## Set a censored observation to Inf
dd$day2 <- ifelse(dd$day2==9, Inf, dd$day2)

############################################################

## Let's check the dataset after this first transformation
head(dd)
```

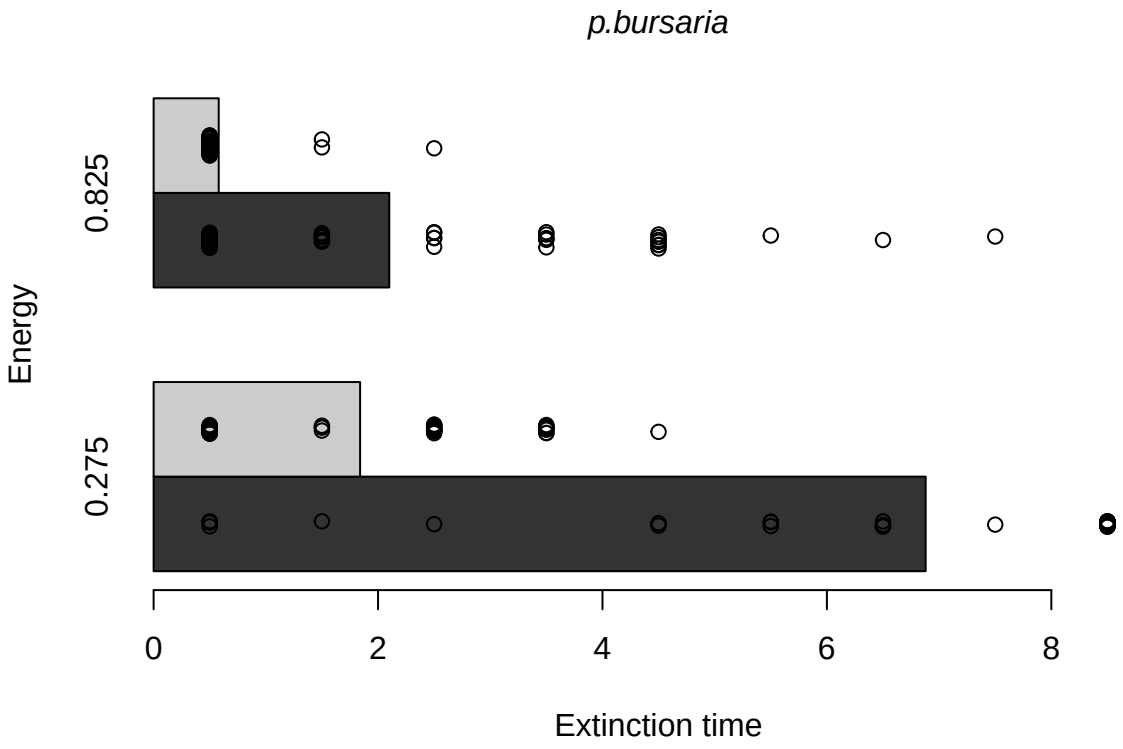
```
##   jar temp energy     species week.persist dd.fm temp.var energy.var day1 day2
## 1   1   15  0.275 blepharisma          8.5  20.0        0          0    8  Inf
## 2   2   20  0.275 blepharisma          8.5  20.0        1          0    8  Inf
## 3   3   20  0.275 blepharisma          4.5   4.5        1          0    4    5
## 4   4   20  0.825 blepharisma          3.5   3.5        1          1    3    4
## 5   5   15  0.825 blepharisma          6.5   6.5        0          1    6    7
## 6   6   20  0.825 blepharisma          1.5   1.5        1          1    1    2
```

```r
## Some summary of the dataset
all.species <- sort(unique(dd$species))
all.temp    <- sort(unique(dd$temp))
all.energy  <- sort(unique(dd$energy))
```

## SOME PLOTTING

```r
## Here we define a function to plot the extinction times distribution of one species.
## The function takes as input the data, the species we want to consider
## and some graphical paramateres
Plot.data.bars <- function(dd, spp,
                           shaydz=grey(c(0.2, 0.8)),
                           horiz=T) {
  spp.dd <- subset(dd, species==all.species[spp])
  mean.week <- tapply(spp.dd$week.persist,
                      list(spp.dd$temp,
                           spp.dd$energy),
                      mean)
  if(horiz==F) {
    mids <- barplot(mean.week, beside=T,
                    ylab="Extinction time",
                    xlab="Energy",
                    ylim=c(0,9),
                    col=shaydz, horiz=horiz)
    for(i in 1:2)
      for(j in 1:2)
        points(jitter(rep(mids[i,j], 50)),
               spp.dd[spp.dd$energy==colnames(mean.week)[j] &
                        spp.dd$temp==rownames(mean.week)[i],5])
    mtext(side=3, line=1, text=all.species[spp], font=3)
  }
```

```
  if(horiz==T) {
    mids <- barplot(mean.week, beside=T,
                    xlab="Extinction time",
                    ylab="Energy",
                    xlim=c(0,9),
                    col=shaydz, horiz=horiz)
    for(i in 1:2)
      for(j in 1:2)
        points(spp.dd[spp.dd$energy==colnames(mean.week)[j] &
                        spp.dd$temp==rownames(mean.week)[i],5],
               jitter(rep(mids[i,j], 50)))
    mtext(side=3, line=1, text=all.species[spp], font=3)
  }
}

## lets test our function!

## select a species number
ss <- 13 #corresponds to species P.Bursarium

## Plot the extinction times distributions
Plot.data.bars(dd,ss,
               shaydz=grey(c(0.2, 0.8)),
               horiz=T)
```




```
## observe the plot
## what can you tell about the effect of temperature and nutrient concentration (energy)
## on the extinction weeks of the species?

## THINK ABOUT HOW TO PLOT the same disributions for all the 17 species
```

```
## THINK ABOUT HOW TO MAKE THESE PLOTS MORE BEAUTIFUL!
## THINK ABOUT ABOUT OTHER WAYS TO PLOT HISTOGRAMS OF MTE DISTRIBUTIONS
##
##
##
##
##
##
```

## Survival model with exponential distribution

The probability for a species with extinction rate $\lambda$ to survive until time $t$ is given by the exponential

$$S(t, \lambda) = e^{-\lambda t} \tag{1}$$

Thus, the probability of the species to go extinct within time interval $[t_{i-1}, t_i]$ (measured in weeks) is given by

$$p_i(\lambda) = S(t_{i-1}, \lambda) - S(t_i, \lambda) \tag{2}$$

Assume that the species we considered has the same extinction rate in all the $N = 200$ replicates we have. We want to calculate the probability (the likelihood) that the species goes extinct in $y_1$ replicates during the time interval $[t_0, t_1]$, in $y_2$ replicates during the time interval $[t_2, t_3]$ etc. Those replicates where the species survives for the whole experiment (which lasts $n = 8$ weeks) are assumed to go extinct during the interval $[t_n, \infty]$, and this happens with probability $p_n = S(t_n)$. Due to the independence of the extinction process in the different replicates, the counts of the exinction events durig the different time intervals $(y_1, \cdots, y_n)$ are assumed to be independent and therefore the probability for observing such counts is then simply given by the product

$$p_1^{y_1}(\lambda) \cdot p_2^{y_2}(\lambda) \cdot \cdots \cdot p_n^{y_n}(\lambda). \tag{3}$$

Since we do not distinguish replicates, we have to multiply this product by a so-called multinomial coefficient, which counts the number of ways to put $N$ individuals into $n + 1$ buckets, with $n_1$ ending up in the first bucket, $n_2$ in the second etc. Thus, the likelihood function, for model parameter $\lambda$, given an output $\mathbf{y}$ of our survival model, is described by the so called multinomial distribution

$$L_{\text{survival}}(\lambda, \mathbf{y}) = \frac{N!}{y_1! \cdots y_{n+1}!} p_1^{y_1}(\lambda) \cdot p_2^{y_2}(\lambda) \cdot \cdots \cdot p_n^{y_n}(\lambda) \tag{4}$$

which the likelihood of our survival model.

- Try to use another distribution in 1 e.g. a Weibull given by:

$$S(t; \lambda, k) = \frac{k}{\lambda} \left( \frac{t}{\lambda} \right)^{k-1} e^{-(t/\lambda)^k} \tag{5}$$

We will use likelihood function 4 (or more elaborated versions of it) to compute the extinction rate $\lambda$ (or similar related paramaters) for the selected species (or for the whole food web) in a Bayesian framework, using priors on the model parameters. A first, straightforward choice for the prior of $\lambda$ is the univariate lognormal. Note that in JAGS, the lognormal distribution is parameterized using the mean $m$ and and precision $\tau = 1/s^2$ of its normally distributed logarithm. Therefore, to get a lognormally distributed extinction rate $\lambda \sim LN(\mu_\lambda, \sigma_\lambda)$, with mean $\mu_\lambda$ and standard deviation $\sigma_\lambda$, we will use the conversion formulas given by

$$m = log(\mu_\lambda) - \frac{\sigma_\lambda^2}{2}, \quad s = \sqrt{\frac{log(1 + \sigma_\lambda^2)}{\mu_\lambda^2}}, \tag{6}$$

for the mean and standard deviation respectively.

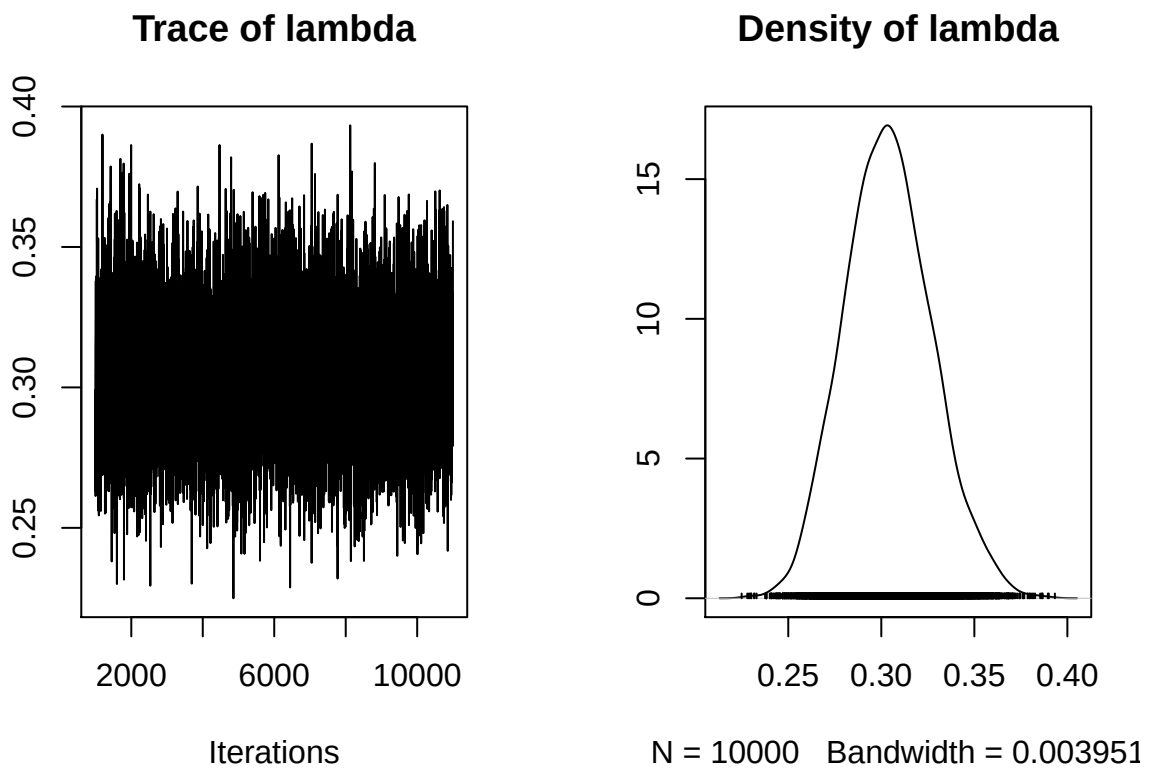
- Think about the meaning of these parameters. What does the two means and standard deviations represent? Make a graphical investigation of these lognormal distributions.
- Do we have any other options for the priors? e.g. a gamma distribution would allow us to write analytical expressions for the posterior, as it is a conjugate prior for the family of exponential distributions.
- Make a prior posterior analysis and compare the different prior choices inlcuding uniform priors (one could also use Jeffrey's priors for an exponential model as it is the most uninformative prior).
- Search the literature and/or ask information to the experts (Owen and Nick) about the biology of the protists and the correspondent parameters (eg growth rates, extinction rates) try to parametrize the priors accordingly.

```
## Let's subset the dataset for the species we considered Paramecium bursaria
spp.dd <- subset(dd, species==all.species[ss])

## define a JAGS object for our model
file.jags.model  <- "survival.jags"

## define the number of iterations of the Markov Chain
sampsize <- 10000

## transform the data to get the counts of the extinction events

data.survival.frame           <- data.frame(matrix(ncol=3,nrow=8,0))
colnames(data.survival.frame) <- c("var","t","data")



#### count the extinction events during the different weeks
ww <- 1
for(ww in 1:8){
  data.survival.frame[ww,"var"] <-paste("Extinction ",ww,sep="")
  data.survival.frame[ww,"t"]   <-ww
  data.survival.frame[ww,"data"]<-length(which(spp.dd$day2==ww))
}

## compile the data

## nuber of replicates
N <- 200

## count of ectinction events
y <- c(data.survival.frame$data,N-sum(data.survival.frame$data))

## time intervals
## note that we set t=1000 for the replicates where the species survived
t <- c(0,data.survival.frame$t,1000)

## write JAGS model definition file:
cat("model {\n",
    "  mulog_lambda <- log(mu_lambda)-0.5*sdlog_lambda^2\n",
    "  sdlog_lambda <- sqrt(log(1+sd_lambda^2/mu_lambda^2))\n",
    "  lambda ~ dlnorm(mulog_lambda,1/sdlog_lambda^2)\n",
```

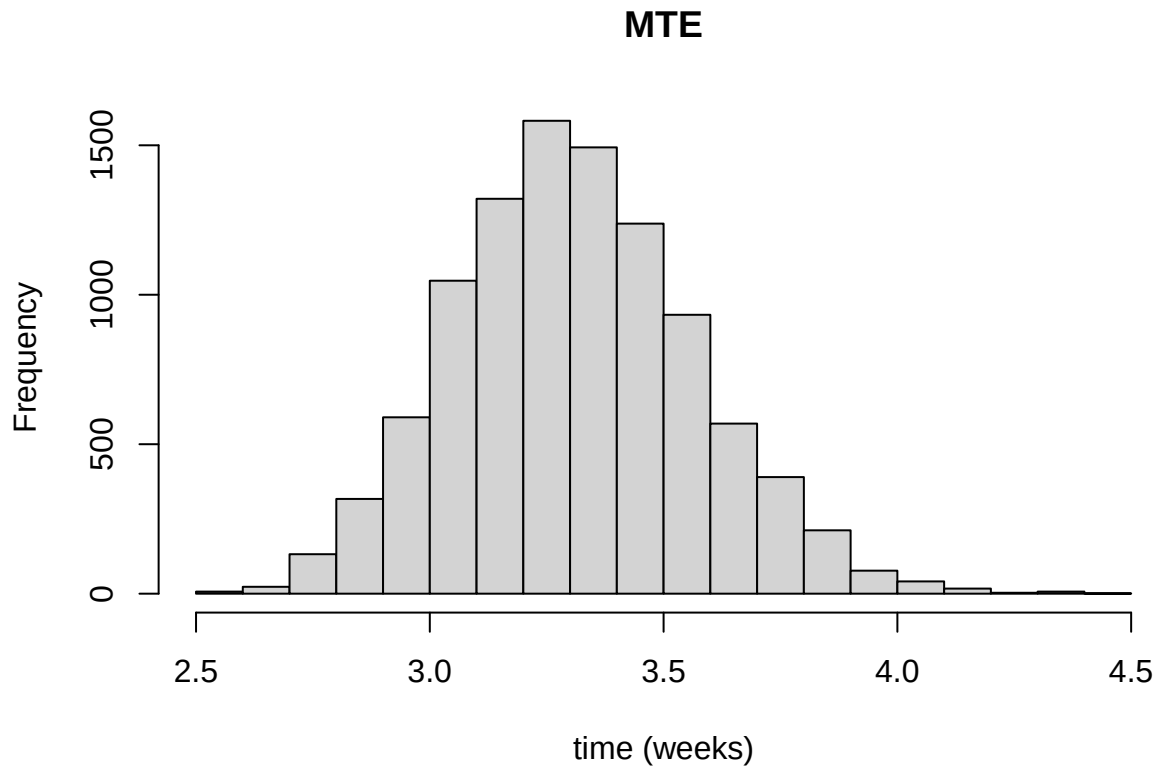
```
    "  for ( i in 1:n ) {\n",
    "    p[i] <- exp(-lambda*t[i])-exp(-lambda*t[i+1])\n",
    "  }\n",
    "  y ~ dmulti(p,N)\n",
    "}\n",
    sep="",
    file=file.jags.model)

### Note how we transformed the mean and sd of the prior.

## run JAGS:
jags.obj <- jags.model(file.jags.model,
                       data=list(mu_lambda = 0.2, ### mean of the prior
                                 sd_lambda = 0.2, ### standard deviation of the prior
                                 n         = nrow(data.survival.frame)+1,
                                 N         = sum(y),
                                 t         = t,
                                 y         = y))
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 1
##    Unobserved stochastic nodes: 1
##    Total graph size: 60
##
## Initializing model
```

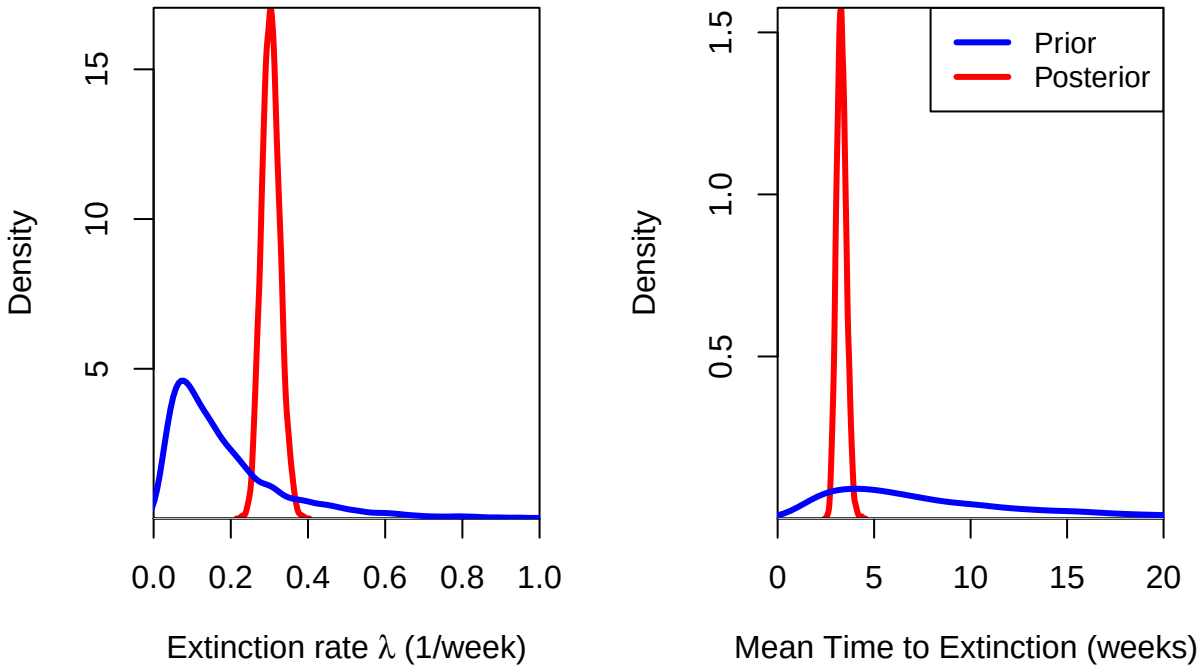
```
## extract and plot results:
jags.res <- coda.samples(jags.obj,c("lambda"),sampsize)
plot(jags.res)
```




```
## compute and plot the distribution of the inferred Mean time to extinction
MTE <- 1/as.matrix(jags.res)
hist(MTE,main="MTE",xlab="time (weeks)")
```



```r
## COMPARE THIS RESULT WITH THE PLOTS OBTAINED BEFORE
## WHAT CAN YOU TELL ABOUT THE OBSERVED EXTINCTION TIME?
## TRY TO CHANGE THE PRIORS PARAMETERS, AND SEE HOW THIS AFFECTS THE MTE.

### first attempt to make a DENSITY PLOT to compare prior and posterior DOUBLE CHECK!!

## conversion functions that give the mean and sd of associated normal (to put into R)
mfun <- function(m,s){log(m)-log(1+s^2/m^2)/2}
sdfun <- function(m,s){sqrt(log(1+(s/m)^2))}

mu_lambda <- 0.2
sd_lambda <- 0.2

### get the mean and sd in log scale
mul_lambda  <- mfun(mu_lambda,sd_lambda)
sdl_lambda  <- sdfun(mu_lambda,sd_lambda)

### sample from R
prior_lambda <- rlnorm(sampsize,mul_lambda,sdl_lambda)
prior_MTE    <- 1/prior_lambda

### posteriros from JAGS output
posterior_lambda <- as.matrix(jags.res)
posterior_MTE    <- 1/posterior_lambda

### some simple plots
par(mfrow=c(1,2))

### prior posterior plots of extinction rates
plot(density(posterior_lambda), col="red",
     lwd=3,lty=1,xaxs="i",yaxs="i",
     main="",
     xlab=expression(paste("Extinction rate ", lambda , " (1/week)",sep="")),
     #xaxt="n",
     #ylim=c(0,MAX_DENSITY),
     xlim=c(0,1)
     )
lines(density(prior_lambda), col="blue",lwd=3,lty=1)


### prior posterior plots of MTE
plot(density(posterior_MTE), col="red",
     lwd=3,lty=1,xaxs="i",yaxs="i",
     main="",
     xlab=expression(paste("Mean Time to Extinction (weeks) ",sep="")),
     #xaxt="n",
     #ylim=c(0,MAX_DENSITY),
     xlim=c(0,20)
     )
lines(density(prior_MTE), col="blue",lwd=3,lty=1)

legend("topright",legend=c("Prior","Posterior"),lwd=3,cex=0.9,col=c("blue","red"))
```




```
#### how can we make these plots more beautiful?
```

- The R package `survival` uses the Kaplan Meier estimator and similar methods to get survival curves. How does this relates to the survival probability defined in 1? check the methods used in the R black boxes and figure out how they are related to our survival model.
- How could we incorporate time dependent covariates? Let's explore Cox proportional hazard models and Landmarking analysis and try to repeat the analysis for the single species.
- How does the result change if we replace the current survival model with a more mechanistic model e.g., Stochastic Lotka Volterrra equations? This could be a very good direction, but probably not feasible for the workshop. In fact, a full mechanistic (and stochastic) model would require much more time and a lot of computational power for calibration and inference.
- How about a Bayesian Hierarchial model then? This could be a good option to imporve the survival model in a semi-mechanistic way. The information about similarity of species could be integrated in the model via overarchicng distributions across trophic groups and/or traits and environmental factors. Think about possible ways to build hierarchies in the survival model, also in order to include the food web structure.
- A realistic aim for the workshop (and for the potential publication) is to build and test a joint likelihood that includes all the data in a meaningful semi-mechanistic way, as well as a robust choice and paramaterization of the priors of the corresponding model parameters.

## References

Cardillo, M. 2005. "Multiple Causes of High Extinction Risk in Large Mammal Species." *Science* 309 (5738): 1239–41. https://doi.org/10.1126/science.1116030.

Dunne, Jennifer A., and Richard J. Williams. 2009. "Cascading Extinctions and Community Collapse in Model Food Webs." *Philosophical Transactions of the Royal Society B: Biological Sciences* 364 (1524): 1711–23. https://doi.org/10.1098/rstb.2008.0219.

Paine, Robert T. 1966. "Food Web Complexity and Species Diversity." *The American Naturalist* 100 (910): 65–75. https://doi.org/10.1086/282400.

Solan, M. 2004. "Extinction and Ecosystem Function in the Marine Benthos." *Science* 306 (5699): 1177–80. https://doi.org/10.1126/science.1103960.

Worsfold, Nicholas. 2007. "The Consequences of Extinction in Experimental Aquatic Communities." Ph.D., University of Sheffield. https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.489745.