Department of Mathematics

Bayesian Statistic Project
Mayfly sensitivity to salt

Written by

GUIDJIME ADINSI Ahouahounko Télesphore
BAH Mamadou Oury
MAGOUJOU Grace

Submitted to : Prof. **KON KAM KING** Guillaume

2022-2023

List of Figures

Contents

March 14, 2023

Introduction

In this report we will be interested in assessing the toxicity of different salts (NaCl, CaCl2, and Commercial) to the mayflies. In cold regions around the world, and specifically in North America, salts are used the de-ice the roads. This, on the upside helps countries avoid their economies coming to a halt whenever there are major blizzards. However, on the downside, once the vast amounts of salt(15million tons every winter according to Vox) dissolves, it washes out to the rivers and streams. This in turn increases salinity in the rivers which can be deadly to freshwater species. Mayfly larvae are a particularly sensitive species. Thus we are going to asses if using different types of salts may be less harmful to the mayflies.

Substance toxicity to a species is usually done by performing bioassays. These are analytical methods used to determine potency of substances by its effect on living species. In our case the species in question are mayflies, and the substances are the salts. The procedure was done by collecting some mayflies in the wild and distributing them into several water tanks. Then pouring a controlled amount of salt in each tank to make sure that the concentration is known, and finally measure the survival after a fixed period of time(96h). This in turn gives us a dataset. Below is a glimpse of it.

A tibble: 99 × 4

N0 <dbl>	Nsurv <dbl>	conc <dbl>	Salt <chr>
9	9	1	NaCl
4	4	2	NaCl
5	5	4	NaCl
6	6	8	NaCl
9	7	16	NaCl
4	4	32	NaCl
9	2	64	NaCl
10	0	128	NaCl
7	0	256	NaCl
7	0	512	NaCl

1-10 of 99 rows Previou

Figure 1: A view of the dataset

As the above figure shows, the dataset contains 4 columns. The first, **N0**, is the number of mayflies in the tank at the beginning. The second column ,**Nsurv**, is the number of mayflies that survived after (96h) at the concentration **conc** of the the salt **salt**. Now to assess the toxicity of each salt to the mayflies, we will first posit a model on the observed phenomena, and then infere the parameters of the model using statistical methods. There are ,in general, two ways of inferring the parameters of the model. The classical method (through maximum likelihood estimation), and the Bayesian method (through the posterior of the parameters). In this report we will follow the latter.

The Bayesian approach is based on Bayes theorem. It involves assigning priors to the parameters (i.e before seeing the data), and then updating this knowledge about the parameters using information from the observed data (through the likelihood function). Which is known as the posterior distribution. Thus to begin the inference we need the model in the form of priors and the likelihood.

1 The Model

First we begin by looking at the proportion of surviving mayflies under each salt plotted against each concentrations. Below is the graph (the x axis is set in log-scale) .

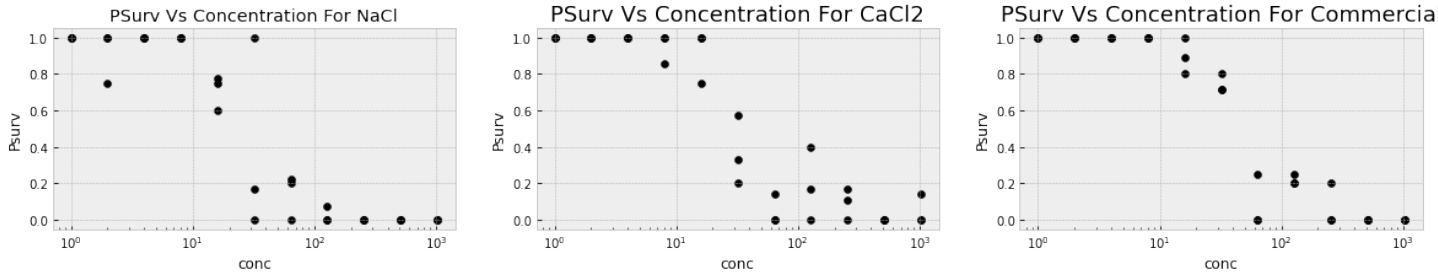

Figure 2: Survival probability under different salts and concentrations

Now as we see above there appears to be a certain pattern showing up in the graphs. This can be modeled by the Log-logistic model. Namely :

$$p_{conc}^{surv} = \frac{d-c}{1+(\frac{conc}{e})^b} + c \ \textit{with}\ e > 0 \ \textit{and}\ b > 0$$

where p_{conc}^{surv} is a function of $conc$ with parameters $d, c, e,$ and b. d is the survival probability at concentration $conc = 0$, c is the survival probability at $conc = \infty$, b is related to slope and e is the inflexion point.
We can begin by looking at the parameters, and since d is the survival probability when $conc = 0$, it is not absurd to assume that all the mayflies are alive at that point (i.e $d = 1$). Moreover, since mayflies are particularly sensitive species, it is not absurd to assume that when $conc = \infty$ that the proportion of living mayflies is 0. We are thus left with a simpler functional form for p_{conc}^{surv}

$$p_{conc}^{surv} = \frac{1}{1+(\frac{conc}{e})^b} \tag{1.0.1}$$

Now let's look at a few functions of p_{conc}^{surv} to get a feeling on the functions.

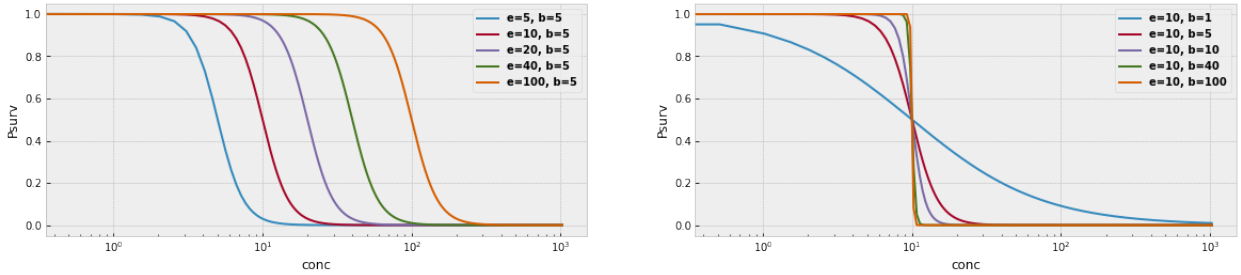

Figure 3: Shape of different survival functions

In figure 3, on the left side we observe that for a fixed b, as e increases the graphs are shifted towards the left side. While on the the right side, we observe that for a fixed e, the corresponding value of e the graphs become steeper as b increases. Moreover, since (it is easy to observe) for $conc = e$ $p^{surv}_{conc} = \frac{1}{2}$, meaning that a $conc = e$ half of the mayflies are dead. We can say that for two given salts the one that has higher toxicity will have a smaller e. Now, that have understood the functional form of p^{suv}_{conc}, we can move on to set the generative model. First we will assume that, at the concentration $conc$, after (96h) each of the mayflies survives with a probability p^{surv}_{conc}, and we call the total number of mayflies that survived N^{surv}_{conc}. Therefore, we have :

$$N^{surv}_{conc} \sim \mathcal{B}(N^{0}_{conc},\, p^{surv}_{conc}) \tag{1.0.2}$$

As noted in the introduction, a Bayesian model requires priors to be put on the unknowns(parameters), and since both e and b are positive, we need priors that have support on the positive reals. There are many distributions that satisfy this criterion. However since the log's of these values are on the real line we can put the priors on the log's of the parameters. In our case we will use the following priors:

- $\log e \sim \mathcal{U}(\log(1),\, \log(1000))$
- $\log b \sim \mathcal{T}(df = 2,\, \mu = 0,\, \sigma = 2)$

To justify the choice of these priors let's first see the information they carry by looking at the graphs of their pdf's

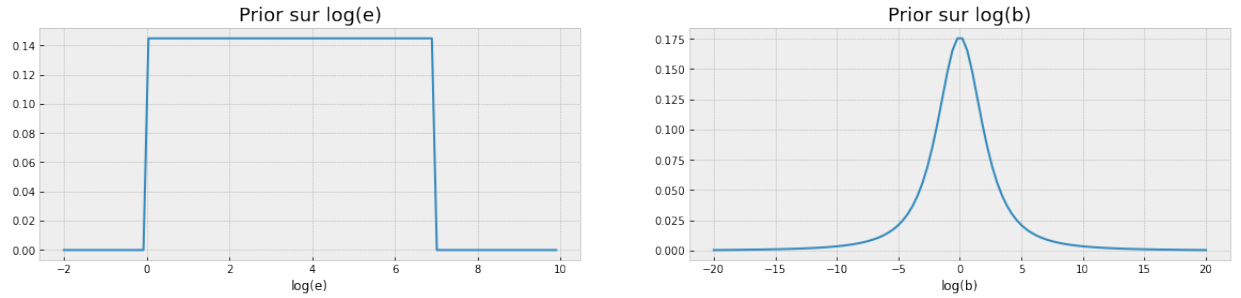

Figure 4: Priors of log(e) and log(b)

The Uniform distribution on the interval $[\log(1), \log(1000)]$ is wide enough to cover most of the used concentration in our bioassay, and also inspired by what we have seen in 3 this allows for a wide range of p^{surv}_{conc} values. Moreover, the Uniform is also less informative(since we are saying that every value in the range is equally likely). On the other hand the student distribution $\mathcal{T}(df = 2,\, \mu = 0,\, \sigma = 2)$ as seen in figure 4 ,on the right hand side, has "probable" values in $[-5,\, 5]$. This translates to values of b in $[\exp(-5),\, \exp(5)]$.Which is somewhere between 0.007 and 148. Moreover, as seen in figure 3(right hand side). This covers a wide range of behaviors.

2 Implementation and Fake Data Check

Doing the inference in a Bayesian modeling problem is straight forward using probabilistic programming language. In our report we use RStan, and also use the python programming language for some visualization tasks. All code will be provided with the report. To assess the implementation and the convergence of the model, we will do a fake data check (parameter recovery). This procedure involves generating fake data from the model with known parameter, and running the inference algorithm to see if recover the parameter that generated the data.
To proceed we simulated four datasets from the model using ($e = 20.4$, $b = 2.01$), ($e = 8.78$, $b = 4.47$), ($e = 45.3$, $b = 1.44$), and ($e = 70.4$, $b = 3.5$), and did the inference to find posterior distribution of the parameters. In the figures below we give the posteriors in blue and the true parameters in red.

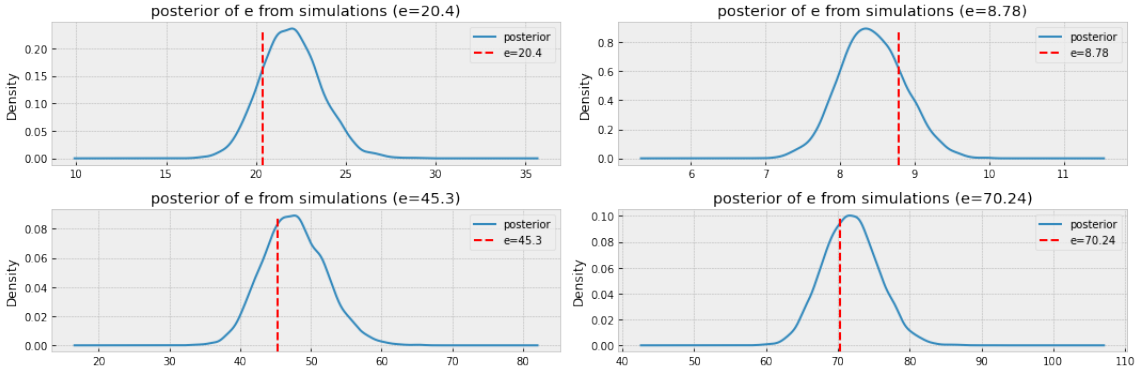

Figure 5: Posteriors of e from simulations

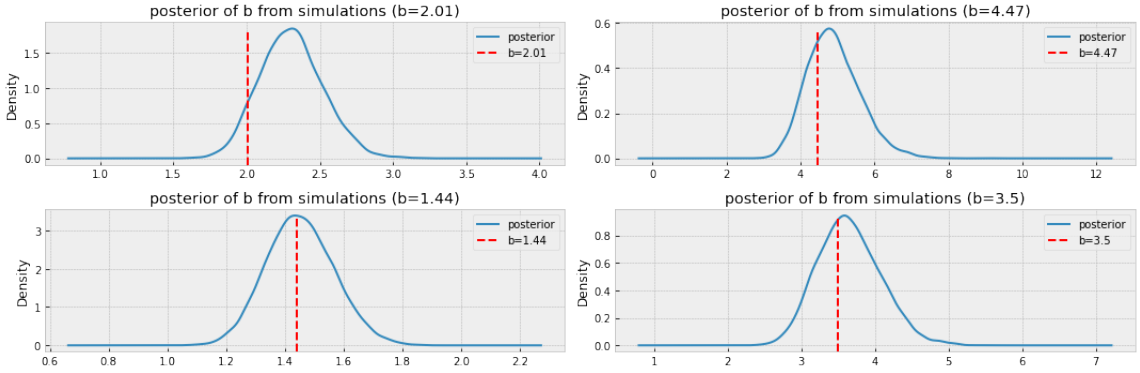

Figure 6: Posteriors of b from simulations

After checking for model convergence (Using $\hat{R}$ statistic, which will be discussed later), we see that we are able to recover the parameter fairly moderately. We therefore move forward and implement the model. In our implementation we assume that each salt has its given e, and b values. Meaning e_1, *and* b_1 for NaCl, e_2, *and* b_2 for CaCl2, and e_3, *and* b_3 for Commercial salt. The prior used were the same as those described before, and we went through with the fitting of the model using Stan.

3 MCMC Convergence Check

The model was fitted using MCMC sampling in R. We chose 4 chains with *iter* = 5000 half of which will be used as warmups(these will be discarded). This is done because we know that at the beginning of the sampling process the chain will be sampling under a different distribution before converging to the posterior distribution, and to also diminish the influence of the starting values [1]. Moreover, using 4 different chains will help us in checking if eventually all chains are sampling from the same distribution. Below is the graph of the traceplot.

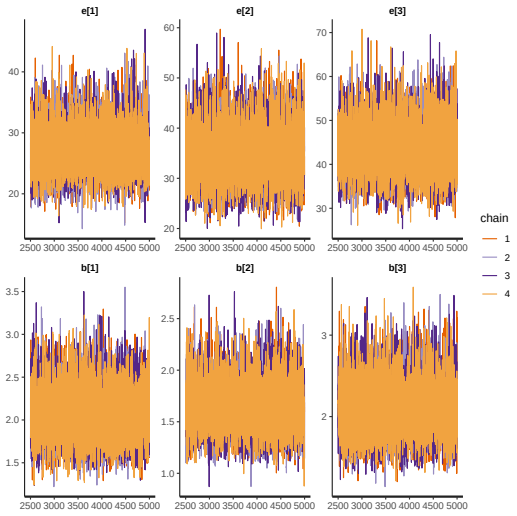

Figure 7: MCMC traceplot

In the above figure, we observe for all 6 graphs on the x-axis we have the iterations (begins after 2500 since the first 2500 were discarded), and on the $y-$ axis we have the value of the corresponding sample at the given iteration. Each color corresponds to a chain. We see that all chains overlay nicely. Moreover, there appears to be no "correlation" structure. We also checked the $\hat{R}$ Statistic which is a way of assessing mixing using between- and within-sequence variances[1]. Gelman et al advises for a value of $\hat{R}$ in the neighborhood of 1. Which is what we observed. Furthermore, to avoid cluttering the report, the mathematical explanations are left out, and the reader can refer to this well explained source [2].

	Rhat
log_e[1]	1
log_e[2]	1
log_e[3]	1
log_b[1]	1
log_b[2]	1
log_b[3]	1
e[1]	1
e[2]	1
e[3]	1
b[1]	1
b[2]	1
b[3]	1
lp__	1

Figure 8: Rhat

4 Model Evaluation:

Now that we have check for convergence, and have confirmed that samples are being drawn from the same target distribution, we can proceed in assessing the model. First, let's look at what has been "learned" from the data by comparing the posteriors and the priors.

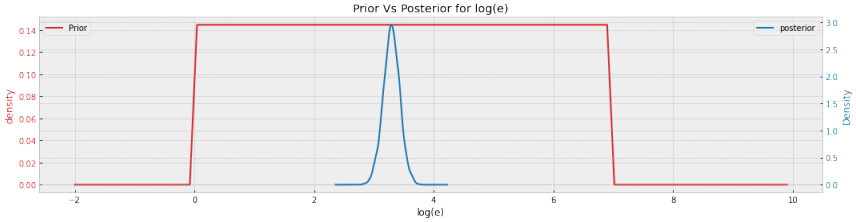

Figure 9: Prior vs posterior of log(e)

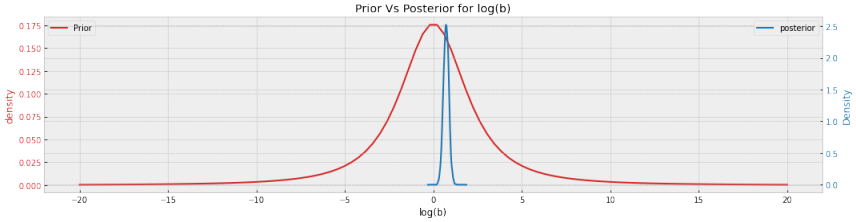

Figure 10: Prior vs posterior of log(b)

In the above two figures we clearly see that the priors on the parameters have gotten narrower through the posterior. This shows that from the information contained in the data, the sampling algorithm has selected the most probable parameters (seen by the smaller support of the posterior). Furthermore, as explained in section 1, there was a certain pattern observed in figure 1 which we modeled by Eq(1.0.1) . Now let's also look at the posterior patterns after having observed the data.

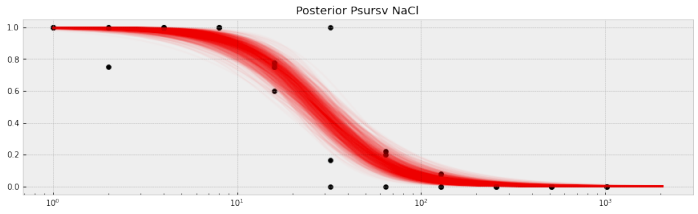

Figure 11: Posterior survival probability under NaCl

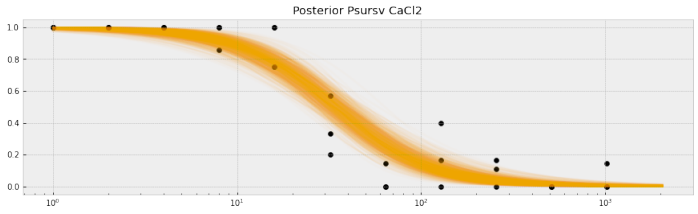

Figure 12: Posterior survival probability under CaCl2

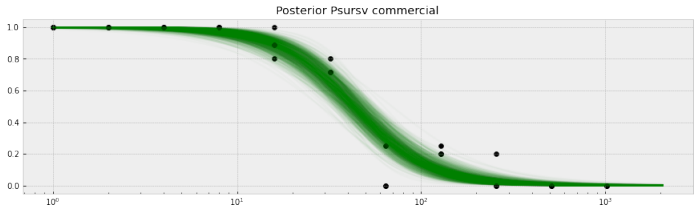

Figure 13: Posterior survival probability under commercial salt

The graphs above were obtained by using the values of e and b sampled from their posteriors to compute the corresponding survival probabilities (p_{conc}^{surv}) for each salt. The values on the $x-$axis range between 1 and 2048 and the axis is in log scale. These possible functions graphically model the observed survival proportions fairly well.

In addition to doing the above check, we also predicted the number of mayflies that survive at each given concentration for every salt using the posterior predictive. In other words we did predictions for our data using the model. We then computed two statistics (for each salt). The first one was the mean number of mayflies that survived, and the second was $\frac{\sum_{i=1}^{33} \mathbb{1}\{N_{salt}^{surv}=1\}}{\sum_{i=1}^{33} \mathbb{1}\{N^{surv}=0\}}$, and we then graphed the comparisons below

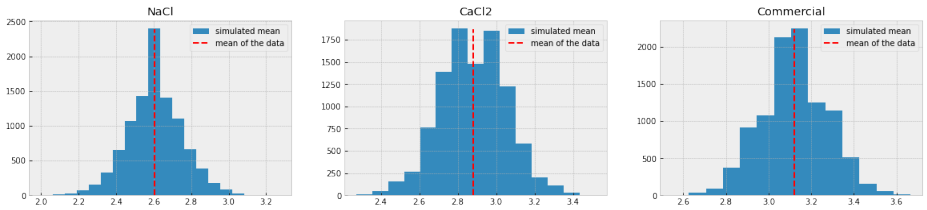

Figure 14: Goodness of fit assessment (1)

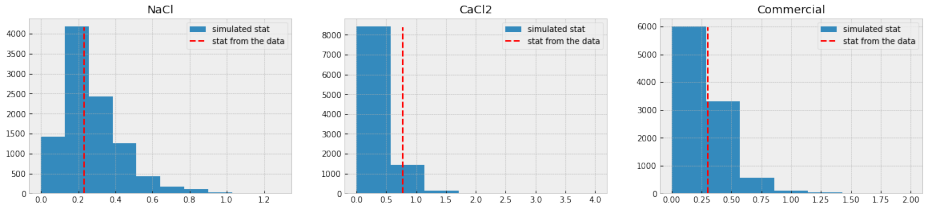

Figure 15: Goodness of fit assessment (2)

and as we see in the graph the statistics are well covered by the histograms.

5 Toxicity of the Salts

Now that we are familiarized with the dataset, implemented the model, and checked for the model validity/evaluation, we are now ready to attack the interpretation. As observed in the introduction and the description of the model, the parameter e is the one that tells the most about the toxicity. We will thus compare the differences of the e's of the three salts. First, let's take a look at the posteriors

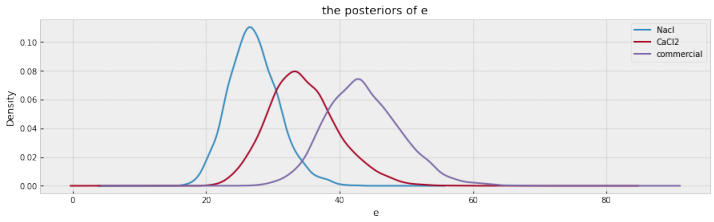

Figure 16: Posteriors of e of different salts

We see that there is great overlap between e_{NaCl} and e_{CaCl2}. Which is also the same for $e_{Commercial}$ and e_{CaCl2}. However, There is almost no overlap between $e_{Commercial}$ and e_{NaCl} except in the tails. Moreover, to test this we can run statistical test. Below is a histogram of the differences.

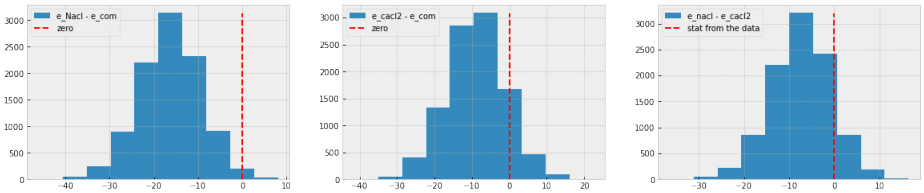

Figure 17: Toxicity comparison between salts

In addition we computed the following posterior probabilities and found that : $\mathbb{P}(e_{NaCl} - e_{NaCl} > 0) = 0.0091$, $\mathbb{P}(e_{CaCl2} - e_{comme} > 0) = 0.1245$, and $\mathbb{P}(e_{CaCl2} - e_{comme} > 0) = 0.1254$. Thus we can say that switching salt from NaCl to Commercial will the most helpful for the mayflies.

6 Conclusion

In this report, we began by looking at the dataset of the mayflies. We then proceeded to look at way of modeling the data generating process. Moreover, we used Bayesian data analysis to infere the parameters of the model. Relating the parameters of the model to the toxicity, we came to conclude that the $NaCl$ salt appears to be more toxic to the mayflies than the commercial salt.

References

[1] Gelman, Andrew, et al. Bayesian data analysis. Chapman and Hall/CRC, 1995..

[2] *Monitoring Convergence*, https://bookdown.org/rdpeng/advstatcomp/monitoring-convergence.html.