# MATH35600 Assessed Practical 2021

**You should work in groups of 3 for this assignment, handing in one report at the end. It is worth 20% of the marks for this unit.**

## Setup

The "frogs" data frame can be loaded with the command:

```
load(file = url("https://mfasiolo.github.io/TOI/frogs.RData"))
head(frogs)

##   size killed
## 1    8      0
## 2    9      2
## 3   10      3
## 4   11      3
## 5   12      6
## 6   13      5
```

Each row contains data from an experiment where 10 *toadis uglibus* frogs of total body length equal to `size` centimeters have been placed in a large enclosed pond. The variable `killed` reports how many of the 10 frogs were killed by predators within three days of the start of the experiment. We are interested in verifying how the size of the frogs effects the probability of being killed by predators. In particular, we consider two different models for the probability of being killed. The first is a modified logistic function

$$P_L(s) = \frac{e^{\varepsilon(\phi - s)}}{1 + e^{\beta\varepsilon(\phi - s)}}, \tag{0.1}$$

where $s$ is the size of the frog. Here $\phi$ is a location parameter such that $P(s = \phi) = 0.5$, $\varepsilon$ controls the rate of change of $P(s)$ with $s$ and $\beta$ controls the asymmetry of the function. An alternative model is the generalized Ricker model

$$P_R(s) = b\left\{\frac{s}{a}\exp\left(1 - \frac{s}{a}\right)\right\}^{\alpha}, \tag{0.2}$$

where $a$, $b$ and $\alpha$ are model parameters. We assume that the number of frogs killed in the $i$-th experiment, $\text{kill}_i$, follows a binomial distribution

$$\text{kill}_i \sim \text{binom}(p_i, n = 10),$$

where $p_i = P(s_i)$ is modelled either via $P_L(s_i)$ or $P_R(s_i)$.

You should do the following, obviously checking results as you go to make sure the answers are sensible.

- Write two functions implementing models (0.1) and (0.2). Write also two functions that evaluate the negative log-likelihood under the two models, using the binomial model given above for $\text{kill}_i$.
- Minimize the negative log-likelihood of the data under both models, using `optim`. Set arguments `method = "BFGS"` and `hessian = TRUE` to store the Hessian at the minimizer. Good initial values for optimizing the parameters of model (0.1) are $\varepsilon = -0.25$, $\beta = 5$ and $\phi = 13$, while for the generalized Ricker model you can use $a = 8$, $b = 0.25$ and $\alpha = 3$. When fitting the Ricker model, you might get warnings saying "`NaNs produced`" but it should be safe to ignore them.
- Visually compare the fitted probabilities $\hat{P}_L(s)$ and $\hat{P}_R(s)$ with the data and discuss which model seems to be fitting the data better. Compare the models in terms of AIC.
- Certain experts on frogs behaviour claim that the probability of predation should monotonically increase with size, because finding good places to hide is more difficult for large frogs. Under model (0.1) and assuming that $\varepsilon < 0$, this happens when $\beta = 1$. Perform a statistical test to assess whether the value $\beta = 1$ is compatible with the data (recall that the Hessian of the negative log-likelihood at its minimiser can be found in the `$hessian` slot of the output of `optim`). **Hint**: do not use a GLRT test here.

- Frogs of specie *toadis uglibus* are currently endangered, hence conservationist are planning to capture some of them in the wild and to keep them in captivity for the duration of the critical period when they are most vulnerable (i.e., when they are too large to hide easily, but still too small to discourage some of their predators). Given that the project budget is limited, they would like to focus on the group of frogs that have the highest probability of being killed, given their size. Provide them with some guidance by:
  - finding the value of $s^*$ that maximises $P_L(s)$. This is, of course, a function $g$ of the model parameters, that is $s^* = g(\boldsymbol{\theta})$ where $\boldsymbol{\theta} = \{\varepsilon, \beta, \phi\}$. You should find $g$ analytically, not numerically.
  - Find the size at which the predation is maximal by plugging the estimated values of the model parameters into the function $g$ you have calculated to obtain $\hat{s}^* = g(\hat{\boldsymbol{\theta}})$.
  - To understand how much they can trust your estimates, conservationists would like to know what is the uncertainty of the estimated size $\hat{s}^*$ you just obtained. Recall that the asymptotic distribution of a linear function of $\hat{\boldsymbol{\theta}}$, such as $\boldsymbol{A}\hat{\boldsymbol{\theta}}$ where $\boldsymbol{A}$ is a known matrix, is simply $N\{\boldsymbol{A}\hat{\boldsymbol{\theta}}, \text{cov}(\boldsymbol{A}\hat{\boldsymbol{\theta}})\}$ where the $\text{cov}(\boldsymbol{A}\hat{\boldsymbol{\theta}}) = \boldsymbol{A}\text{cov}(\hat{\boldsymbol{\theta}})\boldsymbol{A}^T$ and the covariance matrix $\text{cov}(\hat{\boldsymbol{\theta}})$ can be estimated consistently using the Hessian matrix of the negative log-likelihood. Unfortunately $g$ is non-linear, but we can adopt the following linear approximation

    $$\hat{s}^* = g(\hat{\boldsymbol{\theta}}) \approx g(\boldsymbol{\theta}) + \nabla g(\boldsymbol{\theta})^T (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \tag{0.3}$$

    where $\nabla g(\boldsymbol{\theta})$ is the gradient of $g$. Under this approximation (generally called the "delta method") we have that $\mathbb{E}\{g(\hat{\boldsymbol{\theta}})\} = g(\boldsymbol{\theta})$ and

    $$\text{var}(\hat{s}^*) = \nabla g(\boldsymbol{\theta})^T \text{cov}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \nabla g(\boldsymbol{\theta}). \tag{0.4}$$

    Use (0.3) and (0.4) to derive an approximate confidence interval for $s^*$ (note that $\nabla g(\boldsymbol{\theta})$ in (0.4) can be estimated consistently by $\nabla g(\hat{\boldsymbol{\theta}})$).

## What to hand in

You should write, as a group, a concise report of no more than 5 sides of A4 (normal margins $\geq$ 10pt font). The report should be accompanied by an appendix containing well structured, clearly commented R code for performing the analysis. The report should start with a title and the names of all the group participants. The report should be suitable for a statistician, explaining the analysis, allowing them to understand what you have done, what you have concluded and why. You should not assume that they have seen this sheet of instructions, so make sure the work is introduced sensibly. There should be enough detail for a statistical reader to judge the appropriateness of the approach. The report should include appropriate plots. The main body of the report should ideally include no R code (technicalities should be explained with maths, if necessary).

The report should be submitted via the course's BlackBoard page (see "Assessment, submission and feedback" and then "Project") by 12 noon, on the 3rd of May.

## Mark scheme guidance

First class marks will be awarded for work that could be passed on to statistically literate scientists interested in these data, essentially without modification. That is to say the statistics is appropriate and clearly explained, the conclusions appropriately drawn and any limitations are discussed fairly.

Upper second class marks will be awarded for work that could be passed on to the scientists after a round of revision correcting some errors of presentation, interpretation or statistics that are relatively minor.

Lower second class marks will be awarded to work that has some more substantial flaws of presentation, interpretation or statistical reasoning which would require some more work to correct.

Third class marks will be awarded for work that contains some indication of substantive understanding and engagement, but contains more serious errors and misunderstandings.