Accounting for commercial data without reallocationg the catches.

MP Etienne

2021-02-19

Contents

Chapter 1

Model presentation

1.1 Log Normal parametrization

Many different parametrizations of the log normal exist, We have to choose the most convenient for our purpose

1.1.1 Version 1

$$D \sim \mathcal{LN}(\rho, \sigma^2) \iff D = e^{\rho+\sigma N},\ N \sim \mathcal{N}(0,1). \tag{1.1}$$

In this parametrization

$$E(D) = e^{\rho+\frac{\sigma^2}{2}}, \quad Var(D) = \left(e^{\sigma^2} - 1\right) e^{2\rho+\sigma^2}$$

1.1.2 Version 2

$$D \sim \mathcal{LN}(\nu, \sigma^2) \iff D = \mu e^{\sigma N},\ N \sim \mathcal{N}(0,1). \tag{1.2}$$

In this parametrization

$$E(D) = \nu e^{\frac{\sigma^2}{2}}, \quad Var(D) = \nu^2 \left(e^{\sigma^2} - 1\right) e^{\sigma^2}$$

1.1.3 Version 3, the mean of the log normal is the actual parameter

$$D \sim \mathcal{LN}(\mu, \sigma^2) \iff D = \mu e^{\sigma N-\frac{\sigma^2}{2}},\ N \sim \mathcal{N}(0,1). \tag{1.3}$$

In this parametrization

$$E(D) = \mu, \quad Var(D) = \mu^2 \left(e^{\sigma^2} - 1\right).$$

With this parametrization,

$$D \sim \mathcal{LN}\left(E(D), \ln\left\{Var(D)/E(D)^2 + 1\right\}\right).$$

We will use this last parametrization throughout the paper.

1.2 Model

The model proposed by Baptiste has a hierarchical structure with three main components

- the biomass latent field S, representing the spatial repartition of biomass:

$$S(x) = \exp\left\{\alpha_S + \Gamma_S(x)^\top \beta_S + \delta(x)\right\} \tag{1.4}$$

- the preferential sampling density, defined conditionnally to S:

$$\lambda(x) = \exp\left\{\alpha_X + b\log\left\{S(x)\right\} + \Gamma_X(x)^\top \beta_X + \eta(x)\right\} \tag{1.5}$$

- and finally, the observations are defined at the VMS ping scale. Y_i stands for the CPUE at location x_i and is defined conditionnaly on the Poisson process and the biomass latent field S:

$$\begin{aligned} Y_i &= C_i Z_i, \\ C_i &\sim \mathcal{B}(1-p_i), \text{ with } p_i = \exp\left\{-e^{\xi\mu(x_i)}\right\}, \\ Z_i &\sim \mathcal{LN}(\frac{\mu(x_i)}{1-p_i}, \sigma^2), \\ \mu(x_i) &= qS(x_i). \end{aligned}$$

1.3 Potential limitation

The information regarding the catch is not known at the fishing point level (VMS ping) but at the statistical unit, as recorded in the logbook. The available information on the commercial catch is the declaration D_j at the statistical unit level and the VMS ping location $(x_{j1}, ... x_{jn_j})$. D_j is the sum of all individuals catch at the n_j visited fishing points:

$$D_j = \sum_{i \in \mathcal{P}_j} Y_i,$$

$\mathcal{P}_j$ being the vector of all fishing events associated with the logbook entry j.

This diffculty is currently addressed by reallocating the total catch to individual fishing events with a simple proportionality rule:

$$\tilde{Y}_i = \frac{1}{|\mathcal{P}_j|} D_j.$$

This is not completely satisfactory as it might introduce some bias is the total biomass estimation as well as its spatial repetition.

1.4 Is is a real limitation ?

A strategy to answer this question would be

1. Simulation study to understand when this reallocation is problematic
2. A new observation model at the statitical unit level
3. Does this new model improves the estimation ?

Chapter 2

A simulation study to explore the effects of the reallocation

Put the stats project here

Chapter 3

A new model for commrecial data

As mentionned in the introduction, every commercial catch is reported by statistical area/vessel/gear and correspond potentially to several VMS ping. Let's denote D_j a reported catch associated to $|\mathcal{P}_j|$ VMS ping at positions $\left\{x_i, i \in \mathcal{P}_j\right\}$. D_j is the sum of all individuals catch at the n_j visited fishing points:

$$D_j = \sum_{i \in C_j} Y_i, \tag{3.1}$$

C_j begin the vector of all fishing events associated with the logbook entry j.

This diffculty is currently addressed by reallocating the total catch to individual fishing events with a simple proportionality rule:

$$\tilde{Y}_i = \frac{1}{|C_j|} D_j.$$

$\tilde{Y}_i$ is the assumed to follow the observation model defined (3.1).

A more satisfying solution would consist in deriving the distribution of D_j from the distribution of Y_i. However the sum of a mixture of delta lognormal random variables do not resume in a nice known distribution.

An alternative would be to fit a delta lognormal model whose proportion of zero and the two first moments coincide i.e

$$\begin{aligned} P(D_j = 0|S, X) &= \prod_{i \in \mathcal{P}_j} P(Y_i = 0|S, X), \\ &= \exp\left\{ - \sum_{i \in \mathcal{P}_j} e^{\xi S(x_i)} \right\}. \end{aligned} \tag{3.2}$$

The expected catch biomass is defined by $E(D_j) = \sum_{i \in \mathcal{P}_j} E(C_i Z_i)$.

As C_i and Z_i are assumed to be independant

$$\begin{aligned} E(D_j) &= \sum_{i \in \mathcal{P}_j} E(C_i Z_i), \\ &= \sum_{i \in \mathcal{P}_j} (1 - p_i) \frac{\mu(x_i)}{1 - p_i}, \\ &= \sum_{i \in \mathcal{P}_j} \mu(x_i). \end{aligned} \tag{3.3}$$

Regarding the variance

$$Var(D_j) = \sum_{i \in \mathcal{P}_j} Var(C_i Z_i).$$

We are the lead to compute the Variance of $C_i Z_i$.

$$\begin{aligned} Var(C_i Z_i) &= E(var(C_i Z_i | C_i)) + Var(E(C_i Z_i | C_i)), \\ &= Var(Z_i) E(C_i^2) + Var(E(Z_i) C_i), \\ &= Var(Z_i)(1 - p_i) + E(Z_i)^2 p_i (1 - p_i) \end{aligned}$$

3.1 Numerical illustration

3.1.1 Precising the notation

As mentioned in Equation (1.3), $Y \sim \mathcal{LN}(\mu, \sigma^2)$ stands a lognormal distribution such that $E(Y) = \mu$ and $Var(log(Y)) = \sigma^2$.

An alterntive commun specificication denotes $Y \sim LN(\rho, \sigma)$ if $E(log(Y)) = \rho$ and $Var(log(Y)) = \sigma^2$

Look at the the sum of n lognormal distribution

The code below is an helper to go back and forth between the different parametrizations.

```
log_normal_variance2 <- function(rho, sigma){
  ( exp(sigma^2) - 1 ) * exp(2* rho  + sigma^2)
}

LN_rho2mu <- function(rho, sigma){
   exp(rho+ sigma^2 /2 )
}

LN_mu2rho <- function(mu, sigma){
   log(mu) - sigma^2 /2
}

log_normal_variance1 <- function(mu, sigma){
  log_normal_variance2(LN_mu2rho(mu, sigma), sigma)
}

log_normal_mean2 <- function(rho, sigma){
   exp( rho + sigma^2/2 )
}
```

3.1.2 n = 2

```
set.seed(1234)
n <- 2
rho <- rnorm(n, mean = 0, sd = 1 )
sigma <- 01
```

We are ready to look at the sum of 2 lognormal random variables with mean 0.493, 2.176 and variance 0.418, 8.135 .

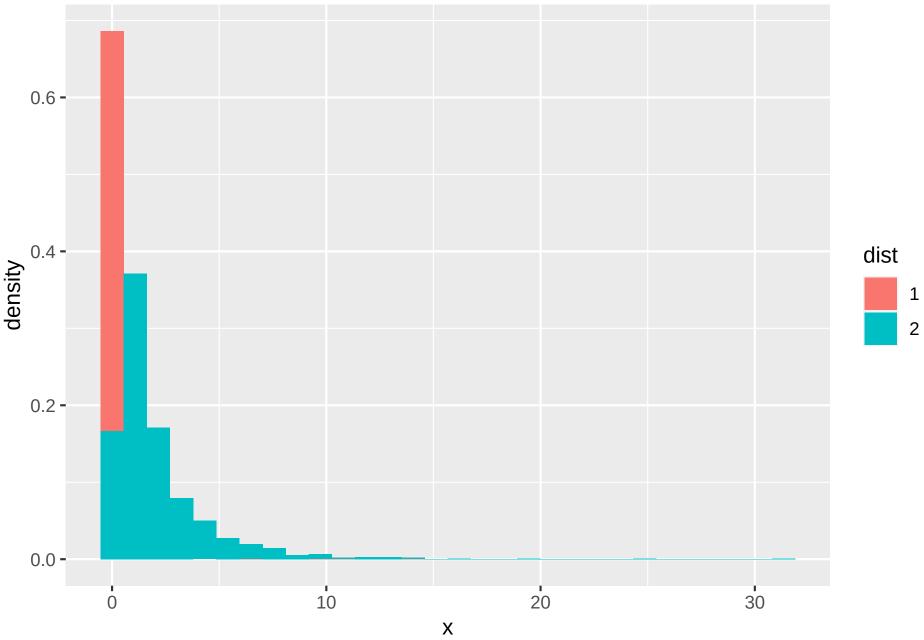
```
n_sim <- 1000
dta <- rho %>% map(function(x){ exp(rnorm(n = n_sim, mean = x, sd = sigma))}) %>%
   map_df(
         ~ data_frame(x = .x),
         .id = "dist"
     )
```

```
## Warning: `data_frame()` is deprecated as of tibble 1.1.0.
## Please use `tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
p <- dta %>% ggplot() + geom_histogram(aes(x=x, y=..density.., fill = dist), position="identity",

p
```

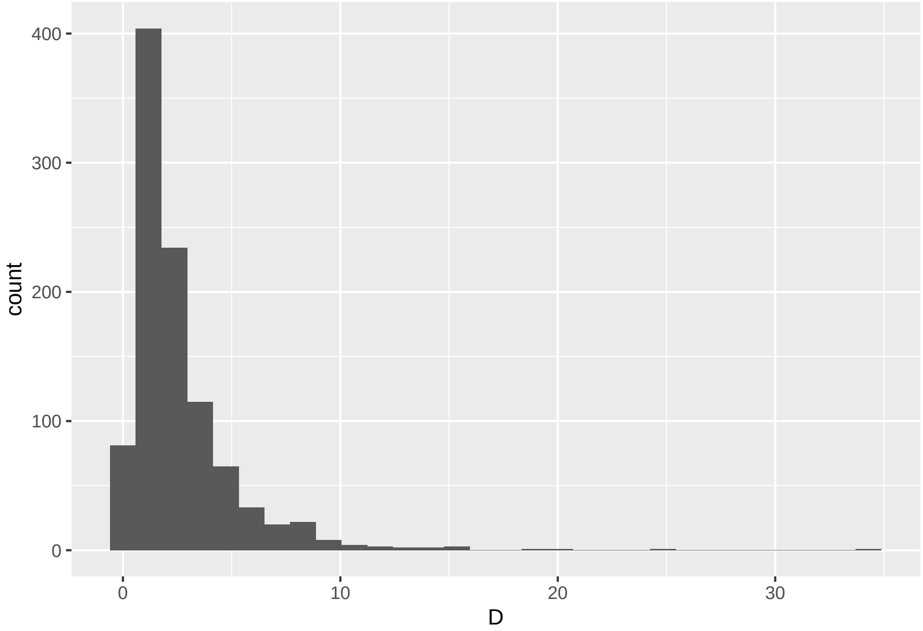
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
sum_dta <- dta %>% group_by(dist) %>%
  mutate(id = row_number()) %>%
  pivot_wider(names_from = dist,    names_glue = "dist_{dist}",
 values_from =x) %>%
  mutate( D = rowSums(.[grep("dist_", names(.))]))

sum_dta %>% ggplot() + geom_histogram(aes(x=D), position="identity")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

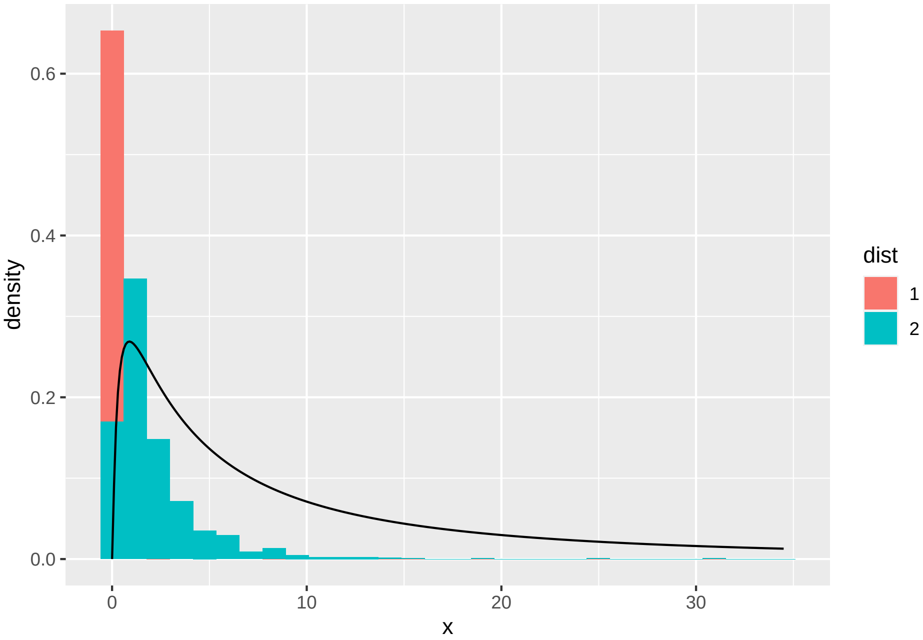


```
mu_sum <- sum(LN_rho2mu(rho, sigma))
S2_sum <- sum(log_normal_variance2(rho, sigma))
sigma_sum = sqrt( S2_sum/mu_sum^2 +1)

sum_density <- tibble(x  = seq(0.001, max(sum_dta$D), 0.1)) %>%
  mutate(log_x = log(x),
         dens = dnorm(log_x,
                      mean=LN_mu2rho(mu_sum, sigma_sum),
                      sd = sigma_sum))

p + geom_line(data= sum_density, aes(x=x, y =dens))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



La moyenne théorique attendue est 2.6689604 et la moyenne empirique vaut 2.6173234, la variance théorique vaut 57.2084894 et la variance empirique vaut 7.0105572.