Elements of statistical inference for Markov Chain models in Biology

José Miguel Ponciano: josemi@ufl.edu
University of Florida, Biology Department

Acknowledgements

- NIMBiOS for hosting this tutorial and taking care of travel and taking care of lodging.
- Linda Allen for the invitation
- University of Florida, Biology department for funds for travel.

Motivation

Gymnogyps californianus

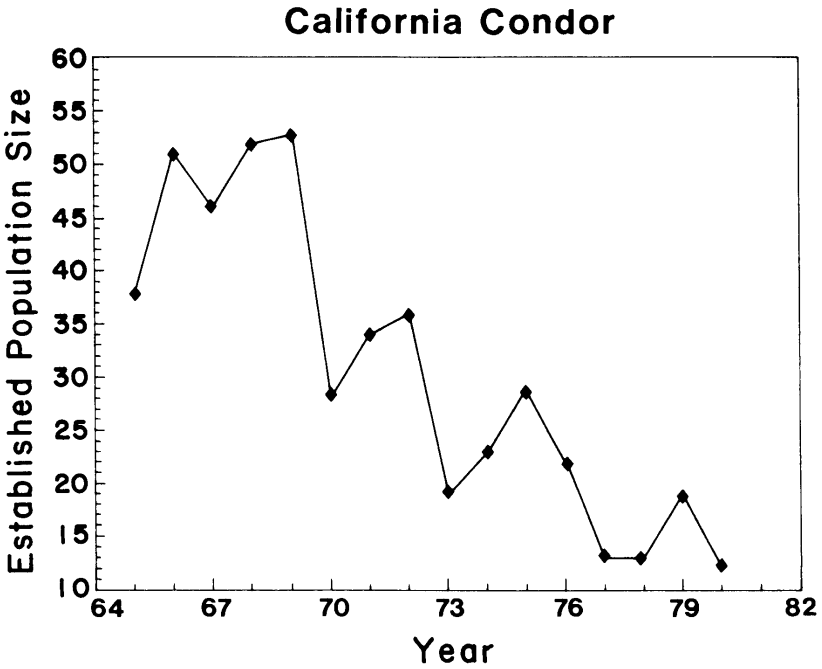


FIG. 7. Estimated total wild population of the California Condor, 1965–1980. Data are from October surveys as listed by Wilbur (1980) and Snyder and Johnson (1985).

Dendroica kirtlandii

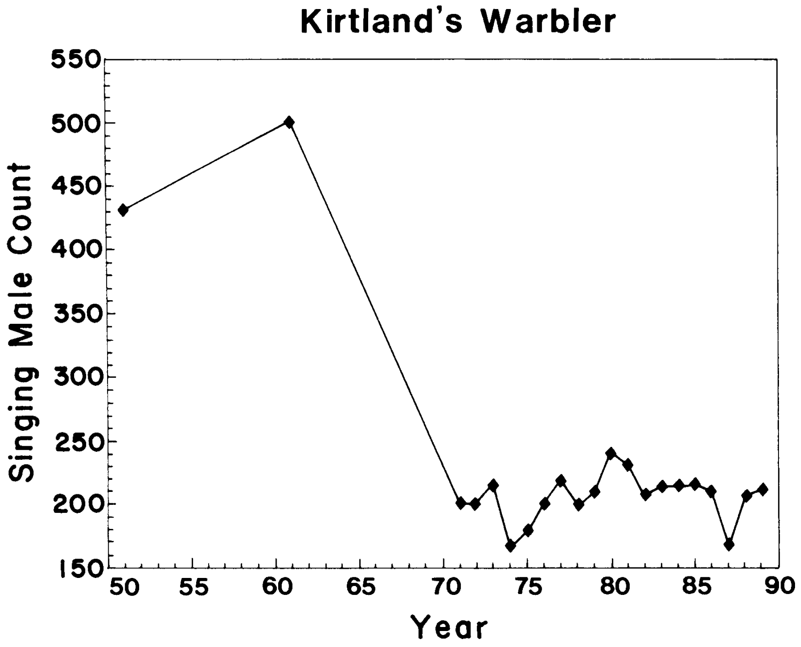
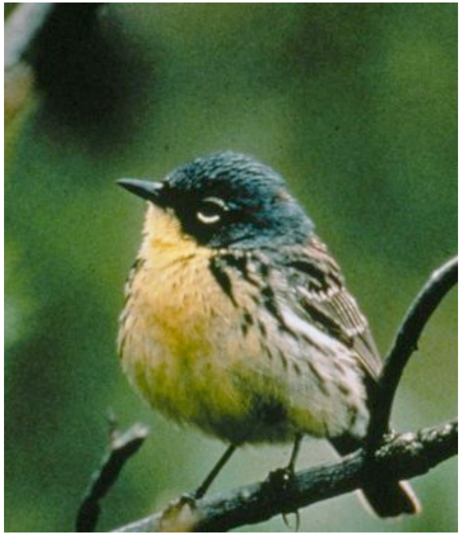


FIG. 6. Total count of Kirtland's Warbler singing males, 1951–1989. Data are from Walkinshaw (1983), supplemented by more recent counts.



Ursus arctos

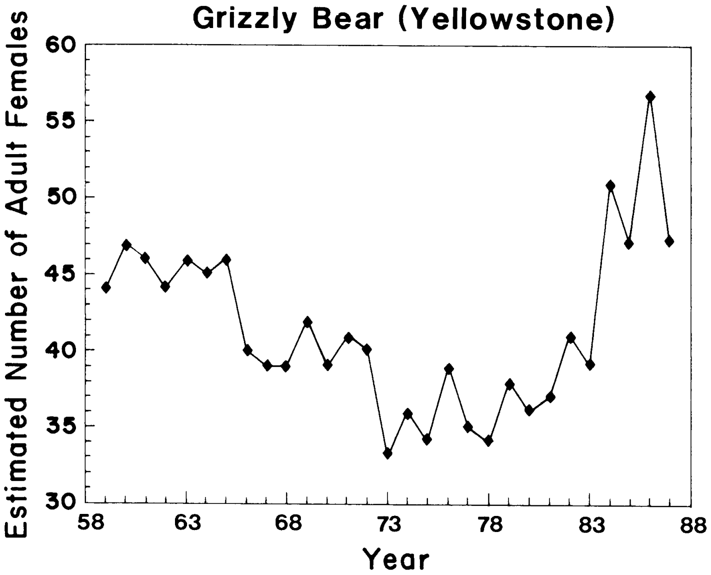


FIG. 5. Estimated number of adult females in the Yellowstone National Park grizzly bear population, 1959–1987. Data, listed by Eberhardt et al. (1986) and supplemented by recent figures, consist of a 3-yr moving sum of the yearly number of adult females seen with cubs.

Grus americana

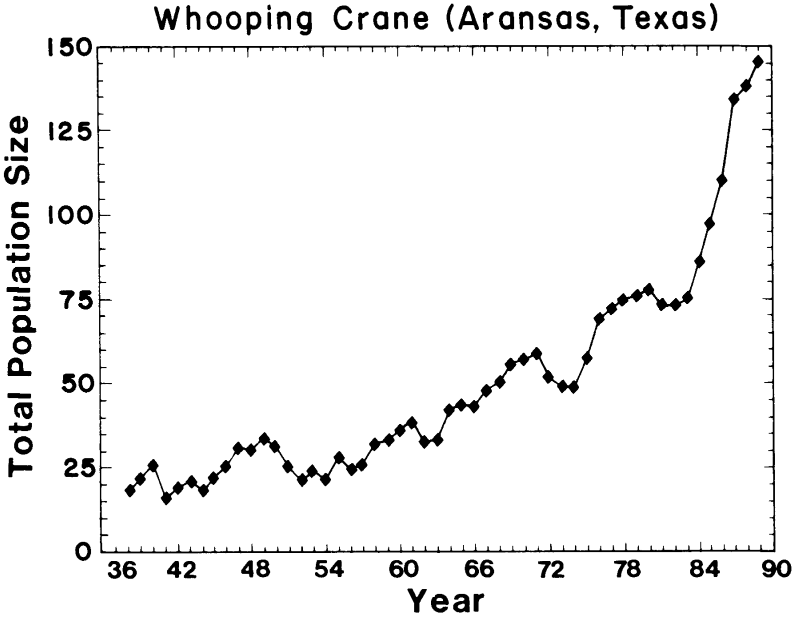
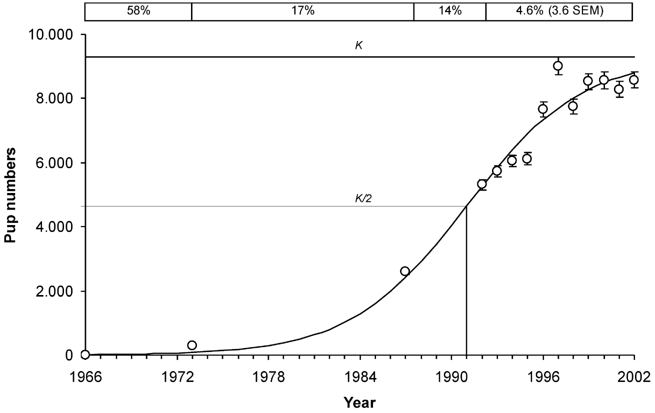
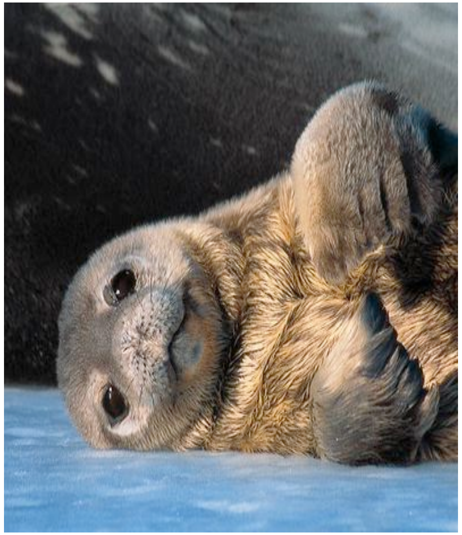


FIG. 4. Total size of the Aransas/Wood Buffalo Whooping Crane population, from 1938–1988. Data are from Boyce (1987), supplemented by more recent counts.

Arctocephallus gazella

Fig. 4 Antarctic fur-seal pup production at Cape Shirreff and San Telmo Islets, South Shetlands (1966–2002) with 3% error bars. The *fitted line* corresponds to the logistic model parameterized by $K = 9294$; $t_{50} = 1991$; $r = 0.2625$. Also shown in *boxes* is the percent rate of increase for different periods and the standard error of the mean (*SEM*) for the series ranging from 1992 to 2002





Respiratory Syncytial Virus

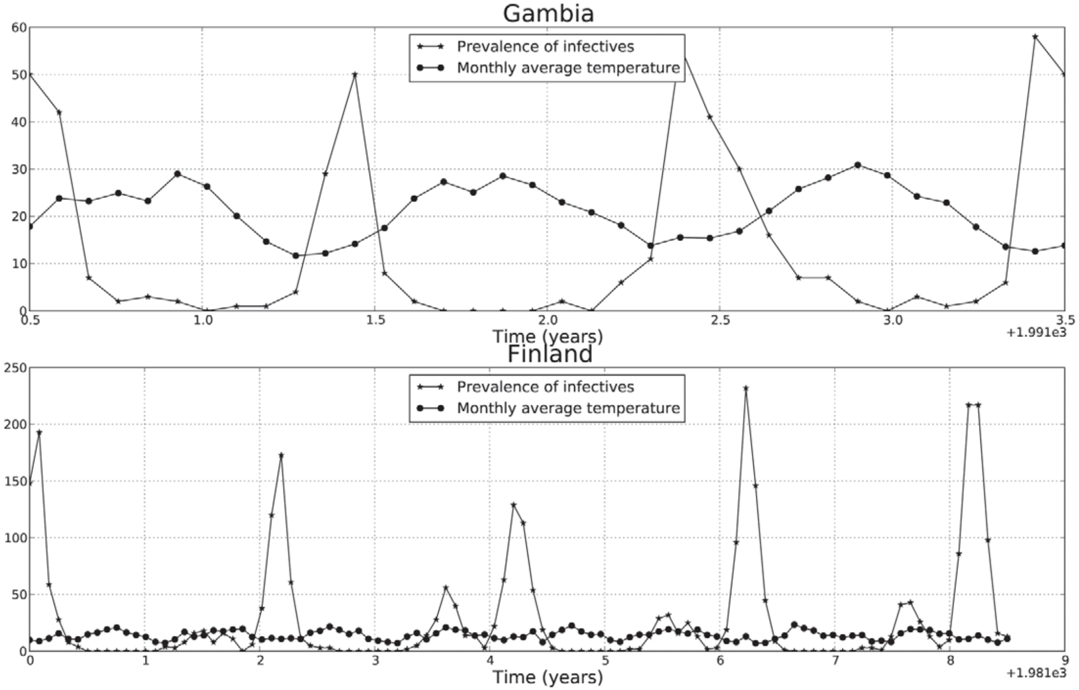
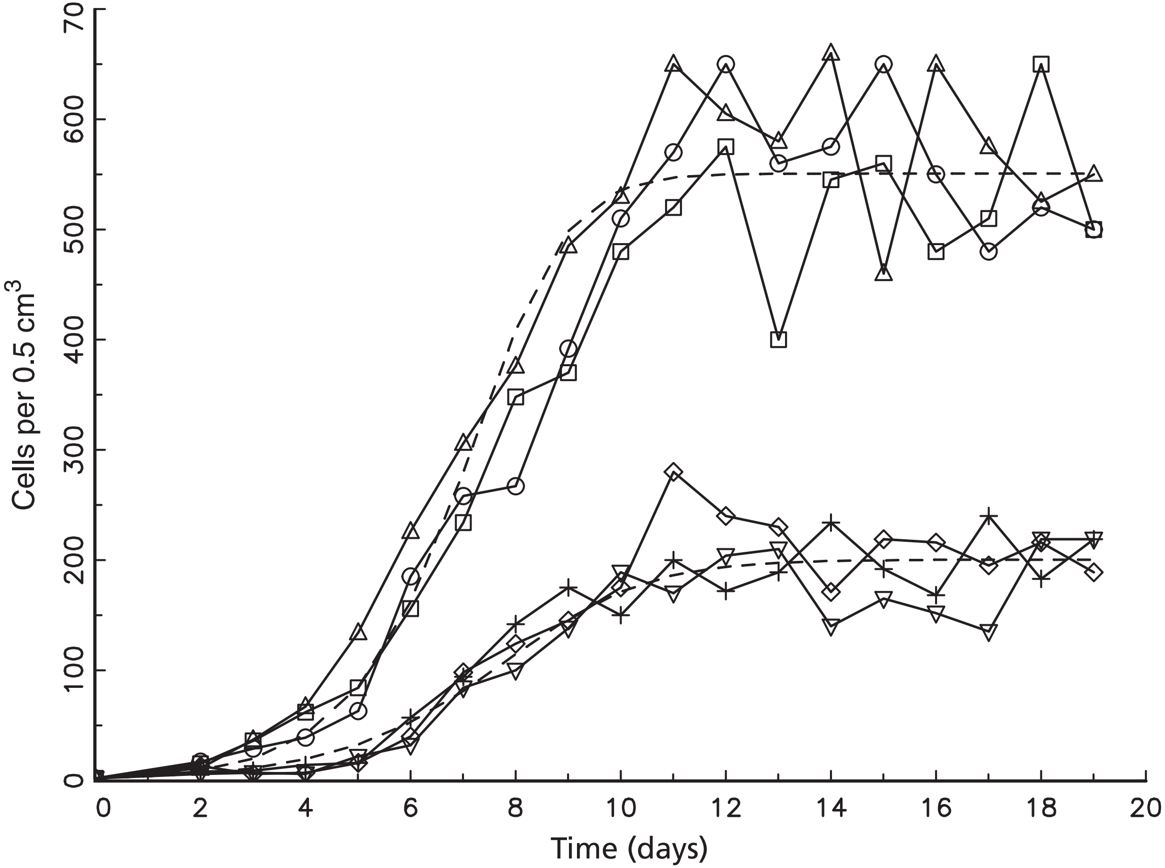


Figure 1. Observed time series of infected individuals in Gambia and Finland. Plotted are the monthly number of reported syncytial virus cases in two cities: Banjul in Gambia (from October 1991 to September 1994) and Turku in Finland (from October 1981 to March 1990). Plotted also is the mean monthly temperature range for both localities, for the same time spans.

doi:10.1371/journal.pcbi.1001079.g001

Gause’s Paramecia experiment



Talk central question

- QUESTION: Can we use stochastic population models to improve management strategies for a population of interest and better understand the biological processes driving the dynamics?

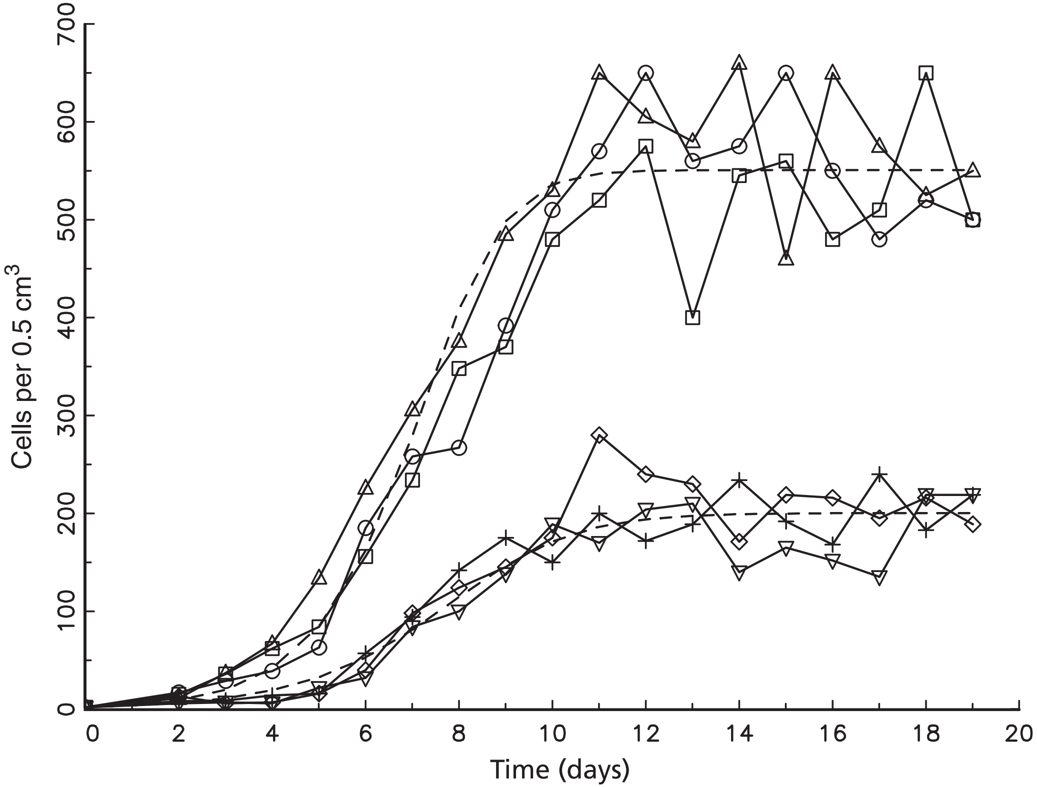
Talk central question

- QUESTION: Can we use stochastic population models to improve management strategies for a population of interest and better understand the biological processes driving the dynamics?
- ANSWER: Probably yes, provided we build those models to seek first biological understanding of a population of interest, rather than mathematical convenience.

Talk central question

- Question: Can we use stochastic population models to improve management strategies for a population of interest and better understand the biological processes driving the dynamics?
- Answer: Probably yes, provided we build those models to seek first biological understanding of a population of interest, rather than mathematical convenience.
- The statistical methodology should therefore:
  1. be informed by the nature of the data and
  2. be informed by and inform the probabilistic model-building process using Markov Chains

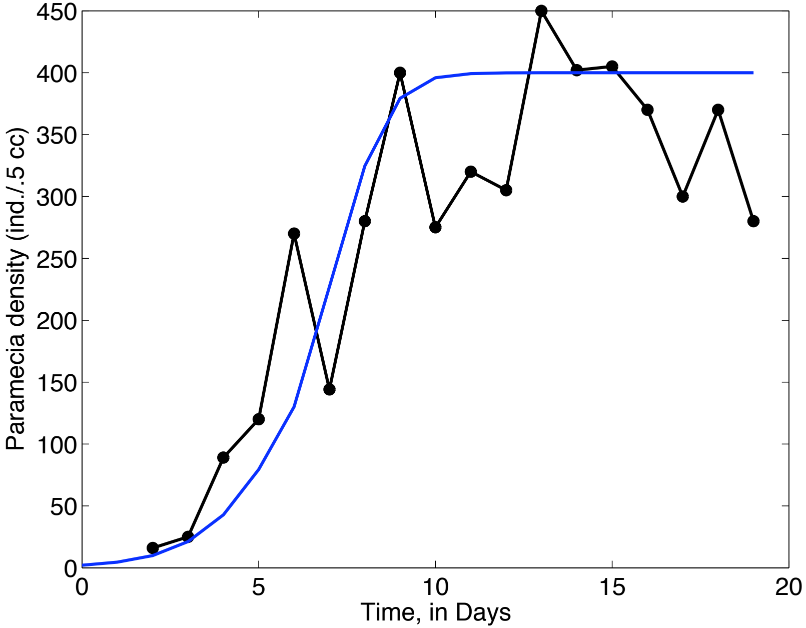
Gause's experiment: explaining deviations from deterministic model



Motivating example: population growth

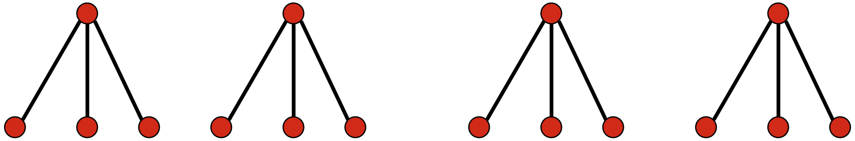
The Stochastic Ricker Model (Dennis and Taper 1994):

$$N_{t+1} = N_t \exp\left[a + b\, N_t + \sigma Z_t\right] \quad \text{where} \quad Z_t \sim \text{iid}\, \mathsf{N}(0,1)$$



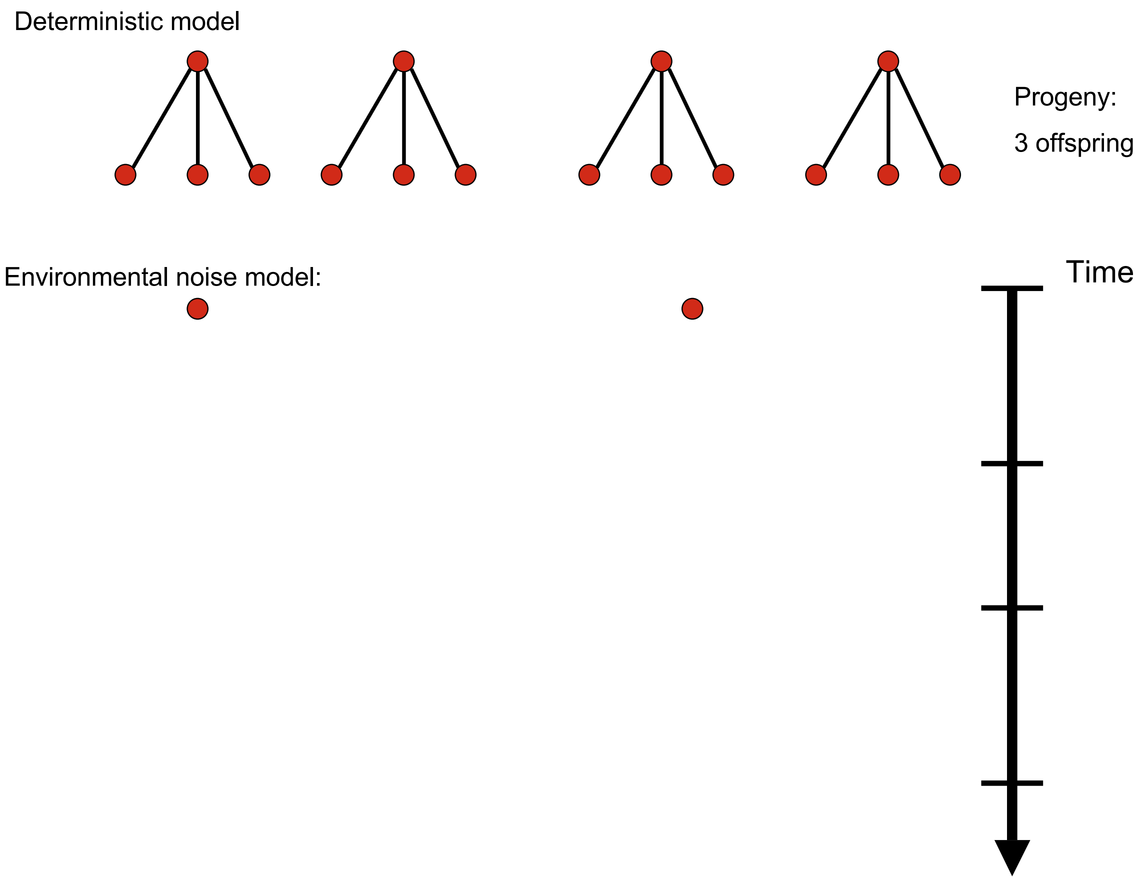
Deterministic model

Deterministic model



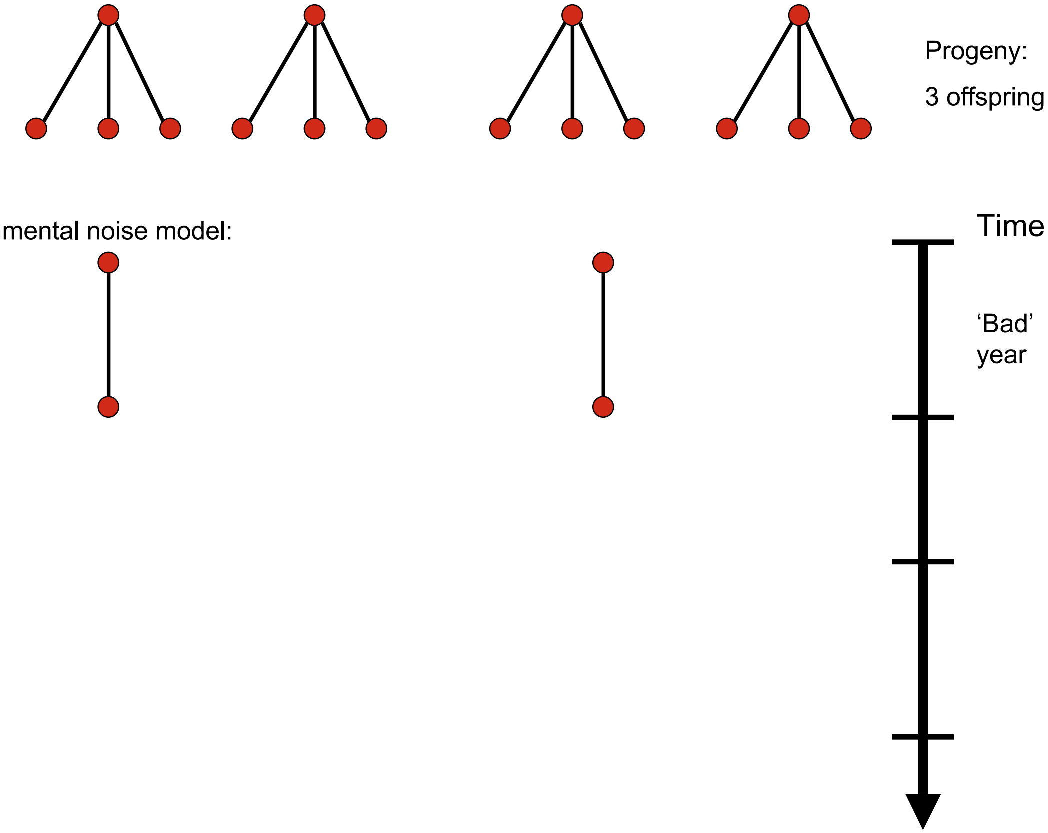
Progeny:

3 offspring


Deterministic model
Progeny:
3 offspring
Environmental noise model:
Time

Deterministic model

Environmental noise model:

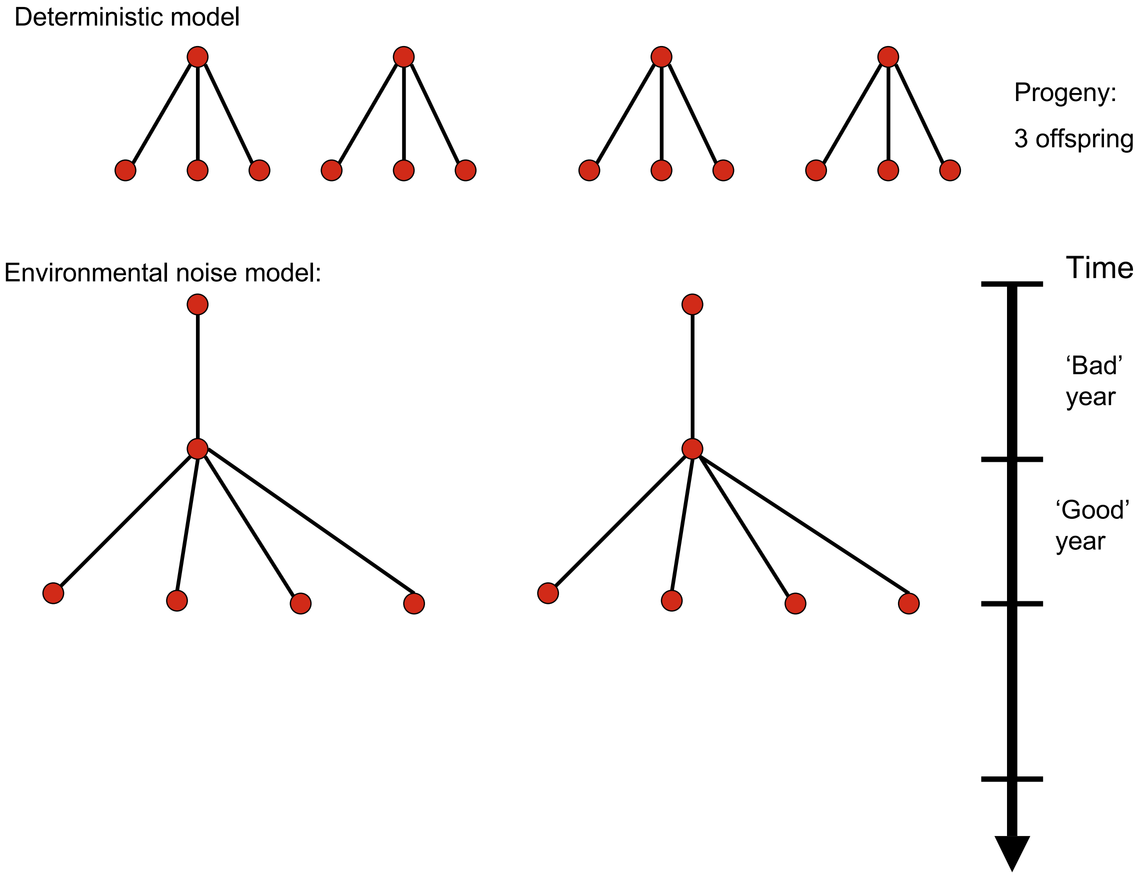


Progeny:

3 offspring

Time

‘Bad’ year



Deterministic model

Progeny:

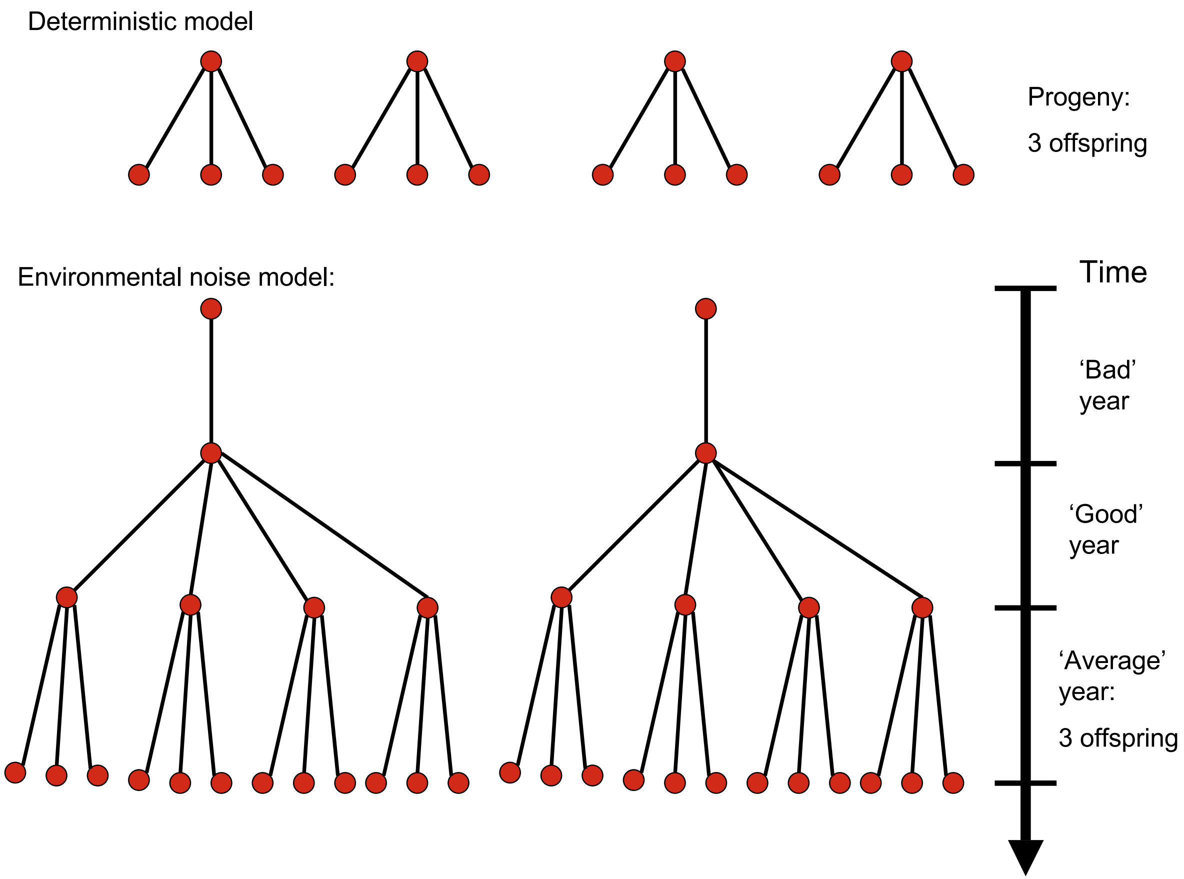
3 offspring

Environmental noise model:

Time

‘Bad’
year

‘Good’
year


Deterministic model
Progeny:
3 offspring
Environmental noise model:
Time
'Bad' year
'Good' year
'Average' year:
3 offspring

Assumptions of the Stochastic Ricker Model

“Cartoon” assumptions:

- This is a population model: “all individuals are equal” (same offspring production, same survival).
- All individuals reproduce and survive independently of each other.
- Environmental noise is non-autocorrelated /phenomenological.

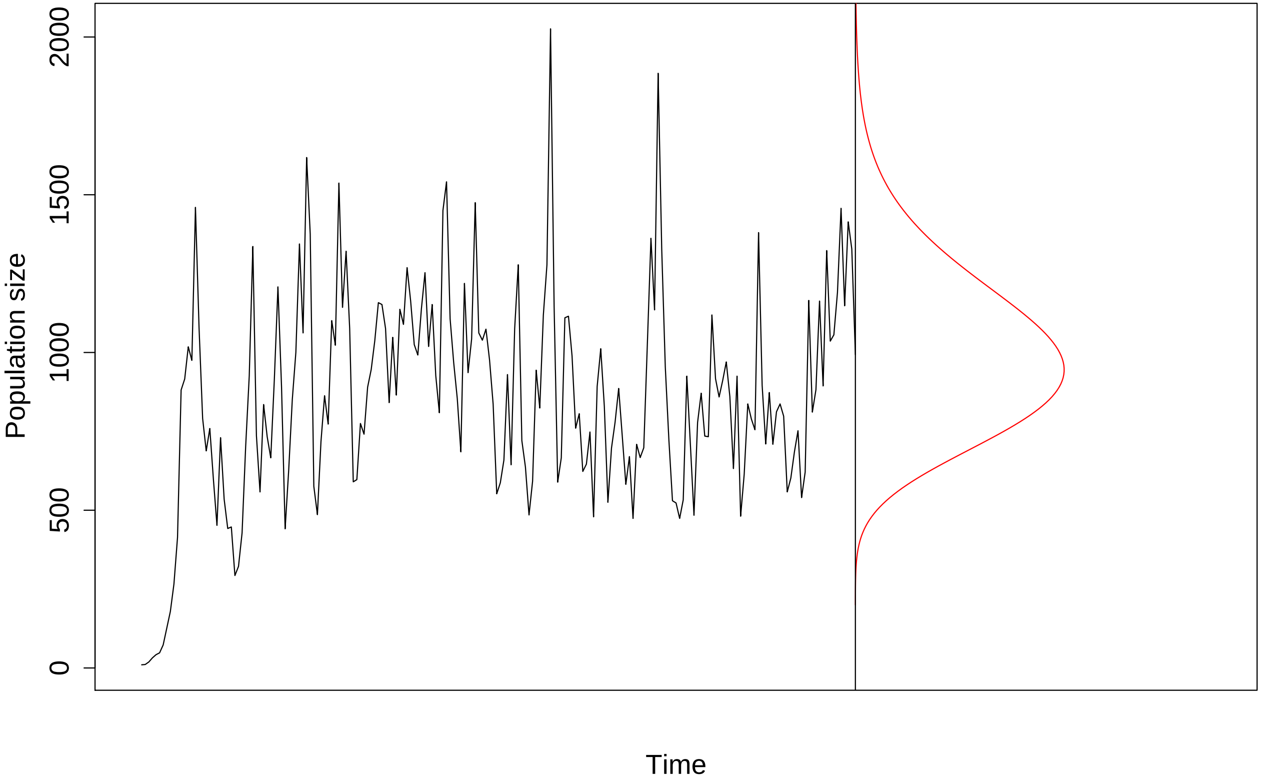
Assumptions of the Stochastic Ricker Model

"Cartoon" assumptions:

- This is a population model: "all individuals are equal" (same offspring production, same survival).
- All individuals reproduce and survive independently of each other.
- Environmental noise is non-autocorrelated /phenomenological.

Biologically useful assumptions:

- The growth rate of the population varies randomly from year to year. The environment affects (equally) every individual in the population (good years, bad years).
- Density-dependence: instead of reaching a carrying capacity point, the population reaches a stationary distribution, a cloud of points around which it fluctuates.

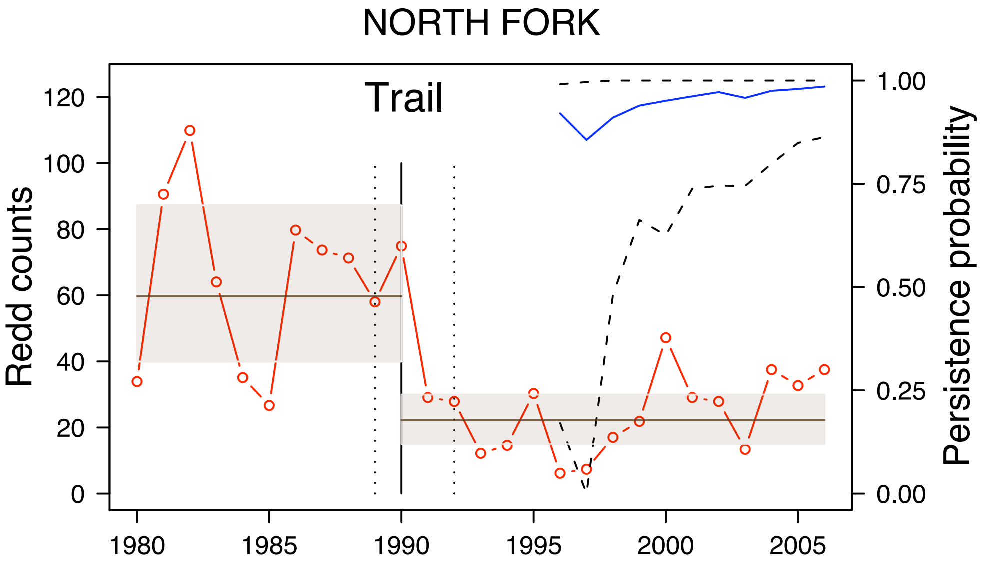

Population size
0
500
1000
1500
2000
Time

However simple, the density independent Stochastic Ricker model $N_{t+1} = N_t\{a + \sigma E_t\}$ allows us to do “Population Viability Analysis”:

Dennis, Munholland and Scott. 1991. Estimation of growth and extinction parameters from endangered species. Ecol. Monogr. 61:115-143

- Explicit expression for the probability of extinction within s years using a diffusion approximation (Stochastic Differential Equations).
- Explicit expression for the expected time until extinction.

Viability Population Monitoring and estimating trends in P(extinction)



Taper, Ponciano, Shepard, Muhlfeld and Staples. Risk-based viable population monitoring of the upper Flathead bull trout. Submitted to Ecol. Applications

Demographic stochasticity: ‘starting from scratch’

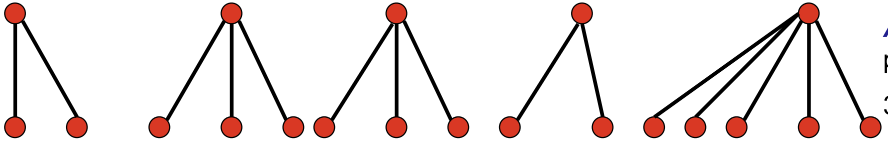
- Demographic stochasticity models variability in demographic traits, like reproduction and survival.
- It is not obvious how to combine demographic stochasticity with environmental noise in a general way.
- This problem lead us to try to formulate/understand a model of environmental noise plus demographic sotchasticity from scratch.

Demographic stochasticity model: each individual has the same offspring distribution

Average progeny
per parent per year:

3 offspring

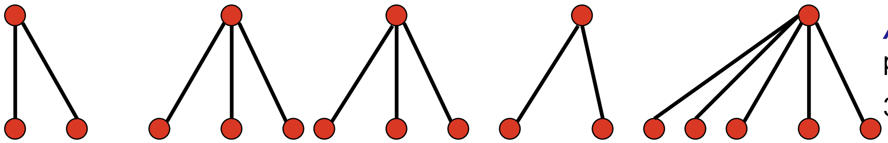
Demographic stochasticity model: each individual has the same offspring distribution



Average progeny per parent per year:

3 offspring

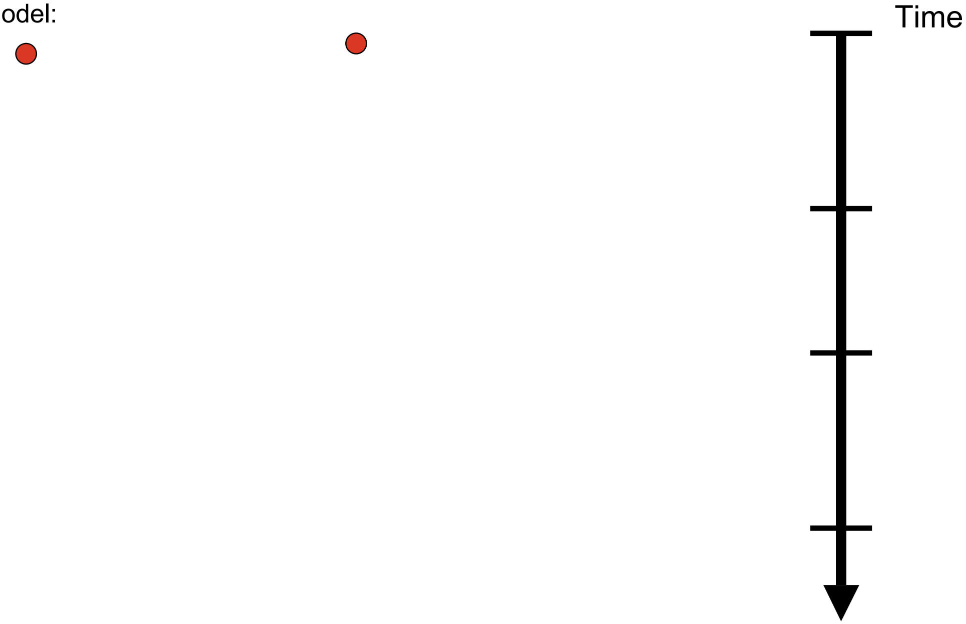
Demographic stochasticity model: each individual has the same offspring distribution



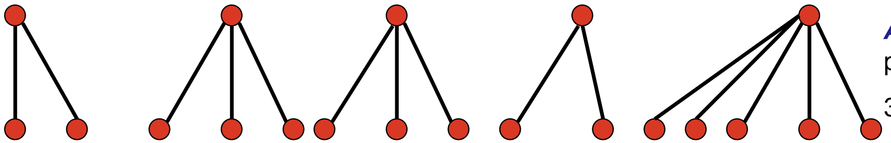
Average progeny per parent per year:

3 offspring

Environmental noise and demographic stochasticity model:



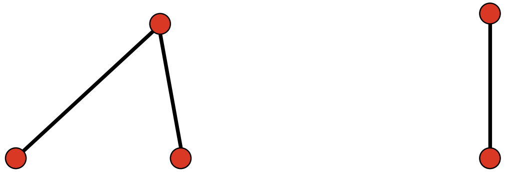
Demographic stochasticity model: each individual has the same offspring distribution



Average progeny per parent per year:

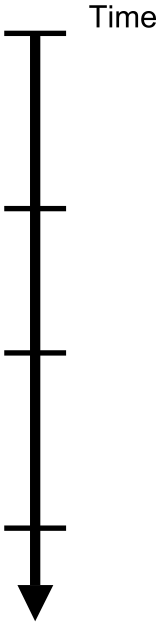
3 offspring

Environmental noise and demographic stochasticity model:

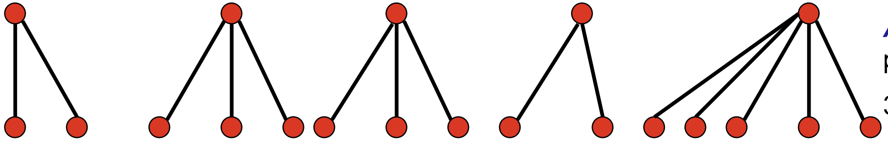


‘Bad’ year:

Average of progeny distribution is depressed by a random quantity



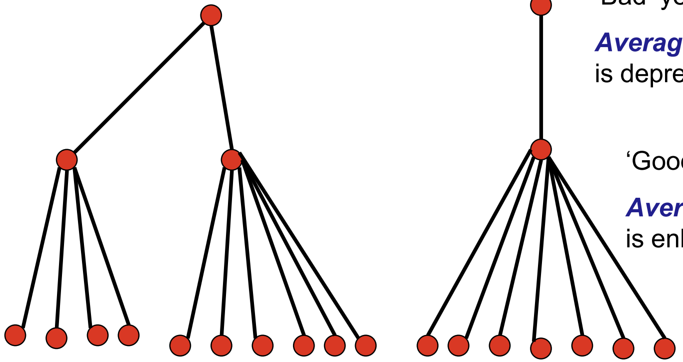
Demographic stochasticity model: each individual has the same offspring distribution



Average progeny per parent per year:

3 offspring

Environmental noise and demographic stochasticity model:

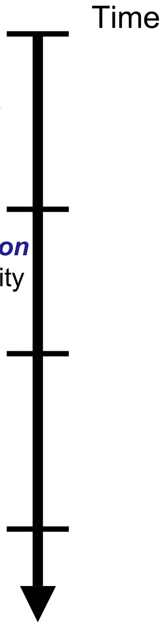


‘Bad’ year:

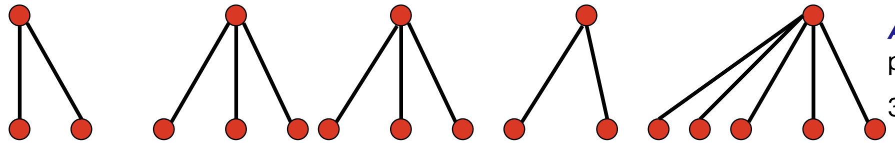
Average of progeny distribution is depressed by a random quantity

‘Good’ year:

Average of progeny distribution is enhanced by a random quantity



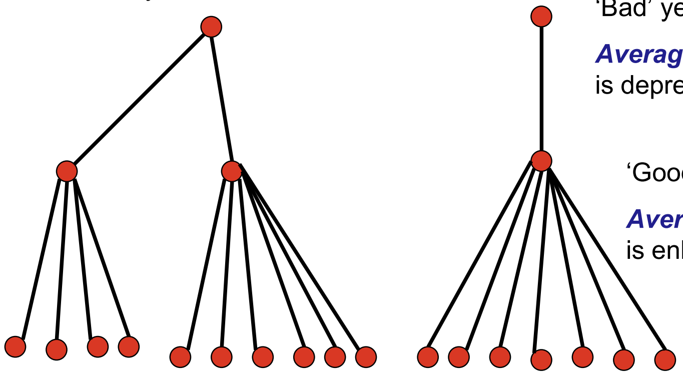
Demographic stochasticity model: each individual has the same offspring distribution



Average progeny per parent per year:

3 offspring

Environmental noise and demographic stochasticity model:

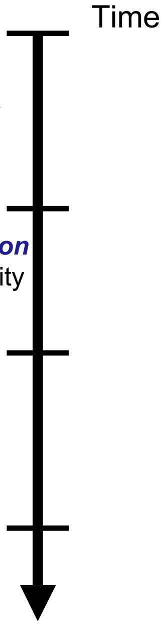


‘Bad’ year:

Average of progeny distribution is depressed by a random quantity

‘Good’ year:

Average of progeny distribution is enhanced by a random quantity



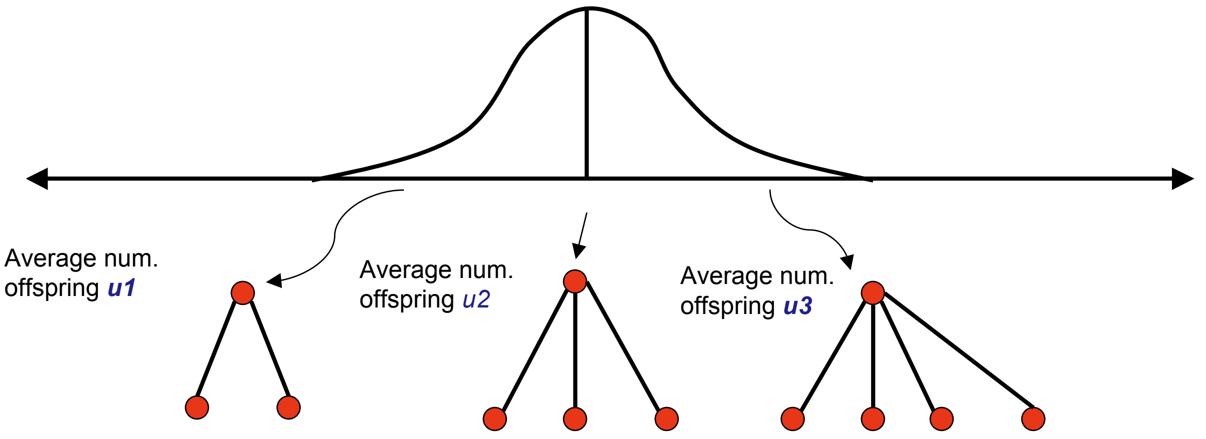
So a model with Environmental noise and demographic stochasticity is by nature a hierarchical stochastic model, where the ***mean*** of the demographic process becomes itself ***a random variable*** when environmental noise is introduced.

Demographic variability and genetic heterogeneity:

Average 'u' progeny per parent *is different, it is*

-a quantitative character -

that can be seen as drawn from a population probability distribution.

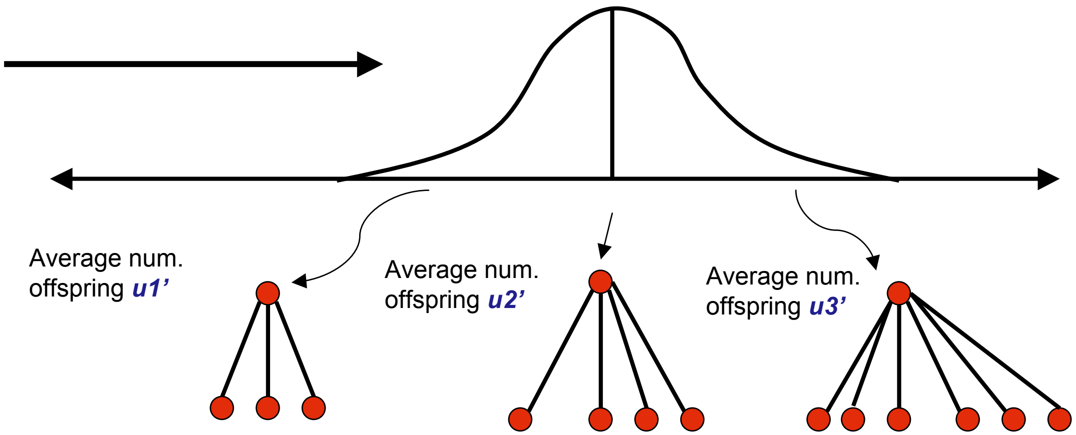

Average num. offspring ***u1***
Average num. offspring *u2*
Average num. offspring ***u3***

Environmental noise, demographic variability and genetic heterogeneity:

Average 'u' progeny per parent *is different, it is*

-a quantitative character -

that can be seen as drawn from a population probability distribution. This distribution is shifted by the enviro. noise



Average num. offspring ***u1'***

Average num. offspring ***u2'***

Average num. offspring ***u3'***

'Good' year:

Average of every progeny distribution is enhanced by a random quantity *or alternatively,*

"what is a good year for some is a bad year for others"

Setting:

- (N_t) be a discrete-time, discrete state stochastic process that models the (density-dependent) growth of a population. Furthermore, let n_t denote population size at time t.

Setting:

- (N_t) be a discrete-time, discrete state stochastic process that models the (density-dependent) growth of a population. Furthermore, let n_t denote population size at time t.
- Let X_i, $i = 1, 2, \ldots, n_t$ be iid random variables denoting the number of offspring born to individual i (non-overlapping gens.), and $g(x)$, $x = 0, 1, 2, \ldots$ be the pmf of X_i with mean and variance $\mathrm{E}[X_i] = \lambda$ and $\mathrm{V}[X_i] = \phi^2$, respectively.

Setting:

- (N_t) be a discrete-time, discrete state stochastic process that models the (density-dependent) growth of a population. Furthermore, let n_t denote population size at time t.
- Let X_i, $i = 1, 2, \ldots, n_t$ be iid random variables denoting the number of offspring born to individual i (non-overlapping gens.), and $g(x)$, $x = 0, 1, 2, \ldots$ be the pmf of X_i with mean and variance $\mathrm{E}[X_i] = \lambda$ and $\mathrm{V}[X_i] = \phi^2$, respectively.
- Let $Y_t = \sum_{i=1}^{n_t} X_i$ be the total number of offspring born between times t and $t + 1$.

Setting:

- (N_t) be a discrete-time, discrete state stochastic process that models the (density-dependent) growth of a population. Furthermore, let n_t denote population size at time t.

- Let X_i, $i = 1, 2, \ldots, n_t$ be iid random variables denoting the number of offspring born to individual i (non-overlapping gens.), and $g(x)$, $x = 0, 1, 2, \ldots$ be the pmf of X_i with mean and variance $\mathrm{E}[X_i] = \lambda$ and $\mathrm{V}[X_i] = \phi^2$, respectively.

- Let $Y_t = \sum_{i=1}^{n_t} X_i$ be the total number of offspring born between times t and $t+1$.

- Finally, let p_t be the density dependent probability of survival of each offspring born at time t. For ex.: $p_t = \exp\{-b\, n_t\}$(Ricker), Gompertz model: $p_t = \exp\{-b \ln n_t\}$, Theta-Ricker model, $p_t = \exp\left\{-b\, n_t^{\theta}\right\}$, and the Hassell model $p_t = 1/(1 + b\, n_t)^c$.

Setting:

- (N_t) be a discrete-time, discrete state stochastic process that models the (density-dependent) growth of a population. Furthermore, let n_t denote population size at time t.

- Let X_i, $i = 1, 2, \ldots, n_t$ be iid random variables denoting the number of offspring born to individual i (non-overlapping gens.), and $g(x)$, $x = 0, 1, 2, \ldots$ be the pmf of X_i with mean and variance $\mathrm{E}[X_i] = \lambda$ and $\mathrm{V}[X_i] = \phi^2$, respectively.

- Let $Y_t = \sum_{i=1}^{n_t} X_i$ be the total number of offspring born between times t and $t+1$.

- Finally, let p_t be the density dependent probability of survival of each offspring born at time t. For ex.: $p_t = \exp\{-b\,n_t\}$(Ricker), Gompertz model: $p_t = \exp\{-b \ln n_t\}$, Theta-Ricker model, $p_t = \exp\left\{-b\,n_t^{\theta}\right\}$, and the Hassell model $p_t = 1/(1 + b\,n_t)^c$.

- Each individual survives independently from each other w.p. p_t.

Demographic Stochasticity:

Conditional on $Y_t = y$, the total number of survivors for next generation is binomially distributed with parameters y and p_t. It follows that the moments of the conditional process $(N_{t+1}|N_t = n_t)$ are

$$\begin{aligned}
\mathrm{E}[N_{t+1}|N_t = n_t] &= \mathrm{E}\left[\mathrm{E}\left[N_{t+1}|(N_t = n_t, Y_t)\right]\right] = \lambda n_t p_t, \\
\mathrm{V}[N_{t+1}|N_t = n_t] &= \mathrm{E}\left[\mathrm{V}\left[N_{t+1}|(N_t = n_t, Y_t)\right]\right] + \mathrm{V}\left[\mathrm{E}\left[N_{t+1}|(N_t = n_t, Y_t)\right]\right] \\
&= \left[\lambda p_t(1-p_t) + \phi^2 p_t^2\right] n_t.
\end{aligned} \tag{1}$$

Example: $X_i \sim \mathrm{Poisson}(\lambda) \Rightarrow (N_{t+1}|N_t = n_t) \sim \mathrm{Poisson}(\lambda n_t p_t)$.

Environmental Stochasticity:

- Defined as the case wherein one or more of the vital rates, say, the mean of the offspring distribution, varies randomly *over time*.

Environmental Stochasticity:

- Defined as the case wherein one or more of the vital rates, say, the mean of the offspring distribution, varies randomly *over time*.
- In the absence of demographic noise, within a single year, all the individuals in the population get the same vital rate value.

Environmental Stochasticity:

- Defined as the case wherein one or more of the vital rates, say, the mean of the offspring distribution, varies randomly *over time*.
- In the absence of demographic noise, within a single year, all the individuals in the population get the same vital rate value.
- In the presence of demographic noise and enviro. noise, the offspring distribution that characterizes *all* the individuals in the population changes its location parameter every year.

Environmental Stochasticity:

- Defined as the case wherein one or more of the vital rates, say, the mean of the offspring distribution, varies randomly *over time*.
- In the absence of demographic noise, within a single year, all the individuals in the population get the same vital rate value.
- In the presence of demographic noise and enviro. noise, the offspring distribution that characterizes *all* the individuals in the population changes its location parameter every year.
- That is, during “good years” the mean of the offspring distribution of the individuals in the population increases and during “bad years” it decreases.

Environmental Stochasticity:

- Defined as the case wherein one or more of the vital rates, say, the mean of the offspring distribution, varies randomly *over time*.
- In the absence of demographic noise, within a single year, all the individuals in the population get the same vital rate value.
- In the presence of demographic noise and enviro. noise, the offspring distribution that characterizes *all* the individuals in the population changes its location parameter every year.
- That is, during "good years" the mean of the offspring distribution of the individuals in the population increases and during "bad years" it decreases.
- In that sense, the biological justification of the formulation of an environmental noise model is to allow for changes over time on the location of the offspring distribution.

Environmental Stochasticity:

- Defined as the case wherein one or more of the vital rates, say, the mean of the offspring distribution, varies randomly *over time*.
- In the absence of demographic noise, within a single year, all the individuals in the population get the same vital rate value.
- In the presence of demographic noise and enviro. noise, the offspring distribution that characterizes *all* the individuals in the population changes its location parameter every year.
- That is, during “good years” the mean of the offspring distribution of the individuals in the population increases and during “bad years” it decreases.
- In that sense, the biological justification of the formulation of an environmental noise model is to allow for changes over time on the location of the offspring distribution.
- Yet, because for very few probability distributions the mean is not a function of the variance, it is difficult to conceive practical models where only the mean of the offspring distribution is affected and not its variance.

Population size moments under demographic variability and environmental noise

Let W_t be a r.v. for the value of the mean of the offspring distribution at time t. At a given time t, $W_t = w_t$

Population size moments under demographic variability and environmental noise

Let W_t be a r.v. for the value of the mean of the offspring distribution at time t. At a given time t, $W_t = w_t$

- $X_i \sim g(x, w_t)$ is each individual's offspring distribution.

Population size moments under demographic variability and environmental noise

Let W_t be a r.v. for the value of the mean of the offspring distribution at time t. At a given time t, $W_t = w_t$

- $X_i \sim g(x, w_t)$ is each individual's offspring distribution.
- $\mathrm{E}[X_i | W_t = w_t] = w_t$ and $\mathrm{V}[X_i | W_t = w_t] = \phi^2(w_t)$.

Population size moments under demographic variability and environmental noise

Let W_t be a r.v. for the value of the mean of the offspring distribution at time t. At a given time t, $W_t = w_t$

- $X_i \sim g(x, w_t)$ is each individual's offspring distribution.
- $\mathrm{E}[X_i|W_t = w_t] = w_t$ and $\mathrm{V}[X_i|W_t = w_t] = \phi^2(w_t)$.
- If $Y_t = \sum_{i=1}^{n_t} X_i$, then, conditioning on $W_t = w_t$ (keep that in mind), $\mathrm{E}[Y_t] = w_t n_t$ and $\mathrm{V}[Y_t] = n_t \phi^2(w_t)$.

Population size moments under demographic variability and environmental noise

Let W_t be a r.v. for the value of the mean of the offspring distribution at time t. At a given time t, $W_t = w_t$

- $X_i \sim g(x, w_t)$ is each individual's offspring distribution.
- $\mathrm{E}[X_i|W_t = w_t] = w_t$ and $\mathrm{V}[X_i|W_t = w_t] = \phi^2(w_t)$.
- If $Y_t = \sum_{i=1}^{n_t} X_i$, then, conditioning on $W_t = w_t$ (keep that in mind), $\mathrm{E}[Y_t] = w_t n_t$ and $\mathrm{V}[Y_t] = n_t \phi^2(w_t)$.
- Now assume that $(N_{t+1}|N_t = n_t, W_t = w_t, Y_t) \sim \mathrm{Binomial}(Y_t, p_t)$.

Population size moments under demographic variability and environmental noise

Let W_t be a r.v. for the value of the mean of the offspring distribution at time t. At a given time t, $W_t = w_t$

- $X_i \sim g(x, w_t)$ is each individual's offspring distribution.
- $\mathrm{E}[X_i | W_t = w_t] = w_t$ and $\mathrm{V}[X_i | W_t = w_t] = \phi^2(w_t)$.
- If $Y_t = \sum_{i=1}^{n_t} X_i$, then, conditioning on $W_t = w_t$ (keep that in mind), $\mathrm{E}[Y_t] = w_t n_t$ and $\mathrm{V}[Y_t] = n_t \phi^2(w_t)$.
- Now assume that $(N_{t+1} | N_t = n_t, W_t = w_t, Y_t) \sim \mathrm{Binomial}(Y_t, p_t)$.
- Averaging over all the possible values of Y_t (given that w_t is fixed), then, for that particular year t we get that

Environmental and demographic noise continued

$$\begin{aligned}
\mathrm{E}[N_{t+1}|N_t = n_t, W_t = w_t] &= \mathrm{E}\left[\mathrm{E}\left[N_{t+1}|(N_t = n_t, W_t = w_t, Y_t)\right]\right] = w_t n_t p_t, \\
\mathrm{V}[N_{t+1}|N_t = n_t, W_t = w_t] &= \mathrm{E}\left[\mathrm{V}\left[N_{t+1}|(N_t = n_t, W_t = w_t, Y_t)\right]\right] \\
&\quad + \mathrm{V}\left[\mathrm{E}\left[N_{t+1}|(N_t = n_t, W_t = w_t, Y_t)\right]\right] \\
&= \left[w_t p_t(1 - p_t) + \phi^2(w_t) p_t^2\right] n_t.
\end{aligned} \tag{2}$$

Environmental and demographic noise continued

$$\begin{aligned}
\mathrm{E}[N_{t+1}|N_t = n_t, W_t = w_t] &= \mathrm{E}\left[\mathrm{E}\left[N_{t+1}|(N_t = n_t, W_t = w_t, Y_t)\right]\right] = w_t n_t p_t, \\
\mathrm{V}[N_{t+1}|N_t = n_t, W_t = w_t] &= \mathrm{E}\left[\mathrm{V}\left[N_{t+1}|(N_t = n_t, W_t = w_t, Y_t)\right]\right] \\
&+ \mathrm{V}\left[\mathrm{E}\left[N_{t+1}|(N_t = n_t, W_t = w_t, Y_t)\right]\right] \\
&= \left[w_t p_t(1-p_t) + \phi^2(w_t) p_t^2\right] n_t.
\end{aligned} \tag{3}$$

Now, averaging over all the the possible values of the Environmental process, we get that the general moments of $(N_{t+1}|N_t = n_t)$ are:

$$\begin{aligned}
\mathrm{E}[N_{t+1}|N_t = n_t] &= \mathrm{E}\left[\mathrm{E}\left[N_{t+1}|(N_t = n_t, W_t)\right]\right] = \mathrm{E}[W_t] n_t p_t, \\
\mathrm{V}[N_{t+1}|N_t = n_t] &= \mathrm{E}\left[\mathrm{V}\left[N_{t+1}|(N_t = n_t, W_t)\right]\right] + \mathrm{V}\left[\mathrm{E}\left[N_{t+1}|(N_t = n_t, W_t)\right]\right] \\
&= \left[\mathrm{E}[W_t] p_t(1-p_t) + \mathrm{E}[\phi^2(W_t)] p_t^2\right] n_t + (n_t p_t)^2 \mathrm{V}[W_t].
\end{aligned} \tag{4}$$

An example with exact transition probability mass function:

If we let $\lambda \sim \mathrm{Gamma}(k, \alpha)$ represent the environmental noise and let each individual have a Poisson offspring distribution then we get that

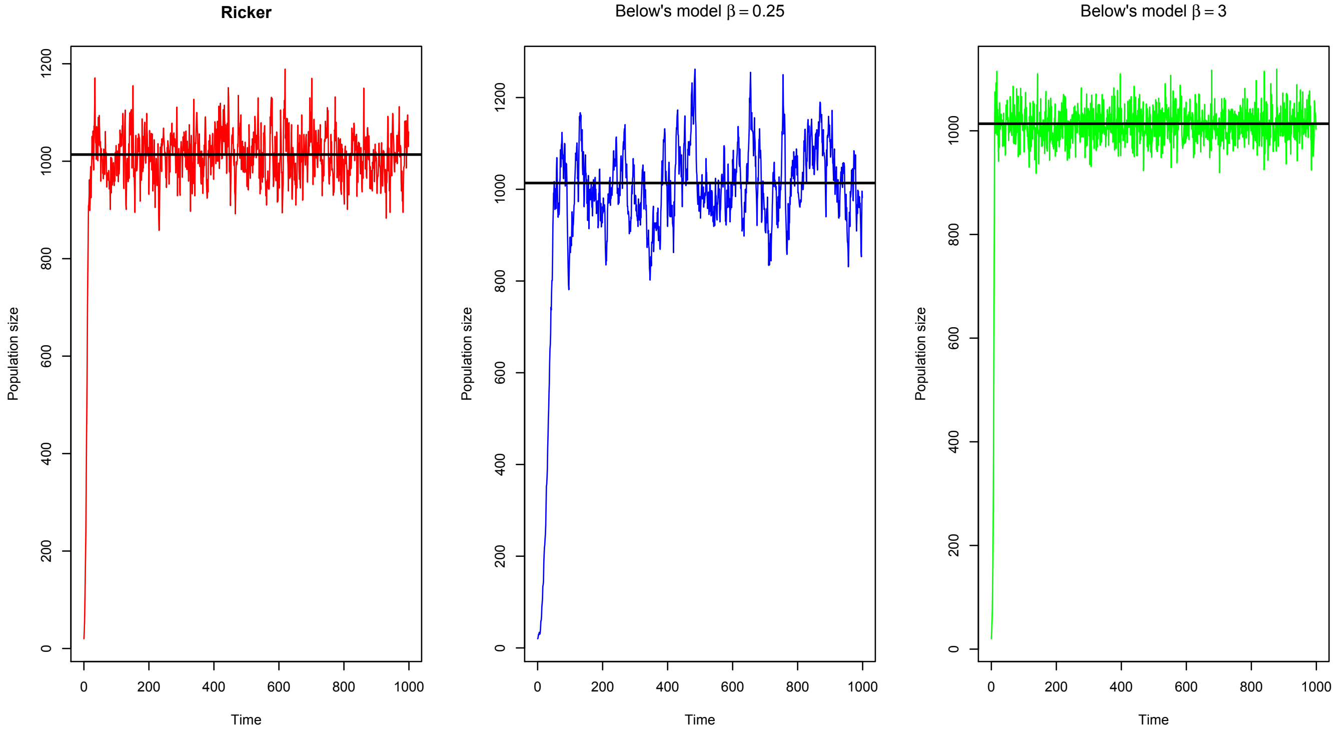
$$P(N_{t+1} = n_{t+1} | N_t = n_t) = \frac{\Gamma(n_{t+1} + k)}{\Gamma(k) n_{t+1}!} \left(\frac{\alpha}{n_t p_t + \alpha} \right)^k \left(\frac{n_t p_t}{n_t p_t + \alpha} \right)^{n_{t+1}},$$

where

$$p_t = \begin{cases} e^{-b n_t} & \text{for Ricker model} \\ \exp\left\{-b\, n_t^{\theta}\right\} & \text{for theta-Ricker model} \\ \exp\{-b \ln n_t\} & \text{for Gompertz model} \\ 1/(1 + b\, n_t)^c & \text{for Hassell's model} \\ 1/(1 + (a-1)(n_t/K)^{\beta}) & \text{for Below's model} \end{cases}$$

Ponciano, J.M. et al *in prep.* Demographic stochasticity, environmental noise and sampling error: implications for conservation biology.

Simulation Example: Demographic and Environmental Stochasticities



Statistical Inference for Markovian population models models

- Discrete state, discrete time Markov processes
- Discrete state, continuous time
- Continuous state, discrete time
- Accounting for sampling error
- Continuous time, continuous states: next talk

Introducing the likelihood function: a Chain-Binomial model

- Field work: Monthly census of extant individuals from a closed population that reproduces every 5 years, for 24 months
- No reproduction occurs during those 24 months
- Data: Number of survivors at the end of each one of the 24 months (no sampling error): $n_1, n_2, \ldots, n_{24}$, starting at $n_0 =$ known cst.
- We want to study the survival process during those 24 months.

Introducing the likelihood function: a Chain-Binomial model

- Probabilistic model of the biological process: consider a discrete time, discrete state Markov process $\{N\}_t$ that models only the survival process from one unit of time to the other (from one month to the next).
- Let $p_{ij} = P(N_{t+1} = j | N_t = i)$, assume n_0 is a fixed quantity and let

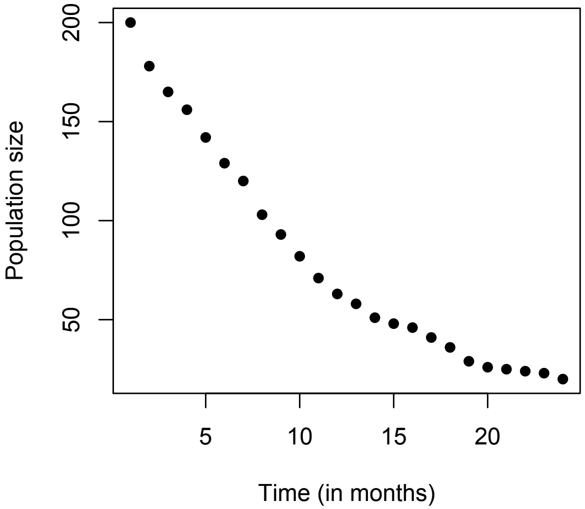
$$p_{ij} = \binom{i}{j} p^i (1-p)^{i-j}, \quad j = 0, 1, \ldots, i$$

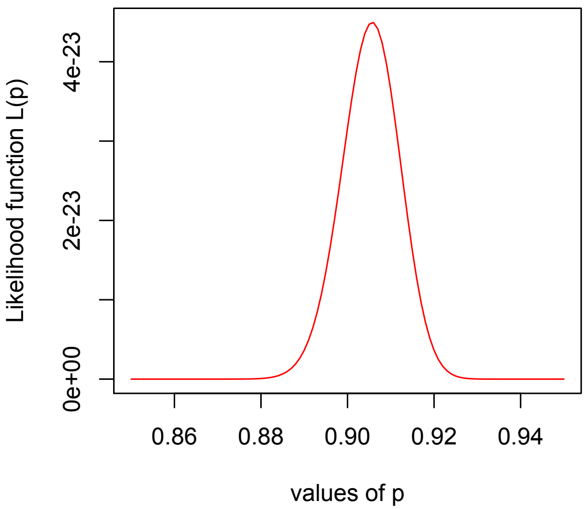
- We have a complete probabilistic description of the observations, except we don't know p!
- Biological questions of interest: Does p changes from one year to the other? From season to season? Between sexes or ages?

The likelihood function

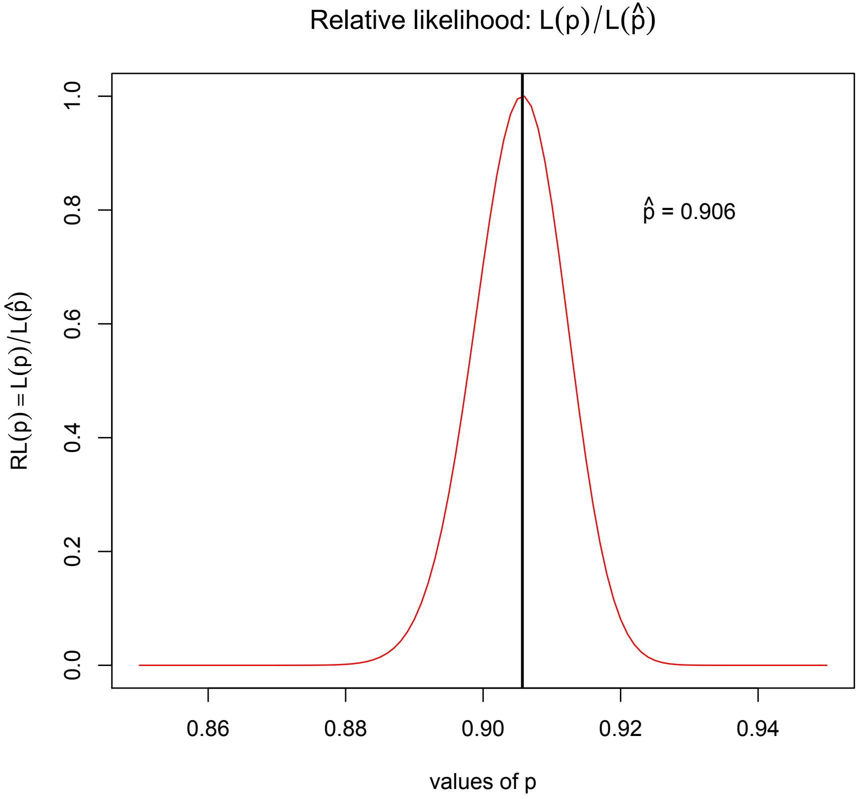
It is the joint probability of the observations N_t evaluated at the recorded data (the n_t), which, according to the Markov property is:

$$\begin{aligned} L(p) = P(N_1 = n_1, N_2 = n_2, \ldots, N_{24} = n_{24}) &= \prod_{i=1}^{24} P(N_i = n_i | N_{i-1} = n_{i-1}) \\ &= \prod_{i=1}^{24} \binom{n_{i-1}}{n_i} p^{n_i} (1-p)^{n_{i-1}-n_i} \end{aligned}$$





The relative likelihood function



Maximizing $L(p)$: set $\frac{dL(p)}{dp} = 0$, solve for p

Amounts to set $\frac{1}{L(p)}\frac{dL(p)}{dp} = 0$, solve for p. That is,

$$\frac{d\,\ln L(p)}{dp} = \frac{d}{dp}\left[\sum_{i=1}^{24} \ln \binom{n_{i-1}}{n_i} p^{n_i}(1-p)^{n_{i-1}-n_i}\right]$$

$$\Rightarrow \frac{d\,\ln L(p)}{dp} \propto \frac{\sum_{i=1}^{24} n_i}{p} - \frac{\sum_{i=1}^{24} n_{i-1}-n_i}{(1-p)} = 0$$

$$\Rightarrow \hat{p} = \frac{\sum_{i=1}^{24} n_i}{\sum_{i=1}^{24} n_{i-1}} = 0.906$$

The likelihood function for the model with environmental and demographic stochasticities:

Let $\lambda \sim \mathrm{Gamma}(k, \alpha)$ represent the environmental noise and let each individual have a Poisson offspring distribution then we saw that the transition pdf was

$$P(N_{t+1} = n_{t+1} | N_t = n_t) = \frac{\Gamma(n_{t+1} + k)}{\Gamma(k) n_{t+1}!} \left(\frac{\alpha}{n_t p_t + \alpha} \right)^k \left(\frac{n_t p_t}{n_t p_t + \alpha} \right)^{n_{t+1}},$$

where

$$p_t = \exp\left\{ -b\, n_t^{\theta} \right\} \quad \text{for theta-Ricker model.}$$

Given a time series data set consisting of the (exact) counts $n_0, n_1, \ldots, n_q$, then the likelihood function for the parameters $\boldsymbol{\theta} = [k, \alpha, b, \theta]'$is again the joint pmf of the population sizes $N_1, \ldots, N_q$ evaluated at the data at hand:

$$L(\boldsymbol{\theta}) = \prod_{t=0}^{q-1} P(N_{t+1} = n_{t+1} | N_t = n_t).$$

If N_0 is an observation from the stationary distribution of the process, then

$$L(\boldsymbol{\theta}) = P(N_0 = n_0) \times \prod_{t=0}^{q-1} P(N_{t+1} = n_{t+1} | N_t = n_t).$$

Estimating parameters of a continuous time, discrete states MC

Consider a pure birth process where

$$\begin{aligned}
P[N(t+\delta t) = n+1 | N(t) = n] &= (\delta t)\lambda_n \\
P[N(t+\delta t) = n | N(t) = n] &= 1 - (\delta t)\lambda_n \\
P[\text{more than 1 birth in time } \delta t] &= \text{negligible},
\end{aligned}$$

where $\lambda_n = \lambda n$. This is an exponential-type growth rate model due to births. Observations of $N(t)$ at times $0 < t_1 < t_2 < \ldots < t_q$ yield the pairs

$$(t_0, n_0), (t_1, n_1), (t_2, n_2), \ldots, (t_q, n_q).$$

Remember that the transition pmf is a translated negative binomial

$$p_n(t) = P[N(t) = n | N(0) = n_0] = \binom{n-1}{n_0-1} \left(\exp^{-\lambda t}\right)^{n_0} \left(1 - \exp^{-\lambda t}\right)^{n-n_0}, n = n_0, n_0{+}1, n_0{+}2, . .$$

To get the total likelihood of the realized observations we write down the transition pmf of each step and use the Markov property.

Estimating parameters of a continuous time, discrete states MC

Transition pmf:

$$\begin{aligned} P[N(t_i) &= n_i|N(t_{i-1}) = n_{i-1}] = \binom{n_i-1}{n_{i-1}-1}\left(\exp^{-\lambda(t_i-t_{i-1})}\right)^{n_{i-1}}\left(1-\exp^{-\lambda(t_i-t_{i-1})}\right)^{n_i-n_{i-1}} \\ &= f(n_i, t_i - t_{i-1}|n_{i-1}) \end{aligned}$$

Estimating parameters of a continuous time, discrete states MC

Transition pmf:

$$\begin{aligned} P[N(t_i) &= n_i | N(t_{i-1}) = n_{i-1}] = \binom{n_i - 1}{n_{i-1} - 1} \left(\exp^{-\lambda(t_i - t_{i-1})}\right)^{n_{i-1}} \left(1 - \exp^{-\lambda(t_i - t_{i-1})}\right)^{n_i - n_{i-1}} \\ &= f(n_i, t_i - t_{i-1} | n_{i-1}) \end{aligned}$$

Let $\tau_1 = t_1 - 0, \tau_2 = t_2 - t_1, \ldots, \tau_q = t_q - t_{q-1}$ (not necessarily evenly spaced).

Estimating parameters of a continuous time, discrete states MC

Transition pmf:

$$\begin{aligned} P[N(t_i) &= n_i|N(t_{i-1}) = n_{i-1}] = \binom{n_i-1}{n_{i-1}-1} \left(\exp^{-\lambda(t_i-t_{i-1})}\right)^{n_{i-1}} \left(1-\exp^{-\lambda(t_i-t_{i-1})}\right)^{n_i-n_{i-1}} \\ &= f(n_i, t_i - t_{i-1}|n_{i-1}) \end{aligned}$$

Let $\tau_1 = t_1 - 0, \tau_2 = t_2 - t_1, \ldots, \tau_q = t_q - t_{q-1}$ (not necessarily evenly spaced).

Then, the likelihood function necessary to connect the model with data is given by

$$\begin{aligned} L(\lambda) = f(n_1, n_2, \ldots, n_q|n_0) &= f(n_1, \tau_1|n_0) f(n_2, \tau_2|n_1) \ldots f(n_q, \tau_q|n_{q-1}) \\ &= \prod_{i=1}^{q} \binom{n_i-1}{n_{i-1}-1} \left(\exp^{-\lambda(\tau_i)}\right)^{n_{i-1}} \left(1-\exp^{-\lambda(\tau_i)}\right)^{n_i-n_{i-1}} \end{aligned}$$

Confronting multiple birth process models to data using the likelihood

Model: a pure birth process with $\lambda_n = \theta + \lambda n$ (immigration + births).

Confronting multiple birth process models to data using the likelihood

Model: a pure birth process with $\lambda_n = \theta + \lambda n$ (immigration + births).

Case 1: suppose $\lambda = 0$ (nothing but immigrations), $n_0 = 0$. Then $p_n(t) = \frac{e^{-\theta t}(\theta t)^n}{n!}$.

Confronting multiple birth process models to data using the likelihood

Model: a pure birth process with $\lambda_n = \theta + \lambda n$ (immigration + births).

Case 1: suppose $\lambda = 0$ (nothing but immigrations), $n_0 = 0$. Then $p_n(t) = \frac{e^{-\theta t}(\theta t)^n}{n!}$.

Case 2: $\lambda > 0, n_0 = 0$ (Immigration and births). Then

$$p_n(t) = \binom{\frac{\theta}{\lambda} + n - 1}{n} \left(e^{-\lambda t}\right)^{\frac{\theta}{\lambda}} \left(1 - e^{-\lambda t}\right)^n .$$

Confronting multiple birth process models to data using the likelihood

Model: a pure birth process with $\lambda_n = \theta + \lambda n$ (immigration + births).

Case 1: suppose $\lambda = 0$ (nothing but immigrations), $n_0 = 0$. Then $p_n(t) = \frac{e^{-\theta t}(\theta t)^n}{n!}$.

Case 2: $\lambda > 0, n_0 = 0$ (Immigration and births). Then

$$p_n(t) = \binom{\frac{\theta}{\lambda} + n - 1}{n} \left(e^{-\lambda t}\right)^{\frac{\theta}{\lambda}} \left(1 - e^{-\lambda t}\right)^n .$$

Case 3: let $\lambda < 0$ and $-\frac{\theta}{\lambda}$ be an integer so that $\lambda_n = \theta + \lambda n$ if $n < \frac{\theta}{\lambda}$ and 0 if $n \geq -\frac{\theta}{\lambda}$. Then

$$p_n(t) = \binom{-\frac{\theta}{\lambda}}{n} \left(1 - e^{\lambda t}\right)^{n_0} \left(e^{\lambda t}\right)^{-\frac{\theta}{\lambda} - n} .$$

Confronting multiple birth process models to data using the likelihood

Model: a pure birth process with $\lambda_n = \theta + \lambda n$ (immigration + births).

Case 1: suppose $\lambda = 0$ (nothing but immigrations), $n_0 = 0$. Then $p_n(t) = \frac{e^{-\theta t}(\theta t)^n}{n!}$.

Case 2: $\lambda > 0, n_0 = 0$ (Immigration and births). Then

$$p_n(t) = \binom{\frac{\theta}{\lambda} + n - 1}{n} \left(e^{-\lambda t}\right)^{\frac{\theta}{\lambda}} \left(1 - e^{-\lambda t}\right)^n .$$

Case 3: let $\lambda < 0$ and $-\frac{\theta}{\lambda}$ be an integer so that $\lambda_n = \theta + \lambda n$ if $n < \frac{\theta}{\lambda}$ and 0 if $n \geq -\frac{\theta}{\lambda}$. Then

$$p_n(t) = \binom{-\frac{\theta}{\lambda}}{n} \left(1 - e^{\lambda t}\right)^{n_0} \left(e^{\lambda t}\right)^{-\frac{\theta}{\lambda} - n} .$$

In any case, if the observations $(t_0, n_0), (t_1, n_1), (t_2, n_2), \ldots, (t_q, n_q)$ are recorded, the likelihood function is written as

$$L(\lambda, \theta) = p(n_1, \tau_1 | n_0) p(n_2, \tau_2 | n_1) \ldots p(n_q, \tau_q | n_{q-1})$$

Example: a continuous time stochastic SIS model

SIS ODE model:

$$\frac{dI}{dt} = \frac{\beta}{N} S(I + \epsilon) - gI$$

$S = N - I = \#$ of susceptibles, $N =$ total pop. size (cst.)

β is the contact rate,

$\epsilon =$import of infection from an external source ($\epsilon = 0$ if pop. is isolated)

$g =$ recovery rate

Analogous to Levins 1969 metapopulation model (Hosts are empty (S) or occupied(I)).

Example: a continuous time stochastic SIS model

SIS ODE model:

$$\frac{dI}{dt} = \frac{\beta}{N} S(I + \epsilon) - gI$$

$S = N - I = \#$ of susceptibles, $N =$ total pop. size (cst.)

β is the contact rate,

$\epsilon =$import of infection from an external source ($\epsilon = 0$ if pop. is isolated)

$g =$ recovery rate

Analogous to Levins 1969 metapopulation model (Hosts are empty (S) or occupied(I)).

Stochastic version: the states are $I = 0, 1, \ldots, N$. At $t = 0$, $I(0) = k$. So $P(I(0) = k) = 1$ and

$$P(I(t) = i) = p_i(t) = P(I(t) = i | X(0) = k).$$

Kolmogorov-Forward equation

If the process is in state i at time t, then at time $t+\Delta t$ it will be either at state $i+1, i-1$ or i (Δt chosen so that at most 1 event occur). Therefore,

$$\begin{aligned} p_i(t+\Delta t) &= p_{i-1}(t)(\Delta t)\left[\tfrac{\beta}{N}S(t)(I(t)+\epsilon)\right] + p_{i+1}(t)(\Delta t)gI(t) \\ &\quad + p_i(t)\left[1-(\Delta t)\tfrac{\beta}{N}S(t)(I(t)+\epsilon)+gI(t)\right]. \end{aligned}$$

Hence

$$\begin{aligned} \tfrac{p_i(t+\Delta t)-p_i(t)}{\Delta t} &= p_{i-1}(t)\left[\tfrac{\beta}{N}S(t)(I(t)+\epsilon)\right] + p_{i+1}(t)(\Delta t)gI(t) \\ &\quad - p_i(t)\left[\tfrac{\beta}{N}S(t)(I(t)+\epsilon)+gI(t)\right], \end{aligned}$$

and since $S = N - I \,\forall\, t$, and letting $\Delta t \to 0$ we get

$$\frac{dp_i(t)}{dt} = p_{i-1}(t)\frac{\beta}{N}(N-i+1)(i-1+\epsilon) + p_{i+1}(t)g(i+1) - p_i(t)\left[\frac{\beta}{N}(N-i)(i+\epsilon)+gi\right]$$

The transition rates matrix Q

In vector notation,

$$\frac{dp_i(t)}{dt} = p_{i-1}(t)\frac{\beta}{N}(N-i+1)(i-1+\epsilon) + p_{i+1}(t)g(i+1) - p_i(t)\left[\frac{\beta}{N}(N-i)(i+\epsilon) + gi\right]$$

becomes $\frac{d\mathbf{p}}{dt} = \mathbf{p}Q$, where $\dim(\mathbf{p}) = 1 \times (N+1)$ and $\dim(Q) = (N+1) \times (N+1)$:

The transition rates matrix Q

In vector notation,

$$\frac{dp_i(t)}{dt} = p_{i-1}(t)\frac{\beta}{N}(N-i+1)(i-1+\epsilon) + p_{i+1}(t)g(i+1) - p_i(t)\left[\frac{\beta}{N}(N-i)(i+\epsilon)+gi\right]$$

becomes $\frac{d\mathbf{p}}{dt} = \mathbf{p}Q$, where $\dim(\mathbf{p}) = 1 \times (N+1)$ and $\dim(Q) = (N+1)\times(N+1)$:

$$Q = \begin{bmatrix} -\beta\epsilon & \beta\epsilon & 0 & 0 & \dots \\ g & -\left[\frac{\beta}{N}(N-1)(1+\epsilon)+g\right] & \frac{\beta}{N}(N-1)(1+\epsilon) & 0 & \dots \\ 0 & 2g & -\left[\frac{\beta}{N}(N-2)(2+\epsilon)+2g\right] & \frac{\beta}{N}(N-2)(2+\epsilon) & \dots \\ 0 & 0 & 3g & -\left[\frac{\beta}{N}(N-3)(3+\epsilon)+3g\right] & \dots \\ 0 & 0 & 0 & 4g & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

The transition rates matrix Q

In vector notation,

$$\frac{dp_i(t)}{dt} = p_{i-1}(t)\frac{\beta}{N}(N-i+1)(i-1+\epsilon) + p_{i+1}(t)g(i+1) - p_i(t)\left[\frac{\beta}{N}(N-i)(i+\epsilon) + gi\right]$$

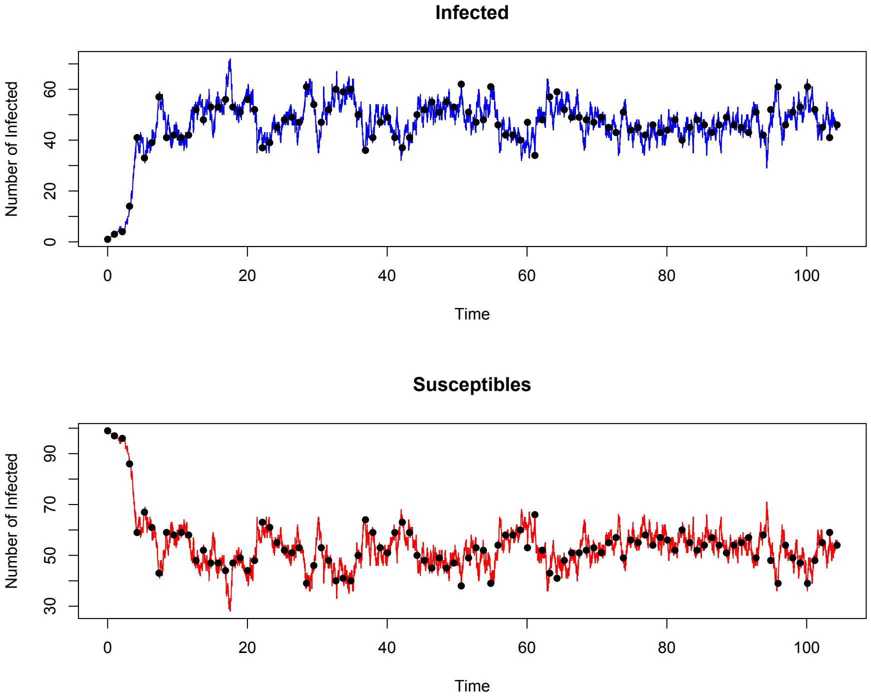
becomes $\frac{d\mathbf{p}}{dt} = \mathbf{p}Q$, where $\dim(\mathbf{p}) = 1 \times (N+1)$ and $\dim(Q) = (N+1) \times (N+1)$:

$$Q = \begin{bmatrix} -\beta\epsilon & \beta\epsilon & 0 & 0 & \ldots \\ g & -\left[\frac{\beta}{N}(N-1)(1+\epsilon)+g\right] & \frac{\beta}{N}(N-1)(1+\epsilon) & 0 & \ldots \\ 0 & 2g & -\left[\frac{\beta}{N}(N-2)(2+\epsilon)+2g\right] & \frac{\beta}{N}(N-2)(2+\epsilon) & \ldots \\ 0 & 0 & 3g & -\left[\frac{\beta}{N}(N-3)(3+\epsilon)+3g\right] & \ldots \\ 0 & 0 & 0 & 4g & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Solution to the system of ODEs: if $\mathbf{p}(0) = \mathbf{p_0}$, then $\mathbf{p_t} = \mathbf{p_0}\exp\{Qt\}$

Observing a realization of the process

- Observations at times $t_1 < t_2 < \ldots < t_{q-1} < t_q$. Let $\tau_i = t_i - t_{i-1}$ as before.
- States: $i_1, i_2, \ldots, i_{q-1}, i_q$



The likelihood function

$$L(\boldsymbol{\theta}) \ = \ P(I(t_1) = i_1, I(t_2) = i_2, \ldots, I(t_{q-1}) = i_{q-1}, I(t_q) = i_q)$$

The likelihood function

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= P(I(t_1) = i_1, I(t_2) = i_2, \ldots, I(t_{q-1}) = i_{q-1}, I(t_q) = i_q) \\
&= P(I(t_1) = i_1) \times P(I(t_2) = i_2 | I(t_1) = i_1) \times P(I(t_q) = i_q | I(t_{q-1}) = i_{q-1})
\end{aligned}
$$

The likelihood function

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= P(I(t_1) = i_1, I(t_2) = i_2, \ldots, I(t_{q-1}) = i_{q-1}, I(t_q) = i_q) \\
&= P(I(t_1) = i_1) \times P(I(t_2) = i_2 | I(t_1) = i_1) \times P(I(t_q) = i_q | I(t_{q-1}) = i_{q-1}) \\
&= \{\mathbf{p_{t_1}}\}_{i_1} \times \{\exp((t_2 - t_1)Q)\}_{i_1,i_2} \times \{\exp((t_3 - t_2)Q)\}_{i_2,i_3} \times \ldots
\end{aligned}
$$

The likelihood function

$$\begin{aligned}
L(\boldsymbol{\theta}) &= P(I(t_1) = i_1, I(t_2) = i_2, \ldots, I(t_{q-1}) = i_{q-1}, I(t_q) = i_q) \\
&= P(I(t_1) = i_1) \times P(I(t_2) = i_2 | I(t_1) = i_1) \times P(I(t_q) = i_q | I(t_{q-1}) = i_{q-1}) \\
&= \{\mathbf{p_{t_1}}\}_{i_1} \times \{\exp((t_2 - t_1)Q)\}_{i_1,i_2} \times \{\exp((t_3 - t_2)Q)\}_{i_2,i_3} \times \ldots \\
&= \{\mathbf{p_{t_1}}\}_{i_1} \times \textstyle\prod_{k=2}^{q} \{\exp(\tau_k Q)\}_{i_{k-1},i_k}
\end{aligned}$$

The likelihood function

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= P(I(t_1) = i_1, I(t_2) = i_2, \ldots, I(t_{q-1}) = i_{q-1}, I(t_q) = i_q) \\
&= P(I(t_1) = i_1) \times P(I(t_2) = i_2 | I(t_1) = i_1) \times P(I(t_q) = i_q | I(t_{q-1}) = i_{q-1}) \\
&= \{\mathbf{p_{t_1}}\}_{i_1} \times \{\exp((t_2 - t_1)Q)\}_{i_1,i_2} \times \{\exp((t_3 - t_2)Q)\}_{i_2,i_3} \times \ldots \\
&= \{\mathbf{p_{t_1}}\}_{i_1} \times \textstyle\prod_{k=2}^{q} \{\exp(\tau_k Q)\}_{i_{k-1},i_k} \\
&= \{\mathbf{p_{t_1}}\}_{i_1} \times \textstyle\prod_{k=2}^{q} \{\mathbf{I}_{i_{k-1}} \times \exp(\tau_k Q)\}_{i_k}, \text{ where}
\end{aligned}
$$

$\mathbf{I}_j$ is a vector that has zeros everywhere, except in the j^{th} position where it has a 1.

The likelihood function

$$\begin{aligned}
L(\boldsymbol{\theta}) &= P(I(t_1)=i_1, I(t_2)=i_2, \ldots, I(t_{q-1})=i_{q-1}, I(t_q)=i_q) \\
&= P(I(t_1)=i_1) \times P(I(t_2)=i_2|I(t_1)=i_1) \times P(I(t_q)=i_q|I(t_{q-1})=i_{q-1}) \\
&= \{\mathbf{p_{t_1}}\}_{i_1} \times \{\exp((t_2-t_1)Q)\}_{i_1,i_2} \times \{\exp((t_3-t_2)Q)\}_{i_2,i_3} \times \ldots \\
&= \{\mathbf{p_{t_1}}\}_{i_1} \times \textstyle\prod_{k=2}^{q}\{\exp((t_k-t_{k-1})Q)\}_{i_{k-1},i_k} \\
&= \{\mathbf{p_{t_1}}\}_{i_1} \times \textstyle\prod_{k=2}^{q}\{\mathbf{I}_{i_{k-1}} \times \exp(\tau_k Q)\}_{i_k}, \text{ where}
\end{aligned}$$

$\mathbf{I}_j$ is a vector that has zeros everywhere, except in the j^{th} position where it has a 1.

Notes: Computing $\exp(\tau Q)$ can be done using a matrix exponentiation algorithm (only once per each iteration of the maximization routine if all τ_k's are equal). However, can greatly reduce computations by calculating $\mathbf{I}_j \exp(\tau Q)$ (a vector) instead of $\exp(\tau Q)$ (a matrix). These are the so-called Krylov space methods. (Citation: On 19 dubious ways. . .)

Matrix exponentiation

A matrix exponentiation is achieved using a T.S. expansion:

$$\exp(\tau Q) = \mathbf{I} + \tau Q + \frac{\tau^2}{2!}Q^2 + \frac{\tau^3}{3!}Q^3 + \ldots$$

Matrix exponentiation

A matrix exponentiation is achieved using a T.S. expansion:

$$\exp(\tau Q) = \mathbf{I} + \tau Q + \frac{\tau^2}{2!}Q^2 + \frac{\tau^3}{3!}Q^3 + \ldots$$

Hence

$$\mathbf{I}_j \exp(\tau Q) = \mathbf{I}_j\mathbf{I} + \mathbf{I}_j\tau Q + \mathbf{I}_j\frac{\tau^2}{2!}Q^2 + \mathbf{I}_j\frac{\tau^3}{3!}Q^3 + \ldots$$

Matrix exponentiation

A matrix exponentiation is achieved using a T.S. expansion:

$$\exp(\tau Q) = \mathbf{I} + \tau Q + \frac{\tau^2}{2!}Q^2 + \frac{\tau^3}{3!}Q^3 + \dots$$

Hence

$$\mathbf{I}_j \exp(\tau Q) = \mathbf{I}_j\mathbf{I} + \mathbf{I}_j\tau Q + \mathbf{I}_j\frac{\tau^2}{2!}Q^2 + \mathbf{I}_j\frac{\tau^3}{3!}Q^3 + \dots$$

- However this is definitely NOT the way to compute it because of accumulation of numerical round-off errors!

Matrix exponentiation

A matrix exponentiation is achieved using a T.S. expansion:

$$\exp(\tau Q) = \mathbf{I} + \tau Q + \frac{\tau^2}{2!} Q^2 + \frac{\tau^3}{3!} Q^3 + \ldots$$

Hence

$$\mathbf{I}_j \exp(\tau Q) = \mathbf{I}_j \mathbf{I} + \mathbf{I}_j \tau Q + \mathbf{I}_j \frac{\tau^2}{2!} Q^2 + \mathbf{I}_j \frac{\tau^3}{3!} Q^3 + \ldots$$

- However this is definitely NOT the way to compute it because of accumulation of numerical round-off errors!
- A program to simulate and estimate parameters for this model using R will be reviewed in the computer session in the afternoon.
- About simulation: with certain computer intensive methods for parameter estimation, all we need is to simulate realizations from the process *conditioned on the ending point*. To do that, use Hobolth and Stone (2009), Annals of Applied Statistics).

Matrix exponentiation

A matrix exponentiation is achieved using a T.S. expansion:

$$\exp(\tau Q) = \mathbf{I} + \tau Q + \frac{\tau^2}{2!} Q^2 + \frac{\tau^3}{3!} Q^3 + \ldots$$

Hence

$$\mathbf{I}_j \exp(\tau Q) = \mathbf{I}_j \mathbf{I} + \mathbf{I}_j \tau Q + \mathbf{I}_j \frac{\tau^2}{2!} Q^2 + \mathbf{I}_j \frac{\tau^3}{3!} Q^3 + \ldots$$

- However this is definitely NOT the way to compute it because of accumulation of numerical round-off errors!
- A program to simulate and estimate parameters for this model using R will be reviewed in the computer session in the afternoon.
- About simulation: with certain computer intensive methods for parameter estimation, all we need is to simulate realizations from the process *conditioned on the ending point*. To do that, use Hobolth and Stone (2009), Annals of Applied Statistics).
- Sampling error?

The likelihood function for waiting times

Model: "Case 1" above. We count the number of births up to time t, where $p_n(t) = \frac{e^{-\theta t}(\theta t)^n}{n!}$. That is,

$$N(t) \sim \mathrm{Poisson}(\theta t).$$

Now let $S > 0$ be a continuous random variable modeling the waiting time until the first birth occurs. Let s denote a realization of S. Consider the following two events:

$$[N(s) = 0] \text{ and } [S > s]$$

The likelihood function for waiting times

Model: “Case 1” above. We count the number of births up to time t, where $p_n(t) = \frac{e^{-\theta t}(\theta t)^n}{n!}$. That is,

$$N(t) \sim \text{Poisson}(\theta t).$$

Now let $S > 0$ be a continuous random variable modeling the waiting time until the first birth occurs. Let s denote a realization of S. Consider the following two events:

$$[N(s) = 0] \text{ and } [S > s]$$

These two events are in fact the same event, so $P[N(s) = 0] = P[S > s]$, which implies that

$$P(N(s) > 0) = 1 - P(N(s) = 0) = P(S \leq s).$$

The likelihood function for waiting times

Model: "Case 1" above. We count the number of births up to time t, where $p_n(t) = \frac{e^{-\theta t}(\theta t)^n}{n!}$. That is,

$$N(t) \sim \text{Poisson}(\theta t).$$

Now let $S > 0$ be a continuous random variable modeling the waiting time until the first birth occurs. Let s denote a realization of S. Consider the following two events:

$$[N(s) = 0] \text{ and } [S > s]$$

These two events are in fact the same event, so $P[N(s) = 0] = P[S > s]$, which implies that

$$P(N(s) > 0) = 1 - P(N(s) = 0) = P(S \leq s).$$

Using our Poisson model for N(s), we then have that

$$F(s) = P(S \leq s) = 1 - P(N(s) = 0) = 1 - e^{-\lambda s}, \quad (0 < s < \infty), \text{ which is the cdf of } S.$$

The likelihood function for waiting times

Model: "Case 1" above. We count the number of births up to time t, where $p_n(t) = \frac{e^{-\theta t}(\theta t)^n}{n!}$. That is,

$$N(t) \sim \text{Poisson}(\theta t).$$

Now let $S > 0$ be a continuous random variable modeling the waiting time until the first birth occurs. Let s denote a realization of S. Consider the following two events:

$$[N(s) = 0] \text{ and } [S > s]$$

These two events are in fact the same event, so $P[N(s) = 0] = P[S > s]$, which implies that

$$P(N(s) > 0) = 1 - P(N(s) = 0) = P(S \leq s).$$

Using our Poisson model for N(s), we then have that

$$F(s) = P(S \leq s) = 1 - P(N(s) = 0) = 1 - e^{-\lambda s}, \quad (0 < s < \infty), \text{ which is the cdf of } S.$$

Graphing this function we see that $F(s)$ is simply telling us that if we wait long enough, we are almost certain to see a birth. Using the cdf we can answer other questions (next slide).

The likelihood function for waiting times (continuous random variables)

Let Δs represent a small positive change in a realized waiting time, so that $(s, s + \Delta s)$ is a small time interval. Then, according to the above calculation we have that

$$P(s < S \leq s + \Delta s) = F(s + \Delta s) - F(s)$$

and dividing both sides of the equation by Δt we get a measure of the *density* of probability over the interval $(s, s + \Delta s)$.

$$\frac{P(s < S \leq s + \Delta s)}{\Delta s} = \frac{F(s + \Delta s) - F(s)}{\Delta s}.$$

The likelihood function for waiting times (continuous random variables)

Let Δs represent a small positive change in a realized waiting time, so that $(s, s+\Delta s)$ is a small time interval. Then, according to the above calculation we have that

$$P(s < S \leq s + \Delta s) = F(s + \Delta s) - F(s)$$

and dividing both sides of the equation by Δt we get a measure of the *density* of probability over the interval $(s, s+\Delta s)$.

$$\frac{P(s < S \leq s + \Delta s)}{\Delta s} = \frac{F(s + \Delta s) - F(s)}{\Delta s}.$$

As $\Delta s \to 0$, the ratio above converges to the derivative of $F(s)$, denoted by $f_S(s)$:

$$\lim_{\Delta s \to 0} \frac{P(s < S \leq s + \Delta s)}{\Delta s} = \frac{dF(s)}{ds} = f_S(s) = \lambda e^{-\theta s}.$$

The derivative of $F(s)$, $f_S(s)$ is the associated probability distribution function of the random variable S. It is the continuous distribution's equivalent to the probability mass function. Thus, by analogy with the discrete case this is the mathematical object that will be used to define the likelihood function, needed to estimate the parameter λ,

The likelihood function for waiting times (continuous random variables)

Let Δs represent a small positive change in a realized waiting time, so that $(s, s + \Delta s)$ is a small time interval. Then, according to the above calculation we have that

$$P(s < S \leq s + \Delta s) = F(s + \Delta s) - F(s)$$

and dividing both sides of the equation by Δt we get a measure of the *density* of probability over the interval $(s, s + \Delta s)$.

$$\frac{P(s < S \leq s + \Delta s)}{\Delta s} = \frac{F(s + \Delta s) - F(s)}{\Delta s}.$$

As $\Delta s \to 0$, the ratio above converges to the derivative of $F(s)$, denoted by $f_S(s)$:

$$\lim_{\Delta s \to 0} \frac{P(s < S \leq s + \Delta s)}{\Delta s} = \frac{dF(s)}{ds} = f_S(s) = \lambda e^{-\theta s}.$$

The derivative of $F(s)$, $f_S(s)$ is the associated probability distribution function of the random variable S. It is the continuous distribution's equivalent to the probability mass function. Thus, by analogy with the discrete case this is the mathematical object that will be used to define the likelihood function, needed to estimate the parameter λ, but not so fast!

The likelihood function for waiting times (continuous random variables)

Suppose we have a collection of observations of the waiting times until the first birth $s_1, s_2, \ldots, s_n$ and we wish to use this info. to estimate the average number of births per unit of time, θ.

The likelihood function for waiting times (continuous random variables)

Suppose we have a collection of observations of the waiting times until the first birth $s_1, s_2, \ldots, s_n$ and we wish to use this info. to estimate the average number of births per unit of time, θ.

Before, we've defined the likelihood as "the joint probability of the observations evaluated at the data at hand". In continuous time, such probabilities are 0!

How do we write down the likelihood function then?

The likelihood function for waiting times (continuous random variables)

Suppose we have a collection of observations of the waiting times until the first birth $s_1, s_2, \ldots, s_n$ and we wish to use this info. to estimate the average number of births per unit of time, θ.

Before, we've defined the likelihood as "the joint probability of the observations evaluated at the data at hand". In continuous time, such probabilities are 0!

How do we write down the likelihood function then?

Suppose that the precision of time-measuring instrument is $\epsilon > 0$. Then, we may calculate the exact likelihood function (Kalbfleisch 1985, Sprott 2000, Pawitan 2001), which consists, for a single observation s_1, of the probability measure over a small interval surrounding the observation:

$$P\left(s_1 - \frac{\epsilon}{2} < S \leq s_1 + \frac{\epsilon}{2}\right) = F\left(s_1 + \frac{\epsilon}{2}\right) - F\left(s_1 - \frac{\epsilon}{2}\right).$$

The likelihood function for waiting times (continuous random variables)

Using the mean value theorem,

$$P\left(s_1 - \frac{\epsilon}{2} < S \leq s_1 + \frac{\epsilon}{2}\right) = F\left(s_1 + \frac{\epsilon}{2}\right) - F\left(s_1 - \frac{\epsilon}{2}\right) \approx \epsilon f(s_1)$$

The likelihood function for waiting times (continuous random variables)

Using the mean value theorem,

$$P\left(s_1 - \frac{\epsilon}{2} < S \leq s_1 + \frac{\epsilon}{2}\right) = F\left(s_1 + \frac{\epsilon}{2}\right) - F\left(s_1 - \frac{\epsilon}{2}\right) \approx \epsilon f(s_1),$$

and the likelihood for the set of recorded waiting times until the first birth is

$$P\left(s_1 - \frac{\epsilon}{2} < S_1 \leq s_1 + \frac{\epsilon}{2}, s_2 - \frac{\epsilon}{2} < S_2 \leq s_2 + \frac{\epsilon}{2}, \ldots, s_n - \frac{\epsilon}{2} < S_n \leq s_n + \frac{\epsilon}{2}\right) =$$

$$P\left(s_1 - \frac{\epsilon}{2} < S_1 \leq s_1 + \frac{\epsilon}{2}\right) P\left(s_2 - \frac{\epsilon}{2} < S_2 \leq s_2 + \frac{\epsilon}{2}\right) \ldots P\left(s_n - \frac{\epsilon}{2} < S_n \leq s_n + \frac{\epsilon}{2}\right),$$

which can be approximated with

$$f_S(s_1) f_S(s_2) \ldots f_S(s_n) \epsilon^n.$$

The likelihood function for waiting times (continuous random variables)

Using the mean value theorem,

$$P\left(s_1 - \frac{\epsilon}{2} < S \leq s_1 + \frac{\epsilon}{2}\right) = F\left(s_1 + \frac{\epsilon}{2}\right) - F\left(s_1 - \frac{\epsilon}{2}\right) \approx \epsilon f(s_1),$$

and the likelihood for the set of recorded waiting times until the first birth is

$$P\left(s_1 - \frac{\epsilon}{2} < S_1 \leq s_1 + \frac{\epsilon}{2}, s_2 - \frac{\epsilon}{2} < S_2 \leq s_2 + \frac{\epsilon}{2}, \ldots, s_n - \frac{\epsilon}{2} < S_n \leq s_n + \frac{\epsilon}{2}\right) =$$

$$P\left(s_1 - \frac{\epsilon}{2} < S_1 \leq s_1 + \frac{\epsilon}{2}\right) P\left(s_2 - \frac{\epsilon}{2} < S_2 \leq s_2 + \frac{\epsilon}{2}\right) \ldots P\left(s_n - \frac{\epsilon}{2} < S_n \leq s_n + \frac{\epsilon}{2}\right).$$

Which can be approximated with

$$f_S(s_1) f_S(s_2) \ldots f_S(s_n) \epsilon^n.$$

Some comments are in order

- Usual undergrad math/stats books: “likelihood for continuous models is the pdf evaluated at the observations”.

Comments. . .

- The above arguments show that such definition is in fact an approximation to the *exact likelihood function*

Comments. . .

- The above arguments show that such definition is in fact an approximation to the *exact likelihood function*
- Some critique the likelihood function because the profile based on the pdf approximation can have singularities

Comments...

- The above arguments show that such definition is in fact an approximation to the *exact likelihood function*
- Some critique the likelihood function because the profile based on the pdf approximation can have singularities
- However, when the *exact likelihood* is computed, such numerical problems disappear: the exact likelihood it is a product of cdf values, hence always bounded between 0 and 1!! (Exemplified in Montoya et al 2009)

Comments. . .

- The above arguments show that such definition is in fact an approximation to the *exact likelihood function*
- Some critique the likelihood function because the profile based on the pdf approximation can have singularities
- However, when the *exact likelihood* is computed, such numerical problems disappear: the exact likelihood it is a product of cdf values, hence always bounded between 0 and 1!! (Exemplified in Montoya et al 2009)
- To estimate the parameter θ, one has to maximize $f_S(s_1)f_S(s_2)\dots f_S(s_n)\epsilon^n$ with respect to θ

Comments...

- The above arguments show that such definition is in fact an approximation to the *exact likelihood function*

- Some critique the likelihood function because the profile based on the pdf approximation can have singularities

- However, when the *exact likelihood* is computed, such numerical problems disappear: the exact likelihood it is a product of cdf values, hence always bounded between 0 and 1!! (Exemplified in Montoya et al 2009)

- To estimate the parameter θ, one has to maximize $f_S(s_1)f_S(s_2)\ldots f_S(s_n)\epsilon^n$ with respect to θ

- Amounts to maximizing $f_S(s_1)f_S(s_2)\ldots f_S(s_n)$ only if ϵ does not depend on θ

Comments. . .

- The above arguments show that such definition is in fact an approximation to the *exact likelihood function*
- Some critique the likelihood function because the profile based on the pdf approximation can have singularities
- However, when the *exact likelihood* is computed, such numerical problems disappear: the exact likelihood it is a product of cdf values, hence always bounded between 0 and 1!! (Exemplified in Montoya et al 2009)
- To estimate the parameter θ, one has to maximize $f_S(s_1)f_S(s_2)\ldots f_S(s_n)\epsilon^n$ with respect to θ
- Amounts to maximizing $f_S(s_1)f_S(s_2)\ldots f_S(s_n)$ only if ϵ does not depend on θ
- Such dependence can occur if the size of the mean number of events per unit of time affects the precision of the instrument (exhausted batteries?)

The ML estimate of θ from waiting times data

Finally, if ϵ does not depend on θ, then

$$L(\theta) \propto f_S(s_1) f_S(s_2) \dots f_S(s_n).$$

and the log-likelihood is

$$\begin{aligned} \ln L(\theta) &\propto \ln \left(\theta^n \exp -\theta \textstyle\sum_{i=1}^n s_i\right) \\ &= n \ln \theta - \theta \textstyle\sum_{i=1}^n s_i, \end{aligned}$$

which allows us to comput the ML estimate of θ:

$$\begin{aligned} \tfrac{d \ln \ell(\theta)}{d\theta} &= \tfrac{n}{\theta} - \textstyle\sum_{i=1}^n s_i = 0 \\ &\Rightarrow \hat{\theta} = \tfrac{n}{\sum_{i=1}^n s_i} = \tfrac{1}{\bar{s}}. \end{aligned}$$

If the data consists of the waiting times until the k^{th} event, denoted S_k, then the preceding argument can be extended.

Waiting time until the $k^{\textbf{th}}$ event

As before, we count the number of births up to time t, where $p_n(t) = \frac{e^{-\theta t}(\theta t)^n}{n!}$. That is,

$$N(t) \sim \mathrm{Poisson}(\theta t).$$

Waiting time until the k^{th} event

As before, we count the number of births up to time t, where $p_n(t) = \frac{e^{-\theta t}(\theta t)^n}{n!}$. That is,

$$N(t) \sim \text{Poisson}(\theta t).$$

Now let $S_k > 0$ be a continuous random variable modeling the waiting time until the k^{th} birth occurs. Let s denote a realization of S. Consider the following two events:

$$[N(s) < k] \text{ and } [S_k > s]$$

Waiting time until the k^{th} event

As before, we count the number of births up to time t, where $p_n(t) = \frac{e^{-\theta t}(\theta t)^n}{n!}$. That is,

$$N(t) \sim \mathrm{Poisson}(\theta t).$$

Now let $S_k > 0$ be a continuous random variable modeling the waiting time until the k^{th} birth occurs. Let s denote a realization of S. Consider the following two events:

$$[N(s) < k] \text{ and } [S_k > s]$$

These two events are in fact the same event, so $P[N(s) < k] = P[S_k > s]$ and we can use our Poisson model for $N(s)$ to find that

$$F_k(s) = P(S_k \leq s) = 1 - P(N(s) < k) = 1 - \sum_{x=0}^{k-1} \frac{e^{-(\theta s)}(\theta s)^n}{n!}, \quad (0 < s < \infty),$$

which is the cdf of S_k.

Waiting time until the $k^{\textbf{th}}$ event

As before, we count the number of births up to time t, where $p_n(t) = \frac{e^{-\theta t}(\theta t)^n}{n!}$. That is,

$$N(t) \sim \text{Poisson}(\theta t).$$

Now let $S_k > 0$ be a continuous random variable modeling the waiting time until the k^{th} birth occurs. Let s denote a realization of S. Consider the following two events:

$$[N(s) < k] \text{ and } [S_k > s]$$

These two events are in fact the same event, so $P[N(s) < k] = P[S_k > s]$ and we can use our Poisson model for $N(s)$ to find that

$$F_k(s) = P(S_k \leq s) = 1 - P(N(s) < k) = 1 - \sum_{x=0}^{k-1} \frac{e^{-(\theta s)}(\theta s)^n}{n!}, \quad (0 < s < \infty),$$

which is the cdf of S_k. Just as with the exponential model, we can find the probability density function of the waiting time until capturing the k^{th} by taking the derivative of $F_k(s)$ with respect to s

Waiting time until the kth event

$$f_{S_k}(s) = \tfrac{d}{ds}\left[1 - \textstyle\sum_{n=0}^{k-1} \tfrac{e^{-(\theta s)}(\theta s)^n}{n!}\right]$$

Waiting time until the k^{th} event

$$f_{S_k}(s) = \tfrac{d}{ds}\left[1 - \textstyle\sum_{n=0}^{k-1} \tfrac{e^{-(\theta s)}(\theta s)^n}{n!}\right] \; = \; -\textstyle\sum_{n=0}^{k-1}\left[-\tfrac{\theta e^{-(\theta s)}(\theta s)^n}{n!} + \tfrac{e^{-(\theta s)} n s^{x-1}\theta^n}{n!}\right]$$

Waiting time until the k^{th} event

$$f_{S_k}(s) = \tfrac{d}{ds}\left[1 - \textstyle\sum_{n=0}^{k-1} \frac{e^{-(\theta s)}(\theta s)^n}{n!}\right] \;=\; -\textstyle\sum_{n=0}^{k-1}\left[-\frac{\theta e^{-(\theta s)}(\theta s)^n}{n!} + \frac{e^{-(\theta s)} n s^{x-1}\theta^n}{n!}\right]$$

$$=\; \textstyle\sum_{n=0}^{k-1}\left[\frac{\theta e^{-(\theta s)}(\theta s)^n}{n!} - \frac{e^{-(\theta s)} n s^{n-1}\theta^n}{n!}\right]$$

Waiting time until the k^{th} event

$$
\begin{aligned}
f_{S_k}(s) = \tfrac{d}{ds}\left[1 - \textstyle\sum_{n=0}^{k-1} \frac{e^{-(\theta s)}(\theta s)^n}{n!}\right] \;&=\; -\textstyle\sum_{n=0}^{k-1}\left[-\frac{\theta e^{-(\theta s)}(\theta s)^n}{n!} + \frac{e^{-(\theta s)} n s^{x-1}\theta^n}{n!}\right] \\
&=\; \textstyle\sum_{n=0}^{k-1}\left[\frac{\theta e^{-(\theta s)}(\theta s)^n}{n!} - \frac{e^{-(\theta s)} n s^{n-1}\theta^n}{n!}\right] \\
&=\; \textstyle\sum_{n=0}^{k-1}\left[\frac{\theta^{n+1} e^{-(\theta s)} s^n}{n!}\right] - \textstyle\sum_{n=1}^{k-1} \frac{\theta^n s^{n-1} e^{-\theta s}}{(n-1)!}
\end{aligned}
$$

Waiting time until the k^{th} event

$$f_{S_k}(s) = \frac{d}{ds}\left[1 - \sum\nolimits_{n=0}^{k-1} \frac{e^{-(\theta s)}(\theta s)^n}{n!}\right] \; = \; -\sum\nolimits_{n=0}^{k-1}\left[-\frac{\theta e^{-(\lambda s)}(\theta s)^n}{n!} + \frac{e^{-(\theta s)} n s^{x-1}\theta^n}{n!}\right]$$

$$= \; \sum\nolimits_{n=0}^{k-1}\left[\frac{\theta e^{-(\theta s)}(\theta s)^n}{n!} - \frac{e^{-(\theta s)} n s^{n-1}\theta^n}{n!}\right]$$

$$= \; \sum\nolimits_{n=0}^{k-1}\left[\frac{\theta^{n+1} e^{-(\theta s)} s^n}{n!}\right] - \sum\nolimits_{n=1}^{k-1}\frac{\theta^n s^{n-1} e^{-\theta s}}{(n-1)!}$$

Now, factor out the term $e^{-\theta s}$ and explicitly write down the sums:

$$f_k(s) = e^{-\theta s}\left\{\begin{array}{l} \theta + \theta^2 s + \frac{\theta^3 s^2}{2!} + \ldots + \frac{\theta^{k-1}s^{k-2}}{(k-2)!} + \frac{\theta^k s^{k-1}}{(k-1)!} \\ \\ \\ \end{array}\right.$$

Waiting time until the $k^{\textbf{th}}$ event

$$\begin{aligned} f_{S_k}(s) = \tfrac{d}{ds}\left[1 - \textstyle\sum_{n=0}^{k-1} \frac{e^{-(\theta s)}(\theta s)^n}{n!}\right] \; &= \; -\textstyle\sum_{n=0}^{k-1}\left[-\frac{\theta e^{-(\lambda s)}(\theta s)^n}{n!} + \frac{e^{-(\theta s)} n s^{x-1}\theta^n}{n!}\right] \\ &= \; \textstyle\sum_{n=0}^{k-1}\left[\frac{\theta e^{-(\theta s)}(\theta s)^n}{n!} - \frac{e^{-(\theta s)} n s^{n-1}\theta^n}{n!}\right] \\ &= \; \textstyle\sum_{n=0}^{k-1}\left[\frac{\theta^{n+1} e^{-(\theta s)} s^n}{n!}\right] - \textstyle\sum_{n=1}^{k-1} \frac{\theta^n s^{n-1} e^{-\theta s}}{(n-1)!} \end{aligned}$$

Now, factor out the term $e^{-\theta s}$ and explicitly write down the sums:

$$f_k(s) = e^{-\theta s}\begin{cases} \theta + \theta^2 s + \frac{\theta^3 s^2}{2!} + \ldots + \frac{\theta^{k-1}s^{k-2}}{(k-2)!} + \frac{\theta^k s^{k-1}}{(k-1)!} \\ \\ -\theta - \theta^2 s - \frac{\theta^3 s^2}{2!} - \ldots - \frac{\theta^{k-1}s^{k-2}}{(k-2)!} \end{cases}$$

Waiting time until the $k^{\textbf{th}}$ event

$$f_{S_k}(s) = \tfrac{d}{ds}\left[1 - \textstyle\sum_{n=0}^{k-1} \frac{e^{-(\theta s)}(\theta s)^n}{n!}\right] \;=\; -\textstyle\sum_{n=0}^{k-1}\left[-\frac{\theta e^{-(\lambda s)}(\theta s)^n}{n!} + \frac{e^{-(\theta s)} n s^{x-1}\theta^n}{n!}\right]$$

$$= \;\textstyle\sum_{n=0}^{k-1}\left[\frac{\theta e^{-(\theta s)}(\theta s)^n}{n!} - \frac{e^{-(\theta s)} n s^{n-1}\theta^n}{n!}\right]$$

$$= \;\textstyle\sum_{n=0}^{k-1}\left[\frac{\theta^{n+1} e^{-(\theta s)} s^n}{n!}\right] - \textstyle\sum_{n=1}^{k-1} \frac{\theta^n s^{n-1} e^{-\theta s}}{(n-1)!}$$

Now, factor out the term $e^{-\theta s}$ and explicitly write down the sums:

$$f_k(s) = e^{-\theta s}\begin{cases} \theta + \theta^2 s + \frac{\theta^3 s^2}{2!} + \ldots + \frac{\theta^{k-1} s^{k-2}}{(k-2)!} + \frac{\theta^k s^{k-1}}{(k-1)!} \\ \\ -\theta - \theta^2 s - \frac{\theta^3 s^2}{2!} - \ldots - \frac{\theta^{k-1} s^{k-2}}{(k-2)!} \end{cases}$$

We get a *telescoping sum*: all the terms cancel except the last one!

Waiting time until the k^{th} event

$$\begin{aligned} f_{S_k}(s) = \tfrac{d}{ds}\left[1 - \textstyle\sum_{n=0}^{k-1} \tfrac{e^{-(\theta s)}(\theta s)^n}{n!}\right] &= -\textstyle\sum_{n=0}^{k-1}\left[-\tfrac{\theta e^{-(\lambda s)}(\theta s)^n}{n!} + \tfrac{e^{-(\theta s)} n s^{x-1}\theta^n}{n!}\right] \\ &= \textstyle\sum_{n=0}^{k-1}\left[\tfrac{\theta e^{-(\theta s)}(\theta s)^n}{n!} - \tfrac{e^{-(\theta s)} n s^{n-1}\theta^n}{n!}\right] \\ &= \textstyle\sum_{n=0}^{k-1}\left[\tfrac{\theta^{n+1} e^{-(\theta s)} s^n}{n!}\right] - \textstyle\sum_{n=1}^{k-1} \tfrac{\theta^n s^{n-1} e^{-\theta s}}{(n-1)!} \end{aligned}$$

Now, factor out the term $e^{-\theta s}$ and explicitly write down the sums:

$$f_k(s) = e^{-\theta s}\begin{cases} \theta + \theta^2 s + \tfrac{\theta^3 s^2}{2!} + \ldots + \tfrac{\theta^{k-1}s^{k-2}}{(k-2)!} + \tfrac{\theta^k s^{k-1}}{(k-1)!} \\ \\ -\theta - \theta^2 s - \tfrac{\theta^3 s^2}{2!} - \ldots - \tfrac{\theta^{k-1}s^{k-2}}{(k-2)!} \end{cases}$$

We get a *telescoping sum*: all the terms cancel except the last one! Therefore, the above equation reduces to

$$f_k(s) = \frac{e^{-\theta s}\theta^k s^{k-1}}{(k-1)!}, \quad 0 < s < \infty, \text{ which is a Gamma pdf.}$$

Waiting time until the $k^{\textbf{th}}$ event

We got that

$$f_k(s) = \frac{e^{-\theta s}\theta^k s^{k-1}}{(k-1)!}, \quad 0 < s < \infty, \text{ which is a Gamma pdf.}$$

Therefore, the likelihood function for a series of independent observations of the waiting times until the k^{th} birth, $s_1, s_2, \ldots, s_n$ is (if ϵ does not depend on θ)

$$L(\theta) \propto f_{S_k}(s_1) f_{S_k}(s_2) \ldots f_{S_k}(s_n).$$

Waiting time until the k^{th} event

We got that

$$f_k(s) = \frac{e^{-\theta s}\theta^k s^{k-1}}{(k-1)!}, \quad 0 < s < \infty, \text{ which is a Gamma pdf.}$$

Therefore, the likelihood function for a series of independent observations of the waiting times until the k^{th} birth, $s_1, s_2, \dots, s_n$ is (if ϵ does not depend on θ)

$$L(\theta) \propto f_{S_k}(s_1) f_{S_k}(s_2) \dots f_{S_k}(s_n).$$

Note: Using the general formulation of the gamma pdf we get that

$$\begin{aligned} P(S \leq s) &= \textstyle\int_0^s \frac{\theta^k}{\Gamma(k)} s^{k-1} e^{-\theta s} ds \\ &= 1 - \textstyle\sum_{n=0}^{k-1} \frac{e^{-s\theta}(\theta s)}{n!} \\ &= \textstyle\sum_{n=k}^{\infty} \frac{e^{-s\theta}(\theta s)}{n!}, \quad \text{which is the right tail of the initial Poisson model.} \end{aligned}$$

Thus, the right tail (*i.e.* from k to ∞) of our initial probabilistic model of the number of births during a period of time s is in fact identical to the left tail of the resulting gamma model of the waiting time until the k^{th} birth occurs.

The likelihood function and further inference questions

- A biologist might not necessarily be interested in estimating the parameters but rather, in knowing which scenario best explains the data (i.e. Does p vary per sex, season, year, according to rain,. . .)
- How can the likelihood function help us decide amongst a suite of models?
  - Answer, case 1: pairwise model selection
  - Answer, case 2: multiple models
- Example: Are the non-linearities introduced by the theta-Ricker model necessary to explain a time series with demographic and environmental stochasticities?

The likelihood function for the model with environmental and demographic stochasticities:

Let $\lambda \sim \mathrm{Gamma}(k, \alpha)$ represent the environmental noise and let each individual have a Poisson offspring distribution then we saw that the transition pdf was

$$P(N_{t+1} = n_{t+1} | N_t = n_t) = \frac{\Gamma(n_{t+1} + k)}{\Gamma(k) n_{t+1}!} \left(\frac{\alpha}{n_t p_t + \alpha} \right)^k \left(\frac{n_t p_t}{n_t p_t + \alpha} \right)^{n_{t+1}},$$

where

$$p_t = \exp\left\{ -b\, n_t^{\theta} \right\} \quad \text{for theta-Ricker model.}$$

Given a time series data set consisting of the (exact) counts $n_0, n_1, \ldots, n_q$, then the likelihood function for the parameters $\boldsymbol{\theta} = [k, \alpha, b, \theta]'$is again the joint pmf of the population sizes $N_1, \ldots, N_q$ evaluated at the data at hand:

$$L(\boldsymbol{\theta}) = \prod_{t=0}^{q-1} P(N_{t+1} = n_{t+1} | N_t = n_t).$$

If N_0 is an observation from the stationary distribution of the process, then

$$L(\boldsymbol{\theta}) = P(N_0 = n_0) \times \prod_{t=0}^{q-1} P(N_{t+1} = n_{t+1} | N_t = n_t).$$

The likelihood function and further inference questions

- A biologist might not necessarily be interested in estimating the parameters but rather, in knowing which scenario best explains the data (i.e. Does p vary per sex, season, year, according to rain,. . .)
- How can the likelihood function help us decide amongst a suite of models?
  - Answer, case 1: pairwise model selection
  - Answer, case 2: multiple models
- Example: Are the non-linearities introduced by the theta-Ricker model necessary to explain a time series with demographic and environmental stochasticities?

$$H_0 : \theta = 1, \quad \text{and we let } k, \alpha, b \text{ vary freely}$$

$$H_1 : \theta \neq 1. \quad \text{We let all the parameters vary freely.}$$

A generalization of the theta-Ricker modeling question: Large sample Likelihood Ratio Tests

Consider the following setting, where the null (restricted) hypothesis is given by

$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta_0} = [c, \theta_2, \theta_3, \ldots, \theta_r]$.

$L_{H_0}(\boldsymbol{\theta_0})$ is maximized at $\tilde{\boldsymbol{\theta}}_{\mathbf{0}} = [c, \tilde{\theta}_2, \tilde{\theta}_3, \ldots, \tilde{\theta}_r]$.

and the alternative (unrestricted) hypothesis is

$H_1 : \boldsymbol{\theta} = \boldsymbol{\theta_0} = [\theta_1, \theta_2, \theta_3, \ldots, \theta_r]$

$L_{H_1}(\boldsymbol{\theta})$ is maximized at $\widehat{\boldsymbol{\theta}} = [\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \ldots, \hat{\theta}_r]$

A generalization of the theta-Ricker modeling question: Large sample Likelihood Ratio Tests

Consider the following setting, where the null (restricted) hypothesis is given by

$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta_0} = [c, \theta_2, \theta_3, \ldots, \theta_r]$.

$L_{H_0}(\boldsymbol{\theta_0})$ is maximized at $\tilde{\boldsymbol{\theta_0}} = [c, \tilde{\theta}_2, \tilde{\theta}_3, \ldots, \tilde{\theta}_r]$.

and the alternative (unrestricted) hypothesis is

$H_1 : \boldsymbol{\theta} = \boldsymbol{\theta_0} = [\theta_1, \theta_2, \theta_3, \ldots, \theta_r]$

$L_{H_1}(\boldsymbol{\theta})$ is maximized at $\widehat{\boldsymbol{\theta}} = [\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \ldots, \hat{\theta}_r]$

Alternatively, H_0 specifies $\boldsymbol{\theta}$ as depending on $q < r$ underlying parameters:

$$\begin{aligned} \theta_1 &= h_1(\xi_1, \xi_2, \ldots, \xi_q) \\ &\vdots \\ \theta_r &= h_r(\xi_1, \xi_2, \ldots, \xi_q) \end{aligned}$$

Generalized Likelihood Ratio Test

Theorem (Samuel Wilks): under regularity conditions, if H_0 is true, then the statistic

$$G^2 = -2\ln\Lambda = -2\ln\left[\frac{L_{H_0}(\tilde{\boldsymbol{\theta}}_{\mathbf{0}})}{L_{H_1}(\widehat{\boldsymbol{\theta}})}\right] \xrightarrow{\text{d}} \chi^2_{(s)},$$

where $s =$ number of restrictions $= r - q =$ number of parameters estimated under $H_1 -$ number of parameters estimated under H_0

- The parameters under the null can be made a function of q other parameters $(q < r)$.
- Regularity conditions are the same as those of ML estimation (See Dennis & Taper 1994)!
- The alternative model is not restricted
- Decision rule: Reject H_0 in favor of H_1 if $G^2_{\text{obs}} \geq \chi^2_{(s)}(\alpha)$, where $\alpha =$ significance level.

Generalized Likelihood Ratio Tests and Confidence Intervals

A $100(1-\alpha)$% CI for θ_1 is the set of all c's for which H_0 would not be rejected at a significance level α.

Generalized Likelihood Ratio Tests and Confidence Intervals

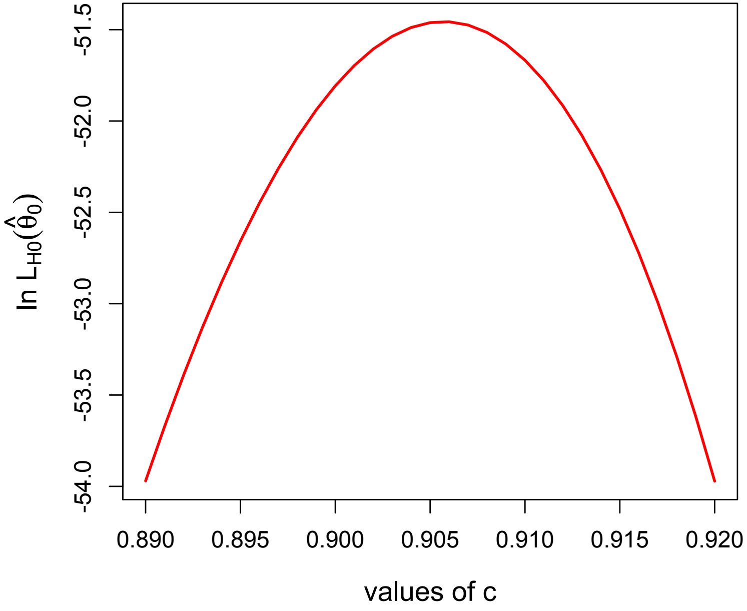
A $100(1-\alpha)$% CI for θ_1 is the set of all c's for which H_0 would not be rejected at a significance level α. Reject H_0 if $G^2_{\text{obs}} \geq \chi^2_{(1)}(\alpha)$

Generalized Likelihood Ratio Tests and Confidence Intervals

A $100(1-\alpha)$% CI for θ_1 is the set of all c's for which H_0 would not be rejected at a significance level α. Reject H_0 if $G^2_{\text{obs}} \geq \chi^2_{(1)}(\alpha)$,

$$\begin{aligned}
\Rightarrow -2\ln\left[\frac{L_{H_0}(\tilde{\boldsymbol{\theta}}_{\mathbf{0}})}{L_{H_1}(\widehat{\boldsymbol{\theta}})}\right] &\geq \chi^2_{(1)}(\alpha) \\
\Rightarrow -2\left[\ln L_{H_0}(\tilde{\boldsymbol{\theta}}_{\mathbf{0}}) - \ln L_{H_1}(\widehat{\boldsymbol{\theta}})\right] &\geq \chi^2_{(1)}(\alpha) \\
\Rightarrow \ln L_{H_1}(\widehat{\boldsymbol{\theta}}) - \ln L_{H_0}(\tilde{\boldsymbol{\theta}}_{\mathbf{0}}) &\geq \frac{\chi^2_{(1)}(\alpha)}{2} \\
\Rightarrow \ln L_{H_1}(\widehat{\boldsymbol{\theta}}) - \frac{\chi^2_{(1)}(\alpha)}{2} &\geq \ln L_{H_0}(\tilde{\boldsymbol{\theta}}_{\mathbf{0}})
\end{aligned}$$

Generalized Likelihood Ratio Tests and Confidence Intervals



Remember that $\tilde{\boldsymbol{\theta}}_{\mathbf{0}} = [c, \ \underbrace{\tilde{\theta}_2, \tilde{\theta}_3, \ldots, \tilde{\theta}_r}_{\text{maximize } r-1 \text{ params.}} \ ]$

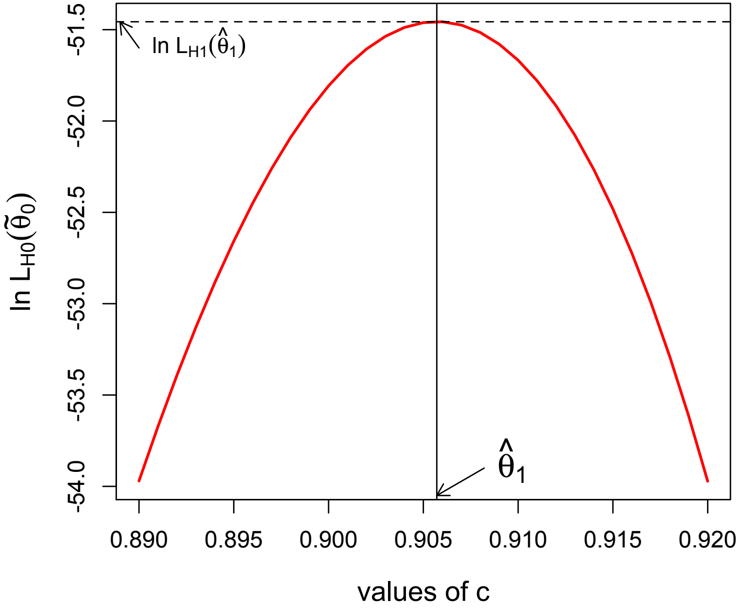
and that $\widehat{\boldsymbol{\theta}} = \underbrace{[\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \ldots, \hat{\theta}_r]}_{\text{maximize } r \text{ params.}}$.

Now, $\frac{\chi^2_{(1)}(\alpha)}{2} = \frac{3.843}{2} = 1.9215$ so reject H_0 if

$$\ln L_{H_1}(\widehat{\boldsymbol{\theta}}) - \frac{\chi^2_{(1)}(\alpha)}{2} \geq \ln L_{H_0}(\tilde{\boldsymbol{\theta}}_{\mathbf{0}})$$

$$\Rightarrow \ln L_{H_1}(\widehat{\boldsymbol{\theta}}) - 1.9215 \geq \ln L_{H_0}(\tilde{\boldsymbol{\theta}}_{\mathbf{0}})$$

Generalized Likelihood Ratio Tests and Confidence Intervals



Remember that $\tilde{\boldsymbol{\theta}}_{\mathbf{0}} = [c, \ \underbrace{\tilde{\theta}_2, \tilde{\theta}_3, \ldots, \tilde{\theta}_r}_{\text{maximize } r-1 \text{ params.}} \ ]$

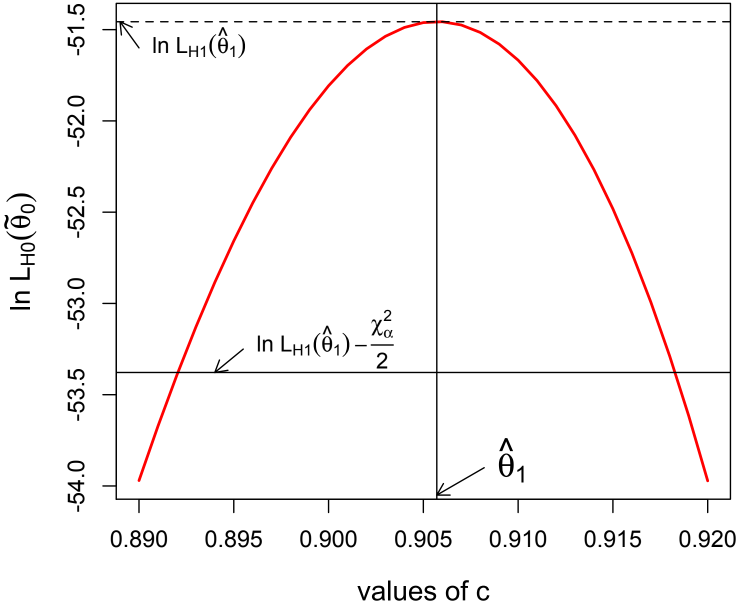
and that $\widehat{\boldsymbol{\theta}} = \underbrace{[\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \ldots, \hat{\theta}_r]}_{\text{maximize } r \text{ params.}}.$

Now, $\frac{\chi^2_{(1)}(\alpha)}{2} = \frac{3.843}{2} = 1.9215$ so reject H_0 if

$$\ln L_{H_1}(\widehat{\boldsymbol{\theta}}) - \frac{\chi^2_{(1)}(\alpha)}{2} \geq \ln L_{H_0}(\tilde{\boldsymbol{\theta}}_{\mathbf{0}})$$

$$\Rightarrow \ln L_{H_1}(\widehat{\boldsymbol{\theta}}) - 1.9215 \geq \ln L_{H_0}(\tilde{\boldsymbol{\theta}}_{\mathbf{0}})$$

Generalized Likelihood Ratio Tests and Confidence Intervals



Remember that $\tilde{\boldsymbol{\theta}}_{\mathbf{0}} = [c, \underbrace{\tilde{\theta}_2, \tilde{\theta}_3, \ldots, \tilde{\theta}_r}_{\text{maximize } r-1 \text{ params.}}]$

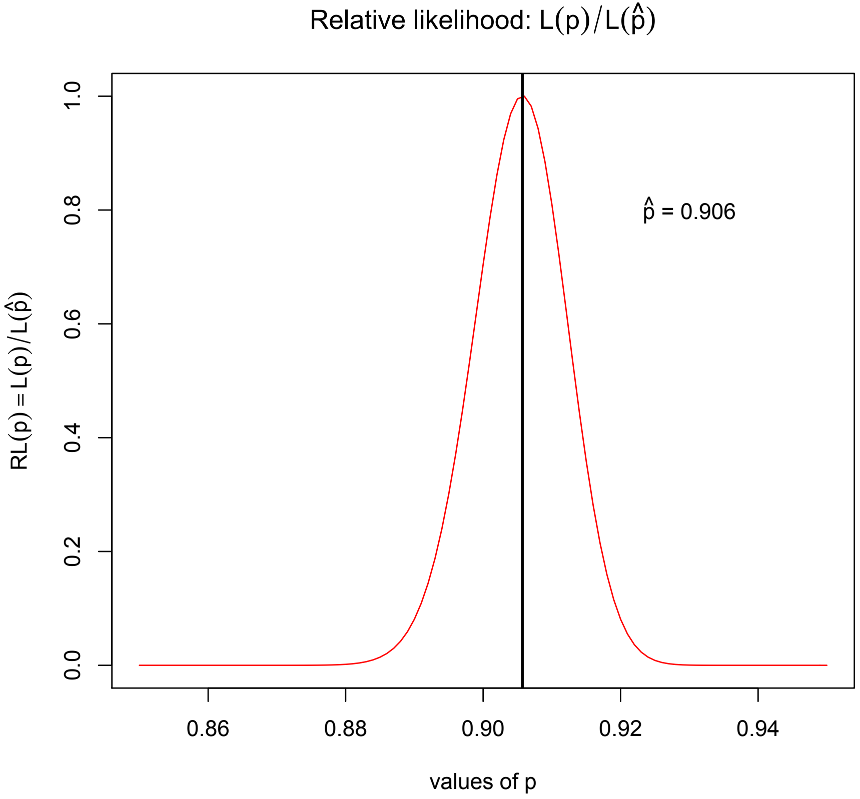
and that $\widehat{\boldsymbol{\theta}} = \underbrace{[\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \ldots, \hat{\theta}_r]}_{\text{maximize } r \text{ params.}}$.

Now, $\frac{\chi^2_{(1)}(\alpha)}{2} = \frac{3.843}{2} = 1.9215$ so reject H_0 if

$$\ln L_{H_1}(\widehat{\boldsymbol{\theta}}) - \frac{\chi^2_{(1)}(\alpha)}{2} \geq \ln L_{H_0}(\tilde{\boldsymbol{\theta}}_{\mathbf{0}})$$

$$\Rightarrow \ln L_{H_1}(\widehat{\boldsymbol{\theta}}) - 1.9215 \geq \ln L_{H_0}(\tilde{\boldsymbol{\theta}}_{\mathbf{0}})$$

The relative likelihood function



Maximizing $L(p)$: set $\frac{dL(p)}{dp} = 0$, solve for p

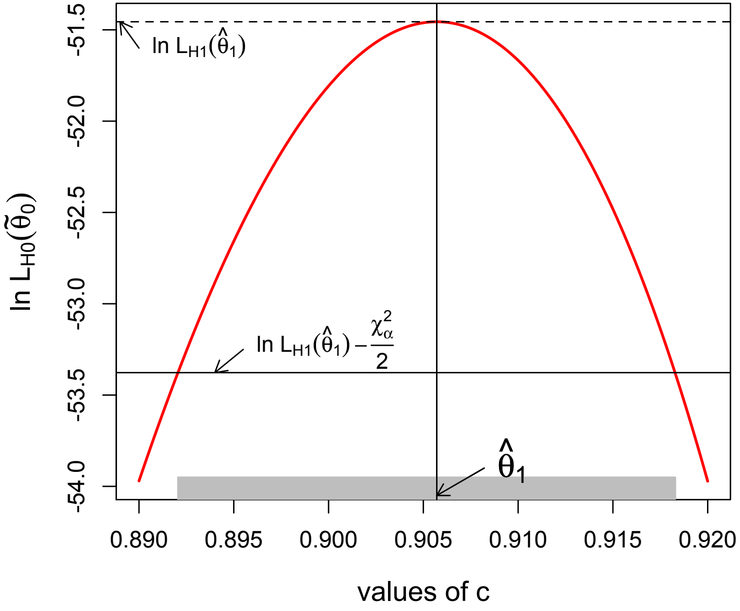
Amounts to set $\frac{1}{L(p)}\frac{dL(p)}{dp} = 0$, solve for p. That is,

$$\frac{d\ln L(p)}{dp} = \frac{d}{dp}\left[\sum_{i=1}^{24} \ln \binom{n_{i-1}}{n_i} p^{n_i}(1-p)^{n_{i-1}-n_i}\right]$$

$$\Rightarrow \frac{d\ln L(p)}{dp} \propto \frac{\sum_{i=1}^{24} n_i}{p} - \frac{\sum_{i=1}^{24} n_{i-1}-n_i}{(1-p)} = 0$$

$$\Rightarrow \hat{p} = \frac{\sum_{i=1}^{24} n_i}{\sum_{i=1}^{24} n_{i-1}} = 0.906$$

Generalized Likelihood Ratio Tests and Confidence Intervals



Remember that $\tilde{\boldsymbol{\theta}}_{\mathbf{0}} = [c,\ \underbrace{\tilde{\theta}_2, \tilde{\theta}_3, \ldots, \tilde{\theta}_r}_{\text{maximize } r-1 \text{ params.}}\ ]$

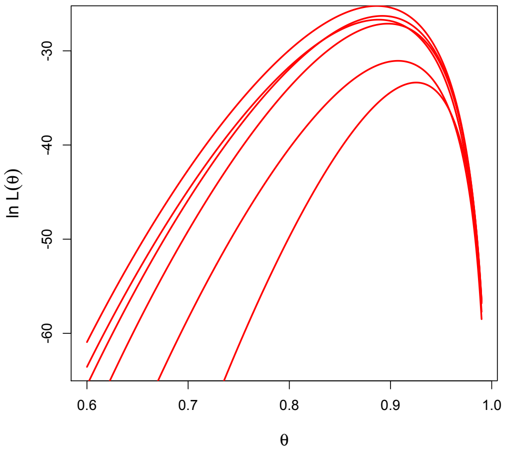
and that $\widehat{\boldsymbol{\theta}} = \underbrace{[\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \ldots, \hat{\theta}_r]}_{\text{maximize } r \text{ params.}}.$

Now, $\frac{\chi^2_{(1)}(\alpha)}{2} = \frac{3.843}{2} = 1.9215$ so reject H_0 if

$$\ln L_{H_1}(\widehat{\boldsymbol{\theta}}) - 1.9215 \geq \ln L_{H_0}(\tilde{\boldsymbol{\theta}}_{\mathbf{0}})$$

Profile likelihood CI: the set of values of c for which we fail to reject H_0!

Fisher's information and asymptotic Wald's C.I.



Let $X_1, X_2, \ldots, X_n$ be a sample of size n, X_i iid or ind.

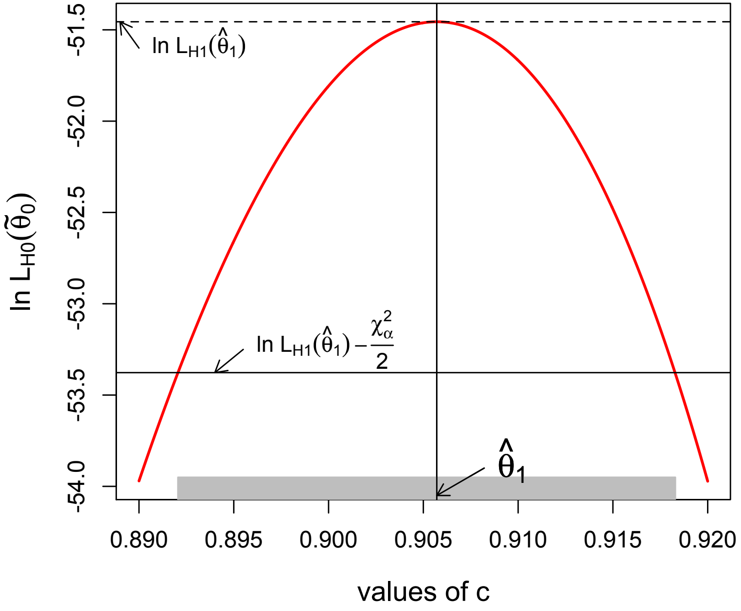
Likelihood: $f(\mathbf{x}; \theta) = f(x_1, x_2, \ldots, x_n; \theta)$ and if X_i discrete

$$= P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n).$$

Now define $\mathcal{I}(\theta) = E_{\mathbf{X}} \left(\left[\frac{\partial}{\partial \theta} \ln f(\mathbf{x}; \theta) \right]^2 \right)$.

Under certain conditions $\mathcal{I}(\theta) = -E_{\mathbf{X}} \left(\left[\frac{\partial^2}{\partial \theta^2} \ln f(\mathbf{x}; \theta) \right]^2 \right)$.

Generalized Likelihood Ratio Tests and Confidence Intervals



Remember that $\tilde{\boldsymbol{\theta}}_{\mathbf{0}} = [c, \ \underbrace{\tilde{\theta}_2, \tilde{\theta}_3, \ldots, \tilde{\theta}_r}_{\text{maximize } r-1 \text{ params.}} \ ]$

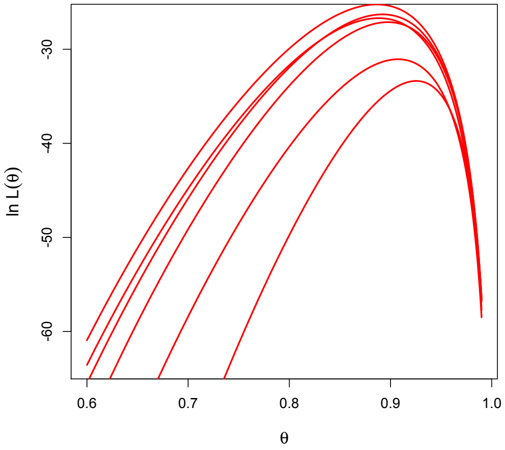
and that $\widehat{\boldsymbol{\theta}} = \underbrace{[\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \ldots, \hat{\theta}_r]}_{\text{maximize } r \text{ params.}}$.

Now, $\frac{\chi^2_{(1)}(\alpha)}{2} = \frac{3.843}{2} = 1.9215$ so reject H_0 if

$$\ln L_{H_1}(\widehat{\boldsymbol{\theta}}) - 1.9215 \geq \ln L_{H_0}(\tilde{\boldsymbol{\theta}}_{\mathbf{0}})$$

Profile likelihood CI: the set of values of c for which we fail to reject H_0!

Fisher's information and asymptotic Wald's C.I.



Let $X_1, X_2, \ldots, X_n$ be a sample of size n, X_i iid or ind.

Likelihood: $f(\mathbf{x};\theta) = f(x_1, x_2, \ldots, x_n; \theta)$ and if X_i discrete

$$= P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n).$$

Now define $\mathcal{I}(\theta) = E_{\mathbf{X}}\left(\left[\frac{\partial}{\partial\theta}\ln f(\mathbf{x};\theta)\right]^2\right)$.

Under certain conditions $\mathcal{I}(\theta) = -E_{\mathbf{X}}\left(\left[\frac{\partial^2}{\partial\theta^2}\ln f(\mathbf{x};\theta)\right]^2\right)$.

Theorem (Abraham Wald): The random variable $\hat{\theta}\xrightarrow{\text{d}}\mathrm{N}\left(\theta, [\mathcal{I}(\theta)]^{-1}\right)$ as $n \rightarrow \infty$. It follows that an approximate $(1-\alpha)100\%$ C.I. for θ is given by $\hat{\theta} \pm z_{\alpha/2}\sqrt{\left[\mathcal{I}(\hat{\theta})\right]^{-1}}$. As sample size grows large, Wald's C.I.'s and the profile likelihood C.I.'s are equivalent. Coverage properties!

Fisher's Information: 2 or more parameters

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_1 \\ \vdots \\ \theta_r \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}.$$

The likelihood is written as the joint pdf of $X_1, \ldots, X_n$ evaluated at the observations $x_1, \ldots, x_n$ and is denoted as $L(\boldsymbol{\theta}) = f(\mathbf{x}; \boldsymbol{\theta})$. The ML estimates $[\hat{\theta}_1, \hat{\theta}_1, \ldots, \hat{\theta}_r]$ are the values of the parameters that **jointly** maximize $L(\boldsymbol{\theta})$, *i.e.* the roots of

$$\begin{cases} \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \theta_1} & = & 0 \\ \\ \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \theta_2} & = & 0 \\ \vdots \\ \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \theta_r} & = & 0. \end{cases}$$

Fisher's Information for 2 or more parameters

In the multivariate case, Fisher's information is written as $\mathcal{I}(\boldsymbol{\theta}) = -E[H(\boldsymbol{\theta})]$, where

$$H(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \theta_1^2} & \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_r} \\ \\ \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_1} & \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \theta_2^2} & \cdots & \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_r} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \theta_r \partial \theta_1} & \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \theta_r \partial \theta_2} & \cdots & \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \theta_r^2} \end{bmatrix}.$$

The Hessian matrix evaluated at the ML estimates and multiplied by -1 is called the "Observed information matrix",

$$J(\hat{\boldsymbol{\theta}}) = \left\{ -\frac{\partial^2 \ln L(\hat{\boldsymbol{\theta}})}{\partial \theta_i \partial \theta_j} \right\} \quad i, j = 1, 2, \ldots, r.$$

Either $\left[\mathcal{I}(\hat{\boldsymbol{\theta}})\right]^{-1}$ or $\left[J(\hat{\boldsymbol{\theta}})\right]^{-1}$ are statistically consistent estimates of the variance of $\hat{\boldsymbol{\theta}}$

Wald's theorem and regularity conditions

Under regularity conditions on $L(\boldsymbol{\theta})$, the random variable $\hat{\boldsymbol{\theta}} \xrightarrow{\text{d}} \mathrm{N}\left(\boldsymbol{\theta}, [\mathcal{I}(\boldsymbol{\theta})]^{-1}\right)$ and an approximate $(1-\alpha)100$% C.I. for θ_i is given by $\hat{\theta}_i \pm z_{\alpha/2}\sqrt{\left\{\left[\mathcal{I}(\hat{\boldsymbol{\theta}})\right]^{-1}\right\}_{i,i}}$.

Regularity conditions roughly say that:

1. $\boldsymbol{\theta}$ cannot be on the boundary of the parameter space.
2. The range of the X_i's cannot depend on θ
3. When multi-modal likelihoods appear, all bets are off!! (And this happens very often.)

Model Selection: Akaike's Information Criterion

Let $f(x)$ and $g(x)$ be two joint pdf's (pmf's) -the likelihood- modeling in two different ways a biological phenomenon. Then the ratio $f(x)/g(x)$ gives us an idea of how much more likely is one model relative to the other one.

Model Selection: Akaike's Information Criterion

Let $f(x)$ and $g(x)$ be two joint pdf's (pmf's) -the likelihood- modeling in two different ways a biological phenomenon. Then the ratio $f(x)/g(x)$ gives us an idea of how much more likely is one model relative to the other one. Now, the Kullback-Leibler Divergence

$$K(f(x), g(x)) = E_X \left[\ln \left(\frac{f(x)}{g(x)} \right) \right]$$

tells us, on average, how much more likely is $g(x)$ relative to $f(x)$. Suppose $f(x)$ is an (unknown) stochastic mechanism that generates the data, the truth.

Model Selection: Akaike's Information Criterion

Let $f(x)$ and $g(x)$ be two joint pdf's (pmf's) -the likelihood- modeling in two different ways a biological phenomenon. Then the ratio $f(x)/g(x)$ gives us an idea of how much more likely is one model relative to the other one. Now, the Kullback-Leibler Divergence

$$K(f(x), g(x)) = E_X \left[\ln \left(\frac{f(x)}{g(x)} \right) \right]$$

tells us, on average, how much more likely is $g(x)$ relative to $f(x)$. Suppose $f(x)$ is an (unknown) stochastic mechanism that generates the data, the truth. Now suppose $g(x)$ is the model that we are trying to use to describe the data. Then, the above expectation expresses how far away from the truth is the model.

Model Selection: Akaike's Information Criterion

Let $f(x)$ and $g(x)$ be two joint pdf's (pmf's) -the likelihood- modeling in two different ways a biological phenomenon. Then the ratio $f(x)/g(x)$ gives us an idea of how much more likely is one model relative to the other one. Now, the Kullback-Leibler Divergence

$$K(f(x), g(x)) = E_X\left[\ln\left(\frac{f(x)}{g(x)}\right)\right]$$

tells us, on average, how much more likely is $g(x)$ relative to $f(x)$. Suppose $f(x)$ is an (unknown) stochastic mechanism that generates the data, the truth. Now suppose $g(x)$ is the model that we are trying to use to describe the data. Then, the above expectation expresses how far away from the truth is the model. Some properties of the K-L distance are

1. $K(f(x), g(x)) = 0 \Leftrightarrow f(x) = g(x)$
2. $K(f(x), g(x)) \geq 0$,
3. Value of θ that minimizes $K(f(x), g(x, \theta))$ is the MLE of θ, $\hat{\theta}$.

K-L divergence as a model comparison tool

The difference in K-L divergence between each of two different models and the truth,*i.e.*,

$$K(f(x), g_1(x; \theta_1)) - K(f(x), g_2(x; \theta_2)).$$

can be used to compare one model against the other, but don't know $f(x)$!

K-L divergence as a model comparison tool

The difference in K-L divergence between each of two different models and the truth,*i.e.*,

$$K(f(x), g_1(x;\theta_1)) - K(f(x), g_2(x;\theta_2)).$$

can be used to compare one model against the other, but don't know $f(x)$! However, Akaike showed that a statistically consistent estimate of

$$K(f(x), g_i(x;\theta_i)) \text{ is given by } AIC_i = -2\ln L(\hat{\theta}_i) + 2 \times p_i,$$

where $p_i = \#$ of model parameters estimated with the data.

K-L divergence as a model comparison tool

The difference in K-L divergence between each of two different models and the truth,*i.e.*,

$$K(f(x), g_1(x;\theta_1)) - K(f(x), g_2(x;\theta_2)).$$

can be used to compare one model against the other, but don't know $f(x)$! However, Akaike showed that a statistically consistent estimate of

$$K(f(x), g_i(x;\theta_i)) \text{ is given by } AIC_i = -2\ln L(\hat{\theta}_i) + 2 \times p_i,$$

where $p_i = \#$ of model parameters estimated with the data.

- If you have a series of models, the decision rule is to pick the model for which the AIC is the smallest.

K-L divergence as a model comparison tool

The difference in K-L divergence between each of two different models and the truth,*i.e.*,

$$K(f(x), g_1(x;\theta_1)) - K(f(x), g_2(x;\theta_2)).$$

can be used to compare one model against the other, but don't know $f(x)$! However, Akaike showed that a statistically consistent estimate of

$$K(f(x), g_i(x;\theta_i)) \text{ is given by } AIC_i = -2\ln L(\hat{\theta}_i) + 2 \times p_i,$$

where $p_i = \#$ of model parameters estimated with the data.

- If you have a series of models, the decision rule is to pick the model for which the AIC is the smallest.
- AIC is a frequentist concept: over hypothetical repeated sampling, it is a consistent estimate of the expected, relative K-L distance between the generating model and the proposed model.

K-L divergence as a model comparison tool

The difference in K-L divergence between each of two different models and the truth,*i.e.*,

$$K(f(x), g_1(x;\theta_1)) - K(f(x), g_2(x;\theta_2)).$$

can be used to compare one model against the other, but don't know $f(x)$! However, Akaike showed that a statistically consistent estimate of

$$K(f(x), g_i(x;\theta_i)) \text{ is given by } AIC_i = -2\ln L(\hat{\theta}_i) + 2 \times p_i,$$

where $p_i = \#$ of model parameters estimated with the data.

- If you have a series of models, the decision rule is to pick the model for which the AIC is the smallest.
- AIC is a frequentist concept: over hypothetical repeated sampling, it is a consistent estimate of the expected, relative K-L distance between the generating model and the proposed model.
- Other information criteria and future research questions with this topic will be covered in next talk.

Observation Error (Real life happens. . .)

General model accounting for sampling error: State-space models

- Let X_t be a d.t. Markov process. Let the conditional density function of X_t given $X_{t-1} = x_{t-1}$ be $g(x_t|x_{t-1}, \theta)$.

General model accounting for sampling error: State-space models

- Let X_t be a d.t. Markov process. Let the conditional density function of X_t given $X_{t-1} = x_{t-1}$ be $g(x_t|x_{t-1}, \theta)$.
- Conditional on X_t, the observations process Y_t is another random variable with pdf given by $f(y_t|x_t, \phi)$:

$$\begin{aligned} \text{(state equation):} \qquad & X_t|X_{t-1} \sim g(x_t|x_{t-1}, \theta), \\ \text{(observation equation):} \qquad & Y_t|X_t \sim f(y_t|x_t, \phi). \end{aligned}$$

General model accounting for sampling error: State-space models

- Let X_t be a d.t. Markov process. Let the conditional density function of X_t given $X_{t-1} = x_{t-1}$ be $g(x_t|x_{t-1}, \theta)$.
- Conditional on X_t, the observations process Y_t is another random variable with pdf given by $f(y_t|x_t, \phi)$:

$$\begin{aligned} \text{(state equation):} \qquad & X_t|X_{t-1} \sim g(x_t|x_{t-1}, \theta), \\ \text{(observation equation):} \qquad & Y_t|X_t \sim f(y_t|x_t, \phi). \end{aligned}$$

- If both g and f are linear Gaussian conditional distributions then the resulting model is called a *linear state-space model (LSSM)*, or *dynamic linear model (DLM)*.

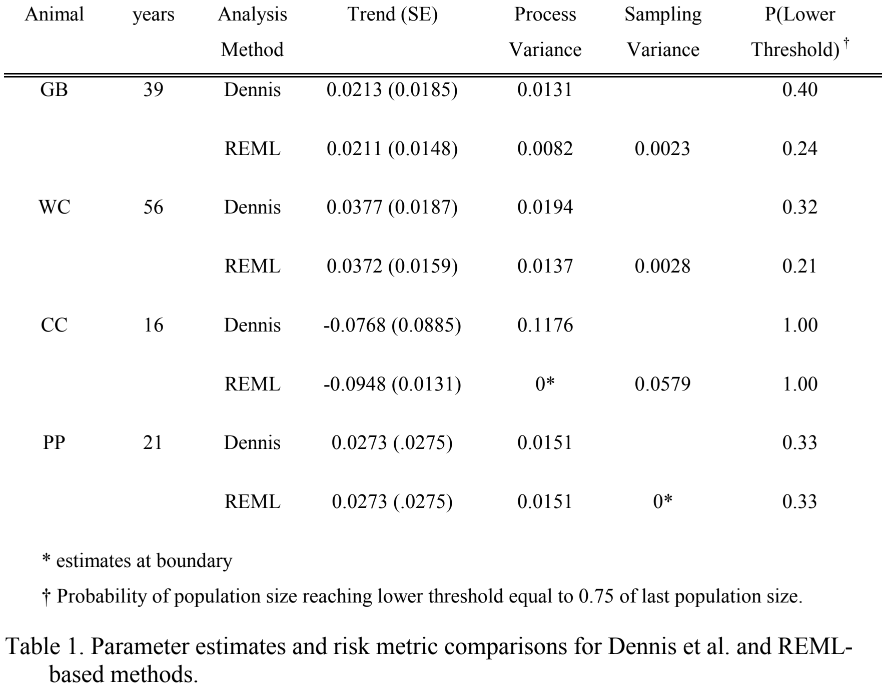
General model accounting for sampling error: State-space models

- Let X_t be a d.t. Markov process. Let the conditional density function of X_t given $X_{t-1} = x_{t-1}$ be $g(x_t|x_{t-1}, \theta)$.
- Conditional on X_t, the observations process Y_t is another random variable with pdf given by $f(y_t|x_t, \phi)$:

$$\begin{array}{rc} \text{(state equation):} & X_t|X_{t-1} \sim g(x_t|x_{t-1}, \theta), \\ \text{(observation equation):} & Y_t|X_t \sim f(.|x_t, \phi). \end{array}$$

- If both g and f are linear Gaussian conditional distributions then the resulting model is called a *linear state-space model (LSSM)*, or *dynamic linear model (DLM)*.
- In general $L(\theta, \phi) = \int f(\mathbf{y}|\mathbf{X}, \phi)g(\mathbf{x}; \theta)d\mathbf{X}$.
- Need computer intensive methods to calc. the likelihood for non-linear, non-gaussian models.

Observation error & density-independence



Animal	years	Analysis Method	Trend (SE)	Process Variance	Sampling Variance	P(Lower Threshold) †
GB	39	Dennis	0.0213 (0.0185)	0.0131		0.40
		REML	0.0211 (0.0148)	0.0082	0.0023	0.24
WC	56	Dennis	0.0377 (0.0187)	0.0194		0.32
		REML	0.0372 (0.0159)	0.0137	0.0028	0.21
CC	16	Dennis	-0.0768 (0.0885)	0.1176		1.00
		REML	-0.0948 (0.0131)	0*	0.0579	1.00
PP	21	Dennis	0.0273 (.0275)	0.0151		0.33
		REML	0.0273 (.0275)	0.0151	0*	0.33

* estimates at boundary

† Probability of population size reaching lower threshold equal to 0.75 of last population size.

Table 1. Parameter estimates and risk metric comparisons for Dennis et al. and REML-based methods.

Staples, Taper and Dennis, 2004. Estimating population trend and process variation for PVA in the presence of sampling error. Ecology 85:923-929.

Observation error and density-dependence

The stochastic Gompertz model

$$N_t = N_{t-1}e^{[(a+b\ln(N_{t-1})+\sigma E_t]}$$

Let $x_t = \ln(n_t)$ and take $c = b + 1$, then we have a first-order autoregressive process (Reddingius, 1971, Dennis and Taper 1994):

$$\begin{aligned} X_t &= X_{t-1} + a + bX_{t-1} + E_t \\ &= a + cX_{t-1} + E_t \end{aligned}$$

Density independence is expressed through $b = 0$ or $c = 1$. For $|c| < 1$ the stationary distribution exists and:

$$\begin{aligned} E[X_\infty] &= \lim_{t\to\infty} E[X_t] = \frac{a}{1-c} \\ Var[X_\infty] &= \lim_{t\to\infty} Var[X_t] = \frac{\sigma^2}{1-c^2} \end{aligned}$$

Stochastic Gompertz with observation error:

- Let Y_t be the estimated logarithmic population abundance, such that:

$$\begin{aligned} Y_t &= X_t + F_t \\ &= a + cX_{t-1} + E_t + F_t \\ &= a + c(Y_{t-1} - F_{t-1}) + E_t + F_t, \end{aligned}$$

  where $F_t \sim \mathrm{N}(0, \tau^2)$.

- The Markov property is lost: it is an ARMA model (Autorregresive Moving Average process).
- There is extra info. in the autocorrelation structure about σ^2 and τ^2.
- The ML parameter estimates are obtained via the Kalman filter (lots of conditioning) or using MVN:

The Multivariate Normal model:

No observation error: we have a series of recorded observations

$$x_0, x_1, \ldots x_q.$$

Assuming X_0 arises from the stationary distribution, the joint pdf of $X_0, X_1, \ldots X_q = \mathbf{X}$ has the following distribution:

$$\mathbf{X} \sim \mathbf{MVN}(\mu, \mathbf{\Sigma})$$

where

$$\Sigma = \frac{\sigma^2}{1-c^2} \begin{pmatrix} 1 & c & c^2 & \ldots & c^q \\ c & 1 & c & \ldots & c^{q-1} \\ c^2 & c & 1 & \ldots & c^{q-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c^q & c^{q-1} & c^{q-2} & \ldots & c \end{pmatrix}$$

and

$$\mu = \frac{a}{1-c}\mathbf{j},$$

$\mathbf{j}$ being a $(q+1) \times 1$ vector of ones.

The Multivariate Normal model:

With observation error: given the observations,$y_0, y_1, \ldots y_q$, the joint pdf of $Y_0, Y_1, \ldots Y_q$ is multivariate normal: writing $\mathbf{Y} = \mathbf{X} + \mathbf{F}$, we get

$$\mathbf{Y} \sim \mathbf{MVN}(\mu, \mathbf{V})$$

where $\mu = \frac{a}{1-c}\mathbf{j}$, $\mathbf{j}$ being a $(q+1) \times 1$ vector of ones, and $\mathbf{V} = \mathbf{\Sigma} + \tau^2\mathbf{I}$. The variance covariance matrix of the process is:

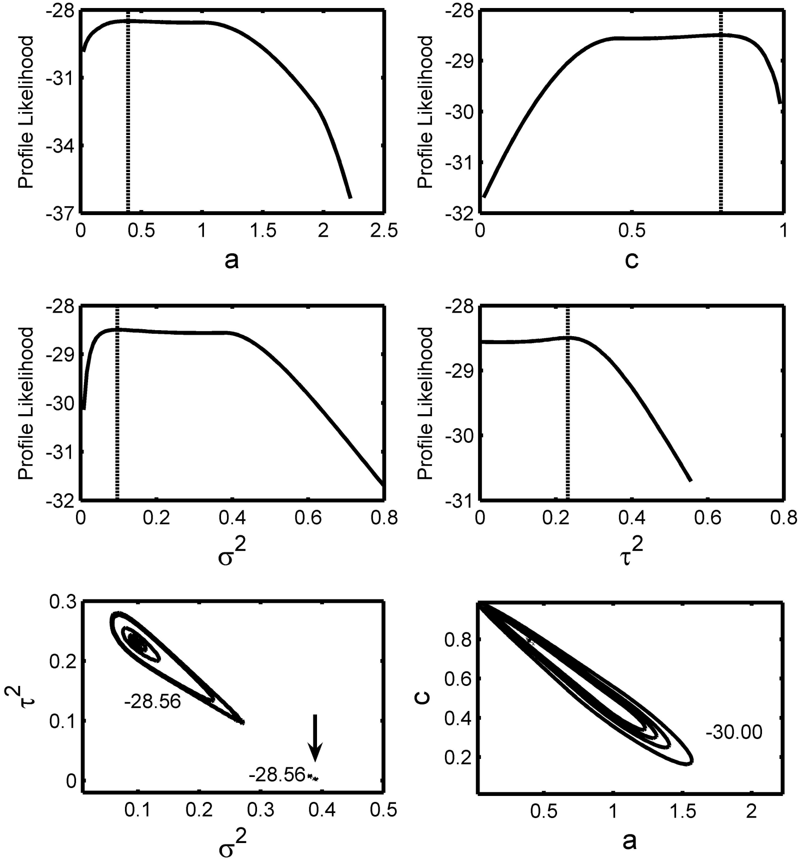
$$\mathbf{V} = \begin{bmatrix} \frac{\sigma^2}{1-c^2} + \tau^2 & \frac{c\sigma^2}{1-c^2} & \frac{c^2\sigma^2}{1-c^2} & \cdots & \frac{c^q\sigma^2}{1-c^2} \\ \frac{c\sigma^2}{1-c^2} & \frac{\sigma^2}{1-c^2} + \tau^2 & \frac{c\sigma^2}{1-c^2} & \cdots & \frac{c^{q-1}\sigma^2}{1-c^2} \\ \frac{c^2\sigma^2}{1-c^2} & \frac{c\sigma^2}{1-c^2} & \frac{\sigma^2}{1-c^2} + \tau^2 & \cdots & \frac{c^{q-2}\sigma^2}{1-c^2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{c^q\sigma^2}{1-c^2} & \frac{c^{q-1}\sigma^2}{1-c^2} & \frac{c^{q-2}\sigma^2}{1-c^2} & \cdots & \frac{\sigma^2}{1-c^2} + \tau^2 \end{bmatrix}.$$

Therefore, the log-likelihood needed for parameter estimation is:

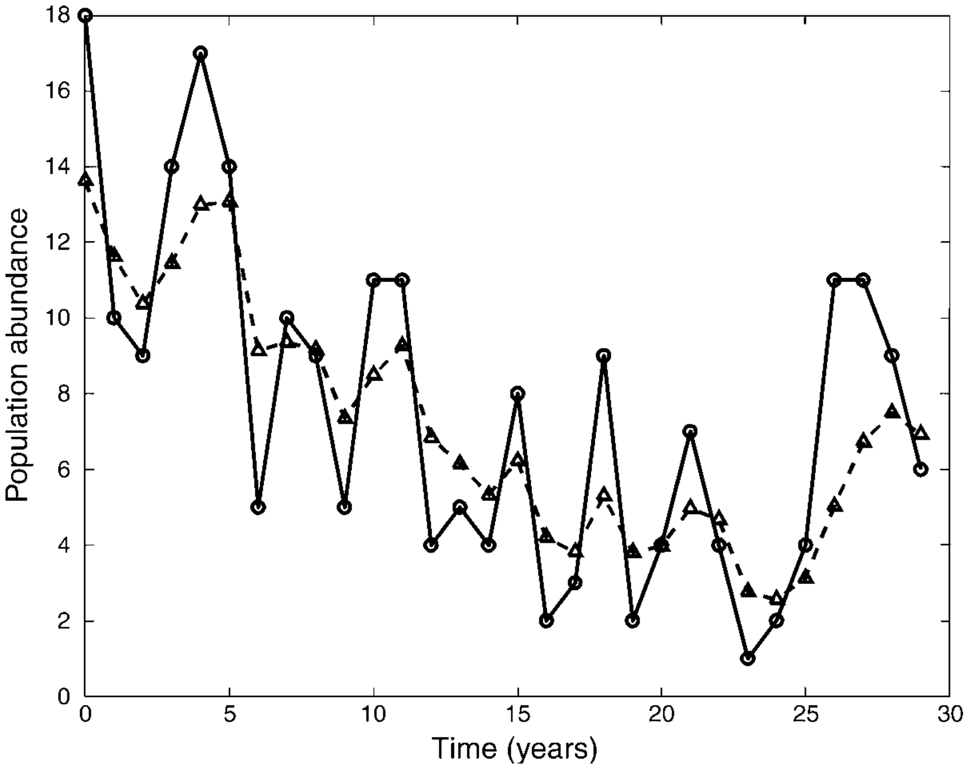
$$\ln L(a, c, \sigma^2, \tau^2) = -\frac{q+1}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mu)'\mathbf{V}^{-1}(\mathbf{y} - \mu)$$

(First differences log-likelihood -REML- can also be obtained and behave nicely)

Log-profile likelihoods

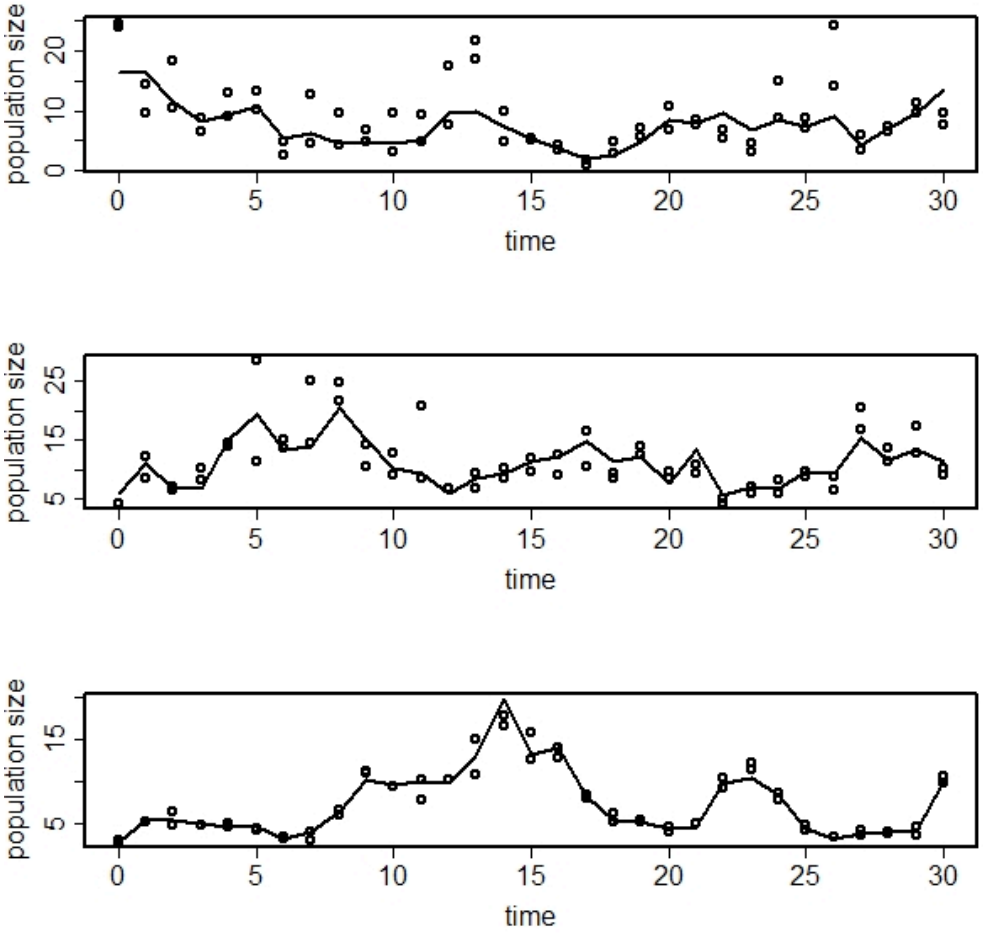


Estimated proportion of observation error: ≈ 70 %



Dennis, B., Ponciano, J.M., Lele, S., Taper, M.L., Staples, D.F. 2006. Estimating density-dependence, process noise and observation error. Ecol. Monogr. 76: 323-341

Replicated Sampling



Replicated Sampling for the GSS model

- Let Y_t be the estimated logarithmic population abundance, such that:

$$\begin{aligned} Y_t &= X_t + F_t \\ &= a + cX_{t-1} + E_t + F_t \\ &= a + c(Y_{t-1} - F_{t-1}) + E_t + F_t, \end{aligned}$$

- If at time step t, p_t replicates are taken yielding observations $\mathbf{Y}_t = [Y_{1t}, Y_{2t}, Y_{3t}, \ldots Y_{pt}]'$, then we write:

$$\mathbf{Y}_t = \mathbf{j}_t X_t + \mathbf{F}_t,$$

where $\mathbf{j}_t$ is a $p_t \times 1$ vector of ones, $\mathbf{F}_t \sim MVN(\mathbf{0}, \tau^2 \mathbf{I}_t)$ and $\mathbf{I}_t$ is a $p_t \times p_t$ identity matrix.

- The likelihood of the observations from $t = 0$ to $t = q$ is the joint pdf of $\mathbf{Y_t}$ given $\mathbf{Y}_{t-1} = \mathbf{y}_{t-1}, \mathbf{Y}_{t-2} = \mathbf{y}_{t-2}, \ldots, \mathbf{Y}_0 = \mathbf{y}_0$.

Replicated Sampling for the GSS model

- The likelihood is multivariate normal and its mean and variance changes with time. The Kalman recursions can also be used here.
- Let $\mathbf{J}_t$ be a $p_t \times p_t$ matrix of ones and, let $\mathbf{j}_t$ be a $p_t \times 1$ vector of ones and $\mathbf{I}_t$ be the $p_t \times p_t$ identity matrix.
- Using the stationary distribution for $X_0 \sim N(\mu_0, \psi^2)$, it is found that $E[\mathbf{Y}_0] = \mathbf{j}_0\mu_0 = \mathbf{m}_0$ and that $Var[\mathbf{Y}_0] = \psi^2\mathbf{J}_0 + \tau^2\mathbf{I}_0$

Replicated Sampling for the GSS model

- The Kalman recursions are:

$$\begin{aligned}
\mu_t &= a + c\left[\mu_{t-1} + \mathbf{j}'_{t-1}\psi^2_{t-1}\mathbf{V}^{-1}_{t-1}(\mathbf{y}_{t-1} - \mathbf{m}_{t-1})\right], \\
\psi^2_t &= c^2\psi^2_{t-1}\left[1 - \psi^2_{t-1}\mathbf{j}'_{t-1}\mathbf{V}^{-1}_{t-1}\mathbf{j}_{t-1}\right] + \sigma^2, \\
\mathbf{m}_t &= \mathbf{j}_t\mu_t, \\
\mathbf{V}_t &= \mathbf{J}_t\psi^2_{t-1} + \tau^2\mathbf{I}_t.
\end{aligned}$$

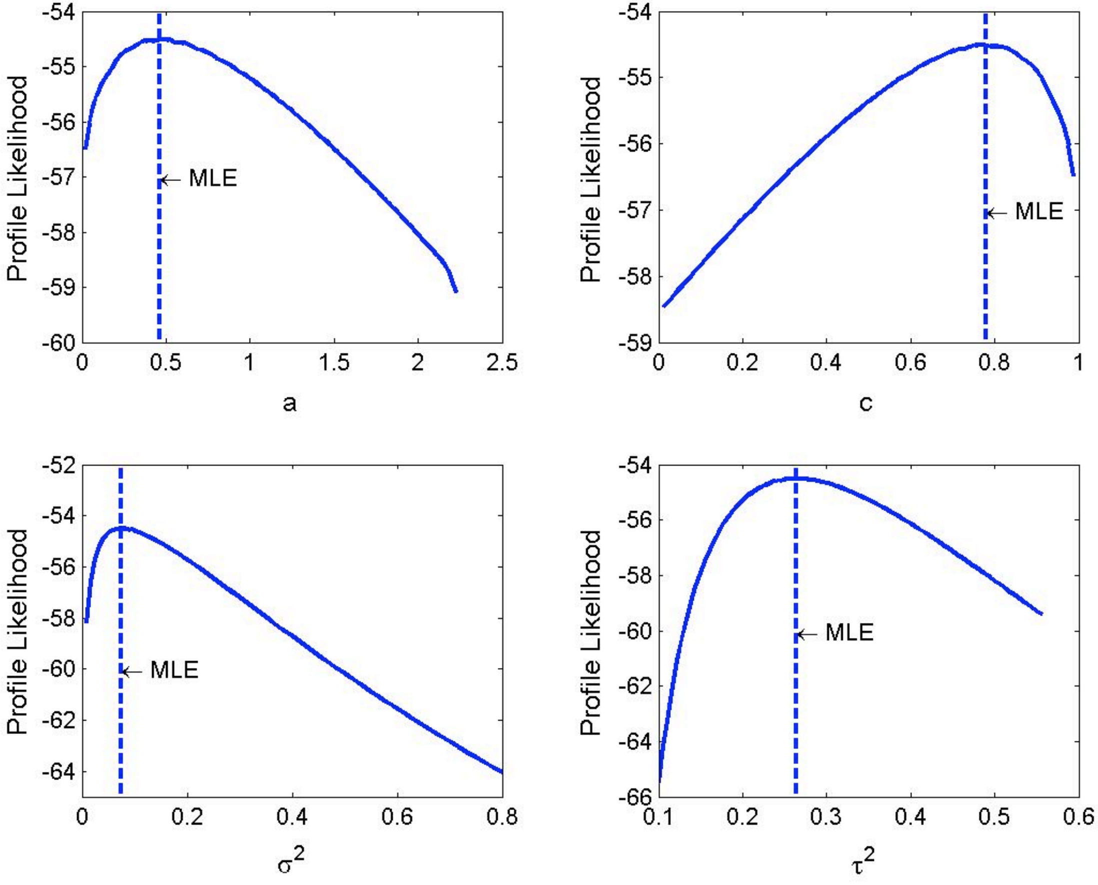
- And the full likelihood function (assuming we start at the stationary distribution) is:

$$L(a, c, \sigma^2, \tau^2) = L(\mathbf{y}_0)L(\mathbf{y}_1|\mathbf{y}_0)L(\mathbf{y}_2|\mathbf{y}_1, \mathbf{y}_0)\ldots(\mathbf{y}_q|, \mathbf{y}_{q-1}, ...\mathbf{y}_0)$$

$$= (2\pi)^{-p/2}(|\mathbf{V}_0||\mathbf{V}_1|\ldots|\mathbf{V}_q|)^{-1/2}\exp\left[-\frac{1}{2}\sum_{t=0}^{q}(\mathbf{y}_t - \mathbf{m}_t)'\mathbf{V}^{-1}_t(\mathbf{y}_t - \mathbf{m}_t)\right],$$

where $p = p_0 + p_1 + \ldots + p_q$.

Log-profile likelihoods



MCMC and computer intensive methods

Next time!