

# **Model Comparison**

# **A General Problem**

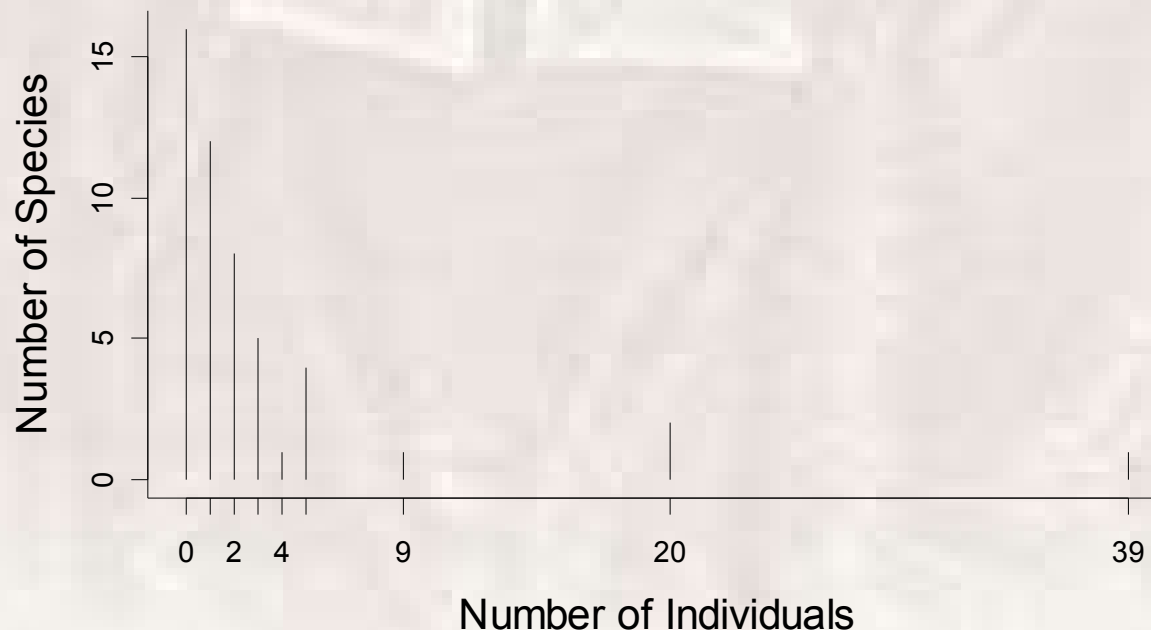
- We can fit several models to data
- But how do we know which is the best model?

# An Ecological Example

- What is the distribution of abundance?
  - A long-standing problem in ecology
- Two (of many) suggestions:
  - gamma
  - lognormal
- The lognormal has some theoretical support, the gamma is more flexible

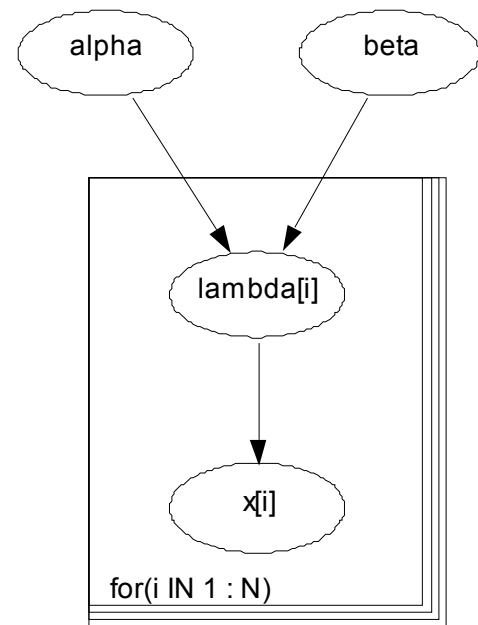
# Some Data

- Mexican Butterflies
  - 0,5,2,0,3,0,0,5,1,4,9,0,5,1,1,2,5,1,1,0,0,2,2,2,2,1,1,2,2  
0,1,0,0,39,0,0,2,3,20,3,1,0,1,1,1,3,0,0,0,0,3
- Which distribution fits best?



# The Models

- For both models, we assume the number of individuals follows a Poisson distribution
  - $x[i] \sim \text{dpois}(\text{lambda}[i])$
- 2 models for the mean
  - Gamma
    - $\text{lambda}[i] \sim \text{dgamma}(\text{alpha}, \text{beta})$
  - Log-normal
    - $\log(\text{lambda}[i]) \sim \text{dnorm}(\mu, \tau)$
- Then give priors for alpha, beta or mu, tau



# Comparing Models

- Several approaches in the Bayesian literature
  - Bayes factors
  - rjMCMC
  - DIC
- In reality there is not true model
  - “All models are wrong, but some are useful”
  - calculating the probability that a model is correct is over-kill
  - instead, take a less formal approach

# DIC

- Classical statistics has AIC
  - Akaike's Information Criterion
  - Deviance + 2 × Number of Parameters
- Based on a similar argument, Bayesians have DIC
  - Deviance Information Criterion
  - $\bar{D} + 2 \times pD$
- Model with the smallest DIC has the highest chance of predicting a replicate data set

# What is DIC?

- We calculate the deviance,  $D$ 
  - $D = -2 \times \log(P(y | \theta)) = -2 \times \text{likelihood}$
- Take the average over the posterior

$$\overline{D} = -\int 2 \log(P(y | \theta)) d\theta$$

- Or, the average deviance from the MCMC
- Measure of goodness of fit

# pD

- AIC penalises the deviance with a term involving the number of parameters
  - more complex models get a higher penalty
- DIC penalises with pD
  - “effective number of parameters”
- Calculated as  $\bar{D} - \hat{D}$ 
  - $\hat{D}$  is the deviance at the posterior mean of the parameters

$$\hat{D} = -2 \log(P(y | \bar{\theta}))$$

# Using DIC

- If we have several models, we can try and find the one with the lowest DIC
  - measured on the same data!
- If values are close (say, within 5), then the different models are similar
  - makes little difference which one is used
- $pD$  can be interpreted as an estimate of the number of parameters
  - almost always positive

# Diversity and DIC

- For the butterflies, we get this:
  - Gamma: DIC = 74.16, pD = 28.91
  - log-Normal: DIC = 77.05, pD = 27.36
- The Gamma fits better, but not greatly
- The complexities (pD) are similar
  - about the number of non-zero observations
- Further investigation revealed that neither model is able to fit well
  - most common species too common