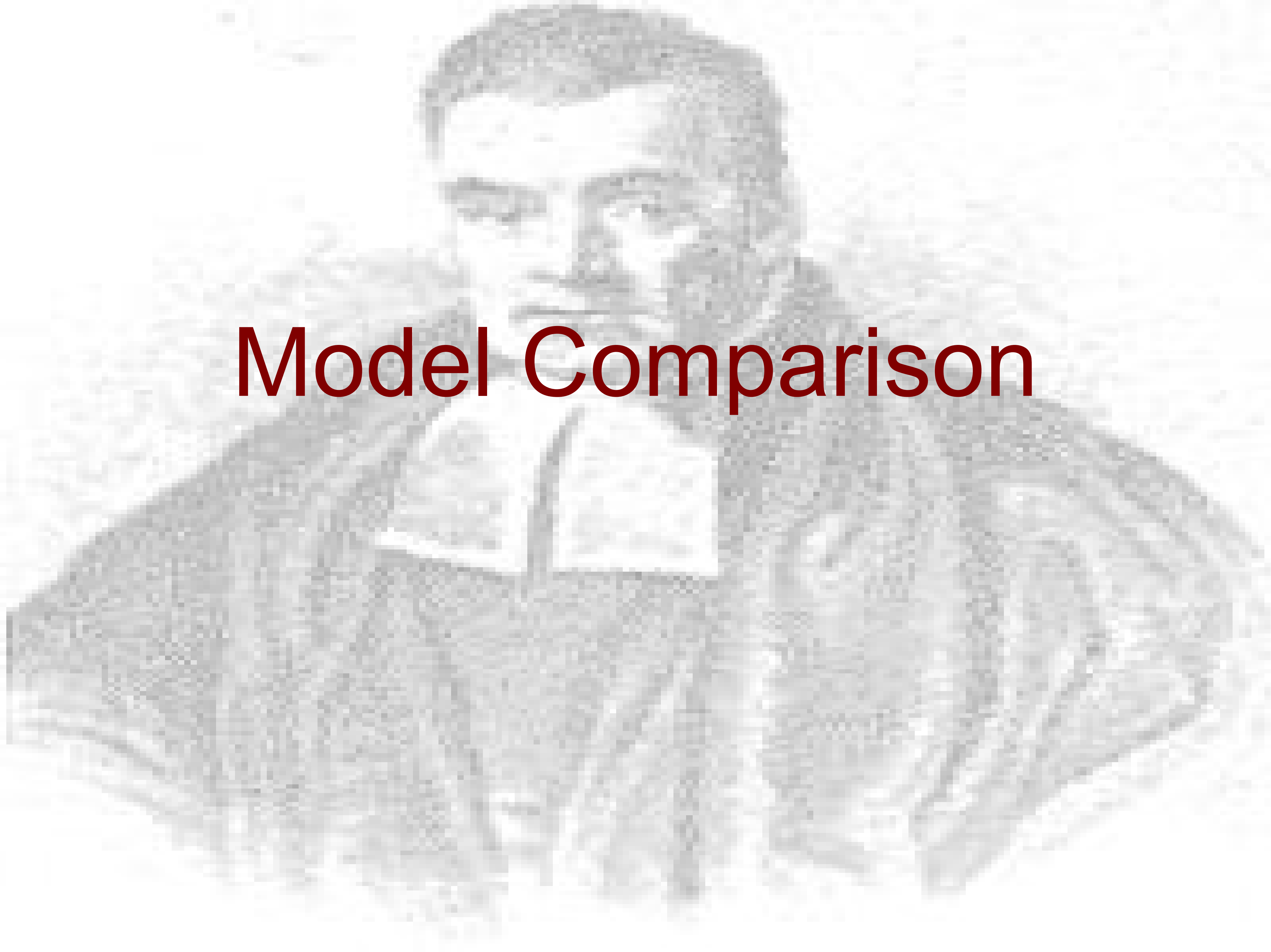
Model Comparison



A General Problem

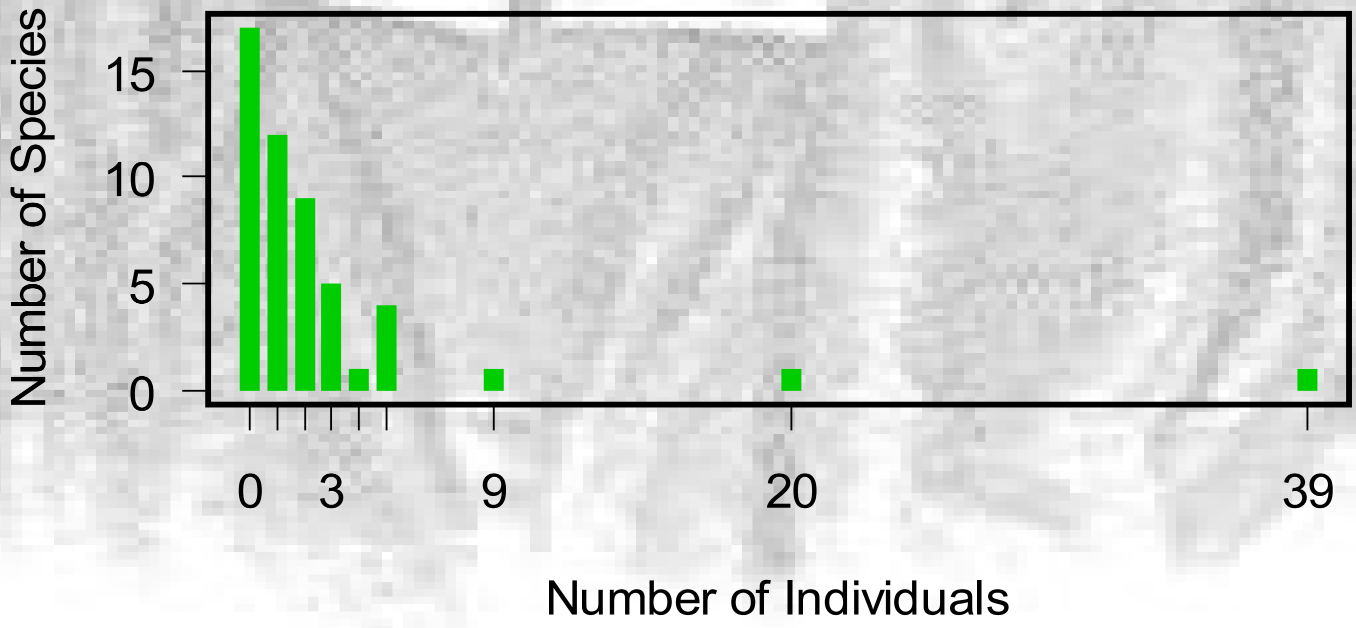
- We can fit several models to data
- But how do we know which is the best model?
- There are several solutions, here is one which is of some use
  - not always the best!

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

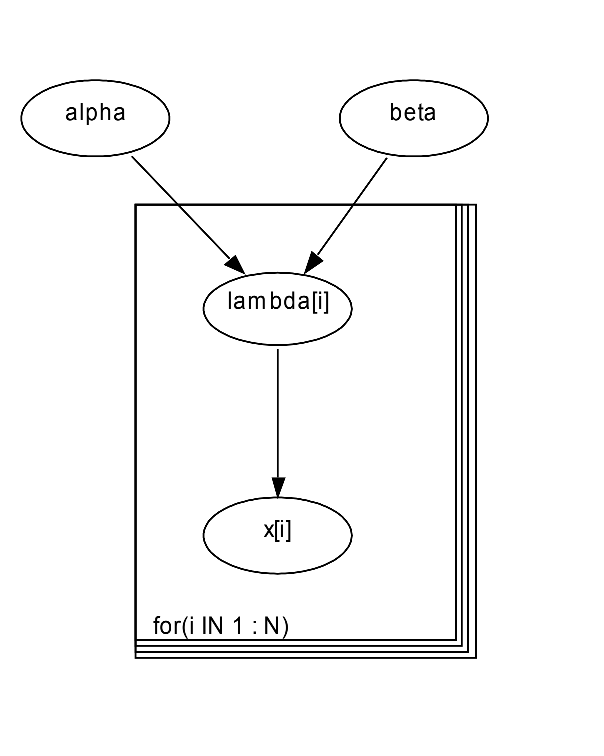
Some Data

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



The Models

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



Comparing Models

- Several approaches in the Bayesian literature
  - Bayes factors
  - rjMCMC
  - DIC
  - SSVS, GVS, CSM, ...
- 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
- From a similar argument, Bayesians have DIC
  - Deviance Information Criterion
  - Dbar + 2 × 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 × log(P(X | θ)) = -2 × likelihood
- Take the average over the posterior

$$\bar{D} = -\int 2\log\left(P\left(X|\theta\right)\right) d\theta$$

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

p_D

- AIC penalises the deviance with a term involving the number of parameters
  - more complex models get a higher penalty
- DIC penalises with p_D
  - "effective number of parameters"
- Calculated as Dbar – Dhat
  - Dhat is the deviance at the posterior mean of the parameters

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

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