Prediction and Missing Data

Summarising Distributions

- Models are often large and complex
- Often only interested in some parameters
  - e.g. not so interested in the intercept
- Need to deal with the other parameters

Marginal Distributions

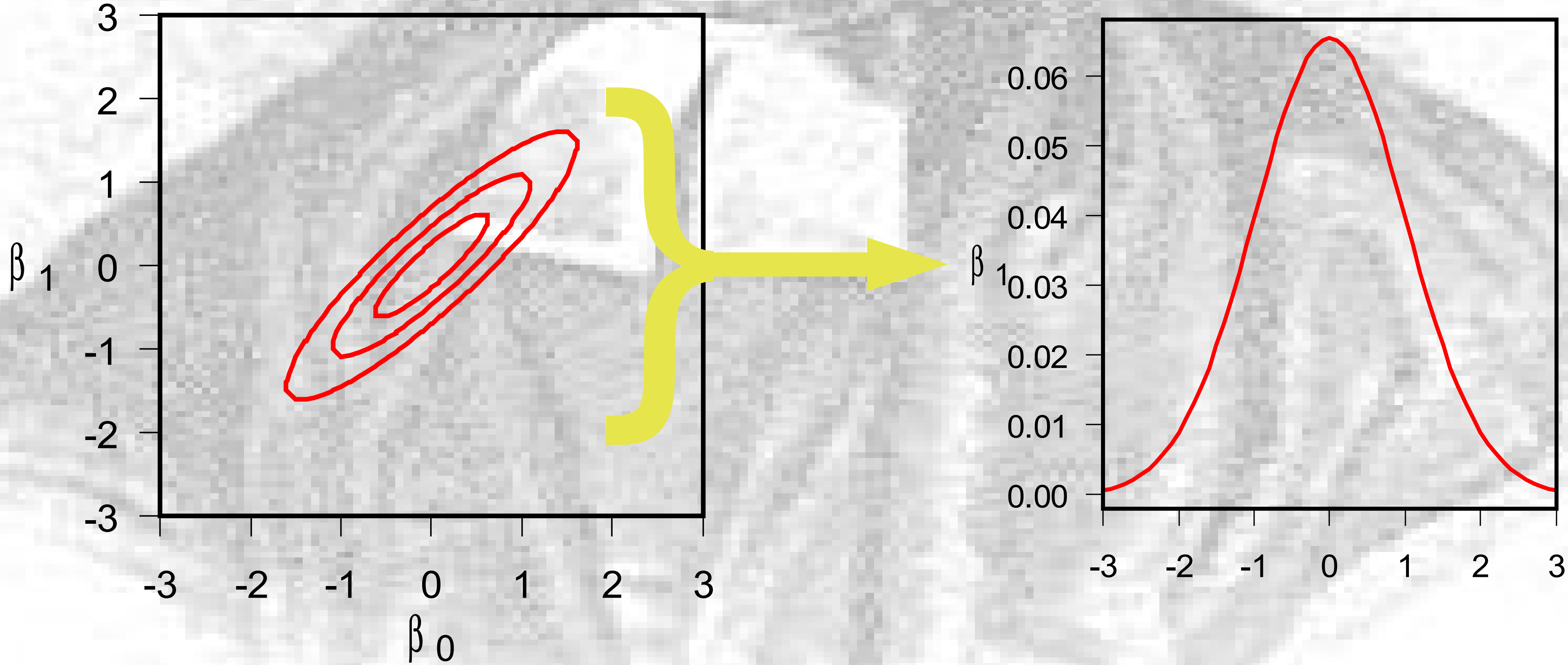
- A model with two parameters, θ_1 and θ_2
- Suppose we are only interested in θ_1
- Calculate the marginal distribution:

$$P(\theta_1|X) = \int P(\theta_1, \theta_2|X) P(\theta_2|X) d\theta_2$$

- Weighted sum of the joint distribution
- In practice, use MCMC, just look at the output for θ_1 and ignore θ_2

Example: Regression

- For each value of β_0, take the distribution of β_1 and add them together



Marginal Distributions

- Marginal distributions include the uncertainty in the parameters they are marginalised over
- The Bayesian approach: take the uncertainty into account
  - don't know parameter values, only estimate them
- We condition on what we know: the data
  - this is measured
  - the parameters are not

Prediction

- We sometimes want to predict new data
  - e.g. population viability analysis
- We have to use estimated parameter values
  - but these are not certain
- Want to make predictions which include the uncertainty in the parameters
  - natural in the Bayesian approach

Prediction: The Maths

- If we have data, X, and parameters, θ, then the posterior is $P(\theta \mid X)$
- We want to predict some new data, X_{new}
- We have a model to do this: $P(X|\theta)$
  - the likelihood
- If we want to make predictions, we should make predictions based on what we know
  - i.e. the data
  - so, we want $P(X_{new}|X)$

How Do We Get P(X_{new}|X)?

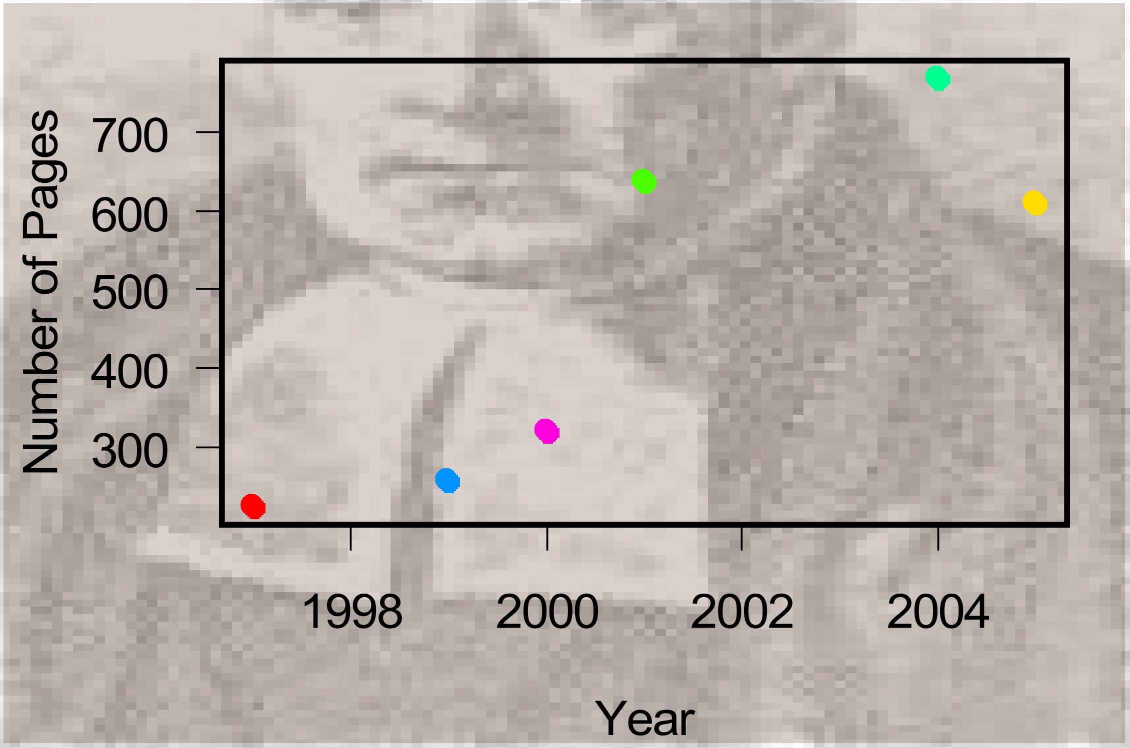
- For each value of θ make a prediction of X_{new}, using P(X|θ)
- We can then take a weighted sum so that the values of θ that are more likely contribute more to the prediction
  - i.e. take P(X_{new}|θ)×P(θ | X) and add them up
- Mathematical statement of this:

$$P(X_{new}|X) = \int P(X_{new}|\theta)P(\theta|X)\,d\theta$$

In Practice: MCMC

- If we do MCMC, then we draw a lot of values from the posterior
  - Each value is equally likely
  - the parameter regions with higher densities are represented by more values
- So, each prediction based on a value is equally likely
- Therefore, we can take each value from the posterior, and simulate the new data using the likelihood

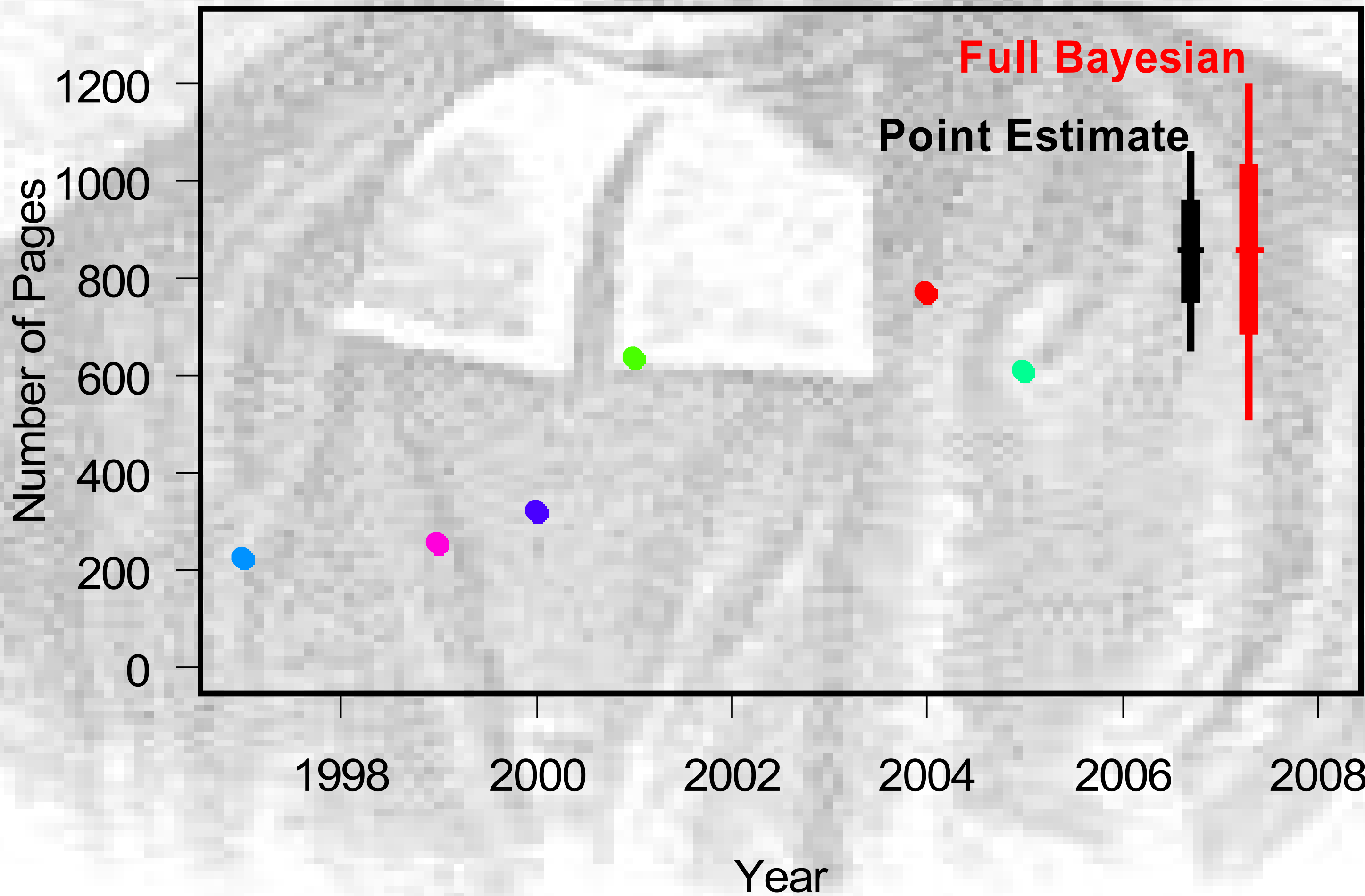
How long with the next Harry Potter book be?



- Model: $P_i \sim N(\alpha + \beta Y_i, \sigma^2)$
- Predict for $Y_7 = 2007$

Posterior Predictive Distribution

- Plot of the Posterior Predictive Distribution



Missing Data

- In many real problems we do not observe all of the data
- It may be unobservable
  - e.g. patients who survive beyond the duration of a medical trial
- The observation may have been lost
  - e.g. someone stood on your tray of samples
- How should we deal with this?

Missing Data Example

- Suppose we do not know when the fourth Harry Potter book was published
  - missing covariate
- We could drop the data point, but that loses information
  - even worse when there are many covariates, with some data missing in all
- We do know the range when it could have been published
  - between books 3 and 5!

Estimation of Missing Data

- The data does tell us something about the missing covariates:
  - the observed covariates
  - the data (through the relationship with the covariates)
- Therefore we can learn about the missing covariates from the observed data
- We can easily formalise this

Missing Data

- If some covariates are not observed, they can be estimated
- Treat them as extra parameters

Inference

- Making the missing data extra parameters means we can get a joint posterior
  - $P(\sigma^2, \alpha, \beta, Y_i)$,
    - *i* indexes missing data
- But we are not interested in Y_i
- Being Bayesians, we can integrate it out:

$$P(\sigma^2, \alpha, \beta | P) = \int P(\sigma^2, \alpha, \beta, Y_i | P) P(Y_i | P) \, dY_i$$

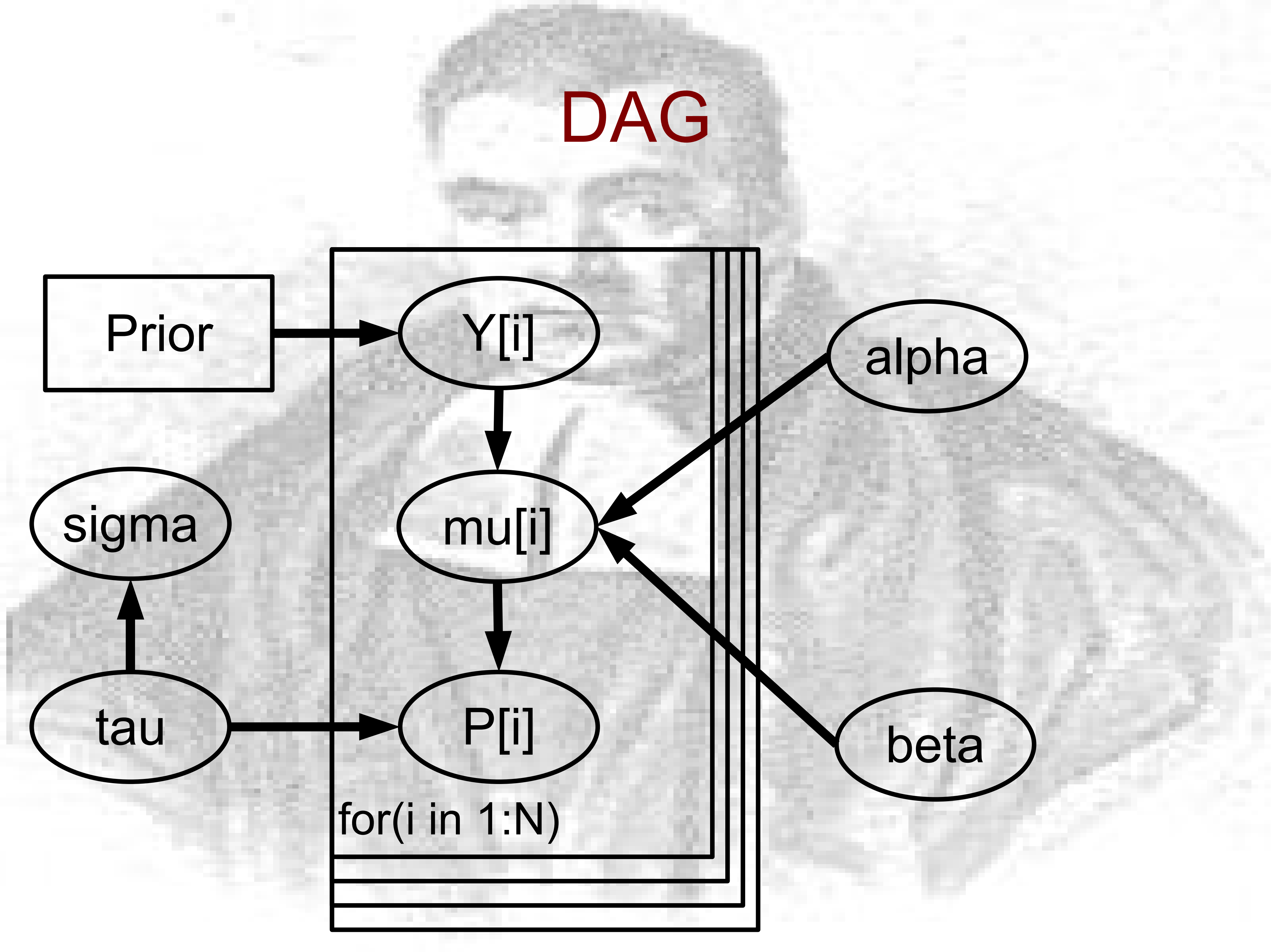
- With MCMC: just drop the Y_i estimates

Multiple Imputation

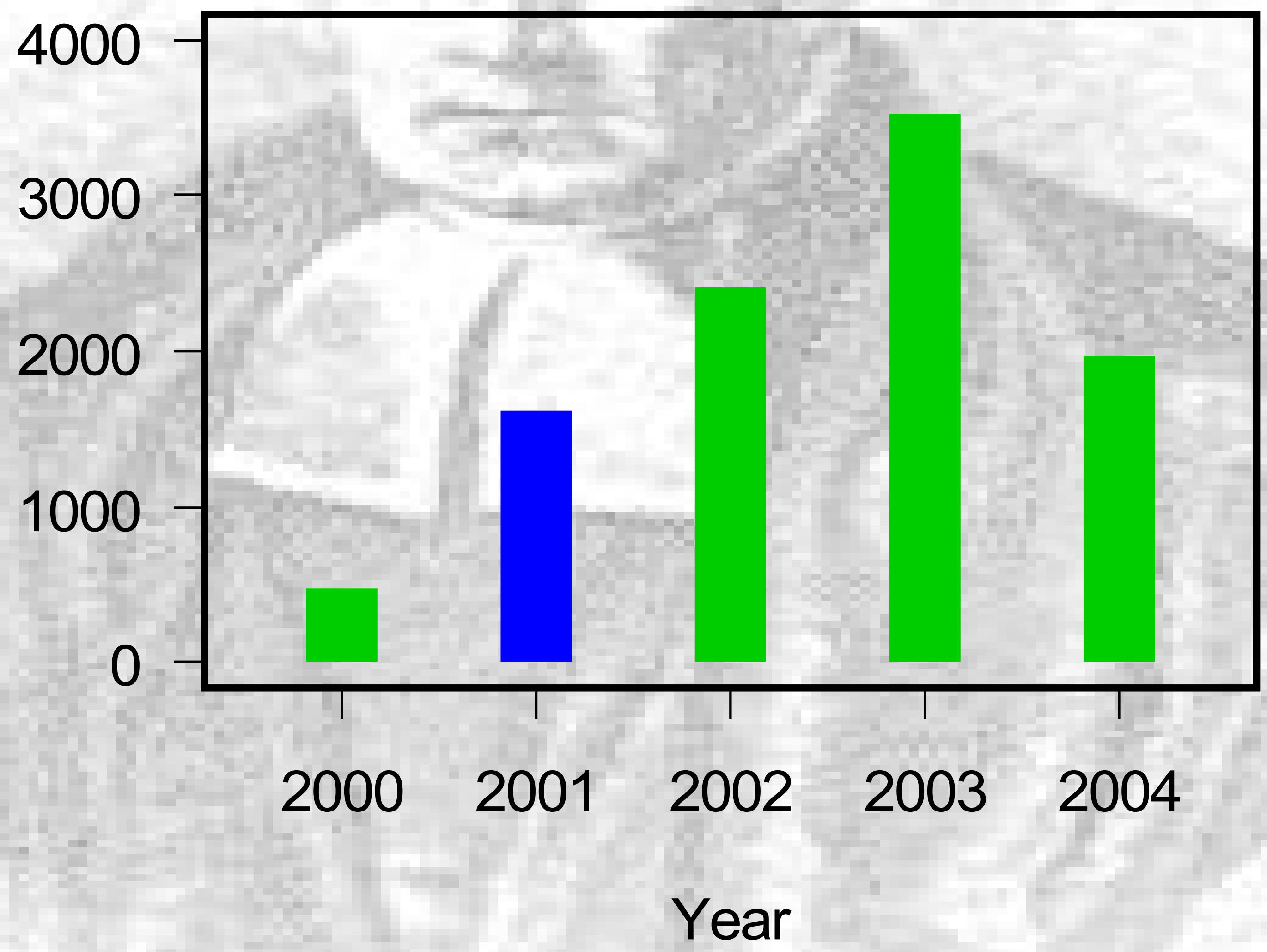
- With MCMC we are repeatedly estimating the parameters
  - drawing them from their posteriors
  - same applies to missing data
- Mirror of an older approach to missing data
    - Called “multiple imputation”

A Model

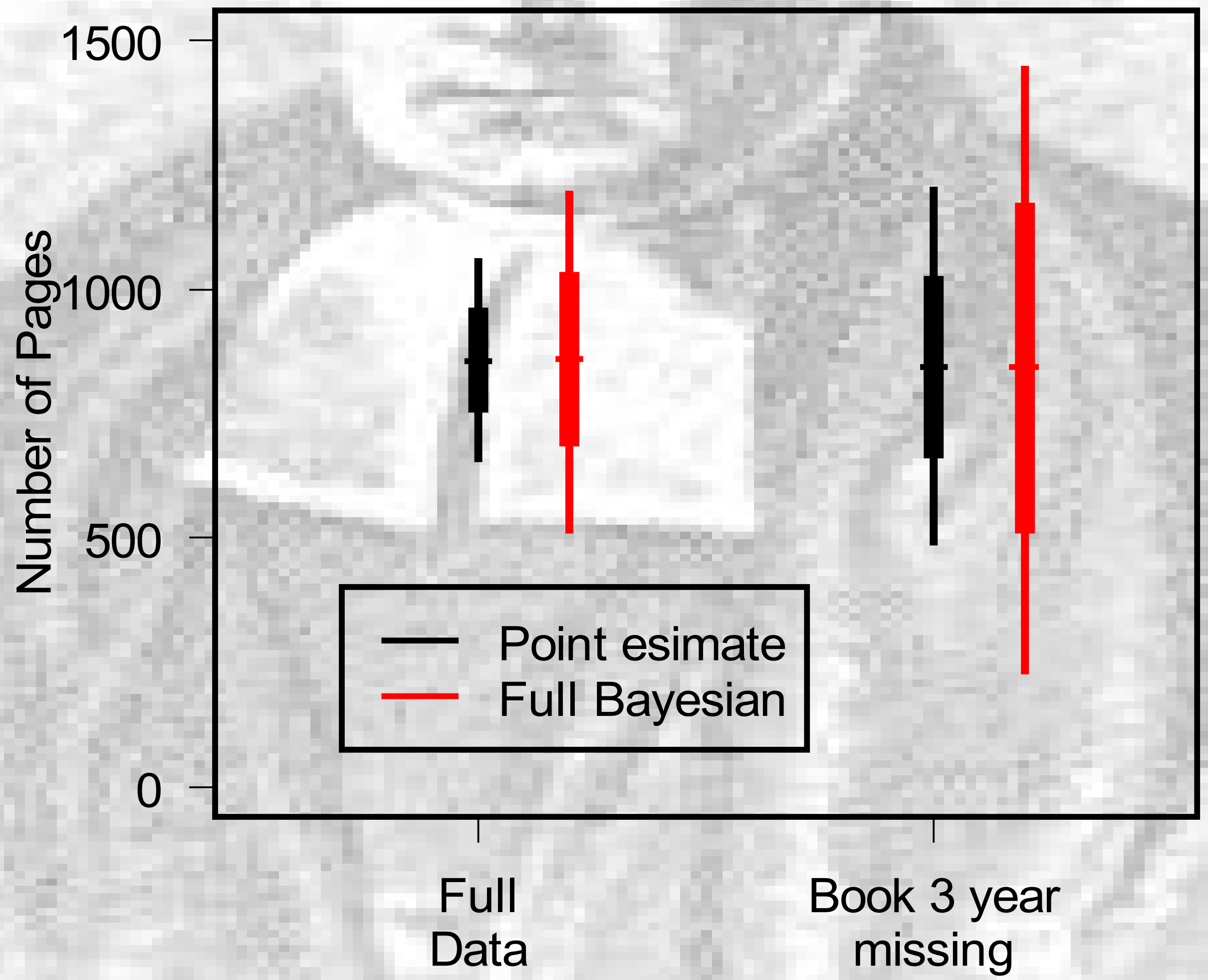
- $P_i \sim N(\alpha + \beta Y_i, \sigma^2)$
- Assume Y_i is random
  - put a simple prior on it
  - but could use a more complex model
- For us, $Y_i \sim U(Y_{i-1}, Y_{i+1})$
  - and round to the nearest integer


DAG
Prior
Y[i]
alpha
sigma
mu[i]
tau
P[i]
beta
for(i in 1:N)

Predicted of Year of Publication



Prediction Length of Book 7



Making it Work

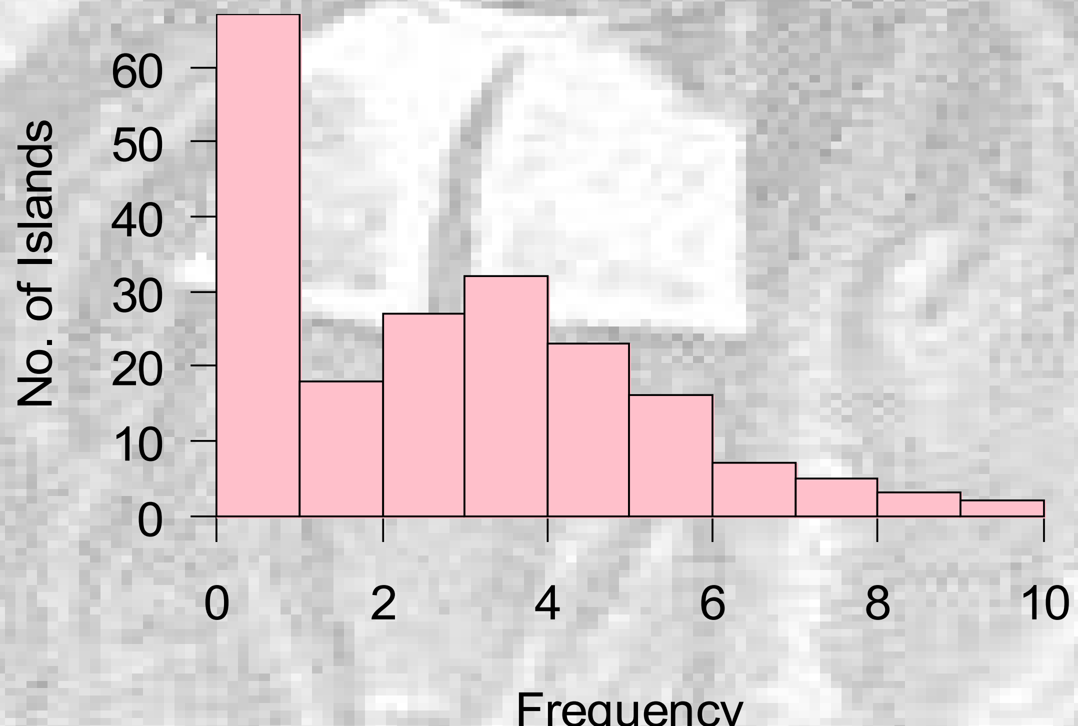
- In this case, the missing data does not affect the point estimates
  - but the uncertainty is higher
- In other cases it can have an effect
  - especially if the data is not missing at random
- With several covariates, removing data points with missing covariates will remove information
- Rather than remove the data points with missing data, estimate the missing data
  - increases precision

When We Need The Missing Data

- Sometimes we need to include the missing data
- e.g. censored data
  - there may be a reason why someone survived the experiment
- We might need to model how the data becomes missing
  - e.g. model that a person survived to the end of the experiment

When Missing Data Helps

- Sample beetles on 200 islands
- For one species, get these abundances:



- Might be a Poisson distribution, but too many zeroes!

The Model: ZIP

- If the rate of capture on each island was constant, then we would expect a Poisson distribution:

$$Pr(N=r)=\frac{\lambda^{r}e^{-\lambda}}{\lambda!}$$

- But we have too many zeroes.
- One explanation: the species only occurs on some islands
- Model: occurrence is binomially distributed
- If species occurs, follows a Poisson distribution

Zero Inflated Poisson

- We end up with a Zero Inflated Poisson Distribution (ZIP)
- Probability:

$$Pr(N=r)=\begin{cases} p+(1-p)e^{-\lambda} & r=0 \\ (1-p)\dfrac{\lambda^{r}e^{-\lambda}}{\lambda!} & r=1,2,3,\ldots \end{cases}$$

- Two parameters: λ and p
- How do we fit this?
  - not a standard distribution

An Indirect Approach

- We do not have to fit the distribution directly
- . Instead we can split the distribution into two:
  - I = 1 if the species is present, else I = 0 and N = 0
  - P(I=1) = p
  - I has a Bernoulli distribution
    - Binomial with 1 trial
- If the species is present, N follows a Poisson distribution
- We augment the data with the un-observed I
  - treat it as missing data

Data Augmentation

- Data augmentation is a common technique
- Makes estimation easier
- But uses more parameters
- Works because we can integrate out the extra parameters
  - take the marginal distribution

The Punchline

- For the missing data, we simulate to estimate the posterior
- We use the posterior to simulate the predicted data
- We could think about the predicted data as missing data, and use data augmentation
- Conceptually, little difference
  - only that the predictions have no observed data after them
- It's all the same framework!