

**Part 10:  
Latent variables and  
missing data**

# Missing data

The Bayesian approach to missing data problems in statistics is straightforward

- since missing data may be treated the same as any other unknown parameter
  - ▶ with a full conditional for GS
  - ▶ or via MH

# Data augmentation

Dealing with missing data is so easy that sometimes it makes sense to *imagine* that there is missing data to make inference more computationally tractable

- such imaginary missing quantities are sometimes called **latent variables**

This technique is called data augmentation, and is especially useful when latent variables

- allow full conditionals for the other unknown parameters to be derived
- and have full conditionals themselves

## Example: Genetic linkage

The genetic linkage of 197 animals is allocated to one of four categories

$$Y = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$$

with probabilities

$$(1/2 + \theta/4, (1 - \theta)/4, (1 - \theta)/4, \theta/4)$$

where  $\theta$  is unknown

## Example: Intractable posterior

Suppose we place a  $\text{Beta}(a, b)$  prior on  $\theta$

Then

$$\pi(\theta|y) \propto \underbrace{\left(\frac{1}{2} + \frac{\theta}{4}\right)^{y_1} (1 - \theta)^{y_2 + y_3} \theta^{y_4}}_{\text{multinomial likelihood}} \times \theta^{a-1} (1 - \theta)^{b-1}$$

multinomial likelihood

$$\propto (2 + \theta)^{y_1} (1 - \theta)^{y_2 + y_3 + b - 1} \theta^{y_4 + a - 1}$$

How can we sample from this?

- One option: MH
- Another option: Data augmentation

# Example: Data augmentation

Consider data augmentation with the data set

$$X = (x_1, x_2, x_3, x_4, x_5)$$

with  $x_1 + x_2 = y_1$ ,  $x_3 = y_2$ ,  $x_4 = y_3$ ,  $x_5 = y_4$

In other words, divide the first cell, with multinomial probability  $(1/2 + \theta/4)$ , into two cells with probabilities  $1/2$  and  $\theta/4$

# Example: Data augmented posterior

Then let  $X = (Y, Z)$  with “missing” or latent data

$$Z: z = x_1 \Rightarrow x_2 = y_1 - z$$

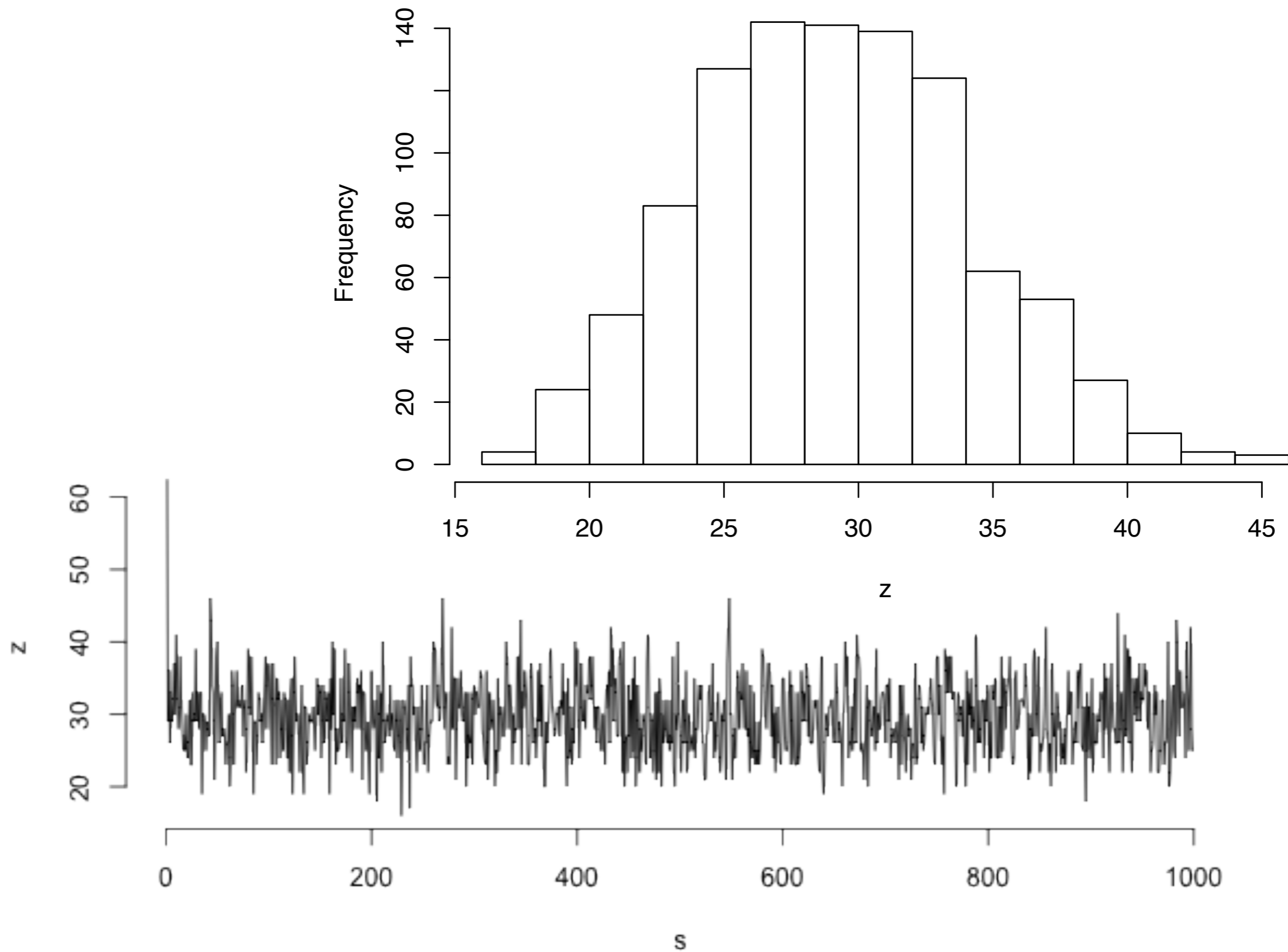
so that

$$\pi(\theta, Z|Y) \propto \binom{y_1}{z} \left(\frac{1}{2}\right)^{y_1-z} \left(\frac{\theta}{4}\right)^z (1-\theta)^{y_2+y_3+b-1} \theta^{y_4+a-1}$$

$$\theta|Z, Y \sim \text{Beta}(z + y_4 + a, y_2 + y_3 + b)$$

$$Z|\theta, Y \sim \text{Bin}\left(y_1, \frac{\theta}{2 + \theta}\right)$$

# Example: latent variable sampling



# Example: latent variable sampling

