# Data missing by accident and by design

Tom Hobbs

June 11, 2024

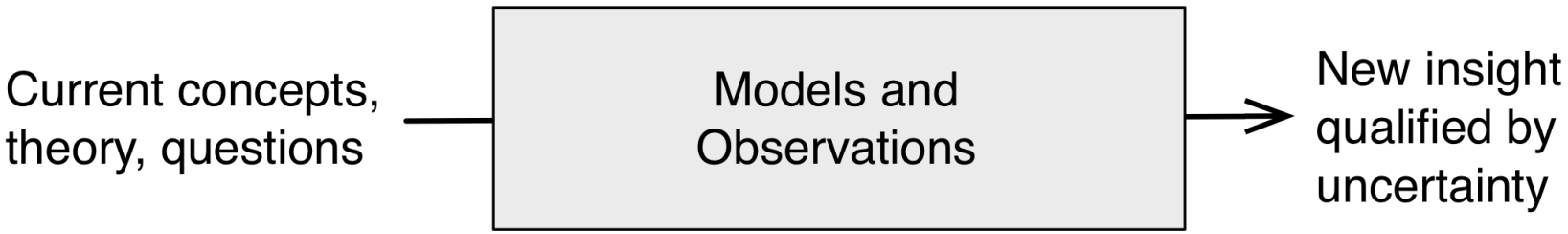







DEB 2042028

# Mevin and Becky would be out of work if there were no missing data.



# Classes of missing data

- By accident–there are observations that we intended to collect that are not collected because of errors in recording, non-response in surveys, instrument failure, tag loss, demonic intrusions etc.
- By design– there are observations that are in the population we seek to learn about that are not included in our sample or experiment.

The question of ignorability asks "When do we need to include information about the data collection process in the model we use for analysis of the data?"

# Missing data are random variables

Remember that Bayesian analysis treats all unobserved quantities as random variables. We see to understand the marginal posterior distributions that give rise to the unobserved quantities conditional on the observed ones. Missing data are treated in the same way as observed data (before they are observed), parameters, and latent variables.

# Data missing by accident

Define an $n$ length vector of responses, $\mathbf{y}$. We will index observations with i=1,2,. . . ,n. We partition the vector $\mathbf{y}$ into observed and missing values, $\mathbf{y} = (\mathbf{y}^{obs}, \mathbf{y}^{mis})'$ and define the n length vector $\mathbf{q}$ as a binary indicator variable

$$q_i = \begin{cases} 0: & y_i \text{ missing from collected data} \\ 1: & y_i \text{ observed} \end{cases}$$

We also have a $n \times J$ matrix of covariates $\mathbf{X}$, where $x_{ij}, j = 1, .., J$ indicates the *jth* covariate for observation $y_i$.

# The joint distribution of observed and missing data

We let $\boldsymbol{\theta}$ be a vector of parameters in an ecological model and $\boldsymbol{\phi}$ be a vector of parameters in a model representing the mechanism that causes the data to be missing, which we will call the missing data model. The joint *likelihood* of the full data, including missing and non-missing observations is

$$[\mathbf{y}, \mathbf{q} \,|\boldsymbol{\theta}, \boldsymbol{\phi}] = [\mathbf{y}^{obs}, \mathbf{y}^{mis}, \mathbf{q} \,|\, \boldsymbol{\theta}, \boldsymbol{\phi}]$$

.

# The joint distribution of observed and missing data

We can factor the joint likelihood

$$[\mathbf{y}^{obs}, \mathbf{y}^{mis}, \mathbf{q} \mid \theta, \phi] = [\mathbf{q} \mid \mathbf{y}^{obs}, \mathbf{y}^{mis}, \theta, \phi][\mathbf{y}^{obs}, \mathbf{y}^{mis} \mid \theta, \phi],$$

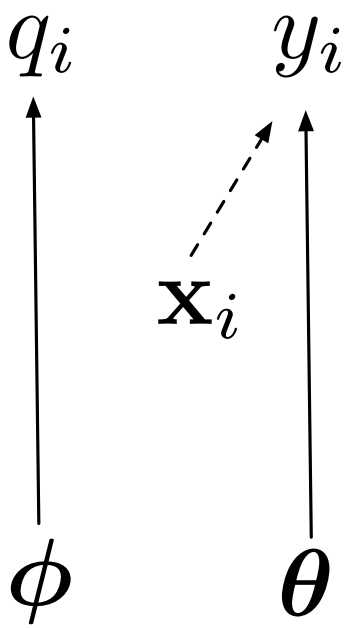
which simplifies to

$$[\mathbf{y}^{obs}, \mathbf{y}^{mis}, \mathbf{q} \mid \theta, \phi] = \overbrace{[\mathbf{q} \mid \mathbf{y}^{obs}, \mathbf{y}^{mis}, \phi][\mathbf{y}^{obs}, \mathbf{y}^{mis} \mid \theta]}^{\text{missing data model ecological model}}$$

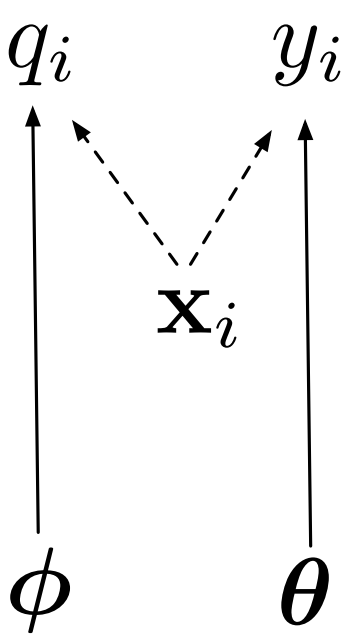
assuming that $\theta$ and $\phi$ are conditionally independent.

# The types of missing data

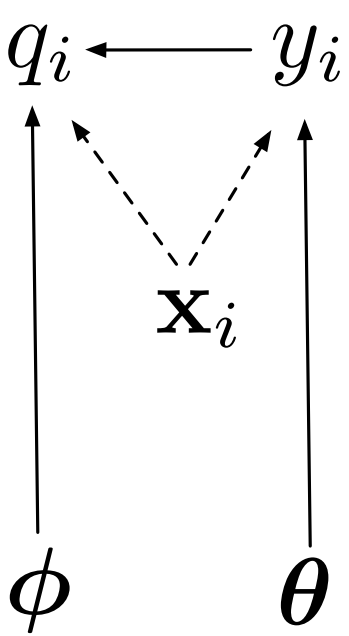
A. Missing completely at random



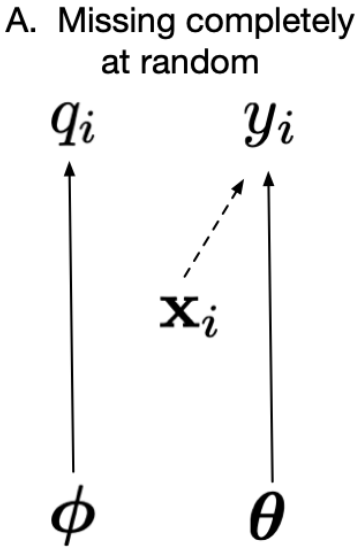

B. Missing at random




C. Missing not at random



# Missing completely at random




$$\overbrace{[\mathbf{y}^{obs}, \mathbf{q}\,|\theta, \phi]}^{\text{likelihood of observed data}} = \overbrace{[\mathbf{q}|\phi]}^{\text{missing data model}} \times \overbrace{[y^{obs}|\theta]}^{\text{ecological model}}$$

The missing data mechanism is *ignorable*. The probability that an observation is missing is the same for all observations. You may assign NA's to missing data and go home and have a beer.

$$y_i \sim [y_i|g(\theta, \mathbf{x}_i), \sigma^2]$$

# Data imputation

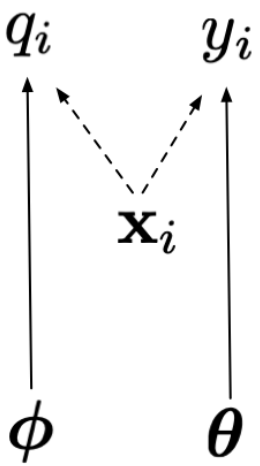
Missing values are *imputed* from the likelihood. If $y_i$ is NA, we make a draw using

$$y_i^{mis(k)} \sim [y_i | g(\theta^{(k)}, \mathbf{x}_i), \sigma^{2(k)}],$$

at each MCMC iteration indexed by $(k)$, which requires that covariates are *not* missing. More about that soon.

# Missing at random

B. Missing at random




$$\overbrace{[\mathbf{y}^{obs}, \mathbf{q} \,|\theta, \phi]}^{\text{likelihood of observed data}} = \overbrace{[\mathbf{q}|\phi, \mathbf{X}]}^{\text{missing data model}} \times \overbrace{[\mathbf{y}^{obs} \,|\, \theta, \mathbf{X}]}^{\text{ecological model}}$$

The missing data mechanism is ignorable *conditional on the covariates*. You may assign NA's to missing data if the ecological model contains covariates explaining the missingness.

# Examples: Ignorable or non-ignorable?

You are studying the relationship between the number of non-native, invasive species of herbaceous plants and native species diversity in a series of plots on a mountain landscape. You have a limited budget.

1) You give the leader of your field crew a list of ordered, randomly selected plot numbers. You instruct the leader to complete observation protocols on as many as possible during a day. Some plots are not done.

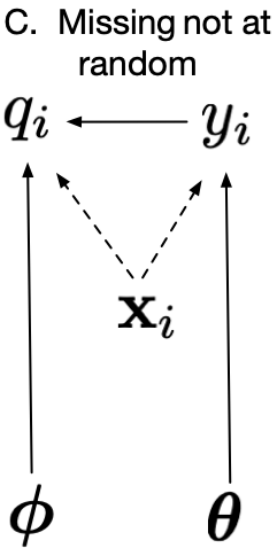
2) You simply tell the field leader to do as many plots as possible without providing a ordered, random list. He or she decides to forgo sampling plots on steep slopes because they require longer travel time.

# Examples: Ignorable or non-ignorable?

You annually measure the length of a species fish in streams soon after peak runoff in a national park.

1) Constraints on funding cause you to miss three out of 10 years.
2) You missed three years because stream run-off was too high to allow plots to be sampled during the period when your research permit allowed access to your sample sites.

# Missing not at random




$$\overbrace{[\mathbf{y}^{obs}, \mathbf{q} \,|\theta, \phi]}^{\text{likelihood of observed data}} = \overbrace{[\mathbf{q}|\mathbf{y}^{obs}, \mathbf{X}, \phi]}^{\text{missing data model}} \times \overbrace{[\mathbf{y}^{obs} \mid \theta, \mathbf{X}]}^{\text{ecological model}}$$

In this case the missing data mechanism is not ignorable and must be explicitly modeled.

# When do you *know* the missing data mechanism is ignorable?

*You don't.* Good judgement is required, in the same way that judgement is required in choosing priors.

- Plot **q** against available covariates, even those that you might not include in your ecological model. Pattern in covariates related to missingingness indicates they should be included in the model and should not be eliminated by model selection.
- Plot mean of **q** against the mean of **y** to determine if there are data missing not at random, which is our next topic. (This requires replication over time or space.)

# Summary up to now

How to handle missing data:

1) Missing completely at random: Assign NA to missing values
2) Missing at random: Assign NA to missing values after assuring that covariates that influence missingness and the ecological process are included in the model
3) Missing not at random: Assign NA to missing values and include a model of the missing data mechanism that exploits the $\mathbf{q}$ vector

In all cases inferences on parameters and posterior predictions are reliable. Posterior predictive checks will not be exact for data missing at random or missing not at random

# Posterior predictive checks

- $T(\mathbf{y}, \boldsymbol{\theta})$ is a test statistic calculated from the observed data.
- $T(\mathbf{y}^{new}, \boldsymbol{\theta})$ is the corresponding statistic based on the posterior predictive realizations.
- Calculate:

$$p_b = \Pr(T(\mathbf{y}^{new}, \boldsymbol{\theta}) \geq T(\mathbf{y}, \theta) \mid \mathbf{y})$$

- If $p_b$ is very large or very small (i.e., close to 1 or 0), **it indicates lack of fit**.

# The problem and a solution using imptuation

When data are missing at random or not at random the *data* will have missing values but the *simulated data* will not because these missing values have been imputed. This means the comparison of statistics between the two sets will not be proper. Moreover, you will not be able to compute a statistic on the data in JAGS if it contains missing values. On the R side, think about what na.rm = TRUE is doing).

This is probably not a big problem if you do your posterior predictive checks on the R side and use na.rm = TRUE because approximate Bayesian P values are probably sufficient. However, if you can't sleep at night without an exact value, see the handout "Posterior predictive checks using multiple imputation."

# Missing covariates

Including NA's for the $y_i$ will automatically model them because the likelihood is a model of how the the $y_i$ arise. JAGS will hum merrily along without complaint

Covariates: Cannot be NA. Missing data must be modeled. In the simplest case,

$$x_i \sim [x_i \mid \mu_x, \sigma_x^2]$$

+ priors on $\mu_x$ and $\sigma_x^2$.

# The model for a single covaiate

$$\mu_i = \exp(\beta_0 + \beta_1 x_i) \tag{1}$$

$$y_i \sim \text{lognormal}(\log(\mu_i), \sigma^2) \tag{2}$$

$$\beta_0 \sim \text{normal}(0, 1000) \tag{3}$$

$$\beta_1 \sim \text{normal}(0, 1000) \tag{4}$$

$$\sigma^2 \sim \text{inverse gamma}(.001, .001) \tag{5}$$

$$x_i \sim \text{gamma}\left(\frac{\mu_x^2}{\sigma_x^2}\frac{\mu_x}{\sigma_x^2}\right) \tag{6}$$

$$\mu_x \sim \text{gamma}\left(\frac{100}{100}\frac{10}{100}\right) \tag{7}$$

$$\sigma_x \sim \text{uniform}(0, 20) \tag{8}$$

$$[\beta, \sigma^2, \mu_x, \sigma_x \mid \mathbf{y}, \mathbf{x}] \propto \prod_{i=1}^{n} [y_i|\mu_i, \sigma_i][x_i|\mu_x, \sigma_x] \times \text{priors} \tag{9}$$

# Missing covaraites

With multiple covariates (which will be the usual case):

$$\mathbf{x}_i \sim \text{multivariate normal}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma})$$

with priors on the $\boldsymbol{\mu}_x$ and $\Sigma$.

# Missing covariates

The previous cases assume that a model of the mean is adequate for modeling the missing covariates. More detailed models may be needed. The same principles apply to covariates as we developed for responses. You must model covariates to account for non-random missingness, using "covariates to predict covariates."

# Example

Suppose nitrogen mineralization rate of soil cores is an important covariate for a model predicting soil fertility. Mineralzation rates tend to be higher in moist soils. You suspect that values from cores taken from particularity wet soil were missing because the assay failed for these cores. You also have a measure of soil moisture for each core. How would you model the x?

# Best practices

- Think about missing values when you design your study. Proper randomization can assure that missing values occur completely at random. There is often a trade-off between the best, ordered randomization scheme and efficiency in the field.
- Always explore missing values if you have many of them. Do the plots described above and think about how missing values might arise.
- It is good to model missing $x's$ to avoid throwing away data when there is a single missing value in a vector of values.
- You may need to conduct calibration studies to inform priors in explicit missing data models needed in the missing not at random case.

# Data missing by design

Experimental and sampling designs are recipes for excluding observations from the population we seek to understand. Why? Because we can't observed everything.

The same general principles apply to data missing by design as missing by accident except now the vector $q$ indexes the *complete data*,

$$q_i = \begin{cases} 1, & \text{if } y_i \text{ observed} \\ 0, & \text{if } y_i \text{ not observed but observable} \end{cases}$$

# Classes of missing by design

- Missing completely at random applies to completely random samples or completely randomized experiments. This is the only case where the way the data were collected can be ignored (with the caveat below).
- Missing at random (and known) applies to all other commonly used designs that have restrictions on the randomization (e.g. stratified random samples, randomized complete block experiments, repeated measures). These are ignorable conditional on covariates or, equivalently, proper subscripting of parameters.
- Missing not at random is an advanced case. See Gelman, A., J. B. Carlin, H. S. Stern, D. Dunson, A. Vehhtari, and D. B. Rubin, 2013. Bayesian data analysis. Chapman and Hall / CRC, London, UK. Chapter 8

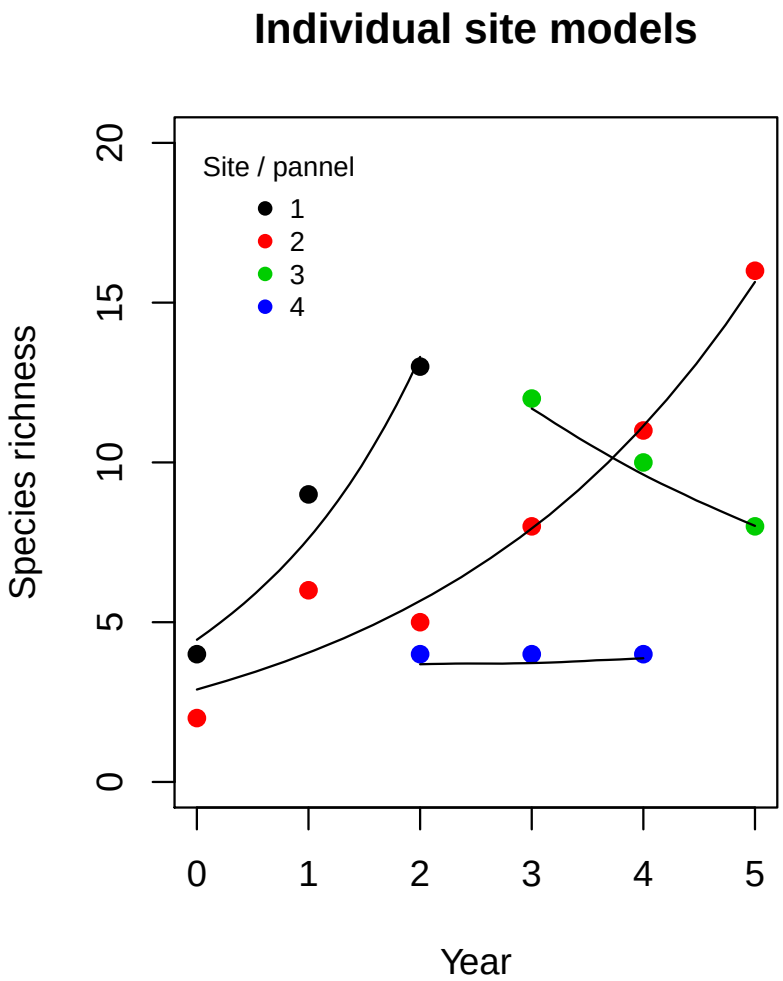
# Caveat

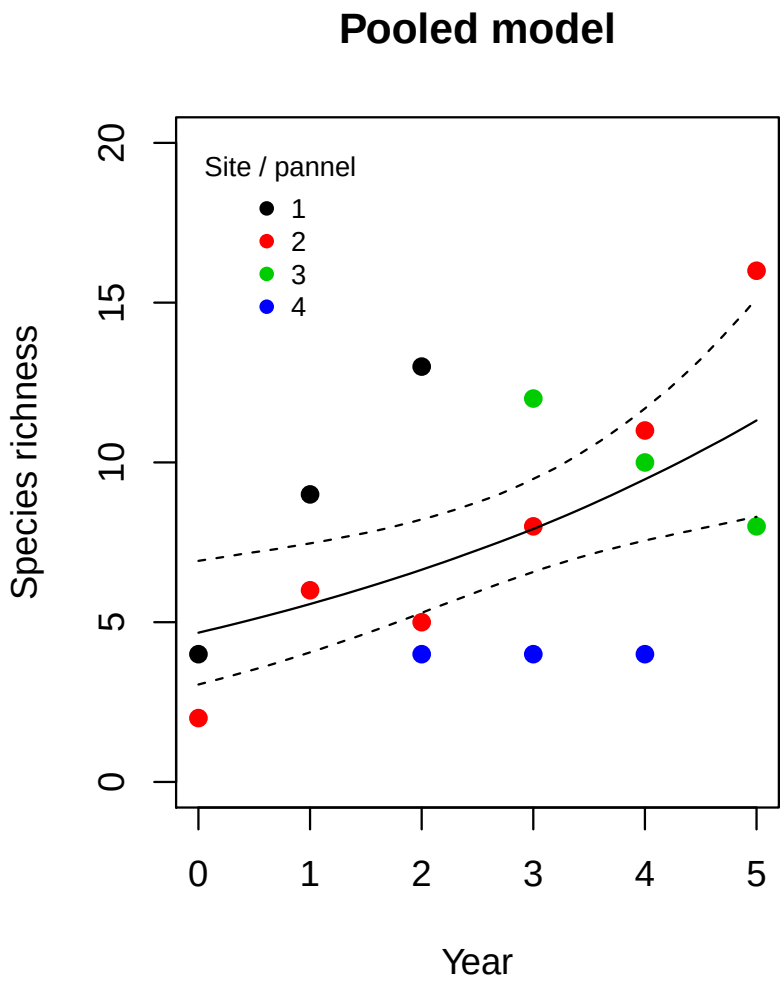
Even completely random designs are not ingornable if the population is finite. More about that soon.

# Topics

- Ignorable conditional on proper model specification by subscripting (briefly now and in lab)
- Inference from stratified random samples (lab)
- Inference on repeated measures (lab)
- Finite population sampling (now and in lab)

# Proper model specification



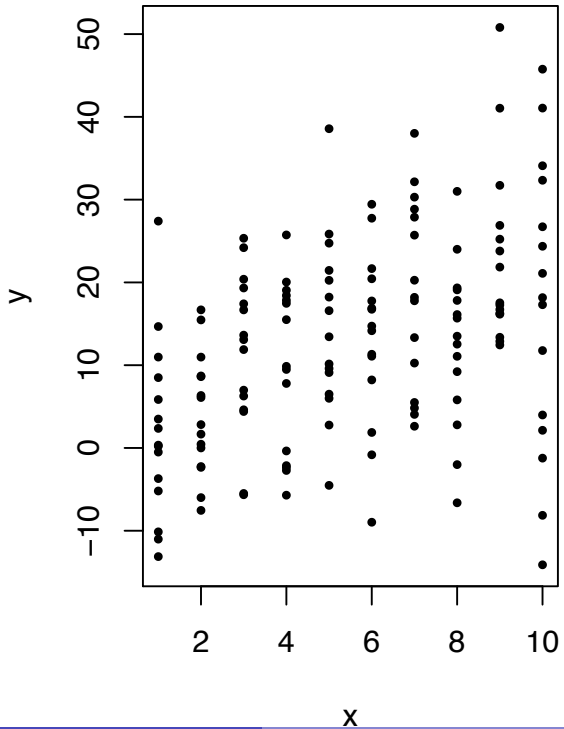



# Proper model specification

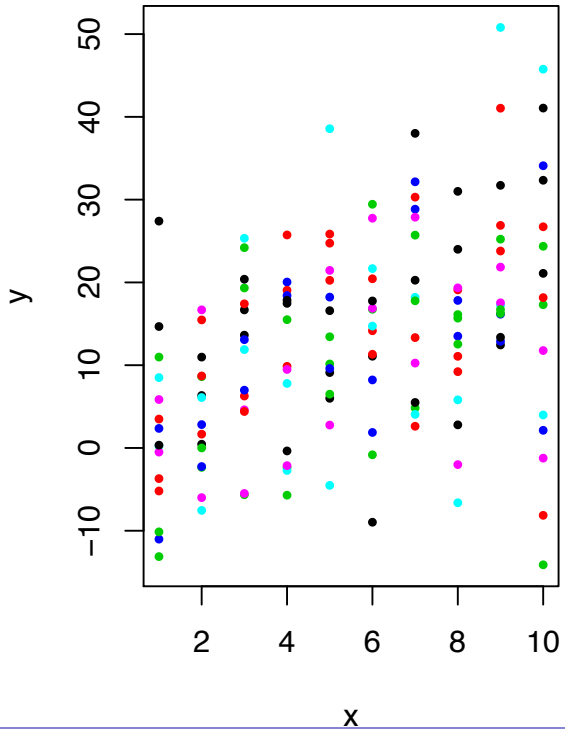
$$\mu_i = \beta_0 + x_i$$
$$y_i \sim \text{normal}(\mu_i, \sigma^2)$$
$$i = 1, ..., n$$

$$\mu_{ij} = \beta_{0_j} + x_{ij}$$
$$y_{ij} \sim \text{normal}(\mu_{ij}, \sigma^2)$$
$$\beta_{0_j} \sim \text{normal}(\mu_{\beta_0}, \sigma^2_{\beta_0})$$
$$j = 1, ..., J$$
$$i = 1, ..., n_j$$

**Completely random design**




**Grouped design**



# Example: Finite population sampling

Our usual assumption is that we take a sample of $n$ observations that is far smaller than the number of possible samples in the population we seek to understand $n \ll N$. However, in many cases this assumption does not hold. Proper inference requires including information on $n$ and $N$ in computation of means, medians, and their credible intervals but not on the "superpopulation" parameters in our model $\theta$.

# Recall the posterior predictive distribution

$$[y^{miss}|\mathbf{y}] = \int_{\theta} [y^{miss} \mid \boldsymbol{\theta}][\boldsymbol{\theta} \mid \mathbf{y}] d\boldsymbol{\theta}$$

# Algorithm for Monte Carlo integration:

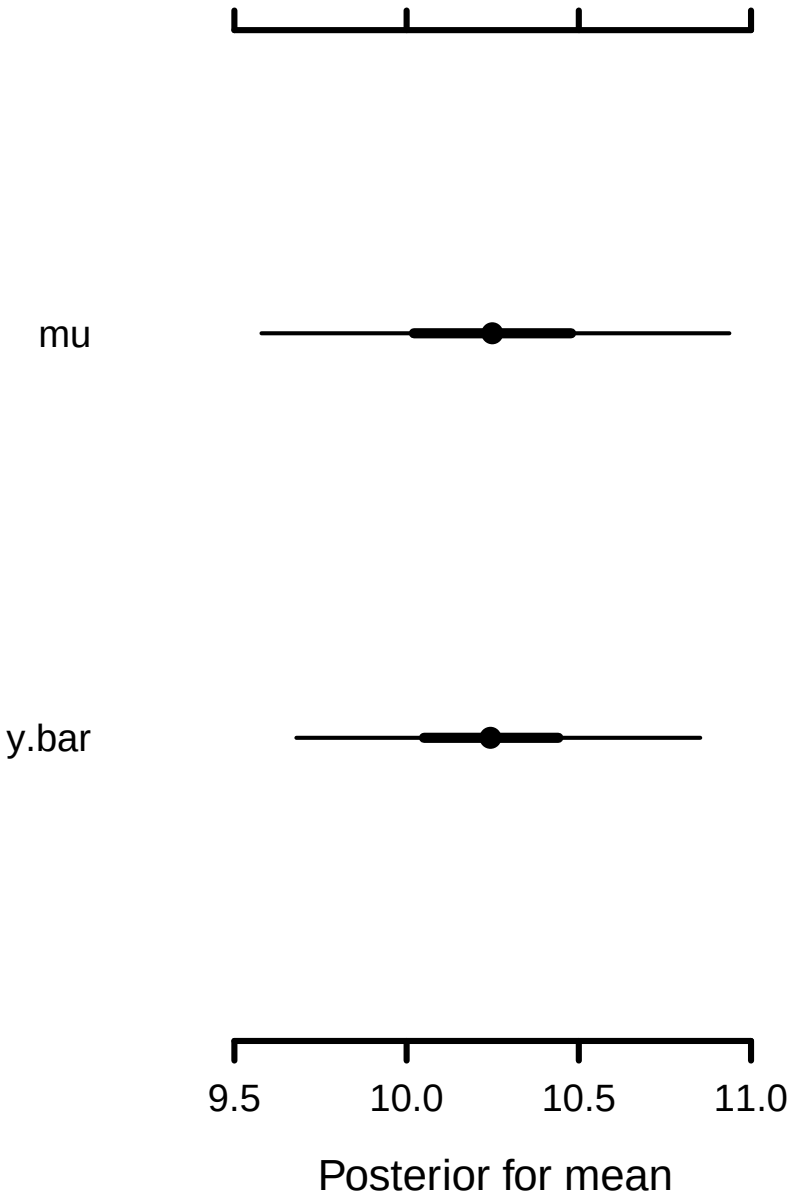
1. Fit the model to find values for $\theta$.
2. At each MCMC iteration (indexed by $t$) make draws of $i = 1, ..., N - n$ *unobserved* values $y^t_{miss,i} \sim [y^t_{miss,i} \mid \theta^t, x_i]$
3. Compute $\bar{y}^t = \frac{n}{N}\text{mean}(\mathbf{y}_{obs}) + \frac{N-n}{N}\text{mean}(\mathbf{y}^t_{miss})$

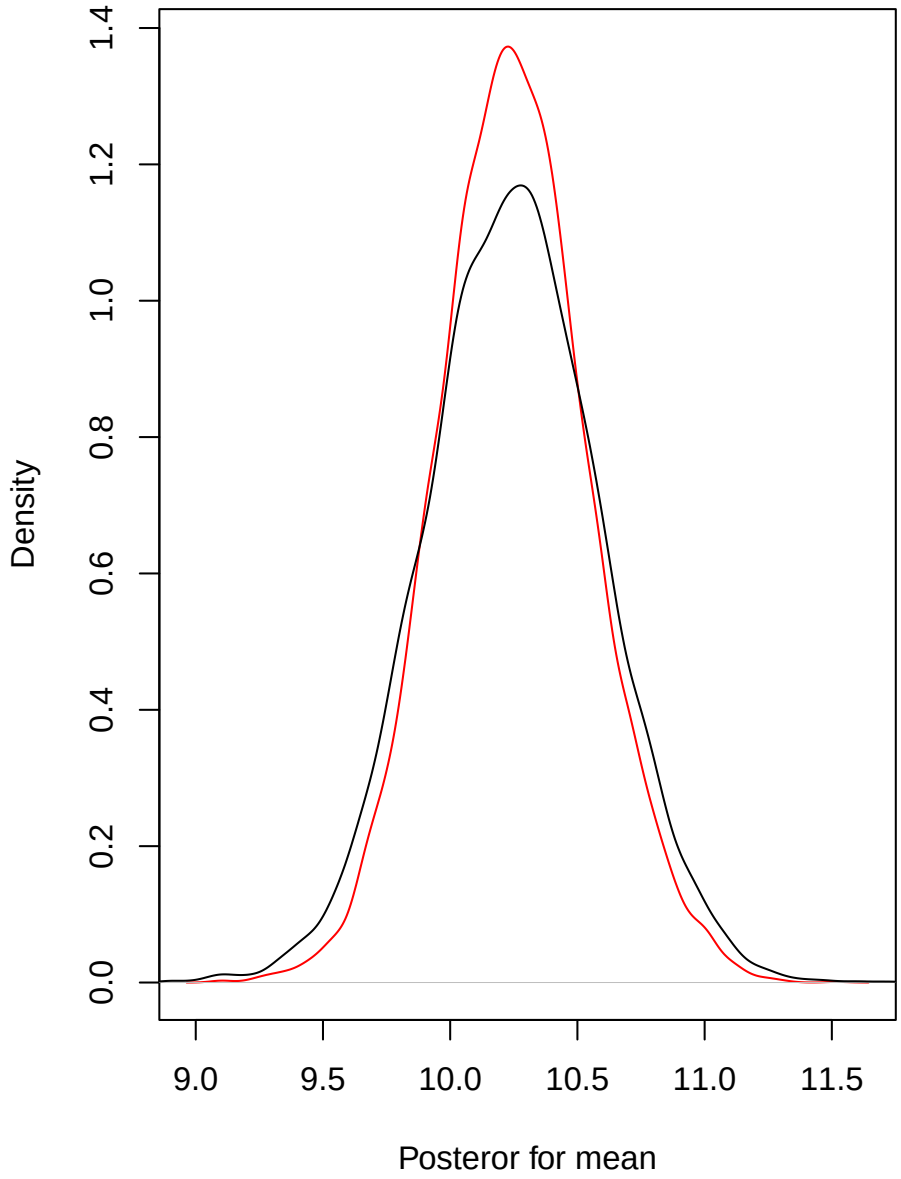
Could compute any function of the simulated and obserevd vectors of observations: median of $\mathbf{y}$, mean of $\log(\mathbf{y})$, etc.

# Code for algorithm

```
{
sink("FiniteMean.R")
cat("
model{
mu ~ dnorm(0,.00001)
sigma ~ dunif(0,20)
for(i in 1:length(y.obs)){
  y.obs[i] ~ dnorm(mu, sigma^-2)
}
for(i in 1:(N-n)){
  y.miss[i] ~ dnorm(mu, sigma^-2) #simulate missing data
}
y.bar = n/N * mean(y.obs) + (N-n)/N * mean(y.miss)
} #end of model
",fill=TRUE)
sink()
}
```

# Output: Note shrikage of finite mean (red line)



# When do you need to consider finite samples?

If $n$ is small and $N$ is large, $\bar{y}$ converges on $\mu$.

$$\bar{y}|y_{obs} \approx \text{normal}\left(\bar{y}, \frac{1}{n} - \left(\frac{1}{N}\right)\sigma^2\right)$$

You can see that $N$ as small as 1000 with a relatively small $n$ (50) will cause less than a 2% change in the variance.

# Further study of ignorability and missing data


- L. J. Zachmann, E. M. Borgman, D. L. Witwicki, M. C. Swan, C. McIntyre, and N. T. Hobbs. Bayesian models for analysis of inventory and monitoring data with non-ignorable missingness. Journal of Agricultural Biological and Environmental Statistics, https://doi.org/10.1007/s13253-021-00473-z, 2021.
- A. Gelman, J. B. Carlin, H. S. Stern, D. Dunson, A. Vehhtari, and D. B. Rubin. Bayesian data analysis. Chapman and Hall / CRC, London, UK, 2013. Chapter 8.
- A. Gelman and J. Hill. Data analysis using regression and multilievel / hierarchical modeling. Cambridge University Press, Cambridge, UK, 2009. Chapter 25
- http://www.columbia.edu/~cjd11/charles_dimaggio/DIRE/styled-4/styled-11/code-10/

# Further study of ignorability and missing data


- http://www.bias-project.org.uk/Missing2012/Lectures.pdfhttp://www.bias-project.org.uk/Missing2012/MissingIndex.htm
- http://www.bias-project.org.uk/Missing2012/Lectures.pdf
- https://web.as.uky.edu/statistics/users/pbreheny/701/S13/notes/4-23.pdf
- W. A. Link and R. J. Barker. Bayesian inference with ecological applications. Academic Press, 2010. Section 8.5
- Schwob, M.R., M.B. Hooten, T. McDevitt-Gales. (2023). Dynamic population models with temporal preferential sampling to infer phenology. Journal of Agricultural, Biological, and Environmental Statistics, 28: 774-791.