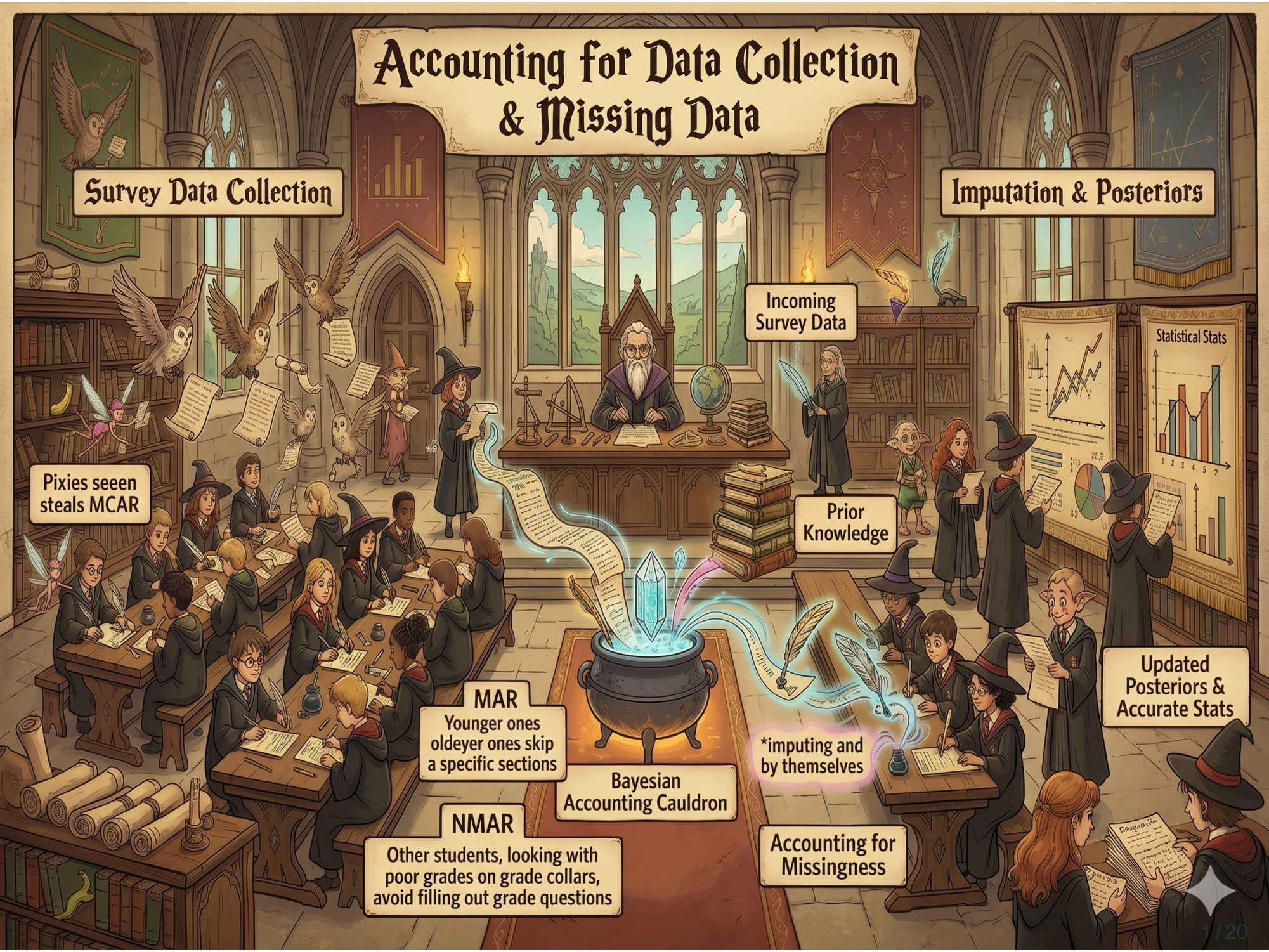
Accounting for Data Collection & Missing Data
Survey Data Collection
Imputation & Posteriors
Incoming Survey Data
Statistical Stats
Pixies seeen steals MCAR
Prior Knowledge
Updated Posteriors & Accurate Stats
MAR
Younger ones oldeyer ones skip a specific sections
*imputing and by themselves
Bayesian Accounting Cauldron
NMAR
Other students, looking with poor grades on grade collars, avoid filling out grade questions
Accounting for Missingness

Overview

- The information describing data collection should be included in the analysis
- We distinguish between:
  - Complete data: includes both observed and missing components
  - Observed data: what we actually see
  - Missing data: unobserved but relevant
- Different data collection mechanisms require different probabilistic models
- This lecture covers:
  1. Ignorability: When we can ignore the missing data mechanism
  2. Missing not at random: When we cannot ignore the mechanism
  3. Censoring and truncation: Special cases of incompleteness

Missing data

- The information describing data collection should be included in the analysis.
- We will consider the complete data to consist of both observed data and missing data.
- Typically, we can use a probabilistic model to characterize the joint distribution observed and missing data under certain assumptions.

Examples:

- In sampling, the complete data consist of N units in the population, and observed data consist of n units in the sample.
- In clinical trials, the complete data consist of outcomes under all treatments for all units, and observed data consist of outcomes under the observed treatments.
- In many applications where we cannot observe the true response variable, e.g., the Slovenia poll example, the complete data is the binary answers from each respondent, and the observed data is the observed (binary plus missing) answers from each respondent.

Notation

- Let $Y = (y_1, ..., y_N)$ be the complete data, where each y_i is a vector.
- Let $I = (I_1, ..., I_N)$ be the indicator matrix of the same dimension as Y such that

$$I_{ij} = \begin{cases} 1 & \text{if } y_{ij} \text{ is observed} \\ 0 & \text{if } y_{ij} \text{ is missing} \end{cases}$$

- We usually consider I as a random variable, i.e., observations are not made deterministically.
- But typically the realized indicator matrix I is fully observed, i.e., we know which samples are observed and which are missing.
- Consider a model for the complete data $p(Y|\theta)$, and a model describing missing data mechanism $p(I|Y, \phi)$. The joint distribution of Y and I becomes

$$p(Y, I|\theta, \phi) = p(Y|\theta)p(I|Y, \phi)$$

We call this the complete-data likelihood.

Notation

- With slight abuse of notation, write $Y = (y_{obs}, y_{mis})$, where y_{obs} denote the collection of y_{ij}'s with $I_{ij} = 1$ and y_{mis} the collection of y_{ij}'s where $I_{ij} = 0$.
- The entire collection of observed data is (y_{obs}, I).
- The observed-data likelihood is thus

$$p(y_{obs}, I|\theta, \phi) = \int p(Y|\theta)p(I|Y, \phi)dy_{mis} = \int p(y_{obs}, y_{mis}|\theta)p(I|y_{obs}, y_{mis}, \phi)dy_{mis}$$

- Note θ and ϕ can be dependent through hyperpriors or even deterministically related.
- Notice that even though we know which samples are actually observed in the particular sample, i.e., I is fully known, I is still a random variable and part of the observed (random) data.
- But in some problems, we can make inference using the "usual" likelihood $p(y_{obs}|\theta)$, which will require additional assumptions.

An implicit assumption

- The first observation of this factorization

$$p(Y, I|\theta, \phi) = p(Y|\theta)p(I|Y, \phi)$$

  is that Y is generated 'before' I, i.e., whether or not a measurement is recorded does not affect the outcome of any units.
- This type of assumption is usually called 'stability' assumption.
- Stability assumptions can be violated in certain applications, where the outcome may change depend on which samples are observed.
  - Taking a measurement of the soil changes the moisture of surrounding soil.
  - By going to the hospital and taking test for a disease, the patient is more likely to contact the disease.
- Violation of stability assumption can be mitigated by careful modeling of the interference process, but we are not going to discuss that here. It is an active area of research in causal inference.

Adding more simplifying assumptions

- Often we do not care directly about ϕ. We only want to infer θ,

$$p(\theta|y_{obs}, I) \propto \int p(\theta, \phi)p(y_{obs}, I|\theta, \phi)d\phi$$
$$= \int\int p(\theta, \phi)p(y_{obs}, y_{mis}|\theta)p(I|y_{obs}, y_{mis}, \phi)dy_{mis}d\phi$$

- If we make the following assumptions:
  1. $p(I|y_{obs}, y_{mis}, \phi) = p(I|y_{obs}, \phi)$
  2. $p(\theta, \phi) = p(\theta)p(\phi)$
- Then we can simplify it into

$$p(\theta|y_{obs}, I) = p(\theta)\int\int p(\phi)p(y_{obs}, y_{mis}|\theta)p(I|y_{obs}, \phi)dy_{mis}d\phi$$
$$= p(\theta)p(y_{obs}|\theta)p(I|y_{obs})$$
$$\propto p(\theta|y_{obs})$$

- When $p(\theta|y_{obs}, I) = p(\theta|y_{obs})$, we say the missing data mechanism or sampling design is ignorable.

Ignorability

- The two conditions on the previous slide are sufficient for ignorability:
  1. Missing at random (MAR): $p(I|y_{obs}, y_{mis}, \phi) = p(I|y_{obs}, \phi)$
  2. Distinct parameters: $p(\theta, \phi) = p(\theta)p(\phi)$
- In many situations we have covariates x. The same discussion extends to conditional ignorability $p(\theta|y_{obs}, I, x) = p(\theta|y_{obs}, x)$.
- Ignorability condition means we can make inference about θ using only the observed data.
- A stronger assumption to MAR is 'missing completely at random' (MCAR), which requires $p(I|y_{obs}, y_{mis}, \phi) = p(I|\phi)$.

Simple random sampling

- In simple random sampling and randomized experiments, the units in the sample are selected with $p(I|x,y) = p(I)$.
- Ignorability is satisfied, so we usually do not even need to explicitly differentiate y_{obs} and y_{miss}. Inference on θ is performed based on the observed-data likelihood.
- For a simple continuous outcome, if we are interested in the population mean $\bar{y}$. From a finite population perspective

$$\bar{y} = \frac{n}{N}\bar{y}_{obs} + \frac{N-n}{N}\bar{y}_{mis}$$

  When n is small relative to N, $\bar{y} \approx \bar{y}_{mis}$.
- With a simple normal model $y|\mu \sim N(\mu, \sigma^2)$, the posterior distribution

$$p(\mu|y_{obs}, I) \quad \propto \quad p(\mu|y_{obs})$$

  and then draw from the posterior predictive distributions

$$p(\bar{y}_{mis}|y_{obs}) = \int p(\bar{y}_{mis}|\mu)p(\mu|y_{obs})d\mu$$

  where $\bar{y}_{mis}|\mu \sim N(\mu, \frac{1}{N-n}\sigma^2)$.

Stratified sampling

- In stratified sampling with K strata and n_k allocated samples in each stratum, we perform simple random sampling within each stratum of the population. Let x denote the stratification variable,

$$p(I|x, y) = p(I|x)$$

- Consider data collection stratified by a categorical variable X representing K age groups, and a simple model

$$y_i|x_i = k \sim N(\mu_k, \sigma_k^2)$$
$$p(I|X) \propto 1_{\sum_{i:x_i=k} I_i = n_k, k=1,\ldots,K}$$

  Then $p(\mu, \sigma|y_{obs}, X, I) \propto p(\mu, \sigma|y_{obs}, X)p(I|X) \propto p(\mu, \sigma|y_{obs}, X)$.

- Consider again if we are interested in the population mean $\bar{y}$, when $n_k \ll N_k$, $\bar{y} \approx \sum_k \frac{N_k}{N} \bar{y}_{k,mis}$ which can be obtained via posterior predictive distribution of $y_{mis}|y_{obs}$ and known strata size N_k's.

More about stratified sampling

- What if the sample within each stratum are not simple random samples, but depend on the value of y_i?
- E.g., if we perform simple random sample on any observations with $y_i > 10$.

$$y_i|x_i = k \sim N(\mu_k, \sigma_k^2)$$
$$p(I|X,Y) \propto 1_{\sum_{i:x_i=k, y_i>10} I_i = n_k, k=1,\ldots,K}$$

Then missing at random is not satisfied, and

$$\begin{aligned} p(\mu,\sigma|y_{obs},X,I) &\propto \int p(\mu,\sigma,y_{obs},y_{mis},X,I)dy_{mis} \\ &\propto \int p(y_{obs},y_{mis},I|\mu,\sigma,X)p(\mu,\sigma)dy_{mis} \\ &= \int p(I|y_{obs},y_{mis},X)p(y_{obs},y_{mis}|\mu,\sigma,X)p(\mu,\sigma)dy_{mis} \end{aligned}$$

Note that $p(I|y_{obs},y_{mis},X) \neq p(I|y_{obs},X)$ so the above does not simplify to be proportional to $p(y_{obs}|\mu,\sigma,X)p(\mu,\sigma)$.

The Slovenia poll example revisited

- Recall that we assume that the counts with DKs can be redistributed like

$$(n_{1\star}^{(11)}, n_{1\star}^{(10)}) \mid \theta \sim \text{Mult}(n_{1\star}, (\frac{\theta_{11}}{\theta_{11}+\theta_{10}}, \frac{\theta_{10}}{\theta_{11}+\theta_{10}}))$$

- For each of the i-th individual, switch notation to use y_{i1} and y_{i2} to denote the vote for the two questions, and an equivalent model:

$$y_{i1}|\phi \sim Bern(\phi_1), \quad y_{i2}|y_{i1}, \psi \sim Bern(\psi_{y_{i1}})$$

  i.e., $\theta_{11} = \phi_1\psi_{11}$, or equivalently $\psi_{11} = \theta_{11}/(\theta_{11}+\theta_{10})$.

The Slovenia poll example revisited

- If we ignore DK responses, the naive estimator of $p(y_{i2})$:

$$\sum_y p(y_{i2}|y_{i1} = y, I_{i2} = 1)p(y_{i1} = y|I_{i2} = 1)$$

- If we assume the second question is MAR, then $I_{i2} \perp y_{i2} \mid y_{i1}$ so

$$p(y_{i2}|y_{i1} = y, I_{i2} = 1) = p(y_{i2}|y_{i1} = y) = \psi_y$$

- The naive estimator above is still biased because $p(y_{i1} = y|I_{i2} = 1) \neq p(y_{i1} = y)$, i.e., *whether* one answers the second question is not independent of *how* one answers the first question.
- On the other hand, for the Gibbs sampler we derived previously, we impute the missing y_{i2} instead, this sampling step is valid under MAR since

$$p(y_{i2}|y_{i1}, I_{i2}) \propto p(y_{i2}, I_{i2}, y_{i1}) = p(y_{i2}|y_{i1})p(I_{i2}|y_{i1})p(y_{i1}) \propto p(y_{i2}|y_{i1})$$

- That is, conditional on all (observed and sampled) y_{i1}, we can sample $y_{i2}|y_{i1} \sim Bern(\psi_{y_{i1}})$ without conditioning on I.

Examples of missing not at random

- Missing depend on y deterministically:
  - Weights over/under certain threshold are not recorded completely because the scale has an upper/lower limit.
  - Evaluate survival rate for patients on a treatment plan of taking the pill for 6 months, by evaluating the patients who finished the 6 month treatment.
- Missing depend on y stochastically:
  - An instrument measuring distance more likely to fault when measuring distant objects.
  - Evaluate app crash rate by analyzing app crash report sent by the app, when the report can fail to be produced when certain type of crash happens.
- Missing depend on confounders:
  - Evaluate the effectiveness of COVID vaccine when only people at higher risk were offered the vaccine.
  - When surveying course evaluations, only students who are present in class are given the survey.

Censored data

- Consider IID draws from $N(\theta, 1)$. Suppose we are interested in estimating θ. But we know the measurement process is problematic. Any observations over C is recorded as C instead of the actual value.
- Denote z_i as the true value and y_i the observed value, then $y_i = \min(z_i, C)$.
- Denote $I_i = 1_{y_i < C}$ as the indicator that the i-th observation is not censored at C.
- Missingness now depends on the unknown outcome y.
- Then we have the posterior given the observed data

$$
\begin{aligned}
p(\theta|y, I) &\propto p(\theta) \prod_i p(y_i, I_i|\theta) \\
&= p(\theta) \prod_{I_i=1} p(y_i, I_i|\theta) \prod_{I_i=0} \int p(y_i, z_i, I_i|\theta) dz_i \\
&= p(\theta) \prod_{I_i=1} p(y_i|\theta) p(I_i|y_i) \prod_{I_i=0} \int p(z_i|\theta) p(y_i, I_i|z_i) dz_i
\end{aligned}
$$

- Missing mechanism is deterministic, how do we deal with the probabilities involving I?

Censored data

- For observations smaller than C, they have to be uncensored $p(I_i = 1|y_i) = 1$.
- For observations over C, they have to be censored $p(I_i = 0|y_i = C) = 1$.
- Equivalently, for censored observations, $p(y_i = C, I_i = 0|z_i)$ is 1 if $z_i \geqslant C$ and 0 otherwise. Therefore,

$$
\begin{aligned}
p(\theta|y, I) &\propto p(\theta) \prod_{I_i=1} p(y_i|\theta)p(I_i|y_i) \prod_{I_i=0} \int p(z_i|\theta)p(y_i, I_i|z_i)dz_i \\
&= p(\theta) \prod_{I_i=1} p(y_i|\theta)p(I_i|y_i) \prod_{I_i=0} \int p(z_i|\theta)1_{y_i=C, I_i=0, z_i \geqslant C} dz_i \\
&= p(\theta) \prod_{I_i=1} p(y_i|\theta) \prod_{I_i=0} \int_C^\infty p(z_i|\theta)dz_i \\
&= p(\theta) \prod_{I_i=1} N(y_i; \theta, 1) \prod_{I_i=0} (1 - \Phi(C; \theta, 1))
\end{aligned}
$$

More clarifications on the notation

- Here we parameterize y_{obs} and y_{miss} into y and z because none of the y is 'missing', i.e., they are recorded at C.
- In this parameterization, y does not follow the normal distribution. Only z follows the normal distribution.
- The alternative parameterization y_{obs} and y_{mis} can also be useful. For notation clarity, consider the case that y_i is not recorded at all when $I_i = 0$. Let y_{obs} be $\{y_i : I_i = 1\}$ and y_{mis} be $\{y_i : I_i = 0\}$, then $y \sim N(\theta, 1)$ becomes the correct model,

$$
\begin{aligned}
p(\theta|y_{obs}, I) &\propto p(\theta) \prod_{I_i=1} p(y_i, I_i|\theta) \prod_{I_i=0} p(I_i|\theta) \\
&= p(\theta) \prod_{I_i=1} p(y_i|\theta)p(I_i|y_i) \prod_{I_i=0} \int p(I_i|y_i)p(y_i|\theta)dy_i \\
&= p(\theta) \prod_{I_i=1} p(y_i|\theta) \prod_{I_i=0} \int \mathbf{1}_{y_i>C} p(y_i|\theta)dy_i \\
&= p(\theta) \prod_{I_i=1} N(y_i; \theta, 1) \prod_{I_i=0} (1 - \Phi(C; \theta, 1))
\end{aligned}
$$

Estimation

- Working out the marginal distribution $p(\theta|y, I)$ is useful for understanding what distribution we are working with.
- But computationally, it is often difficult to direct sample $\theta|y, I$.
- In Bayesian analysis, missing data and parameters play the same role. We can use MCMC to sample from the joint distribution $p(\theta, z|y, I)$.
  - Sample $\theta|z, y, I$ in this case reduces to sampling from $p(\theta|z)$, which is straightforward.
  - Sample $z_i|\theta, y_i, I_i = 0$ from

$$p(z_i|\theta, y_i, I_i = 0) \quad \propto \quad N(z_i; \theta, 1)1_{z_i > C} = TruncNormal(z_i; \theta, 1, [C, \infty))$$

  - $z_i|\theta, y_i, I_i = 1$ does not need to be sampled as $z_i = y_i$.

Truncated data

- Suppose that instead of receiving the censored measurements, we receive only the y_i's below C, and do not know how many values there are above C. This is called (right) truncated data.
- For example, if we want to study the time between exposure to a virus and symptom onset among those infected, and we collect data on people who have shown symptoms before today. We will miss people who still have not shown symptoms by today. So the naive estimates will be biased downwards.
- In this case, we can model the total amount of people who are infected, N, together with the rest of the unknown parameters

$$p(\theta, N|y, I) \propto \quad p(\theta, N)(1 - \Phi(C; \theta, 1))^{N-n} \prod_{I_i=1} N(y_i; \theta, 1)$$

- The prior for N needs to be chosen based on the context.

Summary: Key Concepts

- Complete-data framework: Include both observed and missing data
- Ignorability: When we can use only observed data
  - MAR + distinct parameters $\Rightarrow$ can ignore missingness mechanism
- Censoring vs. Truncation:
  - Censoring: observations exists but not recorded
  - Truncation: observations do not exist
- Computation: Treat missing/censored values as random variables and perform data augmentation via MCMC.