Short course on Missing Data

Demetris Athienitis

University of Florida
Department of Statistics
Contents

1 An Introduction to Missing Data ........................................... 5
  1.1 Introduction ......................................................... 5
  1.2 Types of Missing Data ............................................... 5
  1.3 Causes of Missing Data ............................................. 7
  1.4 Simple Solutions .................................................... 8
  1.5 What About Multiple Imputation? ............................... 9
  1.6 Comparable Methods ............................................... 10

2 Introduction to Multiple Imputation ................................. 11
  2.1 Ignorable and Non-ignorable ...................................... 11
  2.2 Introduction to Bayesian Analysis .............................. 12
  2.3 The Gibbs Sampler .................................................. 15
  2.4 Generating Imputations ............................................ 16
  2.5 Estimation ................................................................ 17
  2.6 Variation and Rubin’s Rules ....................................... 17
  2.7 Variance Ratios ........................................................ 18
  2.8 Degrees of Freedom .................................................. 18
  2.9 Scalar Inferences ...................................................... 19
  2.10 How Many Imputations? .............................................. 19
3 Generating Imputed Datasets ........................................ 21

3.1 How we got here .................................................. 21

3.2 Multiple imputation under a normal linear model ...... 24
1. An Introduction to Missing Data

1.1 Introduction

Missing data, or missing values, occur when no data value is stored for the variable in an observation. This happens for a variety of reasons to be discussed later, but we now have techniques that can compute values to “fill in” for the missing entries, aiding in the statistical analysis of datasets with missing values. Our focus will be on multiple imputation, specifically finding algorithms to analyze different distributions, and applications of that approach.

There are many reasons data can be missing data from a dataset, namely:

- Data wasn’t collected
- Subjects declined to provide values
- Information excised
- Technical malfunction
- Respondents drop out
- Subjects unable to respond

1.2 Types of Missing Data

The reality is that most data analysts will run into missing data at some point. Donald Rubin, a leader in the statistics community who has helped shape mainstream missing data techniques, theorized that every data point has some likelihood of being missing. With this understanding, let’s breakdown the different mechanisms in which data can be missing:

1. **Missing Completely At Random (MCAR):** A variable is considered MCAR if the probability of being missing is the same for all cases. Missing data is not related to any factor, known or unknown.

   Ex) A participant flips a coin to decide whether to fill out depression survey.

   Ex) Random sample from population where each member has same chance of inclusion.
2. **Missing At Random (MAR):** A variable is considered MAR if the probability of being missing is NOT completely random, specifically where the probability of being missing can be fully accounted for by the observed data. MAR is more realistic than MCAR. Modern missing data methods often use a MAR assumption. MCAR implies MAR, but NOT conversely.

**Ex)** Male participants less likely to fill out depression survey, because they are male, not because they are depressed.

**Ex)** Random sample from a population, where probability to be included depends on some known property.

3. **Missing Not At Random (MNAR):** (aka non-ignorable nonresponse) If neither MCAR nor MAR, then the data are MNAR. The value of the variable that’s missing is related to the reason it’s missing. The missing observations are related to values of unobserved data, and the probability of being missing varies for reasons unknown to us.

**Ex)** Participants failed to fill out depression survey because of their level of depression.

(Van Buuren, Thomas)

Using probability notation, we can get a clearer picture on the differences between MCAR, MAR, and MNAR. To lay out some notation commonly used when analyzing missing data, let \( Y = \) the \( n \times p \) matrix containing the data values on \( p \) variables for all \( n \) units in the sample. We then define a response indicator, \( R \), as an \( n \times p \) matrix where each entry is either a 0 if the corresponding entry in the \( Y \) matrix is missing, or a 1 if the \( Y \) entry is observed. Additionally, let \( Y_{\text{obs}} \) denote all observed entries, and \( Y_{\text{mis}} \) all missing entries, where \( Y = (Y_{\text{obs}}, Y_{\text{mis}}) \).

The distribution of \( R \) may depend on \( Y = (Y_{\text{obs}}, Y_{\text{mis}}) \), either by design, or by happenstance. This relationship can be described by a general expression of the missing data model, or the probability a value will be missing:

\[
P(R \mid Y_{\text{obs}}, Y_{\text{mis}}, \psi) = \frac{e^{Y_1} \psi_1 + e^{Y_2} \psi_2}{1 + e^{Y_1} \psi_1 + e^{Y_2} \psi_2}
\] (1.1)

where \( \psi \) contains the parameters for the response indicator, \( R \), in the missing data model.

**Example 1.1** Let \( Y = (Y_1, Y_2) \) be a standard bivariate normal distribution, where \( Y_1 \) and \( Y_2 \) have a correlation of 0.5 between them. Missing data can be created in \( Y_2 \) using the missing data model:

\[
P(R_2 = 0) = \psi_0 + \frac{e^{Y_1}}{1 + e^{Y_1}} \psi_1 + \frac{e^{Y_2}}{1 + e^{Y_2}} \psi_2
\] (1.2)

where \( R_2 \) is the response indicator for \( Y_2 \). We can use different parameter settings for \( \psi = (\psi_0, \psi_1, \psi_2) \), depending on the type of missing data we want to create; as long as \( \psi \) is chosen to give probabilities between (0, 1).

1. The missing data are considered **MCAR** if

\[
P(R = 0 \mid Y_{\text{obs}}, Y_{\text{mis}}, \psi) = Pr(R = 0 \mid \psi)
\] (1.3)

The probability of being missing depends only on some parameter \( \psi \). It does not depend on observed data or the missing data itself. In this example, a possible value for \( \psi_{\text{MCAR}} \) could be (0.5, 0, 0).
2. The missing data are considered **MAR** if

\[
P(R = 0 \mid Y_{obs}, Y_{mis}, \psi) = P(R = 0 \mid Y_{obs}, \psi)
\]

(1.4)

The probability of being missing depends only some observed information. It does not happen randomly. In this example, a possible value for $\psi_{MAR}$ could be $(0, 1, 0)$. $\psi_0$ will always be zero when the data are missing at random, as well as parameters originating from incomplete datasets, like $\psi_2$ in this example.

3. The missing data are considered **MNAR** if it is neither MCAR nor MAR. Therefore,

\[
P(R = 0 \mid Y_{obs}, Y_{mis}, \psi)
\]

(1.5)

does not simplify. The probability of being missing depends on unobserved information, including the missing data itself. In this example, a possible value for $\psi_{MNAR}$ could be $(0, 0, 1)$. But, because of the uncertainty in determining the cause the missing data, $\psi_0$ and $\psi_1$ could also hold non-zero values.

(Van Buuren)

### 1.3 Causes of Missing Data

Now that we understand the different mechanisms in which data can be missing, let’s try to understand why the data are missing the way they are. Missing data can either be intentionally or unintentionally missed, depending on the situation. Additionally, the data that are missing can either be isolated to individual items, or even entire units.

The item/unit nonresponse distinction says *how much* information is missing, while the distinction between intentional/unintentional says *why* some information is missing.

**Intentional:** Missing data that was planned to be missing by the data collector.

**Unintentional:** Unplanned and unforeseen. Not under the control of the data collector.

**Unit Nonresponse:** Respondent skipped one or more of the items in the survey.

**Item Nonresponse:** All outcome data is missing for this respondent.

<table>
<thead>
<tr>
<th>Table 1.1: Examples of reasons for missingness</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intentional</strong></td>
</tr>
<tr>
<td><strong>Unit Nonresponse</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Item Nonresponse</strong></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
1.4 Simple Solutions

There are simple solutions to filling in missing data that may or may not be effective. These solutions are not preferred to *multiple imputation* in any context, but may provide computational relief in specific situations.

1. **Listwise Deletion**: (aka complete case analysis) Simply removing all records (cases) from the analysis if any single value is missing is the default way of handling incomplete data in many statistical packages. In order to achieve unbiased results with this method, the data must hold the MCAR assumption.

   **Pros**: Convenient when missing data rate is small and data are MCAR.

   **Cons**: If not MCAR, listwise deletion can severely bias estimates. Even with the MCAR assumption, standard errors and significance levels are often larger relative to all available data.

2. **Pairwise Deletion**: (aka available-case analysis) Similar to listwise deletion, in that all missing entries are removed from analysis. The difference, though, is that this method doesn't remove the entire record if there is a missing entry. Instead, it calculates the means and (co)variances on all observed data, and uses these summary statistics in a statistical analysis. Again, missing data must be assumed to be MCAR to produce unbiased results.

   **Pros**: Attempts to remedy data loss problem of listwise deletion

   **Cons**: Simplicity is lost in this approach when either (a) missing data are not MCAR (b) the correlation matrix is not positive definite

3. **Mean Imputation**: A quick fix to replace the missing data is to simply replace the missing value on a certain variable by the mean of the available cases. Avoid unless there are only a handful of missing observations, and missing data are assumed to be MCAR.

   **Pros**: Maintains sample size and is easy to use.

   **Cons**: Often causes biased estimates. Underestimates standard deviations and variances.

4. **Regression Imputation**: In this method, the imputed value is predicted from a regression equation. The information in the complete observations is modeled without error terms and used to predict the values of the missing observations (Eekhout). Unlikely to provide significant results, really only works when prediction is close to perfect.

   **Pros**: Unbiased estimates of mean and regression weights under MCAR.

   **Cons**: Overestimates correlations. Underestimates (co)variances.

5. **Stochastic Imputation**: An attempt to refine regression imputation by adding the extra step of augmenting each predicted score with a residual term. The residual term is distributed normally with a mean of zero and variance equal to the residual variance from the regression of the predictor on the outcome (Eekhout). This main idea is an important step forward and will form the basis of more advanced imputation techniques.
Pros: Improves regression imputation. Preserves regression weights and correlation between variables.

Cons: Underestimates the standard error.

6. **LOCF and BOCF**: Both of the techniques, last observation carried forward (LOCF) and baseline observation carried forward (BOCF), require longitudinal data, or an experiment over a period of time. These techniques impute the missing data with the last observation of the individual, thereby assuming that they haven’t changed. Not recommended in most cases.

Pros: Generates a complete dataset.

Cons: Produces biased results, even under assumption of MCAR.

<table>
<thead>
<tr>
<th>Method</th>
<th>Unbiased Mean</th>
<th>Unbiased Reg. Weights</th>
<th>Unbiased Correlation</th>
<th>Standard Error?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listwise Deletion</td>
<td>MCAR</td>
<td>MCAR</td>
<td>MCAR</td>
<td>Too large</td>
</tr>
<tr>
<td>Pairwise Deletion</td>
<td>MCAR</td>
<td>MCAR</td>
<td>MCAR</td>
<td>Complicated</td>
</tr>
<tr>
<td>Mean Imputation</td>
<td>MCAR</td>
<td>–</td>
<td>–</td>
<td>Too small</td>
</tr>
<tr>
<td>Regression Imputation</td>
<td>MCAR</td>
<td>–</td>
<td>–</td>
<td>Too small</td>
</tr>
<tr>
<td>Stochastic Imputation</td>
<td>MAR</td>
<td>MAR</td>
<td>MAR</td>
<td>Too small</td>
</tr>
<tr>
<td>Listwise Deletion</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>Too small</td>
</tr>
</tbody>
</table>

1.5 **What About Multiple Imputation?**

Multiple imputation is a Markov Chain Monte Carlo (MCMC) technique in which the missing values are replaced by \( m > 1 \) simulated versions, where \( m \) is typically between 3-10. There are three main steps in multiple imputation, *imputation*, *analysis*, and *pooling*.

1. **Incomplete Data**: We start with a dataset consisting of ALL observations, including the missing values.
2. **Imputation:** We then create $m$ complete versions of the data where the missing values have been replaced, or imputed, by plausible data values. At the end of this step, we should have $m$ completed datasets, each with identical entries for the observed values, but (probably) different entries for the missing values. The magnitude of the differences of the imputed values reflects our uncertainty about which value to impute.

3. **Analysis:** Next, we analyze each of the individual datasets we just completed in step two, using an appropriate methodology. At the end of this step, we should have $m$ different analyses, each with different parameter estimates.

4. **Pooling:** Lastly, we pool the $m$ parameter estimates together to create one, single estimate. The pooled variance combines within-imputation (conventional) variance and between-imputation variance (extra variance caused by missing data).

### 1.6 Comparable Methods

Multiple imputation is now accepted as the best general method to deal with incomplete data, but this was not always the case. It wasn’t until the 1970’s that Donald Rubin developed the modern day multiple imputation approach for the US Census Bureau. While multiple imputation is thought of as the superior technique in constructing and analyzing missing datasets, other methods are comparable in effectiveness, but may be more difficult to execute. For example,

1. **Prevention:** The prevention of missing data is the most obvious way to improve analysis. While this may be impossible in some situations, stricter guidelines and more specific testing could help prevent the actual occurrence of missing data.

2. **Weighting procedures:** Popular method used to reduce bias in estimates of sample surveys. Used only when the probability of being selected is different between respondents. Weights are fixed, and are inversely proportional to the respondents probability of being selected.
   
   (a) **Pros:** Simple, only one set of weights needed for all incomplete variables.
   
   (b) **Cons:** Discards data by listwise regression. Can’t handle partial responses.

3. **Likelihood-based approaches:** (aka direct likelihood or full information maximum likelihood (FIML)) A likelihood-based method defines a specialized model for the observed data only, skipping over the missing data. Inferences are based on the posterior distribution under the posited model. Parameters are estimated by maximum likelihood and similar methods.

   Multiple imputation is an extension of a likelihood-based approach because it adds the extra step of imputing values for the missing entries. This is advantageous because it is easier to calculate standard errors and less information is lost overall (PSU).
GOAL The ultimate goal of multiple imputation is to estimate and evaluate specific parameters of a dataset with missing values. Of course, we would prefer these results be unbiased and representative of the overall population. Let $\theta$ represent a vector of parameters of the population of interest, and $\hat{\theta}$ represent the estimates of the parameters of $\theta$. Our goal is to find values of $\hat{\theta}$, such that

$$E(\hat{\theta}|Y) = \theta$$

(2.1)

2.1 Ignorable and Non-ignorable

In Section 1.1 we discussed the role of $\psi$, the vector of parameters that describe the response indicator, $R$. While these parameters can be used to create a model that estimates the probability of being missing, they have no intrinsic scientific interest. We are interested in the parameters $\theta$ of the overall population we’re studying. In order to make statements about $\theta$ and the population using a sample containing missing information, we must consider the joint density function of actually observed data and $R$:

$$f(y_{obs}, R|\theta, \psi)$$

(2.2)

which together depends on parameters $\theta$ for the full data $Y$ that we are interested in, and $\psi$ for the response indicator that we aren’t interested in.

The practical importance of the distinction between MCAR, MAR, and MNAR is that it clarifies the conditions in which we can accurately estimate the scientifically interesting parameters without the need to know $\psi$. So, when can we determine $\theta$ without knowing $\psi$? Or, equivalently, how can we determine $\theta$ without knowing the mechanism that created the missing data? The assumption of ignorability is essentially the belief that the available data are sufficient to correct the missing data.

The missing data mechanism, or $\psi$, is ignorable if:

1. the missing data are Missing at Random (MAR)

2. the parameters $\theta$ and $\psi$ are distinct. Essentially, $P(\psi, \theta) = P(\psi)P(\theta)$
Proving (b) can sometimes be difficult, and the MAR requirement is generally considered to be the more important condition for ignoring $\psi$.

There are crucial mathematical implications of ignoring $\psi$. If $\psi$ is ignorable, then,

$$P(Y_{mis}|Y_{obs}, R) = P(Y_{mis}|Y_{obs})$$  

(2.3)

and therefore,

$$P(Y|Y_{obs}, R = 1) = P(Y|Y_{obs}, R = 0)$$  

(2.4)

implying that the distribution of data in $Y$ is the same for both response ($R = 1$) and nonresponse ($R = 0$). Thus, we can use the posterior distribution $P(Y|Y_{obs}, R = 1)$ from the observed data and use this model to create imputations for the missing data.

If $\psi$ is not ignorable then it is non-ignorable, or

$$P(Y|Y_{obs}, R = 1) \neq P(Y|Y_{obs}, R = 0)$$  

(2.5)

There are two main strategies we might employ if the response mechanism is non-ignorable:

1. **Expand the missing data in the hope of making the missing data mechanism closer to MAR.**
   For example, let’s say there is a study on obesity that measures children’s weights over time. Because it becomes a class event, the heavier children decide not to participate. Our missing data is now MNAR, and not the desired MAR. To combat this, we could record their waist circumference instead, and hopefully get more participation, and therefore data that are MAR.

2. **Formulate and fit a non-ignorable imputation model and perform a sensitivity analysis on the critical parameters.** If ignorability does not hold, we need to model the distribution $P(Y,R)$ instead of $P(Y)$. Different strategies to accomplish this include the selection model and the pattern-mixture model, which won’t be discussed here.
Chapter 2. Introduction to Multiple Imputation

<table>
<thead>
<tr>
<th>Frequentist</th>
<th>Bayesian</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Data are a repeatable random sample - there is a frequency</td>
<td>• Data are observed from the realized sample</td>
</tr>
<tr>
<td>• Underlying parameters remain constant during this repeatable process</td>
<td>• Parameters are unknown and described probabilistically</td>
</tr>
<tr>
<td>• Parameters are fixed</td>
<td>• Data are fixed</td>
</tr>
<tr>
<td>• No information prior to the model specification</td>
<td>• Prior information is utilized in the form of prior distributions $p(\theta)$</td>
</tr>
</tbody>
</table>

Bayesians treat unobserved data and unknown parameters in similar ways. They describe each with a probability distribution. For their model, Bayesians specify:

• A *joint density function*, which describes the form of the distribution of the full sample of data (given the parameter values).

• A *prior distribution*, which describes the behavior of the parameter(s) unconditional on the data. The prior could reflect uncertainty about a parameter, or the variety of values that the parameter could take.

A fundamental idea in Bayesian statistics is conditional probability. Let us denote separate events by the letters $A, B, C, \ldots$. Recall the probability of event $A$, *given* event $B$.

$$ P(A|B) = \frac{P(A, B)}{P(B)} \quad (2.6) $$

Additionally, the primary originator of Bayesian statistics, Thomas Bayes, has his own formula for understanding conditional probabilities. In its simplest form, with two events $A$ and $B$, Bayes’ Law relates the conditional probabilities $P(A|B)$ and $P(B|A)$.

$$ P(A|B) = \frac{P(A, B)}{P(B)} \quad \text{or} \quad P(B|A) = \frac{P(B, A)}{P(A)} $$

Therefore,

$$ P(A, B) = P(A|B)P(B) = P(B|A)P(A) $$

and

$$ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{similarly} \quad P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad (2.7) $$

**Definition 2.1** In the future, we will deal with both discrete and continuous random variables. In general, let $p(\cdot)$ denote the probability distribution (or probability density function (pdf), or probability mass function (pmf)) of a random variable. If $X$ and $Y$ are random variables, then $p(X)$ is the *marginal distribution* of $X$ and $p(X, Y)$ is the *joint distribution* of $X$ and $Y$. If $X$ and $Y$ are independent, then $p(X, Y) = p(X)p(Y)$. 
2.2. Introduction to Bayesian Analysis

Recall that the expected value of any function \( h(X) \) of \( X \) is:

\[
E[h(X)] = \begin{cases} 
\sum_{x \in \chi} h(x)p(x) & X \text{ is discrete} \\
\int_{\chi} h(x)p(x) & X \text{ is continuous}
\end{cases}
\]

where \( \chi \) denotes the space of all possible values of the random variable. Typically the distribution of \( X \) depends on some parameter(s), \( \theta \), so, in fact \( p(X) = p(X|\theta) \) (Hitchcock).

In order to make probability statements about \( \theta \) given \( y \), we must start with a model to describe a joint probability distribution for \( \theta \) and \( y \). This model can be written as:

\[
p(\theta, y) = p(y|\theta)p(\theta)
\]

where the joint probability distribution \( p(\theta, y) \) is equal to the product of the sampling distribution (or data distribution) \( p(y|\theta) \) and prior distribution \( p(\theta) \).

Simply conditioning on the known value of the data \( y \), using Bayes’ Law mentioned earlier, yields the posterior distribution:

\[
p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}
\]

where

\[
p(y) = \begin{cases} 
\sum_{\theta} p(\theta)p(y|\theta) & X \text{ is discrete} \\
\int_{\theta} p(\theta)p(y|\theta)d\theta & X \text{ is continuous}
\end{cases}
\]

and \( p(y) \) is often called the marginal distribution of \( y \), but a more informative name is the prior predictive distribution. Prior, because it is not conditional on a previous observation of the process, and predictive because it is the distribution for a quantity that is observable. After the data \( y \) have been observed, we can predict an unknown observable, say \( \tilde{y} \), from a similar process (Gelman).

Another interesting distribution is called the posterior predictive distribution. Posterior because it is conditional on the observed \( y \) and predictive because it is a prediction for an observable \( \tilde{y} \):

\[
p(\tilde{y}|y) = \int_{\theta} p(\tilde{y}, \theta|y)d\theta = \int_{\theta} p(\tilde{y}|\theta, y)p(\theta|y)d\theta
\]

For a missing data model we obtain the posterior \( P(\theta, Y_{\text{mis}}|Y_{\text{obs}}) \), but inference is made based on the marginal posterior distribution of parameter of interest, known as the observed-data posterior, which in typical Bayesian form can be expressed as

\[
P(\theta|Y_{\text{obs}}) \propto P(Y_{\text{obs}}|\theta)p(\theta)
\]

and inference about \( \theta \) is summarised by the posterior distribution, e.g. mean and quantiles.

Remark 2.1. Not to be confused with the complete-data posterior distribution \( P(\theta|Y) \) used for (posterior) complete-case analysis.

We note that

\[
P(\theta, Y_{\text{mis}}|Y_{\text{obs}}) = P(\theta|Y)p(Y_{\text{mis}}|Y_{\text{obs}})
\]
and integrating over $Y_{mis}$ yields

$$P(\theta|Y_{obs}) = \int P(\theta|Y)P(Y_{mis}|Y_{obs})dY_{mis} \quad (2.8)$$

Also note that

$$P(\theta, Y_{mis}) = P(Y_{mis}|\theta, Y_{obs})P(\theta|Y_{obs})$$

and integrating over $\theta$ we get

$$P(Y_{mis}|Y_{obs}) = \int_{\theta} P(Y_{mis}|\theta, Y_{obs})P(\theta|Y_{obs})d\theta \quad (2.9)$$

### 2.3 The Gibbs Sampler

A technique used to construct and sample from arbitrary posterior distributions is the Gibbs sampler. This technique utilizes Markov chain Monte Carlo (MCMC) methods to draw values of the parameter vector $\theta$ from approximate distributions, and then correcting those draws to better represent the target posterior distribution, $p(\theta|y)$. The sampling is done sequentially, with a distribution of the sampled draws depending on the last value drawn. These draws then form a Markov chain (Gelman).

**Definition 2.2** A Markov chain is a sequence of random variables $\theta_1, \theta_2, \ldots, \theta_t$ which, for any $t$, the distribution of $\theta_t$, given all previous $\theta$’s depends only on the most recent value, $\theta_{t-1}$. The set in which the $\theta_t$ take values is called the state space of the Markov chain.

The underlying logic of MCMC sampling is that we can estimate any desired expectation by ergodic averages. That is, we can compute any statistic of a posterior distribution as long as we have $m$ simulated samples from that distribution:

$$E[f(\theta)]_P \approx \frac{1}{N} \sum_{i=1}^{N} f(\theta^{(i)}) \quad (2.10)$$

where $P$ is the posterior distribution of interest, $f(\theta)$ is the desired expectation, and $f(\theta^{(i)})$ is the $i^{th}$ simulated sample from $P$. This is the Monte Carlo estimate of $f(\theta)$, which by the Strong Law of Large Numbers (SLLN) (Walsh)

$$\frac{1}{N} \sum_{i=1}^{N} f(\theta^{(i)}) \xrightarrow{w.p.1} E[f(\theta)]_P$$

if $\{\theta^{(i)}\}$ are i.i.d. so a sufficient lag will be necessary to remove dependency from previous value.

Likewise, the Monte Carlo estimate for any function of $n$ variables $(\theta_1, \ldots, \theta_n)$ is given by

$$E[f(\theta_1, \ldots, \theta_n)]_P \approx \frac{1}{N} \sum_{i=1}^{N} f(\theta_1^{(i)}, \ldots, \theta_n^{(i)}) \quad (2.11)$$

So how do we generate these samples? Let us consider an example.
Example 2.1 Consider the random variables \( X \) and \( Y \) and we are interested in the joint distribution \( p(x,y) \) that might be hard to sample from directly but it is easy to sample from \( p(x|y) \) and from \( p(y|x) \). The Gibbs sampler algorithm works by:

1. Set some initial values \((x^{(0)}, y^{(0)})\).
2. Generate \( x^{(1)} \sim p(x|y^{(0)}) \). Current state: \((x^{(1)}, y^{(0)})\)
3. Generate \( y^{(1)} \sim p(y|x^{(1)}) \). Current state: \((x^{(1)}, y^{(1)})\)
4. Generate \( x^{(2)} \sim p(x|y^{(1)}) \). Current state: \((x^{(2)}, y^{(1)})\)

\[ \vdots \]

Continue \( N \) times. This procedure defines a sequence of pairs of random variables \([X^{(i)}, Y^{(i)}]\) that is a Markov Chain. The conditional distribution of \([X^{(i)}, Y^{(i)}]\) given all of the previous pairs depends only on \([X^{(i-1)}, Y^{(i-1)}]\).

We are not directly sampling from the posterior distribution itself. Rather, we simulate samples by sweeping through all the posterior conditionals, one random variable at a time. To obtain the desired total of \( N \) sample points, one must sample the chain

(i) after a sufficient burn-in period to remove the effects of the initial values,

(ii) at set time points (say every \( n \) samples) following the burn-in, i.e. lag.

We will go into detail about these requirements later, but the final point is that the Gibbs sequence converges to a stationary distribution that is independent of the starting values, and by construction this stationary distribution is the target distribution we are trying to simulate (Walsh).

2.4 Generating Imputations

Our goal is to generate data from the target distribution the observed data posterior \( p(\theta|Y_{\text{obs}}) \). So for (2.8) and (2.9) we need to generate from the target distribution \( P(\theta, Y_{\text{mis}}|Y_{\text{obs}}) \). The chain is created for \( t = 1, 2, \ldots \)

- Start with initial value \( \theta^{(0)} \)
- \( Y_{\text{mis}}^{(t+1)} \sim P(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(t)}) \)
- \( \theta^{(t+1)} \sim P(\theta|Y_{\text{obs}}, Y_{\text{mis}}^{(t+1)}) \)

From the sequence we have

- pairs \( \{(\theta^{(t)}, Y_{\text{mis}}^{(t)})|Y_{\text{obs}}\} \) that is Markov Chain whose stationary distribution is \( P(\theta, Y_{\text{mis}}|Y_{\text{obs}}) \)
- sequence \( \{\theta^{(t)}|Y_{\text{obs}}\} \) with stationary distribution \( P(\theta|Y_{\text{obs}}) \)
2.5 Estimation

There is uncertainty in the true value of $\theta$ because some of the population data are missing. The possible values of $\theta$, given our knowledge of the data in $Y_{obs}$, are captured by the posterior distribution $P(\theta|Y_{obs})$ in (2.8).

To find the expected value of $\theta$ given $Y_{obs}$,

$$E(\theta|Y_{obs}) = E[E(\theta|Y)|Y_{obs}]$$

which is the average of the posterior means of $\theta$ over the repeatedly imputed data. This equation is special case of (2.10) and suggests a specific procedure for combining the results of repeated imputations. We can find the combined estimate, $\bar{\theta}$, by using the first of Rubin’s Rules:

$$\bar{\theta} = \frac{1}{m} \sum_{\ell=1}^{m} \hat{\theta}_\ell$$

where $m$ is the total number of imputations and $\hat{\theta}_\ell$ is an estimate of the $\ell^{th}$ repeated imputation. $\hat{\theta}_\ell$ contains $k$ parameters and is represented as a $k \times 1$ column vector.

2.6 Variation and Rubin’s Rules

The posterior variance of $\theta$ given $Y_{obs}$ is the sum of two variance components:

$$V(\theta|Y_{obs}) = E[V(\theta|Y)|Y_{obs}] + V[E(\theta|Y)|Y_{obs}]$$

where the first component is the average of the repeated complete data posterior variances of $\theta$, also known as the within-variance, or $\bar{U}$. The second component is the variance between the complete data posterior means of $\theta$, also known as the between-variance, or $B$.

We can calculate $\bar{U}$ by finding the average of the complete data variances, where $U_\ell$ is the variance-covariance matrix of $\hat{\theta}_\ell$ obtained for the $\ell^{th}$ imputation:

$$\bar{U} = \frac{1}{m} \sum_{\ell=1}^{m} U_\ell$$

We can calculate $B$ using a standard unbiased estimate of the variance between the $m$ complete data estimates:

$$B = \frac{1}{m-1} \sum_{\ell=1}^{m} (\hat{\theta}_\ell - \bar{\theta})(\hat{\theta}_\ell - \bar{\theta})'$$

To find the total variance, $T$, of $\theta$, we must combine $\bar{U}$, $B$, and a finite simulation error, coalescing into the second of Rubin’s Rules:

$$T = \bar{U} + B + \frac{B}{m}$$

To summarize, the total variance, $T$, stems from three sources:

1. $\bar{U} \rightarrow$ Variance caused by the fact this is a sample from a population rather than the entire population being observed.
2.7 Variance Ratios

2. \( B \rightarrow \) Extra variance caused by missing values.

3. \( \frac{B}{m} \rightarrow \) Extra finite simulation variance, because it is an average

<table>
<thead>
<tr>
<th>Incomplete Sample, ( Y_{obs} )</th>
<th>Complete Sample, ( Y = (Y_{obs}, Y_{mis}) )</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\theta} \rightarrow )</td>
<td>( \hat{\theta} \rightarrow )</td>
<td>( \theta )</td>
</tr>
<tr>
<td>( \hat{U} \rightarrow )</td>
<td>( U = V(\hat{\theta}) )</td>
<td></td>
</tr>
<tr>
<td>( B = V(\hat{\theta}) )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

where “\( \rightarrow \)” means “is an estimate of”

2.7 Variance Ratios

- **Proportion of the variation attributable to the missing data**

  \[ \lambda = \frac{B + \frac{B}{m}}{T} \]

  - \( \lambda = 0 \) if the missing data don’t add extra variation to sampling variance.
  - \( \lambda = 1 \) if all variation is caused by the missing data.
  - \( \lambda > 0.5 \) if the influence of the imputation model on the final result is larger than that of the complete data model.

- **Relative increase in variance due to nonresponse**

  \[ r = \frac{B + \frac{B}{m}}{U} \]

  or alternatively, \( r = \frac{\lambda}{1 - \lambda} \)

- **Estimated rate of missing information**

  \[ \gamma = \frac{r + 2}{\nu + 3} \left( \frac{1}{1 + r} \right) \]

  where \( \nu \) is an estimate of the degrees of freedom given in (2.19).

2.8 Degrees of Freedom

The calculation of the degrees of freedom for statistical tests needs some attention because part of the data is missing. The “old” formula, written by Rubin in 1987, for the degrees of freedom is:

\[ \nu_{old} = (m - 1) \left( 1 + \frac{1}{r^2} \right) = \frac{m - 1}{\lambda^2} \quad (2.18) \]

In 1999, Rubin noted that this equation can produce values that are larger than the degrees of freedom in the complete data. The new formula incorporates the old formula, along
with a error correction calculated from the observed data. Let $v_{com}$ be the degrees of freedom of $\bar{\theta}$ in the hypothetically complete data, where $n =$ sample size and $k =$ number of parameters.

$$v_{com} = n - k$$

The estimated observed data degrees of freedom that accounts for the missing information is:

$$v_{obs} = \frac{v_{com} + 1}{v_{com} + 3} v_{com}(1 - \lambda)$$

Finally, the adjusted degrees of freedom to be used for testing in multiple imputation is:

$$v = \frac{v_{old} v_{obs}}{v_{old} + v_{obs}} \quad (2.19)$$

### 2.9 Scalar Inferences

Single parameter inference applies if $k = 1$, or if $k > 1$ and the test is repeated for each of the $k$ components. Univariate tests are based on the approximation (assuming enough imputations are computed, or data are normal):

$$\frac{\bar{\theta} - \theta}{\sqrt{T}} \sim t_v \quad (2.20)$$

where $t_v$ is the Student’s t-distribution with $v$ degrees of freedom as shown in (2.19), and $T$ is the variance as shown in (2.17). Consequently, the 100(1 $- a$)% confidence interval of a $\bar{\theta}$ is:

$$\bar{\theta} \pm t_{1-a/2,v} \sqrt{T} \quad (2.21)$$

Suppose we want to test the null hypothesis $\theta = \theta_0$ for some specified value $\theta_0$. We can calculate the p-value of the test as the probability:

$$p\text{-value} = P \left[ F_{1,v} \geq \frac{(\theta_0 - \bar{\theta})^2}{T} \right] \quad (2.22)$$

where $F_{1,v}$ is the F-distribution with 1 numerator degree of freedom and $v$ denominator degrees of freedom.

### 2.10 How Many Imputations?

Research has shown that only a few imputations are needed to adequately reduce simulation error and get analyzable results. The efficiency, essentially a measure of quality that takes the form of a standardized value from 0-100 in this context, of an estimate based on $m$ imputations is approximately

$$\left( 1 + \frac{\gamma}{m} \right)^{-1} \quad (2.23)$$

where $\gamma$ is the rate of missing information for the quantity being estimated. The efficiencies achieved for various values of $m$ and rates of missing information are:
2.10. How Many Imputations?

Unless the rate of missing information is very high, there is little advantage to producing and analyzing more than a few imputed datasets.

<table>
<thead>
<tr>
<th>$m$</th>
<th>$\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td>3</td>
<td>97</td>
</tr>
<tr>
<td>5</td>
<td>98</td>
</tr>
<tr>
<td>10</td>
<td>99</td>
</tr>
<tr>
<td>20</td>
<td>100</td>
</tr>
</tbody>
</table>
3. Generating Imputed Datasets

3.1 How we got here

To explain multiple imputation, we start with an example. Suppose we have a univariate dataset with some missing values. Let’s assume these missing values are proven to be missing at random (MAR). To aid in our analysis of the data, we decide to impute these missing values.

**Notation** For univariate $Y$ we write lowercase $y$ for $Y$. Any predictors in the imputation model are collected in $X$. Symbol $X_{obs}$ indicates the subset of $n_1$ rows of $X$ for which $y$ is observed, and $X_{mis}$ is the complementing subset of $n_0$ rows of $X$ for which $y$ is missing. The vector containing the $n_1$ observed data in $y$ is denoted by $y_{obs}$, and the vector of $n_0$ imputed values in $y$ is indicated by $\hat{y}$.

To replace the missing values, we could simply calculate the regression line using the observed data, take the expected value for the associated explanatory variable, and then impute that value into our dataset with the missing values.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_{mis}$$  \hfill (3.1)

Although this may be the “best” value, in that it is the most likely under this regression model, it does not reflect the uncertainty in the actual outcome. To combat this, we could add an error term to our linear regression model. One that is normally distributed, with a mean of 0 and a standard deviation of $\hat{\sigma}$. By drawing a random value, we can more accurately represent the fact that we cannot perfectly predict the future.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_{mis} + \hat{\epsilon} \quad \hat{\epsilon} \sim N(0, \hat{\sigma}^2)$$  \hfill (3.2)

While a using regression model equipped with an error term to impute missing values is logical, it neglects another important uncertainty. The true value of the parameters of our regression model are typically unknown, and cannot be calculated with certainty from a dataset with missing values. This parameter uncertainty needs to be included in the imputations. Bayesian methods are the most popular, and our best bet, at finding parameter values that fit our contextual needs.
Let $y_1, \ldots, y_k$ be the variables containing missing data entries and $x_1, \ldots, x_q$ the completely observed variables in a sample.

Starting with an initial imputation $Y_{mis}^{(0)}$, a sequence of imputations $Y_{mis}^{(1)}, \ldots, Y_{mis}^{(N)}$ is generated by successively generating the imputations $Y_{mis}^{(i)}$ conditional on the observed data and the most recently imputed data of $Y_{mis}(j)$, $j \neq i$. In particular, $Y_{mis}^{(t)}(j)$ is generated from the predictive distribution

$$P(Y_{mis}(j) | Y_1^{(t)}, \ldots, Y_{j-1}^{(t)}, Y_{obs}(j), Y_{j+1}^{(t-1)} \ldots, Y_k^{(t-1)}, X_1, \ldots, X_q)$$

In this predictive distribution, $Y_1^{(t)}, \ldots, Y_{j-1}^{(t)}$ represent the completed data for $y_1, \ldots, y_{j-1}$ in the current iteration. $Y_{obs}(j)$ represents the observed data for $y_j, Y_{j+1}^{(t-1)} \ldots, Y_k^{(t-1)}$ represent the completed data for $y_{j+1}, \ldots, y_k$ in the previous iteration. $X_1, \ldots, X_q$ represent the completely observed data for $x_1, \ldots, x_q$.

When a regression model of $y_i$ on $y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_k, x_1, \ldots, x_q$ is specified and its parameters represented by $\phi_j$ are known, the predictive distribution of $Y_{mis}^{(t)}(j)$ can be defined. However, $\phi_j$ can only be estimated from complete data. To reflect the uncertainty about $\phi_j$ given the complete data, a value for $\phi_j$ is drawn from an appropriate posterior distribution for $\phi_j$ conditional on the most recently completed data, and this value is used to generate $Y_{mis}^{(t)}(j)$. Iteration $t$ of the Gibbs sampler for generating $Y_{mis}^{(t)}$ from $Y_{mis}^{(t-1)}$ is then given by:

$$\phi_1^{(t)} \sim P(\phi_1 | [Y_1^{(t-1)}, \ldots, Y_k^{(t-1)}, X_1, \ldots, X_q]_{obs(1)})$$

$$Y_{mis}^{(1)}(1) \sim P(Y_{mis}(1) | Y_2^{(t-1)}, \ldots, Y_k^{(t-1)}, X_1, \ldots, X_q; \phi_1^{(t)})$$

$$\phi_2^{(t)} \sim P(\phi_2 | [Y_1^{(t)}, Y_2^{(t-1)}, \ldots, Y_k^{(t-1)}, X_1, \ldots, X_q]_{obs(2)})$$

$$Y_{mis}^{(2)}(2) \sim P(Y_{mis}(2) | Y_1^{(t)}, Y_3^{(t-1)}, \ldots, Y_k^{(t-1)}, X_1, \ldots, X_q; \phi_2^{(t)})$$

$$\vdots$$

$$\phi_j^{(t)} \sim P(\phi_j | [Y_1^{(t)}, \ldots, Y_{j-1}^{(t)}, Y_j^{(t)}, \ldots, Y_k^{(t-1)}, X_1, \ldots, X_q]_{obs(j)})$$

$$Y_{mis}^{(j)}(j) \sim P(Y_{mis}(j) | Y_1^{(t)}, \ldots, Y_{j-1}^{(t)}, Y_j^{(t-1)}, \ldots, Y_k^{(t-1)}, X_1, \ldots, X_q; \phi_j^{(t)})$$

$$\vdots$$

$$\phi_k^{(t)} \sim P(\phi_k | [Y_1^{(t)}, \ldots, Y_{k-1}^{(t)}, Y_k^{(t-1)}, X_1, \ldots, X_q]_{obs(k)})$$

$$Y_{mis}^{(t)}(k) \sim P(Y_{mis}(k) | Y_1^{(t)}, \ldots, Y_{k-1}^{(t)}, Y_k^{(t-1)}, X_1, \ldots, X_q; \phi_k^{(t)})$$

**Example 3.1** In this example, we will be using a dataset containing information from a study concerning the effects of the drug Dobutamine. Inspection of the data reveals missing information. For our analysis, we will determine which variables are good indicators of surgery candidates for horses with colic. We will use the variables resting systolic blood pressure ($rs$), resting diastolic blood pressure ($rd$), peak effort systolic blood pressure ($ps$), and peak effort diastolic blood pressure ($pd$). We see that variables contain 100, 99, and 173, and 0 missing values, respectively.
Chapter 3. Generating Imputed Datasets

Let’s set up a missing data model so that we can generate imputations for the missing values. For simplicity, we will assume that all the variables are linearly related.

\[
rs = \beta_{10} + \beta_{11}rd + \beta_{12}ps + \beta_{13}pd + \epsilon_1 ; \quad \epsilon_1 \sim N(0, \sigma_1^2)
\]

\[
rd = \beta_{20} + \beta_{21}rs + \beta_{22}ps + \beta_{23}pd + \epsilon_2 ; \quad \epsilon_2 \sim N(0, \sigma_2^2)
\]

\[
ps = \beta_{30} + \beta_{31}rs + \beta_{32}rd + \beta_{33}pd + \epsilon_3 ; \quad \epsilon_3 \sim N(0, \sigma_3^2)
\]

Reordering the variables according to their fractions of missing data results in the sequence \(rd, rs, ps\). We can simplify the model further by only including complete or completely imputed variables.

\[
rd = \beta_{20} + \beta_{23}pd + \epsilon_2 ; \quad \epsilon_2 \sim N(0, \sigma_2^2)
\]

\[
rs = \beta_{10} + \beta_{11}rd + \beta_{13}pd + \epsilon_1 ; \quad \epsilon_1 \sim N(0, \sigma_1^2)
\]

\[
ps = \beta_{30} + \beta_{31}rs + \beta_{32}rd + \beta_{33}pd + \epsilon_3 ; \quad \epsilon_3 \sim N(0, \sigma_3^2)
\]

\(\epsilon_1, \epsilon_2, \text{ and } \epsilon_3\) are normally distributed error terms, each of which is independent of the explanatory variables in the corresponding regression model.

Applying the earlier equation to generate imputations, and setting \(y_1 = rs, y_2 = rd, y_3 = ps, k = 3, x_1 = pd, q = 1, \phi_i = (\beta_{i0}, \beta_{i1}, \beta_{i2}, \beta_{i3}, \sigma_i), \text{ and } i = 1, \ldots, 3\). Iteration \(t\) the Gibbs sampler is given by

**STEP 1**

1. draw \(\beta_{20}^{(0)}, \beta_{23}^{(0)}, \sigma_2^{(0)}\) from a posterior distribution of \(\beta_{20}, \beta_{23}, \sigma_2\), conditional on the observed data for \(rd\) and \(pd\) in cases where these two variables are simultaaneously observed.

2. impute each missing \(rd_i\) by \(rd_i^{(0)} = \beta_{20}^{(0)} + \beta_{23}^{(0)}pd_i + \epsilon_{2i}^{(0)}\), with each error term \(\epsilon_{2i}^{(0)}\) independently drawn from \(N(0, \sigma_2^{(0)}2)\).

**STEP 2**

1. draw \(\beta_{10}^{(0)}, \beta_{11}^{(0)}, \beta_{13}^{(0)}, \sigma_1^{(0)}\) from a posterior distribution of \(\beta_{10}, \beta_{11}, \beta_{13}, \sigma_1\), conditional on the completed data for \(rd\) in step 1 and the observed data for \(rs\) and \(pd\), restricted to the cases for which \(rs\) is observed.

2. impute each missing \(rs_i\) by \(rs_i^{(0)} = \beta_{10}^{(0)} + \beta_{11}^{(0)}rd_i^{(0)} + \beta_{13}^{(0)}pd_i + \epsilon_{1i}^{(0)}\), with each error term \(\epsilon_{1i}^{(0)}\) independently drawn from \(N(0, \sigma_1^{(0)2})\).

**STEP 3**

1. draw \(\beta_{30}^{(0)}, \beta_{31}^{(0)}, \beta_{32}^{(0)}, \beta_{33}^{(0)}, \sigma_3^{(0)}\) from a posterior distribution of \(\beta_{30}, \beta_{31}, \beta_{32}, \beta_{33}, \sigma_3\), conditional on the completed data for \(rd\) and \(rs\) in steps 1 and 2, and the observed data for \(ps\) and \(pd\), restricted to the cases for which \(ps\) is observed.

2. impute each missing \(ps_i\) by \(ps_i^{(0)} = \beta_{30}^{(0)} + \beta_{31}^{(0)}rs_i^{(0)} + \beta_{32}^{(0)}rd_i^{(0)} + \beta_{33}^{(0)}pd_i + \epsilon_{3i}^{(0)}\), with each error term \(\epsilon_{3i}^{(0)}\) independently drawn from \(N(0, \sigma_3^{(0)2})\).
3.2 Multiple imputation under a normal linear model

The substeps in (a) correspond to the posterior draws of \( \phi_j^{(t)} \) and the substeps in (b) correspond to the draws of \( Y_{mis}^{(t)}(j) \) (Brand).

3.2 Multiple imputation under a normal linear model

Multiple imputation algorithm under the normal linear model.

1. Calculate the cross-product matrix \( S = X_{obs}'X_{obs} \)
2. Calculate \( V = (S + diag(S)\kappa)^{-1} \), with some small \( \kappa \)
3. Calculate regression weights from observed data \( \hat{\beta} = VX_{obs}'y_{obs} \)
4. Draw a random variable \( \hat{g} \sim \chi^2_v \) with \( v = n_{obs} - q \)
5. Calculate specific variance \( \hat{\sigma}^2 = (y_{obs} - X_{obs}\hat{\beta})'(y_{obs} - X_{obs}\hat{\beta})/\hat{g} \)
6. Calculate \( V^{1/2} \) by Cholesky Decomposition
7. Draw \( q \) independent \( N(0,1) \) variates into a vector \( \hat{z}_1 \)
8. Calculate \( \hat{\beta} = \hat{\beta} + \hat{\sigma}\hat{z}_1 V^{1/2} \)
9. Draw \( n_0 \) independent \( N(0,1) \) variates in vector \( z_2 \)
10. Calculate and impute missing \( n_0 \) values \( \hat{y} = X_{mis}\hat{\beta} + z_2\hat{\sigma} \)