

---

# Bayes and Multiple Imputation

ESC 2024 summer Session 6주차  
김상민, 안유빈, 오경주, 정슬아



# Contents

---

## 1. Bayesian

1.0 Bayesian vs Frequentist

1.1 Bayesian의 한계

1.2 MCMC

## 2. Bayesian Iterative Simulation Methods

2.0 Introduction

2.1 Data Augmentation (10.1.1)

2.2 The Gibbs' Sampler (10.1.2)

2.3 Assessing Convergence of Iterative Simulations (10.1.3)

2.4 Some Other Simulation Methods (10.1.4)

## 3. Multiple Imputation

3.0 Multiple Imputation Tutorial

3.1 Large-Sample Bayesian Approximations (10.2.1)

3.2 Other Methods for Creating Multiple Imputations (10.2.3)

3.3 Chained-Equation Multiple Imputation (10.2.4)

# 1

## Bayesian

---

1.0 Bayesian vs Frequentist

1.1 Bayesian의 한계

1.2 MCMC

# 1.0 bayesian vs frequentist

---

## 베이지안

모수를 확률변수로 본다  
→ 불확실성의 정량화가 가능해진다

사전지식+데이터로 모수를 추정한다.  
→ 데이터가 적을때 사전지식이 있으면 더좋다.

## 빈도주의자

모수를 미지의 상수(unknown const)로 본다

데이터로만 모수를 추정한다

# 1.0 bayesian vs frequentist

## <빈도주의자의 추정방법>

1.  $\theta$ 는 unknown 상수, 값을 알아내는것이 우리의 목표

2.  $X$ 의 pdf를  $f(X|\theta)$ 로 가정 (우리가 가진 데이터는  $X$ 확률변수)

3.  $\hat{\theta} = T(X_1, \dots, X_n)$  통계량으로  $\theta$ 의 점추정량을 구하기

: 여기서 여러가지 통계량이 나올 수 있음, 사실 확률 표본들( $X_1, \dots, X_n$ )의 함수는 모두다 통계량이자 추정량

: but 좋은 통계량(추정량)이 있음, 여러가지 평가요소(불편성, 효율성), 대표적으로 좋은 추정량 MLE

4. 점추정이 끝x, 통계량으로 구간추정과 가설검정을 진행

: 이때 통계량(추정량)은 확률변수이기 때문에 특정한 분포를 가짐, 이를 통해 구간추정 및 가설검정

: but 우리는 통계량의 정확한 분포를 모르기에 극한분포를 구함(asymptotic property), 대표적으로 CLT

## 베이지안

모수를 확률변수로  
둔다.

사전지식+데이터로  
모수를 추정한다.

## 빈도주의자

모수를 미지의 상수  
로 둔다

데이터로만 모수를  
추정한다

# 1.0 bayesian vs frequentist

<베이지안의 추정방법>

1.  $\theta$ 는 확률변수,  $\theta|X$  의 분포를 알아내는 것이 우리의 목표

2  $X$ 의 pdf를  $f(X|\theta)$ 로 가정 (우리가 가진 데이터는  $X$  확률변수)

+  $\theta$ 의 분포를 가정(사전분포)

3. 데이터가 주어졌을 때의 세타의 분포를 구하기(사후분포)

$$P(\theta | \text{data}) = \frac{P(\text{data}|\theta) \cdot P(\theta)}{P(\text{data})}$$

$$P(\theta | x_1, x_2, \dots, x_n) = \frac{\prod_{i=1}^n P(x_i | \theta) \cdot P(\theta)}{\int \prod_{i=1}^n P(x_i | \theta) \cdot P(\theta) d\theta}$$

4. 모수의 분포(사후분포)로 점추정, 가설검정, 신뢰구간 구하기 (쉬움)

: 점추정은 MAP, 신뢰구간은 확률적으로 배정

5. 데이터 주어질 때마다 사후분포를 사전분포로 두고 업데이트 진행

베이지안

빈도주의자

모수를 확률변수로  
둔다.

사전지식+데이터로  
모수를 추정한다.

모수를 미지의 상수  
로 둔다

데이터로만 모수를  
추정한다

# 1.0 bayesian vs frequentist

## 1. 모수를 확률변수/상수로 볼때의 차이점

- 모수가 확률변수이면 모수에 대한 불확실성을 정량화 할 수 있다.
- 모수에 대한 불확실성을 정량화 한다는 것  
= 불확실성을 확률적으로 설명할 수 있다

- confidence intervals (빈도주의적 신뢰구간)

ex. 신뢰구간  $[a, b]$  95%

같은 추정방법으로 100개의 신뢰구간을 구하면 95개정도는 모수를 포함할 수 있다. 그 100개 중의 하나가  $[a, b]$

- credible interval(베이지안적 신뢰구간)

ex. 신뢰구간  $[a, b]$

모수가 95% 확률로 이구간안에 있다.

베이지안	빈도주의자
모수를 확률변수로 둔다.	모수를 미지의 상수로 둔다
사전지식+데이터로 모수를 추정한다.	데이터로만 모수를 추정한다

# 1.0 bayesian vs frequentist

2. 사전지식+데이터/데이터로만 모수를 추정

베이지안

빈도주의자

모수를 확률변수로  
둔다.

모수를 미지의 상수  
로 둔다

사전지식+데이터로  
모수를 추정한다.

데이터로만 모수를  
추정한다

베이지안의 추정량(MAP)

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} (P(\text{data} \mid \theta) \cdot P(\theta))$$

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} \left( \prod_{i=1}^n P(x_i \mid \theta) \cdot P(\theta) \right)$$

빈도주의자의 추정량(MLE)

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} P(\text{data} \mid \theta)$$

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} \prod_{i=1}^n P(x_i \mid \theta)$$

## 1.1 bayesian의 한계

---

<베이지안의 한계>

$$P(\theta \mid x_1, x_2, \dots, x_n) = \frac{\prod_{i=1}^n P(x_i \mid \theta) \cdot P(\theta)}{\int \prod_{i=1}^n P(x_i \mid \theta) \cdot P(\theta) d\theta}$$

- 1. 분모의 적분부분을 구하기 어렵다
- 어떻게 해결할 것인가?
- 적분부분은 결국 상수일뿐 (정규화를 시켜줄) 결국 분자에 있는 함수만 생각하면됨 , 얘는 정확히 무슨함수인지 알고 있다.
- 이 함수에서 표본을 뽑아 히스토그램을 그려서 이것을 통해 사후분포의 개형을 알아냄
  
- but 복잡한 함수에서 sampling 하는것도 생각보다 어려움
- 이를 위한 sampling 기법이 MCMC
  
- 2. 사전분포 설정의 어려움
- 굉장히 주관적

## 1.2 MCMC

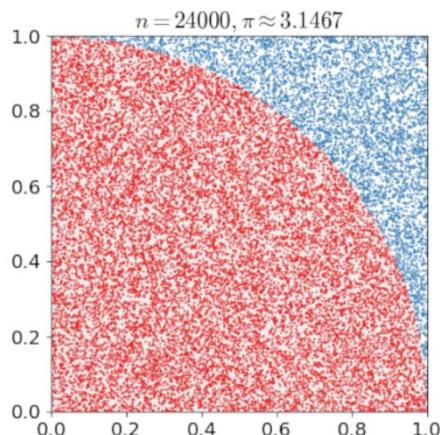
---

### <MCMC>

- 복잡한 확률분포로부터의 SAMPLING 알고리즘 (베이지안에서만 쓰이는것이 아님)
- MCMC= 마르코프체인Markov Chain + 몬테칼로Monte Carlo
- 대표적인 MCMC알고리즘 : 메트로폴리스 헤이스팅스 알고리즘, 갑스샘플러

### <몬테칼로 알고리즘>

- 수학적인 결과를 얻기 위해 반복적으로 무작위 샘플링 방법을 이용하는 컴퓨팅 알고리즘



## 1.2 MCMC

---

### <Markov chain>

- 마르코프체인의 정의: '마르코프 성질'을 가진 '이산 확률 과정'
- 마르코프성질의 정의: '과거와 현재 상태가 주어졌을 때, 미래 상태의 조건부 확률분포가 과거 상태에 영향을 받지 않고 독립적으로 현재 상태로만 결정되는 것을 의미'
- 확률과정의 정의: 시간t에 따라 변화하는 확률변수들의 집합

$$p(\mathbf{z}^{(m+1)} | \mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m)}) = p(\mathbf{z}^{(m+1)} | \mathbf{z}^{(m)}).$$

## 1.2 MCMC

---

<MCMC 샘플링 알고리즘>

1. **목표**: 복잡한 함수  $f(x_1, \dots, x_n)$ 에서 sampling하는 것이 목표
2. **샘플링**: MCMC 샘플링 알고리즘 (메트로폴리스헤이스팅스, 갑스샘플러, SLICE SAMPLING,,)
3. **샘플링 반복**: 반복을 통해 t번째 샘플  $(X_1^t, \dots, X_n^t)$  생성
4. **수렴필요**: 반복횟수 t가 어느정도 커야 샘플  $(X_1^t, \dots, X_n^t)$ 을 함수  $f(x_1, \dots, x_n)$ 에서 sampling했다고 볼 수 있다.(목표분포로의 수렴)
5. **수렴여부 진단** : 겔만 루빈 통계량/ACF

# 2

## Bayesian Iterative Simulation Methods

---

2.0 Introduction

2.1 Data Augmentation (10.1.1)

2.2 The Gibbs' Sampler (10.1.2)

2.3 Assessing Convergence of Iterative Simulations (10.1.3)

2.4 Some Other Simulation Methods (10.1.4)

## 2.0 Introduction

---

### 베이지안 접근법의 필요성

Maximum Likelihood (ML)와 같은 방법은 빈도주의의 관점에서 “Small Sample”일 때 문제가 될 수 있다.

따라서 표본의 크기가 작은 경우에는 빈도주의 접근법보다 베이지안 접근법이 유용하다. (week4 내용)

→ 관심 있는 모수에 대한 적절한 prior distribution을 포함시키고, 이로부터 posterior distribution을 도출하는 베이지안 접근법 활용

### 베이지안 접근법 with missingness

1. Data Augmentation (DA) 
2. The Gibbs' Sampler 
3. Sampling Importance Resampling (SIR)
4. Rejection Sampling
5. Metropolis-Hastings Algorithm
6. Bridge Sampling
7. Markov Normal

...

## 2.0 Introduction

---

시뮬레이션의 반복이 필요 없는 경우 : 결측 무시 가능 & 사전 독립

결측을 무시할 수 있을 때의 Posterior Distribution

$$p(\theta|Y_{(0)}, M) \equiv p(\theta|Y_{(0)}) = \text{const.} \times \underbrace{p(\theta)}_{\text{prior distribution}} \times f(Y_{(0)}|\theta)$$

Factored Posterior Distribution

베이지안 분석에서 모수들끼리 사전 독립일 때, Likelihood function을 완전한 데이터 구성 요소들로 분해하는 게 가능하다. (chapter 7 내용)

$$L(\phi|Y_{(0)}) = \prod_{q=1}^Q L_q(\phi_q|Y_{(0)})$$

그렇다면 모수들끼리 사전 독립 (prior independent)일 때,

1. 결합된 prior distribution이 모수 각각의 prior distribution으로 분해되고,
2. 결합된 density는 likelihood function과 동일하며, 이는 모수 각각의 likelihood function으로 분해될 수 있다.

따라서 posterior distribution도 분해가 가능해지고, 각각의 posterior distribution에서 샘플을 추출할 수 있다.

시뮬레이션의 반복이 필요한 경우 ➡ DA, Gibbs' Sampler 등의 방법들을 활용 !

## 2.1 Data Augmentation

---

### Data Augmentation (DA, Tanner and Wong 1987)

: posterior distribution를 반복적으로 시뮬레이션하는 방법으로, EM algorithm과 multiple imputation의 특징을 결합했다.  
다만 EM과는 달리 소규모 샘플 기반이며, 분포로부터 직접 샘플링한다.

우선 posterior distribution의 근사치로부터  $\theta^{(0)}$ 을 추출하며, 이 초기값은 첫 번째 샘플로 간주된다.

**I Step** : Imputation (EM의 Expectation)

$p(Y_{(1)}|Y_{(0)}, \theta^{(t)})$ 에서  $Y_{(1)}^{(t+1)}$  샘플링

**P Step** : Posterior (EM의 Maximization)

$Y_{(1)}^{(t+1)}$ 로 업데이트한  $p(\theta|Y_{(0)}, Y_{(1)}^{(t+1)})$ 에서  $\theta^{(t+1)}$  샘플링

위의 과정을 여러 번 반복하면 결합된 posterior distribution에서 샘플링한 것에 수렴한다.

## 2.2 The Gibbs' Sampler

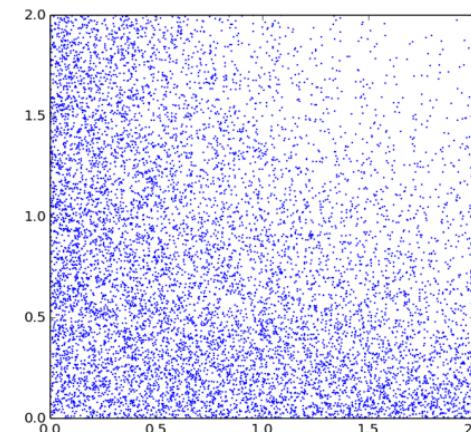
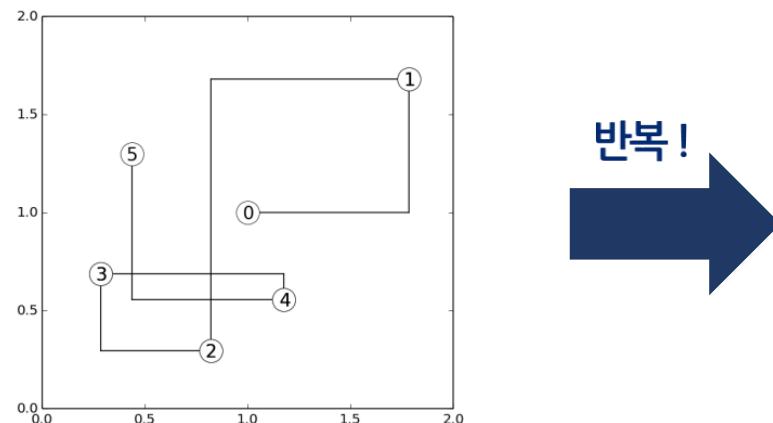
### The Gibbs' Sampler

: posterior distribution를 반복적으로 시뮬레이션하는 방법으로, ECM algorithm과 유사한 방식이다.

다만 모든 단계에서 확률변수의 샘플링을 진행하므로 이해하기 더 쉽다. (ECM algorithm은 각 단계에서 서로 다른 조건부 최대화를 순차적으로 진행 한다. week5 내용)

Gibbs' Sampler는  $X_1, X_2, \dots, X_J$  와 같은  $J$  개의 확률변수로 이루어진 결합 분포  $P(x_1, \dots, x_J)$ 에서 직접 샘플링하는 것이 어려울 때, 비교적 계산이 쉬운 조건부 분포  $p(x_j | x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_J)$ 에서 샘플링하는 방법이다.

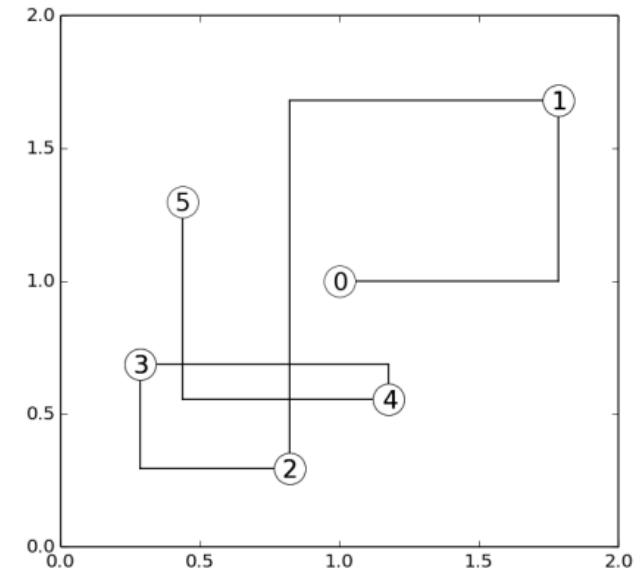
조건부 분포에서 반복해서 샘플링할 경우, 결합 분포에서 샘플링한 것에 수렴한다. (= 결합 분포에서 샘플링하는 것과 비슷한 효과를 낸다.)



## 2.2 The Gibbs' Sampler

우선 각 확률변수에서  $x_1^{(0)}, x_2^{(0)}, \dots, x_J^{(0)}$  와 같은 초기값을 선택한다. 이때 초기값 선택은 어떤 방법을 사용하든 상관없다.

1.  $x_1^{(t+1)} \sim p(x_1 | x_2^{(t)}, x_3^{(t)}, \dots, x_J^{(t)})$
2.  $x_2^{(t+1)} \sim p(x_2 | \underline{x_1^{(t+1)}}, x_3^{(t)}, \dots, x_J^{(t)})$
3.  $x_3^{(t+1)} \sim p(x_3 | \underline{x_1^{(t+1)}}, \underline{x_2^{(t+1)}}, x_4^{(t)}, \dots, x_J^{(t)})$
- ...
- J.  $x_J^{(t+1)} \sim p(x_J | \underline{x_1^{(t+1)}}, \underline{x_2^{(t+1)}}, \dots, \underline{x_{J-1}^{(t+1)}})$



$J$ 개의 확률변수에 대해 모두 샘플링해야 비로소  $\mathbf{x}^{(t+1)} = (x_1^{(t+1)}, \dots, x_J^{(t+1)})$  와 같은 완전한 하나의 샘플을 얻게 된다.

위의 과정을 여러 번 반복하면 결합된 posterior distribution에서 샘플링한 것에 수렴한다.

## 2.2 The Gibbs' Sampler

---

### The Gibbs' Sampler when $J=2$

$J$ 가 2일 때, 즉 확률변수가 2개일 때 이는 Data Augmentation과 본질적으로 동일하다.

$X_1 = Y_{(1)}$ ,  $X_2 = \theta$ 이며, 분포는  $Y_{(0)}$  하에 설정된다고 가정할 때 이는 DA의 작동 방식과 동일하다.

$t$  번째 반복 단계에서  $d$  번째 imputed 데이터 셋에 대해 다음과 같은 과정이 진행된다.

$$1. \quad Y_{(1)}^{(d,t+1)} \sim p(Y_{(1)} | Y_{(0)}, \theta^{(d,t)})$$

$$2. \quad \theta^{(d,t+1)} \sim p(\theta | Y_{(1)}^{(d,t+1)}, Y_{(0)})$$

여기서는 Gibbs' Sampler 자체를 독립적으로  $D$  번 실행했다.

결과적으로  $D$  개의 imputed 데이터 셋이 생성되며,  $Y_{(1)}$ 의 값들은 결측치에 대한 multiple imputations이다.

## 2.2 The Gibbs' Sampler

---

### 예시 : Gibbs' Sampler in Multivariate Normal with Missingness

다음 다변량 정규분포에서  $n$  개의 독립적인 관측을 가지고 있다고 가정한다.

$$y_i \sim \text{iid Multivariate Normal}(\theta, \Sigma)$$

우리가 구하고자 하는 것은 다음과 같다.  $Y_{(1)}, \theta, \Sigma$

이 값들을 추정하기 위해서 Gibbs' Sampler를 활용하고자 한다. 이를 위해서는 각각의 조건부 posterior distribution을 구해야 한다.

$$\theta^{(t+1)} \sim p(\theta | Y_{(0)}, Y_{(1)}^{(t)}, \Sigma^{(t)})$$

$$\Sigma^{(t+1)} \sim p(\Sigma | Y_{(0)}, Y_{(1)}^{(t)}, \theta^{(t+1)})$$

$$Y_{(1)}^{(t+1)} \sim p(Y_{(1)} | Y_{(0)}, \theta^{(t+1)}, \Sigma^{(t+1)})$$

## 2.2 The Gibbs' Sampler

우선,  $Y_{(1)}^{(0)}, \theta^{(0)}, \Sigma^{(0)}$  와 같은 초기값을 설정한다.

Step 1.  $\theta^{(t+1)} \sim p(\theta|Y_{(0)}, Y_{(1)}^{(t)}, \Sigma^{(t)})$

결과적으로 위의 분포는 다면량 정규 분포를 따른다.

prior distribution을 설정하고, posterior distribution을 구하면 다음과 같이 도출된다.

평균:  $(\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y})$       공분산 행렬:  $(\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y})$

Step 2.  $\Sigma^{(t+1)} \sim p(\Sigma|Y_{(0)}, Y_{(1)}^{(t)}, \theta^{(t+1)})$

결과적으로 위의 분포는 역위샤트 분포를 따른다.

prior distribution을 설정하고, posterior distribution을 구하면 다음과 같이 도출된다.

$$\text{inverse-Wishart}(\nu_0 + n, [S_0 + S_\theta]^{-1})$$

## 2.2 The Gibbs' Sampler

---

Step 3.  $Y_{(1)}^{(t+1)} \sim p(Y_{(1)}|Y_{(0)}, \theta^{(t+1)}, \Sigma^{(t+1)})$

결과적으로 위의 분포는 다면량 정규 분포를 따른다.

우선, 위의 분포를 각 관측에서 결측에 대한 조건부 분포의 곱으로 나타낼 수 있다는 다음의 결과를 활용한다.

$$p(Y_{(1)}|Y_{(0)}, \theta, \Sigma) \propto \prod_{i=1}^n p(y_{i,(1)}|y_{i,(0)}, \theta, \Sigma)$$

나누어진 각 조건부 분포도 당연하게 다변량 정규 분포를 따르는데, 평균과 공분산 행렬은 다음과 같이 구해진다.

평균:  $\theta_{b|a} = \theta_{[b]} + \Sigma_{[b,a]} (\Sigma_{[a,a]})^{-1} (y_{[a]} - \theta_{[a]})$

공분산 행렬:  $\Sigma_{b|a} = \Sigma_{[b,b]} - \Sigma_{[b,a]} (\Sigma_{[a,a]})^{-1} \Sigma_{[a,b]}$

이때,  $a$ 는 관측된 변수의 집합이고  $b$ 는 결측된 변수의 집합이다.

## 2.3 Assessing Convergence of Iterative Simulations

---

### 수렴을 평가하는 것의 필요성

만약 DA나 Gibbs' Sampler이 충분히 반복되지 않았다면, **타겟 분포를 잘 대표하지 못할 수 있다.**

따라서 우리가 진행하고 있는 시뮬레이션이 타겟 분포에 잘 수렴하고 있는지를 확인해볼 필요가 있다.

그러나 각 반복 시뮬레이션마다 '단일 목표값'이 존재하지 않기 때문에 수렴을 평가하는 것은 쉽지 않다.

### 수렴 평가의 기초 아이디어

모수의 공간 전체에 분산된  $D > 1$  개의 시퀀스를 시뮬레이션하는 방법이다. 앞서 Gibbs' Sampler 자체를 반복하여 여러 데이터 셋을 생성하는 예시를 살펴보았다. 이 경우 진행 과정은 동일하지만 '**서로 다른 초기값**'에서 시작한다는 것이 큰 차이일 것이다.

즉, 타겟 분포에 수렴하는지 평가하기 위해서 서로 다른 초기값에서 시작한 여러 시퀀스를 생성하겠다는 것이다.

→ 시퀀스 간 분산도(variation between sequences)와 시퀀스 내 분산도(variation within sequences)가 거의 동일해질 때 수렴한다.

## 2.3 Assessing Convergence of Iterative Simulations

---

### Potential Scale Reduction Statistic (= Gelman-Rubin Statistic)

: 앞의 아이디어를 발전시켜 만든 모니터링 통계량으로, 다음과 같이 표현된다.

$$\sqrt{\widehat{R}} = \sqrt{\frac{\widehat{\text{Var}}^+(\psi|Y_{(0)})}{\bar{V}}}$$

- $\psi$  : 우리가 관심을 갖고 확인, 추정하려는 스칼라 값 (estimand)
- $\psi_{d,t}$  :  $d$  번째 시퀀스의  $t$  번째 샘플 값
- $D$  : 총 시퀀스의 수
- $T$  : 시퀀스 내에서의 총 반복 횟수
- $\bar{\psi}_{d\cdot}$  :  $d$  번째 시퀀스의 평균,  $\frac{1}{T} \sum_{t=1}^T \psi_{d,t}$
- $\bar{\psi}_{..}$  : 모든 시퀀스의 평균,  $\frac{1}{D} \sum_{d=1}^D \bar{\psi}_d$
- $s_d^2$  :  $d$  번째 시퀀스의 분산,  $\frac{1}{T-1} \sum_{t=1}^T (\psi_{d,t} - \bar{\psi}_{d\cdot})^2$

시퀀스 간 분산(variance between sequences)

$$B = \frac{T}{D-1} \sum_{d=1}^D (\bar{\psi}_{d\cdot} - \bar{\psi}_{..})^2$$

시퀀스 내 분산(variance within sequences)

$$\bar{V} = \frac{1}{D} \sum_{d=1}^D s_d^2$$

## 2.3 Assessing Convergence of Iterative Simulations

---

주변 사후 분산(marginal posterior variance)의 추정

$$\widehat{\text{Var}}^+(\psi|Y_{(0)}) = \frac{T-1}{T}\bar{V} + \frac{1}{T}B$$

marginal:  $p(\theta_1|y)$

conditional:  $p(\theta_1|\theta_2, y)$

:  $\text{Var}(\psi|Y_{obs})$ 는 관심 있는 특정 모수에 대한, 사후 분포의 분산(= 타겟 분포의 분산)이다.

유한한  $T$ 에 대해  $\bar{V}$ 는 주변 사후 분산의 과소 추정량이 될 수 있다. 개별 시퀀스는 타겟 분포의 전체 범위를 탐색할 시간이 충분하지 않았기 때문이다. 하지만 각 시퀀스가 서로 다른 초기값에서 시작하기 때문에, 시퀀스 간의 차이는 클 수 있다. 결과적으로,  $\bar{V}$ 는  $B$  보다 작다.

$T \rightarrow \infty$  일 때는 개별 시퀀스가 타겟 분포의 전체 범위를 충분히 탐색하게 되고, 각 시퀀스 간의 차이도 줄어들게 된다. 즉,  $\bar{V}$ 는 점차 커지며 타겟 분포의 분산을 잘 반영하게 되고,  $B$ 는 점차 작아지며 상대적으로 작은 영향을 가지게 된다. 그리고  $\bar{V}$ 와  $B$ 는 거의 동일해진다.

위의 내용들을 바탕으로 편향을 보정하기 위해 가중치를 부여한 추정량을 설계했다.

## 2.3 Assessing Convergence of Iterative Simulations

---

Potential Scale Reduction Statistic (= Gelman-Rubin Statistic)

$$\sqrt{\widehat{R}} = \sqrt{\frac{\widehat{\text{Var}}^+(\psi|Y_{(0)})}{\overline{V}}}$$

$T \rightarrow \infty$  일 때,  $\overline{V}$ 는 주변 사후 분산에 점차 가까워지게 된다. 결국에는 주변 사후 분산과  $\overline{V}$ 를 비교해보아야 수렴을 확인할 수 있다. 즉, 위의 값이 1에 가까울수록 시뮬레이션이 수렴했음을 의미하며, 1보다 크다면 아직 수렴하지 않았다는 의미이다.

이때 우리가 확인하고 싶은 값( $\psi$ ) 모두에 대해 1에 가까워지지 않는다면 시뮬레이션을 계속해서 진행해야 하며, 경우에 따라 알고리즘 자체를 수정할 필요가 있다.

→ 모두에 대해 1에 가까워진다면, 그 이후 시퀀스에서 얻은 모든 샘플들을 타겟 분포의 샘플로 간주할 수 있다.

\* 일반적으로 1.1~1.2 이하의 값을 기준으로 두지만 상황에 따라 달라질 수 있다.

## 2.4 Some Other Simulation Methods

---

### Sampling Importance Resampling (SIR)

타겟 분포  $f(\theta)$ 에서 직접 샘플링하는 것이 계산 문제로 어려울 때, 동일한 support를 가진 근사 분포  $g(\theta)$ 에서 샘플을 얻을 수 있다.

이때 중요도 가중치(importance weight)인  $R_d = f(\theta_d)/g(\theta_d)$ 를 적용하여, 샘플의 중요도에 따라 가중치를 부여한다.

예시)  $f(\theta)$ 에서 10개의 샘플을 뽑고자 한다. 우선  $g(\theta)$ 에서 100개를 뽑고, 중요도 가중치에 따라 확률 비례하게 최종 10개를 뽑는다.

### Rejection Sampling

중요도 가중치  $R_d$ 를 특정 상수와 비교하여 샘플을 순차적으로 수락하거나 거부하는 방법

### Metropolis-Hastings Algorithm

Rejection Sampling을 Gibbs' Sampler에 포함시키는 방법

### Bridge Sampling

잘못된(=목표와 다른) 분포에서 샘플을 얻어 타겟 분포로 가는 “bridge”를 만드는 방법

### Markov Normal

MCMC 방법을 사용해 여러 개의 초기 샘플을 생성하고, 이를 분석함으로써 타겟 분포에 더 효율적으로 수렴할 수 있도록 돋는 방법

# 3

## Multiple Imputation

---

3.0 Multiple Imputation Tutorial

3.1 Large-Sample Bayesian Approximations of the Posterior Mean and Variance  
Based on a Small Number of Draws (10.2.1)

3.2 Other Methods for Creating Multiple Imputations (10.2.3)

3.3 Chained-Equation Multiple Imputation (10.2.4)

# Overview

---

0. 베이지안의 접근 방식

1. MCMC 알고리즘을 사용한 방법론

- Data Augmentation
- Gibb's sampler

2. **Multiple Imputation**에 베이지안 알고리즘을 사용하자!

- Multiple Imputation이란?
- 베이지안 알고리즘과 multiple imputation (10.1, 10.3)
- Multiple Imputation Chained-equation (10.4)

→ 지난 내용들을 먼저 살펴 볼 예정

# 3.0 Multiple Imputation Tutorial

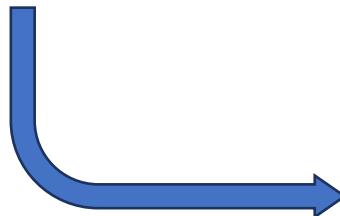
---

## Data analysis의 흐름

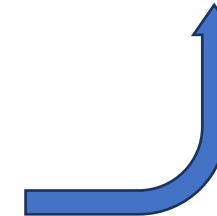
데이터를 얻는다(sampling)



얻은 데이터를 기반으로 적절한 방법으로  
통계적 분석을 한다



분석에 들어가기 전 데이터 자체를 살펴보고,  
좋은 분석이 가능한 형태로 만든다!



- 데이터에 결측치가 발견됨
- Missing value들을 어떻게 처리하지?  
(Missing value가 포함된 data set은 어떻게 처리하지?)

# 3.0 Multiple Imputation Tutorial

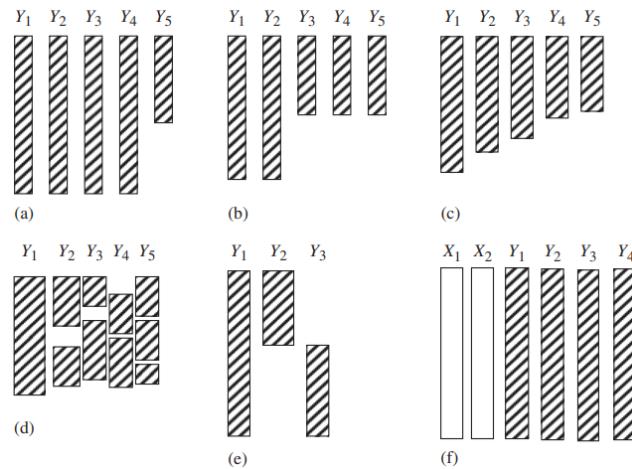
- Missing value들을 어떻게 처리하지?  
(Missing value가 포함된 data set은 어떻게 처리하지?)

일단 먼저,

## 1. 어떤 형태로 Missing 되어 있는지 살펴보자(Missing Pattern)

다양한 패턴이 있음!

- 특정 상황에서는 특정 패턴으로만 missing됨
- 특정 패턴은 좀 더 간편하게 처리가 가능함



## 2. 어떻게 Missing이 된 건지 살펴보자(Missing Mechanism)

MCAR : 정말 random한 missing

MAR : 어떤 값이 Missing된 것이 **observed**와만 관련이 있고  
Missing된 값의 원래 값과는 관련이 없다

MNAR : 어떤 값이 Missing된 것이 **observed**와도 관련이 있고  
**Missing된 값**의 원래 값과도 관련이 있다

- 결국 어떤 값이 비어 있는 이유가 무엇인지 생각해보는 것!

- 1) MCAR이나 MAR의 경우 (합리적인 방법으로) **모델링**을 통해 적당한 분석이 가능함!  
(MNAR의 경우 missing mechanism을 분석하여 모델링 자체가 어려움)
- 2) MAR의 경우 특정 상황에서 **Missing Mechanism을 무시**할 수 있는 경우가 생겨  
간단한 모델링이 가능함!  
+ missing mechanism을 고려하지 않을 경우 (그래서 틀린 경우) 분석에 문제가 발생함

## 3.0 Multiple Imputation Tutorial

---

☆ MAR의 경우 Missing Mechanism을 무시할 수 있는 경우가 생긴다 ☆

→ Multiple imputation이 용이해지는 이유

(모델이 간소화되면서 간단한 모델 구축이 가능해짐! – 채워 넣을 때 결측 패턴을 반영하지 않아도 됨)

- Missing mechanism을 무시할 수 있는 조건 (Week 4)

(a) Missing mechanism: MAR (+MCAR)

(b) Missing mechanism & Data의 parameter  $\psi, \theta$ 가 independent

→ 뒤에서 간단하게 다룰 예정!

# 3.0 Multiple Imputation Tutorial

---

## Multiple Imputation

1) Why imputation? (Week 2)

2) Why Multiple Imputation? (Week 3)

# 3.0 Multiple Imputation Tutorial

---

## Multiple Imputation

### 1) Why imputation?

#### Deletion vs Imputation

	X	Y
1	12	7
2	8	?
3	10	11
4	9	?
5	15	9
6	20	25
7	5	5
8	11	16
9	4	6
10	18	8

#### (Row) Deletion

- 결측이 있는 행(unit)을 제거하는 방법
- 정보의 손실
- 특정 Missing mechanism(MCAR)을 제외하면 거의 대부분 bias가 발생한다

#### Imputation

- 결측이 있는 부분을 특정 값으로 채워 넣는 방법
- Imputation method에 따라 bias가 발생하기도 한다
- Single imputation의 경우 모수의 uncertainty가 underestimate된다

# 3.0 Multiple Imputation Tutorial

## Multiple Imputation

### 2) Why Multiple Imputation?

Single imputation vs multiple imputation

X	Y	Z	W
3	-	1	4
6	5	2	8
3	-	-	5
-	2	5	9

Single imputation

- Mean imputation
  - 채워 넣는 값 자체에 uncertainty가 부여되지 않음
  - 서로 다른 변수의 관계를 고려하지 않음
  - bias가 발생
- Regression imputation
  - 채워 넣는 값에는 uncertainty가 부여됨 but 채워 넣은 data set을 활용한 estimate에는 uncertainty가 부여되지 않음

Multiple imputation

- 채워 넣는 값에도 uncertainty를 부여하고
- 채워 넣은 data set을 활용한 estimate에도 uncertainty가 부여됨

Uncertainty (불확실성)

- 내가 채워 넣은 값이 틀렸을 수도 있다는 사실을 고려했는가?

# 3.0 Multiple Imputation Tutorial

## Multiple imputation

- 채워 넣는 값에도 uncertainty를 부여하고
- 채워 넣은 data를 활용한 estimate에도 uncertainty가 부여됨

	X	Y
1	9	11
2	10	?
3	12	15
4	14	?
5	17	17
6	19	15
7	7	21
8	5	5
9	21	19
10	25	23

- 주어진 Data set
- X는 complete하고 Y의 2, 4번째 unit에 missing이 존재
- multiple imputation을 활용하여 종속변수 **Y의 평균( $\theta$ )**을 추정하고 싶음!

How?

- **Y의 평균( $\theta$ )**의 평균 추정치 = Y의 표본평균
- **Y의 평균( $\theta$ )**의 분산 추정치 = Y의 표본평균의 분산

	(1)	(2)	(3)	(4)
2	11	14	13	18
4	16	14	20	18

- 위의 표는 Y의 결측값을 대체할 Data set
- 각 값은 다양한 기법으로 구할 수 있다!  
(채워 넣는 값에 uncertainty 부여)
- 각 값을 대입한 4개의 complete data를 생성  
(채워 넣은 data set을 활용한 estimate에 uncertainty 부여)

# 3.0 Multiple Imputation Tutorial

	X	Y		X	Y		X	Y		X	Y		X	Y
1	9	11	1	9	11	1	9	11	1	9	11	1	9	11
2	10	11	2	10	14	2	10	13	2	10	18	2	10	18
3	12	15	3	12	15	3	12	15	3	12	15	3	12	15
4	14	16	4	14	14	4	14	20	4	14	18	4	14	18
5	17	17	5	17	17	5	17	17	5	17	17	5	17	17
6	19	15	6	19	15	6	19	15	6	19	15	6	19	15
7	7	21	7	7	21	7	7	21	7	7	21	7	7	21
8	5	5	8	5	5	8	5	5	8	5	5	8	5	5
9	21	19	9	21	19	9	21	19	9	21	19	9	21	19
10	25	23	10	25	23	10	25	23	10	25	23	10	25	23

sample mean **15.3**

**15.4**

**15.9**

**16.2**

sample variance **28**

**26.3**

**28.5**

**26.6**

variance of  
sample mean **2.8**

**2.63**

**2.85**

**2.66**

이렇게 구하고 난 뒤, multiple imputation의 방법론을 이용하여 **Y의 평균**과 관련된 추정치를 다음과 같이 계산할 수 있음!

평균 = 각 complete data set에서의 **표본평균들의 평균**

분산1 = 각 complete data set에서의 **표본평균의 분산들의 평균**

분산2 = 각 complete data set에서의 **표본평균들의 표본분산**

총분산 = 분산1 + (1+1/4) x 분산2

즉,

평균 =  $(15.3 + 15.4 + 15.9 + 16.2)/4 = 15.7$

분산1 =  $(2.8 + 2.63 + 2.85 + 2.66)/4 = 2.735$

분산2 =  $\frac{(15.3-15.7)^2+(15.4-15.7)^2+(15.9-15.7)^2+(16.2-15.7)^2}{4-1} = 0.18$

총분산 =  $2.735 + (1 + \frac{1}{4})0.18 = 2.96$

# 3.0 Multiple Imputation Tutorial

$\hat{\theta}_d, d = 1, \dots, D$ 를  $D$ 개의 complete-data에서 계산한  $\theta$ 의 추정치라고 한다면, 이를 이용하여  $\bar{\theta}_D$ 를 다음과 같이 정의할 수 있다.

$$\bar{\theta}_D = \frac{1}{D} \sum_{d=1}^D \hat{\theta}_d$$

...

$$(\theta - \bar{\theta}_D) T_D^{-1/2} \sim t_v$$
$$v = (D-1) \left(1 + \frac{1}{D+1} \frac{\bar{W}_D}{B_D}\right)^2$$

또한  $W_d, d = 1, \dots, D$ 를 각 complete-data에서 구한  $\theta$ 의 분산의 추정치라고 한다면, 이를 이용하여  $\theta$ 의 총변동  $T_D$ 를 다음과 같이 정의할 수 있다.

$$T_D = \bar{W}_D + \frac{D+1}{D} B_D$$

$$\bar{W}_D = \frac{1}{D} \sum_{d=1}^D W_d$$

$$B_D = \frac{1}{D-1} \sum_{d=1}^D (\hat{\theta}_d - \bar{\theta}_d)^2$$

여기서  $\bar{W}_D$ 는 average within-imputation variance,  $B_D$ 는 between-imputation component라고 부른다.

Multiple imputation으로 인해  
생겨난 분산

대표본인 경우  $v$ 의 자유도를 가진  $t$  분포를 사용하여 구간추정 및 검정을 진행할 수 있다.

1) 왜  $(1+1/D)$ 를 곱할까?

-  $D$ 가 작은 경우 분산이 작게 추정되는 것을 보완하기 위한 조정

2) 왜 대표본인데  $t$ 분포일까?

-  $D$ 가 작은 경우 분산을 과소추정하지 않기 위해

3) 왜  $D$ 가 작을까...?

- Rubin 박사님의 multiple imputation 아이디어는 1970년대 초에 고안되었음

-  $D$ 가 클 경우, 큰 계산량과 메모리가 소모됨

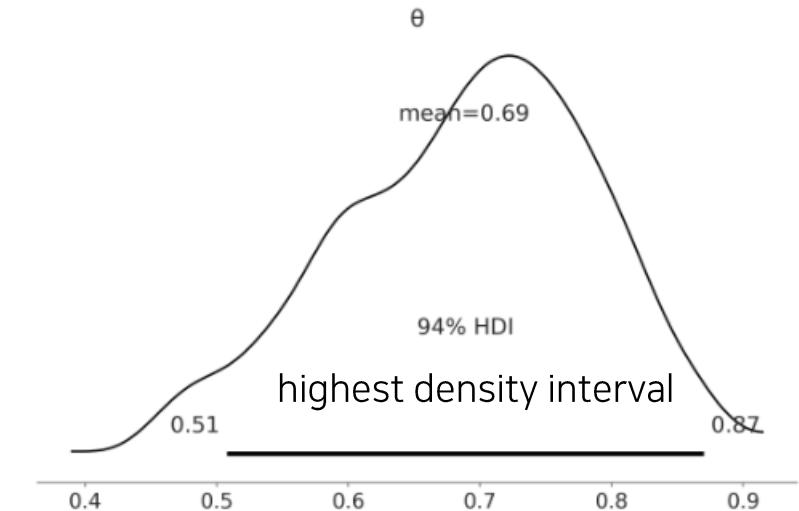
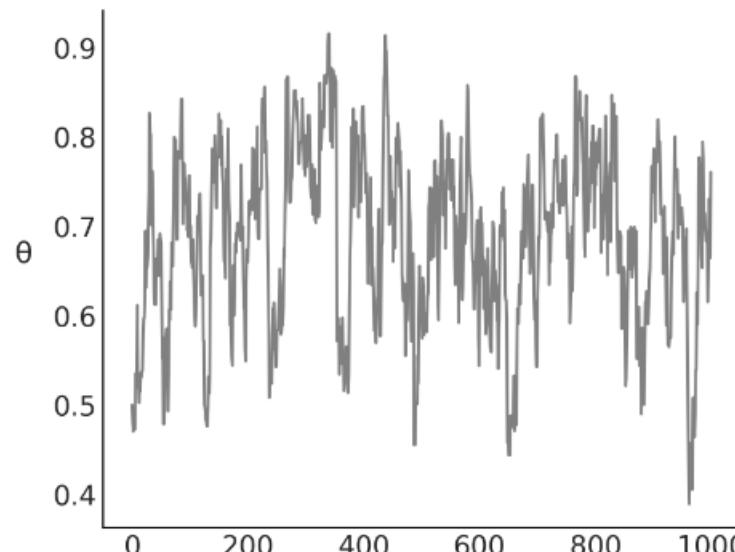
-  $D=5$  정도만으로 충분하다는 결론 → 뒤에서 논의

## 3.1 Large-Sample Bayesian Approximations of the Posterior Mean and Variance Based on a Small Number of Draws

---

앞에서 발표한 10.1절의 내용

- bayesian iterative method는 결국  $\theta$ 의 posterior 분포에서의 draw를 만들어 냄을 알 수 있었음
- 유의미한  $\theta$ 들을 draw했다고 했을 때,  $\theta$ 의 값들을 empirical distribution으로서 사용하여  $\theta$ 에 대한 추론을 하려면  
**몇 천개** 정도의 독립적인 draw가 요구됨



## 3.1 Large-Sample Bayesian Approximations of the Posterior Mean and Variance Based on a Small Number of Draws

---

$$p(\theta|Y_{(0)})$$

- 그러나 observed-data posterior distribution의 정규근사를 가정한다면  
몇 백 이하의 draw만으로 posterior distribution의 평균과 분산을 추정할 수 있음
- 예를 들어 처음부터 적당한 수의 draw( $\theta$ )들을 t분포와 같은 model에 fitting한다면  
엄청나게 많은 draw없이도 우리가 실제로 구한 empirical distribution은 posterior distribution의 추정에 사용하기 충분해지게 됨

더 적은 수의 draw?

## 3.1 Large-Sample Bayesian Approximations of the Posterior Mean and Variance Based on a Small Number of Draws

---

$$p(\theta|Y_{(0)}, Y_{(1)})$$

만약 missing의 부분이 적고 complete-data posterior distribution에 대한 추론이 multivariate normal이나 t distribution을 기반으로 한다면  $\theta$ 의 posterior mean과 posterior variance는 매우 작은 수의 draw (5개-10개) 만으로 유의미한 추정을 할 수 있음!!

How? → multiple imputation의 방법론을 사용

---

Q: multiple imputation은 어떻게 5개만으로 유의미한가? (D=5)

<https://stefvanbuuren.name/fimd/sec-howmany.html> → 이에 대한 구체적인 논의가 담겨있음

- Von Hippel (2009) : *the number of imputations should be similar to the percentage of cases that are incomplete*
- impute 횟수(D) = 100 x missing의 비율 ex) 5%의 missing → 5번의 imputation → 이 때 좋은 성질들은 가지고 있더라
- 이론상 큰 D(impute 횟수)가 항상 더 좋은 것은 맞지만, 추정치가 매우 불확실하진 않거나, 꼭 full distribution을 구하고 싶은 것이 아니거나, 모수의 추정치가 실제 모수와 너무 다른 것이 아니라는 판단을 할 수 있을 경우 적당한 missingness에 대해서는 5-20회 정도로 충분함

## 3.1 Large-Sample Bayesian Approximations of the Posterior Mean and Variance Based on a Small Number of Draws

---

multiple imputation의 절차

– D개의 complete data set을 만들고, 각 complete data set에서의  $\theta$ 와 관련된 값을 구한 후, combining rule에 따라 마무리

- 1) Complete data set 만들기 – by sampling  $Y_{(1)}$
- 2) Complete data set을 이용하여  $\theta$ 의 추정치 구하기 – by sampling  $\theta$
- 3) D번 반복하기 – D개의  $(\theta, Y_{(1)})$  set을 만들어 냄
- 4) Bayesian Approximation과 Combining rule에 따라  $\theta$ 의 posterior mean과 posterior variance를 구함 + 구간추정

# 3.1 Large-Sample Bayesian Approximations of the Posterior Mean and Variance Based on a Small Number of Draws

모델을 설명하기 전 이 단원에서 다루는 posterior model은 전부 missing mechanism이 ignorable한 경우를 다룸

- MAR에서 missing mechanism이 무시되는 이유를 간단하게 살펴보고자 함!

notation

$L_{full}(\theta, \psi | y_{(0)}, m) = \int f_Y(y_{(0)}, y_{(1)} | \theta) f_{M|Y}(m | y_{(0)}, y_{(1)}, \psi) dy_{(1)}$  L full은 missing mechanism을 적용한  $\theta, \psi$  의 full likelihood를 의미

$L_{ign}(\theta | y_{(0)}) = \int f_Y(y_{(0)}, y_{(1)} | \theta) dy_{(1)}$  L ign은 missing mechanism을 무시한  $\theta$ 만의 likelihood를 의미

Ignorable condition ( $\theta, \psi$ 는 독립이라고 가정)

$\frac{L_{full}(\theta)}{L_{full}(\theta^*)} = \frac{L_{ign}(\theta)}{L_{ign}(\theta^*)}$  임의의  $\theta, \theta^*$ 에 대하여 L full의 likelihood ratio와 L ign의 likelihood ratio가 항상 같다면  
우리는 굳이 우도함수로 L full을 사용할 필요가 없음!

Under MAR ( $\theta, \psi$ 는 독립이라고 가정)

$L_{full}(\theta, \psi | y_{(0)}, m) = \int f_Y(y_{(0)}, y_{(1)} | \theta) f_{M|Y}(m | y_{(0)}, y_{(1)}, \psi) dy_{(1)}$   
 $= f_{M|Y}(m | y_{(0)}, \psi) \times \int f_Y(y_{(0)}, y_{(1)} | \theta) dy_{(1)}$  MAR은 m의 값(missingness)가  $y_{(1)}$ (관측되지 않은 값)에 의존하지 않음  
 $= f_{M|Y}(m | y_{(0)}, \psi) \times f(y_{(0)} | \theta)$  Likelihood ratio로 나타냈을 때, L ign 부분만 남음을 볼 수 있음! → 위의 조건을 만족함

## 3.1 Large-Sample Bayesian Approximations of the Posterior Mean and Variance Based on a Small Number of Draws

---

$$p(\theta, \psi | y_{(0)}, m) \propto p(\theta, \psi) \times L_{full}(\theta, \psi | y_{(0)}, m)$$
$$p(\theta | y_{(0)}) \propto p(\theta) \times L_{ign}(\theta | y_{(0)})$$

따라서 만약 missing mechanism이 MAR(이고  $\theta, \psi$ 는 독립)이라면  
bayesian inference를 위한 posterior distribution을 구축할 때  
Missing mechanism이 포함된 위의 식이 아닌  
Missing mechanism을 무시한 아래의 식을 이용할 수 있게 됨!!

$$p(\theta | Y_{(0)}, M) = \underline{p(\theta | Y_{(0)})} = \text{const.} \times p(\theta) \times f(Y_{(0)} | \theta)$$

따라서 이 단원에서는 이 형태에만 관심을 가짐

## 3.1 Large-Sample Bayesian Approximations of the Posterior Mean and Variance Based on a Small Number of Draws

---

다시 전으로 돌아가서, bayesian multiple imputation을 하기 위한 모델을 구축하고자 함

- Complete posterior distribution  $p(\theta|Y_{(0)}, Y_{(1)})$  이 multivariate normal이나 t분포를 따른다는 가정하에 MI를 사용한다고 했기 때문에
- 우리는  $p(\theta|Y_{(0)})$  (observed posterior distribution)와  $p(\theta|Y_{(0)}, Y_{(1)})$  (Complete posterior distribution)을 연관시키는 아이디어 이용
- 이에 다음과 같은 식을 쓸 수 있음!  
→여기서 sampling하는 것이 쉽기 때문

$$p(\theta|Y_{(0)}) = \int p(\theta, Y_{(1)}|Y_{(0)})dY_{(1)} = \int p(\theta|Y_{(0)}, Y_{(1)})p(Y_{(1)}|Y_{(0)})dY_{(1)}$$

위의 식을 살펴보면  $p(\theta|Y_{(0)})$ 는  $Y_{(1)}$ 을  $p(Y_{(1)}|Y_{(0)})$ 에서 draw하고 그 값을 적용한 complete-data posterior distribution인  $p(\theta|Y_{(0)}, Y_{(1)})$ 에서  $\theta$ 를 draw하여 구할 수 있음을 생각해볼 수 있다. 즉,

첫 번째 draw에서는  $Y_{(1)}^{(1)}$ 을  $p(Y_{(1)}|Y_{(0)})$ 에서,  $\theta^{(1)}$ 을  $p(\theta|Y_{(0)}, Y_{(1)}^{(1)})$ 에서 draw,

두 번째 draw에서는  $Y_{(1)}^{(2)}$ 을  $p(Y_{(1)}|Y_{(0)})$ 에서,  $\theta^{(2)}$ 을  $p(\theta|Y_{(0)}, Y_{(1)}^{(2)})$ 에서 draw,

세 번째 draw에서는  $Y_{(1)}^{(3)}$ 을  $p(Y_{(1)}|Y_{(0)})$ 에서,  $\theta^{(3)}$ 을  $p(\theta|Y_{(0)}, Y_{(1)}^{(3)})$ 에서 draw...

를 반복하여  $D$ 개의 draw를 얻게 된다.

하지만 결국 궁금한 것은  $p(\theta|Y_{(0)})$ 의 평균과 분산!

→ approximation 공식으로 해결!

## 3.1 Large-Sample Bayesian Approximations of the Posterior Mean and Variance Based on a Small Number of Draws

---

$$p(\theta|Y_{(0)}) = \int p(\theta, Y_{(1)}|Y_{(0)})dY_{(1)} = \int p(\theta|Y_{(0)}, Y_{(1)})p(Y_{(1)}|Y_{(0)})dY_{(1)}$$

위의 식을 사용하면 observed-data posterior distribution의 mean과 variance를 다음과 같이 간단히 작성할 수 있다.

$$E(\theta|Y_{(0)}) = E[E(\theta|Y_{(0)}, Y_{(1)})|Y_{(0)}]$$

$$Var(\theta|Y_{(0)}) = E[Var(\theta|Y_{(0)}, Y_{(1)})|Y_{(0)}] + Var[E(\theta|Y_{(0)}, Y_{(1)})|Y_{(0)}]$$

Multiple imputation에서는 missing value에 대한 적분을 다음과 같은 평균으로 근사할 수 있다.

$$p(\theta|Y_{(0)}) \approx \frac{1}{D} \sum_{d=1}^D p(\theta|Y_{(1)}^{(d)}, Y_{(0)})$$

이 근사를 이용하여 위에서 정리한  $p(\theta|Y_{(0)})$ 의 평균과 분산또한 근사할 수 있음!

여기서  $Y_{(1)}^{(d)} \sim p(Y_{(1)}|Y_{(0)})$ 는  $Y_{(1)}$ 의 posterior distribution으로부터의 draw이다

## 3.1 Large-Sample Bayesian Approximations of the Posterior Mean and Variance Based on a Small Number of Draws

---

같은 방식으로 평균과 분산에 대해서도 근삿값을 구할 수 있다.

$$E(\theta|Y_{(0)}) \approx \int \theta \frac{1}{D} \sum_{d=1}^D p(\theta|Y_{(1)}^{(d)}, Y_{(0)}) d\theta = \frac{\sum_{d=1}^D E(\theta|Y_{(1)}^{(d)}, Y_{(0)})}{D} = \bar{\theta}$$

여기서  $E(\theta|Y_{(1)}^{(d)}, Y_{(0)}) = \hat{\theta}_d$ 로 표기할 수 있고, 이는  $d$ 번째 complete data set에서  $\theta$ 의 평균을 의미한다.

$$Var(\theta|Y_{(0)}) \approx \frac{1}{D} \sum_{d=1}^D V_d + \frac{1}{D-1} \sum_{d=1}^D (\hat{\theta}_d - \bar{\theta})^2 = \bar{V} + B$$

여기서  $V_d = Var(\theta|Y_{(1)}^{(d)}, Y_{(0)})$ 으로,  $d$ 번째 complete data set에서  $\theta$ 의 분산을 의미한다.  $\bar{V}$ 는  $D$ 개의 posterior variance, 즉  $V_d = Var(\theta|Y_{(1)}^{(d)}, Y_{(0)})$ 의 평균,  $B$ 는 imputation 간의 분산이다. 즉 데이터 내에서의 변동과, imputation 간의 변동이 합쳐진 형태이다.

작은  $D$ 에 대해서 posterior mean은 위의 평균 공식을 사용하지만 posterior variance는 improved approximation을 적용하는데 그 식은 다음과 같다.

$$Var(\theta|Y_{(0)}) \approx \bar{V} + (1 + D^{-1})B$$

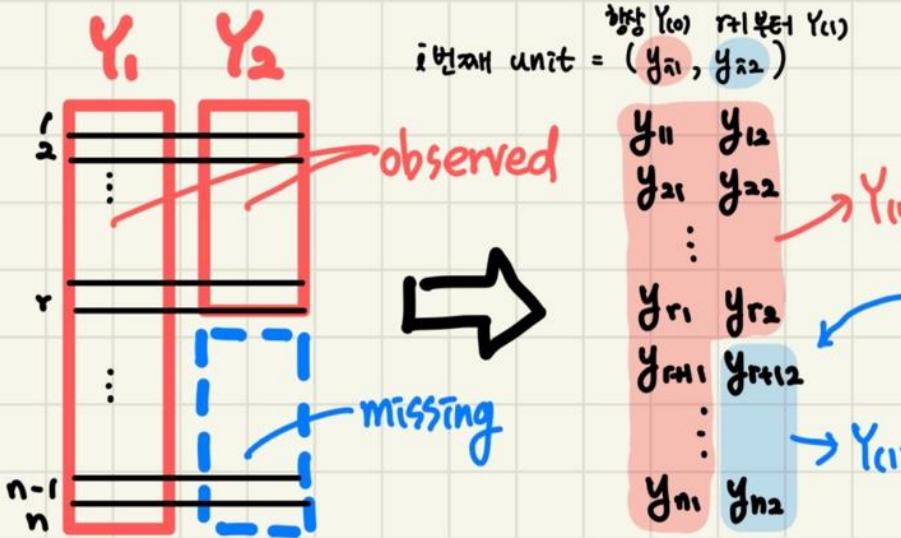
작은  $D$ 에 대해 조정을 주는 이유는 분산의 과소추정 방지와 reference distribution을 normal에서 t로 대체하기 위함이다. 이때의 자유도는  $v = (D-1)(1 + \frac{D}{D+1} \frac{\bar{V}}{B})^2$ 이다.

앞에서 살펴본 내용과 사실상 같은 내용

Bayesian simulation + multiple imputation 방법론

# 3.1 Large-Sample Bayesian Approximations of the Posterior Mean and Variance Based on a Small Number of Draws

간단한 예시



Model of  $Y = [Y_1, Y_2]$  : Multivariate normal

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim N(\mu = \begin{bmatrix} M_1 \\ M_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \delta_{11} & \delta_{12} \\ \delta_{21} & \delta_{22} \end{bmatrix})$$

$$\theta = (M_1, M_2, \delta_{11}, \delta_{12}, \delta_{22})^T$$

$M_2$   $\hat{=}$  estimate  $\bar{M}_2$

$$P(\theta | Y_{(0)}) = \int P(\theta | Y_{(0)}, Y_{(1)}) P(Y_{(1)} | Y_{(0)}) dY_{(1)}$$

② normal  
or t

$$\Rightarrow (Y_{(11)}^{(1)}, \hat{\theta}^{(1)}) = (\hat{M}_1^{(1)}, \hat{M}_2^{(1)}, \delta_{11}^{(1)}, \delta_{12}^{(1)}, \delta_{22}^{(1)})$$

$$(Y_{(11)}^{(2)}, \hat{\theta}^{(2)})$$

$$(Y_{(11)}^{(3)}, \hat{\theta}^{(3)})$$

$$(Y_{(11)}^{(4)}, \hat{\theta}^{(4)})$$

$$(Y_{(11)}^{(5)}, \hat{\theta}^{(5)})$$

$$\therefore \tilde{M}_2 = \frac{\bar{y}_2^{(1)} + \bar{y}_2^{(2)} + \bar{y}_2^{(3)} + \bar{y}_2^{(4)} + \bar{y}_2^{(5)}}{5}$$

$\therefore \text{Var}(\tilde{M}_2)$

$$= \frac{1}{5} \sum_{d=1}^5 \frac{s_{22}^{(d)}}{n} + \left(1 + \frac{1}{5}\right) \frac{1}{4} \sum_{d=1}^5 (\bar{y}_2^{(d)} - \tilde{M}_2)^2$$

$\therefore 95\% \text{ CI of } M_2$

$$\tilde{M}_2 \pm t_{v, 0.025} \cdot \sqrt{\text{Var}(\tilde{M}_2)}$$

$$\left( \because \frac{M_2 - \tilde{M}_2}{\sqrt{\text{Var}(\tilde{M}_2)}} \sim t_v \right)$$

## 3.2 Other Methods for Creating Multiple Imputations

---

앞서 살펴본 예시를 살펴보면 무언가 걸리는 것이 있는데 바로  $Y_{(1)}^{(d)} \sim p(Y_{(1)}|Y_{(0)})$

$$p(\theta|Y_{(0)}) = \int p(\theta, Y_{(1)}|Y_{(0)})dY_{(1)} = \int p(\theta|Y_{(0)}, Y_{(1)})\underline{p(Y_{(1)}|Y_{(0)})}dY_{(1)}$$

이 식은  $\theta$ 와 무관한 식

앞의 예시에서  $p(Y_{(1)}|Y_{(0)})$ 에 대해서 생각해보면  $Y_2|Y_1$ 의 분포에서 sampling하면 되겠다는 생각을 할 수 있는데,  
 $Y_2|Y_1$ 의 분포는  $\theta$ 가 주어졌을 때 정규분포를 따르는 것이기 때문에  $\theta$ 에 의존하고 있음을 알 수 있음

실제로  $p(Y_{(1)}|Y_{(0)})$ 는  $p(Y_{(1)}, \theta|Y_{(0)})$ 를  $\theta$ 에 대해서 적분한 형태로, 실제로는 구하기 쉽지 않음 (predictive distribution이라고 함)

앞 절에서 소개한 Data augmentation은 반복적인 drawing으로

$\theta, Y_{(1)}$ 의 joint posterior distribution  $p(Y_{(1)}, \theta|Y_{(0)})$ 로의 수렴분포에서 추출하는 효과를 냄으로써 이를 해결

난 5개만 뽑을 건데,,,

→  $p(Y_{(1)}|Y_{(0)})$ 로부터의 적당한 근사 draw를 이용하고, Multiple Imputation의 combining rule을 이용하면 그래도 유효한 추론을 할 수 있음

→ 이에 책(교재 p248)에서는 7가지 방법을 제안 ex)  $Y_{(1)}^{(d)} \sim p(Y_{(1)}|Y_{(0)}, \tilde{\theta})$ , 여기서  $\tilde{\theta}$ 는  $\theta$ 의 ML estimate ( $\theta$ 의 uncertainty가 고려되지 않음)

### 3.3 Chained-Equation Multiple Imputation

---

그 중, 연쇄적으로 값을 update하는 아이디어를 이용하여 Multiple Imputation을 수행하는 방법이

Chained-Equation Multiple Imputation 또는 Multiple Imputation by Chained-Equation(MICE)임

여러 변수의 Joint distribution은 구하기 어렵지만

각 변수에 대한 conditional distribution은 구할 수 있는 경우 사용하기 좋음

	X1	X2	Z
1	2	1	3
2	3	-	7
3	-	-	0
4	8	7	5
5	-	9	1
6	-	-	6

Step 1: missing 부분을 (특정 방법으로) 채워 넣는다

Step 2: X1의 3, 5, 6번째 unit을 나머지 정보로 update (draw)

Step 3: X2의 2, 3, 6번째 unit을 나머지 정보로 update (draw)

→ Step 2부터 반복 (수렴할 때까지) → 1개의 complete data set이 완성

→ D개의 초기값으로 진행하여 D개의 complete set을 만들어 활용할 수 있음!

Update하는 방법 – regression, pmm(R의 MICE 패키지의 default)

### 3.3 Chained-Equation Multiple Imputation - 참고

missing<sup>0</sup> 포함된 변수  $X_1, X_2, \dots, X_K$  와 완전히 관측된 변수  $Z$ 가 있다고 할 때,  $x_{(0)}$ 은  $X_1, X_2, \dots, X_K$ 에서 관측된 데이터의 집합이고  $x_{j(1)}$ 은  $j$  번째 변수인  $X_j$ 의 missing data의 집합이라고 하자.

(a)  $x_{1(1)}^{(0)}, x_{2(1)}^{(0)}, \dots, x_{K(1)}^{(0)}$ 에 어떤 근사 과정을 거쳐 값을 채워 넣는다

(b)  $t$  번째 반복에서의 imputed 값들이  $x_{1(1)}^{(t)}, x_{2(1)}^{(t)}, \dots, x_{K(1)}^{(t)}$ 라고 할 때,

$t+1$  번째에는 각 값이 다음과 같은 predictive distribution에서의 draw로서 update된다

$$x_{1(1)}^{(t+1)} \sim p(x_{1(1)} | x_{(0)}, z, x_{1(1)}^{(t)}, x_{2(1)}^{(t)}, x_{3(1)}^{(t)}, \dots, x_{K-1(1)}^{(t)}, x_{K(1)}^{(t)}),$$

⋮

$$x_{j(1)}^{(t+1)} \sim p(x_{j(1)} | x_{(0)}, z, x_{1(1)}^{(t+1)}, \dots, x_{j-1(1)}^{(t+1)}, x_{j+1(1)}^{(t)}, \dots, x_{K-1(1)}^{(t)}, x_{K(1)}^{(t)}),$$

⋮

$$x_{K(1)}^{(t+1)} \sim p(x_{K(1)} | x_{(0)}, z, x_{1(1)}^{(t+1)}, x_{2(1)}^{(t+1)}, x_{3(1)}^{(t+1)}, \dots, x_{K-1(1)}^{(t+1)})$$

▷  $\{x_{j(1)}^{(t+1)}\}$  을 draw하는 과정

1)  $\theta_j^{(t+1)}$  을 먼저  $(x_{(0)}, z, x_{1(1)}^{(t+1)}, \dots, x_{j-1(1)}^{(t+1)}, x_{j(1)}^{(t)}, \dots, x_{K-1(1)}^{(t)}, x_{K(1)}^{(t)})$  가 주어졌을 때의  $\theta_j$ 의 posterior distribution에서 draw한다.

2) 그 후  $\{x_{j(1)}^{(t+1)}\}$  는  $(x_{(0)}, z, x_{1(1)}^{(t+1)}, \dots, x_{j-1(1)}^{(t+1)}, x_{j(1)}^{(t)}, \dots, x_{K-1(1)}^{(t)}, x_{K(1)}^{(t)})$  와  $\theta_j^{(t+1)}$  가 주어졌을 때의  $\{x_{i(1)}^{(t+1)}\}$  의 posterior predictive distribution에서 draw한다.

X1	...	Xj	...	XK	Z
3	...	-	...	4	5
5	...	3	...	-	1
-	...	6	...	-	4
-	...	1	...	1	9

만약 변수들에 대한 conditional distribution들과 joint distribution이 잘 정의되어 있는 경우 이 알고리즘은 **Gibb's sampler**가 되고 수렴이 보장됨

### 3.3 Chained-Equation Multiple Imputation - 참고

[https://www.gerkovink.com/miceVignettes/Ad\\_hoc\\_and\\_mice/Ad\\_hoc\\_methods.html](https://www.gerkovink.com/miceVignettes/Ad_hoc_and_mice/Ad_hoc_methods.html)

```
> nhanes
  age bmi hyp chl
  1  NA NA NA NA
  2 22.7 1 187
  3  NA 1 187
  4 3  NA NA NA
  5 1 20.4 1 113
  6 3  NA NA 184
  7 1 22.5 1 118
  8 1 30.1 1 187
  9 2 22.0 1 238
 10 2  NA NA NA
 11 1  NA NA NA
 12 2  NA NA NA
 13 3 21.7 1 206
 14 2 28.7 2 204
 15 1 29.6 1  NA
 16 1  NA NA NA
 17 3 27.2 2 284
 18 2 26.3 2 199
 19 1 35.3 1 218
 20 3 25.5 2  NA
 21 1  NA NA NA
 22 1 33.2 1 229
 23 1 27.5 1 131
 24 3 24.9 1  NA
 25 2 27.4 1 186

> imp<-mice(nhanes,m=4,maxit=7) > result <- complete(imp,"broad")
> result
  iter imp variable  age.1 bmi.1 hyp.1 chl.1 age.2 bmi.2 hyp.2 chl.2 age.3 bmi.3 hyp.3 chl.3 age.4 bmi.4 hyp.4 chl.4
  1  1 1 26.3 1 118 1 30.1 1 238 1 27.2 1 131 1 35.3 1 186
  2  2 2 22.7 1 187 2 22.7 1 187 2 22.7 1 187 2 22.7 1 187 2 22.7 1 187
  3  1 3 29.6 1 187 1 30.1 1 187 1 30.1 1 187 1 30.1 1 187 1 30.1 1 187
  4  3 25.5 1 204 3 20.4 1 238 3 27.2 2 184 3 26.3 2 204
  5  1 20.4 1 113 1 20.4 1 113 1 20.4 1 113 1 20.4 1 113 1 20.4 1 113
  6  3 21.7 2 184 3 24.9 1 184 3 27.4 2 184 3 27.5 1 184
  7  1 22.5 2 187 1 22.5 1 118 1 22.5 1 118 1 22.5 1 118 1 22.5 1 118
  8  1 30.1 2 4 1 187 1 187 1 30.1 1 187 1 30.1 1 187 1 30.1 1 187
  9  2 22.0 3 1 238 2 22.0 1 238 2 22.0 1 238 2 22.0 1 238 2 22.0 1 238
 10  2  NA 3 2  NA 2 27.5 1 184 2 21.7 1 131 2 22.0 1 187 2 22.7 2 229
 11  1  NA 3 2 2  NA 1 27.4 1 238 1 24.9 1 113 1 27.2 1 131 1 28.7 1 131
 12  2  NA 3 3 2  NA 1 22.7 1 187 2 20.4 2 187 2 28.7 1 184 2 20.4 2 187
 13  3 21.7 3 4 1 206 1 206 3 21.7 1 206 3 21.7 1 206 3 21.7 1 206 3 21.7 1 206
 14  4 1 28.7 2 204 2 204 2 28.7 2 204 2 28.7 2 204 2 28.7 2 204 2 28.7 2 204
 15  4 2 29.6 1  NA 2 28.7 2 204 2 28.7 2 204 2 28.7 2 204 2 28.7 2 204
 16  4 3 29.6 1  NA 1 29.6 1 238 1 29.6 1 187 1 29.6 1 199 1 29.6 1 187
 17  5 1 29.6 1 284 1 29.6 1 131 1 28.7 2 113 1 27.4 1 187 1 22.0 1 238
 18  5 2 27.2 2 199 3 27.2 2 284 3 27.2 2 284 3 27.2 2 284 3 27.2 2 284
 19  5 3 26.3 2  NA 1 35.3 1 218 1 35.3 1 218 1 35.3 1 218 1 35.3 1 218
 20  5 4 26.3 2  NA 2 26.3 2 199 2 26.3 2 199 2 26.3 2 199 2 26.3 2 199
 21  6 1 25.5 1 229 3 25.5 2 204 3 25.5 2 199 3 25.5 2 229 3 25.5 2 229
 22  6 2 25.5 1 131 1 30.1 1 238 1 27.4 1 131 1 20.4 1 187 1 30.1 1 238
 23  6 3 25.5 1  NA 1 33.2 1 229 1 33.2 1 229 1 33.2 1 229 1 33.2 1 229
 24  6 4 25.5 1  NA 1 27.5 1 131 1 27.5 1 131 1 27.5 1 131 1 27.5 1 131
 25  7 1 25.5 1 186 1 24.9 1 204 3 24.9 1 218 3 24.9 1 186 2 27.4 1 186
 26  7 2 25.5 1 186 2 27.4 1 186 2 27.4 1 186 2 27.4 1 186 2 27.4 1 186
```

---

감사합니다