베이지안 통계 <기초>

- 평일 저녁 신촌역 편의점 아르바이트 상황에서

가장 편한 요일 고르기

기초

베이지안 통계 <기초>

박민지, 박채은,박희경, 배시예

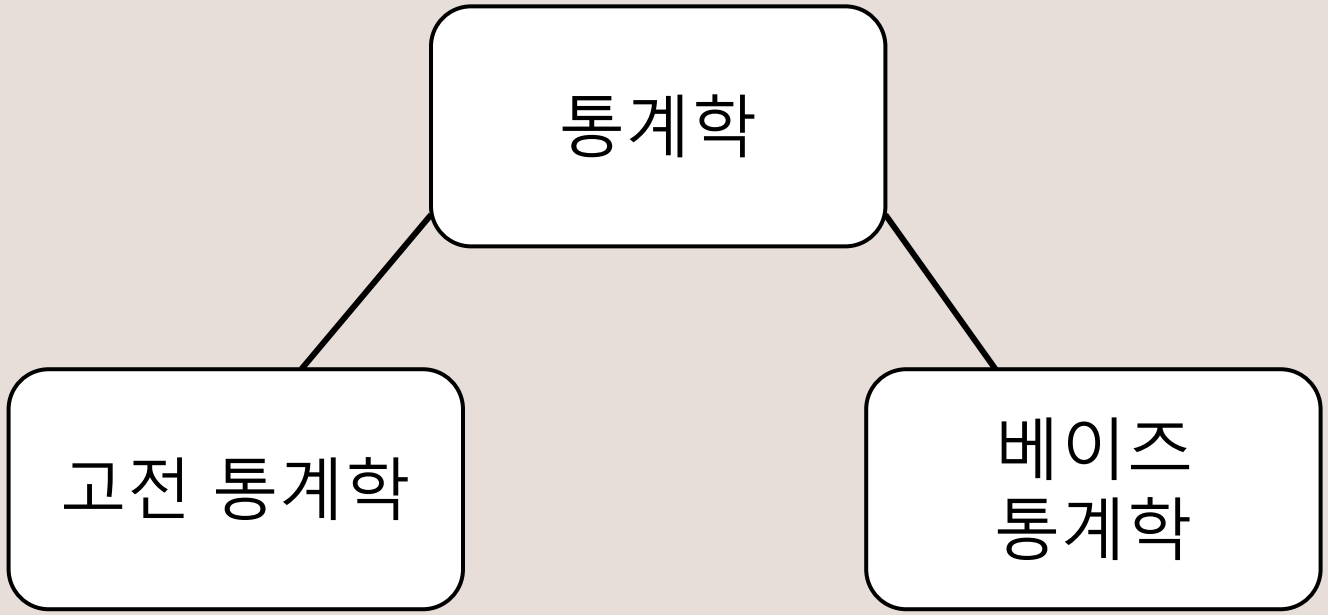
목차

The Table of Contents

베이즈 통계

Bayesian Statisttics

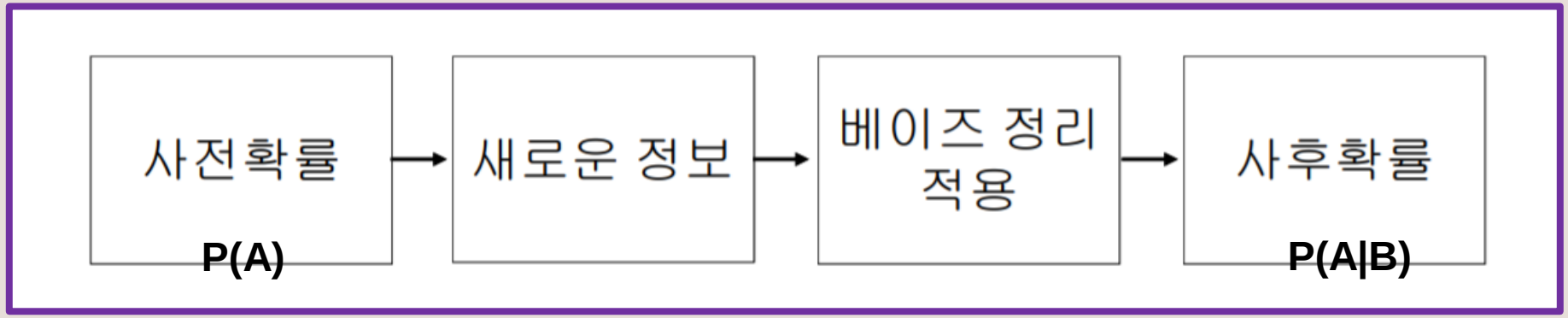
- 베이즈가 주창한 통계적 방법 => 고전 통계학과 구분됨.
- 표집에서 얻은 정보뿐만 아니라 연구자가 가지고 있는 사전 지식이나 주관적 의견 또는 신념과 같은 정보도 포함시키는 추리 통계의 한 방법



베이즈 정리

Bayes' Theorem

- 확률 변수의 사전 확률과 사후 확률 사이의 관계를 나타내는 정리
- "결과로부터 원인을 추론하기"



(ex) A: 병에 걸린 사건, B: 증상이 있을 사건
P(A) -> P(A|B) : 증상이 있을 때 병에 걸렸을 확률

베이즈 정리

Bayes' Theorem

베이즈 정리 계산하기

$$\Pr(A \mid B) = \frac{\Pr(B \mid A) \cdot \Pr(A)}{\Pr(B)}$$

$$= \frac{\Pr(B \mid A) \cdot \Pr(A)}{\Pr(B \mid A) \cdot \Pr(A) + \Pr(B \mid A^c) \cdot \Pr(A^c)}.$$

Bayes theorem for continuous distributions

continuous random variable θ

conditional density for θ given y

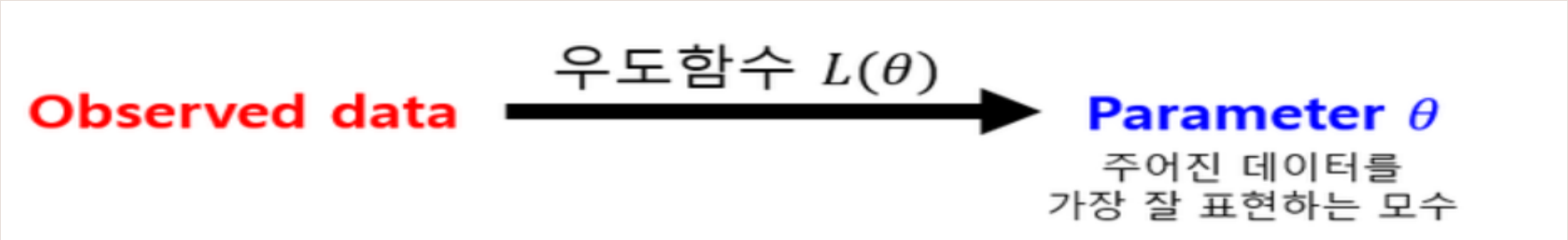
$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{\int f(y|\theta)f(\theta)d\theta}.$$

Likelihood

Likelihood

- **우도 (Likelihood) = P (확률분포 D / 관측값 X)**

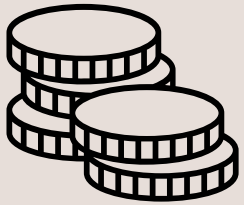
  (ex) 동전을 10번 던져서 앞면이 7번 나왔다. 이때의 모수(확률 분포)는?



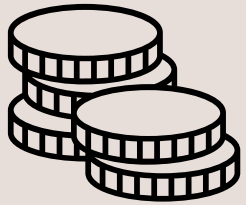
Prior & Posterior

Prior & Posterior

0. Example



50%



70%

- θ = { 50% 동전, 70% 동전}
- 결과는 이미 5번 던졌을 때 앞면 2번, 뒷면 3번으로 정해짐
- X ~ Bin (5, ?) (X = 동전을 던져 나온 앞면의 수)
- ‘동생은 이때까지 60%로 70% 동전을 썼어!’

Prior & Posterior

Prior & Posterior

1. Prior

- 사전확률 : 개인의 믿음, 사전 정보, 도메인 지식
- 주관적
- 0, 1로 두는 것은 좋지 않음

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{\sum_{i=1}^{m} P(B|A_i)P(A_i)}$$

(ex) 이전에 남동생이 60%로 70%동전을 사용했다는 나의 믿음

Prior & Posterior

Prior & Posterior

2. Posterior

- 사후확률
- Information in prior + Information in data
- 데이터가 매우 많다면, prior의 영향력 少

(ex) 사전 확률과 동생과의 내기결과를 바탕으로 새롭게 구한
70%동전을 사용했을 확률 분포

Predictive distribution

Predictive distribution

Prior predictive distribution(사전 예측 분포)

$$f(\tilde{x}) = \int f(\tilde{x}, \theta) d\theta = \int f(\tilde{x}|\theta) f(\theta) d\theta$$

Posterior predictive(사후 예측 분포)

$$f(\tilde{x}|x) = \int f(\tilde{x}, \theta|x) d\theta = \int f(\tilde{x}|\theta, x) f(\theta|x) d\theta$$

If x and $\tilde{x}$ are independent,

$$f(\tilde{x}|x) = \int f(\tilde{x}|\theta, x) f(\theta|x) d\theta = \int f(\tilde{x}|\theta) f(\theta|x) d\theta$$

MCMC

Markov Chain Monte Carlo

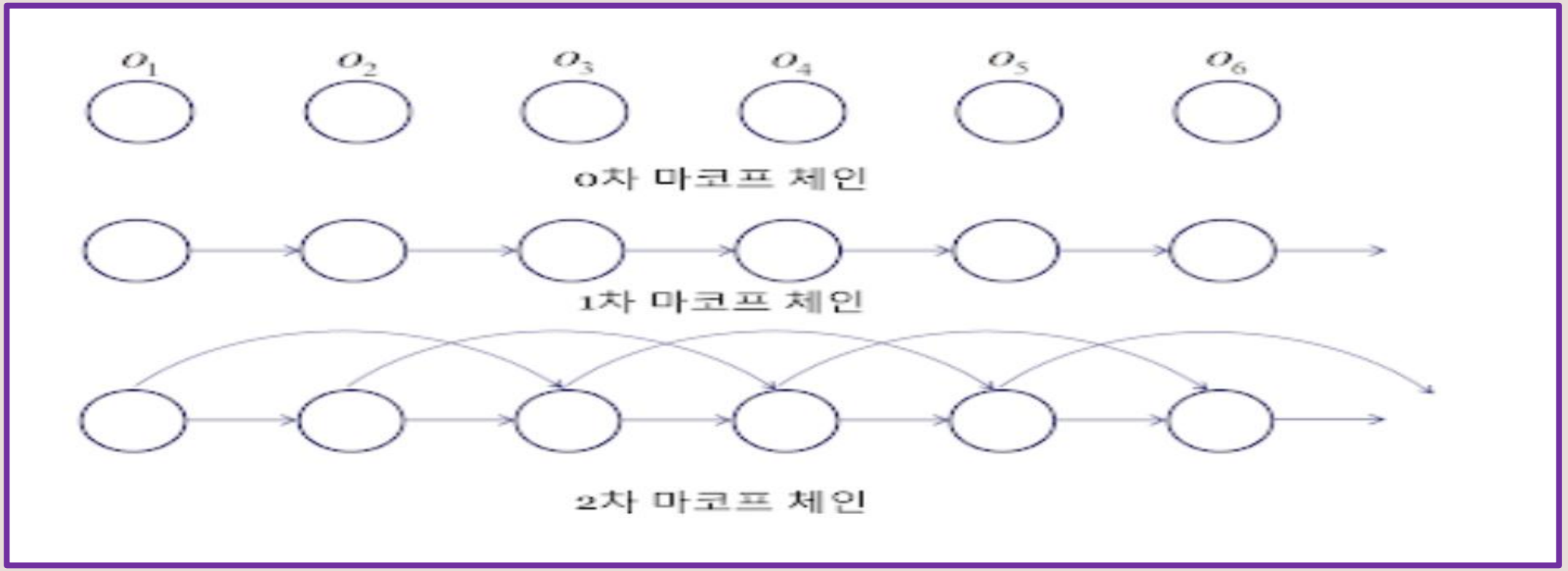
0. MCMC

마르코프 연쇄 몬테카를로 방법(Markov Chain Monte Carlo, MCMC)은 마르코프 연쇄의 구성에 기반한 확률 분포로부터 원하는 분포의 정적 분포를 갖는 표본을 추출하는 알고리즘의 한 분류이다.

MCMC

Markov Chain Monte Carloc

1. Markov Chain : 각 상태는 바로 이전의 상태에만 영향을 받는다.



MCMC

Markov Chain Monte Carloc

1. Markov Chain

- (가정1) Markov assumption: X_{t+1} 에서의 확률분포는 X_t에 의해서만 결정되어야 한다.

$$p(X_{t+1}|X_t, X_{t-1}, \ldots, X_2, X_1) = p(X_{t+1}|X_t) \quad \text{for all } t=2,\ldots,n$$

$$p(X_1, X_2, \ldots, X_n) = p(X_1) \cdot p(X_2|X_1) \cdot p(X_3|X_2, X_1) \cdot \ldots \cdot p(X_n|X_{n-1}, X_{n-2}, \ldots, X_2, X_1)$$

$$p(X_1, X_2, \ldots X_n) = p(X_1) \cdot p(X_2|X_1) \cdot p(X_3|X_2) \cdot p(X_4|X_3) \cdot \ldots \cdot p(X_n|X_{n-1})$$

- (가정 2) 전이 확률은 시간에 따라 바뀌지 않는다.

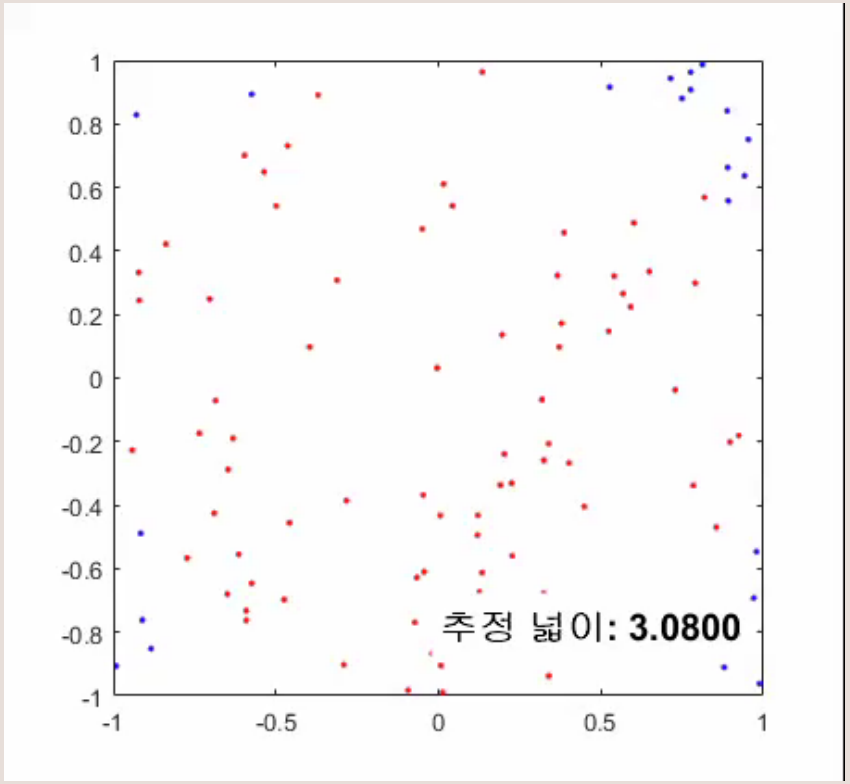
MCMC

Markov Chain Monte Carloc

2. Monte Carlo

무작위 추출된 난수(random number)를 이용하여 함수의 값을 계산하는 통계학적 방법

(ex)원을 넓이를 계산하는 시뮬레이션



Conjugate

Conjugate

- Prior와 Posterior가 같도록 설정하는 것
- 베이지안으로 posterior를 계산할 때,
  계산이 너무 복잡한 문제를 해결하기 위해 등장

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{\int f(y|\theta)f(\theta)d\theta}.$$

likelihood $p(x \mid \theta)$	conjugate prior $p_0(\theta)$	posterior $p(\theta \mid x)$
$\text{Normal}(\theta, \sigma)$	$\text{Normal}(\mu_0, \sigma_0)$	$\text{Normal}(\mu_1, \sigma_1)$
$\text{Binomial}(N, \theta)$	$\text{Beta}(r, s)$	$\text{Beta}(r + n, s + N - n)$
$\text{Poisson}(\theta)$	$\text{Gamma}(r, s)$	$\text{Gamma}(r + n, s + 1)$
$\text{Multinomial}(\theta_1, \ldots, \theta_k)$	$\text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$	$\text{Dirichlet}(\alpha_1 + n_1, \ldots, \alpha_k + n_k)$

Gibbs Sampling

Gibbs Sampling

1. Gibbs Sampling 이란?

- Initialize θ_0, φ_0
- For i=1, ... , m, repeat:

  1) using φ_{i-1}, draw $\theta_i \sim P(\theta|\ \varphi_{i-1}, y)$

  get $(\theta_i\ ,\ \varphi_{i-1})$

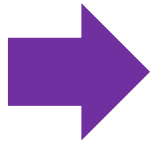
  2) using θ_i , draw $\varphi_i \sim P(\varphi|\ \theta_i\ , y)$

  get $(\theta_i\ ,\ \varphi_i)$

Gibbs Sampling

Gibbs Sampling

0. Set (x_0, y_0, z_0) to some starting value.
1. Sample $x_1 \sim p(x|y_0, z_0)$.
   Sample $y_1 \sim p(y|x_1, z_0)$.
   Sample $z_1 \sim p(z|x_1, y_1)$.
2. Sample $x_2 \sim p(x|y_1, z_1)$.
   Sample $y_2 \sim p(y|x_2, z_1)$.
   Sample $z_2 \sim p(z|x_2, y_2)$.
   $\vdots$



$$v^1 \mid v^2, v^3, \ldots, v^d$$
$$v^2 \mid v^1, v^3, \ldots, v^d$$
$$\vdots$$
$$v^d \mid v^1, v^2, \ldots, v^{d-1}$$

Gibbs Sampling

Gibbs Sampling

2. Gibbs Sampling의 특징

- (θ_1 , φ_1), (θ_2 , φ_2), (θ_3 , φ_3),..., (θ_m , φ_m) 은 dependent
- (θ_t , φ_t)는 $(\theta_{t-1}, \varphi_{t-1})$에 의존 : Markov chain
- Posterior distribution은 Markov chain의 stationary distribution
- (θ_m , φ_m)의 분포는 posterior에 approaches

Hierarchical Modeling

Hierarchical Model

0. Poisson Example

Hierarchical Modeling

Hierarchical Model

150개

$$y_i|\lambda \overset{iid}{\sim} Pois(\lambda) \qquad i = 1, 2 \cdots 150$$

위치 간의 잠재적인 차이점과
같은 위치의 쿠키가 서로 더 유사할 가능성이 있다는 사실을 무시함.

30개 X 5

$$y_i|\lambda_1 \overset{iid}{\sim} Pois(\lambda) \qquad i = 1, 2 \cdots 30$$
$$y_i|\lambda_2 \overset{iid}{\sim} Pois(\lambda) \qquad i = 1, 2 \cdots 30$$
$$y_i|\lambda_3 \overset{iid}{\sim} Pois(\lambda) \qquad i = 1, 2 \cdots 30$$
$$y_i|\lambda_4 \overset{iid}{\sim} Pois(\lambda) \qquad i = 1, 2 \cdots 30$$
$$y_i|\lambda_5 \overset{iid}{\sim} Pois(\lambda) \qquad i = 1, 2 \cdots 30$$

다른 위치의 데이터를 무시함.

Hierarchical Modeling

Hierarchical Model

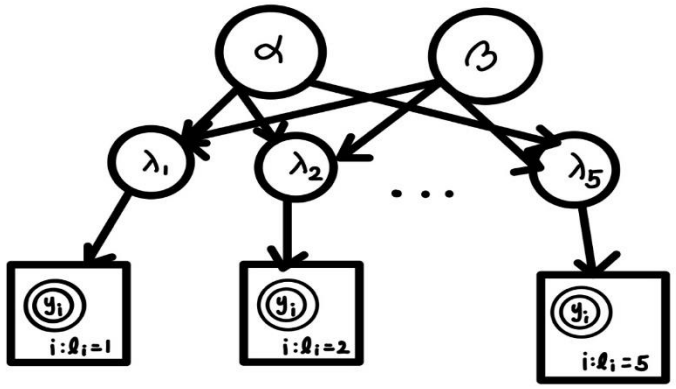
1. Hierarchical Modeling - Correlated Data

$$y_i | \ell_i, \lambda_\ell \overset{iid}{\sim} Pois(\lambda_{\ell_i}) \qquad \ell_i \in \{1.2.3.4.5\}$$

$$i = 1, 2, 3 \cdots, 150$$

$$\lambda_\ell | \alpha, \beta \overset{iid}{\sim} \Gamma(\alpha.\beta)$$

$$\alpha \sim p(\alpha) \quad \beta \sim p(\beta)$$



목차

The Table of Contents

1. 신촌역 세븐일레븐 17-19시,
평일 중 어떤 요일에 아르바이트를 하는 것이 좋을까?

- A요일이 B요일 보다 사람이 많을 확률
- 각 요일마다 18000명 이상일 확률

2. 사용한 데이터

- 서울교통공사 연도별 일별 시간대별 역별 승하차 인원
(2017년~2019년 데이터)
- Column: 날짜, 호선, 역번호, 역명, 승/하차 구분, 1시간 단위 승/하차 인원
- 전처리 후 사용한 데이터 총 739개

날짜	호선	역번호	역명	구분	06시 이전	06 ~ 07	07 ~ 08
2019-01-01	1호선	150	서울역	승차	348	321	348
2019-01-01	1호선	150	서울역	하차	222	821	808
2019-01-01	1호선	151	시청	승차	87	98	143
2019-01-01	1호선	151	시청	하차	48	237	323
2019-01-01	1호선	152	종각	승차	669	318	217
2019-01-01	1호선	152	종각	하차	68	179	293

3. 사용한 모델

- Hierarchical model, MCMC
- 승차 + 하차 인원(승객 수) ~ poisson
- Lambda의 분포 가정

Lambda: 요일 별 평균 승객 수

Uniform distribution

Gamma distribution

Exponential distribution

4. 분석: prior distribution

Uniform distribution

Gamma distribution

Exponential distribution

4. 분석: prior distribution

Uniform distribution

$$y_{i,j} \mid \lambda_j \overset{ind}{\sim} Pois(\lambda_j)$$

$$\lambda_j \mid \alpha, \beta \sim U(\alpha, \beta)$$

$j = 1, 2, 3, 4, 5$

$i = 1, \ldots, n$

$$mu \sim U(0, 1e6)$$

$$sig \sim U(0, 1e6)$$

$$\alpha = \frac{2mu - \sqrt{12sig^2}}{2}$$

$$\beta = \frac{2mu + \sqrt{12sig^2}}{2}$$

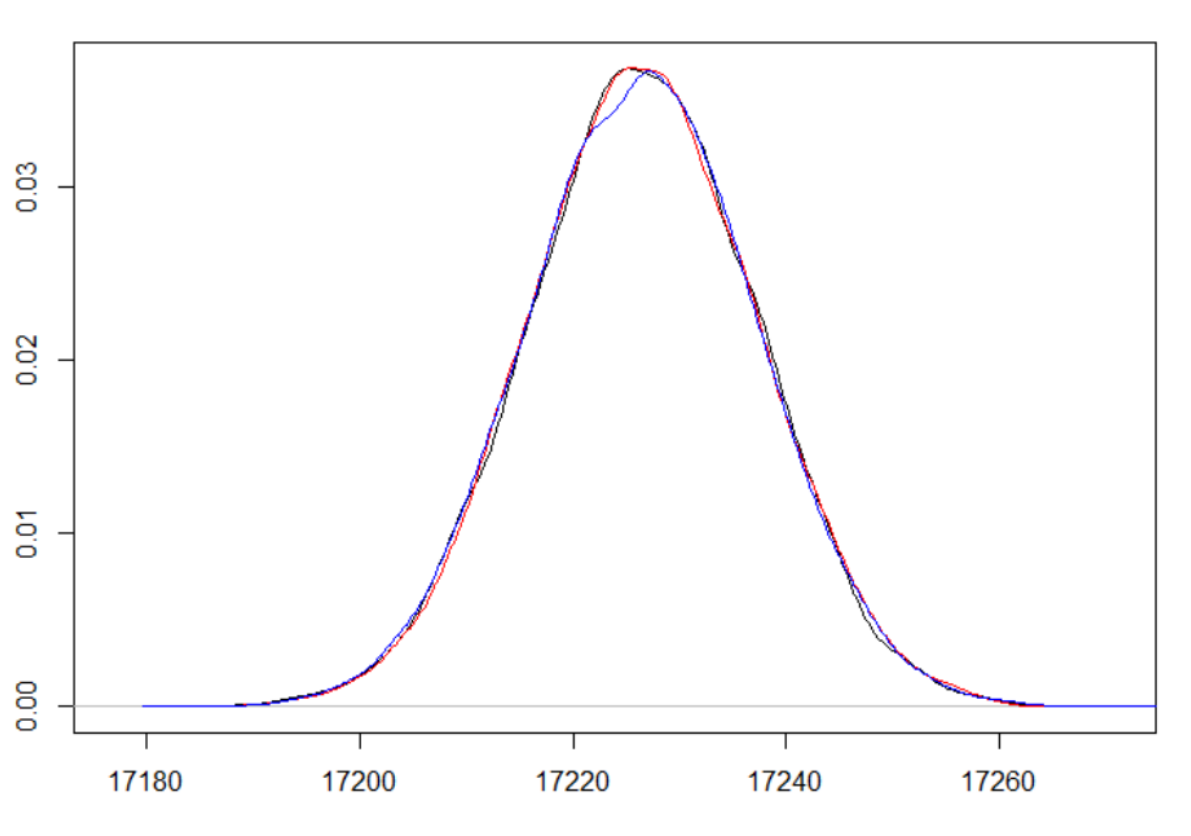
Exponential distribution

4. 분석: posterior distribution

Uniform distribution

Gamma distribution

Exponential distribution



Density of lam[1]

4. 분석: posterior distribution

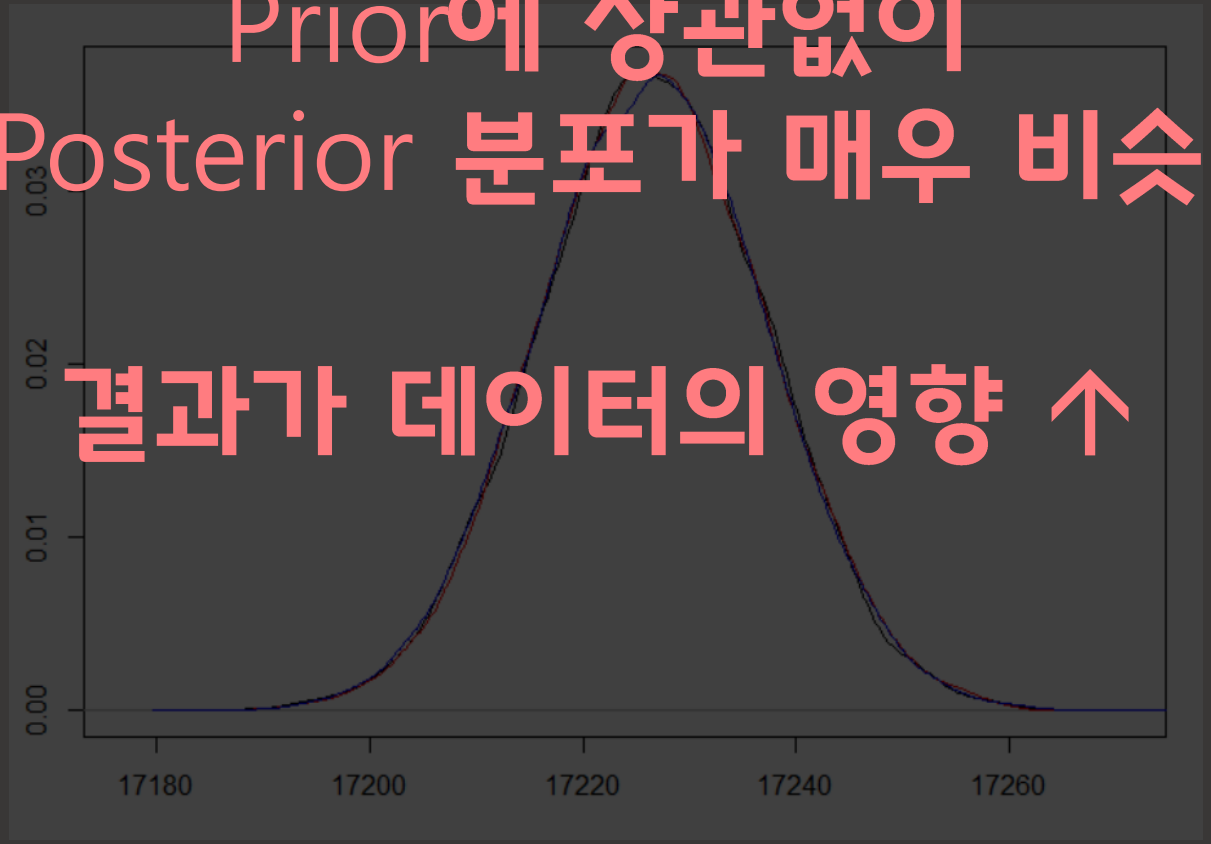
Uniform distribution

Gamma distribution

Exponential distribution

Prior에 상관없이
Posterior 분포가 매우 비슷

결과가 데이터의 영향 ↑



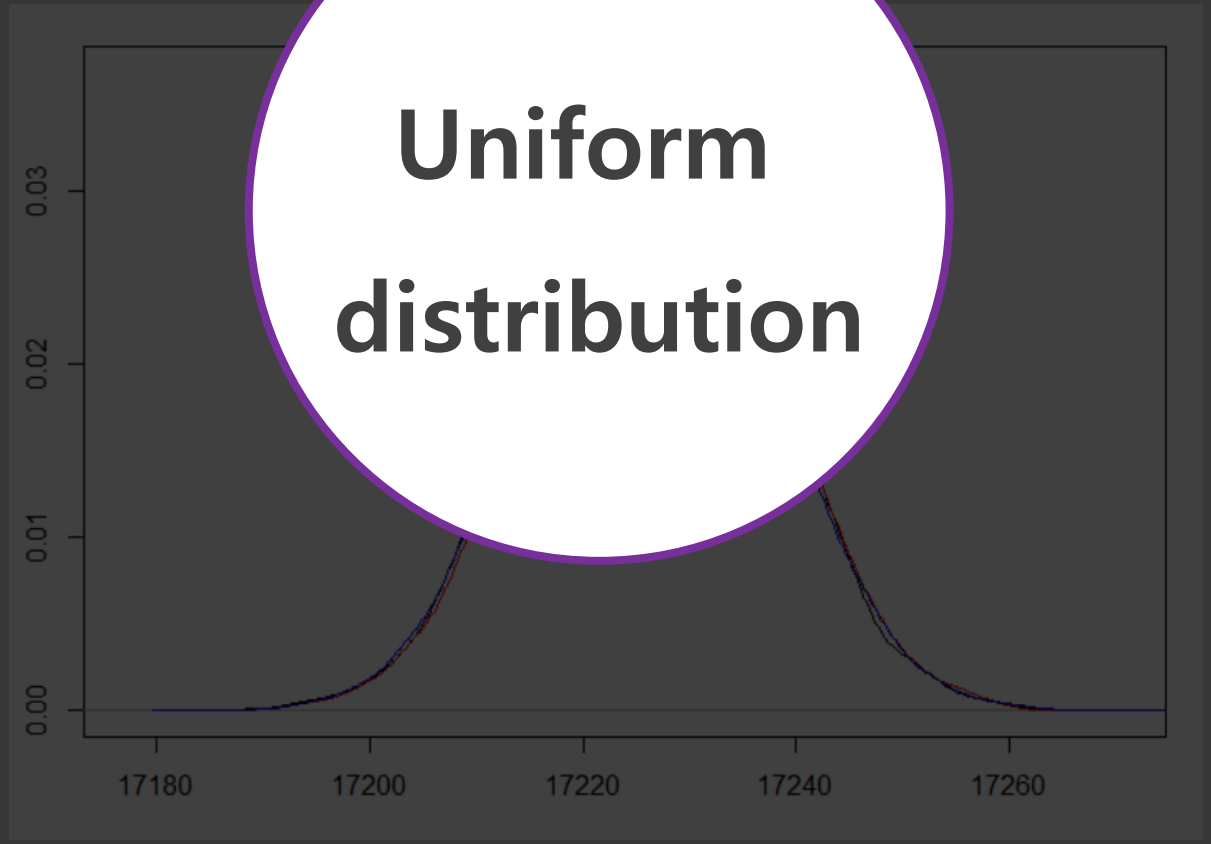
Density of lam[1]

4. 분석: posterior distribution

Uniform distribution

Gamma

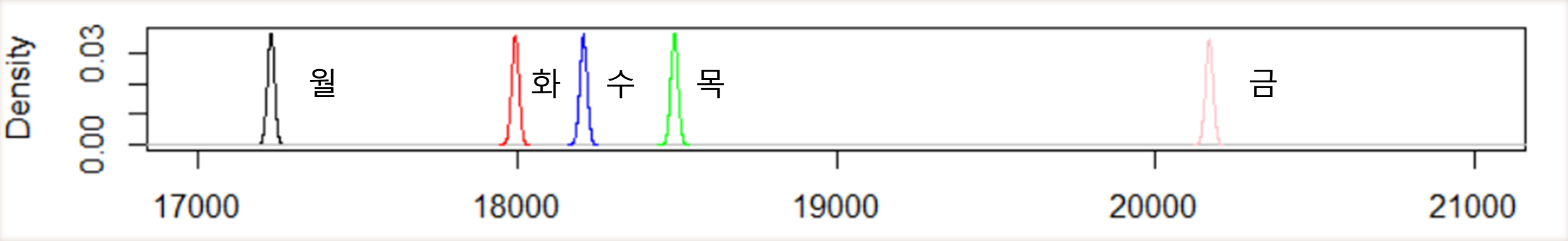
Exponential distribution



Uniform distribution

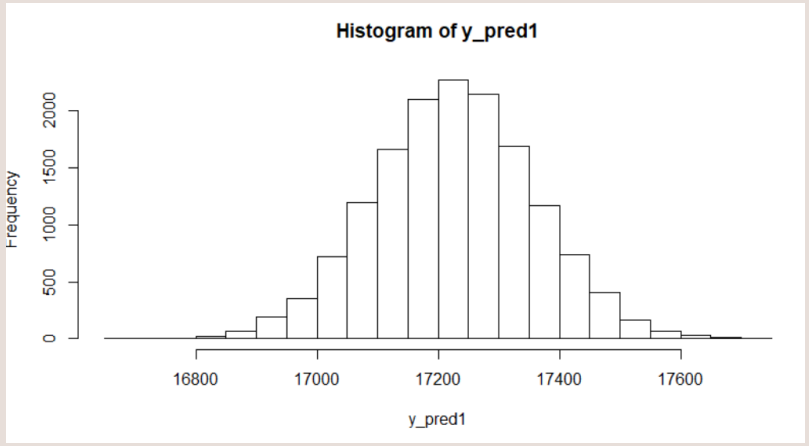
Density of lam[1]

4. 분석: 요일 별 lambda의 posterior distribution

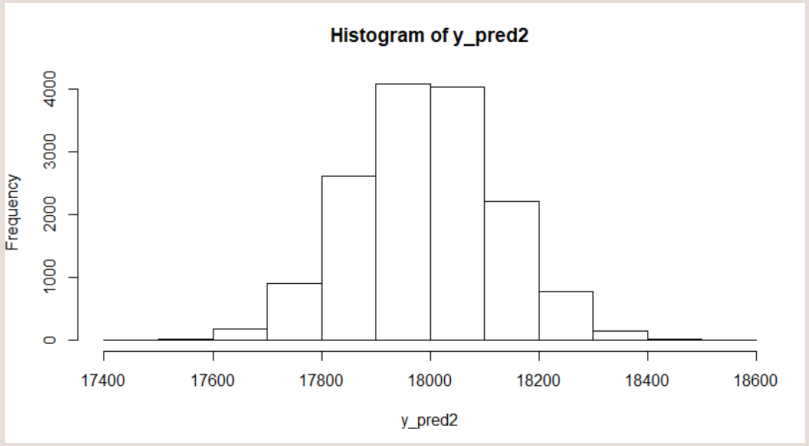


월요일에 평균 승객이 가장 적을 것이다.

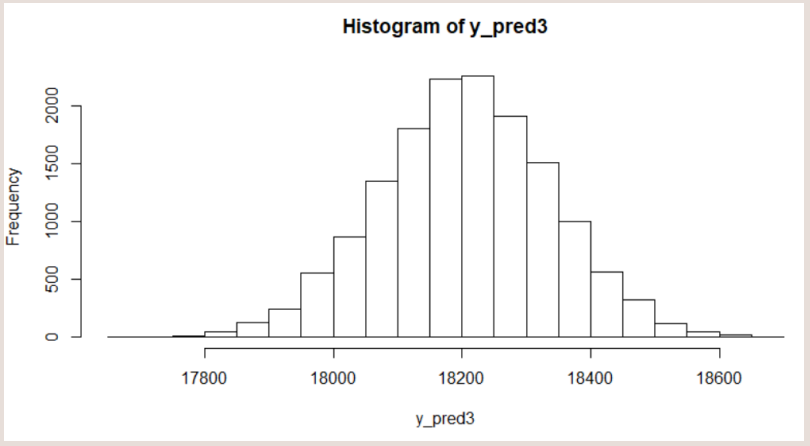
4. 분석: 요일 별 승객 수 예측



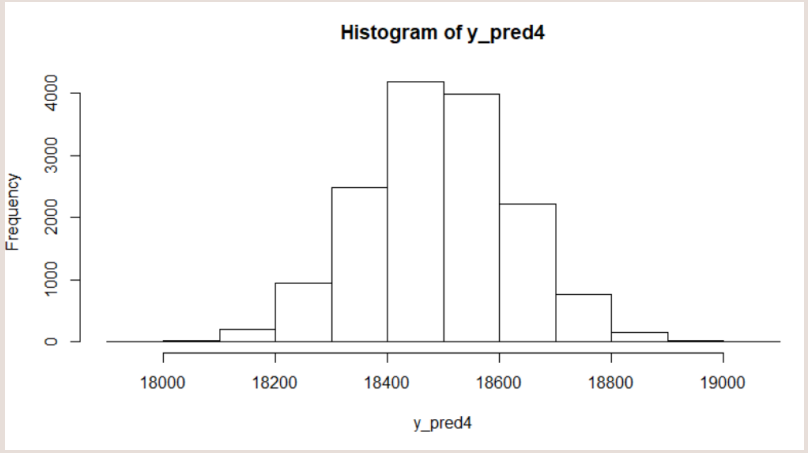
월



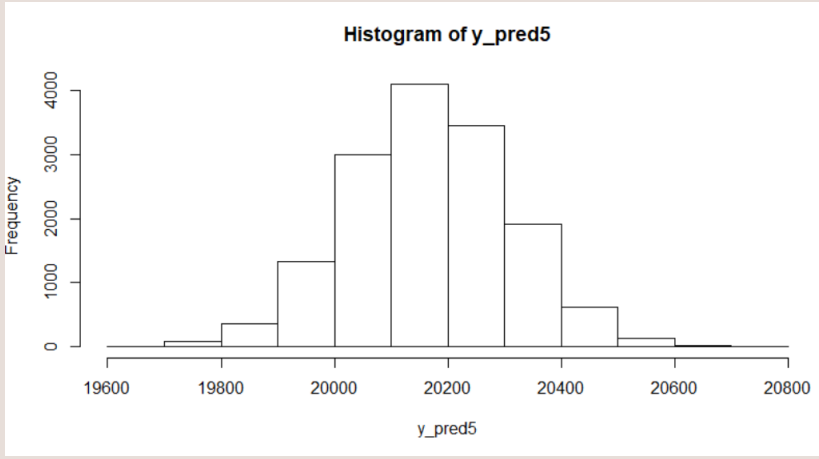
화



수



목



금

4. 분석: 요일 별 승객 수 비교

P(월<화) = 1
P(월<수) = 1
P(월<목) = 1
P(월<금) = 1

P(화<수) = 0.8661
P(화<목) = 0.9947
P(화<금) = 1

P(수<목) = 0.9365
P(수<금) = 1

P(목<금) = 1

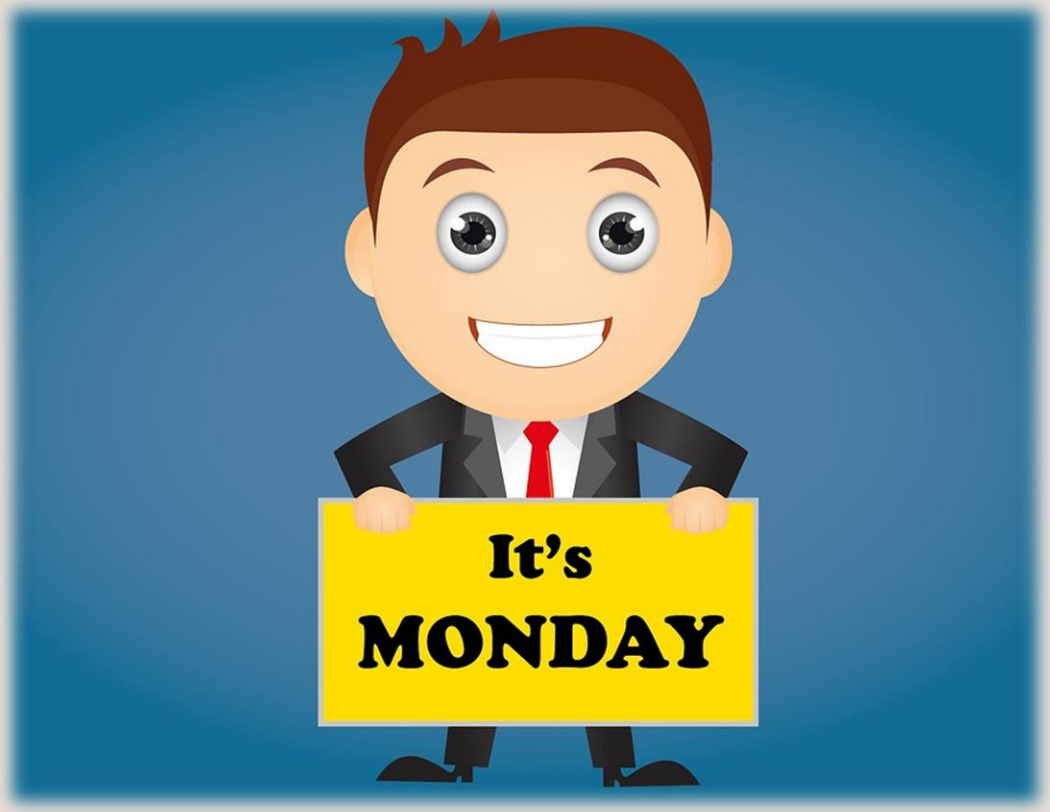
월요일 < 화요일 < 수요일 < 목요일 < 금요일

4. 분석: 요일 별 승객 수가 18000명 이상일 확률

각 요일에 18000명 이상이 지하철을 이용할 확률

월	0
화	0.4795
수	0.9339
목	0.9998
금	1

5. 결론: 어느 요일에 아르바이트를 하는 것이 가장 좋을까?



6. 한계점

- 데이터가 포아송 분포를 따른다는 가정
- 데이터 수의 부족

THANK YOU