

# 베이지안 통계학 PBL

---

정규분포 사후분포 예측

# 정규데이터

## 2-1 데이터, 분석목적, 필요성

- 데이터: 1. 국방부\_공군 신체측정정보(남) <655>  
1분기(3월 4일)<325> 2분기(6월 16일)<330> <공공데이터 포털>  
<배꼽수준 허리둘레, 엉덩이 둘레를 평균 낸 데이터 사용>
- 분석목적: 분기별 데이터에서 하의수치를 정하는 대표적인 수치(허리둘레, 엉덩이 둘레)로  
다음 분기 때 입대하는 사람들의 하의 수치를 대략적으로 파악.
- 필요성: 국방부 보급분야에서 어느 하의수치에 비중을 두고 보급을 해야 될 지 판단이 가능

배꼽수준 허리둘레	엉덩이 둘레	Average _m1
84	102.3	93.2
85.3	101.1	93.2
89.5	107.2	98.4
79.5	95.9	87.7
72.6	90.1	81.4
90.8	104	97.4
83.6	98.9	91.3
76.8	91.5	84.2
79.5	97.3	88.4
73.5	93.4	83.5
83.1	100.4	91.8
91.6	108.3	100

1분기 데이터

배꼽수준 허리둘레	엉덩이 둘레	Average _m2
67.5	89.1	78.3
75.7	93.9	84.8
85.2	104.3	94.8
78.5	96.6	87.6
74.2	93.1	83.7
80	102.9	91.5
76.9	97.2	87.1
95.3	106.4	100.9
92.6	108.4	100.5
90.2	108.9	99.6
74.7	93.5	84.1
84.2	94.5	89.4

2분기 데이터

# 정규데이터

## 2-2-1 정규성 검정

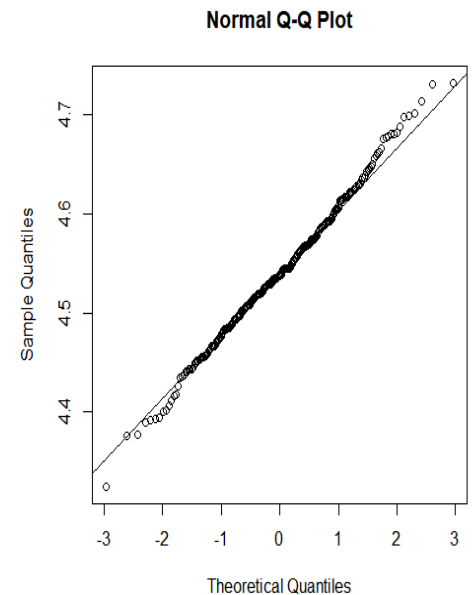
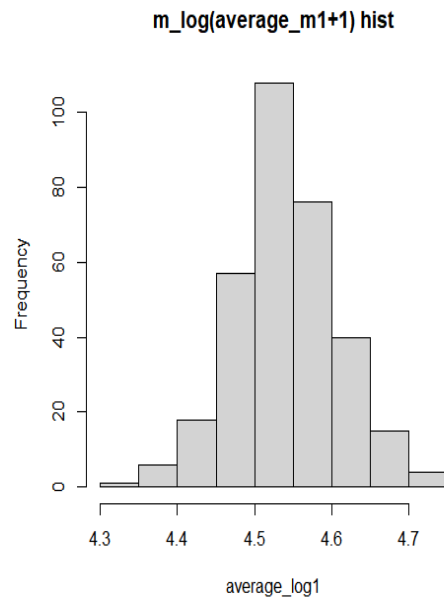
- 국방부\_공군 신체측정정보(남) 1분기(3월 4일)

Min. 1st Qu. Median Mean 3rd Qu. Max.  
74.50 88.80 92.60 93.02 96.80 112.60  
표준편차 6.34

### > 로그 변환 후 정규성 검정(Shapiro Test)

Shapiro-Wilk normality test

data: average\_m1(1분기)  
W = 0.99592, p-value = 0.5641



# 정규데이터

## 2-2-2 정규성 검정

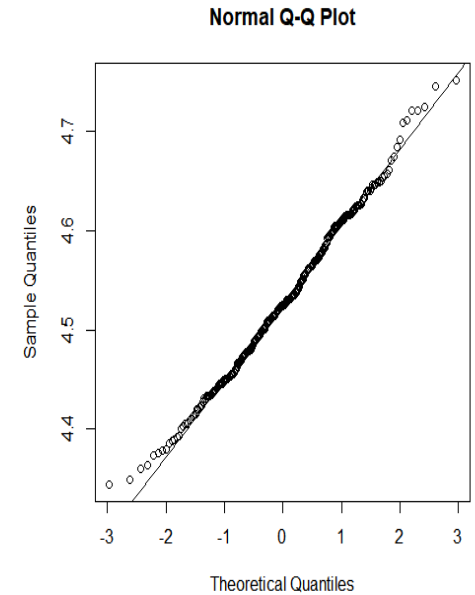
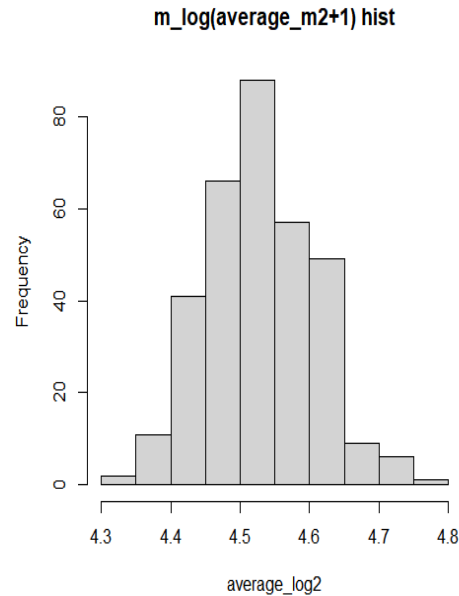
- 국방부\_공군 신체측정정보(남) 2분기(6월 16일)

Min. 1st Qu. Median Mean 3rd Qu. Max.  
76.00 86.70 91.25 91.79 96.38 114.80  
표준편차 7.2

### > 로그 변환 후 정규성 검정(Shapiro Test)

Shapiro-Wilk normality test

data: average\_m2(2분기)  
W = 0.99417, p-value = 0.2382



# 정규데이터

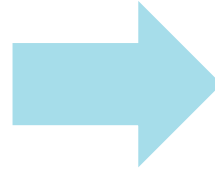
## 2-2-3 분석 과정

\* Log 변환 된 데이터들이 정규성을 따르므로 Log변환 된 값으로 분석 진행 \*

사전분포 1분기 데이터
사전분포 $\theta \sim N(\mu_0, \sigma_0^2) = N(4.53, 1.85^2)$

2분기 데이터 (현재 데이터로 사용) <330개의 관측>
$N = 330$ , (관측치) $\bar{x} = 4.52$ , $s = 1.97$

$\bar{X} \theta \sim N(\theta, s^2)$
$\bar{X} \theta \sim N(\theta, s^2/n) = N(\theta, 1.97^2/330)$ $\bar{x} = 4.52$



$$w_n = \left(1 + \frac{\sigma^2}{n\sigma_0^2}\right)^{-1} = \frac{\frac{1}{\sigma^2/n}}{\frac{1}{\sigma^2/n} + \frac{1}{\sigma_0^2}}$$

$$\mu_n = \frac{\bar{x} + \mu_0 \left(\frac{\sigma^2}{n\sigma_0^2}\right)}{1 + \frac{\sigma^2}{n\sigma_0^2}} = w_n \bar{x} + (1 - w_n) \mu_0$$

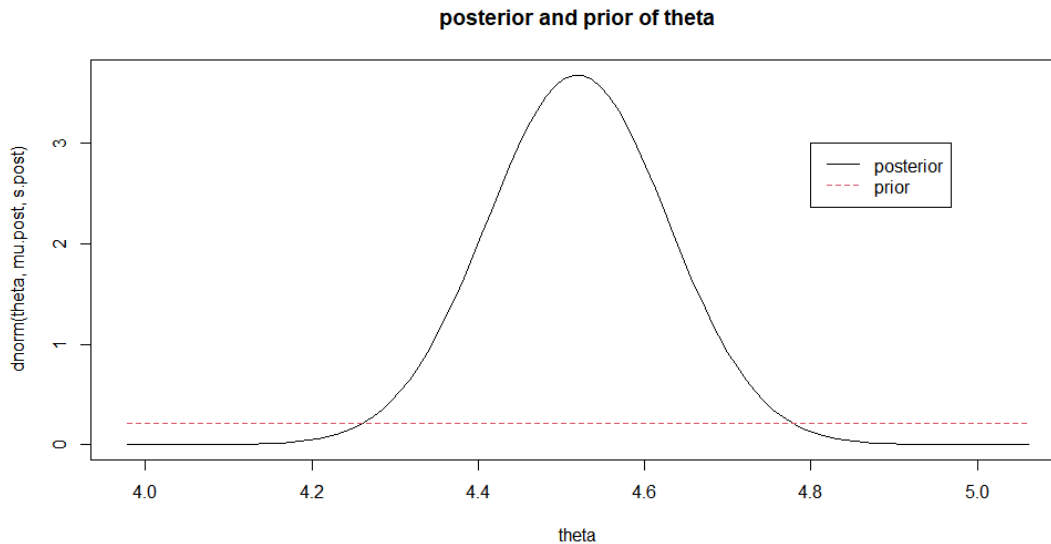
$$\sigma_n^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1} = w_n \cdot \frac{\sigma^2}{n}$$

\* 위 데이터로 3분기 사후분포 예측  $\rightarrow \theta \sim (\mu_n, \sigma_n^2) = (4.52, 0.11^2)$

# 정규데이터

## 2-2-4 분석 과정

- 사전분포(1분기 데이터 이용)  $\theta \sim N(\mu_0, s_0) = N(4.53, 1.85^2)$
- 사후분포(3분기 예측)  $\theta \sim (\mu_n, \delta_n^2) = N(4.52, 0.11^2)$
- 사전분포의 분산이 표본평균  $\bar{x}$ 의 분산에 비해 상당히 큰 값으로 사전분포의 영향이 미미하다.  
→ 그러므로 사전밀도함수는 거의 균일분포에 가깝다.



# 정규데이터

## 2-2-5 분석 과정

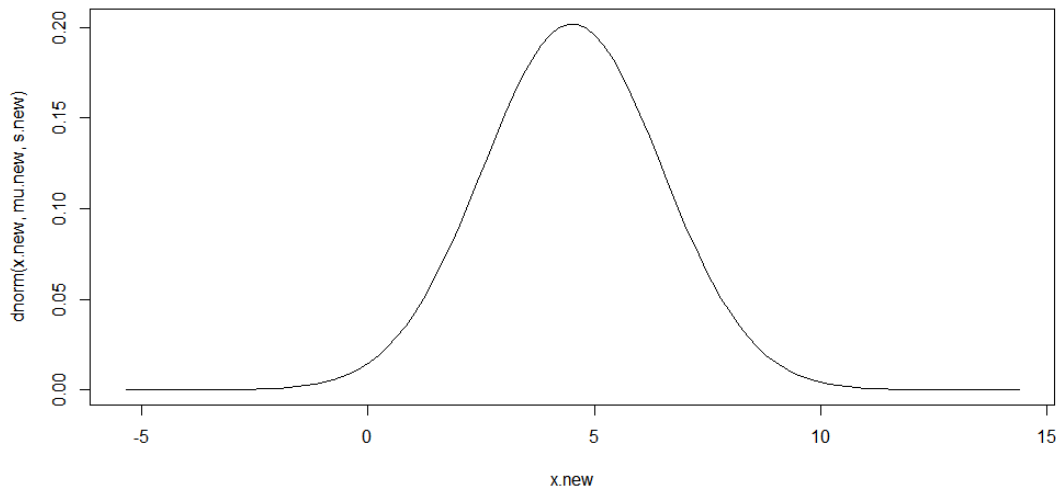
- 예측분포

예측분포 $\sim N(\mu_n, \delta_n^2 + s^2)$
$N(4.12, 0.11 + 1.97^2)$

$$E(X_{n+1}|x_1, \dots, x_n) = \mu_n = E(\theta|x_1, \dots, x_n)$$

$$\begin{aligned} \text{Var}(X_{n+1}|x_1, \dots, x_n) &= \sigma^2 + \sigma_n^2 = \text{Var}(X_{n+1}|\theta) + \text{Var}(\theta|x_1, \dots, x_n) \\ &\geq \text{Var}(\theta|x_1, \dots, x_n) \end{aligned}$$

predictive density of x(new)



# 정규데이터

## 2-2-6 분석 과정

- 예측분포(지수변환 한 값, 베이지안 추정치)

관측데이터(330)

$N(91.79, 7.20^2)$

사전분포  $\sim N(\exp(\mu_0), \exp(s_0))$

$N(93.02, 6.34^2)$

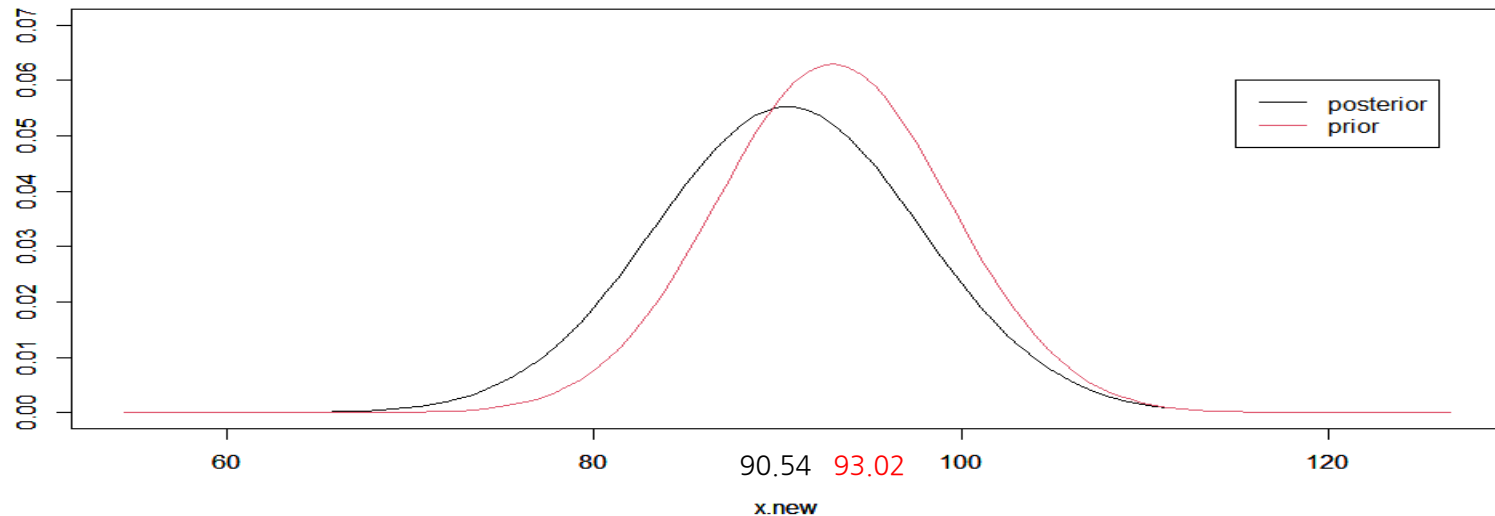


예측분포  $\sim N(\exp(\mu_n), \exp(\delta_n^2 + s^2))$

$N(90.54, 7.23^2)$

- 3분기 때는 하의 사이즈를 더 작게 보급해야 한다는 것을 알 수 있다.

predictive density of x(new) compare





# 정규데이터

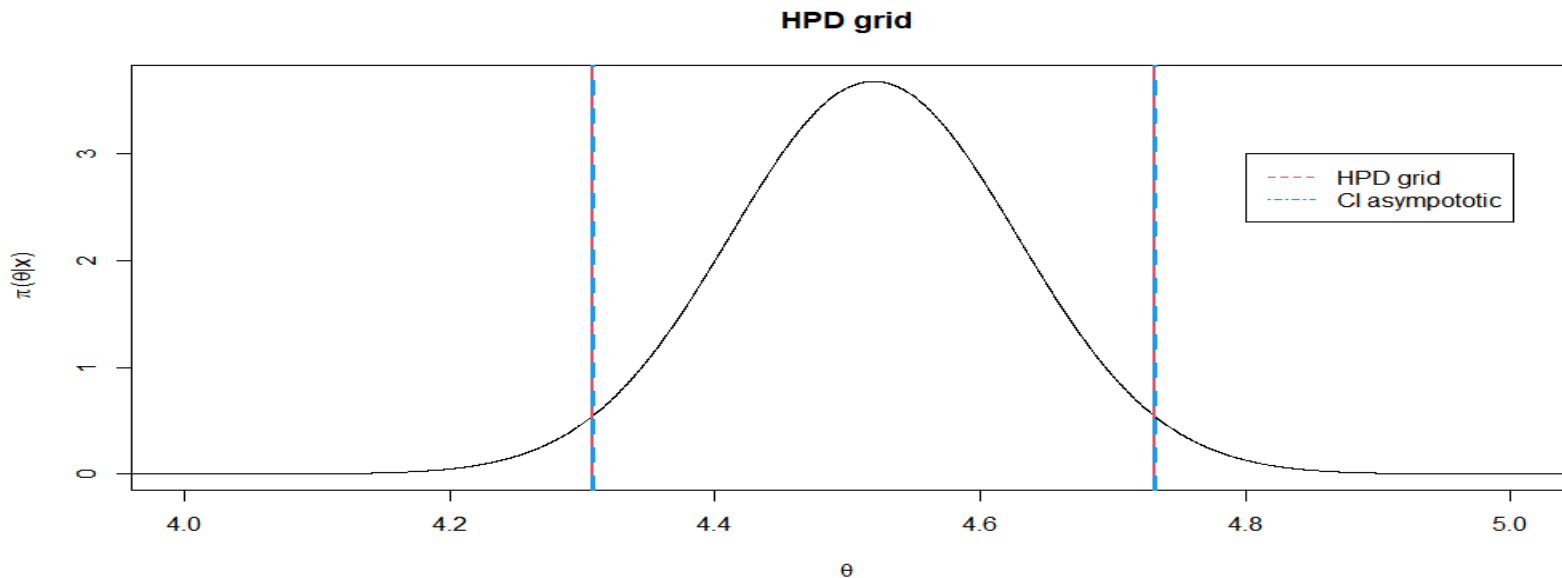
## 2-3-1 HPD Interval

- HPD Interval

격자점 이용한 HPD Interval
4.3072 4.7318

사후분위수 이용한 HPD Interval
4.307248 4.731949

대략 0.000048 0.000149 차이  
= 차이가 없다



# 정규데이터

## 2-3-2 HPD Interval vs 고전적 신뢰구간

- HPD Interval

격자점 이용한 HPD Interval
4.3072 4.7318

고전적 신뢰구간
4.307244 4.731953

사후분위수 이용한 HPD Interval
4.307248 4.731949

정규분포의 모형을 따르므로 고전적 신뢰구간하고 같다.

