베이지안을 이용한 정규분포 데이터 예측

백대환[1], 정규분포 베이지안 예측

요약

정규분포 - 정규분포를 따르는 분기별 하의 수치 데이터 집단을 사용하여 다음의 분기에 하의 수치를 예측하는 베이지안 추론을 하였다. 본 논문에서는, 정규사전분포와 관측데이터의 정규분포를 이용하여 정규사후분포와 예측분포를 구하였다. 분석하는 과정을 통해 보급분야에 기여가 되는 결론을 도출해냈다.

주요용어 : 정규분포, 격자점, 사후분위수, 우도함수, 베이지안

1. 서론

분석하고자 하는 연구의 배경과 목적을 서술

정규분포 - 과거 자료 및 현재 자료를 가지고 베이지안 추론 방법을 이용하여 추정하고 분석을 한 후에 서로 그 결과를 비교하여 판단을 해보는 과정을 거쳤다. 본 논문에서는 하의 수치 1분기 데이터의 분포인 정규분포를 사전분포로 정의를 하였고, 2분기 데이터도 정규분포 자료이고 이 데이터도 사용하여 분석을 실시했다. 사후분포 구하는 식을 이용하여 사후분포를 구하였고, 사후분포를 이용해 예측분포 및 최대사후구간(격자점, 사분위수)를 구하였고, 고전적 신뢰구간과 비교를 해보았다. 예측분포를 이용해 결론을 도출해 내면서 논문을 마무리한다.

2. 데이터

2.3 정규 데이터

분석에 사용되는 데이터를 설명한다.

데이터는 공공데이터 포털에서 가져온 국방부_공군_신체측정정보(1분기, 2분기)를 이용하였다. 이 데이터 안에서 배꼽 수준 허리둘레, 엉덩이둘레를 평균을 낸 하의 수치 데이터를 사용하였다. 1분기, 2분기 데이터가 Shapiro-test를 하였을 때, p-value가 0.05를 넘지 않아 정규성을 가지지 못했고, 정규성을 가지기 위한 과정으로 log변환(data <- log(data+1))을 진행하였다. 진행 후 Shapiro-test결과 두 데이터 셋의 p-value 모두 0.05를 넘겼다. 그렇기 때문에 log변환 된 데이터를 이용해 분석을 진행하였다. 분석을 진행하고 최종적으로 보일 때에는 지수변환(exp)를 이용해 값을 나타냈다.

3. 분석 모형

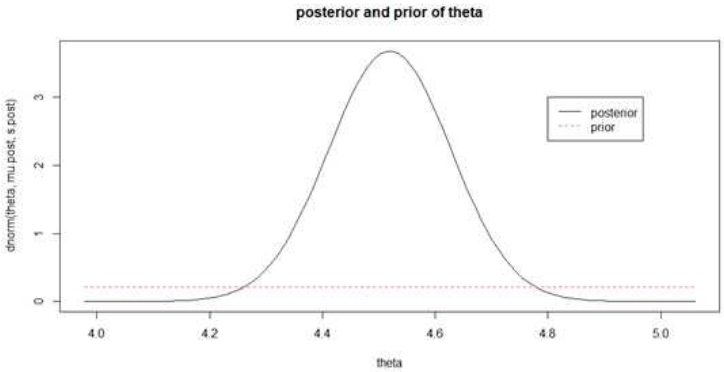
3.3 정규분포에 대한 베이지안 추론

사후분포와 예측분포를 구하기 위해서 사전분포로는 1분기 데이터의 분포를 이용하였다. 1분기 데이터의 분포는 정규분포를 따른다($\theta \sim N(\mu_0, \sigma_0^2)$). 2분기 데이터는 추가된 관측 데이터(2분기 데이터)로 330개의 데이터가 있고, 마찬가지로 정규분포를 따른다(($\theta \sim N(\bar{x}, \sigma^2)$). 위 두 데이터에서 나온 수치로 아래의 식에 대입해 사후분포를 구한다($N(\mu_n, \sigma_n^2)$).

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

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

사전분포의 분산이 표본평균 $\bar{x}$의 분산(N으로 나눈 값)에 비해 상당히 큰 값으로 사전분포의 영향이 미미해진다. 그러므로 사전밀도함수는 거의 균일분포에 가까워지게 된다. 아래는 사전분포(빨간색), 사후분포(검은색) 그래프로 나타내 보았다.



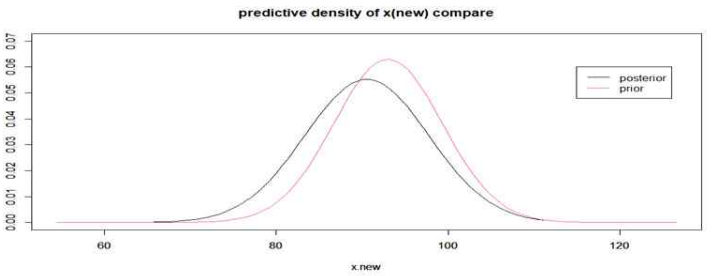
구해진 사후분포의 평균과 분산을 아래의 식에 대입하여 예측분포를 구하는 과정을 진행했다.

$$E(X_{n+1}|x_1, \ldots, x_n) = \mu_n = E(\theta|x_1, \ldots, x_n)$$
$$Var(X_{n+1}|x_1, \ldots, x_n) = \sigma^2 + \sigma_n^2 = Var(X_{n+1}|\theta) + Var(\theta|x_1, \ldots, x_n)$$
$$\geq Var(\theta|x_1, \ldots, x_n)$$

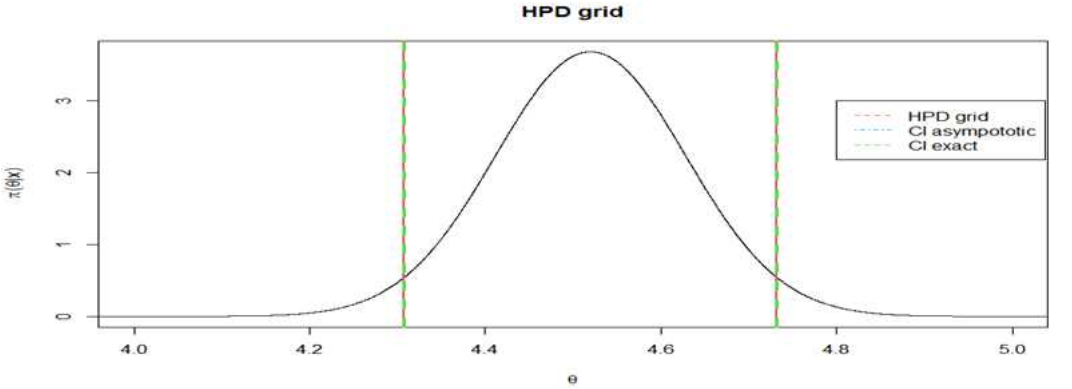
구해진 예측분포(log된 값)을 실제 수치로 보여주기 위해 지수변환(exp)를 해주었다. 이후, 사후분포에 대해 격자점, 사후분위수 이용한 최대사후구간을 구했고, 고전적 신뢰구간과 비교하였다.

4. 분석 결과

4.3 정규분포에 대한 베이지안 추론



왼쪽 그림은 사전분포와 예측분포의 그래프이다(빨간색 예측분포, 검은색 사전분포). 사전분포와 관측데이터 330개를 이용해 모형을 만들어 사후분포와 예측분포(베이지안 추정치)를 예측하였다. 예측분포의 평균값은 사전분포보다 작게 나타나고, 분산 값은 더 크게 나타난다. 그 이유는 추가된 관측 데이터(2분기) 330개의 분포가 사전분포보다 평균값이 더 작게 나타났고, 분산 값이 더 크게 나타났기 때문이다. 여기서 예측분포는 3분기를 예측한 값이고 3분기에 들어오는 군인의 하의수치는 더 작을 것이라고 판단이 가능하다.



위 그림은 격자점과 사후 분위수를 이용한 최대사후구간과 고전적 신뢰구간을 비교한 그래프이다. 빨간색, 파란색, 초록색 점선이 겹쳐있는 것을 볼 수 있다. 모두 95%신뢰구간을 기준으로 구하였고, 최대사후구간을 구한 값이랑 고전적 신뢰구간을 구한 값이 거의 같은 값으로 나타났다. 사후분포가 정규분포를 따르고 좌우대칭이기 때문에 위와 같은 결과가 나타났다.

5. 결론

보고서의 결론을 서술한다.

1분기 데이터(사전분포)와 2분기 데이터(관측데이터)를 이용하여 3분기 예측분포(베이지안 추정치)를 예측해본 결과, 1분기에 비해 더 작아진 값으로 나오는 것을 볼 수 있다. 이는 관측데이터(2분기 데이터)가 더 작은 값을 가지고 있었기 때문이다. 결과적으로, 1분기, 2분기 데이터를 이용해 3분기에는 하의 수치가 더 작게 예측이 됐고, 3분기에 하의를 보급할 때 2분기보다 더 작은 사이즈의 하의를 보급해야 할 것이라고 판단이 가능하다.

참고문헌

혼합 정규분포의 베이지안 ROC 곡선 추정. Bayesian ROC curve estimation with a normal mixture distribution.(2009). 박주원 (한양대학교 대학원 응용수학과 국내석사)