확률통계 및 공통 알고리즘

4. 통계적 추론

임채영

서울대학교

이번강의에서 다룰 내용

- 통계적추론의 개념
- 점추정의 개념, 추정량의 평가, 구간추정의 개념
- 최대가능도법
- 베이지안 추론
- 유의성 검정

통계적 추론 (Statistical Inference)

- 표본으로부터의 정보를 이용하여 모집단에 관한 추측이나 결론을 이끌어내는 과정
- 추정(Estimation)
- 유의성 검정(Significance test, 또는 가설 검정(Hypothesis test))

추정 (Estimation)

표본으로부터 모집단의 특성값(모수)에 대한 추측값과 오차를 제시

- 모수(Population parameter, θ) : 모집단의 특징을 나타내는 대표값 (예 : 모평균 μ, 모분산 σ^2)
- 랜덤표본 : 서로 독립이고 동일한 확률분포를 따르는 확률변수들을 말하며, 실제로 표본을 추출하여 얻은 값들을 관측값(Observation)이라 한다.

모수의 추정 (Parameter estimation) - 점추정 (Point estimation), 구간추정 (Interval estimation)

점추정(Point estimation)

표본으로부터 계산한 모수의 추정값을 제시

추정량의 예

- 모평균의 추정량 : 표본평균 $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- 모분산의 추정량 : 표본분산 $\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

추정량의 평가

추정량(Estimator)을 평가하는 몇 가지 기준이 있다.

(1) 편향 (Bias):

- Bias ($\hat{\theta}$) = $E(\hat{\theta}) - \theta$
- 불편추정량 (Unbiased estimator)

  $E(\hat{\theta}) = \theta$를 만족하는 추정량 $\hat{\theta}$

  예 : 표본평균과 표본분산은 각각 모평균과 모분산의 불편추정량이다.

(2) 표준오차 (Standard error), $SE(\hat{\theta})$

- 추정량의 표준편차

  예 : $X_1, \cdots, X_n \overset{i.i.d.}{\sim} (\mu, \sigma^2)$,

$$se(\hat{\mu}) = sd(\hat{\mu}) = \sqrt{var(\bar{X})} = \frac{\sigma}{\sqrt{n}}$$

- $MSE(\hat{\theta})$: Bias와 SE를 동시에 고려한 평가 기준

$$\begin{aligned} MSE(\hat{\theta}) &= Var(\hat{\theta}) + (Bias(\hat{\theta}))^2 \\ &= (SE(\hat{\theta}))^2 + (Bias(\hat{\theta}))^2 \end{aligned}$$

구간추정 (Interval estimation)

- 모수를 추정값을 구간으로 제공
- 구간추정의 일반적 방법: 신뢰구간 (Confidence Interval, CI)
- 신뢰수준 (Confidence level)이 $100(1-\alpha)\%$인 신뢰구간 (L, U)는 다음을 만족한다.

$$P(L \leq \theta \leq U) = 1 - \alpha$$

  - L, U는 표본으로부터 구해짐. 즉, $L \equiv L(X_1, \cdots, X_n)$, $U \equiv U(X_1, \cdots, X_n)$
  - 따라서 (L, U) 는 확률 변수로 이루어진 구간 (random interval)
- $1-\alpha$ 는 포함확률(coverage probability)이라고 부름

신뢰구간 예시

모분산 σ^2를 알 때 정규모집단의 모평균 μ의 구간추정

- 신뢰수준 $100(1-\alpha)\%$인 μ의 신뢰구간

$$\left(\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\ \bar{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

- $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$로부터

$$\begin{aligned} 1-\alpha &= P\left(-Z_{1-\frac{\alpha}{2}} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le Z_{1-\frac{\alpha}{2}}\right) \\ &= P\left(\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

- 오차의 한계: 신뢰구간에서 허용하는 가장 큰 오차. 신뢰구간 길이의 $\frac{1}{2}$. 즉, $Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$

신뢰구간의 의미

μ의 100(1-α)% 신뢰구간 : 100번의 표본 추출을 통해 얻어진 100개의 신뢰구간

$$\left(\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

에서, 100(1-α)% 개 정도의 신뢰구간이 모평균을 포함할 거라 기대함

- $n = 25$, σ = 10일 때 모평균의 90% 신뢰구간
- $(\bar{X} - Z_{0.95}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{0.95}\frac{\sigma}{\sqrt{n}})$
  $= (\bar{X} - 1.645\frac{10}{\sqrt{25}}, \bar{X} + 1.645\frac{10}{\sqrt{25}})$ $(\because Z_{0.95} \simeq 1.645)$
- 표본을 새로 추출할 때마다 $\bar{X}$가 달라지므로 신뢰구간 역시 길이를 유지한 채 표본에 따라 값들이 달라짐

오차의 한계와 표본의 크기

- 오차의 한계가 일정수준 이하가 되게 하기 위해 필요한 표본의 수를 계산
- 모분산을 알고 있는 경우의 정규모집단의 모평균 신뢰구간의 오차의 한계: $Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
- 오차의 한계를 d 이하로 하는 n :

$$
\begin{aligned}
Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} &\leq d, \\
\left(\frac{Z_{1-\frac{\alpha}{2}} \sigma}{d} \right)^2 &\leq n
\end{aligned}
$$

예제

중앙아메리카의 저소득층 원주민을 대상으로 49 명의 표본조사를 한 결과 혈청 내의 콜레스테롤 양이 평균 157.02(mg/L) 이었다고 한다. 이들 원주민 전체에서 혈청 내의 콜레스테롤 양이 정규분포이고, 표준편차가 30(mg/L)라고 할 때, 원주민 전체에서 혈청 내의 콜레스테롤 양의 평균에 대하여 95% 신뢰구간을 구해보자. 그리고, 모평균 μ에 대한 95% 추정의 오차 한계를 5(mg/L) 이하로 하기를 원한다면, 몇 명 이상의 원주민을 표본으로 조사해야 하는지 구해보자. (단, $Z_{0.975} = 1.96$)

최대 가능도 추정법 (Maximum Likelihood Estimation)

가능도 (likelihood)를 최대가 되게 하는 모수 값을 그 추정량으로 하는 방법

- 가능도:
  - 확률변수의 관측값이 주어졌을때 해당 확률의 정도
  - 모수를 가지는 확률 분포의 경우 모수값이 가능한 정도
  - 일반적으로, 확률밀도함수에 확률변수의 관측값을 대입한 값

- 예제: $Ber(p)$를 따르는 확률변수 X_1, X_2, X_3가 독립적으로 다음과 같이 관측되었다고 하자. $X_1 = 1, X_2 = 0, X_3 = 0$.
  - 결합 확률질량함수에 관측값을 대입한 값은 $p(1,0,0) = p_1(1)p_2(0)p_3(0) = p \times (1-p) \times (1-p)$
  - 따라서 성공확률 p에대한 가능도는 $L(p) = p(1-p)^2$, $0 \le p \le 1$.
- 일반적으로, 모수 θ에 대한 가능도함수는

$$L(\theta) = L(\theta|X_1, \cdots, X_n) = f(X_1, \cdots, X_n; \theta)$$

- 이산확률변수인 경우 가능도는 관측값이 주어졌을때의 해당 확률이 된다.
- 확률변수들이 독립적으로 관측되었을 경우 가능도는 다음과 같이 주변확률분포의 곱으로 표현.

$$L(\theta) = L(\theta | X_1, \cdots, X_n) = \prod_{i=1}^{n} f(X_i; \theta)$$

최대가능도 추정의 원리 - 예

- 공장 A의 한 생산라인의 불량률 (p)을 추정하기 위해 10개의 제품을 랜덤 추출하였다고 하자. 불량인 경우를 1로 할 경우, 10개의 제품의 불량정보는 $(0, 0, 0, 1, 0, 1, 0, 0, 0, 0$이라고 하자.
- 각 제품의 불량 유무를 1또는 0을 가지는 확률변수 X로 봤을때 $X_1, \cdots, X_{10} \overset{i.i.d.}{\sim} Ber(p)$

- 불량률 p에 대한 가능도: $L(p) = p^2(1-p)^8, 0 \le p \le 1$

  즉, 불량률이 p일때 $(0, 0, 0, 1, 0, 1, 0, 0, 0, 0)$를 관측할 확률 (정도)
- $L(0.5) = 0.5^2 0.5^8 = 0.00098$: 불량률이 $p = 0.5$일때 $(0, 0, 0, 1, 0, 1, 0, 0, 0, 0)$를 관측할 확률(정도)
- $L(0.2) = 0.2^2 0.8^8 = 0.0067$: 불량률이 $p = 0.2$일때 $(0, 0, 0, 1, 0, 1, 0, 0, 0, 0)$를 관측할 확률(정도)
- $(0, 0, 0, 1, 0, 1, 0, 0, 0, 0)$를 관측할 확률(정도)를 최대가 되게 하는 p

최대가능도 추정의 원리

- 주어진 관측값에서 모수의 가능도를 최대가 되게 하는 값으로 모수를 추정하는것이 최대가능도 추정법(maximum likelihood estimation, MLE)이다.

$$\hat{\theta} = \arg\max_{\theta \in \Omega} L(\theta | X_1, \cdots, X_n)$$

로그 가능도

- 가능도 $L(\theta)$는 θ에 관하여 복잡한 함수의 형태로 나타나고, 모수들의 집합인 Ω (Parameter space)가 조건에 따라 bounded set인 경우도 있다.
- 로그변환을 시킬 경우 좀 더 다루기 편한 함수가 되기도 한다.
- 독립인 관측값들이 주어졌을때,

$\ell(\theta) = \log(L(\theta)) = \log(\prod f(X_i;\theta)) = \sum_{i=1}^{n} \log(f(X_i,\theta)).$

$$\hat{\theta} = \arg\max_{\theta\in\Omega} L(\theta) = \arg\max_{\theta\in\Omega} \ell(\theta)$$

예제 1

- 앞에서 예시로 소개하였던 문제에서 제품의 불량률 p의 최대가능도 추정량을 구하여라.

예제 2

- 구간 $[0, \theta]$에서의 균일분포 $U[0, \theta](0 < \theta < \infty)$로부터의 랜덤표본을 $X_1, \cdots, X_n$이라 하자. 이때 모수 θ의 MLE를 구하여라.

최대가능도 추정량의 계산

- 많은 경우 가능도가 최대가 되게 하는 θ를 찾기 위해 미분을 이용
- 즉, $\dot{\ell}(\theta; x) = 0$ (가능도 방정식, likelihood equation) 의 해 중에 최대가능도 추정량을 찾음.
- 가능도방정식을 통해 추정량의 형태을 직접 구하거나
- 그렇지 않은 경우, 수치적으로 근을 찾는 방법을 사용.
- 수치적으로 근을 찾는 경우, 다양한 최적화 방법을 사용하게 된다. Newton-Rapson method, Fisher scoring method, Constraint optimization, EM algorithm, 등등.

예제 3

- 치즈회사에서 새로운 치즈(A) 를 선보여 다른 두 경쟁 제품 (B,C)과의 선호도 차이를 알아보고자 50명에게 blind test를 실시 하여 선호도를 조사한 결과, 20, 18, 12 명이 각각의 치즈를 골랐다. 선호비율을 p_1, p_2, p_3라고 할때 주어진 자료를 가지고 최대가능도 추정량을 구하여라

MLE 의 성질

몇가지 가정하에,

- 일치성:

$$\hat{\theta} \to \theta \text{ in probability}$$

- 점근적 정규성:

$$\sqrt{n}(\hat{\theta} - \theta) \to \mathcal{N}(0, \mathcal{I}^{-1}(\theta)) \text{ in distribution,}$$

  여기서 $\mathcal{I}(\theta) = E\left[\left(\frac{d}{d\theta} \log f(X;\theta)\right)^2\right]$는 피셔정보 (*Fisher Information*)라고 부른다.

- 즉, 점근적으로 (n이 커질때), MLE는 평균이 θ, 분산이 $\frac{1}{n\mathcal{I}(\theta)}$인 정규분포를 따른다.

Fisher Information:

- 정의

$$\mathcal{I}(\theta) = E\left[\left(\frac{d}{d\theta}\log f(X;\theta)\right)^2\right]$$

- 확률밀도함수 f 가 적당한 조건을 만족할때,

$$\mathcal{I}(\theta) = -E\left(\frac{d^2}{d\theta^2}\log f(X;\theta)\right)$$

- 앞의 피셔정보는 하나의 확률변수(또는 하나의 데이터)에 대한 피셔정보이다.
- n개의 확률변수(또는 n개의 데이터)에 대한 피셔정보는

$$\mathcal{I}_n(\theta) = -E\left(\frac{d^2}{d\theta^2}\log f(X_1,\cdots,X_n;\theta)\right)$$

- 서로 독립이고 같은 분포로부터 나온 데이터인 경우

$$\begin{aligned}\mathcal{I}_n(\theta) &= -E\left(\frac{d^2}{d\theta^2}\log f(X_1,\cdots,X_n;\theta)\right)\\ &= -E\left(\sum_{i=1}^{n}\frac{d^2}{d\theta^2}\log f(X_i;\theta)\right) = n\mathcal{I}(\theta)\end{aligned}$$

혼합분포를 이용한 MLE

가우시안 혼합분포를 통한 예제

- 데이터 $X_1, \cdots, X_n$가 두 개의 구성원을 갖는 가우시안 혼합분포를 따른다고 가정하자.
- 두개의 군집이 있는 데이터를 모델링하는데 사용할 수 있다.
- 즉, $X_1, \cdots, X_n \overset{i.i.d.}{\sim} f(x|\boldsymbol{\theta})$,
  $f(x|\boldsymbol{\theta}) = w_0 \frac{1}{\sigma_0^2} \phi\left(\frac{x-\mu_0}{\sigma_0}\right) + (1-w_0) \frac{1}{\sigma_1^2} \phi\left(\frac{x-\mu_1}{\sigma_1}\right)$,
  $\boldsymbol{\theta} = (\mu_0, \sigma_0^2, \mu_1, \sigma_1^2, w_0)$, $\sigma_0^2, \sigma_1^2 > 0, 0 \le w_0 \le 1$.

- 로그 가능도함수는 다음과 같다.

$$\begin{aligned}
\ell(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log f(X_i|\boldsymbol{\theta}) \\
&= \sum_{i=1}^{n} \log \left(w_0 \frac{1}{\sigma_0^2} \phi\left(\frac{X_i - \mu_0}{\sigma_0}\right) + (1 - w_0) \frac{1}{\sigma_1^2} \phi\left(\frac{X_i - \mu_1}{\sigma_1}\right)\right) \\
&= \sum_{i=1}^{n} \log \left(w_0 \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(X_i - \mu_0)^2} \right. \\
&\qquad \left. + (1 - w_0) \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(X_i - \mu_1)^2} \right)
\end{aligned}$$

- 최대가능도를 최대가 되게 하는 $\boldsymbol{\theta}$를 찾기가 어렵다.

$i-$번째 데이터 X_i가 혼합분포의 어느 구성원으로부터 왔는지를 나타내는 잠재 변수(latent variable) K_i를 도입해 보자.

- $K_i = 0$이면 X_i는 $\frac{1}{\sigma_0^2}\phi\left(\frac{X_i-\mu_0}{\sigma_0}\right)$로부터 왔고 (즉, $N(\mu_0, \sigma_0^2)$),
- $K_i = 1$이면 $\frac{1}{\sigma_1^2}\phi\left(\frac{X_i-\mu_1}{\sigma_1}\right)$로부터 온 것 (즉, $N(\mu_1, \sigma_1^2)$)을 나타낸다.
- K_i는 $i-$번째 데이터가 어느 군집에 속하는지 알려주는 label로 생각할 수 있다.
- 잠재변수 K_i는 $P(K_i = 0) = w_0$, $P(K_i = 1) = 1 - w_0$인 확률변수이다.

- $X_i|K_i = 0 \sim N(\mu_0, \sigma_0^2)$
- $X_i|K_i = 1 \sim N(\mu_1, \sigma_1^2)$
- 잠재변수와 조건부 확률분포들을 이용히면 X_i의 확률분포가 처음에 가정했던 혼합븐포가 됨을 알수 있다.

$$
\begin{aligned}
f(X_i) &= f(X_i|K_i = 0)P(K_i = 0) + f(X_i|K_i = 1)P(K_i = 1) \\
&= w_0 \frac{1}{\sigma_0^2} \phi\left(\frac{X_i - \mu_0}{\sigma_0}\right) + (1 - w_0) \frac{1}{\sigma_1^2} \phi\left(\frac{X_i - \mu_1}{\sigma_1}\right)
\end{aligned}
$$

- 한편, X_i, K_i의 결합확률분포는 다음과 같다.

$$\begin{aligned} f(X_i, K_i) &= f(X_i|K_i)f(K_i) \\ &= \frac{1}{\sigma_{K_i}^2}\phi\left(\frac{X_i-\mu_{K_i}}{\sigma_{K_i}}\right) w_0^{1-K_i}(1-w_0)^{K_i} \\ &= \frac{1}{\sqrt{2\pi\sigma_{K_i}^2}} e^{-\frac{1}{2\sigma_{K_i}^2}(X_i-\mu_{K_i})^2} w_0^{1-K_i}(1-w_0)^{K_i} \end{aligned}$$

- 만약 K_i를 안다면 (어느 군집으로부터 왔는지의 정보가 있다면), (X_i, K_i), $i = 1, \cdots, n$을 이용하여 로그 가능도 함수는 다음과 같이 나타낼수 있다.

$$\begin{aligned}
\ell(\boldsymbol{\theta}|X_1, K_1, \cdots, X_n, K_n) &= \sum_{i=1}^{n} \log f(X_i, K_i|\boldsymbol{\theta}) \\
&= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\log(\sigma_{K_i}^2) - \sum_{i=1}^{n}\left(\frac{1}{2\sigma_{K_i}^2}(X_i - \mu_{K_i})^2\right) \\
&+ \left(n - \sum_{i=1}^{n} K_i\right)\log(w_0) + \left(\sum_{i=1}^{n} K_i\right)\log(1 - w_0)
\end{aligned}$$

- X_i들만 가지고 있을때의 로그가능도 함수보다 덜 복잡하다.
- 하지만 K_i는 실제로 관측한 값이 아니므로 (X_i, K_i)를 이용한 로그가능도 함수를 바로 사용 할 수는 없다.
- (X_i, K_i)를 이용한 로그가능도 함수는 나중에 소개할 EM 알고리즘에서 사용된다.

베이지안 추론

- 통계적 추론의 한 방법으로, 추론해야 하는 대상의 사전 확률에서 데이터 관측을 통해 해당 대상의 사후 확률을 업데이트하여 추론하는 방법
- 베이즈 확률론을 기반으로 하며, 이는 추론하는 대상을 확률변수로 보아 그 변수의 확률분포를 추정하는 것을 의미

빈도주의 v.s. 베이지안

빈도주의 **(Frequentist approach)**

- 확률변수가 특정한 분포를 따른다고 가정하고 그 분포의 모수를 추정한다. 이 때, 모수는 고정된 상수이다.
- 확률은 ‘무한히 많은 시행’에서의 상대적인 빈도로 정의된다.
- 통계적 추론은 모수를 추정하는 것에 목적을 둔다.

베이지안 (Bayesian approach)

- 자료가 특정한 분포에서 나왔다고 할때, 그 분포의 모수를 고정된 상수가 아니라 확률변수로 가정된다. 즉, 모수도 분포를 가지는 것으로 간주한다.
- 확률을 빈도나 어떤 시스템의 물리적 속성으로 여기는 빈도주의와는 달리, 베이지안들은 주관주의 확률이론에 따라 확률을 어떤 사람이 특정한 순간에 주어진 명제나 사건에 대해 갖는 믿음의 정도(degree of belief)로 정의한다. 따라서 모수의 분포를 추정할 때 현재 관찰된 자료뿐만 아니라 이전의 자료나 연구자의 믿음 등도 고려된다.
- 새로운 자료가 수집되면 모수에 대한 추정이 업데이트 된다.

“동전 하나를 던졌을 때 앞면이 나올 확률이 **50**퍼센트이다.”

빈도주의

- 동전 하나 던지기를 수천, 수만번 하면 그중에 50퍼센트는 앞면이 나오고, 50퍼센트는 뒷면이 나온다
- 객관적 확률로 해석

베이지안

- 동전 하나 던지기의 결과가 앞면이 나올 것이라는 확신은 50 퍼센트이다.
- 주관적 확률로 해석

베이지안 추론에 필요한 요소

베이지안 추론은 다음의 세 가지 요소가 필요하다.

- 사전분포 : 모수 θ의 분포로 자료를 보기전 분석자의 θ에 관한 불확실성을 나타낸다. $\pi(\theta)$로 나타낸다.
- 확률모형 : 데이터의 분포에 관한 모형 $x \mid \theta \sim f(x \mid \theta)$ 또는 $\pi(x \mid \theta)$로 나타낸다.
- 사후분포 : 데이터가 주어졌을 때, θ의 확률분포로 데이터를 본 후의 분석자의 θ 에 관한 불확실성을 나타낸다. $\pi(\theta|x)$로 나타낸다.

- 확률모형 $f(x \mid \theta)$, 사후분포 $\pi(\theta \mid x)$는 조건부 확률분포로 해석한다.
- 즉, $f(x \mid \theta) = f(x, \theta)/\pi(\theta)$, $\pi(\theta \mid x) = f(x, \theta)/f(x)$.
- 여기서는 $f(\cdot)$ 또는 $\pi(\cdot)$를 확률밀도함수(또는 확률질량함수)로 혼용해서 사용한다.
- 베이즈 정리를 이용하면, 사후확률은 다음과 같다.
- $\pi(\theta \mid x) = f(x \mid \theta)\pi(\theta)/f(x) \;\propto\; f(x \mid \theta)\pi(\theta)$
- 즉, 사후분포 $\propto$ 가능도 $\times$ 사전분포

- 사후분포에서 평균, 중앙값, 최빈값등을 모수 θ의 베이지안 추정값으로 사용할 수 있다.
- 사후분포의 평균 (Posterior mean)
- 사후분포의 중앙값 (Posterior median)
- 사후분포의 최빈값 (Maximum a Posteriori, MAP)

정규분포 예제

- 분산 σ^2가 알려져 있을 때, 평균 θ, 분산 σ^2인 정규분포를 생각해 보자.

$$f(x|\theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\}, -\infty < \theta < \infty$$

- 모수 θ의 사전분포도 정규분포로 가정한다.

  $\theta \sim N(m, s^2)$

- 즉, 자료 관측 이전에 θ가 대략 m이라 믿으며, 그 불확실성이 대략 s만큼인 정규분포를 따른다고 믿는 것을 뜻한다.

- 데이터 $X = x$가 관측되었다고 할때, 베이즈정리에 의한 사후분포는

$$f(\theta|x) \propto \frac{1}{\sqrt{2\pi s^2}} exp\left\{-\frac{(\theta - m)^2}{2s^2}\right\} \times \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{(x - \theta)^2}{2\sigma^2}\right\}$$

- 이 식을 정리하면

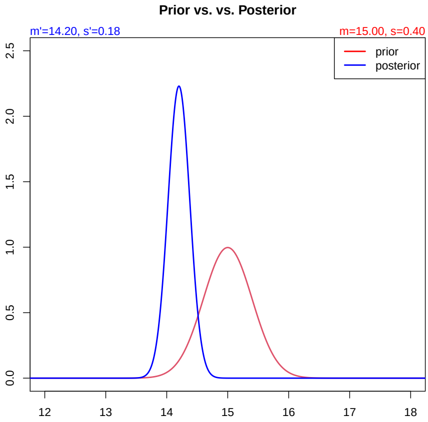
$$\theta \,|\, X = x \sim N\left(\frac{\frac{x}{\sigma^2} + \frac{m}{s^2}}{\frac{1}{\sigma^2} + \frac{1}{s^2}}, \left(\frac{1}{\sigma^2} + \frac{1}{s^2}\right)^{-1}\right)$$

- 사후분포의 평균 $\frac{\frac{x}{\sigma^2} + \frac{m}{s^2}}{\frac{1}{\sigma^2} + \frac{1}{s^2}}$은 데이터의 평균 (이 경우 x) 과 사전분포의 평균(m)의 가중평균으로 볼 수 있다.

정규분포 예제- 데이터 이용

배터리를 만드는 어떤 공장이 있다. 이 공장에서 만드는 배터리의 불량률(%)은 정규분포를 따르는데, 분산은 0.04이고 평균은 알려져있지 않다. 그런데 몇 년간의 자료를 토대로 보았을 때 불량률의 평균은 평균이 15%, 분산이 0.16인 정규분포를 따르는 것으로 확인되었다. 만약 하루 동안 공장을 가동하여 그 날의 불량률이 14%로 관측되었다면, 이를 통하여 사후분포를 구하고 불량률의 베이즈 추정을 구해보자.

- μ에 대한 사전분포는 $N(15, 0.16)$이다.
- 베이즈 정리에 의하여 $X = 14$와 알려져있는 분산 $\sigma^2 = 0.04$를 토대로 μ의 사후분포는 $N(14.2, 0.032)$를 따름을 알 수 있다.
- 사전확률분포와 사후확률분포



- θ에 대한 사전믿음(정규분포를 따른다는)에 관측값 $X = 14$의 정보가 더해져 θ에 대한 사후분포가 만들어짐을 확인할 수 있다.
- 관측값에 의하여 평균이 다소 작아진 정규분포를 따르는 사후분포를 띄게 되었다.
- 또한 사후분포의 분산이 작아졌음을 알 수 있는데, 이는 θ가 사후분포의 평균 근방의 값을 가질 것에 대한 믿음이 더 커진 것이라 생각할 수 있다.

베이지안 결정이론 (Bayesian Decision Theory)

- 에러를 최소화하고 가장 위험이 작은 결정을 내리는 원칙.
- 에러마다 다른 가중치를 주는 손실함수 (loss function)을 고려하여
- '기대손실'이 가장 작게끔 하도록 결정을 내리는 원칙을 뜻한다.
- 베이지안 결정이론은 패턴 분류 문제에 대한 기본적인 통계학적 접근방법으로 볼 수 있다.
- 베이지안 결정이론에 의한 베이지안 추정량의 예 - 제곱손실함수 (quadratic loss function)의 기댓값을 최소화하는 추정량은 사후분포의 평균 (posterior mean) 이다

베이지안 추론에 필요한 분포

- 베이지안 추론을 위해 데이터의 분포와 사전분포로 많이 쓰이는 분포들은 다음과 같다.
- 다항분포, 디리클레분포, 베타분포, 감마분포, 다변량 가우시안, 지수족, 가우시안 믹스쳐, 위샤트 분포등

사전분포의 종류

- 사전분포는 데이터를 보기전에 모수 θ의 불확실성을 나타내는 확률분포이다.
- 켤레사전분포 **(Conjugate prior)**: 사전분포와 사후분포가 같은 분포족 (distribution family)에 속하게 되는 사전분포를 말한다.

  예) 데이터가 정규분포를 따를때, 평균의 사전분포도 정규분포를 따른다고 할 경우, 평균의 사후 분포도 정규분포를 따름
- **Informative prior (subjective prior), non-informative prior (diffuse prior, objective prior)**

  모수에 대해 구체적 정보를 주는 경우와 일반적 정보 또는 아예 정보가 없는 경우로 나눠서 생각할 수 있다.

  예) $\mu \sim N(1, 0.1^2)$ v.s. $\mu \propto 1$.

- **Proper prior** v.s. **Improper prior**

  사전분포가 잘 정의된 확률분포일때 (proper prior)와, 잘 정의되지 않은 확률분포(improper prior), 즉, 적분 또는 합이 유한하지 않을때로 나누어 생각할 수 있다.

- 예를들어, $X \sim N(\mu, \sigma^2)$ 인 자료가 있다고 하자. 이때 분산은 알려져 있다고 가정한다. 평균에 대한 특별한 정보(믿음)가 없는경우 평균이 어떤 값을 가지던지 동일한 정도의 정보를 주도록 하고 싶으면 μ 가 균일분포를 따른다고 가정할 수 있다. 다만 평균의 범위가 $(-\infty, \infty)$이므로 이러한 분포는 적분가능하지 않다.

- 사전분포가 improper prior이어도 사후분포가 proper prior가 될수 있다. 다만 항상 되는것은 아니기 때문에 확인이 필요하다.

사후분포 예시: 이항분포와 베타사전분포

- 이항분포

  연속된 n번의 독립적 시행에서 각 시행이 확률 θ (단, $0 \le \theta \le 1$)일때 성공횟수는 이항분포를 따른다.

- 이항분포의 확률질량함수는 다음과 같다.

$$f(x|\theta) = \frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{n-x}, 0 \le \theta \le 1$$

- 사전분포: 모수 θ 에 대한 사전분포로 베타분포를 고려해 보자.
  $\theta \sim Beta(\alpha, \beta)$, $\alpha, \beta > 0$
- 베타분포 의 확률밀도함수: $f(\theta|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$
- 데이터 X가 관측되었다고 할때, 베이즈 정리를 이용한 θ의 사후분포를 구하여 보자.

$$\begin{aligned} f(\theta|X) &= Lik(\theta|X)\,\pi(\theta) \\ &\propto \frac{n!}{X!(n-X)!}\theta^{X}(1-\theta)^{n-X} \times \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{\alpha+X-1}(1-\theta)^{\beta+n-X-1} \\ &\propto \text{pdf of } Beta(\alpha+X, \beta+n-X) \end{aligned}$$

- θ의 사후분포는

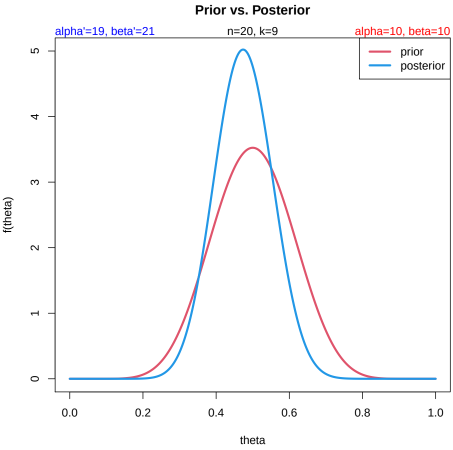
$$\theta \,|\, X \sim Beta(\alpha+X, \beta+n-X)$$

- 이항분포인 Likelihood에 대하여 베타사전분포를 사용하면 사후분포 역시 베타분포임을 알 수 있다.
- 베타사전분포는 이항가능도함수에 대한 켤레 사전분포가 된다.
- 데이터 X가 추가(관측)되면서 모수 θ의 분포가 수정(업데이트) 되었다.

성공확률의 베이지안 추정

- 어떤 도시 시장의 지지율 θ를 추정하는 상황을 생각하자.
- 20명으로부터의 랜덤 표본을 추출한 결과 (지지하는 경우 1)는 $(1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0)$ 이다
- 이 경우, $n = 20$, $X = 9$가 되고, 가능도함수는 이항분포를 따른다. 이때 지지율 θ에 대하여 베이지안 추정을 하여 보자.

- θ에 대한 사전분포로는 평균이 0.5인 $Beta(10, 10)$을 사용한다.
- θ의 사후분포는 $Beta(19, 21)$을 따른다.
- 그래프를 통해 사전확률분포와 사후확률분포를 비교할 수 있다.



- 처음에 부여한 θ의 사전확률분포는 지지율 θ가 0.5 근방에서 가장 높은 확률을 띄는 대칭분포이다.
- 데이터가 없는 상태에서 지지율이 0.5근방일 것으로 가정한 것이다.
- 데이터가 추가되고 나서, 이에대한 Likelihood로 인하여 사후분포가 업데이트되면서, 사후확률분포는 지지율이 0.5 보다 다소 작은 값 근방에서의 확률이 가장 높은 것으로 나옴. 사후분포의 평균은 0.475
- 그래프의 모양으로 보았을 때 분산이 작아졌고 이는 사후분포가 그 분포의 평균 근방의 값을 가질 것에 대한 믿음이 더 커진 것이라 해석할 수 있다. 자료(정보)가 추가됨으로서 모수정보 대한 불확실성정도가 줄어든것으로 해석 할 수 있다.

베이지안 분석 응용 - 나이브 베이즈를 이용한 분류

- N개의 특성(feature)들을 가지고 K개의 클래스 $(C_1, \cdots, C_K)$ 중 하나로 분류시키는 문제를 생각해보자.
- 이 경우 하나의 관측값은 N개의 특성들을 모아놓은 벡터로 생각할 수 있고, 이 벡터를 확률벡터 $\boldsymbol{X} = (X_1, \cdots, X_N)$라고 본다면,
- 주어진 문제는 조건부 확률 $P(C_k|\boldsymbol{X})$를 계산하여 값이 가장 큰 클래스로 정하는 문제로 볼 수 있다.
- 즉, $\arg\max_{k \in \{1, \cdots, K\}} P(C_k|\boldsymbol{X})$

- $P(C_k|\boldsymbol{X}) = f(\boldsymbol{X})$를 $\boldsymbol{X}$에 대한 함수로 본다면, 클래스를 아는 여러개의 $\boldsymbol{X}$들을 관측하여 $f(\boldsymbol{X})$를 추정하는 문제로 볼 수 있다.
- 나이브 베이즈는 이러한 방법 중 하나로, 주어진 클래스 상에서 특성들이 (X_j) 서로 독립이라는 가정을 통해 (실제로 독립이 아닐지라도) $P(C_k|\boldsymbol{X})$를 다음과 같이 단순화 시킨다.

$$
\begin{aligned}
P(C_k|\boldsymbol{X}) &= \frac{P(\boldsymbol{X}|C_k)P(C_k)}{P(\boldsymbol{X})} \quad \text{베이즈 정리 이용} \\
&= \frac{1}{P(\boldsymbol{X})}P(C_k)P(X_1, X_2, \cdots, X_N|C_k) \\
&= \frac{1}{P(\boldsymbol{X})}P(C_k)\prod_{j=1}^{N} P(X_j|C_k)
\end{aligned}
$$

- $P(C_k|\boldsymbol{X})$에서 $\boldsymbol{X}$는 주어진것으로 보기때문에, 나이브 베이즈 분류기는 다음과 같다.

$$\arg\max_{k\in\{1,\cdots,K\}} P(C_k|\boldsymbol{X})$$

$$\overset{NaiveBayes}{=} \arg\max_{k\in\{1,\cdots,K\}} \frac{1}{P(\boldsymbol{X})} P(C_k) \prod_{j=1}^{N} P(X_j|C_k)$$

$$\arg\max_{k\in\{1,\cdots,K\}} P(C_k) \prod_{j=1}^{N} P(X_j|C_k)$$

- 여기서 $P(C_k)$는 k 번째 클래스에 속할 사전확률로 아무 사전 정보가 없다면 $P(C_k) = \frac{1}{K}$로 놓을수 있다.
- $P(X_j|C_k)$는 클래스 k에 속했을때의 j번째 특성 X_j의 확률로 데이터를 통해 추정한다.

- 예를들어 클래스 k에 속하는 n_k개의 관측값 $\boldsymbol{X}_1^{(k)}, \cdots, \boldsymbol{X}_{n_k}^{(k)}$이 있을때, j번째 특성값의 데이터는 각 $\boldsymbol{X}_i^{(k)}$의 j번째 원소를 뽑은 $X_{1j}^{(k)}, \cdots, X_{n_k j}^{(k)}$들이고, j번째 특성값들이 정규분포를 따른다면, 해당 정규분포의 μ_{kj}, σ_{kj}^2을 $X_{1j}^{(k)}, \cdots, X_{n_k j}^{(k)}$를 이용하여 추정하여 사용한다.
- 즉, $\hat{P}(X_j|C_k) = \frac{1}{\sqrt{2\pi\hat{\sigma}_{kj}^2}} e^{-\frac{1}{2\hat{\sigma}_{kj}^2}(X_j - \hat{\mu}_{kj})^2}$

유의성 검정

- 기존의 이론이나 법칙을 부정하는 것으로 보이는 현상이 관측되었을 때, 이를 유지할지 부정할지를 결정하는데 사용.
- 반증을 찾기 위해 설정된 가설 (주로 '기존의 가설'): 귀무가설 (Null hypothesis, H_0)
- 귀무가설의 대안으로 상정되는 가설: 대립가설(Alternative hypothesis, H_1)
- 귀무가설에 대한 반증의 강도를 제공하는 과정을 유의성 검정 (Test of significance)이라 한다.

예제1

건물의 소화용으로 사용되는 살수장치가 섭씨 55도에서 작동되도록 제조하려고 한다. 제조공정의 이상 여부를 판단하기 위해 생산품 중에서 표본을 추출하여 작동 시작 온도를 조사하고자 한다. 이러한 조사에서 공정에 이상이 있다는 증거가 뚜렷하면 후속되는 기술적 조치를 하려 한다.

- 살수장치의 평균 작동 시작 온도 : μ
- 공정에 이상이 있는 경우 : $\mu \neq 55$
- $H_0 \ : \ \mu = 55, H_1 \ : \ \mu \neq 55$

만약, 9개의 표본을 관측한 결과 표본평균 $\bar{x} \ = \ 55.63$ 이라는 결과가 나왔다면, 어떻게 해석해야 할까?

예제 2

- 흰색 또는 빨간색 구슬 10개가 주머니에 들어있다. 흰구슬이 5개 또는 7개가 들어있다고 알려져 있다.
- 구슬을 10번 복원추출하여 나온 값으로 흰 구슬이 5개인지 7개인지 결정하고자 한다.
- 이때 생각할수 있는 가설 두 개는 다음과 같다.

$$
\begin{aligned}
H_0 &: \quad p = 0.5 \ (= p_0) \\
H_1 &: \quad p = 0.7 \ (= p_1)
\end{aligned}
$$

  여기서 p는 한번 추출할때 흰구슬이 나올 확률이다.
- 만약 10번의 복원추출 결과 2개가 흰색이 나왔다면, 어느 가설을 골라야 할까?

예제 2- 계속

- X를 10번의 복원추출한 후의 흰구슬의 개수라고 하자. 이를 이용하여 가설을 고르는 규칙(test or rule)을 만들어 보자.
- 먼저 임의로 test 1을 $X \geq 6$이면 귀무가설을 기각하고 대립가설을 채택하는것으로 정하자.
- 만약 관측값이 2인경우, 즉 $X = 2$, test 1에 의하면 귀무가설을 채택한다.
- 이제 test 2를 $X \geq 1$이면 귀무가설을 기각하는것으로 정하자.
- 이 경우, 관측값 $X = 2$는 test 2에 의하여 귀무가설을 기각하고 대립가설을 채택한다.
- 어떤 test를 골라야 할까?

필요한 용어의 정리

- 단순가설 (simple hypothesis): $\theta = \theta_0$, $p = 0.5$와 같이 모수를 특정값으로 가정하는 가설
- 복합가설 (composite hypothesis): 모수값이 하나보다 많은경우를 가정하는 가설
- 단측가설 (one-sided): $\theta > \theta_0$, 또는 $\theta < \theta_0$와 같이 비교하는 값의 한 쪽에 대해서만 제시되는 가설
- 양측가설 (two-sided): $\theta \neq \theta_0$와 같이 양 쪽에 대해서 제시되는 가설

- 1종 오류 (type I error): 귀무가설이 옳은 상황에서 귀무가설을 기각함으로 인해 생기는 오류
- 2종 오류(Type II error): 귀무가설이 틀린 상황에서 귀무가설을 기각하지 못함으로 인해 생기는 오류
- 두가지의 오류를 동시에 작게 하기 어렵기 때문에, 보통 1종 오류가 일어날 확률을 정하고 (controlling type I error), 그중에 2종 오류가 일어날 확률이 적은 test를 고려한다.

실제현상 / 검정결과	H_0 참	H_1 참
H_0 채택	옳은 결정	**제 2종 오류**
H_1 채택	**제 1종 오류**	옳은 결정

- 유의수준 (significance level): 1종 오류가 일어날 확률, α
  - $\alpha = 0.05$라 함은, 귀무가설이 참인데 기각할 오류를 5% 이하로 하겠다는 것이다.
- 검정력 (power): 귀무가설이 거짓일때 test가 귀무가설을 기각할 확률
  - 2종 오류가 일어날 확률을 β라고 하면, 검정력은 $1 - \beta$
- 주어진 유의수준하에서 검정력이 가장 큰 (2종 오류가 제일 작은) test를 most powerful test라고 부른다.

- 검정통계량 (Test statistics): 가설 검정에 사용되는 통계량
- 기각역 (Critical region): 귀무가설 H_0을 기각시킬 수 있는 검정통계량의 관측값의 영역.
- 앞의 예제2에서 검정 통계량과 기각역, 1종오류 확률등을 찾아보자.

- 유의확률 (P 값): 검정 통계량의 관측값을 가지고 귀무가설이 기각되게하는 가장작은 유의수준.
  - 또는 검정통계량의 관측값을 포함하는 기각역의 최소확률
- 표본으로부터 구한 검정통계량의 관측값으로 구한 유의확률이 지정된 유의수준 이하로 나타나면 통계적으로 유의하다라고 표현
- 예제2에서 관측값이 $X = 2$인 경우에 대한 유의확률을 구해보자.

가설검정에서 알아두어야 할 사항

- 검정통계량이 다르면 다른 test
- 기각역의 형태는 대립가설의 영향을 받음
- P값(유의확률) 또는 기각역을 구하기 위해서는 귀무가설하에서 검정통계량의 분포를 알아야 함
- 어떤 가설을 귀무가설로?
  - 기존에 믿어오던(알려져 있던) 사실
  - 오류의 위험이 더 큰 경우를 1종 오류가 되도록 정함

예제3- 정규모집단에서 모평균의 가설 검정 (분산을 아는경우)

- $H_0 \ : \ \mu = \mu_0$
- 검정통계량: $Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$
- 검정통계량의 관측값 : z_0
- $P(N(0,1) \leq z_\alpha) = \alpha$

대립가설 H_1	유의확률	유의수준 α의 기각역
$\mu > \mu_0$	$P(Z > z_0)$	$Z > z_{1-\alpha}$
$\mu < \mu_0$	$P(Z < z_0)$	$Z < -z_{1-\alpha}$
$\mu \neq \mu_0$	$P(\lvert Z \rvert > \lvert z_0 \rvert)$	$\lvert Z \rvert > z_{1-\alpha/2}$

유의성 검정의 절차

- 귀무가설, 대립가설, 유의수준을 설정한다.
- 표본을 추출하고 검정통계량의 값을 계산한다.
- 가설을 기각할 수 있는지 없는지를 판단하고, 결론을 이끌어낸다.
  - 유의수준으로부터 기각역을 찾아 검정통계량의 값이 기각역에 속하는지 또는
  - 검정통계량의 값으로 유의확률을 계산하여 유의수준과 비교

예제

A사에서 생산중인 고양이 사료 캔의 열량은 평균이 1,200kcal, 표준편차가 100kcal로 알려져 있다. 이제, 사료의 열량을 늘리기 위해 재료를 일부 변경하여 만든 시제품을 25개 생산하여 조사한 결과 평균 열량이 $\bar{x} = 1240$kcal이었다. 새로운 재료로 만든 사료 열량의 표준편차가 100kcal로 유지된다고 할 때, 이 조사 결과는 사료의 열량을 늘리기 위한 재료 변경이 성공적임을 뜻하는가? 유의 수준 0.05에서 검정해보자.

모평균에 관한 추론

- 모평균 μ에 관한 추론을 할 때는 일반적으로 표본평균 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 을 사용
- 모분산이 알려진 경우, 정규분포 이용

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

- 모분산이 알려져있지 않은 경우에는, t분포 이용 (데이터가 정규분포를 따를 때)

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

모평균에 대한 유의성 검정 (모분산을 모르는 정규모집단)

- t검정 (one sample t-test)
- 귀무가설: $H_0 \; : \; \mu = \mu_0$
- 검정통계량: $T \; = \; \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$
- 검정통계량의 관측값: t_0

$T \sim t(k)$일때, $P(T \leq t_p(k)) = p$

대립가설 H_1	유의확률	유의수준 α의 기각역
$\mu > \mu_0$	$P(T > t_0)$	$T > t_{1-\alpha}(n-1)$
$\mu < \mu_0$	$P(T < t_0)$	$T < -t_{1-\alpha}(n-1)$
$\mu \neq \mu_0$	$P(\|T\| > \|t_0\|)$	$\|T\| > t_{1-\alpha/2}(n-1)$

예제

반도체 제조과정에서 사용되는 실리콘 다이오드는 0.60볼트의 가동전압이 요구되며, 이러한 목표에서 벗어나면 불순물의 양을 조절하는 장치에 대한 조정이 필요하다고 한다. 이러한 실리콘 다이오드를 제조하는 공장에서 120개를 랜덤추출하여 조사한 결과 가동전압(볼트)의 평균과 표준편차가 각각 다음과 같았다.
$\bar{x} = 0.62, \ s = 0.11$
적당한 가설을 세우고, 유의수준 5% 에서 가설을 검정해보자.

모분산에 관한 추론

- 모집단의 분산 σ^2의 추론은 일반적으로 표본분산 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ 을 사용
- $X_1, X_2, \cdots, X_n$이 정규분포 $N(\mu, \sigma^2)$에서의 랜덤표본일 때 $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$
- 카이제곱분포를 이용한 모분산의 추론은 정규모집단 가정에 민감하므로 정규성 가정에 대해 충분히 검토한 후 사용

모분산에 관한 유의성 검정

- 귀무가설 H_0 : $\sigma^2 = \sigma_0^2$
- 검정통계량 : $\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$
- 검정통계량의 관측값 : χ_0^{22}

$chi^2 \sim\sim \chi^2(k)$일때, $P(\chi^2 \leq \chi_p^2(k)) = p$

대립가설 H_1	유의확률	유의수준 α의 기각역
$\sigma^2 > \sigma_0^2$	$P(\chi^2 > \chi_0^2)$	$\chi^2 > \chi_{1-\alpha}^2(n-1)$
$\sigma^2 < \sigma_0^2$	$P(\chi^2 < \chi_0^2)$	$\chi^2 < \chi_{\alpha}^2(n-1)$
$\sigma^2 \neq \sigma_0^2$	$2P(\chi^2 > \chi_0^2)$ 또는 $2P(\chi^2 < \chi_0^2)$에서 1보다 작은 값	$\chi^2 > \chi_{1-\alpha/2}^2(n-1)$ 또는 $\chi^2 < \chi_{\alpha/2}^2(n-1)$

예제

플라스틱판을 생산하는 한 공장이 있다. 판 두께의 표준편차가 1.5mm를 상회하면 공정에 이상이 있는 것으로 간주한다. 어느 날 10개의 판을 랜덤추출하여 두께를 측정한 결과가 다음과 같다.

$$\{226, 228, 226, 225, 232, 228, 227, 229, 225, 230\}$$

과거의 공정관리 기록에 의하면, 이러한 판 두께의 분포는 정규분포라고 해도 무방할 때, 공정에 이상이있는가를 유의수준 5%에서 검정해 보자.

- 표본의 수 n = 10, 표본표준편차 s = 2.27