

베이즈 추론은 빈도주의 추론과 상호보완적 역할을 하며 컴퓨팅의 발달과 더불어 급속히 통계학의 주요학파로 자리잡았다. 특히 머신러닝과 다양한 응용분야에서 베이즈 통계는 핵심 역할을 수행하지만 동시에 종종 오용이 되기도 한다. 이 장에서는 베이지안에 대한 기본적인 리뷰와 함께 빈도주의에 어떻게 영향을 주는지 알아보자.

베이지안과 빈도주의 모두 통계적 추론의 가장 기본적인 단위는 다음과 같이 정의되는 확률밀도함수들의 family 이다.

$$\mathcal{F} = \{f_{\mu}(x); x \in \mathcal{X}, \mu \in \Omega\};$$

여기서 x 는 관측된 자료를 의미하며 표본공간 \mathcal{X} 상의 한 점이다. 반면에 관측되지 않는 모수 μ 는 모수공간 Ω 상의 한 점으로 간주할 수 있다. 보통 우리는 $f_{\mu}(x)$ 로 부터 생성된 x 를 관측하고 이를 바탕으로 μ 에 관한 추론을 한다.

확률밀도 함수의 가장 대표적인 예로 정규분포를 들 수 있다.

$$f_{\mu} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2}$$

여기서 \mathcal{X} 와 Ω 모두 $\mathcal{R}^1 = (-\infty, \infty)$ 임을 알 수 있다. 또다른 (이산분포의) 예로 포아송 분포를 들 수 있다..

$$f_{\mu} = e^{-\mu} \mu^x / x!$$

여기서 $\mathcal{X} = \{0, 1, \dots, \dots\}$ 이고 $\Omega = (0, \infty)$ 라는 것을 쉽게 알 수 있다.

베이즈 추론에서는 \mathcal{F} 에 추가로 사전분포(prior distribution)에 관한 가정을 한다.

$$g(\mu), \quad \mu \in \Omega;$$

여기서 $g(\mu)$ 는 μ 에 관한 사전 정보를 나타내며 사전정보란 데이터를 관측하기 전에 알고 있는 정보를 의미한다. 예를 들면 위의 정규분포 예제에서 우리가 μ 가 양수이고 과거 경험에 비추어서 10을 넘지 않는다고 믿는다면 $g(\mu) = 1/10$ 로 0과 10사이의 정의된 균등분포로 고려할 수 있다.

베이즈 정리는 $g(\mu)$ 안에 있는 이러한 선험지식과 관측치의 정보를 결합하는 방식이다. $g(\mu | x)$ 가 사후분포 (posterior distribution), 즉 x 를 관측한 후 이를 바탕으로 사전분포 $g(\mu)$ 를 업데이트 한 분포, 를 나타낸다고 하자. 베이즈 정리는 다음과 같이 사후분포 $g(\mu | x)$ 를 $g(\mu)$ 와 \mathcal{F} 를 이용하여 표현한 식이다.

$$\text{Bayes' Rule: } g(\mu | x) = g(\mu)f_{\mu}(x)/f(x), \quad \mu \in \Omega, \quad (1)$$

여기서 $f(x)$ 는 x 의 marginal density이다.

$$f(x) = \int_{\Omega} f_{\mu}(x)g(\mu)d\mu.$$

위의 베이지 정리 1에서는 빈도주의와 반대로 x 는 고정된 값(관측값)이고 μ 가 모수공간에서 다양한 값을 가진다는 것을 의미한다, 이 의미를 강조하기 위해 식 1를 다음과 같이 표현할 수 있다.

$$g(\mu | x) = c_x L_x(\mu)g(\mu) \quad (2)$$

여기서 $L_x = f_{\mu}(x)$ 는 우도함수(likelihood function)이며 c_x 는 normalizing constant로 $g(\mu | x)$ 의 적분값이 1이 되도록 하게 만드는 역할을 한다.

Note 우도함수에 임의의 상수 c_0 를 곱하는 것은 c_x 의 값을 변화하는 것으로 간주할 수 있기때문에 사실 식2에는 아무런 영향을 주지 않는다. 예를 들자면 포아송 분포에서 $x!$ 을 무시하고 $L_x(\mu) = e^{-\mu}\mu^x$ 로 표현할 수 있다. 이렇게 x 에만 관련된 항을 마치 상수처럼 취급함으로서 종종 베이지 통계에서 각종계산을 간편하게 할 수 있게 된다.

모수공간 Ω 상의 임의의 두점 μ_1 과 μ_2 에 대해서 사후분포의 밀도함수 비를 식 (1)를 이용하여 다음과 같이 표현할 수 있다.

$$\frac{g(\mu_1 | x)}{g(\mu_2 | x)} = \frac{g(\mu_1)f_{\mu_1}(x)}{g(\mu_2)f_{\mu_2}(x)} \quad (3)$$

위의 식의 장점은 marginal 분포를 사용하지 않고 사후분포밀도함수의 비를 정의할 수 있다는 점이다. 위의 식이 의미하는 바는 사후분포의 밀도함수의 오즈비(odds ratio)는 사전분포의 오즈비 곱하기 우도함수의 오즈비라는 것으로 베이지 법칙을 또다른 방식으로 표현하는 것으로 생각할 수 있다.

3.1 Two Example

쌍둥이의 경우 1/3은 일란성이며 2/3는 이란성 쌍둥이로 알려져 있다. 쌍둥이를 임신한 물리학자가 초음파 검사를 통해 태아의 성별을 알아본 결과 쌍둥이의 성별이 같았다. 이 물리학자는 이렇게 추가적인 정보를 얻은 경우 이 쌍둥이들이 일란성인 확률은 얼마인지 궁금해 한다고 가정하자.

일단 일란성 쌍둥이의 경우 항상 성별이 같지만 이란성 쌍둥이의 경우 성별이 같을 확률은 절반이다. 이 사실과 식 (3)을 이용하여 물리학자의 질문에 대답해보자.

$$\begin{aligned} \frac{g(\text{Identical} | \text{Same})}{g(\text{Fraternal} | \text{Same})} &= \frac{g(\text{Identical})}{g(\text{Fraternal})} \cdot \frac{f_{\text{Identical}}(\text{Same})}{f_{\text{Fraternal}}(\text{Same})} \\ &= \frac{1/3}{2/3} \cdot \frac{1}{1/2} = 1. \end{aligned}$$

즉 사후 오즈는 1이다. 따라서 물리학자의 쌍둥이가 일란성일 확률은 0.5로 사전확률인 1/3에서 증가했다는 것을 알 수 있다. 위의 과정을 표를 이용하여 설명해보자 표 3.1에서 셀 a, b, c, d 는 초음파 검사결과(관측치 x)와 일란성 여부(모수 μ)에 따라 나올 수 있는 4가지 경우를 나타낸다. 셀 b 의 경우 발생할 수 없는 경우이기 때문에

Sonogram shows:

		<i>Same sex</i>	<i>Different</i>	
		<i>a</i>	<i>b</i>	
<i>Identical</i>	Twins are:	1/3	0	1/3
<i>Fraternal</i>		<i>c</i>	<i>d</i>	2/3
		1/3	1/3	

Physicist

Doctor

그림 3.1 쌍둥이 문제 분석

확률은 0이라는 것을 알 수 있다. 셀 *c*와 *d*의 경우 이란성 쌍둥이의 동일 성별인 경우와 아닌 경우는 같은 확률로 간주할 수 있고 의사의 사전분포에 따르면 이 두 확률의 합이 2/3이라는 점을 활용하면 각각의 셀에 1/3의 확률을 부여할 수 있다. 마지막으로 *a*의 확률은 *b*의 확률이 0이기 때문에 의사의 사전분포에 의해서 자동적으로 1/3이 된다.

쌍둥이 문제의 경우 사전분포에 관해서 믿을 만한 근거 (예를 들자면 출생기록)가 있기때문에 이를 바탕으로 계산한 사후분포의 값은 대부분의 사람들이 쉽게 받아들인다. 하지만 많은 경우 사전분포를 정하는 것은 쉽지 않을 수 있다. 다음 예제를 통해서 이런 경우에 대해서 알아보자.

표 3.1은 22명 학생의 역학(mechanics)과 벡터과목의 성적이다 두 과목 성적사이의 표본상관계수 $\hat{\theta}$ 는 아래 공식을 사용하여 계산한 결과 0.498이다.

$$\hat{\theta} = \frac{\sum_{i=1}^{22} (m_i - \bar{m})(v_i - \bar{v})}{\left[\sum_{i=1}^{22} (m_i - \bar{m})^2 \sum_{i=1}^{22} (v_i - \bar{v})^2 \right]^{1/2}}$$

여기서 m 과 v 는 각각 역학과 벡터의 성적을 의미하고 \bar{m} 과 \bar{v} 는 해당과목의 평균이다.

만약 우리가 모상관계수 θ 에 관해 사후분포를 구하고 싶다면 어떻게 베イズ 추론을 할 수 있을까?

표 3.1: 22 학생들의 **mechaincs**와 **vectors**과목 시험성적

	1	2	3	4	5	6	7	8	9	10	11
mechaincs	7	44	49	59	34	46	0	32	49	52	44
vectors	51	69	41	70	42	40	40	45	57	64	61

	12	13	14	15	16	17	18	19	20	21	22
mechaincs	36	42	5	22	18	41	48	31	42	46	63
vectors	59	60	30	58	51	63	38	42	69	49	63

우선 (m, v) 가 이변량 정규분포(bivariate normal distribution)를 따른다고 가정하자. 이 경우 표본상관계수의 확률밀도함수는 다음과 같이 주어진다.

$$f_{\theta}(\hat{\theta}) = \frac{(n-2)(1-\theta^2)^{(n-1)/2} (1-\hat{\theta}^2)^{(n-4)/2}}{\pi} \int_0^{\infty} \frac{dw}{(\cosh w - \theta\hat{\theta})^{n-1}}$$

앞의 표현방식에 따르자면 관측치 x 에 해당하는 것이 $\hat{\theta}$ 이며 모수 μ 는 θ 에 대응되고 \mathcal{F} 에 해당하는 것이 위의 확률밀도함수이다. 이 경우 $\mathcal{X} = \Omega = [-1, 1]$ 이다. 여기서 표본상관계수는 θ 의 최대우도추정량(MLE)이며 이 경우 4장에서 배운 MLE의 성질을 활용한다면 다음과 같은 결과를 얻을 수 있다.

$$\sqrt{n}(\hat{\theta} - \theta) \rightarrow N(0, (1 - \theta^2)^2), \text{ as } n \rightarrow \infty.$$

위의 결과를 바탕으로 2장에 배운 내용을 활용하여 다음과 같은 추론을 할 수 있다.

- **Plug-in:** $\hat{\theta}$ 의 표준오차는 $(1 - \theta^2)/\sqrt{n}$ 으로 주어지므로 $\hat{se}(\hat{\theta}) = (1 - \hat{\theta}^2)/\sqrt{n}$ 을 사용할 수 있다.
- **Delta method:** 다음과 같은 Fisher's z-transformation을 고려해보자.

$$m(\hat{\theta}) = \frac{1}{2} \log \left(\frac{1 + \hat{\theta}}{1 - \hat{\theta}} \right)$$

이 경우 delta method에 의해

$$\sqrt{n}(m(\hat{\theta}) - m(\theta)) \rightarrow N(0, 1), \text{ as } n \rightarrow \infty,$$

임을 쉽게 보일 수 있다. 이 경우 분산이 상수이므로 더 이상 plug-in을 사용하지 않아도 된다.

이제 베이지 추론을 대해서 생각해보자. 베이지 추론에서 가장 어려운 부분은 사전분포를 정하는 문제이다. 사실 많은 경우 사전분포를 정의할 만한 과거의 경험이 충분하지 않는 것이 일반적이다. 이럴 경우 가장 쉽게 생각할 수 있는 것이 Laplace의 insufficient reason에 기반한 "flat prior"이다. 즉 θ 가 모수공간 Ω 에 균등하게 분포되어 있다고 가정하는 것이다.

$$g(\theta) = \frac{1}{2} \quad \text{for } -1 \leq \theta \leq 1.$$

그림 3.2에서 검은 실선은 θ 의 사전분포로 flat prior를 사용한 경우 사후분포를 보여주고 있다. 이 사후분포는 사실 θ 의 MLE 값($\hat{\theta}^{mle}=0.498$)에서의 (적분시 1이 되게끔 정규화한) 우도함수 $f_{\theta}(0.498)$ 과 동일하다.

그림 3.2 붉은 파선으로 표현되는 사후분포는 다음과 같이 주어지는 *Jeffreys' prior*를 사용할 경우 생성되는 사후분포이다.

$$g^{\text{Jeff}}(\theta) = 1/(1 - \theta^2), \quad (4)$$

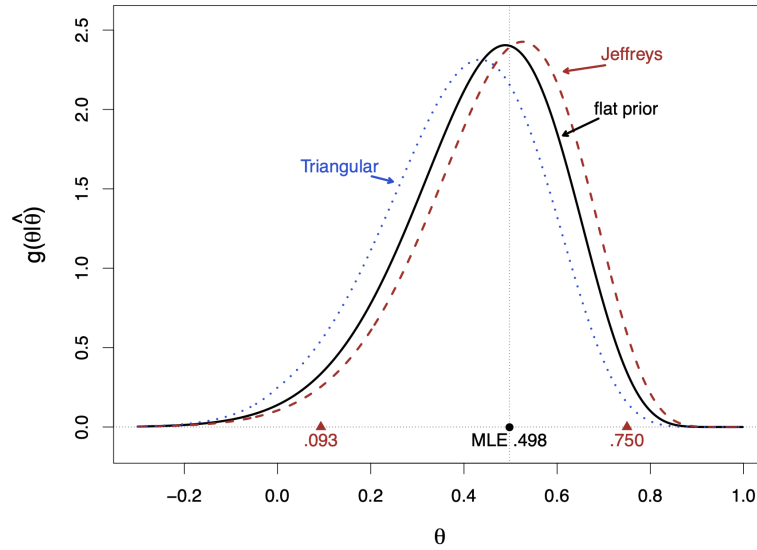


그림 3.2 학생들 시험성적 자료; 3가지 사전 분포에 대응하는 상관계수의 사후분포

위의 공식은 noninformative prior라는 이론에서 기반한 것으로 큰 θ 의 절대값이 큰 경우 (즉 ± 1)에 높은 가능성에 두고 있다.

여기서 주목할 점은 위의 사전분포는 확률밀도함수가 아니다! 즉 $\int_{-1}^1 g(\theta)d\theta = \infty$ 이며 이러한 사전분포를 improper prior라고 한다. 이러한 경우에도 사후분포는 확률밀도함수가 될 수 있으며 반드시 베이즈 법칙을 사용하여 계산된 사후분포의 적분값이 1임을 보여야한다. 이럴 경우 사후분포는 proper하다고 한다.

그림 3.2에서 청색 점선은 다음과 같은 삼각형 형태의 사전분포 사용시 생성되는 사후분포를 나타낸다.

$$g(\theta) = 1 - |\theta|$$

위의 분포는 축소 사전분포(shrinkage prior)의 기본적인 예로서 θ 가 작은 값을 가지는 것을 선호한다. 이러한 사전분포의 선택에 여기에 대응되는 사후분포가 다른 사후분포들에 비해 왼쪽을 조금 옮긴 형태로 나타나는 것으로 확인할 수 있다. 축소 사전분포는 건초더미에서 바늘을 찾는 것과 같은 large-scale estimation과 testing에서 중요한 역할을 하고 있다는 것을 이 과목 후반부에서 확인할 수 있다.

3.2 Uninformative Prior Distributions

만약 우리가 괜찮은 사전분포를 가지고 있다면 베이즈 추론을 사용하는 것이 아마도 빈도주의 추론을 사용하는 것보다 만족할 만한 결과를 가져올 것이다. 하지만 사전분포에 관한 유용한 정보를 얻는 것은 생각보다 쉬운 일은 아니다. 이렇게 적절한 경험이 없는 경우 사전분포를 어떻게 정하는 지가 베이즈 추론의 오랜 숙제로 남겨져 있었다.

이러한 문제의 해결책으로 noninformative prior가 제시되었다. 여기서 noninformative의 의미는 사용된 사전 분포가 추론의 결과에 은연중에 영향을 미치지 않는다는 것을 의미한다. Laplace의 insufficient reason 즉 flat

prior를 사용하는 것이 이 목적에 부합된다고 오랫동안 여겨져 왔다. 하지만 1860대에 벤다이어그램의 벤, 1920년대에는 Fisher에 의해 다음과 같은 flat prior의 심각한 문제점이 지적되었다.

앞의 상관계수 예제를 살펴보면 θ 의 flat prior $g^{\text{flat}} = c$ 는 $\gamma = e^\theta$ 에 대해서는 더 이상 uniform prior가 아니다.

$$g(r) = g^{\text{flat}} |\partial r / \partial \theta| = c e^\theta = c \gamma.$$

그렇다면 다음과 같은 사후확률을 계산한다고 가정해보자.

$$\Pr(\theta > 0 \mid \hat{\theta}) = \Pr(\gamma > 1 \mid \hat{\theta})$$

여기서 우리는 θ 또는 γ 중 어느 모수에 대한 flat prior를 사용하는지 여부에 따라 사후확률이 달라짐을 알 수 있다. 따라서 flat prior는 더 이상 noninformative라고 할 수 없다

여기서 잠깐 주의를 Fisher information로 전환해보자. 모수가 하나이고 모수공간 $\Omega = \mathcal{R}^1$ 인 경우 Fisher information은 다음과 같이 정의된다.

$$\mathcal{I}_\mu = \mathbb{E}_\mu \left[\left(\frac{\partial}{\partial \mu} \log f_\mu(x) \right)^2 \right]$$

예를 들면 포아송 분포의 경우

$$\frac{\partial}{\partial \mu} (\log f_\mu) = \frac{x}{\mu} - 1$$

이며 따라서

$$\mathcal{I}_\mu = \mathbb{E}_\mu \left(\frac{x}{\mu} - 1 \right)^2 = \frac{1}{\mu}$$

Jeffreys' prior는 Fisher information을 이용하여 다음과 같이 정의된다.

$$g^{\text{Jeff}}(\mu) = \mathcal{I}_\mu^{-1/2}.$$

사실 $1/\mathcal{I}_\mu$ 는 $\hat{\mu}^{mle}$ 의 분산 σ_μ^2 의 근사값이기 때문에

$$g^{\text{Jeff}}(\mu) = 1/\sigma_\mu$$

라고 할 수 있다.

Jeffreys' prior는 변환에 관계없이 각 모수에 대한 같은 형태의 prior를 제공하기 때문에 Fisher와 Venn이 지적한 문제점을 해결할 수 있다. 식 (2)를 이용해서 설명해보자.

$$g^{\text{Jeff}}(\theta \mid x) = c_x L_x(\theta) g^{\text{Jeff}}(\theta) = c_x L_x(\theta) (\mathcal{I}_\theta)^{-1/2}$$

이제 미분가능한 매끄러운 변환 h 를 통해 새로운 모수 $\gamma = h(\theta)$ 를 정의하자. 이 경우

$$g^{\text{Jeff}}(\gamma | x) = c_x L_x(\gamma) g^{\text{Jeff}}(\gamma) = c_x L_x(\gamma) (\mathcal{I}_\gamma)^{-1/2}$$

여기서

$$\mathcal{I}_\gamma = \mathbb{E}_\gamma \left[\left(\frac{\partial}{\partial \gamma} \log f_\gamma(x) \right)^2 \right]$$

이고 또한

$$\frac{\partial}{\partial \gamma} \log f_\gamma(x) = \frac{\partial \theta}{\partial \gamma} \frac{\partial}{\partial \theta} \log f_\theta(x)$$

이다. 따라서,

$$\mathcal{I}_\gamma = \left(\frac{\partial \theta}{\partial \gamma} \right)^2 \mathcal{I}_\theta,$$

임을 보일 수 있다. 이제

$$g(\theta | x) = g(\gamma | x) \left| \frac{\partial \gamma}{\partial \theta} \right| = c_x L_x(\gamma) g(\gamma) \left| \frac{\partial \gamma}{\partial \theta} \right|$$

에서 $g^{\text{Jeff}}(\gamma)$ 를 prior로 사용할 경우 위의 식이 $g^{\text{Jeff}}(\theta | x) = c_x L_x(\theta) g^{\text{Jeff}}(\theta)$ 됨을 보이자.

$$\begin{aligned} g(\theta | x) &= c_x L_x(\gamma) g^{\text{Jeff}}(\gamma) \left| \frac{\partial \gamma}{\partial \theta} \right| \\ &= c_x L_x(\gamma) (\mathcal{I}_\gamma)^{-1/2} \left| \frac{\partial \gamma}{\partial \theta} \right| \\ &= c_x L_x(\gamma) \left(\left(\frac{\partial \theta}{\partial \gamma} \right)^2 \mathcal{I}_\theta \right)^{-1/2} \left| \frac{\partial \gamma}{\partial \theta} \right| \\ &= c_x L_x(\theta) \left| \frac{\partial \gamma}{\partial \theta} \right| \left| \frac{\partial \theta}{\partial \gamma} \right| (\mathcal{I}_\theta)^{-1/2} \\ &= c_x L_x(\theta) g^{\text{Jeff}}(\theta) \end{aligned}$$

상관계수 예제의 경우 Jeffreys prior를 사용해보자. 우선 $\hat{\theta}$ 의 근사 표준편차는 다음과 같다.

$$\sigma_\theta = c(1 - \theta^2)$$

따라서 Jeffreys prior가 식 (4)와 같이 주어진다. 여기서 Jeffreys' prior가 improper prior이기 때문에 상수 c 의 값은 임의로 정할 수 있으며 베이지 법칙을 계산할때 영향을 주지 않는다.

그림 3.2에서 빨간색 삼각형이 표시하는 구간은 Jeffreys' prior를 기반으로 구한 95% credible interval $[0.0093, 0.750]$ 이다. 즉,

$$\int_{0.093}^{0.750} g^{\text{Jeff}}(\theta | \hat{\theta}) = 0.95$$

이 구간은 Neyman의 95% 신뢰구간과 거의 동일하며 one-parameter 경우 베이지안과 빈도주의가 거의 동일한 결론을 내릴 수 있다는 것을 보여준다.

하지만 다차원에서 이와 같은 연결고리가 항상 유지되는 것은 아니다. 우리가 다음과 같이 10개의 서로 독립이고 평균이 다른 정규분포모형에서 각 1개의 자료를 생성한 후 관측한다고 가정하자.

$$x_i \stackrel{ind}{\sim} \mathcal{N}(\mu_i, 1) \quad \text{for } i = 1, 2, \dots, 10,$$

이 경우 Jeffreys' prior는 각각의 모형에서 상수로 주어진다. 하지만 우리가 결합분포 (joint distribution)에서 Jeffreys' prior를 고려하면 역시 다음과 같은 flat prior가 주어진다.

$$g(\mu_1, \mu_2, \dots, \mu_{10}) = \text{constant},$$

나중에 13장에서 논의하겠지만 다차원에서 이런 flat prior는 심각한 문제를 야기할 수 있다.

3.3 Flaws in Frequentist Inference

베이즈 추론은 추론에 있어서 일관성있는 (coherent) 결론을 제시하지만 빈도주의 추론에서는 같은 이야기가 성립하지 않는다. 다음 *meter reader* 예제를 통해서 차이를 알아보자.

Example. 엔지니어가 12 튜브의 전압을 표준화로 보정된 전압계를 이용하여 측정하고자 한다. 전압측정치 x 는 평균이 μ 이고 분산이 1인 정규분포를 따른다. 측정치는 82에서 99사이의 값을 가지며 $\bar{x} = 92$. 하지만 다음날 엔지니어는 전압계가 100이상의 전압이 흐를 경우 100 볼트로 전압을 읽는다는 것을 발견하였다. 사실 그 전날 읽은 12개의 전압 관측치중 100을 넘는 것은 하나도 없었지만 여전히 표본평균을 μ 의 추정치로 사용하는 것이 적절할까?

빈도주의 통계학자는 이 사실을 알고 \bar{x} 는 더 이상 μ 의 unbiased estimator가 아니라고 얘기한 후 censoring을 고려한 다른 대표값을 사용할 것을 권장하였다. 빈도주의 관점에서는 지금 관측된 자료뿐만 아니라 향후 실제 확률모형에서 생성될 미래의 자료에 대해서도 적절한 추정량인지 여부를 고려해야 한다.

반면에 베이지안의 경우 사후분포 $g(\mu | \mathbf{x}) = c_{\mathbf{x}} L_{\mathbf{x}}(\mu) g(\mu)$ 계산시 필요한 정보는 사전분포 $g(\mu)$ 와 실제로 이미 관측된 데이터 \mathbf{x} 이며 실제 관측되지 않은 데이터를 고려하지 않아도 된다. 즉 관측된 자료가 censoring이 되지 않았다면 베이즈 추론에 영향을 주지 않는다. 조금 더 구체적으로 얘기하자면 우도함수의 값은 주어진 데이터를 기준으로 변화가 없기 때문에 베이즈 추론의 결과는 사전분포에 의해서 결정된다는 것이다. 이 예제에서 우도함수는

$$L_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^{12} (\phi_{\mu,1}(x_i))^{1-C_i} (\delta_{100}(x_i))^{C_i}$$

여기서 $\phi_{\mu,1}(x_i)$ 는 평균이 μ 이고 분산이 1인 정규분포의 확률밀도함수이며 $\delta_{100}(x_i)$ 는 $x_i = 100$ 에서 point mass를 의미하며 $C_i(x) = I(x_i > 100)$ 는 censoring여부를 나타낸다.

그림 3.3은 또 순차검정을 통한 다른 예를 제시하고 있다. 매달 실험에서 평균이 μ 이고 분산이 1인 정규분포를 따르는 관측치가 하나씩 생성된다고 하자. 조사관은 매달 누적된 자료를 이용하여 계산된 통계량 Z_i 의 값으로

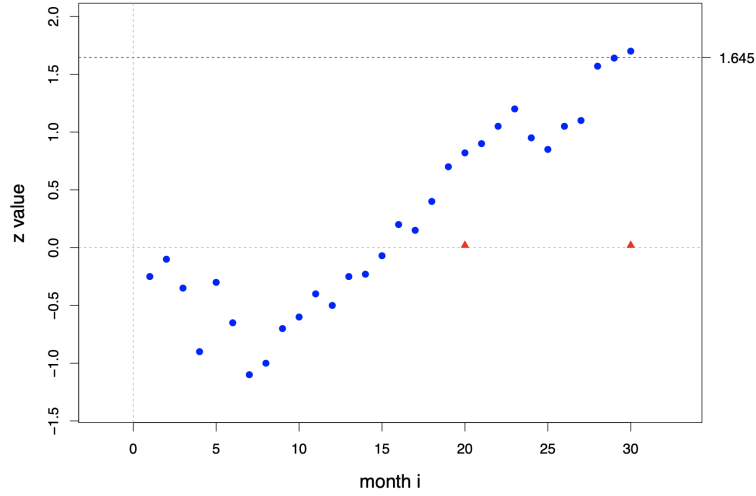


그림 3.3 귀무가설 $\mu = 0$ 하에서 첫번째 달부터 30번째 달까지 Z값

계산하고 그 결과는 3.3에 제시되어 있다.

$$Z_i = \sum_{j=1}^i x_j / \sqrt{i}$$

즉, 통계량 Z_i 는 다음과 같은 분포를 따르며 이 값을 바탕으로 가설검정 $H_0 : \mu = 0$ vs $H_A : \mu > 0$ 을 실시할 수 있다.

$$Z_i \sim \mathcal{N}(\sqrt{i}\mu, 1)$$

실험은 30개월에 종료가 되기로 계획되어 있고 조사관은 그 전에는 데이터를 볼 수 없다. 그림에서 볼 수 있듯이 30번째 달에 Z_i 값이 1.66으로 $z_{0.05} = 1.645$ 를 넘어서 조사관은 귀무가설을 유의수준 0.05에서 기각할 수 있다. 그런데 알고보니 조사관이 실험 포로토콜을 깨고 혹이나 실험을 조기에 멈출 수 있지 않을까 해서 20번째 달 데이터를 들여다 본 것으로 드러났다. 하지만 $Z_{20} = 0.79$ 로 기각역을 넘지 못해서 실험은 예정대로 30개월동안 지속되었다. 하지만 이 경우 실험을 중단하는 규칙이 "30번째 달에서 통계량이 1.645보다 큰 경우"에서 "20번째달 또는 30번째 달에서 통계량이 1.645보다 큰 경우"로 변경되었음을 알 수 있다. 하지만 이 경우 새로운 stopping rule의 확률을 계산하면 다음과 같다.

$$\Pr(\{Z_{20} > 1.645\} \text{ or } \{Z_{20} < 1.645, Z_{30} > 1.645\}) = \Pr(Z_{20} > 1.645) + \Pr(Z_{20} < 1.645, Z_{30} > 1.645) = 0.074$$

따라서 순간의 실수로 인해 stopping rule은 더 이상의 유효하지 않기때문에 이 결과를 바탕으로 귀무가설에 대한 결정을 내리는 것은 불가능하다. 하지만 베이지안의 관점에서는 stopping rule의 변화는 영향을 주지 않는다. 우선 관측된 전체 데이터 $\mathbf{x} = (x_1, x_2, \dots, x_{30})$ 에 대한 우도함수는 다음과 같다.

$$L_x(\mu) = \prod_{i=1}^{30} e^{-\frac{1}{2}(x_i - \mu)^2}$$

위의 우도함수는 stopping rule에 관계없이 항상 동일하며 사후분포는 우도함수를 통해서만 데이터에 의존하기

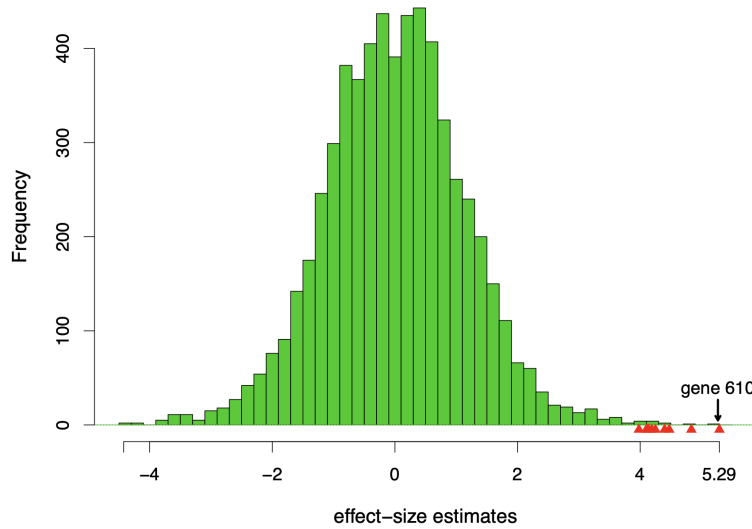


그림 3.4 전립선 암 연구에서 6033개 유전자의 효과크기에 관한 불편추정치. 601번 유전자의 추정치 $x_{610} = 5.29$ 이다. 효과크기는 얼마인가?

때문에 결국 stopping rule은 베이지 추론에 영향을 주지 않는다.

하지만 베이지 통계의 이러한 장점은 고차원에서는 항상 유효한 것은 아니다. 우리가 전립선암의 바이오마커를 찾는데 관심이 있다고 가정하자. 이를 위해서 전립선 암환자 52명과 건강한 일반인 50명을 대상으로 마이크로 어레이를 이용하여 개인별로 $N = 6033$ 유전자의 발현값을 구하였다. 그 후 각 개별 유전자별로 환자그룹과 일반인 그룹의 유전자 발현정도를 비교하는 검정통계량 x 를 계산하였으며 이 검정통계량의 분포는 다음과 같다.

$$x_i \sim \mathcal{N}(\mu_i, 1), \quad i = 1, 2, \dots, N.$$

여기서 μ_i 는 i 번째 유전자의 실제 효과크기(true effect size)를 나타낸다. 그림 3.4는 x_i 의 값을 보여주고 있다.

대부분의 유전자들은 전립선 암과 관련이 없을 것으로 생각되기 때문에 이 경우 μ_i 의 값은 0에 가까울 것으로 짐작된다. 하지만 바이오마커 역할을 하는 비교적 큰 양(또는 음)의 효과크기를 가질 것으로 간주한다.

그림 3.4에서 가장 큰 값은 $x_{610} = 5.29$ 이다. 그렇다면 μ_{610} 의 추정치로 x_{610} 이 적합할까? 개별적으로 볼 때 x_i 들이 μ_i 의 불편추정량이라는 걸 생각한다면 일견 일리있는 이야기이지만 x_{610} 이 6033개의 통계량 값중 가장 큰 값이라는 사실을 고려한다면 x_{610} 을 추정치로 사용하는 것은 과대추정의 문제를 유발할 것이다. 만약 $\epsilon_i = x_i - \mu_i$ 이고 $y = \max_i \epsilon_i$ 라고 하면 $\mathbb{E}(\epsilon_i) = 0$ 이지만

$$\mathbb{E}(y) \leq \sqrt{2 \log N}$$

임을 보일 수 있다. 즉 $y \approx \sqrt{2 \log N}$ 으로 간주할 수 있다. 따라서 어떤 형식으로라도 $\max x_i$ 보다 작은 값을

사용하는 것이 필요하다. 참고로 위의 식은 다음과 같이 증명할 수 있다.

$$\begin{aligned}
 e^{t\mathbb{E}(Y)} &\leq \mathbb{E}(e^{tY}) \\
 &= \mathbb{E}(\max_i e^{tX_i}) \\
 &\leq N \cdot \mathbb{E}(e^{tX}) \\
 &= N \cdot e^{t^2/2}
 \end{aligned}$$

양변에 \log 를 취한 후 t 로 나누면

$$\mathbb{E}(Y) \leq \frac{\log N}{t} + \frac{t}{2}$$

여기서 $t = \sqrt{2\log N}$ 를 대입하면 $\mathbb{E}(Y) \leq \sqrt{2\log N}$ 임을 보일 수 있다,

하지만 베이지스 추론에서는 데이터가 어떻게 뽑혔는지는 중요하지 않고 효과크기 벡터 μ 의 사전분포 $g(\mu)$ 가 베이지스 추론의 결과에 결정적인 역할을 한다. 예를 들자면 flat prior를 사용할 경우 $\hat{\mu}_{610} = x_{610} = 5.29$ 로 바람직하지 않은 과대추정을 하게 될 것이다. 15장에서 배울 보다 적절한 uninformative prior 사용한다면 $\hat{\mu}_{610} = 4.11$ 로 주어진다. 결국 베이지스 추론에서는 사전분포의 선택에 중요한 역할을 하지만 고차원에서 사전분포를 선택하는 것은 쉽지 않은 문제이다.

3.4 A Bayesian/Frequentist Comparison List

베이지스안과 빈도주의의 차이는 그림 3.5가 잘 보여주고 있다. 베이지스의 경우 데이터가 주어진(즉 고정된) 상황에서 수직적으로 모수 μ 가 변화하는 사후분포를 고려하는 반면에 빈도주의에서는 모수가 고정되어 있고 (미래에 생성되는 잠재적인 데이터를 통해서) 데이터의 값이 변화하는 수평적 관점을 가지고 있다. 두 방법의 차이를 비교하자면 다음과 같다

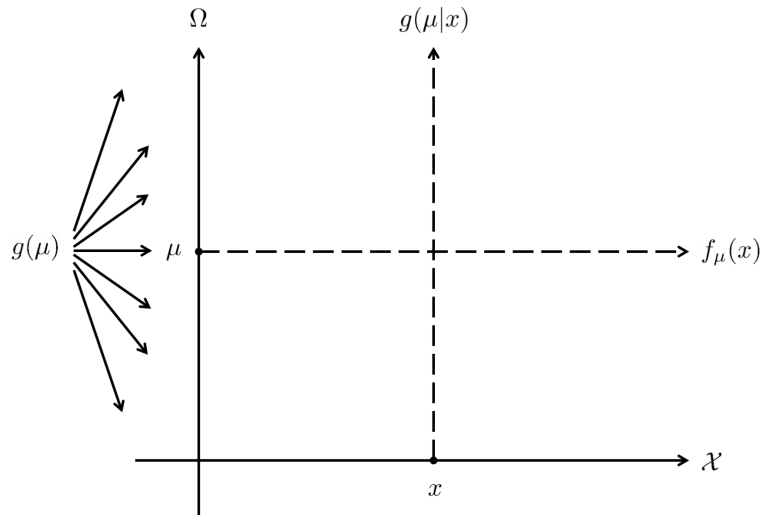


그림 3.5 베이지스 추론은 데이터가 주어진 상황에서 수평적으로 진행되는 반면 빈도주의 추론은 주어진 모수 μ 하에서 수직적으로 진행된다.

- 베이지 추론은 관측된 데이터만을 바탕으로 사전분포에 의해 좌우된다. 반면에 빈도주의에서는 어떤 분석방법 $t(x)$, 즉 알고리즘이 이 역할을 대체한다.
- 베이지 추론의 경우 사후분포를 구한후 다양한 문제에 대한 답변(예를 들자면 $\mathbb{E}(\text{gfr})$ 또는 $\Pr(\text{gfr} < 40)$)을 할 수 있는 반면 빈도주의에서는 각각의 문제마다 다른 방식의 알고리즘을 제사해야 하는 경우가 종종 있다. 예를 들면 아래 식에서 c 의 값은 주어진 문제에 따라 다르게 선택할 수 있다.

$$\sum_{i=1}^n (x_i - \bar{x})^2 / (n - c)$$

- 베이지 룰은 데이터가 순차적으로 업데이트 되거나 여러가지 다른 출처의 통계적 증거를 결합하는데 탁월한 도구이며 빈도주의에서 가장 유사한 방법은 최대우도추정량이다.
- Prior에 관한 적절한 분포가 없을 경우 사전분포의 주관적 선택은 의문부호가 뒤따른다. 이런 이유도 빈도주의가 객관성을 요구하는 과학문제에서 선호되고 있다.
- 고차원 문제에서는 베이지안과 빈도주의의 경계가 흐릿해지며 두가지 방법을 결합한 방법이 많이 사용된다, 경험적 베이지 방법(empirical Bayes methd)이나 lasso가 이러한 예라고 할 수 있다.