ESC Spring 2021 Week1

Intro to Bayesian Statistics and Conjugacy

학술부 임선우, 김수연

March 4, 2021

Overview of Two main Tasks of Statistics : Inference & Prediction

Parameter $\vec{\theta}$: 모집단의 특성을 수치로 나타낸 것. (기존에는 fixed constant vector로 배움)

Basic Setting and Terms

- Gather data of $x_1, ..., x_n$. Assume $X_1, ..., X_n \sim$ iid $f(x; \vec{\theta})$: 여기서는 iid인 random sample만을 다룸.
- **Statistic** $T : [X_1, ..., X_n] \to \mathbb{R}$: 통계적인 목적을 위해 계산됨. 추정 목적 : estimator, 가설검정 목적 : test statistic.
- **Sampling Distribution** : the pdf / pmf of T

Statistic 대표적: $\bar{X}$, median$(X_1, \cdots, X_n)$, S^2

- 만약 $X_1, ...X_n$이 $N(\mu, \sigma^2)$에서 온 random sample, "Sampling Distribution" $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$: 근사가 필요없음. :수능

Two main tasks of Statistics : 대상이 다르다.

linear regression / polynomial / ... : 해석목적

1. **Inference** : Data $x_1, ..., x_n$를 사용하여 모델 생성 후 $\vec{\theta}$에 관한 추론 (underlying true quantity, relationship, etc)
2. **Prediction** : Data : $x_1, ..., x_n$를 사용하여 모델 생성 후 prediction about response of unseen data x_{new}.

Example regarding Numeric Response (house prices)

1. **Inference** : 다른 요소는 동일, 한강 뒷편 view → 한강 view로 옮길 시 얼마나 집값이 오를지 선형회귀
   : **important to have interpretable parameters**
2. **Prediction** : Some attributes given (3000 sqft, 2 bathrooms, ...) → 얼마? : **accurate predictions, no interest in how**

Nominal Response에서도 마찬가지로 logistic regression과 같이 설명력에 강점을 둔 모델 (input 변화에 대한 log odds의 변화가 선형적)이나 SVM, Decision Tree 등 설명력보다는 예측력을 위한 모델이 있겠다.

Overview of Two main approaches in Statistics : Frequentist vs Bayesian

Frequentist view of Statistics :

1. parameter $\vec{\theta}$: **unknown constant (scalar or vector)**
2. Data Gathering → Estimator (주로 $\bar{X}$)와 관련된 **Limiting Distribution**을 활용. sample size ↑→ 극한분포 분산 ↓
3. **Inference** : **data**에서 point estimate $\hat{\theta}$를 구하고 나서
   - **Estimation** : 95% **confidence interval** : size n random sample을 "수많이" 반복하여 각각 얻은 신뢰구간의 95%가량은 참 모수θ 포함.
   - **Hypothesis Test** : Parameter Space를 H_0, H_1의 두 갈래로 양분 → H_0이 옳다 가정 → Get the limiting Distribution of Test Statistic → 얻은 데이터 = 증거는 그 sampling distribution에서 얼마나 H_1쪽으로 극단적인 증거인지 $\alpha = 0.05$와 비교
4. **Prediction** : Use : $X_{new} \sim f(x; \hat{\theta})$.
   참고) Supervised Learning의 예측은 **Bias Variance Tradeoff** : 간단한 모델 : $bias^2$가, 복잡한 모델 : Variance가 ↑

Bayesian view of Statistics :

1. parameter $\vec{\theta}$: **unknown and Random (variable or vector)**
2. Prior Belief of $\vec{\theta}$ represented in probability measure → observe data → "update" the belief into the posterior
3. **Inference** : Posterior distribution 활용한 **estimation**과 **hypothesis test**.
4. **Prediction** : Use the Posterior Predictive Distribution

Framework of Frequentist Approach in Statistics

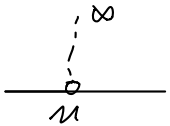
$\sqrt{n}\frac{\bar{X}-\theta}{\sigma} \xrightarrow{d} N(0,1)$

$S^2 \xrightarrow{p} \sigma^2$ & CMT $S \xrightarrow{p} \sigma$ > Slutsky Thm $\Rightarrow \frac{\bar{X}-\theta}{S^2/n} \xrightarrow{d} N(0,1)$

The Central Limit Theorem for Random Sample coming from a distribution having finite variance

1. If $X_1, ..., X_n \sim$ iid $f(x;\vec{\theta})$ came from a dist'n having a finite variance, CLT : $\sqrt{n}(\bar{X}_n - \theta)/\sigma \xrightarrow{d} N(0,1)$
2. "Usage" of CLT : $\bar{X}_n \overset{a}{\approx} N(\mu, \sigma^2/n)$ for n-large, $\bar{X}_n$ 자체의 극한분포?? $\bar{X}_n \xrightarrow{p} \mu$ by WLLN

Dirac Delta Function

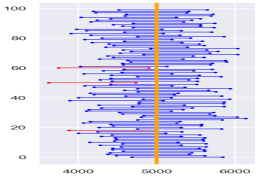


Example of Inference about population mean using CLT

1. $\vec{\Theta}$를 constant vector로 보는 관점에 맞게 하나의 값으로 점추정을 하고 그 점추정량에 관해 신뢰구간으로 나타냄.
2. CLT & Slutsky Theorem $\rightarrow P[-1.96 \geq \frac{\bar{x}-\theta}{\sqrt{s^2/n}} \geq 1.96] \approx 0.95$
3. **Estimation** : Use 95% CI : $[\bar{x} - 1.96\sqrt{s^2/n}, \bar{x} + 1.96\sqrt{s^2/n}]$: 다른 data gathering $\rightarrow \bar{x}$들을 모으면 95%는 μ 포함
4. **Hypothesis Test** : $H_0 : \theta = 100, H_1 : \theta > 100$, "Test Statistic under the Null" $T_{\bar{X}} = \frac{\bar{X}-100}{s/\sqrt{n}} > 1.96$일 때 reject H_0

Bias Variance Tradeoff in Frequentist way of Supervised Learning

1. $y = f(x) + \epsilon$, ϵ has expectation 0, variance σ^2.
2. $E[Y - \hat{Y}(x)]^2 = \sigma^2 + [E(\hat{Y}(x)) - f(x)]^2 + E[(\hat{Y}(x) - E(\hat{Y}(x))]^2 =$ observation variance + $(bias)^2$ + variance



100개씩 신뢰구간들

(a) Example of Confidence Intervals

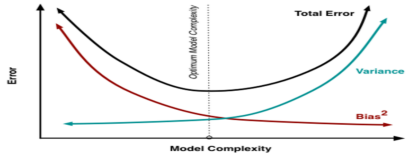


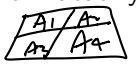
Photo Credit: Scott Fortmann-Roe

(b) Bias Variance Tradeoffs of Model Complexity and Error

Framework of Bayesian Approach in Statistics : Belief and Bayes Theorem

✓ 베이즈통계학에서는 **주관적 믿음**을 **확률**로 표현. 이를 증명하지는 않으나 **belief**의 **axiom**들이 확률의 그것과 겹침.

✓ **Partition** : $A_i, i = 1, ..., n$ form a **partition** of a set A if A_i are mutually exclusive $\&$ and collectively exhaustive : $\cup_{i=1}^{n} A_i = A$ and $A_i \cap A_j = \phi, \forall i \neq j$.



✓ **Rule of Total Probability** : If $\{H_1, ..., H_n\}$ is a partition of H, $p(H) = 1$ and E is a specific event. Then, $P(E) = \sum_{i=1}^{n} P(E \cap H_k) = \sum_{i=1}^{n} P(E|H_k)P(H_k)$.

✓ **Bayes Theorem** : $P(H_i|E) = \frac{P(E|H_i)P(H_i)}{P(E)} = \frac{P(E|H_i)P(H_i)}{\sum_{k=1}^{n} P(E|H_k)P(H_k)}$

Q) 왜 사전확률, 사후확률?

A) posterior : 사건 E가 일어났을 때 **원인**이 H_i였을 확률 = 원인으로 의심되는 H_i가 일어날 확률 $\times$ 원인 의심 H_i가 주어졌을 때 E가 일어날 확률. **사건의 선후관계를 바꿔 역으로 구성해 나감**.

ex1) "범행"이 일어났고 금발머리미녀가 범인일 확률 찾기 (만화 코난)

ex2) 당신이 새벽 강남대로에서 얼핏 어떤 차가 speeding으로 경찰과 추격전을 벌이는 걸 보았다 하자. 그 차는 확실히 Hyudai 차거나 Porsche였다. 무슨 차가 더 likely한 추측일까?

$P(Hyundai|sppeeding) \propto P(Hyundai)P(speeding|Toyota)$.
$P(Porsche|speeding) \propto P(Porsche)P(Speeding|Porsche)$. Weigh four probabilities!

Framework of Bayesian Approach in Statistics : Prior, Likelihood and Posterior

Bayesian Statistics : Framework

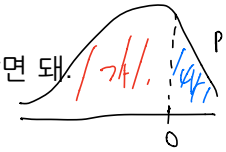
Data sample size ↑ ⇒ posterior inclined to likelihood

prior belief ↑ ⇒ " " " prior

1. Express your belief in **Prior** : $P(\theta)$
2. Gather data of a **Random Sample (iid)** $[x_1, ..., x_n]$. **Likelihood** : $L(data|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$.
3. Update into **Posterior** : $p(\theta|data) = \frac{p(\theta)L(data|\theta)}{\int_\theta p(\theta)L(data|\theta)} \propto p(\theta)L(data|\theta)$. $\int_\theta p(\theta)L(data|\theta)$: **normalizing constant**

Inference

1. $\vec{\Theta}$를 확률변수로 보는 관점에 맞게 모수의 추론이 **posterior distribution**으로 제공.
2. **Estimation** : A $100(1-\alpha)\%$ **credible interval** : interval I satisfying the posterior probability $P(\theta \in I|data) = 1 - \alpha$
   : posterior distribution의 quantile만 알면 돼.
3. **Hypothesis Test** : 역시 posterior distribution의 quantile만 알면 돼.



$P(\theta < 0 | data) = \frac{\text{가}}{\text{가나}} = \text{가}$

Prediction

$p(x_{new}|\theta)$: data generating process

1. **prior predictive** : $p(x_{new,prior}|data) = \int_\theta p(x_{new}|\theta)p(\theta)d\theta$ $p(x_{new}|\theta)$를 모든 θ를 고려하되 $p(\theta)$라는 가중치를 적용한 셈
2. **posterior predictive** : $p(x_{new,post}|data) = \int_\theta p(x_{new}|\theta, data)p(\theta|data)d\theta = \int_\theta p(x_{new}|\theta)p(\theta|data)d\theta$
   두번째에서 마지막 식에 data가 random sample임을 통해 $x_{new} \perp data$ 성질을 활용.

가중치가 $p(\theta|data)$로 !

Comparison between Frequentists vs Bayesians : Example from section 2.7 of BDA3

Setting) 미국의 한 주는 수많은 county로 이루어져 있다. 인구가 n인 한 county를 생각해 보자. X : 해당 county의 1980년대 kidney cancer 사망자 수. Θ : underlying death rate I want to estimate.

ex) $n=1000$, $X=0$ ⋮ ⇓
$X=1$ ⋮ ⇑

Figure: Counties in California

"Naive" Frequentist Estimate : $\hat{\Theta} = \frac{X}{n}$: **비율**로 **확률**을 접근. 인구가 큰 county에서는 reliable, $n \downarrow$: 이 **비율**이 문제가 됨. 가령 $n = 1000$일 때 $X = 0$: $\hat{\theta}$가 비현실적으로 낮음, $X = 1$: $\hat{\theta}$가 비현실적으로 높음.

Bayesian Approach :

1. "Likelihood" : $X|\Theta = \theta \sim Bin(n, \theta)$
2. "Prior" : $\Theta \sim Beta(\alpha, \beta)$. 나중에 밝히지만 Method of Moments로 있는 데이터를 활용해 prior분포의 α, β 값을 정하였다.
3. "Posterior" : $\Theta|X = x \sim Beta(x + \alpha, n - x + \beta)$.

o Stan에서 베이즈 모델링 prior분포의 moment 들을 맞추는 방식.
MoM estimator

디테일 : 나중에 beta binomial model : posterior derivation, code, visualization 등등

Important Concepts needed in Bayesian Statistics : Probability and Independence of "Events"

✓ **Event** : any subset of **Sample Space** Ω : set of all possible results in a random experiment.

✓ Given Ω and associated sigma algebra $\mathbb{B}$, **probability function** is a function P with domain $\mathbb{B}$ satisfying

1. $P(A) \geq 0, \forall A \in \mathbb{B}$.
2. $P(\Omega) = 1$
3. If $A_1, A_2, ... \in \mathbb{B}$ are pairwise disjoint, $P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

✓ Events A and B are **independent** if $P(A \cap B) = P(A)P(B)$, implicitly, $P(A|B) = P(A)$.

✓ $\{A_i\}$ is a sequence of independent events if $P(\cap_{i=1}^{n} A_i) = \prod_{i=1}^{n} P(A_i)$. $\{A_i\}_{i=1}^{\infty}$ is independent if it holds for any n.

✓ 주의) Pairwise independence vs mutual independence

✓ $\{A_i\}_{i=1}^{n}$ are **pairwisely independent** if $P(A_i \cap A_j) = P(A_i)P(A_j), \forall i \neq j$.

✓ $\{A_i\}_{i=1}^{n}$ are **mutually independent** if $\forall k \leq n$ and for every subset $\{B_i\}_{i=1}^{k}$ of $\{A_i\}_{i=1}^{n}$, $P(\cap_{i=1}^{k} B_i) = \prod_{i=1}^{k} P(B_i)$

: A, B, C events가 $P(A \cap B \cap C) = P(A)P(B)P(C)$ 혹은 pairwise independence를 만족하는 것만으로는 mutually independence를 만족하지 않는다!

Important Concepts needed in Bayesian Statistics : Conditional Independence of "Events"

✓ **Conditional Independence** : $P(F \cap G|H) = P(G|H)P(G|H)$.
If $F\&G$ are conditionally independent given H, $P(F|H \cap G) = P(F|H)$.

$P(G|H)P(F|H \cap G) = P(F \cap G|H)$: always!

$= P(F|H)F(G|H)$ if events $F\&G$ are conditionally independent given event H

$\rightarrow P(F|H \cap G) = P(F|H)$.

Interpretation : If I know 1) H : True, 2) $F\&G$ are conditionally independent given H,

information of G does not change the belief about F.

✓ Example : $F =$ { a patient smokes}, $G =$ { a patient has a lung cancer}, $H =$ {smoking causes lung cancer}

Important Concepts needed in Bayesian Statistics : Joint Distributions

Continuous Case

Given a continuous joint cdf $F_{Y_1Y_2}(a,b) = \int_{-\infty}^{a}\int_{-\infty}^{b} p_{Y_1Y_2}(y_1,y_2)dy_2dy_1$, function $p_{Y_1Y_2}$ is the joint density of two continuous random variables Y_1 and Y_2. We have

1. $p_{Y_1}(y_1) = \int_{-\infty}^{\infty} p_{Y_1Y_2}(y_1,y_2)dy_2$
2. $p_{Y_2|Y_1}(y_2|y_1) = \dfrac{p_{Y_1Y_2}(y_1,y_2)}{p_{Y_1}(y_1)}$

Note that $p_{Y_1Y_2} \Leftrightarrow p_{Y_1},\ p_{Y_2|Y_1}$ but **NOT** $p_{Y_1Y_2} \Leftrightarrow p_{Y_1},\ p_{Y_2}$

Bayesian Inference

Here two random variables are now data(y) and parameter(θ).

- $p(\theta)$: beliefs about θ
- $p(y|\theta)$: beliefs about y for each value of θ.

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y,\theta)d\theta} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}$$

Important Concepts needed in Bayesian Statistics : Independence

Q) 왜 θ가 주어졌을 때 (θ를 알 때) $Y_1, Y_2, ..., Y_n$은 independent sample이라고 말할 수 있을까?

A) θ : parameter describing the conditions under which the random variables are generated

θ를 안다는 것 $\rightarrow$ $Y_i, Y_j, ...$들이 어느 분포에서 나왔는지 안다
$\rightarrow$ Y_j에 대해 안다고 해서 Y_i에 대해 더 알 수 있는 게 없음 $\rightarrow$ (θ에 대해 조건부) 독립
$\rightarrow$ 이게 곧 i.i.d. sample !

- conditional independence means that "Y_j gives no additional information about Y_j beyond that in knowing θ"

Suppose $Y_1, ..., Y_n$ are generated in similar ways from a common process.
This suggests that marginal densities are all equal to some common density giving

$$p(y_1, ..., y_n|\theta) = \Big[\prod_{i=1}^{n} p(y_i|\theta)\Big]$$

We say that $Y_1, ..., Y_n$ are conditionally independent and identically distributed (i.i.d.) given θ.

$$Y_1, ..., Y_n|\theta \ \sim \ p(y|\theta)$$

Justification for Constructing Probability Model

이렇게 쓸 수 있는 이유?

$$p(y_1, ..., y_n) = \int p(y_1, ..., y_n|\theta)p(\theta)d\theta$$
$$= \int \Big[\prod_{i=1}^{n} p(y_i|\theta)\Big]p(\theta)d\theta$$

Exchangeability

If $p(y_1, ..., y_n) = p(y_{\pi_1}, ..., y_{\pi_n})$ for all permutations of 1, 2, ..., n then $Y_1, ..., Y_n$ are exchangeable.

data가 exchangeable하다는 것은 각각 data의 label이 아무런 의미가 없다는 것을 말한다.

$\rightarrow$ If $Y_1, ..., Y_n$ are conditionally independent (not marginally) given $\theta \sim p(\theta)$, $Y_1, ..., Y_n$ are (marginally) exchangeable.

de Finetti's Theorem

한마디로 요약하자면, exchangeable observations are conditionally independent relative to some latent variable !

$\rightarrow$ If $Y_1, ..., Y_n$ are exchangeable, $p(y_1, ..., y_n) = \int \Big[\prod_{i=1}^{n} p(y_i|\theta)\Big]p(\theta)d\theta$

Intro to Conjugacy

Bayesian Inference의 과정

1. Likelihood : $p(D|\theta)$를 뭘로 설정할지
2. Prior (주어졌는지? 아님 내가 줘야하는지?) : $p(\theta)$
3. Full Probability model $p(\theta|D) \propto p(\theta)p(D|\theta)$

$$\textbf{(Bayes Thm)} \quad p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}$$

Bayesian Inference의 단점 : Normalizing Constant

해결방법 ?

1. Use conjugate prior
2. MCMC : sample from $p(\theta|D)$
3. approximate $p(\theta|D)$ (ex. Variational Inference)

Binomial Model : Example

1. Data : binomial data

$$\text{n} = 129,\ Y_i = \begin{cases} 1 & \text{if happy :)} \\ 0 & \text{if unhappy :(} \end{cases}$$

outcome : 118 happy, 11 unhappy

Conditional on θ, Y_i's are i.i.d. binary random variables with expectation θ.

$$\begin{aligned} p(y_1, ..., y_{129}|\theta) &= \theta^{\sum_{i=1}^{129} y_i} (1-\theta)^{129-\sum_{i=1}^{129} y_i} \\ &= \theta^{118}(1-\theta)^{11} \end{aligned}$$

2. Prior : uniform prior

$$p(\theta) = 1 \text{ for all } \theta \in [0,1]$$

a.k.a.

$$p(\theta) \sim Beta(1,1)$$

$$p(\theta) = \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)} \theta^{1-1}(1-\theta)^{1-1}$$

Binomial Model : Example

cf) Beta pdf : $\int_0^1 \theta^{a-1}(1-\theta)^{b-1} = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

3. Plug-in to Bayes' rule : 이제 믿음을 업데이트하기 !

$$\begin{aligned} p(\theta|y_1,...,y_{129}) &= \frac{p(y_1,...,y_{129}|\theta)p(\theta)}{p(y_1,...,y_{129})} \\ &= \frac{\theta^{118}(1-\theta)^{11}\times 1}{p(y_1,...,y_{129})} = \frac{\theta^{118}(1-\theta)^{11}}{\int p(y_1,...,y_{129}|\theta)p(\theta)d\theta} \\ &= \frac{\Gamma(131)}{\Gamma(119)\Gamma(12)}\theta^{118}(1-\theta)^{11} \end{aligned}$$

알지만 모른다 치고 $p(y_1,...,y_{129})$를 구해보자 !

$$\begin{aligned} 1 = \int_0^1 p(\theta|y) &= \int_0^1 \frac{p(y|\theta)p(\theta)}{p(y)}d\theta \\ &= \frac{1}{p(y)}\int_0^1 \theta^{118}(1-\theta)^{11}d\theta = \frac{1}{p(y)}\times\frac{\Gamma(119)\Gamma(12)}{\Gamma(131)} \\ p(y) &= \frac{\Gamma(119)\Gamma(12)}{\Gamma(131)} \end{aligned}$$

Binomial Model : Generalization

1. Data $\sim$ Binomial

$Y|\theta \ \sim \ Binom(n, \theta)$

pdf $P(Y = y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$

$E[Y|\theta] = n\theta$

2. Prior $\sim$ Beta(a, b)

$p(\theta) = \dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$ for $0 \leq \theta \leq 1$

$E[\theta] = \dfrac{a}{a+b}$

3. Posterior

$$
\begin{aligned}
p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\
&= \underbrace{\frac{1}{p(y)}}_{\text{n.c}} \times \underbrace{\binom{n}{y}\theta^y(1-\theta)^{n-y}}_{\text{likelihood}} \times \underbrace{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}}_{\text{prior}} \\
&= c(y)\ \theta^{a+y-1}(1-\theta)^{b+n-y-1} \quad \propto p(y|\theta)p(\theta) \\
&= Beta(a+y, b+n-y)
\end{aligned}
$$

$$
E[\theta|y] = \frac{a+y}{a+b+n}
$$

Class of beta priors is conjugate for binomial sampling model

We can see that posterior is a combination of prior and data information. (weighted average!)

$$
E[\theta|y] = \frac{a+y}{a+b+n} = \frac{n}{a+b+n} \times \underbrace{\frac{y}{n}}_{\text{sample mean}} + \frac{a+b}{a+b+n} \times \underbrace{\frac{a}{a+b}}_{\text{prior expectation}}
$$

Conjugacy : more...!

Q) 어떤 경우에 conjugate prior를 써서 posterior distribution을 구할 수 있을까?

Likelihood : class $\mathcal{F}$ of exponential family

$$p(y_i|\theta) = f(y_i)\ g(\theta)\ exp(\phi(\theta)^T\ s(y_i))$$

$$p(y|\theta) = \prod_{i=1}^{N} f(y_i)\ g(\theta)^N\ exp(\phi(\theta)^T \underbrace{\sum_{i=1}^{N} s(y_i)}_{\text{sufficient statistics}_{=t(y)}})$$

Prior : $p(\theta)\ \propto\ g(\theta)^\eta\ exp(\phi(\theta)^T\ \nu)$

Posterior : $p(\theta|y)\ \propto\ g(\theta)^{\eta+N}\ exp(\phi(\theta)^T\ (\nu + t(y)))$

(Prior, Posterior는 모두 θ에 대한 함수이므로 모수는 $\eta \to \eta + N$, $\nu \to \nu + t(y)$로 업데이트됨을 알 수 있다.)