Part II. 통계적 추론 (Statistical Inference)

IV. 확률표본(Random Sample)

A. 어떤 특정한 분포로부터의 '**크기가 n 인 확률표본**'은 n 개의 동일한 특정 분포를 갖는 서로 확률적으로 독립인 확률변수 $X_1, X_2, ... X_n$ 들의 모음

i. 통계량(a statistic): (미지의 모수에 의존하지 않는) 확률표본의 변환(transformation) – 원칙적으로 확률표본의 값들이 관찰될 경우 그 값들에 대응하는 통계량의 값을 계산할 수 있으며, 이를 통계치(a statistic)라고 함

1. 표본평균(Sample mean): $\bar{X} = \dfrac{\sum_1^n X_i}{n}$, n 에 의존함을 강조시 $\bar{X}_n$ 로 표기

$$E\left(X_i\right) = \mu,\ Var\left(X_i\right) = \sigma^2 \ \Rightarrow \ E\left(\bar{X}\right) = \mu,\ Var\left(\bar{X}\right) = \frac{\sigma^2}{n}$$

2. 표본분산(Sample variance): $S^2 = \dfrac{\sum_1^n \left(X_i - \bar{X}\right)^2}{n-1}$ 또는

$$S^{*2} = \frac{\sum_1^n \left(X_i - \bar{X}\right)^2}{n}$$

$$E\left(S^2\right)=\sigma^2,\ E\left(S^{*2}\right)=\frac{n-1}{n}\sigma^2$$

B. $N\left(\mu,\sigma^2\right)$로부터의 확률표본

i. 표본평균(Sample mean) :

$$\bar{X} \sim N\left(\mu,\frac{\sigma^2}{n}\right).$$

ii. 표본분산(Sample variance) :

$$\frac{n-1}{\sigma^2}S^2=\frac{n}{\sigma^2}S^{*2}=\frac{\sum_1^n\left(X_i-\bar{X}\right)^2}{\sigma^2}\sim\chi^2_{n-1}$$: 자유도가 n-1 인 카이제곱분포

iii. $\bar{X}\perp S^2$: 즉 표본평균과 표본분산은 확률적으로 독립이다.

V. 극한분포이론 (Limiting Distribution Theory)

A. 대수의 (약)법칙 ((Weak) Law of Large Number) =확률적 수렴(Convergence in Prob.)

i. $X_1, X_2, ...X_n$ ~ i.i.d. with $E(X_i) = \mu < \infty,\ V(X_i) = \sigma^2 < \infty$,

$\Rightarrow$ 임의의 양의 실수 ε에 대해, $P\left(\left|\bar{X}_n - \mu\right| > \varepsilon\right) \to 0$ as $n \to \infty$.

$\Rightarrow$ $\bar{X}_n$가 μ로 확률적으로 수렴함. 또는 $\bar{X}_n \xrightarrow{p} \mu$ 또는 $\underset{n\to\infty}{p\lim} \bar{X}_n = \mu$

ii. $X_1, X_2, ...X_n$ ~ 서로독립적이지도 않고, 동일한 분포를 하지도 않음.

다만 $m_n = E\left(\sum_1^n X_i\right) < \infty$, $V_n = V\left(\sum_1^n X_i\right) < \infty$, $\frac{V_n}{n^2} \to 0$ as $n \to \infty$.

$\Rightarrow$ $\underset{n\to\infty}{p\lim} \bar{X}_n = \lim_{n\to\infty} \frac{m_n}{n}$

B. 중심극한 정리 (Central Limit Theorem)

i. 분포수렴: 확률변수 X_n의 분포함수의 값이 연속적인 모든 점에서 $n \to \infty$ 에 따라 확률변수 Y 의 분포함수의 값으로 수렴할 때 확률변수 X_n가 확률변수 Y로 분포수렴한다고 함 $\Rightarrow X_n \xrightarrow{D} Y$

ii. $X_1, X_2, \ldots X_n \sim$ i.i.d. $E(X_i) = \mu < \infty,\ V(X_i) = \sigma^2 < \infty$

$$\Rightarrow \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \xrightarrow{D} N(0,1) \text{ as } n \to \infty.$$: 중심극한정리

iii. $X_1, X_2, \ldots X_n$ ~ 서로독립적이지도 않고, 동일한 분포를 하지도 않음.

다만 $\lim_{n\to\infty} nV(\bar{X}_n) = \sigma^2 < \infty$,

$$\Rightarrow \frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{V(\bar{X}_n)}} \xrightarrow{D} N(0,1) \text{ as } n \to \infty.$$: (좀 더 일반화된) 중심극한정리

VI. 점추정 (Point Estimation)

A. 추정량 (Estimator)

i. 모수 θ 의 추정을 위해 적절히 고안된 통계량 $\hat{\theta}(X_1, X_2, ...X_n)$ 을 추정량, 관측된 표본 값에 대응되는 해당 통계량의 값을 추정치(Esitimate)라 함

1. $\hat{\theta}(X_1, X_2, ...X_n)$ 를 줄여서 $\hat{\theta}$ 또는 $\hat{\theta}_n$ 으로 표기

B. 추정량에 요구되는 성질

i. 불편성(Unbiasedness) 또는 점근적 불편성(Asymptotic Unibiasedness)

$E\left(\hat{\theta}\right) = \theta$ (불편추정량) 또는 $\lim_{n\to\infty} E\left(\hat{\theta}_n\right) = \theta$ (점근적 불편추정량)

ii. 일치성(Consistency)

$\hat{\theta}_n \xrightarrow{p} \theta$ (일치추정량) $\Leftarrow$ (점근적) 불편추정량이고, $\lim_{n\to\infty} V\left(\hat{\theta}_n\right) = 0$

iii. 유효성(Efficiency) 또는 점근적 유효성

1. 모수 θ 를 불편 추정량 $\hat{\theta}$ 로 추정할 경우 추정량 $\hat{\theta}$ 의 분산의 이론적인 하한을 크래머-라오 하한(CR Lower Bound: CRLB)이라하며, 불편추정량 $\hat{\theta}$ 의 유효성은

$\left({CRLB}/{V(\hat{\theta})} \right) \times 100\%$로 나타냄

a. 100% 유효성을 갖는 불편추정량을 유효추정량이라고 함

b. 점근적불편추정량의 분산이 점근적으로 CRLB 에 도달할 경우 점근적 유효추정량이라고 함

i. 불편추정량과 점근적불편추정량의 성능(Performance)는 보통 평균제곱오차(Mean Squared Error: MSE)로 비교함

$$MSE(\hat{\theta}) = V(\hat{\theta}) + \left[E(\hat{\theta}) - \theta \right]^2$$

c. 유효추정량은 최소분산불편추정량(Minimum Variance Unibiased Estimator: MVUE) 임

C. 적률방법 (Method of Moments)

i. 수학적 적률(보통 추정하고자 하는 모수의 함수)과 표본적률을 등치시키는 식으로부터 추정량을 구하는 방법

ii. k 번째 (수학적) 적률: $E(X^k) = \frac{\sum_1^n (X_i)^k}{n}$: k 번째 표본적률

iii. 적률방법추정량은 점근적 불편성과 일치성을 가짐

D. 최우추정법(Maximum Likelihood Method)

i. 우도함수 (Likelihood function) : 표본의 결합확률 분포를 추정하고자 하는 모수의 함수로 볼 때, 이를 우도함수라 함

$$L(\theta_1,\theta_2,..\theta_k)\equiv\prod_1^n f\left(x_i\middle|\theta_1,\theta_2,..\theta_k\right)$$

ii. 최우추정법은 우도함수를 극대화 시키는 모수의 값을 해당 모수의 추정치로 삼는 방법이며 이를 확률표본 $X_1,X_2,...X_n$의 함수로 볼 때, 이를 최우추정량(Maximul Likelihood Estimator : MLE)이라고 함

1. 경우에 따라서는 우도함수의 극대화 해(solution)를 찾는 것 보다 우도함수에 자연대수를 취한 $\ln L$ 를 극대화 하는 해를 찾는 것이 용이

iii. $\hat{\theta}_n=\hat{\theta}(X_1,X_2,...X_n)$ 을 θ 의 최우추정량이라고 할 때, 이는 다음과 같은 성질을 갖는다.

1. 즉 MLE 는 점근적 불편성, 일치성, 그리고 점근적 유효성을 갖는다

a. 점근적 유효성: $nV\left(\hat{\theta}_{n\sigma}\right)\to\sigma(\theta)$ as $n\to\infty$, 단 $\dfrac{\sigma(\theta)}{n}$ 은 CRLB.

2. 충분히 큰 n 에 대해 $\hat{\theta}_n \sim N\left(\theta, \frac{\sigma^2(\theta)}{n}\right)$ 으로 근사할 수 있다.

   a. $\sigma(\theta)$에 대한 추정량을 $\sigma\left(\hat{\theta}_n\right)$라고 하면,

   b. $\frac{\hat{\theta}_n - \theta}{\sigma\left(\hat{\theta}_n\right)/\sqrt{n}} \sim N(0,1)$ 또는 $\hat{\theta}_n \sim N\left(\theta, \frac{\sigma^2\left(\hat{\theta}_n\right)}{n}\right)$ 으로 근사하여, 사용할 수 있다.

VII. 신뢰구간 (Confidence Interval: CI) 또는 구간추정(Interval Estimation)

A. 모집단이 $N\left(\mu,\sigma^2\right)$일 때, μ에 대한 CI

i. σ^2가 알려진 경우 100 ×(1-α)% CI: $\bar{X} \pm z_{\alpha/2} \sigma/\sqrt{n}$. 단, $P\left(Z > z_\alpha\right) = \alpha$, $Z \sim N\left(0,1\right)$

ii. σ^2를 모르는 경우 100 × (1-α) %CI: $\bar{X} \pm t_{\alpha/2,n-1} S/\sqrt{n}$. 단, S 는 표본분산의 제곱근. $P\left(T_{n-1} > t_{\alpha,n-1}\right) = \alpha$, T_{n-1}은 자유도 n-1 인 t 분포 확률변수.

iii. σ^2를 모르고, n 이 충분히 클 경우 100 × (1-α) % CI: $\bar{X} \pm z_{\alpha/2} S/\sqrt{n}$

B. 모집단이 $N\left(\mu,\sigma^2\right)$ 일 때, σ^2 에 대한 100 × (1-α) % CI: $\left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}\right)$. 단 $P\left(\chi^2_{n-1} > \chi^2_{\alpha,n-1}\right) = \alpha$, χ^2_{n-1} 은 자유도 n-1 인 χ^2 분포 확률변수.

i. 표준편차(σ)에 대한 CI : $\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}}\right)$

VIII. 가설검정 (Hypothesis Testing)

A. 가설검정의 구성요소

i. 귀무가설(null hypothesis)과 대립가설(alternative hypothesis)

1. 귀무가설(H_0)는 검정 결과 그것을 기각해야 하는 증거가 나타나기 전까지는 사실로 받아들여야 하는 가설

2. 대립가설(H_1)은 검정결과 귀무가설을 기각하는 경우 사실로 받아들여야 하는 가설이며, 따라서 증거로부터 우리가 보이고자 하는 새로운 사실이나 관계를 설명하는 내용을 통상적으로 대립가설로 둠

ii. 검정통계량(Test Statistic)

1. 검정통계량은 귀무가설하에서 (적어도 근사적으로라도) 그 확률 분포가 알려진 통계량 $T \equiv T\left(X_1, X_2 \ldots X_n\right)$

2. 검정통계량의 대립가설하의 분포는 알려져 있을 필요는 없으나 귀무가설하의 분포와 구별되는 분포를 해야 함

iii. 기각역(Rejection Region)

1. 검정통계량의 값이 이 구간에 포함되면 귀무가설을 기각함 (R)

B. 제 1 형 오류와 제 2 형 오류

i. 제 1 형 오류: 모집단에서 귀무가설이 옳음에도 불구하고 표본으로부터 귀무가설을 기각하는 경우

1. 제 1 형 오류를 범할 확률을 검정의 크기(size) 또는 유의수준(significance level)이라고 함 (α)

$$\alpha = P\left(T\left(X_1, X_2 \ldots X_n\right) \in R \middle| H_0 \text{ is true}\right)$$

2. 보통 유의수준은 1%, 5%, 10% 등으로 주어짐

ii. 제 2 형 오류: 모집단에서 귀무가설이 옳지 않음에도 불구하고 귀무가설을 기각하지 못하는 경우

1. 귀무가설이 옳지 않을 때 귀무가설을 기각할 확률을 검정력(Power of test)이라고 하며, 제 2 형 오류의 확률을 β라 하면, 검정력은 1-β

$$1-\beta = P\left(T\left(X_1, X_2 \ldots X_n\right) \in R \middle| H_0 \text{ is not true}\right) \text{ : 검정력 함수}$$

2. 주어진 유의수준에서 가능하면 검정력이 큰 검정이 좋은 검정임

   a. 유의수준이 클수록 검정력은 커진다

   b. 두 가설간의 차이가 뚜렷할수록 검정력은 커진다.

C. CI 에 근거한 검정

i. 각 경우에 있어서의 CI 를 도출하였던 과정을 응용하여 적절한 검정통계량을 고안해낼 수 있음

ii. 실례: $N\left(\mu,\sigma^2\right)$ 일 때, μ 에 대한 $100 \times (1\text{-}\alpha)$ % CI 는 $\bar{X} \pm t_{\alpha/2,n-1} S/\sqrt{n}$

$\Rightarrow p\left(\left|\dfrac{\bar{X}-\mu}{S/\sqrt{n}}\right| > t_{\alpha/2,n-1}\right) = 1-\alpha \;\Rightarrow\; \dfrac{\bar{X}-\mu}{S/\sqrt{n}}$ 가 자유도 n-1 인 t 분포를 하는 것을이용하여, μ 와 관련된 가설검정의 검정통계량으로 이용

1. $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$ (양측검정의 경우) : $\dfrac{\bar{X}-\mu_0}{S/\sqrt{n}} > t_{\alpha/2,n-1}$ 이거나 $\dfrac{\bar{X}-\mu_0}{S/\sqrt{n}} < -t_{\alpha/2,n-1}$ 이면 유의수준 α에서 귀무가설을 기각함

2. $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$ (또는 $H_0 : \mu \leq \mu_0$, $H_1 : \mu > \mu_0$ (단측검정의 경우) : $\dfrac{\bar{X}-\mu_0}{S/\sqrt{n}} > t_{\alpha,n-1}$ 이면 유의수준 α에서 귀무가설을 기각함

3. $H_0 : \mu = \mu_0$, $H_1 : \mu < \mu_0$ (또는 $H_0 : \mu \geq \mu_0$, $H_1 : \mu < \mu_0$

(단측검정의 경우) : $\frac{\bar{X}-\mu_0}{S/\sqrt{n}} < -t_{\alpha,n-1}$ 이면 유의수준 α에서 귀무가설을 기각함