제 2 강 통계적 기초 – 보충 내용

통계적 추론 (Statistical Inference)

o 확률표본(Random Sample) : 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

o 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

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

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

- 표본분산(Sample variance): $S^2 = \frac{\sum_1^n (X_i - \bar{X})}{n-1}$ 또는 $S^{*2} = \frac{\sum_1^n (X_i - \bar{X})}{n}$

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

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

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

- 임의의 양의 실수 ε에 대해, $P\left(\left|\bar{X}_n - \mu\right| > \varepsilon\right) \to 0$ as $n \to \infty$.
- $\bar{X}_n$ 가 μ로 확률적으로 수렴함. 또는 $\bar{X}_n \xrightarrow{p} \mu$ 또는 $\underset{n\to\infty}{p\lim} \bar{X}_n = \mu$

O 중심극한 정리 (Central Limit Theorem)

i. $X_1, X_2, \ldots X_n$ ~ 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

O 추정량 (Estimator) : 모수 θ의 추정을 위해 적절히 고안된 통계량 $\hat{\theta}(X_1, X_2, \ldots X_n)$을 추정량, 관측된 표본 값에 대응되는 해당 통계량의 값을 추정치(Esitimate)라 함

- $\hat{\theta}(X_1, X_2, \ldots X_n)$를 줄여서 $\hat{\theta}$ 또는 $\hat{\theta}_n$으로 표기
- 추정량에 요구되는 성질

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

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

ii. 일치성(Consistency)

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

iii. 유효성(Efficiency) 또는 점근적 유효성

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

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

O 신뢰구간 (Confidence Interval: CI) 또는 구간추정(Interval Estimation)

- 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: $\bar{X} \pm z_{\alpha/2}\,S/\sqrt{n}$

O 가설검정 (Hypothesis Testing)의 구성요소

- 귀무가설(null hypothesis)과 대립가설(alternative hypothesis)
  - i. 귀무가설(H_0)는 검정 결과 그것을 기각해야 하는 증거가 나타나기 전까지지는 사실로 받아들여야 하는 가설
  - ii. 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
- 검정통계량(Test Statistic)
  - i. 검정통계량은 귀무가설하에서 (적어도 근사적으로라도) 그 확률분포가 알려진 통계량
  - ii. 검정통계량의 대립가설하의 분포는 알려져 있을 필요는 없으나 귀무가설하의 분포와 구별되는 분포를 해야 함
- 기각역(Rejection Region)
  - i. 검정통계량의 값이 이 구간에 포함되면 귀무가설을 기각함 (R)

O 제 1 형 오류와 제 2 형 오류

- 제 1 형 오류: 모집단에서 귀무가설이 옳음에도 불구하고 표본으로부터 귀무가설을 기각하는 경우. 제 1 형 오류를 범할 확률을 유의수준(significance level)이라고 함 (α)
  i. 보통 유의수준은 1%, 5%, 10% 등으로 주어짐
- 제 2 형 오류: 모집단에서 귀무가설이 옳지 않음에도 불구하고 귀무가설을 기각하지 못하는 경우
  i. 귀무가설이 옳지 않을 때 귀무가설을 기각할 확률을 검정력(Power of test)이라고 하며, 제 2 형 오류의 확률을 β라 하면, 검정력은 1-β
  ii. 주어진 유의수준에서 가능하면 검정력이 큰 검정이 좋은 검정임

O $N\left(\mu,\sigma^2\right)$일 때, $\dfrac{\bar{X}-\mu}{S/\sqrt{n}}$가 자유도 n-1 인 t 분포를 하는 것을 이용하여, μ와 관련된 가설검정의 검정통계량으로 이용

i. $H_0:\mu=\mu_0$, $H_1:\mu\neq\mu_0$ (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

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

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