문제 1 Central Limit Theorem

중심극한정리는 확률변수의 합 형태(sum of random variables)의 극한분포를 손쉽게 구할 수 있도록 해주기에 통계학에서 가장 자주 사용하는 정리입니다. 이 문제에서는 중심극한정리의 정의와 그 활용에 대해 짚어보겠습니다.

(1-1) ✎ 중심극한정리(Central Limit Theorem)의 정의와 그 의미를 서술하시오.

- 통계학입문(3판) 7장 참고
- Hogg(8판) 4장 2절, 5장 3절 참고

Def: 평균이 μ이고 분산이 σ^2인 모집단에서 랜덤추출한 서로 독립인 확률변수 $X_1, X_2 \cdots X_n$이 존재할 때, $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$은 n이 ∞에 가까워짐에 따라 $N(0,1)$인 표준정규분포를 따른다.

$$\left(\text{As } n \to \infty \quad \bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right) \right)$$

Meaning: 표본의 크기가 충분히 크다면, 모집단이 어떤 분포를 갖고있던지 $N(\mu, \sigma^2)$를 따른다.

→ 따라서 추출한 표본의 통계량을 통해서 모수 $(\mu, \sigma^2 \cdots)$을 합리적인 추정할 수 있다는 점에 의의가 있는 정리이다.

(1-2) ✎ 중심극한정리가 통계적 추론 중 "구간추정"에서 어떻게 유용한지 서술하시오.

> Hogg(8판) 4장 2절 참고

(1-3) ✎ 중심극한정리를 이용하여 모평균에 대한 근사신뢰구간을 만들 때, 표준오차($\sqrt{Var(\bar{X})}$) 부분의 모분산을 표본분산으로 대체할 수 있는 이유를 수식적으로 증명하시오.

> 표본분산 s^2는 모분산 σ^2로 확률수렴한다는 사실을 이용할 수 있습니다.
> Slutsky's theorem을 이용할 수 있습니다.
> Hogg(8판) 5장 1~3절 참고

1-2) 모집단이 정규분포를 따르지 않아도 신뢰구간 추정이 가능하다.
CLT에 의해 표본의 크기가 충분히 크다면 표본평균의 분포는 정규분포를 따르므로, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$을 이용해 근사신뢰구간을 구할수있다.
ex) 신뢰도 95%로 μ(모평균)을 추정하면 $\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$로 구간추정이 가능하다.
(모분산을 모른다고 함)

1-3) $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2 - \bar{X}_n^2\right)$

$= \frac{n}{n-1}\left(E(X_i^2) - \mu^2\right)$

$\lim_{n\to\infty} S_n = \lim_{n\to\infty} \frac{n}{n-1}\left(E(X_i^2) - \mu^2\right) = E(X_i^2) - \mu^2 = \sigma^2$

As $n \to \infty$,

$\therefore$ 표본분산 S_n^2는 모분산 σ^2에 수렴한다.

따라서 CLT가 적용되면 표본분산 S^2이 σ^2에 확률수렴하므로 대신하여 사용할 수 있다.

문제 2 Student's Theorem

스튜던트 정리는 통계적 추정에서 필요한 정리 중 하나로, 표본평균과 표본분산이 어떤 분포를 갖는지 알려줍니다. 이 문제에서는 스튜던트 정리의 내용을 어떻게 수식적으로 유도할 수 있는지 짚어보겠습니다.

스튜던트 정리는 다음과 같이 총 4개의 내용으로 구성되어 있습니다.

① $\overline{X} \approx N(\mu, \frac{\sigma^2}{n})$ → 〈문제 1〉에서 증명함

② 표본평균 $\overline{X}$ 와 표본분산 s^2은 서로 독립이다.

③ ???

④ ???

(2-1) ✎ ③의 내용을 쓰고 증명하시오.

☝ Hogg(8판) 3장 6절 참고

☝ 무작위표본 $X_1, \cdots, X_n$이 독립적으로 동일하게(independently and identically distributed) 평균이 μ이고 분산이 σ^2인 정규분포를 따를 때, 자유도가 n인 카이제곱분포를 따르는 새로운 확률변수 V를 아래와 같이 두어 증명에 활용할 수 있습니다.

$$V = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n)$$

③ $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

Proof: $Z^2 = \left(\frac{X-\mu}{\sigma}\right)^2$, So, $V = \sum_{i=1}^{n} \left(\frac{X_i-\mu}{\sigma}\right)^2 = \sum_{i=1}^{n} Z_i^2$

$V = \sum_{i=1}^{n} \left(\frac{X_i-\bar{X}+\bar{X}-\mu}{\sigma}\right)^2 = \sum_{i=1}^{n} \left(\frac{X_i-\bar{X}}{\sigma}\right)^2 + n\left(\frac{\bar{X}-\mu}{\sigma}\right)^2 = \sum_{i=1}^{n} \left(\frac{X_i-\mu}{\sigma}\right)^2 + \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$

$= S^2(n-1) \times \frac{1}{\sigma^2} + \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$

by ①, $\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right) \sim N(0,1)$ (표준화) 이므로 $\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2(1)$

by ②, $\frac{S^2(n-1)}{\sigma^2}$과 $\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$ 독립이며 $\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$가 자유도가 1인 카이제곱 분포를

따르고 V는 자유도가 n인 카이제곱 분포를 따르며 $\frac{(n-1)S^2}{\sigma^2}$은 자유도가 $(n-1)$인 카이제곱분포를 따른다.

(2-2) ✎ ④의 내용을 쓰고, (2-1)을 이용하여 증명하시오.

> Hogg(8판) 3장 6절 참고
> t분포의 정의에 따르면, 표준정규분포를 따르는 확률변수와 카이제곱분포를 따르는 확률변수를 이용하여 t분포를 유도할 수 있습니다.

---

④ $T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$

proof: $T = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \cdot \frac{\sigma}{S} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \Big/ \sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}}$

by ① $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, by 2-1 $\frac{S^2(n-1)}{\sigma^2} \sim \chi^2(n-1)$

T분포의 정의: ㉮ W, V 독립, $W \sim N(0,1)$, $V \sim \chi^2(r)$

$$T = \frac{W}{\sqrt{V/r}} \sim t(r)$$

따라서 $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = W$, $\frac{(n-1)S^2}{\sigma^2} = V$라 하면, r은 $n-1$이 되며

$$T = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}\cdot\frac{1}{n-1}}} = \frac{W}{\sqrt{V/r}} \sim t(r) \ (= t(n-1))$$

문제 3 t-test

t검정은 모집단이 정규분포를 따르지만 모표준편차를 모를 때, 모평균에 대한 가설검정 방법입니다. 대개 두 집단의 모평균이 서로 차이가 있는지 파악하고자 할 때 사용하며, 표본평균의 차이와 표준편차의 비율을 확인하여 통계적 결론을 도출합니다.

(3-1) 어떤 학우가 DSL 학회원(동문 포함)의 평균 키가 DSL 학회원이 아닌 사람의 평균 키보다 크다는 주장을 하여, 실제로 그러한지 통계적 검정을 수행하려고 합니다. 며칠간 표본을 수집한 결과 다음의 결과를 얻었다고 합시다.

표본 수 : 각 101명
측정에 응한 DSL 학회원들의 평균 키 : 178.5cm / 표준편차 : 7.05cm
측정에 응한, DSL 학회원이 아닌 사람들의 평균 키 : 179.9cm / 표준편차 : 7.05cm

(a) ✎ 귀무가설과 대립가설을 설정하시오. Let DSL 평균키 $=\mu_A$, not DSL 평균키 $:\mu_B$
DSL 표준편차: σ_A, not DSL 표준편차 : σ_B

$H_0: \mu_A - \mu_B = 0$. $H_1: \mu_A - \mu_B > 0$

(b) ✎ 유의수준 5%에서의 가설검정을 수행하고 결론을 도출하시오.

> 통계학입문(3판) 7장 참고
> 어떤 검정통계량이 어떤 분포를 따르는지, 언제 귀무가설을 기각하는지 정해야 합니다.

기호정리: $\bar{X}_A = 178.5$, $\bar{X}_B = 179.9$, $n_A = 101$, $n_B = 101$, $S_A = 7.05$, $S_B = 7.05$, $\alpha = 0.05$
by CLT, 모분산 unknown, 표본수 충분히 크므로 Z검정으로 진행. and $\hat{\sigma}_A^2 = S_A^2$, $\hat{\sigma}_B^2 = S_B^2$ 로 모분산 대체

검정통계량 $Z = \dfrac{(\bar{X}_A - \bar{X}_B) - (\mu_A - \mu_B)}{\sqrt{\hat{\sigma}_A^2/n_A + \hat{\sigma}_B^2/n_B}} = -1.411187$

만약 $Z \geq Z_{0.05}$. reject H_0 . $-1.411187 < 1.645$ 이므로 not reject H_a

⇒ 결론: DSL 학회원들의 평균키가 DSL이 아닌 사람들의 평균키보다 크다고 보기 어렵다.

(3-2) 신촌 연세로를 지나는 버스 노선의 이용객 수가 '차 없는 거리 해제(2022.10.09.)' 이후 유의미하게 증가했는지 파악하기 위해, 우선 2022년 9월의 평균 이용객 수와 2022년 11월의 평균 이용객 수가 유의미한 차이를 보이는지 통계적 검정을 수행하려고 합니다. 〈elementary1.ipynb〉

(a) ✎ 귀무가설과 대립가설을 설정하시오. Let 해제전 평균이용객수 μ_A, 해제후 평균이용객수 μ_B

$H_0: \mu_A - \mu_B = 0$, $H_1: \mu_A - \mu_B < 0$

(b) ⓒ 파이썬 scipy의 stats 패키지를 활용하여 유의수준 5%에서의 가설검정을 수행하고 결론을 도출하시오.