기수 : 10기 이름 : 정 성 오

23-2 DSL 정규세션

기초과제 1 통계적 사고

☑ 본 과제는 「통계학입문」, 「통계방법론」 및 「수리통계학(1), (2)」 일부에 상응하는 내용의 복습을 돕기 위해 기획되었습니다. 평가를 위한 것이 아니므로, 주어진 힌트(☝)를 적극 활용하시고 학회원 간 토론, Slack의 질의응답을 활용하시어 해결해주십시오. 단, 답안 표절은 금지합니다.

☑ 서술형 문제는 ✎로, 파이썬 코딩 문제는 ⓒ로 표기되어 있습니다. 각 문제에서 요구하는 방법에 맞게 해결해주십시오.

☑ 문제1,2,3-1은 따로 작성하시어 pdf로 제출해주시고, 문제3-2,4는 ipynb 파일에 답안을 작성하시어 제출해주십시오. 파일 이름 : [0712]_elementary1_영문이름.ipynb

☑ **7/12(수) 23시 59분까지** Github에 pdf 파일과 ipynb 파일을 모두 제출해주십시오.

☑ 참고 도서 :
통계학입문(3판, 강상욱 외), Introduction to Mathematical Statistics(8판, Hogg et.al.)

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

정의 : $X_1, X_2, \cdots, X_n$이 평균이 μ, 분산이 σ^2인 동일한 분포에서의 관찰값들일 때, $n \to \infty$면,
표본 평균인 확률변수 $\bar{X}$는 $N(\mu, \sigma^2/n)$을 따른다.

의미 : 표본이 정규분포가 아닌 다른 분포에서 추출됐을지라도 CLT에 의해 $\bar{X}$는 정규분포를 따르기 때문에 모집단의 특성을 파악하기 용이하다.

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

☝ Hogg(8판) 4장 2절 참고

확률 변수가 정규분포를 따르지 않는 경우에도, CLT를 활용하면 μ에 대한 근사적인 신뢰구간을 얻을 수 있다.

(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

CLT : $n\to\infty$일 때, $\bar{X} \sim N(\mu, \sigma^2/n)$ (i.e. $\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \xrightarrow{d} N(0,1^2)$)

$P\left(\left|\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}\right| \le z_{\frac{\alpha}{2}}\right) = 1-\alpha$

$= P(\underbrace{\bar{X} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}} \le \mu \le \bar{X} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}}_{\text{모평균 } \mu\text{에 대한 신뢰구간}})$

$\sqrt{Var(\bar{X})} = \sqrt{\frac{\sigma^2}{n}}$ 에서 모분산 σ^2을 표본분산으로 대체할 수 있다.

$s^2 \xrightarrow{P} \sigma^2$

Slutsky's theorem : $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{P} c$ 일 때

i) $X_n + Y_n \xrightarrow{d} X + c$

ii) $X_n Y_n \xrightarrow{d} Xc$

iii) $X_n / Y_n \xrightarrow{d} X/c$

$\frac{\bar{X}-\mu}{s/\sqrt{n}} = \underbrace{\left(\frac{\sigma}{s}\right)}_{①} \underbrace{\frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}}}_{②}$ 에서

① $\frac{\sigma}{s} \xrightarrow{P} \frac{\sigma}{\sigma} = 1$

② $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1^2)$

∴ Slutsky's theorem ii)에 따라 $\frac{\bar{X}-\mu}{s/\sqrt{n}} \xrightarrow{d} N(0,1^2)$

그러므로 $n\to\infty$면 σ^2대신 s^2을 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 $(n-1)S^2/\sigma^2$은 $\chi^2(n-1)$ 분포를 따른다.

〈증명〉

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

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

$$= \sum\frac{(X_i-\bar{X})^2}{\sigma^2} + 2\sum\frac{(X_i-\bar{X})(\bar{X}-\mu)}{\sigma^2} + \sum\frac{(\bar{X}-\mu)^2}{\sigma^2}$$

$$= \sum\frac{(X_i-\bar{X})^2}{\sigma^2} + 2(\bar{X}-\mu)\underbrace{\sum\frac{(X_i-\bar{X})}{\sigma^2}}_{=0} + \sum\frac{(\bar{X}-\mu)^2}{\sigma^2}$$

$$= \frac{(n-1)}{\sigma^2}\underbrace{\sum\frac{(X_i-\bar{X})^2}{n-1}}_{=S^2} + n\cdot\frac{(\bar{X}-\mu)^2}{\sigma^2}$$

$$= (n-1)S^2/\sigma^2 + \frac{(\bar{X}-\mu)^2}{\sigma^2/n} \sim \chi^2(n)$$

서로 독립인 정규분포를 따르는 변수의 제곱(Z^2)들의 합

$\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \xrightarrow{d} N(0,1)$ 이므로 $\frac{(\bar{X}-\mu)^2}{\sigma^2/n} \xrightarrow{d} \chi^2(1)$

②에 따르면 S^2과 $\bar{X}$는 독립이므로 $(n-1)S^2/\sigma^2$과 $\frac{(\bar{X}-\mu)^2}{\sigma^2/n}$도 독립

∴ $(n-1)S^2/\sigma^2$은 $\chi^2(n-1)$ 분포를 따른다고 할 수 O

(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 확률변수 $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ 은 자유도가 $n-1$ 인 Student t분포를 따른다.

〈증명〉

자유도가 ν인 Student t분포의 정의: $T=\frac{Z}{\sqrt{V/\nu}}$ (Z는 표준정규분포, V는 자유도가 ν인 카이제곱 분포)

(2-1)에 따라 $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$ 이므로

V대신 $(n-1)S^2/\sigma^2$을 대입, ν대신 $n-1$을 대입하면 T는 자유도가 $n-1$인 Student t분포가 됨.

$Z = \frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}$

$T = \left(\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}\right) \Big/ \sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)} \sim t(n-1)$

$= (\bar{X}-\mu)\sqrt{\frac{n}{\sigma^2}}\sqrt{\frac{\sigma^2}{S^2}}$

$= \frac{\bar{X}-\mu}{\sqrt{S^2/n}}$

$\therefore T = \frac{\bar{X}-\mu}{\sqrt{S^2/n}} \sim t(n-1)$

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.

귀무가설(H_0): DSL 학회원들의 평균 키가 DSL 학회원이 아닌 사람의 평균 키보다 작거나 같다.

대립가설(H_1): DSL 학회원들의 평균 키가 DSL 학회원이 아닌 사람의 평균 키보다 크다.

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

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

μ_1 : DSL 학회원들의 평균 키 (모평균) $\xrightarrow{\text{sampling}}$ $\overline{X_1} = 178.5$, $S_1 = 7.05$, $n_1 = 101$

μ_2 : DSL 학회원이 아닌 사람들의 평균 키 (모평균) $\longrightarrow$ $\overline{X_2} = 179.9$, $S_2 = 7.05$, $n_2 = 101$

(표본이 정규성, 등분산성, 독립성 만족한다고 가정)

$H_0 : \mu_1 \le \mu_2$

$H_1 : \mu_1 > \mu_2$

$\alpha = 0.05$

합동분산 $S^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)}$

$= \dfrac{100 \times 49.7025 + 100 \times 49.7025}{100 + 100}$

$= 49.7025$

표준오차 $se = \sqrt{S^2 \cdot \left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$

$= 7.05 \times \sqrt{\frac{1}{101}+\frac{1}{101}}$

$\fallingdotseq 0.992$

검정통계량 $t = \dfrac{\overline{X_1}-\overline{X_2}}{se}$

$= \dfrac{178.5-179.9}{0.992}$

$\fallingdotseq -1.411$ (t는 자유도가 200인 t분포 따름 $\because$ 자유도(df) $= (n_1-1)+(n_2-1) = 200$)

$t > t_{0.05,200}$ 이면 H_0 기각

$t = -1.411$

$t_{0.05,200} = 1.645$

$t < t_{0.05,200}$

$\therefore$ H_0을 기각하지 못한다. 즉, DSL 학회원의 평균 키가 DSL 학회원이 아닌 사람의 평균 키보다 크다고 할 수 없다.