기수 : 10 이름 : 임창재

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

(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($\sqrt{Var(\overline{X})}$) 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

1-1. Central Limit Theorem

Let $X_1 \cdots X_n$ denote the observations of a random sample from a distribution that has mean μ and positive variance σ^2

Then the random variable

$$Y_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\,\sigma} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$$

converges in distribution to a random variable that has a normal distribution with mean zero and variance 1.

In other words, $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} N(0, \sigma^2)$

Central Limit Theorem is important in Statistics because it makes us possible to know the approximate distribution of sample mean no matter what the distribution of the population is. (for large n)

1-2. CLT in Interval Estimation.

구간 추정 (Interval Estimate) 이란, 어떤 모집단에 대한 관심모수의 점추정치와 그 추정의 정확도를 모두 고려한 추정 방법이다.

예를 들어, 평균이 μ 이고 분산이 σ^2 인 임의의 분포로부터 n 개의 표본을 추출한다고 하자. $X_1 \cdots X_n$

관심모수 μ 의 95% 신뢰구간이 $(\bar{x} - L, \bar{x} + U)$ 라 하자. 여기서 $\hat{x}, L, U$ 는 모집단에서 추출된 표본으로부터 계산된다. (Slutsky's theorem)

이는 모수 μ 가 신뢰구간 $(\bar{x} - L, \bar{x} + U)$ 에 포함될 확률이 95% 라는 것을 의미한다.

이때 $Pr(\bar{x} - L < \mu < \bar{x} + U) = 0.95$ 를 만족하는 L 과 U 를 계산하기 위해서는 $\bar{x}$ 의 분포를 알아야 한다.

여기서 CLT에 의해 $\bar{X}$ 가 근사적으로 정규분포를 따르기 때문에 구간추정이 가능해진다.

1-3. Slutsky's Theorem in Interval Estimation

Let $X_1 \cdots X_n \overset{iid}{\sim}$ a distribution with mean μ, variance $\sigma^2 < \infty$.

By CLT, $\dfrac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$ approximately

$$Pr\left(-1.96 < \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} < 1.96\right) = Pr\left(\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right) \cong 0.95$$

This cannot be a 95% confidence interval for μ, because σ is unknown

Note, $S_n/\sigma \xrightarrow{P} 1$

Recall, by CLT, $\dfrac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$ approximately

By Slutsky's theorem,

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \cdot \frac{\sigma}{S_n} \xrightarrow{D} N(0,1)$$

$$Pr\left(\bar{X}_n - 1.96\frac{S_n}{\sqrt{n}} < \mu < \bar{X}_n + 1.96\frac{S_n}{\sqrt{n}}\right) \cong 0.95$$

$$\therefore \left(\bar{X}_n - 1.96\frac{S_n}{\sqrt{n}},\ \bar{X}_n + 1.96\frac{S_n}{\sqrt{n}}\right) \text{ is 95\% approximate confidence interval for } \mu.$$

기수 : 10 이름 : 임창재

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

② 표본평균 $\bar{X}$ 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.

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

(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.

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

proof) Let $X_1 \dots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$$\frac{X_i-\mu}{\sigma} \sim N(0,1) \Rightarrow \left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2_{(1)} \Rightarrow V := \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2_{(n)}$$

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

Note. $\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2 = \frac{n-1}{\sigma^2}\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{n-1} = (n-1)\frac{S^2}{\sigma^2} \cdots (1)$

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

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

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

2-2. ④ $T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$

Note. By CLT, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \cdots (1)$

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

Def. Student's t-distribution $\cdots (3)$

Random variable $Z/\sqrt{V/\nu} \sim t(\nu)$ if $Z \sim N(0,1)$, $V \sim \chi^2(\nu)$

$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{(\bar{X}-\mu)/(\sigma/\sqrt{n})}{\sqrt{(n-1)S^2/(\sigma^2(n-1))}} \sim t(n-1) \text{ by (1), (2), (3).}$$ ■

기수 : 10 이름 : 임창재

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.

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

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

(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.

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

3-1. (a) μ_1 := DSL 학회원의 평균 신장, μ_2 := DSL 학회원이 아닌 사람의 평균 신장.

귀무가설 $H_0 : \mu_1 = \mu_2$

대립가설 $H_1 : \mu_1 > \mu_2$

(b) $\bar{x}_1 = 178.5cm$, $S_1 = 7.05cm$, $m_1 = 101$명

$\bar{x}_2 = 179.9cm$, $S_2 = 7.05cm$, $m_2 = 101$명.

대표본이고 모분산을 알지 못한다.

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S_1^2/m_1 + S_2^2/m_2}} = \frac{178.5 - 179.9}{\sqrt{2 \times 7.05/101}} = -3.747 \sim N(0,1) \quad \text{under } H_0$$

Reject H_0 if $Z > 1.64$

$Z = -3.747 < 1.64$ 이므로 not reject H_0.

따라서 유의수준 0.05에서 DSL 학회원의 평균신장이 DSL 학회원이 아닌 사람의 평균 신장보다 크다고 할 수 없다.

문제 4 Linear Regression

회귀분석은 통계분석 방법 중 가장 많이 사용되는 방법으로, 독립변수로 종속변수를 예측할 수 있는 좋은 추세선을 유도하는 것이 핵심입니다. 추세선의 표준오차가 작을수록 유의미한 추세선이 됩니다. 더 자세한 내용은 7/18(화) 세션에서 다뤄질 예정입니다.

(4-1) (통계학입문(3판) 541쪽 15번 변형) 토플(TOEFL) 점수로 토익(TOEIC) 점수를 예측할 수 있는지 검증해보려고 합니다. 토플과 토익은 미국 교육기업 ETS가 서로 다른 목적으로 개발한 시험이기에 문제 내용과 구성이 다르지만, 두 시험 점수 간의 관련성이 있는 것으로 알려져 있습니다.
〈elementary1.ipynb〉

(a) ⓒ 임의로 선정된 20명의 학생에게 토플과 토익을 모두 치르게 하여 얻은 자료가 있습니다. 파이썬 sklearn.linear_model의 LinearRegression 패키지를 활용하여, 토플 점수로 토익 점수를 예측하는 회귀식을 최소제곱법으로 추정하시오.

(b) ⓒ 유의수준 5%에서 각 회귀식의 선형성이 있는지 가설검정을 수행하시오.

(4-2) (통계학입문(3판) 582쪽 19번 변형) 미국 어느 지역의 자동차 판매대수와 광고비, 자동차 전문세일즈맨수, 판매 대리점의 위치 등에 대하여 조사한 자료를 분석하고자 합니다. 〈elementary1.ipynb〉

ⓒ 파이썬 sklearn.linear_model의 LinearRegression 패키지를 활용하여, Akaike information criterion을 기준으로 아래 모델 중 가장 우수한 모델을 선택하시오.

model	종속변수	독립변수
a	자동차 판매대수	광고시간(분), 세일즈맨(명), 도시지역
b		광고시간(분), 세일즈맨(명)
c		광고시간(분), 도시지역
d		세일즈맨(명), 도시지역
e		광고시간(분)
f		세일즈맨(명)

Reference
- 통계학입문(3판, 강상욱 외)
- Introduction to Mathematical Statistics(8판, Hogg et.al)
- 23-1 정규세션 〈통계적 사고〉 (8기 정건우)

DATA
SCIENCE LAB
담당자 : 학술부(이성균)
leesg0104@yonsei.ac.kr