

25-1 DSL 정규 세션

기초과제



- ☑ 본 과제는 「통계학입문」, 「통계방법론」 및 「수리통계학(1)」 일부 내용을 다루며, NumPy와 Pandas의 활용 연습을 돕기 위해 기획되었습니다. 평가를 위한 것이 아니므로, 주어진 힌트(?)를 적극 활용하시고 학회원 간 토론, Slack의 질의응답을 활용하시어 해결해주시고요. 단, 답안 표절은 금지합니다.
- ☑ 서술형 문제는 ✍, 코딩 문제는 © 으로 표기가 되어 있습니다. 각 문제에서 요구하는 방법에 맞게 해결하며, 서술형 문제들은 따로 작성하시어 pdf로 제출해주시고 코드 문제들은 ipynb 파일에 답안을 작성하시어 제출해주시고요.
- ☑ 1/13 (월) 23시 59분까지 Github에 PDF 파일과 ipynb 파일을 모두 제출해주시고요. Github에 제출하는 방법을 모른다면 학술부장 혹은 과제 질의응답을 위한 오픈채팅방을 활용해주시고요.

문제 1 (Weak) Law of Large Numbers

큰 수의 법칙은 표본 평균의 수렴성을 보장하는 법칙으로, 중심극한정리(CLT)와 더불어 통계학에서 중요한 법칙입니다. 예를 들어, 큰 수의 법칙은 몬테카를로(Monte Carlo) 방법론의 이론적 기반을 제공합니다. 이 문제에서는 큰 수의 약한 법칙의 정의를 확인하고, 이를 코드를 통해 확인해 보겠습니다.

1-1 ✍ : 큰 수의 약한 법칙(Weak Law of Large Numbers)의 정의를 서술하시고 증명하시오.

👉 Hogg(8판) 5장 1절

정의) $\{X_n\}$ 이 공통평균 μ 와 공분산 σ^2 ($\sigma^2 < \infty$)을 갖는 iid 확률 변수 일때

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

$$\bar{X}_n \xrightarrow{P} \mu$$
 이다.

증명) $E(\bar{X}_n) = \mu$ $Var(\bar{X}_n) = \frac{\sigma^2}{n}$
 체비셰프 부등식에 의해 임의의 $\varepsilon > 0$ 에 대해

$$P[|\bar{X}_n - \mu| \geq \varepsilon] = P[|\bar{X}_n - \mu| \geq (\varepsilon \sqrt{n}) / (\sigma / \sqrt{n})] \leq \frac{\sigma^2}{n \varepsilon^2} \rightarrow 0$$
 $\Rightarrow \bar{X}_n$ 이 확률적으로 μ 에 수렴

1-2 © : NumPy를 이용하여 큰 수의 약한 법칙(WLLN)을 시뮬레이션으로 확인하시오.

문제 2 Central Limit Theorem

중심극한정리는 확률변수의 합 형태 (Sum of Random Variables)의 극한분포를 손쉽게 구할 수 있도록 해 주기에 통계학에서 가장 자주 사용하는 정리입니다. 이 문제에서는 중심극한정리의 정의와 그 활용에 대해 짚어보겠습니다.

2-1 ✎ : 중심극한정리(Central Limit Theorem)의 정의를 서술하시오.

👉 통계학입문 (3 판) 7 장 참고

👉 Hogg(8 판) 4 장 2 절, 5 장 3 절 참고

모집단이 정규분포가 아니더라도 평균 μ , 분산 σ^2 일때 random한 표본의 크기 n 이 충분히 크면 ($n \geq 30$) \bar{X} 의 분포는 근사적으로 정규분포 $N(\mu, \frac{\sigma^2}{n})$ 을 따른다.
 \Rightarrow 모평균 μ , 모분산 σ^2 , $n \geq 30$ 이면 $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

2-2 ✎ : 중심극한정리가 통계적 추론 중 "구간추정"에서 어떻게 활용되는지 서술하시오.

👉 Hogg(8 판) 4 장 2 절

CLT에 의해 $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$

$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$ 이고, 신뢰수준 $1-\alpha$ 에서 $P(-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}) = 1-\alpha \rightarrow P(\bar{X}_n - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}) = 1-\alpha$

\therefore 모집단 평균 μ 의 신뢰구간은 $(\bar{X}_n - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X}_n + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}})$

모분산이 알려지지 않은 경우, 표본 분산 S^2 을 사용하고 t분포를 적용, $t = \frac{\bar{X}_n - \mu}{S/\sqrt{n}} \sim t_{n-1}$

신뢰수준 $1-\alpha$ 에서 신뢰구간 : $(\bar{X}_n - t_{\alpha/2, n-1} \frac{S}{\sqrt{n}}, \bar{X}_n + t_{\alpha/2, n-1} \frac{S}{\sqrt{n}})$

2-3 © : NumPy 를 이용하여 중심극한정리(CLT)가 적용되는 과정을 확인하시오.

문제 3 모분산에 관한 추론

카이제곱 분포는 모집단의 모분산 추정에 유용하게 쓰이며, 정규분포에서의 랜덤표본에서 표본분산과 관계되는 분포입니다. 표준정규분포를 따르는 서로 독립인 확률변수 $Z_1, Z_2, Z_3, \dots, Z_k$ 가 있을 때, $V = Z_1^2 + Z_2^2 + Z_3^2 + \dots + Z_k^2 \Rightarrow V \sim$ 자유도가 k 인 χ^2 분포를 따른다고 할 수 있습니다. 대개 모분산에 관한 추론에 사용되며, 검정통계량으로 $\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ 가 쓰입니다.

3-1 : 플라스틱 판을 제조하는 공장이 있다. 판 두께의 표준편차가 1.5mm를 넘으면 공정 상에 이상이 있는 것으로 간주합니다. 오늘 아침 10 개의 판을 무작위 추출하여 두께를 측정한 결과가 다음과 같았습니다.

{226, 228, 226, 225, 232, 228, 227, 229, 225, 230}

해당 판 두께의 분포가 정규분포를 따른다고 할 때, 공정에 이상이 있는지를 검정하세요.

a) 귀무가설과 대립가설을 설정하시오.

H_0 : 판 두께의 표준편차는 $\sigma \leq 1.5$

H_1 : 판 두께의 표준편차는 $\sigma > 1.5$

b) 유의수준 5%에서의 가설검정을 수행하고 판 두께의 분산에 대한 90% 신뢰구간을 구하시오.

어떤 검정통계량이 어떤 분포를 따르는지, 언제 귀무가설을 기각하는지 정해야 합니다.

기초통계량: $n=10$ $\bar{X} = 227.6$ $S^2 = 6.04$

Under H_0 , 검정통계량 $\chi^2 = \frac{(n-1)S^2}{\sigma^2} = 24.16$

기각역 $\chi^2_{0.05,9} = 16.92$ 이고 $\chi^2 = 24.16 > \chi^2_{0.05,9} = 16.92$ 이므로 귀무가설을 기각

결론: 유의수준 5%에서 판 두께의 표준편차가 1.5mm를 초과한단 판단할 수 있다.

(양쪽) 90% 신뢰구간: $\left(\frac{(n-1)S^2}{\chi^2_{0.95,9}}, \frac{(n-1)S^2}{\chi^2_{0.05,9}} \right)$. $\chi^2_{0.95,9} = 3.33$ $\chi^2_{0.05,9} = 16.92$ 이므로 대입하면

90% 신뢰구간은 (3.22, 16.34) 이다.

문제 4 통계적 방법론

t 검정은 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. ANOVA Test의 경우 집단이 2 개보다 많은 경우 모평균에 차이가 있는지 파악하고자 할 때 사용되며, 이것은 코드로만 살펴보겠습니다.

4-1 ✎ : 어떤 학우가 DSL 학회원(동문 포함)의 평균 키가 DSL 학회원이 아닌 사람의 평균 키보다 크다고 주장하여, 실제로 그러한지 통계적 검정을 수행하려고 합니다. 며칠간 표본을 수집한 결과 다음과 같은 값을 얻었습니다.

표본 수: 총 250 명, 각 125 명
측정에 응한 DSL 학회원들의 평균 키 : 173.5cm / 표준편차 : 7.05cm
측정에 응한, DSL 학회원이 아닌 사람들의 평균 키 : 171.4cm / 표준편차 : 7.05cm

a) ✎ 귀무가설과 대립가설을 설정하시오.

DSL 학회원의 평균 키: μ_1 , DSL 학회원이 아닌 사람들의 평균 키: μ_2 라 할 때,

$H_0: \mu_1 - \mu_2 \leq 0$, $H_1: \mu_1 - \mu_2 > 0$

b) ✎ 유의수준 5%에서의 가설검정을 수행하고 결론을 도출하시오. (단, 키는 정규분포를 따르며 각 집단의 분산은 같다고 가정한다.)

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

기초통계량: $n_1 = n_2 = 125$, $\bar{X}_1 = 173.5$, $S_1 = 7.05$ / $\bar{X}_2 = 171.4$, $S_2 = 7.05$

Under H₀, 검정통계량: $S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2} = 49.7025 \rightarrow S_p = 7.05$, $t = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = 2.37$

유의수준이 5%이므로 $\alpha = 0.05 \rightarrow t_{0.05, 248} = 1.645$ 이고 $t = 2.37 > t_{0.05, 248} = 1.645$ 이므로 귀무가설을 기각한다.

결론: 유의수준 5%에서 DSL 학회원의 평균 키가 학회원이 아닌 사람들의 평균 키보다 크다고 할 수 있다.

4-2 © : 한 학우가 이번에는 각 학회의 평균 키가 똑같다는 주장을 하였습니다. 해당 학우가 제공한 ESC 학회의 학회원별 키 데이터를 활용해 가설검정을 진행하고자 합니다. 데이터는 heights.csv 파일에 저장되어 있습니다.

a) ✎ 귀무가설과 대립가설을 설정하시오.

ESC 학회원의 평균 키를 μ_3 이라 할 때

$H_0: \mu_1 = \mu_2 = \mu_3$ H_1 : 적어도 하나의 μ_i 는 다른 것과 다르다.

b) © 파이썬의 scipy.stats 을 활용해서 유의수준 5%에서의 가설검정을 수행하고 결론을 도출하시오. 결론은 .ipynb 파일에 쓰셔도 괜찮습니다.

✎ One-way Anova Test 를 활용해서 사용하는 문제입니다.
✎ 활용해야 될 함수는 scipy.stats.f_oneway 입니다.

문제 5 NumPy + Pandas 활용

기초과제.ipynb 파일에 제공된 문제들을 참고하여 수행하시기 바랍니다.

Reference

- 통계학입문(3 판, 강상욱 외)
- Introduction to Mathematical Statistics(8 판, Hogg et.al)
- 23-2 기초과제 1 (9 기 이성균)
- 24-1 기초과제 1 (10 기 신재우)
- 24-2 기초과제 1 (11 기 김현진, 김정우)

Data Science Lab

담당자: 12 기 이정우

leejeongwoo9941@yonsei.ac.kr