기수 : 15기 이름 : 배건우

26-1 DSL 정규 세션

기초과제

- ☑ 본 과제는「통계학입문」,「통계방법론」및「수리통계학(1)」일부 내용을 다루며, NumPy 와 Pandas 의 활용 연습을 돕기 위해 기획되었습니다. 평가를 위한 것이 아니므로, 주어진 힌트(👋)를 적극 활용하시고 학회원 간 토론, Slack 의 질의응답을 활용하시어 해결해주십시오. 단, 답안 표절은 금지합니다.
- ☑ 서술형 문제는 ✎ , 코딩 문제는 © 으로 표기가 되어 있습니다. 각 문제에서 요구하는 방법에 맞게 해결하며, 서술형 문제들은 따로 작성하시어 pdf 로 제출해주시고 코드 문제들은 ipynb 파일에 답안을 작성하시어 제출해주십시오.
- ☑ **1/16 (금) 23 시 59 분까지** Github 에 PDF 파일과 ipynb 파일을 zip 파일로 압축하여 제출해주십시오. Github 에 제출하는 방법을 모른다면 학술부장에게 연락 부탁드립니다.

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(8 판) 5 장 1 절

theorem: let $X_1, \cdots, X_n$ be i.i.d with $E(X_i)=\mu$, $Var(X_i) < \infty$ for $i = 1, \cdots, n$
the sample mean $\bar{X} = \frac{1}{n}\sum X_i$
then, $\lim_{n\to\infty} Pr(|\bar{X}-\mu| < \varepsilon) = 1$

proof: since X_i s are i.i.d $\Rightarrow E(\bar{X}) = E(\frac{1}{n}\sum X_i) = \frac{1}{n}E(\sum X_i) = \frac{1}{n} \cdot n \cdot \mu = \mu$.
$Var(\bar{X}) = Var(\frac{1}{n}\sum X_i) = \frac{1}{n^2}Var(\sum X_i) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{1}{n}\sigma^2$
Putting this into Chebyshev's inequality $Pr(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$,
$Pr(|\bar{X} - E(\bar{X})| \geq \varepsilon) \leq Var(\bar{X})/\varepsilon^2$, $Pr(|\bar{X}-\mu| \geq \varepsilon) \leq \sigma^2/n\varepsilon^2$
$Pr(|\bar{X}-\mu| < \varepsilon) \geq 1 - \sigma^2/n\varepsilon^2$, $\lim_{n\to\infty} Pr(|\bar{X}-\mu| < \varepsilon) \geq \lim_{n\to\infty}(1 - \sigma^2/n\varepsilon^2) \geq 1 - \lim_{n\to\infty} \frac{\sigma^2/\varepsilon^2}{n} \geq 1$.

1-2 © : NumPy 를 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

If you have a large number of i.i.d random variables
from any dist with mean μ and variance σ^2,
the sample mean will approximately normally distributed as $n \to \infty$

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

- Hogg(8 판) 4 장 2 절

모분산이 알려진 정규분포는 정규분포를 이용한 신뢰구간을,
모분산이 알려지지 않은 정규분포로 확률변수를 t-분포를 사용하여
구간에 대한 구간 추정을 수행할 수 있다.

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,…, Z_k 가 있을 때, $V = Z_1, + Z_2 + Z_3 + \cdots + Z_k \Rightarrow V \sim$ 자유도가 k 인 X^2 분포를 따른다고 할 수 있습니다. 대개 모분산에 관한 추론에 사용되며, 검정통계량으로 $X^2 = \frac{(n-1)S^2}{\sigma^2} \sim X^2(n-1)$가 쓰입니다.

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

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

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

a) ✎ 귀무가설과 대립가설을 설정하시오.

H_0 : 판 두께의 표준편차가 1.5mm 는 넘지 않는다 ($\sigma \leq 1.5$)

H_1 : 판 두께의 표준편차가 1.5mm 를 넘는다. ($\sigma > 1.5$)

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

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

$$X^{*2} = \frac{(n-1)s^2}{\sigma^2} = \frac{9 \times 5.15}{2.25} = 20.6 \qquad s^2 = \frac{1}{10-1}\Sigma(x_i - \bar{x})^2 = 5.15$$

$$\chi^2_{0.95,9} \approx 16.92$$

$X^{*2} > \chi^2_{0.95,9}$ 이므로 H_0 기각.

$$CI: \left(\frac{(n-1)s^2}{\chi^2_{0.95,9}}, \frac{(n-1)s^2}{\chi^2_{0.05,9}}\right)$$

$$\rightarrow \left(\frac{9 \cdot 5.15}{16.92}, \frac{9 \cdot 5.15}{3.33}\right)$$

$$\rightarrow (2.73, 17)$$

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 Test 의 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.

H_0 : DSL 학회원의 평균 키가 DSL 학회원이 아닌 사람의 평균 키보다 같거나 작다.

H_1 : DSL 학회원의 평균 키가 DSL 학회원이 아닌 사람의 평균 키보다 크다

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.

$\bar{x}_1 = 173.5$ $\bar{x}_2 = 171.4$ $\sigma_1 = \sigma_2 = 7.05$

$SE = s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = 7.05\sqrt{\frac{2}{125}} \approx 0.892$

$t = \frac{\bar{x}_1 - \bar{x}_2}{SE} = \frac{173.5-171.4}{0.892} \approx 2.35 > t_{0.95,\,248} \approx 1.65$

∴ reject H_0

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.

H_0 : 각 학회의 평균 키가 똑같다.

H_1 : 각 학회의 평균 키가 하나라도 같지 않다.

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)
- 25-1 기초과제 1 (12 기 이정우)
- 25-2 기초과제 1 (13 기 강승우)

Data Science Lab

담당자: 14 기 박창용

qkrckddyd0@yonsei.ac.kr