**기수 :** 13 **이름 :** 서승범

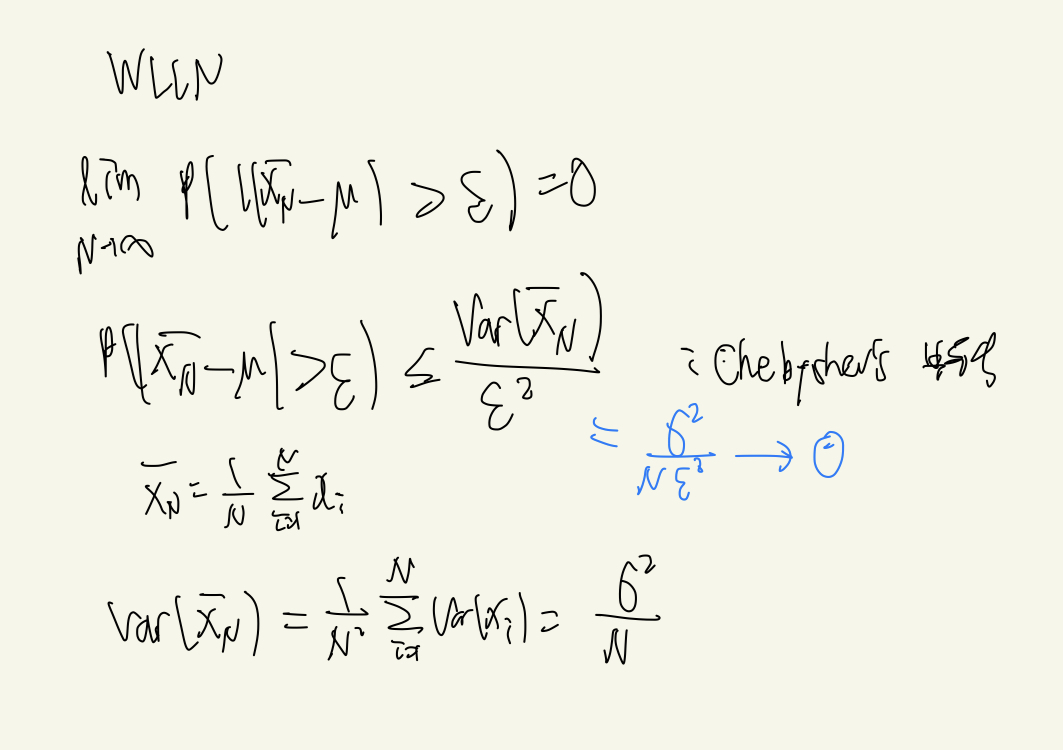
|  |
| --- |
| A blue and white logo  Description automatically generated**25-1 DSL 정규 세션**  **기초과제** |
|  |
| * 본 과제는 「통계학입문」, 「통계방법론」 및 「수리통계학(1)」 일부 내용을 다루며, NumPy와 Pandas의 활용 연습을 돕기 위해 기획되었습니다. 평가를 위한 것이 아니므로, 주어진 힌트(N)를 적극 활용하시고 학회원 간 토론, Slack의 질의응답을 활용하시어 해결해주십시오. 단, 답안 표절은 금지합니다. * 서술형 문제는 ! , 코딩 문제는 © 으로 표기가 되어 있습니다. 각 문제에서 요구하는 방법에 맞게 해결하며, 서술형 문제들은 따로 작성하시어 pdf로 제출해주시고 코드 문제들은 ipynb 파일에 답안을 작성하시어 제출해주십시오. * **1/13 (월) 23시 59분까지** Github에 PDF 파일과 ipynb 파일을 모두 제출해주십시오. Github 에 제출하는 방법을 모른다면 학술부장 혹은 과제 질의응답을 위한 오픈채팅방을 활용해주십시오. |

|  |  |
| --- | --- |
| **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: 확률변수 수열 중에서 n번째까지의 확률 변수의 평균과 모평균 차이의 절대값이 임의의 양수 ε보다 클 확률이 0에 수렴한다.



**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 (3판) 7장 참고 * Hogg(8판) 4장 2절, 5장 3절 참고 |

모집단에서 표본이 충분히 크다면, 이 표본평균의 분포는 모집단의 분포 모양과는 관계없이 정규분포에 근사한다. 즉, n이 클 때 표본평균은 을 따르게 된다.

**2-2** ! **:** 중심극한정리가 통계적 추론 중 “구간추정”에서 어떻게 활용되는지 서술하시오.

|  |
| --- |
| * Hogg(8판) 4장 2절 |

모집단 분포가 정규분포가 아니더라도 표본평균의 분포가 근사적으로 정규분포를 따른다는 것을 이용해 표본평균의 분포를 정규분포로 간주하고 구간추정을 할 수 있다. 또한 모집단의 표준편차를 모르더라도 표본의 표준편차를 사용하여 표준편차의 신뢰구간을 계산할 때 정규분포를 사용하는 근거가 된다. 예를 들어, 신뢰구간 계산 시, 표본평균의 표준오차를 사용해 신뢰구간을 구하게 되는데, 95% 신뢰구간을 계산할 때 표본평균 ± (1.96 × 표본평균의 표준오차)를 이용하는 것이 그 예시이다.

**2-3** © **:** NumPy를 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 가 있을 때, ~ 자유도가 k인 분포를 따른다고 할 수 있습니다. 대개 모분산에 관한 추론에 사용되며, 검정통계량으로 가 쓰입니다. | |

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

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

**a)** ! 귀무가설과 대립가설을 설정하시오.

귀무가설: 판 두께의 두께의 표준편차는 1.5mm이다.

대립가설: 판 두께의 표준편차는 1.5mm보다 크다.

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

|  |
| --- |
| * 어떤 검정통계량이 어떤 분포를 따르는지, 언제 귀무가설을 기각하는지 정해야 합니다. |

표본 표준편차는 S=2.271

이며 이 = 20.63이다. 자유도가 9일 때 카이제곱의 알파 0.05의 단측 검정에서 귀무가설의 신뢰구간은 <16.919이다. 검정통계량이 16.919보다 크므로 귀무가설을 기각한다.

90%의 양측 신뢰구간의 경우 이 3.325와 16.919 사이에 있어야 하므로 판 두께의 분산은 2.74<variance<13.95이다.

|  |  |
| --- | --- |
| **문제 4** | 통계적 방법론 |
| 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. ANOVA Test의 경우 집단이 2개보다 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 250명, 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.

귀무가설: DSL 학회원의 평균 키와 아닌 사람의 평균 키가 동일

대립 가설: DSL 학회원의 평균 키와 아닌 사람의 평균 키에 유의미한 차이가 있음

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

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

Unpaired t-test, equal variance를 수행하면 차이는 2.1이라는 것을 이용해 t-test의 검정통계량을 구하면T=-2.355이다. 따라서 대략 p=0.02 수준으로 p<0.05 수준에서 두 집단의 평균이 다르다고 볼 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.

1. ! 귀무가설과 대립가설을 설정하시오.

귀무가설: 세 학회의 평균 키가 같다.

대립가설: 세 학회의 평균 키에 유의미한 차이가 있다.

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

|  |
| --- |
| * One-way Anova Test를 활용해서 사용하는 문제입니다. * 활용해야 될 함수는scipy.stats.f\_oneway입니다. |
| * 결론: 유의미한 차이가 있다. |

|  |  |
| --- | --- |
| **문제 5** | NumPy + Pandas 활용 |
| 기초과제.ipynb 파일에 제공된 문제들을 참고하여 수행하시기 바랍니다. | |

|  |  |
| --- | --- |
| Reference | **Data Science Lab** |
| - 통계학입문(3판, 강상욱 외)  - Introduction to Mathematical Statistics(8판, Hogg et.al)  - 23-2 기초과제 1 ( 9기 이성균 )  - 24-1 기초과제 1 ( 10기 신재우 )  - 24-2 기초과제 1 ( 11기 김현진, 김정우 ) | 담당자: 12기 이정우  [leejeongwoo9941@yonsei.ac.kr](mailto:leejeongwoo9941@yonsei.ac.kr) |