**기수 :** 13기 **이름 :** 박수빈

|  |
| --- |
| 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 |

큰 수의 약한 법칙은 다음과 같다.

을 공통 평균 , 공통 분산 을 가진 i.i.d. 확률변수의 열이라고 하자. 이면, 모든 에 대해,

Pf)

표본평균의 평균과 분산은 다음과 같다.

체비셰프 부등식에 따라 다음과 같이 쓸 수 있다.

따라서 확률의 공리, 함수 극한의 대소관계에 의해, 정의를 만족한다

**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 |

중심극한정리는 다음과 같다.

평균 , 분산 인 분포에서 크기가 인 표본들의 표본평균 는 이 충분히 클 때 다음의 정규분포를 근사적으로 따른다.

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

|  |
| --- |
| * Hogg(8판) 4장 2절 |

구간추정은 추출한 표본의 통계량을 바탕으로 신뢰수준에 따라 모수가 위치할 수 있는 구간인 신뢰구간을 추정하는 방법이다. 중심극한정리에 의해, 표본의 분포가 어떤 형태이든 표본 크기가 충분히 크다면 표본평균의 분포는 정규분포로 근사할 수 있다. 이를 바탕으로 신뢰수준(일반적으로 95%와 99%를 사용)에 따라 신뢰구간을 추정할 수 있다.

**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 표준편차가 1.5 mm 이하이다.

대립가설(): 판 두께의 표준편차가 1.5 mm를 초과한다.

이때 는 판 두께 분포의 표준편차이다.

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

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

주어진 표본의 평균은 227.6, 표본분산은 가 된다. 샘플의 개수가 10이므로, 자유도는 9가 된다. 검정통계량은 가 된다. 한편, 유의수준 5%에 대한 임계값은 약 16.92가 된다. 검정통계량이 임계값보다 크므로 귀무가설을 기각한다. 따라서 공정에 이상이 있다고 추론할 수 있다.

10%와 90%에서의 임계값은 각각 약 4.17, 14.68이 된다. 이를 식에 적용하면 신뢰구간은 (3.16, 11.14)가 된다.

|  |  |
| --- | --- |
| **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 학회원들의 평균 키()는 DSL학회원이 아닌 사람의 평균 키()보다 작거나 같다.

대립가설(): DSL 학회원들의 평균 키()는 DSL학회원이 아닌 사람의 평균 키()보다 크다.

**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. |

키가 정규분포를 따르므로, 검정통계량은 t-분포를 따른다고 가정할 수 있다. 샘플 개수는 각 125개, 측정된 DSL 학회원들의 평균 키는 173.5 cm, 측정된 DSL 학회원들이 아닌 인원들의 평균 키는 171.4 cm이다. 따라서 합동분산은 49.7025, 합동표준편차는 7.05가 된다. t-분포 검정통계량 식 을 적용하면, 검정통계량은 2.35가 된다. 유의수준은 5%이고, 자유도는 248이므로, 임계값은 약 1.65가 된다. 검정통계량이 임계값보다 크므로, 귀무가설을 기각한다. 따라서 DSL 학회원들의 평균 키는 DSL 학회원이 아닌 인원들보다 더 크다고 추론할 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.

귀무가설(): 세 학회의 평균 키는 동일하다.

대립가설(): 적어도 하나의 학회의 평균 키는 나머지 학회의 평균 키와 다르다.

**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. |

|  |  |
| --- | --- |
| **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) |