**기수 :** 12 **이름 : 서지니**

|  |
| --- |
| A blue and white logo  Description automatically generated**24-1 DSL 정규 세션**  **기초과제 1** **통계적 사고** |
|  |
| * 본 과제는 「통계학입문」, 「통계방법론」, 「선형대수」 및 「수리통계학(1)」 일부에 상응하는 내용의 복습을 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. * **7/25 (목) 23시 59분까지** Github에 PDF 파일과 ipynb 파일을 모두 제출해주십시오. Github 에 제출하는 방법을 모른다면 학술부장 혹은 과제 질의응답을 위한 오픈채팅방을 활용해주시오. * 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 해 주기에 통계학에서 가장 자주 사용하는 정리입니다. 이 문제에서는 중심극한정리의 정의와 그 활용에 대해 짚어보겠습니다. | |

**1-1** ! **:** 중심극한정리(Central Limit Theorem)의 정의를 서술하시오.

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

동일한 확률분포를 가진 독립 확률 변수 n개의 평균의 분포는 n이 적당히 크다면 정규분포에 가까워진다는 정리

**1-2** ! **:** 중심극한정리가 통계적 추론 중 “구간추정”에서 어떻게 활용되는지 서술하시오.

|  |
| --- |
| * Hogg(8판) 4장 2절 |

모집단 분포가 정규분포라는 가정이 없을 때 CLT를 활용하면, 모집단이 정규분포 안 따라도 표본평균 분포가 근사적으로 정규분포를 따른다는 사실을 이용해 신뢰구간을 구할 수 있다.

텍스트, 친필, 폰트, 화이트이(가) 표시된 사진

자동 생성된 설명

|  |  |
| --- | --- |
| **문제 2** | Linear Algebra |
| 선형대수학은 머신러닝을 위한 수학 중에서 가장 중요한 요소 중 하나이며, 이 중에서 가장 중요한 것 중에서 하나는 바로 SVD (Singular Value Decomposition, 특이값 분해) 입니다. 이것을 알기 위해서 고유값과 고유벡터를 활용한 Diagonalization 에 대해서 먼저 알아본 후, SVD 를 사용하여 실제로 이미지 압축을 적용해보겠습니다. | |

**2-1** ! **:** [Diagonalization](https://deep-learning-study.tistory.com/341) 의 정의가 다음과 같이 주어졌습니다.

|  |
| --- |
| Diagonalization 이란 정방행렬(A) 를 Eigenvalue, Eigenvector 를 통해서 대각행렬 (D) 를 만드는 것이며, 즉 를 통해서 대각행렬 (D) 를 찾는 것입니다.  조건들 : 1.) A 는 정방행렬 (Square Matrix) 이다. 2.) 는 개의 독립인 고유벡터를 가지고 있다.   1. A 에 대한 고유벡터들을 찾으며, 이것을 각각 으로 놓는다 2. 메트릭스를 만든다 3. 를 구하면 다음과 같은 형태가 나오게 된다   만약에 가 대칭 (Symmetric) 행렬이면 다음과 같은 꼴이 나오게 됩니다. |

다음과 같은 정방행렬에Diagonalization 을 적용시켜서 나오게 되는 대각행렬을 쓰시오.

5 0

0 4

2 0 0

0 2 0

0 0 1

3 0 0

0 1 0

0 0 2

**2-2** ! **:** SVD 의 정의가 다음과 같이 주어졌습니다.

|  |
| --- |
| SVD 란 Diagonalization 과는 달리 모든 행렬 (A) 에 대해서 사용이 가능합니다.   1. 행렬들을 만듭니다. 이것은 항상 대칭 (Symmetric) 행렬이 됩니다. 2. , 으로 대각화를 진행을 하고 나서 정규직교화까지 하게 된다면 와 를 얻게 됩니다. 3. 여기에서 0이 아닌 고유값들이 내림차순으로 나열된 것이 바로 D 가 되며, 이것은 바로 행렬의 대칭 원소들이 됩니다. 4. 결국 의 관계를 가지기 때문에 위에서 구한 와 를 대입시키면 되며 도 3.) 에서 구했던 걸로 대입을 하면 됩니다.   참고 자료 :   * <https://www.youtube.com/watch?v=rziHzFk5JyU> * <https://www.youtube.com/watch?v=HeGdlgB8450> (해당 자료를 참고하여 문제를 풀어주세요.) * <https://angeloyeo.github.io/2019/08/01/SVD.html#google_vignette> |

다음과 같은 행렬에 SVD 를 적용하여 나오는 행렬을 구하시오.

2^2, 0, 0

0, 1, 0

**2-3** © **:** .ipynb 파일에서는 SVD (Singular Value Decomposition, 특이값-분해) 를 이미지 압축에 활용하는 예시를 보여주고 있습니다. 해당 코드를 확인한 후, 새로운 사진에 대해 원본에 비해서 적은 용량을 차지하면서도 원본에 대한 정보를 유지해주는 차원 수가 무엇인지 알아냅시오.

60차원

|  |  |
| --- | --- |
| **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.

1. ! 귀무가설과 대립가설을 설정하시오.

H0: 판 두께의 표준편차는 1.5mm 이하이다. (혹은 =1.5mm)

H1: 판 두께의 표준편차는 1.5mm를 초과한다.

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

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

폰트, 친필, 상징, 라인이(가) 표시된 사진

자동 생성된 설명 귀무가설 하의 (표준편차 = 1.5) 검정통계량 = 20.624 ~ 카이제곱 분포(자유도 9)

유의수준 5%, 자유도 9에서 임계값은 16.010로 검정통계량이 더 크므로 귀무가설 기각.

신뢰구간:

친필, 폰트, 서예, 라인이(가) 표시된 사진

자동 생성된 설명

-> (2.74,13.96)

|  |  |
| --- | --- |
| **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 비율 을 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.

H0: u1 = u2 (두 모집단 평균이 같다.)

H1: u1 > u2 (학회원 평균이 더 큼.)

**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. |

합동 표준편차 Sp = 7.05

검정통계량 t = 두 표본평균 차 / Sp \* (2/n)^2 = 2.364

유의수준 5%, 자유도 250-2 = 248에서 임계값은 (단측검정)

1.645

검정통계량이 더 크므로 귀무가설 기각. 학회원 평균 키가 더 크다는 가설 지지.

**4-2** © **:** 한 학우가 이번에는 각 학회의 평균 키가 똑같다는 주장을 하였습니다. 해당 학우가 제공한 ESC 학회의 학회원별 키 데이터를 활용해 가설검정을 진행하고자 합니다.

**a)**  ! 귀무가설과 대립가설을 설정하시오.

귀무가설: u1 = u2 (두 학회 평균 키 같다.)

대립가설: ,u1 =! U2(같지 않다.)

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

|  |
| --- |
| * One-way Anova Test 를 활용해서 사용하는 문제입니다. * 활용해야 될 함수는scipy.stats.f\_oneway 입니다. |

|  |  |
| --- | --- |
| **문제 5** | © Numpy + Pandas 활용 |
| 기초과제.ipynb 파일에 제공된 문제를 참고하여 수행하기 바랍니다. (네) | |

|  |  |
| --- | --- |
| Reference | **Data Science Lab** |
| - Introduction to Mathematical Statistics(8판, Hogg et.al)  - 23-2 기초과제 1 ( 9기 이성균 )  - 24-1 기초과제 1 ( 10기 신재우 ) | 담당자 : 11기 김현진, 11기 김정우  [Rlaguswls186790@yonsei.ac.kr](mailto:Rlaguswls186790@yonsei.ac.kr)  [kjungwoo@yonsei.ac.kr](mailto:kjungwoo@yonsei.ac.kr) |