**Probability and Statistics, Final Exam**

**Your ID#: [ 2019270664 ]**

**Your Name: [ 윤 재 현 ]**

|  |
| --- |
| **[Instruction] 안내**  **(1)** Please read and answer the questions carefully. Write your answers in Korean or English.  (문제를 잘 읽고 신중하게 답변하십시오. 영어 또는 한국어로 답안을 작성하세요).  **(2)** Please compress (.zip) and submit this report (.docx), and the entire R-code (.R file) you wrote into the BlackBoard.  (이 시험지와 답변에 사용된 R코드를 압축하여 블랙보드에 지정된 시간까지 반드시 제출하십시오.)  Final Exam: 15:00 PM ~ 16:30 PM (90 Min.)  \*I highly recommend that you start submitting to the Blackboard at 16:25PM, at least 5 minutes before the end. If you do not submit by the given time, you will receive 0 points. (시험 끝나기 5분전인 16시 25분에는 블랙보드에 제출을 시작하길 권장합니다. 16시 30분에는 제출할 수 없도록 닫히며, 제출을 못했을 시 당연 0점 처리됩니다.)  **(3)** The R-code should be executable when the TA runs. The submitted compressed file (.zip) must be named Final\_YourID\_Yourname.zip.  (R 코드는 TA가 돌렸을 시 깔끔하게 돌아가야 하며, 제출될 압축파일은 반드시 Final\_YourID\_Yourname.zip 로 명명하여 제출하십시오.)  **(4)** There is a partial score. Even if you can't resolve it completely, I hope you can try it as far as you can.  (부분 점수가 있으므로, 완벽하게 해결하지 못하더라도 할 수 있는 만큼 시도해 보길 바랍니다.)  **(5)** All the questions were asked at a level that, given enough time, a student who followed the class well could solve it 100%. However, since the given time is very short, it would be very insufficient time to solve the given total of 9 problems by referring to other materials.  (모든 문제는 충분한 시간이 주어지면 수업을 잘 따라온 학생이 100% 풀 수 있는 수준으로 출제되었습니다. 그러나 주어진 시간이 매우 짧으므로, 주어진 9개의 문제를 다른 자료를 참고하여 풀기에는 매우 부족한 시간일 것입니다.)  I don't expect you to be able to solve all the problems in any given time. The ability to "select & focus" at a given time in a situation where you are aware of your abilities is also very important in this era. So, find the problem you can resolve quickly and try to resolve it!  (여러분이 주어진 시간에 모든 문제를 해결할 수 있기를 기대하지 않습니다. 자신의 능력을 파악하고 있어, 주어진 시간에 '선택 및 집중' 할 수 있는 능력도 이 시대에서 매우 중요한 능력입니다. 본인의 능력을 고려했을 때, 최대한 빠르게 해결할 수 있는 문제를 찾아서 해결하십시오.)  I wish all of you good results.  (모두들 좋은 결과 있기를 바랍니다.) |

**[Q1 – Q5]** Load the three files ***"Data1\_PS\_2020.txt", "Data1\_PS\_2021.txt", and "Data1\_PS\_2022.txt"*** into the R environment. This is the actual grades of P&S class students collected for 3 years, including your scores so far, at the actual Korea University Sejong Campus. Since the lecture materials and contents have not changed in three years, it can be assumed that the difficulty of the classes has not changed. Also, it is assumed that students in each year are independent of each other. Except for Grade (A+, B+ ,..., F), grades are assumed to be continuous random variables. Based on these data, test the hypotheses below.

**(Kor: "Data1\_PS\_2020.txt", "Data1\_PS\_2021.txt", "Data1\_PS\_2022.txt" 세 파일을 R환경에 불러오십시오. 이것은 실제 고려대학교 세종캠퍼스에서, 여러분들의 지금까지 점수를 포함하여, 3년동안 수집된 P&S수업 수강생들의 실제 성적입니다. 3년동안 강의자료 및 컨텐츠가 바뀌지 않았으므로, 수업의 난이도는 변하지 않았다고 가정할 수 있습니다. 또한, 각각 년도의 수강생들은 서로 독립이라고 가정합니다. 학점 (A+, B+ ,..., F)를 제외한 나머지 성적들은 연속형 확률 변수라 가정합니다. 이 자료들을 바탕으로 하단의 가설들을 검정하십시오.)**

|  |
| --- |
| **[Data Structure]** |

**(Q1)** Dr. Seo, who can't sleep these days, has a lot of trouble with the students of the 2022 P&S class. This is because the significant percentage of students who do not submit assignments has increased rapidly to an unprecedented level. However, no matter how many times I reviewed the quizzes, the difficulty didn't change at all. So, Dr. Seo, who manages national statistics, made this hypothesis. "***For students in 2022, due to the impact of COVID-19 or another reasons, the motivation for studying will be different from the previous years (2020, 2021yrs).***" In this regard, please conduct a hypothesis test based on the appropriate test statistic to arrive at a conclusion at 5% significance level. **(10 Points).**

**(Kor: 요새 잠을 자지 못하는 서교수는 2022 P&S 수업 수강생 때문에 고민이 많습니다. 지금까지 전례를 찾아볼 수 없을 정도로 과제를 아예 제출 안하는 학생의 비율이 급격히 늘어났기 때문입니다. 하지만, 아무리 퀴즈들을 다시 검토해봐도 난이도는 전혀 변하지 않았습니다. 국가 통계를 관리하는 서교수는 그래서 이러한 가설을 세웠습니다. "*2022년도 수강생들은 COVID-19 또는 그 외의 영향으로 인해 공부에 대한 동기부여가 그 이전 년도들 (2020, 2021yrs)와 다를 것이다. (a.k.a. 수듄 차이?)* " 본 가설에 대해서 적절한 검정 통계량을 바탕으로 5% 유의수준에서 가설검정을 수행하고 결론에 이르십시오.)**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **[Answer]**  **\***Please describe the reason for choosing the test statistic & random variables, the results of the hypothesis test, interpretation, and conclusion in a logical manner. (Hint: Hypotheses can be tested on at least four different random variables.)  **(Kor: 통계량과 확률 변수를 선정한 이유, 가설 검정 결과, 해석, 결론 등을 논리성을 갖추어 서술하십시오; 힌트: 최소 4가지 다른 확률변수에 대해 가설 검정 가능합니다.)**  먼저 현재 2022년도 수강생들의 데이터에서 확인할 수 있는 것은 HW1, 2, 3의 점수와 중간고사의 점수이며 총 점수의 합을 확인 할 수 있다. 이전 년도와 비교할 수 있는 데이터는 결국 HW123과 중간고사의 점수이다.  그러므로 가설을 다음과 같이 말할 수 있다.  H1: 2022년도 수강생들은 점수는 이전 년도(2021, 2020)과 다르다.  H0: 2022년도 수강생들은 점수는 이전 년도(2021, 2020)과 같다.  또한 3가지 그룹을 가진 중간고사 이전의 점수들의 비교임으로, multiple group test를 진행해야한다. 이때 3그룹의 크기 모두 30이 넘기때문에 parametric method인 ANOVA test를 시행한다.  pvalue가 유의수준(5%)보다 작기 때문에 null hypthesis를 기각하고, 우리의 alternatvie hypothesis를 채택한다.   |  |  |  | | --- | --- | --- | | **random variable** | **p value** | **reject/no reject** | | HW1 | 5.701e-07 | reject | | HW2 | 1.835e-07 | reject | | HW3 | 8.051e-05 | reject | | Midterm | 0.0001551 | reject |   4 점수 분포 모두 H0을 기각하고, H1을 채택할 수 있었다. 이에 따라 2022년도 수강생들의 중간고사 이전 점수들은 모두 2021, 2020년도 수강생들의 중간고사 점수와 다르다는 것을 알 수 있었고, 이 의미는 2022년도 수강생들의 학업 동기부여가 이전 년도 수강생들과 다르다고 할 수 있다. |

**(Q2)** Please **provide evidence to support the conclusions in [Q1] through visualization**. **(10 Points).**

**(Kor: 시각화를 통해 [Q1]의 결론을 뒷받침하는 증거를 제공하십시오.)**

|  |
| --- |
| **[Answer]**      위 4개의 이미지를 살펴보면 2022년도의 박스의 위치가 모두 2020, 2021년도의 박스의 위치보다 낮게 위치하고 있다는 사실을 알 수 있다. 이를 통해 Q1과 마찬가지로 2022년도의 수강생들의 학업 성취도가 이전 년도 학생들보다 평균적으로 낮다는 것을 알 수 있고, 학업 동기부여가 낮다고 할 수 있다. |

**(Q3)** Which category (random variables; i.e. HW1, HW2, ..., Grade) has the most statistically significant difference in among the three years of 2020, 2021, and 2022? Please choose a statistically testable random variables, and test it at the 5% significance level, and arrive at a conclusion based on an appropriate test statistic. **(10 Points).**

**(Kor: 2020, 2021, 2022 각각 3개년에서 통계적으로 가장 유의한 차이가 나는 항목(확률변수; i.e. HW1, HW2, ..., Grade)는 무엇인가요? 통계적으로 검정이 가능한 확률 변수를 선택하고 적절한 검정 통계량을 바탕으로 5% 유의수준에서 검정 후 결론에 이르십시오.)**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **[Answer]**  **\***Please describe the reason for choosing the test statistic & random variables, the results of the hypothesis test, interpretation, and conclusion in a logical manner.  **(Kor: 통계량과 확률 변수를 선정한 이유, 가설 검정 결과, 해석, 결론 등을 논리성을 갖추어 서술하십시오.**  먼저 2022년도에서 다른 년도와 비교할 수 있는 데이터는 HW1, HW2, HW3, Midterm만 존재한다.  이 확률 변수 4가지를 통해 비교를 시행할 것이다. 하지만 이 비교는 위의 Q1에서 시행을 했다. 이를 통해 살펴보고자 한다.   |  |  |  | | --- | --- | --- | | **random variable** | **p value** | **rank** | | HW1 | 5.701e-07 | 2 | | HW2 | 1.835e-07 | 1 | | HW3 | 8.051e-05 | 3 | | Midterm | 0.0001551 | 4 |   위에 만든 표를 동일하게 가져온 다음 p value의 크기에 따라 random variable을 살펴보면 다음과 같다. 먼저 제일 작은 값은 HW2이고, 그 다음으로 HW2, HW3, Midterm 순이다. pvalue가 작을 수록 유의수준이 작아지더라도 H0을 기각할 확률이 높아진다. 그러므로 가장 유의미한 차이가 나는 확률 변수는 HW2라고 할 수 있다. |

**(Q4)** Please **provide evidence to support the conclusions in [Q3] through visualization**. **(10 Points).**

**(Kor: 시각화를 통해 [Q3]의 결론을 뒷받침하는 증거를 제공하십시오.)**

|  |
| --- |
| **[Answer]**      위 그림을 보면 HW3, Midterm, HW1의 경우 세 그룹은 다르다고 해도, 두 그룹은 비교적 비슷한 위치에 박스의 중앙선(평균)이 위치하고 있다. 하지만 HW2의 경우 3가지 그룹에서 가장 분포가 차이를 보이고 있다. 그러므로 Q3과 마찬가지로 HW2가 가장 큰 차이를 보이는 유의미한 random variable이라고 할 수 있다. |

**(Q5)** Please **predict your final grade (A+, B+, ..., F) this semester through [Q1~Q4] and basic statistics or visualizations that you think are necessary**. The prediction **must be supported by statistical evidence (i.e. on average...)**. Is it possible to make predictions with data for 2020 and 2021? If you think it's possible, you've learned about "*The Necessity of Statistics*" this semester. Of course, providing evidence requires a detailed understanding, but statistics are the foundation of virtually all artificial intelligence and data science techniques. **(10 Points).**

**(Kor: [Q1~Q4] 및 본인이 필요하다고 생각하는 기초 통계량 또는 시각화 등을 통해, 이번 학기 본인의 최종 학점 (A+, B+, ..., F)를 예측해 보십시오. 그 예측에는 반드시 통계적 근거가 뒷받침 되어야 합니다 (예: 평균적으로...?). 2020년도와 2021년도의 자료를 가지고 예측이 가능합니까? 가능하다고 생각이 들면, 당신은 이번 학기를 통해 "*통계의 필요성*" 에 대해 알게 된 것입니다. 물론, 근거를 제공하는 것은 이 수준을 넘어 보다 디테일한 이해가 필요합니다만, 통계는 사실상 모든 인공지능 및 데이터 사이언스 기술의 근간이 됩니다.)**

|  |
| --- |
| **[Answer]** |

**[Q6 – Q9] Load “Data2.txt” file**. This data shows the **number of tardiness** for **9 male and 9 female** students respectively. Based on this data, please perform the suitable hypothesis test about the difference in # of tardiness between male and female students.

**(Kor: "Data2.txt" 파일을 로드하십시오. 이것은 남학생 9명과 여학생 9명의 지각 횟수를 기록한 자료입니다. 이 자료를 바탕으로 남학생과 여학생의 지각 횟수 차이에 대한 적절한 가설 검정을 수행하십시오.)**

**(Q6)** The number of tardies in each gender naturally follows a Poisson distribution.In this situation, graduate student Kim argued that a test statistic of the form below would be superior to known statistical methodologies for the data above:

**(Kor: 각 성별의 지각 횟수는 포아송 분포를 따릅니다. 이 상황에서, 대학원생 Kim은 아래와 같은 형식의 검정 통계량이 위의 자료에 대해 이미 알려진 통계적 방법론에 비해 나을 것이라고 주장하였습니다:)**

Based on the proposed test statistic, **please investigate statistically whether there is a difference in the number of tardiness by gender** **(20 Points).**

**(Kor: 제시된 검정 통계량을 바탕으로 성별에 따라 지각 횟수에 차이가 있는지 통계적으로 조사하십시오.)**

|  |
| --- |
| **[Code, result, interpretation, conclusion, and etc.] \*코드, 결과, 해석, 결론 등 기술**  kimStat은 다음과 같이 구할 수 있다. |

**(Q7)** Please Investigate statistically whether **there is a difference in the number of tardiness by gender** through a representative **nonparametric** method **(10 Points).**

**(Kor: 비모수적 방법을 통해서 성별에 따라 지각 횟수에 차이가 있는지 통계적으로 조사하십시오.)**

|  |
| --- |
| **[Code, result, interpretation, conclusion, and etc.] \*코드, 결과, 해석, 결론 등 기술**  2개의 그룹에 대한 test임으로 wilcoxon test를 시행한다. 이때 가설은 다음과 같이 정의한다.  H1: 남성과 여성의 지각 횟수에 차이가 존재한다.  H0: 남성과 여성의 지각 횟수에는 차이가 없다.  시행을 한결과 p-value는 0.01582가 나왔다. 유의수준을 0.05라고 하면 H0을 기각하고 H1을 채택할 수 있다. 즉 남성과 여성의 지각 횟수 차이가 존재한다. |

**(Q8)** Please Investigate statistically whether **there is a difference in the number of tardiness by gender** through a representative **parametric** method **(10 Points).**

**(Kor: 모수적 방법을 통해서 성별에 따라 지각 횟수에 차이가 있는지 통계적으로 조사하십시오.)**

|  |
| --- |
| **[Code, result, interpretation, conclusion, and etc.] \*코드, 결과, 해석, 결론 등 기술**  2개의 그룹에 대한 test임으로 t-test를 시행한다. 이때 가설은 다음과 같이 정의한다.  H1: 남성과 여성의 지각 횟수에 차이가 존재한다.  H0: 남성과 여성의 지각 횟수에는 차이가 없다.  시행을 한결과 p-value는 0.0120786가 나왔다. 유의수준을 0.05라고 하면 H0을 기각하고 H1을 채택할 수 있다. 즉 남성과 여성의 지각 횟수 차이가 존재한다. |

**(Q9)** As in [Q6-Q8], different methodologies can be applied to resolve the same problem. Please compare/evaluate the three analysis methods and reach a conclusion as to which of the analysis methods you have performed makes the most sense. **(10 Points).**

**(Kor: [Q6-Q8]에서 처럼 같은 문제 해결을 위해 다양한 방법론을 적용할 수 있습니다. 본인이 수행한 분석 방법 중 어느 것이 가장 타당한지, 세 가지 분석 방법을 비교/평가 하여 결론에 이르십시오.)**

|  |
| --- |
| **[Answer] \***within 1 page |

**-At the end-** Well done for one semester!