**Probability & Statistics, Homework #3**

DCCS221(00)

**Your ID#: [2021270682]**

**Your Name: [ Donguk Kim (김동욱) ]**

| **[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)** The logical flow is more important than the amount of answers. However, it is also important to write down enough of what you learn in class.  (논리적 흐름이 답변의 양보다 훨씬 중요합니다. 하지만, 당연히 수업시간에 다룬 내용을 충분히 적는 것이 중요합니다)  **(3)** Please compress 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.)  **(4)** The R-code should be executable when the TA runs. The submitted compressed file (.zip) must be named HW3\_YourID\_Yourname.zip.  (R 코드는 TA가 돌렸을 시 깔끔하게 돌아가야 하며, 제출될 압축파일은 반드시 HW3\_YourID\_Yourname.zip 로 명명하여 제출하십시오.) |
| --- |

**[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]**  **텍스트, 스크린샷이(가) 표시된 사진  자동 생성된 설명** |
| --- |

**(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.

**(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.)**  여러 독립 그룹 간의 평균 차이가 우연에 의한 것인지, 아니면 실제 통계적인 의미가 있는지를 판단하기 위한 방법으로 ANOVA 방법을 사용했다. 그 결과 HW2의 pvalue가 가장 작은것을 확인하여 HW2를 변수로 선택했다.  **Alternative Hypothesis 대립가설**  H1: The average HW2 among those with 2020!=2022  H1: The average HW2 among those with 2021!=2022  **Null Hypothesis 귀무가설**  H0: The average HW2 among those with 2020=2022  H0: The average HW2 among those with 2021=2022  위의 가설을 토대로 표본이 충분히 크기 때문에 표본이 정규분포를 따른다고 볼 수 있고 따라서 모수적 방법인 t.test를 실행해였다. 2022년도의 학생들을 2020그리고 2021학생들과 비교하였을 때 각각 3.852e-08<0.05, 0.0005605<0.05 로 유의미합니다. 따라서 5% 유의 수준 하에서 Null hypothesis는 reject되고 Alternative hypothesis가 select됩니다.  따라서 HW2를 기준으로 검정한 결과 2022 학생들의 수준이 2020, 2021 학생들의 수준보다 떨어진다고 볼 수 있습니다.  이를 뒷받침 하기 위해 2020 그룹과 2021 그룹을 한 그룹으로 묶어 2022그룹과 t.test를 진행해본 결과 p-value가 2.754e–06로 5% 유의 수준 하에서 Null hypothesis는 reject되고 Alternative hypothesis가 select되어 같은 결과가 나왔으므로 이 또한 가설검정을 뒷받침 합니다. |
| --- |

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

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

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

**(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.

**(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.**    여러 독립 그룹 간의 평균 차이가 우연에 의한 것인지, 아니면 실제 통계적인 의미가 있는지를 판단하기 위한 방법으로 ANOVA 방법을 사용했다. 어떤 연속형 확률변수가 가장 유의미한 차이를 보이는지를 조사하기 위해 HW1, HW2, HW3, Midterm 각각에 대해 ANOVA 검정을 진행하고 p-value 값을 얻었다. p-value 값을 확인해보니 5% 유의수준 하에서 모든 변수들이 유의미하게 차이가 나는 것을 확인했다. 그중 HW2에서 p-value가 1.84e-07로 가장 작기 때문에 HW2가 3개년에서 통계적으로 가장 유의한 차이가 나는 항목입니다. |
| --- |

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

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

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

**(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.

**(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]**    표본이 충분히 크기 때문에 두 개의 값이 상관관계를 가지는 변수를 비교하는 모수적 방법인 pearson’s correlation을 사용하였습니다. 성적에 가장 큰 영향을 미치는 중간고사를 변수로 선택했으나 2021년도의 값이 0.05228로 0.05보다 크므로 5%유의수준 하에서 유의하지 못하여 기말고사 문제와 가장 유사한 HW3와 기말고사 점수를 선택하였습니다. 그 결과 HW3와 final의 p-value는 2020,2021년 각각 2.021e-06, 1.917e-06으로 2020,2021년도의 의 HW3 와 final은 5% 유의수준 하에서 양의 상관관계를 가진다고 말할 수 있다.  따라서 HW3를 기준으로 검정한 결과 HW3의 결과가 기말고사 점수에 영향을 끼칠 수 있고 이는 내가 B+그룹의 평균 점수만큼의 HW3 점수를 받을 수 있다면 양의 상관관계에 따라 기말고사에 좋은 성적을 받을 수 있고 따라서 B+ 이상의 성적을 받을 가능성이 높다. |
| --- |

**[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: "Data1.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**.

**(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**      계산된 p-값이 0.001로 5%유의수준 하에, p-값이 0.05보다 낮으므로, 귀무 가설(성별 간의 지각 횟수에 차이가 없다는 가설)을 기각할 충분한 증거가 있습니다. 이는 성별에 따른 지각 횟수의 차이가 통계적으로 유의미하다는 것을 나타냅니다. |
| --- |

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

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

| **[Code, result, interpretation, conclusion, and etc.] \*코드, 결과, 해석, 결론 등 기술**    **Alternative Hypothesis 대립가설**  H1: The average 지각횟수 among those with 남자!=여자  **Null Hypothesis 귀무가설**  H0: The average 지각횟수 among those with 남자=여자  표본이 충분하지 않기 때문에 표본이 정규분포를 따르지 않을 수 있고 따라서 비모수적 방법인 wilcoxon 방법의 결과를 더 신뢰할 수 있다. 성별에 따라 유의미한 차이를 보이는지를 조사하기 위해 wilcoxon 방법을 진행하고 p-value 값을 얻었다. p-value 값을 확인해보니 0.01582325<0.05로 5% 유의수준 하에서 성별간 유의미하게 차이가 나는 것을 확인했다. |
| --- |

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

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

| **[Code, result, interpretation, conclusion, and etc.] \*코드, 결과, 해석, 결론 등 기술**    **Alternative Hypothesis 대립가설**  H1: The average 지각횟수 among those with 남자!=여자  **Null Hypothesis 귀무가설**  H0: The average 지각횟수 among those with 남자=여자  모수적 방법인 t.test를 통해 0.0120786<0.05로 유의미합니다. 따라서 5% 유의 수준 하에서 Null hypothesis는 reject되고 Alternative hypothesis가 select됩니다. 이는 성별에 따른 지각 횟수가 통계적으로 유의미하다는 것을 나타냅니다. |
| --- |

**(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.

**(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  위에 주어진 자료에서 남자 여자의 표본의 크기가 9개씩 있으므로(30이하), 표본의 크기가 충분히 크다고 볼 수 없다. 즉, 모집단이 정규분포를 따를 것이라는 것을 가정할 수 없고 성별에 따른 지각횟수에 차이가 있다는 가설을 검정하기 위해서는 비모수적 방법인 wilcoxon test를 진행하는 편이 더욱 타당하다고 볼 수 있다. t.test의 경우 표본이 충분히 크기 때문에 표본이 정규분포를 따른다고 볼 수 있고 이에 따라서 결과를 신뢰할 수 있지만 이번 문제의 경우 표본이 너무 적기 때문에 비교적 신뢰성이 떨어진다. |
| --- |

**(Q10 ~ Q12)** Please **perform data analysis using the statistical hypothesis test based on the given “*Data3.txt*”**. Based on the given data, **you can make a variety of hypotheses yourself, conduct analysis, and freely describe the newly discovered thing**.

(Kor. 주어진 자료에서 다양한 가설을 여러분이 상정하고 검정하여 무엇인가 새로운 것을 밝혀나가 보십시오. 자료가 여러분에게 무엇을 말하고자 하는지 알아내려고 애써보세요.)

**(Q10)** **Please list the continuous random variables in the order** in which there is a statistically **significant gender difference** based on a hypothesis test in the given data.

(Kor. 주어진 자료에 가설검정을 수행하여 어떤 연속형 확률 변수가 통계적으로 성별 간 유의미한 차이가 있는지 순서대로 나열하십시오.)

| **[Answer] \*\*\*One page limit. You must prepare a basis according to the hypothesis testing principle, and organize it to arrive at a conclusion. (한 페이지 이내로, 가설검정의 원칙의 기본에 충실하여 증거를 마련하고, 그를 잘 정리한 뒤 결론을 이끌어 내십시오.)**    표본이 충분히 크기 때문에 표본이 정규분포를 따른다고 볼 수 있고 t.test의 결과를 신뢰할 수 있다. 성별에 따라 어떤 연속형 확률변수가 가장 유의미한 차이를 보이는지를 조사하기 위해 Age, Height, Weight, sysBP, HR, Resting\_SaO2, BMI, FEV1pp\_utah, FVCpp\_utah, FEV1\_FVC\_utah 각각에 대해 t test를 진행하고 p-value 값을 얻었다. p-value 값을 확인해보니 5% 유의수준 하에서 Height, Weight, sysBP, FEV1\_FVC\_utah , BMI, HR, FEV1pp\_utah, FVCpp\_utah순으로 성별간 유의미하게 차이가 나는 것을 확인했다. 반면, Age, Resting\_SaO2는 p-value 값이 0.05보다 크기 때문에 5% 유의수준 하에서 성별간 유의미하게 차이가 난다고 보기 어렵다. |
| --- |

**(Q11)** **Please perform the same analysis as (Q1) with a nonparametric test and compare the results.**

(Kor. (Q10)과 동일한 분석을 비모수 검정을 통해 수행해보고 두 결과를 비교하십시오.)

| **[Answer] \*\*\*One page limit.**    표본이 충분한지 여부에 상관없이 비모수적 방법인 wilcoxon 방식으로 p값을 통해 유의수준 알아볼 수 있다. 성별에 따라 어떤 연속형 확률변수가 가장 유의미한 차이를 보이는지를 조사하기 위해 Age, Height, Weight, sysBP, HR, Resting\_SaO2, BMI, FEV1pp\_utah, FVCpp\_utah, FEV1\_FVC\_utah 각각에 대해 wilcoxon test를 진행하고 p-value 값을 얻었다. p-value 값을 확인해보니 5% 유의수준 하에서 Height, Weight, sysBP, FEV1\_FVC\_utah , Resting\_SaO2, BMI, HR, FEV1pp\_utah 순으로 성별간 유의미하게 차이가 나는 것을 확인했다. 반면, Age, FVCpp\_utah는 p-value 값이 0.05보다 크기 때문에 5% 유의수준 하에서 성별간 유의미하게 차이가 난다고 보기 어렵다. |
| --- |

**(Q12)** **Please find out which variables have a close association with the severity of COVID-19** through a hypothesis test.

(Kor. 주어진 자료에서 어떤 변수가 COVID-19의 중증도와 강한 연관성이 있는지를 가설검정을 통해 조사해보십시오.)

| **[Answer] \*\*\*One page limit. You must prepare a basis according to the hypothesis testing principle, and organize it to arrive at a conclusion. (한 페이지 이내로, 가설검정의 원칙의 기본에 충실하여 증거를 마련하고, 그를 잘 정리한 뒤 결론을 이끌어 내십시오.)**    여러 독립 그룹 간의 평균 차이가 우연에 의한 것인지, 아니면 실제 통계적인 의미가 있는지를 판단하기 위한 방법으로 ANOVA 방법을 사용했다. 또한 범주형 데이터에서 두 변수 간의 독립성을 테스트하기 위해 일부 변수의 경우 chi-squared test를 진행했다.  Severity\_Group 에 따라 어떤 연속형 확률변수가 가장 유의미한 차이를 보이는지를 조사하기 위해 Age, Height, Weight, sysBP, HR, Resting\_SaO2, BMI, FEV1pp\_utah, FVCpp\_utah, FEV1\_FVC\_utah 각각에 대해 anova test를 했다.  Severity\_Group 에 따라 어떤 categorical 확률변수가 가장 유의미한 차이를 보이는지를 조사하기 위해 Gender, Race, HaveCough 각각에 대해 chisq.test를 했다.  Height\_CM은 p-value 값이 0.05보다 크기 때문에 5% 유의수준 하에서 성별간 유의미하게 차이가 난다고 보기 어렵다. 그외의 변수들에서는 Severity\_Group에 유의미한 차이를 보였다. |
| --- |

**-At the end-** Good job!