고려대학교
KOREA UNIVERSITY

Probability & Statistics, Homework #3

DCCS221(00)

Your ID#: [2021270678]
Your Name: [나강민]

[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 R 코드를 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 TA 가 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 R 환경에 불러오십시오. 이것은 실제 고려대학교 세종캠퍼스에서, 여러분들의 지금까지 점수를 포함하여, 3 년동안 수집된 P&S 수업 수강생들의 실제 성적입니다. 3 년동안 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]

StudentID	HW1	HW2	HW3	Midterm	Final	Attend	Total	Grade
S2020_29	9.090909	10	10	30	30	10	89.09091	A+
S2020_14	9.090909	10	10	24.06977	26.78571	10	79.94639	A+
S2020_24	8.181818	10	9	27.55814	21.42857	10	76.16853	A+
S2020_10	10	7.272727	8	24.4186	24.64286	10	74.33419	A+
S2020_34	10	10	9	22.67442	22.5	10	74.17442	A+
S2020_28	10	10	6	25.46512	22.5	10	73.96512	A+
S2020_27	9.090909	10	9	21.97674	21.42857	10	71.49622	A+
S2020_4	9.090909	10	9	23.37209	19.28571	10	70.74872	A+
S2020_19	10	10	8	25.46512	16.07143	10	69.53654	A+
S2020_18	10	9.090909	9	19.88372	21.42857	10	69.4032	A+
S2020_8	9.090909	8.181818	7	24.06977	20.35714	10	68.69964	A+
S2020_30	9.090909	10	8	26.16279	15	10	68.2537	B+
S2020_5	9.090909	10	9	25.11628	15	10	68.20719	B+
S2020_6	7.272727	8.181818	9	25.46512	17.57143	10	67.49109	B+
S2020_3	10	9.090909	6	23.37209	17.14286	10	65.60586	B+
S2020_26	9.090909	10	7	16.04651	22.5	10	64.63742	B+
S2020_7	9.090909	10	8	21.97674	13.28571	10	62.35337	B+
S2020_31	7.272727	10	9	20.5814	15	8	61.85412	B+
S2020_22	7.272727	8.181818	6	25.46512	12.85714	10	59.7768	B+
S2020_23	9.090909	9.090909	9	16.39535	11.78571	8	55.36288	B+
S2020_11	9.090909	10	5	23.02326	6.857143	10	53.97131	B+
S2020_9	7.272727	10	9	25.11628	2.142857	9	53.53186	C+
S2020_13	0	7.272727	6	23.02326	17.14286	10	53.43884	C+
S2020_35	7.272727	7.272727	7	24.06977	6.428571	10	52.04379	C+
S2020_25	9.090909	10	5	20.5814	6.857143	10	51.52945	C+
S2020_32	9.090909	8.181818	6	21.62791	6.428571	10	51.32921	C+
S2020_21	7.272727	10	8	16.74419	7.5	7	49.51691	C+
S2020_16	9.090909	7.272727	6	18.48837	8.571429	10	49.42344	C+
S2020_2	7.272727	9.090909	6	15	6.428571	10	43.79221	C+
S2020_33	7.272727	4.545455	1	20.23256	5.785714	10	38.83645	C+
S2020_17	9.090909	4.545455	5	10.11628	6.428571	7	35.18121	C+
S2020_15	10	6.363636	0	10.11628	2.142857	9	28.62277	C+
S2020_20	9.090909	0	0	17.44186	0	9	26.53277	D+
S2020_12	7.272727	10	0	4.186047	2.142857	10	23.60163	D+
S2020_1	0	7.272727	1	0	2.142857	7	10.41558	D+

[Data1_PS_2020.txt]
Final scores of P&S in 2020yr

StudentID	HW1	HW2	HW3	Midterm	Final	Attend	Total	Grade
S2021_521	10	10	10	24	24.375	10	88.375	A+
S2021_532	9	9	9	25	26.25	10	88.25	A+
S2021_518	9	10	10	23.33333333	22.875	10	85.208333	A+
S2021_515	10	9	10	26	18.75	10	83.75	A+
S2021_51	10	10	10	12	30	10	82	A+
S2021_517	9	10	10	16.66666667	24.375	10	80.041667	A+
S2021_531	9	10	8	22.33333333	16.875	10	76.208333	A+
S2021_53	10	10	10	15	20.625	10	75.625	A+
S2021_538	8	8	10	20.33333333	19.125	10	75.458333	A+
S2021_520	8	10	10	17	19.5	10	74.5	A+
S2021_536	9	8	3	19.33333333	24.375	10	73.708333	A+
S2021_55	10	8	10	12.33333333	20.625	10	70.958333	A+
S2021_529	9	8	5	22	16.875	10	70.875	B+
S2021_542	9	10	10	30	1.875	10	70.875	B+
S2021_533	10	9	6	15.66666667	18.75	10	69.416667	B+
S2021_519	9	8	9	19.33333333	13.5	10	68.833333	B+
S2021_523	9	7	10	19.66666667	11.25	10	66.916667	B+
S2021_516	9	9	10	9	19.125	10	66.125	B+
S2021_56	10	9	10	11.66666667	15	10	65.666667	B+
S2021_525	9	10	10	7.666666667	18.75	10	65.416667	B+
S2021_537	10	10	8	9.666666667	17.25	10	64.916667	B+
S2021_513	10	10	6	19.66666667	7.5	10	63.166667	B+
S2021_539	9	6	8	18.33333333	11.25	10	62.583333	B+
S2021_58	9	9	10	17.33333333	3.75	10	59.083333	B+
S2021_59	9	10	3	15	11.25	10	58.25	B+
S2021_535	10	7	5	16	9.375	10	57.375	B+
S2021_530	9	6	6	19.33333333	5.625	10	55.958333	C+
S2021_543	9	6	9	10	11.25	10	55.25	C+
S2021_514	9	10	5	13.66666667	7.5	10	55.166667	C+
S2021_541	10	10	9	7	8.625	10	54.625	C+
S2021_510	9	6	6	9.333333333	15	8	53.333333	C+
S2021_522	10	0	0	17.33333333	13.125	10	50.458333	C+
S2021_528	9	6	0	13.33333333	5.625	10	43.958333	C+
S2021_511	9	9	6	3.666666667	5.625	10	43.291667	C+
S2021_526	10	5	0	17	0	10	42	C+
S2021_534	9	9	1	5	7.5	10	41.5	C+
S2021_57	7	0	0	19.33333333	0	10	36.333333	C
S2021_540	8	0	0	8	9.375	10	35.375	C
S2021_524	8	0	0	13.33333333	0	10	31.333333	C
S2021_54	8	0	0	9	0	10	27	C
S2021_527	5	6	0	1.666666667	0	10	22.666667	C
S2021_52	9	0	0	5.666666667	1.875	6	22.541667	D+
S2021_512	0	0	0	10.66666667	0	10	20.666667	F

[Data1_PS_2021.txt]
Final scores of P&S in 2021yr

StudentID	HW1	HW2	HW3	Midterm	Sum_HW123_Midterm
S2022_518	9.5	8.9474	10	29.277108	57.724476
S2022_542	9.5	8.4211	9.5238095	30	57.444862
S2022_516	10	8.9474	9.4285714	25.301205	53.677144
S2022_530	9.5	8.9474	8.5714286	25.662651	52.681475
S2022_56	8	9.3684	8.0952381	26.385542	51.849203
S2022_511	9.3	8.8421	8.5714286	26.746988	51.460251
S2022_517	9.5	8.4211	10	21.686747	49.607796
S2022_58	7.3	7.8947	9.047619	24.939759	49.182141
S2022_541	10	9.4737	9.1428571	19.879518	48.496035
S2022_532	9	9.4737	2.8571429	25.662651	46.993477
S2022_512	10	10	7.3333333	18.072289	45.405624
S2022_528	9.5	8.4211	7.6190476	18.795181	44.33528
S2022_515	9.5	8.9474	3.8095238	21.686747	43.943639
S2022_531	8.5	10	3.3333333	21.325301	43.158634
S2022_543	10	8.4211	9.5238095	11.204819	39.149681
S2022_545	9.5	9.1579	7.1428571	11.204819	37.005571
S2022_53	2	8.4211	4.7619048	21.325301	36.508288
S2022_519	9.3	9.4737	3.6190476	11.927711	34.32042
S2022_52	4.5	4.2105	2.2857143	21.686747	32.682987
S2022_529	7	2.1053	0	21.325301	30.430964
S2022_520	8.8	8.3158	3.8095238	9.0361446	27.961457
S2022_59	9	2.8316	0	13.373494	25.005079
S2022_527	2.5	4.7368	0	17.710843	24.947854
S2022_510	4.5	7.3684	0	12.650602	24.519023
S2022_539	6.1	0	0	18.072289	24.17228
S2022_524	5	7.8947	8.4761905	1.8072289	23.178153
S2022_538	0.5	2.1053	2.3809524	14.819277	19.805496
S2022_54	6	3.6842	0	9.3975904	19.08108
S2022_57	4	5.2632	3.1428571	8.5060241	18.912039
S2022_525	4	0	0	14.819277	18.819277
S2022_522	0	0	0	17.710843	17.710843
S2022_51	0	0.5263	0	11.927711	12.454023
S2022_526	0	0	0	11.927711	11.927711
S2022_540	0	0	0.0952381	9.7590361	9.85427
S2022_544	1.5	2.6316	0.0952381	5.060241	9.287058
S2022_521	0	0	0	8.313253	8.313253
S2022_534	0	0	0	7.9518072	7.951807
S2022_523	0	0	2.8571429	4.6987952	7.555938
S2022_537	0	0	0	7.2289157	7.228915
S2022_547	0	0	0	6.5060241	6.506024
S2022_513	0	0	0	5.7831325	5.783133
S2022_533	0	0.5263	0	0	0.526316
S2022_55	0	0	0	0	0
S2022_514	0	0	0	0	0
S2022_535	0	0	0	0	0
S2022_536	0	0	0	0	0
S2022_546	0	0	0	0	0

[Data1_PS_2022.txt]
Aggregate score to date for P&S class in 2022
This is your score distribution.

(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. "*2022 년도 수강생들은 COVID-19 또는 그 외의 영향으로 인해 공부에 대한 동기부여가 그 이전 년도들 (2020, 2021yrs)와 다를 것이다. (a.k.a. 수준차이?)* " 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 4 가지 다른 확률변수에 대해 가설 검정 가능합니다.)

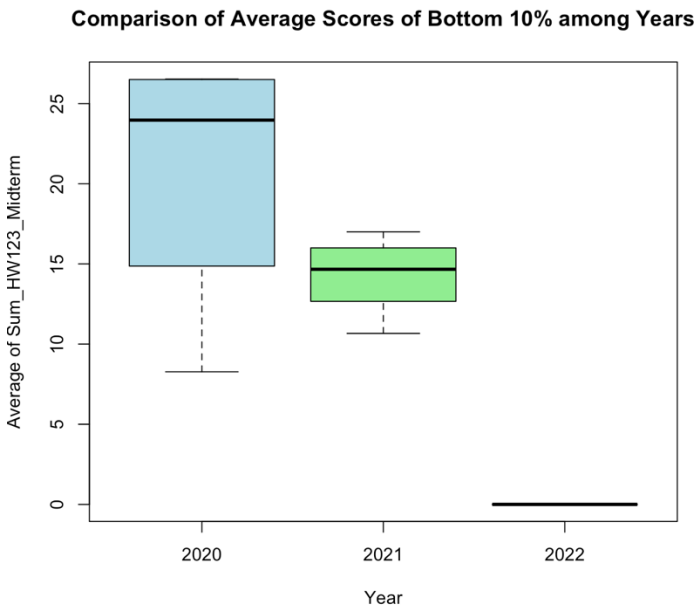
본 분석에서는 세 년도(2020 년, 2021 년, 2022 년)의 학생들의 성적 분포가 같은지를 확인하기 위해 30 개가 넘는 샘플 사이즈를 통해 ANOVA 검정을 수행했습니다.

학생들의 과제 점수(HW1, HW2, HW3)와 중간고사 점수(Midterm)가 포함되어 있습니다. 이 점수들을 합산하여 'Sum_HW123_Midterm'이라는 새로운 확률 변수를 생성했습니다. 직관적으로 상위권에 있는

학생들은 성적에 차이가 크게 없기 때문에 각 년도의 'Sum_HW123_Midterm' 값의 하위 10%를 선택하여 다른 확률 변수인 'Average'를 생성했습니다. 이 변수는 각 년도의 성적 분포의 하위 10%를 대표합니다.
ANOVA 검정을 수행하여 세 년도의 성적 분포가 같은지를 확인했습니다. 귀무가설은 "세 년도의 성적 분포는 동일하다"이며, 대립가설은 "적어도 한 년도의 성적 분포는 다르다"입니다.
ANOVA 검정 결과, p-value는 3.86789809646307e-05로, 이는 0.05보다 훨씬 작습니다. 따라서 귀무가설을 기각하고 대립가설을 채택합니다. 이는 세 년도의 성적 분포가 통계적으로 유의미하게 다르다는 것을 의미합니다. 결론적으로, 이 분석을 통해 학생들의 성적 분포가 년도에 따라 변화하고 있음을 확인할 수 있었습니다.

(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 3 개년에서 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.

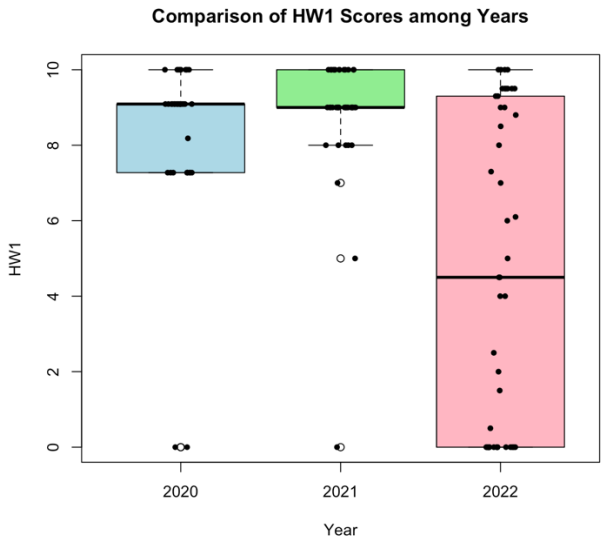
각 항목(HW1, HW2, HW3, Midterm)에 대해 ANOVA 검정을 수행한 이유는, 세 년도(2020년, 2021년, 2022년) 간의 HW1, HW2, HW3, Midterm(4가지 변수밖에 없기 때문) 분포를 파악하기 위함입니다.
ANOVA 검정의 귀무가설은 "세 년도의 성적 분포는 동일하다"이며, 대립가설은 "적어도 한 년도의 성적 분포는 다르다"입니다.
검정 결과, 모든 항목에서 p-value가 0.05보다 훨씬 작게 나왔습니다. 이는 귀무가설을 기각하고 대립가설을 채택하는 것을 의미합니다. 즉, 세 년도 간의 성적 분포는 통계적으로 유의미하게 다르다는 결론을 내릴 수 있습니다.
또한, p-value가 가장 작은 항목은 HW2로, 이 항목에서 가장 큰 차이를 보였습니다. 이는 다른 항목보다 HW2

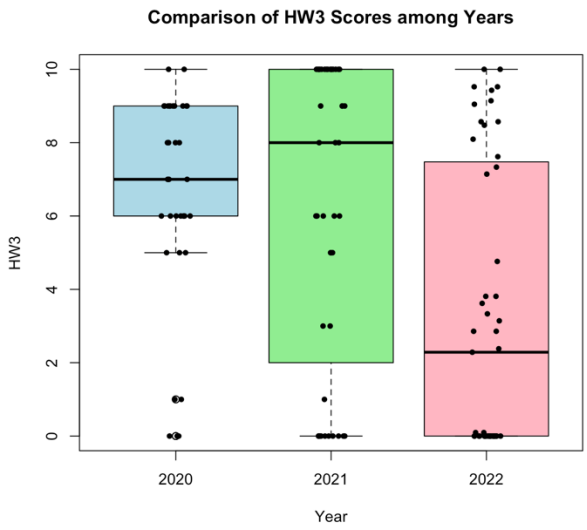
점수에서 년도별 차이가 가장 크게 나타났다는 것을 의미합니다.
따라서, 이 분석을 통해 학생들의 성적 분포가 년도에 따라 변화하고 있음을 확인할 수 있었습니다.

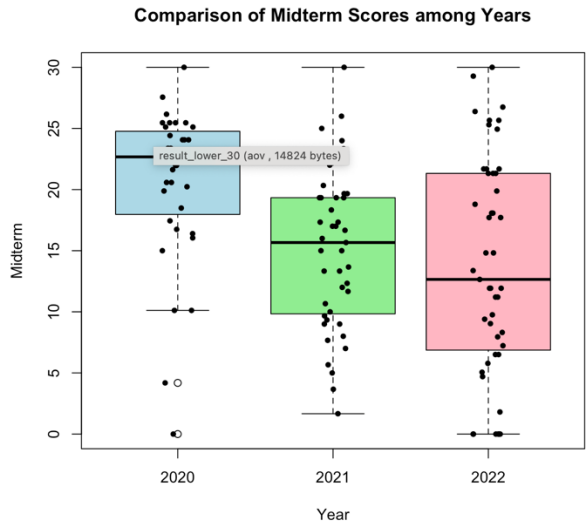
[1] "The p-value of the ANOVA test for HW1 is: 5.70149663555652e-07"
[1] "The p-value of the ANOVA test for HW2 is: 1.83528888313493e-07"
[1] "The p-value of the ANOVA test for HW3 is: 8.05121351880682e-05"
[1] "The p-value of the ANOVA test for Midterm is: 0.00015511516599804 5"

(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...?). 2020 년도와 2021 년도의 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]

나의 최종 학점을 예측하기 위한 2020, 2021 이 가지고 있는 Total 값을 임의로 작성하였습니다.

2020 년과 2021 년 데이터 각각에 대해 선형 회귀 모델을 학습시킨 후, 이를 바탕으로 Total 점수가 60 점일 때 Grade 를 예측해보겠습니다.

회귀 모델에서 예측값은 회귀 계수와 독립변수의 곱으로 계산할 수 있습니다. 따라서 'Total'이 60 일 때 'GradeNum'을 예측하려면, 각 모델의 회귀 계수를 60 과 곱하면 됩니다.

Total 점수가 60 일 때 예상되는 GradeNum 값을 얻을 수 있습니다. 그런 다음 이 값들을 반올림하여 가장 가까운 학점으로 변환하면 됩니다.

예를 들어, 예측된 GradeNum 값이 3.7 이라면 이는 'A0'에 가깝고, 3.3 이라면 'B+'에 가깝습니다. 이렇게 학점을 예측할 수 있습니다.

```
> # 2021 년 모델을 이용한 예측
> predicted_GradeNum_2021 <- model_2021$coefficients[1] + model_2021$coefficients[2] * 60
> print(predicted_GradeNum_2021)
(Intercept)
    3.275069
> # 2020 년 모델을 이용한 예측
> predicted_GradeNum_2020 <- model_2020$coefficients[1] + model_2020$coefficients[2] * 60
> print(predicted_GradeNum_2020)
(Intercept)
    3.456018
```

즉 2021 년과 2020 년 모두 B+에 가깝다는 결론을 도출할 수 있습니다.

중간고사 : 10.42105263
1 차과제 : 7.1
2 차과제 : 8
3 차과제 : 8 (예상)
기말고사 : 15(예상)
Attend : 10(예상)
Total : 60

[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 9 명과 여학생 9 명의 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 Kim 은 아래와 같은 형식의 검정 통계량이 위의 자료에 대해 이미 알려진 통계적 방법론에 비해 나을 것이라고 주장하였습니다:)

$$kimStat = \frac{|\bar{X} - \bar{Y}|}{sd(X) + sd(Y)/8}$$

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

코드:

먼저, `kimStatFunc`라는 함수를 정의했습니다. 이 함수는 두 그룹의 평균 차이를 해당 그룹의 표준 편차의 합으로 나눈 값을 반환하는 함수입니다.

다음으로, 원래의 데이터를 사용하여 `obs_kimStat` 변수에 Kim 의 검정 통계량을 계산했습니다. `Data2_read` 데이터 프레임에서 "Female" 그룹과 "male" 그룹의 "numTardy" 변수를 추출하여 `kimStatFunc` 함수에 적용했습니다.

그 다음, `null_kimStat`이라는 빈 벡터를 생성하고, 1000 번의 반복을 통해 순열 검정을 위한 null 분포를 생성했습니다. 각 반복에서는 `Data2_read` 데이터 프레임의 "Gender" 열을 무작위로 섞은 후, 해당 순열에 대해 `kimStatFunc` 함수를 적용하여 null 분포를 생성했습니다.

마지막으로, 관측된 Kim 의 검정 통계량보다 극단적인 통계량이 나올 확률인 p-value 를 계산했습니다. `null_kimStat` 벡터에서 관측된 통계량보다 절대값이 큰 값을 카운트하고, 이를 반복 횟수와 비교하여 p-value 를 계산했습니다.

결과:
계산된 p-value 는 0.01198801 입니다.

해석:
p-value 는 0.05 보다 작으므로, 통상적인 유의 수준 0.05 에서 귀무 가설을 기각합니다. 따라서, 남성과 여성의 지각 횟수에는 통계적으로 유의미한 차이가 있다고 할 수 있습니다.

결론:
위의 분석을 통해 남성과 여성의 지각 횟수에는 통계적으로 유의미한 차이가 있다는 결론을 얻을 수 있습니다. 이는 Kim 의 검정 통계량을 사용하여 순열 검정을 수행하여 확인되었습니다.

(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

2 개의 그룹에 대한 test 임으로 wilcoxon test 를 시행합니다. 이때 가설은 다음과 같이 정의합니다.
`wilcox.test` 함수가 반환한 결과를 보면, Wilcoxon 순위합 검정의 검정 통계량(W)은 13 이고, p-value 는 약 0.01582 입니다.
p-value 는 귀무 가설(여기서는 두 그룹의 지각 횟수의 분포가 동일하다는 가설)이 참일 경우, 관찰된 통계량 또는 더 극단적인 통계량이 관찰될 확률을 나타냅니다.

이 p-value 는 0.05 보다 작으므로, 통상적인 유의 수준(0.05)에서 귀무 가설을 기각하게 됩니다. 즉, 남성과 여성의 지각 횟수에는 통계적으로 유의미한 차이가 있다고 결론지을 수 있습니다.

(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

`t.test` 함수가 반환한 결과를 보면, 두 표본 t-검정의 검정 통계량(t)은 -2.9013, 자유도(df)는 16 이고, p-value 는 약 0.01041 입니다.
p-value 는 귀무 가설(여기서는 두 그룹의 지각 횟수의 평균 차이가 없다는 가설)이 참일 경우, 관찰된 통계량 또는 더 극단적인 통계량이 관찰될 확률을 나타냅니다.
이 p-value 는 0.05 보다 작으므로, 통상적인 유의 수준(0.05)에서 귀무 가설을 기각하게 됩니다. 즉, 남성과 여성의 지각 횟수에는 통계적으로 유의미한 차이가 있다고 결론지을 수 있습니다.
추가로, 95% 신뢰구간을 확인해보면, 이 구간이 0 을 포함하지 않습니다. 이는 성별에 따른 지각 횟수의 평균 차이가 0 이 아님을 더욱 확증해줍니다.
그리고, 여성 그룹의 평균 지각 횟수는 약 2.44 회, 남성 그룹의 평균 지각 횟수는 약 4.89 회로, 남성 그룹의 지각 횟수가 더 많음을 알 수 있습니다.
따라서, 이 결과는 남성과 여성의 지각 횟수에 차이가 있음을 통계적으로 보여주며, 이 차이는 우연이 아닌 실질적인 차이로 해석될 수 있습니다.

(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

1. 순열 검정: 순열 검정은 두 집단에 대한 차이를 검정하기 위해 모든 가능한 데이터 배치를 고려하는 방법입니다. 이 방법은 데이터의 분포에 대한 가정이 필요하지 않으며, 작은 표본 크기에서도 사용할 수 있습니다. 그러나, 순열 검정은 계산 복잡성이 높은 편이며, 표본 크기가 큰 경우에는 시간이 많이 소요될 수 있습니다.

2. 비모수 검정: 비모수 검정은 데이터의 분포에 대한 가정이 필요 없는 검정 방법입니다. 이 방법은 데이터가 정규 분포를 따르지 않거나, 분산이 동일하지 않은 경우에 유용합니다. 그러나, 비모수 검정은 모수 검정보다 통계적인 검정력이 낮을 수 있습니다.
3. 모수 검정: 모수 검정은 데이터가 특정 분포(대게는 정규 분포)를 따른다는 가정 하에 수행하는 검정 방법입니다. 이 방법은 데이터의 분포가 알려져 있고, 그 분포의 모수(예: 평균, 분산)에 대한 가설을 검정합니다. 모수 검정은 통계적 검정력이 높으며, 데이터의 분포에 대한 정보를 활용하여 더 정확한 결과를 얻을 수 있습니다.

이 세 가지 방법 모두 지각 횟수에 대한 성별 차이를 검정하는데 사용되었고, 모두 유의미한 차이를 발견하였습니다. 그러나 세 방법 중 어느 것이 가장 타당한지는 데이터의 특성과 분석의 목적에 따라 달라집니다. 만약 데이터의 분포가 알려져 있고, 그 분포를 따르는 것이 확인된다면 모수 검정이 가장 강력한 방법일 것입니다.

(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.)**

데이터에서 연속형 변수들은 'Age', 'Height_CM', 'Weight_KG', 'sysBP', 'HR', 'Resting_SaO2', 'BMI', 'FEV1pp_utah', 'FVCpp_utah', 'FEV1_FVC_utah'입니다. 이들 중에서 성별에 따라 통계적으로 유의미한 차이를 보이는 변수를 찾기 위해 가설 검정을 수행하겠습니다.

우선, 각 변수에 대한 가설을 설정합니다. 귀무 가설(H0)은 "성별에 따라 해당 변수의 평균에 차이가 없다"이고, 대립 가설(H1)은 "성별에 따라 해당 변수의 평균에 차이가 있다"입니다.

각 변수에 대해 독립 표본 t-검정을 수행하여 p-value 를 계산합니다. 이 때, p-value 가 0.05 보다 작다면 귀무 가설을 기각하고, 성별에 따라 해당 변수의 평균에는 통계적으로 유의미한 차이가 있다고 결론지을 수 있습니다.

모든 변수에 대해 이러한 과정을 반복합니다. 그리고 나서, p-value 가 0.05 보다 작은 변수들을 p-value 의 오름차순으로 나열합니다. 이렇게 나열된 변수들이 바로 성별에 따라 통계적으로 유의미한 차이를 보이는 연속형 변수들입니다.

성별에 따라 통계적으로 유의미한 차이를 보이는 연속형 변수들은 다음과 같습니다:

1. 'Height_CM' (p-value: 0.000000e+00)
2. 'Weight_KG' (p-value: 2.682517e-169)
3. 'sysBP' (p-value: 8.293869e-15)
4. 'FEV1_FVC_utah' (p-value: 3.482947e-08)
5. 'BMI' (p-value: 6.864698e-07)
6. 'HR' (p-value: 6.826721e-06)
7. 'FEV1pp_utah' (p-value: 1.810843e-02)
8. 'FVCpp_utah' (p-value: 3.993117e-02)

위의 변수들은 모두 p-value 가 0.05 보다 작으므로, 성별에 따라 평균에 통계적으로 유의미한 차이가 있다고 볼 수 있습니다.
반면에 'Resting_SaO2'와 'Age'는 p-value 가 0.05 보다 크므로, 성별에 따른 평균 차이가 통계적으로 유의미하지 않다고 해석할 수 있습니다.
따라서, 'Height_CM', 'Weight_KG', 'sysBP', 'FEV1_FVC_utah', 'BMI', 'HR', 'FEV1pp_utah', 'FVCpp_utah' 이 8 개의 변수는 성별에 따라 통계적으로 유의미한 차이를 보이는 연속형 무작위 변수로 판단할 수 있습니다.

(Q11) <u>Please perform the same analysis as (Q1) with a nonparametric test and compare the results.</u>
(Kor. (Q10)과 동일한 분석을 비모수 검정을 통해 수행해보고 두 결과를 비교하십시오.)

[Answer] ***One page limit.**

비모수 검정(Mann-Whitney U 검정)의 결과와 모수 검정(t-검정)의 결과를 비교해보겠습니다.

비모수 검정의 결과는 다음과 같습니다:

1. 'Height_CM' (p-value: 0.000000e+00)
2. 'Weight_KG' (p-value: 1.421444e-186)
3. 'sysBP' (p-value: 1.677031e-16)
4. 'FEV1_FVC_utah' (p-value: 2.602760e-08)
5. 'HR' (p-value: 8.796487e-07)
6. 'Resting_SaO2' (p-value: 7.169090e-04)
7. 'BMI' (p-value: 1.235464e-02)
8. 'FEV1pp_utah' (p-value: 4.313092e-02)
9. 'FVCpp_utah' (p-value: 9.141779e-02)
10. 'Age' (p-value: 7.284600e-01)

모수 검정의 결과는 다음과 같습니다:

1. 'Height_CM' (p-value: 0.000000e+00)
2. 'Weight_KG' (p-value: 2.682517e-169)
3. 'sysBP' (p-value: 8.293869e-15)
4. 'FEV1_FVC_utah' (p-value: 3.482947e-08)
5. 'BMI' (p-value: 6.864698e-07)
6. 'HR' (p-value: 6.826721e-06)
7. 'FEV1pp_utah' (p-value: 1.810843e-02)
8. 'FVCpp_utah' (p-value: 3.993117e-02)
9. 'Resting_SaO2' (p-value: 4.214585e-01)
10. 'Age' (p-value: 7.785194e-01)

두 결과를 비교해보면, 대부분의 변수에서 비모수 검정과 모수 검정의 결과가 일치합니다. 즉, 성별에 따른 차이가 통계적으로 유의미한 변수들은 대체로 같습니다.

그러나, 'Resting_SaO2'와 'FVCpp_utah'에서는 두 검정의 결과가 다릅니다. 'Resting_SaO2'는 비모수 검정에서는 성별에 따른 차이가 유의미하다고 나타났지만, 모수 검정에서는 그렇지 않았습니다. 반면에 'FVCpp_utah'는 모수 검정에서는 성별에 따른 차이가 유의미하다고 나타났지만, 비모수 검정에서는 그렇지 않았습니다.

(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-19 의 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.)

COVID-19 의 중증도와 연관성이 있는 변수를 찾기 위해, 우리는 Kruskal-Wallis H 검정을 수행하였습니다. 이 검정은 하나의 범주형 변수와 여러 연속형 변수 간의 관계를 분석하는 비모수 통계 방법입니다.

주어진 데이터에 대해 10 개의 연속형 변수('Age', 'Height_CM', 'Weight_KG', 'sysBP', 'HR', 'Resting_SaO2', 'BMI', 'FEV1pp_utah', 'FVCpp_utah', 'FEV1_FVC_utah')에 대해 Kruskal-Wallis H 검정을 수행하였습니다. 그 결과, 아래와 같은 p-value 를 얻었습니다:

- 'Resting_SaO2', 'FEV1pp_utah', 'FVCpp_utah', 'FEV1_FVC_utah': p-value < 0.00001
- 'Age': p-value < 6.671825e-250
- 'HR': p-value < 1.992014e-91
- 'BMI': p-value < 3.786377e-48
- 'Weight_KG': p-value < 2.822723e-41
- 'sysBP': p-value < 0.00001
- 'Height_CM': p-value = 0.4887421

p-value 는 가설검정에서 통계적 유의성을 나타내는 지표로, 값이 작을수록 해당 변수가 'Severity_Group'과 유의미한 관계가 있다는 것을 나타냅니다. 일반적으로 p-value 가 0.05 미만일 때, 해당 변수는 통계적으로 유의미하다고 판단합니다.

따라서, 'Resting_SaO2', 'FEV1pp_utah', 'FVCpp_utah', 'FEV1_FVC_utah', 'Age', 'HR', 'BMI', 'Weight_KG', 'sysBP' 변수들은 모두 p-value 가 0.05 미만으로, 'Severity_Group'과 통계적으로 유의미한 관계가 있다고 볼 수 있습니다. 반면, 'Height_CM' 변수는 p-value 가 0.05 이상으로, 'Severity_Group'과 유의미한 관계가 없다고 판단되었습니다.

결론적으로, COVID-19 의 중증도와 강한 연관성이 있는 변수는 'Resting_SaO2', 'FEV1pp_utah', 'FVCpp_utah', 'FEV1_FVC_utah', 'Age', 'HR', 'BMI', 'Weight_KG', 'sysBP'입니다. 이 변수들을 사용하여 중증도를 예측하는 모델을 만드는 것이 유의미할 것으로 보입니다. 이와 같은 방법으로, 우리는 통계적 기법을 사용하여 COVID-19 의 중증도와 연관성이 있는 변수를 식별할 수 있습니다.

-At the end- Good job!