고려대학교
KOREA UNIVERSITY

**Probability and Statistics, Final Exam**

**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)** Please compress (.zip) 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.)
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 5 분전인 16 시 25 분에는 블랙보드에 제출을 시작하길 권장합니다. 16 시 30 분에는 제출할 수 없도록 닫히며, 제출을 못했을 시 당연 0 점 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 TA 가 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 10 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 10 개의 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 "*Q1_Data.tsv*" file. The number of confirmed cases of COVID-19 from 18 independent institutions was collected (*numCOVID* variable). It is also investigated whether each institution is a statistician group or a computer science group (*Job* variable).
(Kor: "Q1_Data.tsv" 파일을 로드하십시오. 이 자료에는 18 개의 독립적인 기관에서 발생한 COVID-19 확진자 수가 numCOVID 변수에 측정되어 있습니다. 또한, 각 기관이 통계학자 그룹인지 전산학자 그룹인지에 대한 정보가 Job 변수에 측정되어 있습니다.)

**(Q1)** Please perform hypothesis testing on **whether occupations have different numbers of COVID-19 cases** using **one parametric** and **one non-parametric** approaches at **5% significance level.** Everything from clear hypothesis setting to conclusion must be done statistically neatly. **(10 Points).**
(Kor: 유의수준 5%에서 직업 종류에 따라 코로나 확진자 수가 다르다는 가설검정을 하나의 모수적 방법과 하나의 비 모수적 방법으로 수행하십시오. 명확한 가설 설정 부터 결론까지 통계적으로 깔끔하게 이루어져야 합니다.)

**[Answer]**

'Job' 변수는 범주형 변수이고 'numCOVID'는 연속형 변수이므로, 모수적 방법으로는 독립 표본 t-검정, 비모수적 방법으로는 Mann-Whitney U 검정을 사용할 수 있습니다.

1. 모수적 방법: 독립 표본 t-검정
2. 비모수적 방법: Wilcoxon 검정

p-value 가 0.05 미만일 경우, 'Job'에 따라 'numCOVID'에 통계적으로 유의미한 차이가 있다고 판단할 수 있습니다.

t 검정 결과를 해석하면 다음과 같습니다: 모수

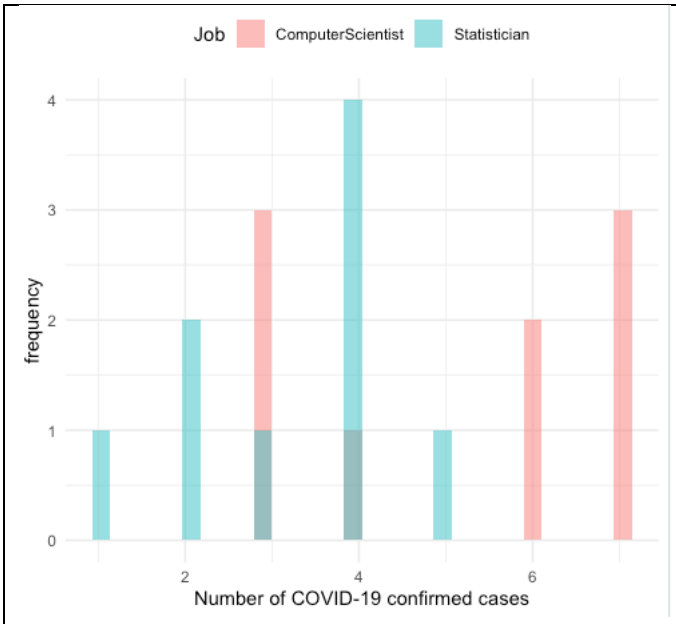
1. t 값은 2.5202 이고, 자유도는 14.432 입니다. 이는 두 직업 그룹 사이의 평균 차이가 표준오차의 2.5202 배라는 것을 의미합니다.
2. p-value 는 0.02406 으로, 0.05 보다 작습니다. 따라서 유의수준 5%에서 귀무가설을 기각하고, 통계학자 그룹과 전산학자 그룹의 COVID-19 확진자 수는 통계적으로 유의미하게 다르다는 결론을 내릴 수 있습니다.
3. 신뢰구간(Confidence Interval)은 0.2859005 부터 3.4918773 까지입니다. 이는 두 직업 그룹의 COVID-19 확진자 수 평균 차이가 이 구간 안에 있을 것이라는 것을 95% 확신한다는 의미입니다.
4. 표본 평균(Sample estimates)을 보면, 'ComputerScientist' 그룹의 평균 확진자 수는 5.111111 이고, 'Statistician' 그룹의 평균 확진자 수는 3.222222 입니다. 이는 'ComputerScientist' 그룹의 평균 확진자 수가 'Statistician' 그룹보다 더 많다는 것을 보여줍니다.

Wilcoxon 순위합 검정 결과를 해석하면 다음과 같습니다: 비모수

1. W 값은 61.5 입니다. 이는 두 그룹 간의 순위합 차이를 나타냅니다.
2. p-value 는 0.06507 로, 0.05 보다 큽니다. 따라서 유의수준 5%에서 귀무가설을 기각할 수 없습니다. 이는 통계학자 그룹과 전산학자 그룹의 COVID-19 확진자 수가 통계적으로 유의미하게 다르다고 볼 수 없다는 것을 의미합니다.

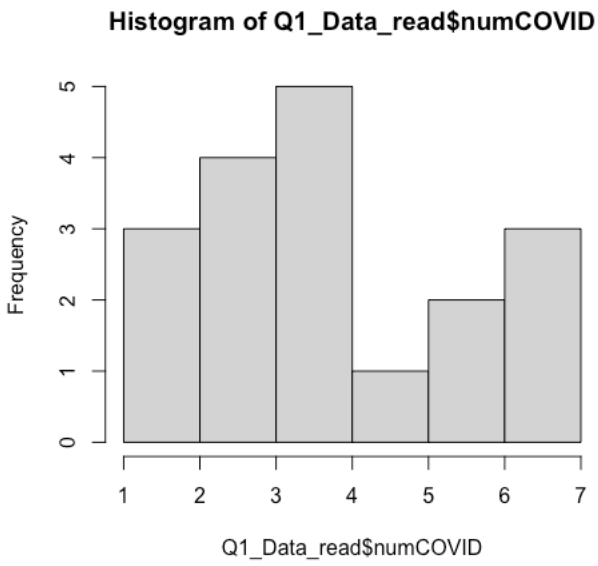
**(Q2)** To investigate if the parametric method performed in (Q1) is valid visually, please examine the distribution indirectly by **visualizing a histogram for the number of numCOVID for each occupation**.
(Kor: (Q1)에서 수행한 모수적 방법이 시각적으로 타당한지 알아보기 위해 직업별 코로나 19 감염자수 히스토그램을 시각화하여 간접적으로 분포를 살펴보시기 바랍니다.)

**[Answer & Plot]**




**(Q3)** Among the hypothesis tests performed in (Q1), **which hypothesis test is preferable? Why?** **(10 Points).**
(Kor: (Q1)에서 수행한 가설 검정 중 어떤 가설검정을 수행하는 것이 바람직합니까? 그 이유는 무엇입니까?)

**[Answer]**




직업별 코로나 19 감염자 수에 대한 가설검정을 선택할 때, 데이터의 특성을 고려해야 합니다.

1. 모수적 방법(t-검정)을 사용할 때는 데이터가 정규분포를 따르는지 확인해야 합니다. 만약 각 직업별 코로나 19 감염자 수가 정규분포를 따른다면, t-검정이 바람직합니다.
2. 비모수적 방법(Mann-Whitney U 검정)은 데이터의 분포에 대한 가정이 필요하지 않습니다. 따라서 데이터가 정규분포를 따르지 않거나, 데이터의 분포를 알 수 없는 경우에 비모수적 방법을 사용하는 것이 바람직합니다.

즉 현재 데이터가 정규분포를 따르지 않기 때문에 현재 예제에서는 비모수적 방법이 더욱 바람직하다고 볼 수 있습니다.

**(Q4)** In "*Q4_Data.R*", there is a code for a hypothesis test method called "Seo-test" developed by Prof. Minseok Seo (***Different from the version in HW#3). The current code is configured to work when **"Q1_Data.tsv" is loaded as a variable called "data"**. Please run the corresponding code and **test the hypothesis at the 10% significance level**. Everything from clear hypothesis setting to conclusion must be done statistically neatly. **(10 Points).**
(Kor: "Q3_Data.R"에는 제가 개발한 "Seo-test"라는 가설 검정 방법에 대한 코드가 들어 있습니다 (***과제 3 의 방법과 다릅니다). 현재 코드는 "Q1_Data.tsv"가 "data"라는 변수로 로드될 때 바로 작동할 수 있도록 구성되어 있습니다. 해당 코드를 적용하여 10% 유의 수준에서 가설 검정 하십시오. 명확한 가설 설정 부터 결론까지 통계적으로 깔끔하게 이루어져야 합니다.)

**[Answer]**

먼저 가설을 설정합니다:

- 귀무가설(H0): 통계학자와 컴퓨터 과학자의 코로나 감염자 수에는 차이가 없다.
- 대립가설(H1): 통계학자와 컴퓨터 과학자의 코로나 감염자 수에는 차이가 있다.

그런 다음 우리가 얻은 p-value 는 0.085 로, 10%의 유의 수준보다 작습니다.

- 따라서, 10%의 유의 수준에서 귀무가설을 기각하고 대립가설을 받아들입니다. 이는 통계학자와 컴퓨터 과학자의 코로나 감염자 수에는 통계적으로 유의미한 차이가 있다는 것을 나타냅니다.

**(Q5) What are the limits of the "*Seo-test statistic*"** that can be used to test for differences between two groups? **Compare it to the format of the statistic you already know. (10 Points).**
(Kor: 두 그룹 간 차이를 테스트하는 데 사용할 수 있는 "Seo-검정통계량의" 한계는 무엇입니까? 이미 알고 있는 통계량의 수식과 비교하여 서술하십시오.)

**[Answer]**

"Seo-검정통계량"은 두 그룹의 중앙값의 차이를 계산하는 비모수 검정 통계량입니다. 이는 두 그룹의 데이터 분포가 심각하게 왜곡되어 있거나, 데이터가 작아서 정규 분포를 따르는지 확인할 수 없는 경우에 유용합니다.

그러나 이 검정통계량은 몇 가지 한계를 가지고 있습니다:

1. 비모수 검정은 데이터의 순위에만 의존하기 때문에, 원래 데이터의 크기나 변동성에 대한 정보를 완전히

활용하지 못합니다. 따라서, 비모수 검정은 모수 검정에 비해 통계적 효과가 약할 수 있습니다.

2. "Seo-검정통계량"은 중앙값의 차이만을 고려합니다. 따라서 두 그룹 사이에 평균이나 분산 등 다른 통계적 특성에 차이가 있을 경우 이를 감지하지 못할 수 있습니다.

3. 또한, "Seo-검정통계량"은 두 그룹의 데이터 분포가 같다는 가정을 내포하고 있습니다. 이 가정이 만족되지 않는 경우, 검정 결과의 신뢰성이 떨어질 수 있습니다.

이와 비교하여, t-검정 같은 모수 검정은 두 그룹의 평균 차이를 비교합니다. 이는 데이터 분포가 정규분포를 따르는 경우에 통계적으로 더 강력하고 정보를 더 많이 활용할 수 있다는 장점이 있습니다. 하지만, 이 가정이 만족되지 않는 경우에는 비모수 검정 방법을 사용하는 것이 바람직합니다.

1. "Seo-검정통계량":
    - 사용되는 통계량: 두 그룹의 중앙값의 차이
    - 수식: |median(Group1) - median(Group2)|

   "Seo-검정통계량"은 두 그룹의 중앙값 차이를 절대값으로 취하여 계산합니다. 이 방법은 두 그룹의 분포가 심각하게 왜곡되어 있거나, 데이터가 작아서 정규 분포를 따르는지 확인할 수 없는 경우에 유용합니다.
2. t-검정:
    - 사용되는 통계량: 두 그룹의 평균 차이
    - 수식: (mean(Group1) - mean(Group2)) / SE

   여기서 SE 는 두 그룹의 표준 오차(Standard Error)를 의미합니다. t-검정은 두 그룹의 평균 차이를 표준 오차로 나누어 표준화합니다. 이 방법은 두 그룹의 데이터 분포가 정규 분포를 따르는 것으로 가정하며, 이 가정이 만족될 경우 통계적으로 더 강력한 검정력을 가집니다.

**[Q6 – Q9]** Load "*Q6_Data.tsv*" file. From a total of 72 people, their disease information (*Disease* variable), severity information (Severity variable), height information (*Height* variable), DNA information (DNA_1,

DNA_2, …, DNA_20 variables), and 1,724 RNA information (Gene_1, Gene_2, …, Gene_1724 variables) were investigated.
(*Tip: All genetic data can be considered as categorical random variable and all genomic data can be considered as continuous random variables.)
(Kor: "Q6_Data.tsv" 파일을 불러 오십시오. 이 자료에는 총 72 명의 질병정보(Disease 변수), 암의 중증도 (Severity 변수), 키 정보(Height 변수), DNA 정보 20 개 (DNA_1,…, DNA_20 변수들), RNA 정보 1,724 개(Gene_1, Gene_2, …, Gene_1724 변수들)이 조사되어 있습니다. )
(*Tip: 모든 유전 데이터 (DNA 데이터)는 범주형 확률변수로, 모든 RNA 데이터는 연속 확률변수로 간주할 수 있습니다.)

**(Q6)** Please find DNA & RNA markers (variables) that differ between cancer and normal groups at the 1% significance level. These are techniques used in 2000 in panels that can actually diagnose cancer early. Additionally, this skill is an essential skill for artificial intelligence, which you will learn later. Please describe how many genes are found. **(10 Points).**
(Kor: 암그룹과 정상그룹 사이에 차이가 나는 DNA, RNA 마커(변수)를 1% 유의수준에서 찾아보세요. 이것은 실제로 암을 조기에 진단할 수 있는 패널에서 2000 년에 사용된 기술입니다. 또한 이 기술은 추후에 여러분이 배울 인공지능에 있어서 필수 기술입니다. 몇 개의 유전자가 찾아지는지 서술해 주세요)

**[Answer]**

318

**(Q7)** Please use hypothesis testing at the 1% significance level to check DNA & RNA markers (variables) related to the severity of cancer. Please explain how many genes were discovered? **(10 Points).**
(Kor: 1% 유의수준에서 가설검증을 이용하여 암의 중증도와 관련된 DNA & RNA 마커(변수)를 확인해주세요. 얼마나 많은 유전자가 발견되었는지 기술하십시오.)

**[Answer]**

21

**(Q8)** Please identify DNA & RNA markers (variables) related to the height using hypothesis testing at 1% significance level. Please describe how many genes are found **(10 Points).**
(Kor: 1% 유의수준에서 가설검정을 이용하여 키와 관련된 DNA 및 RNA 마커들을 확인하세요. 얼마나 많은 유전자가 발견되었는지 기술하십시오.)

**[Answer]**

8

**(Q9)** Perform visualizations to support the results of Q6, Q7, and Q8, and draw overall conclusions. **(10 Points).**
(Kor: Q6, Q7, Q8 의 결과를 뒷받침할 수 있는 시각화를 수행하고, 종합적인 결론을 내리십시오.)

**[Answer]**

**(Q10)** Multiple hypothesis tests on the same data raise certain statistical issues due to their "significance level" concept. This part was not covered in class, but it is one of the most important statistical issues in the era of modern data science. I believe that you can understand this concept by studying it a little if you have followed the course normally. After studying this issue through Googling, **please describe in (Q6-Q8) what is causing the problem in your project.** Also, if possible, look for a solution, then **modify the P-value from your project (Q6-Q8) and describe it.** (*Tip: There are so many correction methods such as Bonferroni, Scheffe, Sidak, and Benjamini & Hochberg methods). **(Bonus 20 points).**
(동일 자료에 대한 복수의 가설 검정은 "유의수준" 컨셉에 의해 특정 통계적 문제를 야기합니다. 이에 대해서는 수업시간에 다루지 않았지만, 현대 데이터 사이언스 시대에서 사용되는 통계적 분석에서 가장 중요한 이슈 중 하나입니다. 본 수업을 정상적으로 따라왔다면, 약간의 구글링을 통해 이를 쉽게 이해할 수 있을 것이라 믿습니다. **구글링을 통해 이 이슈를 공부한 뒤, 여러분이 수행한 (Q6-Q8)에서 어떤점이 이런 문제를 야기시키는지에 대해서 서술**하십시오. 또한, 가능하다면 이를 해결하는 **다양한 방법을 이용해서 여러분이 수행한 (Q6-Q8)의 결과를 수정한 뒤 서술**하십시오. 팁을 주자면, 상당히 많은 솔루션이 존재하며 R 로 구현되어 있습니다. 예: Bonferroni, Scheffe, Sidak, and Benjamini & Hochberg 방법).

**[Answer]** **No page limit.**

본문에서 언급하신 문제는 "다중 비교 문제" 또는 "다중 가설 검정 문제"라고 합니다. 이 문제는 여러 개의 가설을 동시에 검정할 때 발생하며, 특히 각 가설에 대해 별도의 검정을 수행하면서 발생하는 유형 I 오류의 확률을 제어하지 않을 때 발생합니다. 유형 I 오류는 귀무 가설이 참일 때, 귀무 가설을 잘못 기각하는 경우를 말합니다.

Q6-Q8 의 분석에서는 각각의 DNA 와 RNA 마커에 대해 별도의 가설 검정을 수행했기 때문에 이 문제가 발생합니다. 즉, 각 마커가 암의 중증도와 독립적이라는 귀무 가설을 검정했는데, 모든 마커에 대해 동일한 유의수준(0.01)을 적용했기 때문에, 여러 번의 검정을 반복함에 따라 실제로는 중요하지 않은 마커를 중요하다고 잘못 판단할 확률이 증가합니다.

이 문제를 해결하기 위한 방법으로는 본페로니 교정, Scheffe 교정, Sidak 교정, 벤자미니-호크베르그 방법 등이 있습니다. 본페로니 교정은 가장 보수적인 방법으로, 유의수준을 검정의 수로 나누어 교정합니다. 다른 방법들은 보다 복잡하지만, 보다 많은 가설을 기각하면서도 전체적인 유형 I 오류의 확률을 제어합니다.

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