**Probability and Statistics, Final Exam**

**Your ID#: [ 2021270682 ]**

**Your Name: [ Dounguk Kim ]**

| **[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 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 10개의 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 18개의 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]**  **Alternative Hypothesis 대립가설**  H1: 직업종류에 따라 Statistician과 ComputerScientist의 코로나 확진자 수가 다르다  **Null Hypothesis 귀무가설**  H0: 직업종류에 따라 Statistician과 ComputerScientist의 코로나 확자 수가 다르지 않다  모수적 방법으로 t.test, 비모수적 방법으로 wilcoxon 검정을 하였다. t.test의 결과 0.02406<0.05로 유의미합니다. 따라서 5% 유의 수준 하에서 Null hypothesis는 reject되고 Alternative hypothesis가 select됩니다.  이는 직업에 따른 코로나 확진자수의 관계가 통계적으로 유의미하다는 것을 나타냅니다.  하지만 wilcoxon검정 결과 p값이 0.06507으로 0.05보다 크므로 5%유의수준 하에서는 Null Hypothesis가 select 되고, 10% 유의수준 하에서는 Alternative Hypothesis가 select됩니다.  이는 직업 따른 코로나 확진자수의 관계가 통계적으로 유의미하지 않다는 것을 나타냅니다. |
| --- |

**(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 코로나19 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]**  **wilcoxon 검정을 사용하는 것이 더욱 바람직 합니다. 왜냐하면**  표본이 충분하지 않기 때문에 표본이 정규분포를 따르지 않을 수 있고 따라서 비모수적 방법인 wilcoxon 방법의 결과를 더 신뢰할 수 있다. |
| --- |

**(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 (\*\*\*과제3의 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]**  **Alternative Hypothesis 대립가설**  H1: 두 직업군 사이의 COVID-19 감염자 수 중앙값에 차이가 있다.  **Null Hypothesis 귀무가설**  H0: 두 직업군 사이의 COVID-19 감염자 수 중앙값에 차이가 없다.  Seo-test를 이용한 테스트 결과 0.088<0.10로 유의미합니다. 따라서 10% 유의 수준 하에서 Null hypothesis는 reject되고 Alternative hypothesis가 select됩니다.  이는 두 직업군 사이에 따른 COVID-19 감염자 수 중앙값에 차이가 있다. 입니다. |
| --- |

**(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-검정통계량은 적은량의 표본을 통한 시뮬레이션의 결과로 1000개의 랜덤 그룹을 만들어 하는 방법이기 때문에 비모수적 방법이고 따라서 표본의 크기가 크면 클 수록 정규분포를 따르는 모수적 방법의 t.test에 비해 정확성이 떨어질 수 있습니다. |
| --- |

**[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: "Q5\_Data.tsv" 파일을 불러 오십시오. 이 자료에는 총 72명의 질병정보(Disease 변수), 암의 중증도 (Severity 변수), 키 정보(Height 변수), DNA 정보 20개 (DNA\_1,…, DNA\_20 변수들), RNA 정보 1,724개(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 2000년에 사용된 기술입니다. 또한 이 기술은 추후에 여러분이 배울 인공지능에 있어서 필수 기술입니다. 몇 개의 유전자가 찾아지는지 서술해 주세요)**

| **[Answer]**  테스트 결과 1% 유의수준에서 유의미한 마커의 개수는 315개 입니다. |
| --- |

**(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]**  테스트 결과 1% 유의수준에서 유의미한 마커의 개수는 0개 입니다. |
| --- |

**(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]**  테스트 결과 1% 유의수준에서 유의미한 마커의 개수는 1059개 입니다. |
| --- |

**(Q9)** Perform visualizations to support the results of Q6, Q7, and Q8, and draw overall conclusions. **(10 Points).**

**(Kor: Q6, Q7, Q8의 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 R로 구현되어 있습니다. 예: Bonferroni, Scheffe, Sidak, and Benjamini & Hochberg 방법).

| **[Answer] No page limit.**  이 문제는 통계학에서 '다중비교 문제' 또는 '다중검정 문제'라고 알려져 있습니다. 동일한 데이터 세트에 대해 여러 가설 검정을 수행하면, 1종 오류(false positive)의 확률이 증가합니다. 이는 실제로는 차이가 없는데도 불구하고, 우연히 차이가 있다고 판단하는 오류입니다. 이 문제는 통계적 가설검정에서 유의수준을 설정함으로써 일어나는 문제로, 유의수준을 p-value로 해석하면, 그것은 우리가 우연히 1종 오류를 범할 확률을 의미합니다. 따라서 여러 번의 가설 검정을 수행하게 되면, 이 오류를 범할 확률이 누적되어 전체적으로 1종 오류를 범할 확률이 증가하게 됩니다. 이 문제를 해결하기 위한 방법으로는 여러 가지가 있습니다.   1. 보네페로니(Bonferroni) 보정: 가장 간단하고 보수적인 방법으로, 유의수준을 검정의 수로 나누어 보정합니다. 이 방법은 Type I 오류를 통제하는 데 효과적이지만, 검정력을 크게 감소시킬 수 있습니다. 2. 벤자미니-호크베르그(Benjamini-Hochberg) 방법: FDR(False Discovery Rate)을 통제하는 방법으로, 보네페로니 방법보다 덜 보수적이며 검정력을 더 잘 유지합니다. 3. 시닥(Sidak) 방법, 셰페(Scheffe) 방법 등도 있습니다. 이들 방법은 각각 특정 상황에 최적화된 방법들입니다. |
| --- |

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