**Homework #3 (Probability & Statistics)**

**[2019270664 윤재현]**

|  |
| --- |
| **Rationale of the Homework #3**  Finding meaningful variables freely from given data is no longer the exclusive property of the Department of Statistics. If you are a computer major student, you should be able to find meaningful information based on statistical theory more proficiently on larger data. |

|  |
| --- |
| **[Instruction]**  **(1)** Once you **download** the compressed file from Blackboard and release it, **there will be “Data1.txt” and “Data2.tsv” files.**  **(2)** Try to resolve the problems below in this state.  \* You can consult with a friend, but copying the R code is absolutely prohibited (It will be shortcut for the F grade).  **(3)** Please fill in the blanks on the bottom report. **Compress and submit** **this report (.docx), and the entire R-code (.R file)** you wrote.  **\*\*\*The R-code should be executable when the TA runs.**  **\*\*\*The submitted compressed file (.zip) must be named HW3\_YourID\_Yourname.zip.**  Due date: 2022-12-09 11:59PM |

**(Q1 ~ Q4)** based on the **“*Data1.txt*”**.

**[Q1]** Please divide the *Severity\_Group* random variable into a binary random variable containing the {"Healthy", "COVID-19"} and a random variable containing the severity information of COVID-19 {0,1,2,3,4}.

\*\*Tip: If you search for "String split in R", a lot of web page will be searched on Google. Don't forget that data preprocessing requires computer students to do better!

|  |
| --- |
| **[Outcome & Code]**  **Code:**  n <- length(data1$Severity\_Group)  data1$Severity\_Group <- replace(data1$Severity\_Group, data1$Severity\_Group == "Healthy", 0)  for(i in 1:n) {  if(data1$Severity\_Group[i] != "0") {  x <- strsplit(data1$Severity\_Group[i], "COVID19\_Phase")[[1]][2]  data1$Severity\_Group[i] <- x  }  }  data1$Severity\_Group <- as.numeric(data1$Severity\_Group)  Outcome:  [1] 0 2 2 3 2 3 2 2 3 1 3 2 0 4 1 2 2 3 1 3 2 2 2 3 0 4 2 4 4 3 3 1 0 3 4 2 2  [38] 2 1 3 4 3 2 2 2 0 2 2 2 3 0 3 3 3 1 3 3 0 0 0 0 2 4 3 2 4 4 2 2 1 0 0 2 0  [75] 2 2 2 3 4 2 0 0 2 3 2 0 0 0 1 0 2 2 0 2 0 1 0 3 4 3 4 2 2 3 3 0 2 1 1 0 2  [112] 3 3 4 2 3 0 3 2 4 0 3 0 0 4 0 4 0 2 0 0 0 0 0 0 2 3 2 0 4 0 4 0 3 4 4 4 2  [149] 0 0 0 1 0 0 0 3 0 0 0 0 2 0 0 1 1 2 4 1 2 2 0 3 0 0 2 0 3 2 3 0 3 3 1 1 0  [186] 3 1 1 3 1 3 2 0 3 0 1 3 2 0 0 2 0 2 0 0 0 0 0 0 0 0 3 4 2 3 2 2 0 3 0 0 3  [223] 3 1 1 2 0 2 4 0 0 0 0 0 0 2 0 0 0 0 2 0 0 2 0 0 0 0 0 2 3 4 3 4 0 2 0 1 0  [260] 2 3 0 0 0 3 0 3 0 2 0 2 0 4 2 3 0 3 0 0 2 0 0 2 2 2 0 0 0 2 2 2 2 0 4 0 0  [297] 3 0 3 4 2 0 4 4 3 4 0 4 0 0 0 0 2 0 2 0 0 2 0 1 4 0 2 3 3 2 3 0 1 2 0 2 0  [334] 3 2 0 2 0 1 4 0 2 0 3 1 0 3 1 3 2 3 0 0 0 3 1 0 3 2 1 0 3 0 0 0 0 2 2 2 0  [371] 0 2 2 2 2 0 1 1 0 2 2 0 4 2 3 1 3 3 0 0 1 0 3 0 0 4 4 0 0 2 3 2 0 0 1 0 2  [408] 0 2 0 4 2 0 0 3 2 2 0 3 2 4 2 2 2 4 2 4 0 4 2 4 1 1 0 0 0 2 3 2 2 3 1 3 3  [445] 0 0 2 2 2 4 3 0 2 0 0 0 3 4 3 4 3 3 0 4 0 2 2 2 3 2 0 2 1 0 3 4 3 4 3 2 3  [482] 2 2 0 0 4 0 2 0 0 2 3 2 3 4 2 0 2 3 2 2 0 0 0 3 3 0 1 4 3 4 0 0 3 2 2 2 2  [519] 0 2 4 1 0 2 3 2 2 0 2 0 0 3 0 0 0 2 3 1 4 3 2 0 1 1 2 1 0 0 0 0 0 0 0 2 0  [556] 0 1 0 0 3 4 2 2 4 0 3 0 2 2 4 2 2 0 1 0 0 0 0 3 3 0 4 0 0 1 3 1 0 0 0 2 3  [593] 1 2 0 4 0 3 1 2 0 4 1 3 2 3 0 2 3 2 3 0 1 3 2 0 2 2 0 3 4 0 0 4 0 0 3 0 4  [630] 2 1 2 3 3 2 0 0 0 2 0 3 2 4 3 3 0 4 1 2 2 2 2 4 3 0 0 1 0 3 1 0 0 0 4 0 3  [667] 2 2 2 2 2 0 3 3 2 0 0 0 0 4 4 0 1 0 4 0 0 2 1 0 0 0 4 1 4 0 0 0 0 0 0 0 2  [704] 0 0 3 3 0 2 2 0 0 1 0 0 4 2 3 4 3 3 2 0 3 0 0 0 0 2 4 2 1 1 4 0 3 3 0 0 0  [741] 0 2 0 1 0 2 2 0 0 3 1 4 0 2 4 1 0 0 2 0 0 0 0 3 0 0 0 0 1 0 0 3 2 2 1 0 0  [778] 0 0 0 0 3 3 3 4 0 2 2 2 0 2 4 4 2 4 0 2 3 0 2 0 0 4 0 0 0 0 3 3 2 3 3 0 0  [815] 3 0 4 0 2 3 0 0 0 2 0 2 0 1 2 1 0 1 2 2 4 0 0 2 2 0 2 1 1 0 2 0 2 0 4 0 0  [852] 3 0 4 0 0 2 3 0 1 0 1 2 2 0 3 3 2 3 4 2 2 2 4 3 4 2 1 3 2 2 3 2 0 2 0 0 1  [889] 0 3 1 2 2 0 4 2 0 0 0 0 2 0 0 0 3 0 0 2 3 0 3 0 3 2 4 1 1 3 0 0 0 0 4 1 1  [926] 3 0 4 2 2 3 3 3 1 4 3 4 0 2 0 0 0 3 0 3 0 3 4 0 4 0 0 0 2 1 3 2 3 2 0 0 0  [963] 0 2 0 0 3 2 0 0 2 0 3 2 0 2 0 1 0 3 3 4 3 4 4 0 1 3 3 0 0 0 0 3 2 0 3 2 0  [1000] 0  [ reached getOption("max.print") -- omitted 8021 entries ] |

**[Q2]** A total of 15 meaningful variables should have been created through [Q1]. Using these variables, **please search for significant variables related to the Height** variable using the parametric methods at 5% significance level.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **[Answer]** \*\*\*One page limit. You must prepare a basis according to the hypothesis testing principle, and organize it (Table is one of the methods) to arrive at a conclusion.  Height와 다른 random variable과의 연관성을 찾기 위해 Gender, Race, HaveCough는 t-test를 시행하고, Severity\_Group은 ANOVA test를 시행했고, 나머지는 Pearson’s correlation test를 시행했다.  two group test의 가설:  H1 = Gender(or Race, or HaveCough)에서 두 그룹은 Height로 구별된다.  H0 = Gender(or Race, or HaveCough)에서 두 그룹은 Height로 구별되지 않는다.  Severity\_Group test의 가설:  H1 = COVID-19의 민감성은 Height에 따라 차이를 보인다.  H0 = COVID-10의 민감성은 Height와는 관계가 없다.  Pearson’s correlation test의 가설:  H1 = Height와 주어진 변수는 상관 관계가 있다.  H0 = Height와 주어진 변수는 상관 관계가 없다.  각 가설의 검증은 다음과 같이 진행되었다.  p-value를 구한 후 그 p-value가 유의수준(0.05)보다 작은 경우 null hypothesis를 기각한 후 alternative hypothesis를 채택하였다. 반대의 경우 null hypothesis를 기각할 충분한 증거를 갖지 못했기 때문에 기각하지 않았다.   |  |  |  | | --- | --- | --- | | **Random variable** | **p-value** | **Reject / No Reject** | | Age | 3.604859e-21 | Reject | | Gender | 0 < 2.2e-16 | Reject | | Race | 9.438644e-08 | Reject | | Weight | 0 < 2.2e-16 | Reject | | sysBP | 1.554121e-06 | Reject | | HR | 4.271295e-10 | Reject | | Resting\_SaO2 | 0.001010738 | Reject | | BMI | 8.935233e-06 | Reject | | HaveCough | 0.007682601 | Reject | | FEV1pp\_utah | 0.8084664 | No Reject | | FVCpp\_utah | 0.0134378 | Reject | | FEV1\_FVC\_utah | 0.2303211 | No Reject | | Severity\_Group | 0.07215079 | No Reject |   즉 Height과 연관이 있는 signficant variable은 Age, Gender, Race, Weight, sysBP, HR, Resting\_SaO2, BMI, HaveCough, FVCpp\_utah이다. |

**[Q3]** Please perform the same analysis as [Q2] with a **nonparametric test** and compare the results.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **[Answer]** \*\*\*One page limit. You must prepare a basis according to the hypothesis testing principle, and organize it (Table is one of the methods) to arrive at a conclusion.  Q2와 동일한 가설을 사용하고, 가설 검증을 위한 method만 바꿔 시행하였다.  two group test는 Wilcoxon test를, multi gorup test는 Kruskal-Wallis test를, pearson’s correlation test는 Spearman correlation test로 시행하였다.  각 가설의 검증은 다음과 같이 진행되었다.  p-value를 구한 후 그 p-value가 유의수준(0.05)보다 작은 경우 null hypothesis를 기각한 후 alternative hypothesis를 채택하였다. 반대의 경우 null hypothesis를 기각할 충분한 증거를 갖지 못했기 때문에 기각하지 않았다.   |  |  |  |  | | --- | --- | --- | --- | | **Random variable** | **p-value**  **(non-parametric)** | **Reject /**  **No Reject** | **p-value**  **(parametric)** | | Age | 2.357553e-18 | Reject | 3.604859e-21 | | Gender | 0 < 2.2e-16 | Reject | 0 < 2.2e-16 | | Race | 1.007774e-06 | Reject | 9.438644e-08 | | Weight | 0 < 2.2e-16 | Reject | 0 < 2.2e-16 | | sysBP | 2.997111e-09 | Reject | 1.554121e-06 | | HR | 8.780584e-11 | Reject | 4.271295e-10 | | Resting\_SaO2 | 0.3357285 | No Reject | 0.001010738 | | BMI | 0.06309004 | No Reject | 8.935233e-06 | | HaveCough | 0.008430961 | Reject | 0.007682601 | | FEV1pp\_utah | 0.8388484 | No Reject | 0.8084664 | | FVCpp\_utah | 0.05520026 | No Reject | 0.0134378 | | FEV1\_FVC\_utah | 0.8074129 | No Reject | 0.2303211 | | Severity\_Group | 0.4887421 | No Reject | 0.07215079 |   먼저 nonprametric test에서 Height와 연관이 있는 Random variable은 Age, Gender, Race, Weight, sysBP, HR, HaveCough이다. 하지만 일반적으로 parametric method가 nonprametric method보다 statistic power가 크기 때문에 parameteric method를 이용한 결과를 채택한다.  두 방법을 이용해서 나온 p-value를 보면, sysBP, HR의 경우를 제외한 random variable의 p-value는 모두 parametric method의 결과가 더 작았다. 하지만 sysBP, HR의 경우 nonparametric method의 p-value가 더 작게 나왔다. 이러한 결과를 보아 sysBP, HR이 normal distribution을 따르지 않는다고 생각할 수 있다. |

**[Q4]** **Please find out which variables have a close association with the severity of COVID-19** through a hypothesis test.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **[Answer]** \*\*\*One page limit. You must prepare a basis according to the hypothesis testing principle, and organize it (Table is one of the methods) to arrive at a conclusion.  Severity of COVID-19는 multiple group을 가지고 있는 random variable이다. 이에 따라 Gender, Race, HaveCough와의 연관성 test는 chi-squared test를 시행한다. 3가지 random variable을 제외한 다른 경우는 multi group test인 ANOVA test를 시행한다.  ch-squared test의 가설:  H1: COVID-19의 민감도가 Gender(or Race, or HaveCough)에 따라 달라진다.  H0: COVID-19의 민감도가 Gender(or Race, or HaveCough)과 관계가 없다.  multi-group test의 가설:  H1: COVID-19의 민감도는 continuous random variable에 따라 차이를 보인다.  H0: COVID-19의 민감도는 continuous random variable과 관계가 없다.  각 가설의 검증은 다음과 같이 진행되었다.   |  |  |  |  | | --- | --- | --- | --- | | **Random variable** | **statistic** | **p-value** | **association level** | | Age | f = 1131.811 | 7.447016e-234 | 5 | | Gender | X-squared = 20.007 | 0.0004979167 | 11 | | Race | X-squared = 329.33 | 5.078052e-70 | 7 | | Height | f = 3.234158 | 0.07215079 | 13 | | Weight | f = 87.01543 | 1.332812e-20 | 10 | | sysBP | f = 6.124297 | 0.01335188 | 12 | | HR | f = 372.321 | 2.511091e-81 | 6 | | Resting\_SaO2 | f = 2120.405 | 0 < 2.2e-16 | 4 | | BMI | f = 91.76304 | 1.238162e-21 | 9 | | HaveCough | X-squared = 248.81 | 1.176717e-52 | 8 | | FEV1pp\_utah | f = 51312.91 | 0 < 2.2e-16 | 2 | | FVCpp\_utah | f = 6526.356 | 0 < 2.2e-16 | 1 | | FEV1\_FVC\_utah | f = 46461 | 0 < 2.2e-16 | 3 |   p-value가 작을 수록 null hypothesis를 기각할 확률이 높다. 즉, p-value가 낮을 수록 그 random variable과 COVID-19 민감도와 관계가 밀접하다는 것을 의미한다. 이에 근거해association level은 p-value로 먼저 판단한 후 R에서 0으로 나온 값은 statistic값으로 판단하였다.  상위 5 level의 random variable은 FVCpp\_utah, FEV1pp\_utah, FEV1\_FVC\_utah, Resting\_SaO2, Age이다. 즉, 이 5개의 random variable이 COVID-19 민감도와 가장 밀접한 관계를 가지고 있다. |

**(Q5 ~ Q7)** based on the **“*Data2.tsv*”**.

Load “*Q5\_Data.tsv*” file. From a total of 72 people, their disease information (*Disease* variable), height information (*Height* variable), and 1,724 genomic information (Gene\_1, Gene\_2, …, Gene\_1724 variables) were investigated. All genomic data are continuous random variables.

**[Q5]** Please identify the genes that differ between the cancer-group and the normal-group using a hypothesis test at 1% significance level. Please describe how many genes are found. Also, please state clearly which statistical methods were used.

|  |
| --- |
| **[Answer]** \*\*\*One page limit.  암에 걸린 사람들과 건강한 사람들 사이에서 차이를 보이는 유전자를 찾아야하기 때문에 two group test를 시행한다. 이때 조사된 데이터를 보면 총 72명 중 암에 걸린 사람들은 25명, 건강한 사람들은 수는 47명이다. 암에 걸린 사람들의 수가 CLT를 만족하지 않기 때문에 parametric method가 아닌 nonparametric method를 사용한다. Wilcoxon test를 시행한다. 가설검증을 시행할 때 사용한 가설은 다음과 같다.  H1: 암에 걸린 사람들과 건강한 사람들은 (1-1724)유전자 정보에서 차이를 보인다.  H0: 암에 걸린 사람들과 건강한 사람들은 (1-1724)유전자 정보가 동일하다.  이를 이용해 각 유전자에 Wilcoxon test를 수행하고, 그 p-value가 유의수준(0.01)보다 낮은 경우 null hypothesis를 기각하고, alternative hypothesis를 채택한다. 즉, p-value가 유의수준(0.01)보다 낮은 경우 그 유전자는 암에 걸린 사람들과 건강한 사람들에서 차이를 보인다고 할 수 있다.  이러한 근거로, 1% 유의수준에서 암에 걸린 사람들과 건강한 사람들을 구별하는 유전자는 Gene\_3, Gene\_6, Gene\_7, ..., Gene\_1248, Gene\_1249, Gene\_1267등 이며 총 331개의 유전자가 검출되었다. |

**[Q6]** Please check if the top 5 genes and the bottom 5 genes found in [Q5] can actually differentiate between cancer and normal by boxplot visualization.

**\*\*Tip:** If you search "boxplot in R" or "ggpubr" on Google, there are numerous visualization example codes. Great visualization, too, is an area where computer scientists can do better!

|  |
| --- |
| **[Answer]** \*\*\*One page limit.  차이가 큰 5개(상위 5개) 유전자는 395, 391, 13, 19, 231번 유전자이고, 작은 5개(하위 5개) 유전자는 880, 1249, 914, 12, 44번 유전자이다.  상위 유전자 5개      하위 유전자 5개 |

**[Q7]** Please identify genes related to the height using hypothesis testing at 1% significance level. Please describe how many genes are found**.** Also, please state clearly which statistical methods were used.

|  |
| --- |
| **[Answer]** \*\*\*One page limit.  키와 연관되어있는 유전자를 찾아야 하기 때문에 two continuous association test를 시행한다. 이때 데이터의 수는 72임으로 CLT 조건에 따라 normal distribution을 따른다고 할 수 있다. 즉, parametric method인 Pearson’s correlation test를 시행할 수 있다. 가설검증을 시행할 때 사용한 가설은 다음과 같다.  H1: 키와 (1-1724)유전자 정보는 연관성이 있다.  H0: 키와 (1-1724)유전자 정보는 연관성이 없다.  이를 이용해 각 유전자에 Pearson’s correlation test를 수행하고, 그 p-value가 유의수준(0.01)보다 낮은 경우 null hypothesis를 기각하고, alternative hypothesis를 채택한다. 즉, p-value가 유의수준(0.01)보다 낮은 경우 그 유전자는 키와 관련된 유전자라고 생각할 수 있다.  이러한 근거로 1% 유의수준에서 키와 관련된 유전자는 77, 145, 267, 321, 382, 793, 812, 1019, 1130, 1199, 1222, 1265번 유전자이며 총 12개의 유전자가 검출되었다. |