**Homework #2 (Probability & Statistics)**

**[2021270678 # & 나강민]**

|  |
| --- |
| **Rationale of the Homework #1**  (1) Let's apply the statistical hypothesis tests you learned for real medical data via R-programming.  (2) Learn how to import data by referring to Googling or our course textbooks, such as https://rafalab.github.io/dsbook/importing-data.html. The final exam is also a practical exam, so **you must study how to import data**. |

|  |
| --- |
| **[Instruction]**  **(1)** Once you **download** the compressed file from Blackboard and release it, **there will be data.txt**  **(2)** **By placing data.txt file in your Working Directory (setwd(XXX))** and using the fread() function of the data.table package **as following figure**, you can load the data on your computer.      The following feature values ​​can be observed in the data, and these are values ​​that are actually used in clinical practice.    **(3)** 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).  **(4)** 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 HW2\_YourID\_Yourname.zip.** |

**[Q1]** It is known by the NIH of the United States that the systolic blood pressure in normal people follows the following distribution: sysBP ~ N(, =16.5). Let's **perform following hypothesis test** **under the 5% significance level** to find evidence that COVID19 patients are different from the loaded data.

|  |
| --- |
| **(1) Test statistics**  검정 통계량: 3.37622  **(2) P-value**  p-value: 0.0007348903  **(3) Decision Making at 5% significance level**  (\*) Use all of the "Null hypothesis", "Reject", "Alternative hypothesis", "Select" and "Significance level" terminology to make a statistical conclusion.  귀무 가설(H0): 코로나에 걸린 사람들의 수축기 혈압이 정상인의 혈압과 같다. ("Collected Samples are not biased")  대립 가설(H1): 코로나에 걸린 사람들의 수축기 혈압이 정상인의 혈압과 다르다. ("The sample collected will be different from the population")  검정 통계량이 약 3.38이고 p-value는 약 0.00073입니다. 유의 수준은 5%(0.05)로 설정되었습니다. p-value가 유의 수준 0.05보다 작으므로, 귀무 가설을 기각하고 대립 가설을 선택합니다. 이는 '코로나에 걸린 사람들의 수축기 혈압이 정상인의 혈압과 통계적으로 유의미하게 다르다'는 결론을 내리게 됩니다. 즉, 이 결과는 코로나에 걸린 사람들의 수축기 혈압이 정상인의 혈압과 크게 다르다는 것을 의미합니다. 이는 귀무 가설을 기각하였기 때문에, 코로나에 걸린 사람들의 수축기 혈압이 정상인의 혈압과 통계적으로 유의미한 차이가 있다는 점을 기반으로 합니다. 따라서, 통계적으로 깔끔하게 올바른 결론은 '수집된 샘플은 편향되어 있으며, 코로나에 걸린 사람들의 수축기 혈압은 정상인의 혈압과 다르다'라고 할 수 있습니다. 이는 5%의 유의 수준에서 귀무 가설을 기각하고 대립 가설을 선택한 결과입니다. |

**[Q2]** Under the same circumstances as [Step 1], please try totest statistically whether **COVID19 patients show higher systolic blood pressure than known national statistics** at 5% significance level.

\*Tip: subset() or which() function can be employed to extract COVID19 subjects.

|  |
| --- |
| **(1) Alternative Hypothesis**  대립 가설(H1): 코로나19 환자의 수축기 혈압은 알려진 국가 통계보다 높다.  **(2) Null Hypothesis**  귀무 가설(H0): 코로나19 환자의 수축기 혈압은 알려진 국가 통계와 동일하다.  **(3) Test statistics**  검정 통계량: 3.37622  **(4) P-value**  p-value: 0.0003674451  **(5) Decision Making at 5% significance level**  (\*) Use all of the "Null hypothesis", "Reject", "Alternative hypothesis", "Select" and "Significance level" terminology to make a statistical conclusion.  p-value가 유의 수준 0.05보다 작으므로, 귀무 가설을 기각하고 대립 가설을 선택합니다. 이는 '코로나19 환자의 수축기 혈압은 알려진 국가 통계보다 통계적으로 유의미하게 높다'는 결론을 내리게 됩니다.  즉, 이 결과는 코로나19 환자의 수축기 혈압이 알려진 국가 통계보다 높다는 것을 의미합니다. 이는 귀무 가설을 기각하였기 때문에, 코로나19 환자의 수축기 혈압이 알려진 국가 통계보다 통계적으로 유의미한 차이가 있다는 점을 기반으로 합니다.  따라서, 통계적으로 깔끔하게 올바른 결론은 '코로나19 환자의 수축기 혈압은 알려진 국가 통계보다 높다'라고 할 수 있습니다. 이는 5%의 유의 수준에서 귀무 가설을 기각하고 대립 가설을 선택한 결과입니다. |

**[Q3]** According to the National Institutes of Health (NIH) in the United States, **heart rate in men** is known to follow the distribution: HR\_male ~ N(74, =11.1). Based on this, hypothesize and test alternative hypotheses with **three different directions** at 5% significance level, respectively. Also, **please statistically interpret the final outcome clearly through the analysis**.

|  |
| --- |
| **[Answer]**  검정 통계량은 약 1.68입니다. 이는 샘플 평균이 모집단 평균과 얼마나 다른지를 표준 오차 단위로 표현한 것입니다. 이 값이 크면 클수록 샘플 평균이 모집단 평균과 더 크게 다르다는 것을 의미합니다.  '남성의 심박수가 평균보다 높다'는 가설에 대한 p-value는 약 0.047입니다. 이는 5%의 유의 수준에서 귀무 가설을 기각하고 이 가설을 지지하게 만드는 값입니다. 즉, 샘플 데이터를 바탕으로 남성의 심박수가 국립보건원에서 제시한 평균보다 통계적으로 유의미하게 높다고 할 수 있습니다.  '남성의 심박수가 평균보다 낮다'는 가설에 대한 p-value는 약 0.953입니다. 이는 5%의 유의 수준에서 귀무 가설을 기각할 수 없으므로, 이 가설을 지지할 수 없습니다. 즉, 샘플 데이터를 바탕으로 남성의 심박수가 국립보건원에서 제시한 평균보다 통계적으로 유의미하게 낮다고 할 수 없습니다.  ‘남성의 심박수가 평균과 다르다'는 가설에 대한 p-value는 약 0.093입니다. 이는 5%의 유의 수준에서 귀무 가설을 기각할 수 없으므로, 이 가설을 지지할 수 없습니다. 즉, 샘플 데이터를 바탕으로 남성의 심박수가 국립보건원에서 제시한 평균과 통계적으로 유의미하게 다르다고 할 수 없습니다.  따라서, 이 분석을 통해 '남성의 심박수가 국립보건원에서 제시한 평균보다 높다'는 가설을 지지할 수 있습니다. 다시 말해, 샘플 데이터를 바탕으로 남성의 심박수가 국립보건원에서 제시한 평균보다 통계적으로 유의미하게 높다고 결론 내릴 수 있습니다. |

**[Q4]** Perform the same analysis as in [Q3] **at 10% significance level**. After that, Please describe what is different from [Q3] and **what you can learn from it.**

|  |
| --- |
| **[Answer]**  '남성의 심박수가 평균보다 높다'는 가설은 5%의 유의 수준에서도 지지되었으므로, 10%에서도 지지됩니다.  '남성의 심박수가 평균보다 낮다'는 가설은 5%와 10%의 유의 수준에서 모두 지지되지 않습니다.  '남성의 심박수가 평균과 다르다'는 가설은 5%의 유의 수준에서는 지지되지 않았지만, 10%의 유의 수준에서는 지지됩니다.  이로부터 우리는 두 가지 주요한 점을 배울 수 있습니다:  유의 수준을 높이면 귀무 가설을 기각하기 쉬워집니다. 이는 유의 수준이 우리가 귀무 가설을 잘못 기각하는 오류를 감수할 확률을 결정하기 때문입니다. 유의 수준을 높일수록 이런 오류를 더 많이 감수하게 되므로, 귀무 가설을 기각하기 쉬워집니다.  가설 검정의 결과는 사용하는 유의 수준에 따라 달라질 수 있습니다. 따라서, 가설 검정을 수행할 때는 사용하는 유의 수준을 항상 명시해야 합니다.  이에 따라, 통계적 가설 검정은 단순히 p-value가 0.05보다 작은지를 보는 것만으로는 충분하지 않습니다. 분석의 목적과 문맥, 그리고 우리가 감수할 수 있는 오류의 정도를 고려하여 적절한 유의 수준을 선택해야 합니다. |

**[Q5]** In a basic health checkup for Koreans surveyed by Korea's National Health Insurance, resting SaO2 was known to have an average of 96.0245 and a variance of 8.995. Based on this, please do hypothesis test to conclude whether the currently loaded "Data.txt" is an independent population in terms of resting SaO2. Feel free to set the significance level at an acceptable level, but make sure it leads to the correct conclusion statistically cleanly.

|  |
| --- |
| **[Answer]**  분석 결과에 따르면, 검정 통계량은 약 0.44이고 p-value는 약 0.66입니다. 여기서 유의 수준은 5%로 설정하였습니다.  p-value가 유의 수준 0.05보다 크므로, 귀무 가설을 기각하지 않습니다. 이는 '현재 로드된 데이터의 휴식 SaO2 평균이 국민건강보험에서 조사한 평균과 통계적으로 유의미한 차이가 없다'는 결론을 내리게 됩니다.  즉, 이 결과는 현재 로드된 데이터의 휴식 SaO2 평균이 국민건강보험에서 조사한 평균과 크게 다르지 않다는 것을 의미합니다. 다시 말해, 현재 로드된 데이터는 휴식 SaO2 측면에서 국민건강보험에서 조사한 한국인 개체군을 잘 대표하고 있다고 볼 수 있습니다.  그러므로, 통계적으로 깔끔하게 올바른 결론은 '현재 로드된 데이터는 휴식 SaO2 측면에서 국민건강보험에서 조사한 한국인 개체군을 대표하는 독립적인 개체군이다'라고 할 수 있습니다. 이는 귀무 가설을 기각하지 않았기 때문에, 데이터가 국민건강보험에서 조사한 값과 통계적으로 유의미한 차이가 없다는 점을 기반으로 합니다. |