BE 303
Lab Assignment

b19003
SCEE - IIT Mandi

May 6, 2022

1 General remarks

This lab assignment includes a database which is distributed in addition to this tasksheet as an excel sheet. A general task is the data representation.
One corner stone in any report are high-quality graphs that represent the data analysis. Proper labelling of graphical elements and highlighting of important aspects is essential and required. The quality of graphics will be graded. Thus, when preparing the graphs with R, use the tools available to produce high-quality graphs and charts. You may need commands that you have not used before or were not discussed in the lab or class. Use documentation options and manuals or R such as the help documentation. Another part of the submission are all the R-code files. The code must be commented in such a way that an independent reader can understand the program logic.

For each task your report should include:

- Rationale and description of applied methodology
- Results and analysis of the data
- Conclusion

The submission must include:

- Report including analysis parts and graphics, in sequential order to exercises.
- R-Script File, which includes the R-Code that was used to conduct the assignment.
- Your modified excel sheet, if any modification was done (e.g., deleting not required columns).

The code and excel sheet (the original one if you did not change anything) will be checked for functionality. Therefore ensure that the code runs when excel sheet and code is located in the same file. Do not work with absolute pathways.

It is allowed to use code snippets that we have used in the lab or that you may find from other sources and to modify it for your data analysis. Remember to properly comment the functions you are implementing.

Apart from that **any form of copying** will lead to **zero marks** for the Lab assignment for **all involved parties.**

2 Tasks

2.1 Simulations (30 marks)

The task of this part is to analyze the following points using a simulation:

- the application of central limit theorem in statistical analysis
- to investigate the limitations of the assumptions for a parameteric test
- analyse the alternatives if the assumptions are violated.

In the simulation you can control the population parameters. Based on that you can simulate experiments that sample from that population and you can analyse if the conclusion based on the simulated samples are correct.

Do the following steps:

- Generate a population with 1 000 000 subjects using **uniform** distribution with the probability density function: $100 \leq x \leq 200 = 1$.
- Take n number of samples where n is much larger than 30 and conduct a t-test with a reference value where the nullhypothesis is true. Show that the $\alpha-$value represents the type-I error by repeating the experiment multiple times and evaluating the number how often the Nullhypothesis is kept and rejected.
- Show that the t-test is not functioning correctly when the sample number becomes small. Which assumptions are violated?
- Show that in case the sample size become too small, an alternative test can work.

  (Hint: Use high quality graphics to explain and illustrate you results)

2.2 Data analysis - Part I (30 marks)

Conduct the following tasks:

1. Import the population data from the excel sheet "India_Health_Database". To obtain population data for the years 2012 - 2017, extrapolate the total population for each state and each year, with the data from 2011 and 2018 assuming a linear increase (or decrease).
2. Take the dataset of **Malaria** and compute the **mortality** for corresponding years and state. What is the average **mortality** of all states, and what are the values for the confidence interval $\pm$ 1 SD? Compare your values to other resources using other references. Does your confidence interval include this value?
3. Is there a significant difference in the **mortality** for the different years? Fomulate a nullhypothesis, apply the correct statistical test and analyse the results. (Hint: Be careful with small case numbers and death numbers)
4. Does the **mortality** correlate with the urbanisation level? Explain your results.

2.3 Data analysis - Part II (30 marks)

1. In the attached excel sheet "Glucose_BP_levels" is a dataset of glucose levels and the systolic blood pressure. The research question is if there is a relation between glucose levels and high blood pressure. Choose the correct statistical analysis method and evaluate if there is a measurable effect. Please consider the following points:

    - Report and analyse the output. Can we predict values beyond the given datapoints from the dataset? Or in other words is the model valid for any value of x beyond the dataset? Justify your answer.
    - Compute the Confidence intervals with $\alpha = \mathbf{3}$ % and include them in your graphical representation.
    - Analyse the residuals of your model. What do you observe? Discuss the results and potential fallacies.