DS203 Programming for Data Science

Assignment 5 – Statistical Testing

Instructions:

1. Use python and libraries such as numpy, matplotlib, and pandas.
2. Submit your assignment, preferably a single .ipynb file, by September 11, 2022, 11:59pm on Moodle at: https://moodle.iitb.ac.in/mod/assign/view.php?id=92431
3. In your assignment, show not only the final operations that worked, but also the results of one or two other things that you tried that did not work.
4. Include copious comments about your thoughts in choosing the operations for a particular problem, and your observations about what worked and what did not work, including possible reasons.
5. Comment every line of code to demonstrate understanding of how the code works. Write as much code as you can on your own, but you can use functions from the aforementioned libraries.
6. Cite sources referred at the place where they were used, and include a reference section at the end.

Problems:

1. We will model the data from "% of fouses having mud or unburnt brick as material of wall" from the previous assignment based the file NDAP_REPORT_7004.csv.
   a. Examine Q-Q plots for Gaussian and uniform distributions, and visually assess which of the two is a better fit. [2]
   b. Find out MLE parameter estimates for Gaussian distribution. [1]
   c. Find out MLE parameter estimates for uniform distribution. [1]
   d. Between Gaussian and uniform with MLE parameters, determine the one that is more likely to explain the observed data. [3]
2. Assume there is a new type of distribution called IQ (inverted quadratic) given by p(x) = 0.75 $(1-x^2)$ for $|x| \leq 1$ and p(x) = 0 for $|x|>1$.
   a. Write a python function to generate Q-Q plot for input samples with respect to the IQ function and use it for XXX. [3]
   b. Introduce parameters into IQ to allow scaling and shifting (while maintaining an area of 1) and show the derivation. [2]
   c. Derive expressions for the MLE estimates, show the derivation, and write a python function that returns the MLE estimates of the parameters given a column of data. [2]
   d. Write a python function to compute data log likelihood for the parameterized IQ function, and use it to compute the log likelihood of the data from Q1 for the MLE estimate parameters. [3]
3. Determine the correct type of test for each of the following , state the reason for choosing that test, and perform the test using in-built functions in python libraries:
   a. Determine if the percent of houses across districts with "mud or unburnt brick" is really larger for rural versus urban in the housing data. [2]
   b. Determine if either rural or urban percent of houses with "mud or unburnt brick" for the same district can be predicted using the other in the housing data. [2]
   c. Determine if there is a relation between body style and drive wheels in the automobile data. [2]