Amrita VishwaVidyapeetham
Amrita School of Computing, Coimbatore
Lab Evaluation -3
Fifth Semester
Computer Science and Engineering
19CSE304 Foundations of Data Science

Course Outcomes (COs):

CO	Course Outcomes
CO01	Understand the statistical foundations of data science.
CO02	Apply pre-processing techniques over raw data so as to enable further analysis.
CO03	Conduct exploratory data analysis and create insightful visualizations to identify patterns.
CO04	Identify machine learning algorithms for prediction/classification and to derive insights
CO05	Analyse the degree of certainty of predictions using statistical test and models

Duration: One hour Maximum: 20 Marks

The dataset *test.csv* came from a sample of newborns in a large hospital system. You treat it as if it were a simple random sample. The data contains the following variables for 1,174 mother-baby pairs: *the baby's birth weight in ounces, the number of gestational days, the mother's age in completed years, the mother's height in inches, pregnancy weight in pounds, and whether or not the mother smoked during pregnancy.*

The objective of your study is to see whether there is any difference in the ages of the smoking and non-smoking mothers. The claim is that the average age of mothers who smoked is less than those who did not smoke. You are required to employ an appropriate and suitable statistical test. Assume the alpha value as 5%.

Specific Requirements:

1. Draw the histograms of the two distributions in the sample(2), and state your inferences(1).
2. Which statistical test would you employ?(2)
3. State the Null and Alternative Hypothesis (1).
4. What statistic do you recommend to apply for this case(2)?
5. State the observed statistic(2)
6. Simulate the test statistic under the Null that will help make the decision(3).
7. Draw a histogram to depict what the model of random selection predicts about the statistic.(3)
8. Evaluate the empirical p-value of the test(3).
9. What is the outcome of your analysis? (1)
