**AMPBA**

**Statistical Analysis - I**

**Individual Assignment**

**Due Date: 24-09-2023**

Download the file “Data for Assignment” from LMS.

Select a simple random sample of size 1000 from the given data base. Open the Excel sheet “Dataset for Assignment”. Press the function key F9 once or twice and make sure that the numbers in columns A to L change automatically. Copy the entire range A1:L1001 into a new Excel sheet as values. This is your own unique dataset.

The data refers to the performance of beneficiaries in the adult education program. The variables are self-explanatory. The variables “WRITE”, “READ”, “MATH” and “TOTAL” are the scores in respective components and the total score. The “TOTAL” score is obtained by adding the three scores namely “WRITE”, “READ” and “MATH”. Check if there is any inconsistency in the “TOTAL” score and if so, replace it with the sum of the three scores namely “WRITE”, “READ” and “MATH”.

You will have to clean up the data by replacing #NULL! with blank or zero.

Details of the other variables are given below:

**GENDER: T:** Male; **F:** Female

**AGE:** Age in completed years

**CASTE: OT:** Other Caste; **SC:** Scheduled Caste; **ST:** Scheduled Tribe

**RELIGN:** Religion **C:** Christian; **H:** Hindu; **M:** Muslim

**MTOUNGUE:** Mother Tongue**; D:** Urdu; **K:** Kannada; **T:** Tamil; U: Telugu

**OCCU:** Occupation; **A:** Agriculture; **B:** Small Business; **C:** Cooli; **H:** Housewife; **U:** other

**INCOME:** Income in Rs. Per month

**AREA:** Agricultural land in acres

1. Draw a histogram for the TOTAL score (you have to define appropriate class intervals) and comment on the distribution of the TOTAL score.
2. We take a simple random sample of 16 learners. What is the probability that the sample average of this sample () is more than 60?
3. Create a new variable called “PERFORMANCE”. If the score is above the sample average, label the PERFORMANCE as “HIGH”, otherwise, “LOW”. Using the new variable “PERFORMANCE” and GENDER, comment on the relative performance of Male and Female learners.
4. Using the sample that you have just selected, calculate a 95% two-sided confidence interval for the following.

A. Mean of TOTAL Score

B. Mean of INCOME

C. Proportion of “HIGH” performers among Male learners

D. Proportion of “HIGH” performers among Female Learners.

1. If we want to estimate a 95% confidence interval for the population mean of “TOTAL” marks within ± 5 marks, what is the sample size required? Please note that you do not know σ. You may use your present sample of 1000 as a pilot sample.
2. If we want to estimate a 95% confidence interval for the proportion of “HIGH” performers within ± 0.05, what is the appropriate sample size?