SRN □□□□□□□□□□□□

PES University, Bengaluru
(Established under Karnataka Act No. 16 of 2013)

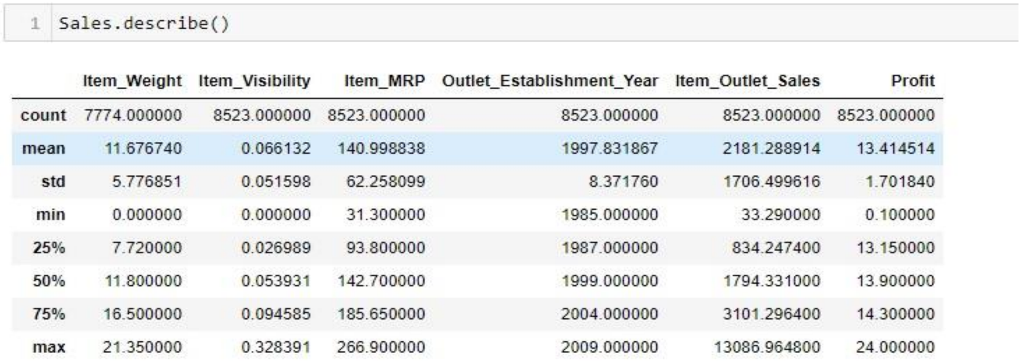
UE20CS902

April 2022: END SEMESTER ASSESSMENT (ESA)
M TECH DATA SCIENCE AND MACHINE LEARNING_ SEMESTER I

UE20CS902 – Statistical Methods for Decision Making

Time: 3 Hrs	Answer All Questions	Max Marks: 80

1 a) Consider the following output file of sales data — 2



```
Sales.describe()
```

```
       Item_Weight  Item_Visibility     Item_MRP  Outlet_Establishment_Year  Item_Outlet_Sales       Profit
count  7774.000000      8523.000000  8523.000000                8523.000000        8523.000000  8523.000000
mean     11.676740         0.066132   140.998838                1997.831867        2181.288914    13.414514
std       5.776851         0.051598    62.258099                   8.371760        1706.499616     1.701840
min       0.000000         0.000000    31.300000                1985.000000          33.290000     0.100000
25%       7.720000         0.026989    93.800000                1987.000000         834.247400    13.150000
50%      11.800000         0.053931   142.700000                1999.000000        1794.331000    13.900000
75%      16.500000         0.094585   185.650000                2004.000000        3101.296400    14.300000
max      21.350000         0.328391   266.900000                2009.000000       13086.964800    24.000000
```

Based on the data and using coefficient of variation, order the high to low in terms of coefficient of variation value (Do not consider outlet establishment year for calculation).

b) Compute the mean of the following data set — 2

field_1 = [10,9,8,10,11,8,9,11,8,9,14,6,7,8,9,8,7,10]

c) Explain the usefulness of central limit theorem in sampling distribution — 2

d) Explain about bootstrapping. — 2

e) Under what conditions or scenario would you prefer using a z test or a t test? — 2

2 a) While conducting t test would you do prefer checking for normality, if so what is the test for normality you would prefer ? — 2

b) What does ANOVA stand for ? State its purpose. — 2

c) Provide the test statistic for Z test when conducting single sample test, also state the notation's used. — 2

d) A bin contains 3 different types of lamps. The probability that a type 1 lamp will give over 100 hours of use is 0.7, with the corresponding probabilities for type 2 and 3 lamps being 0.4 and 0.3 respectively. Suppose that 20 per cent of the lamps in the bin are of type 1, 30 per cent are of type 2 and 50 per cent are of type 3. What is the probability that a randomly selected lamp will last more than 100 hours?

2

e) Identify the possible distribution for the following scenario

i) Number of covid cases per day.

ii) Counting the number of defects from inspected sample of size n.

2

SECTION B – 30 MARKS

3 a) Two catalysts are being analyzed to determine how they affect the mean yield of a chemical process. Specifically, catalysts 1 is currently in use, but catalyst 2 is acceptable. Since catalyst is cheaper, it should be adopted, providing it does not change the process yield. A test is run in the pilot plant and results in the data as shown in the table. Is there any difference in mean yields for an α =.05 and assume equal variances.

Number	1	2	3	4	5	6	7	8
Catlayst1	91.50	94.18	92.18	95.39	91.79	89.07	94.72	89.21
Catalyst2	89.19	90.95	90.46	93.21	97.19	97.04	91.07	92.75

State the hypothesis and type of test to be used (2 marks)

Test the hypothesis and conclude (4 marks)

6

b) Until a certain period of time 1975-1980 the mean price / earnings (P/E) ratio of approximately 1800 stocks was 14.35 and the standard deviation was 9.73. In a sample of 30 randomly chosen NYSE stocks, the mean (P/E) ratio for the year 1981 was 11.77. Using this estimate can we conclude for a significance level of .05 that there is a change in NYSE value from its previous value? Please provide steps in concluding.

State the hypothesis and type of test to be used (2 marks)

Test the hypothesis and conclude (4 marks)

6

3 c The demand for a particular spare part was found to vary from day to day. In a sample study the following information was obtained.

Quantity demanded

Days	Mon	Tue	Wed	Thur	Friday	Saturday
Quantity demanded	1124	1125	1110	1120	1126	1115

i) Write the hypothesis.(2 marks)

6

SRN

		ii. Test the hypothesis at 1% level of significance that the number demanded depends upon the day (4 marks)	
3	d	General hospitals patient account division has compiled data on the age of accounts receivable. The data collected indicate that the age of the accounts follows a normal distribution with μ = 28 days and σ = 8 days. What portion of the accounts is between 20 and 40 days old ? (4 marks) The hospital administrator is interested in sending reminder letters to the oldest 15% of accounts. How many days old should an account be, before a reminder letter is sent. (2marks)	6

3	e	A pharmaceutical company claims that its new tablet is effective in increasing the height of children. The data of heights (in cm) of 7 children is recorded before and after consuming the tablet. Check for normality of the data and perform appropriate test (2 marks). Test the company's claim at a 5% level of significance using the p-value approach (4 marks).
		ht_before = [121, 125, 130, 120, 145, 126, 134]
		ht_after = [130, 129, 148, 122, 147, 130, 148]
			6

SECTION C – 30 MARKS			
4	a	Consider the insurance.csv file and answer the following questions i) Provide a summary statisitcs of all the variables, based on the summary which variable do you think has more variability. (4 marks) ii) Provide a histogram for the variable bmi, based on histogram and calculation of mean, median and mode what would be the closest distribution you would suggest (3 marks) iii) Plot the histogram for the variable charges, also plot the histogram for variable charges based on smoker type, based on these three plots what do you observe, what do you conclude (assume bins = 15). (5 marks) iv) Draw a scatter plot for all of the variables, what is your observation and conclusion for the relationship between age and charges, bmi and charges, would you prefer to further subclassify and develop a scatter plot, if yes or no why ? (3 marks)	15
	b	Consider the insurance.csv file and answer the following questions i) Check whether the BMI data follows normal distribution by using a qq plot and shaipro test (3marks) ii) Construct a hypothesis to prove that smoker have higher charges than non smoker and conclude accordingly (4 marks) iii) Is there a statisitical difference between insurance charges for male and female and conclude accordingly (4 marks) iv) Is the proportion of smokers significantly different in different genders? (4 Marks)	15