SRN | | | | | | | | | |

PES University, Bengaluru

(Established under Karnataka Act 16 of 2013)

END SEMESTER ASSESSMENT(ESA) B. TECH 3RD SEMESTER CSE

(Section D)

UE15CS203

INTRODUCTION TO DATA SCIENCE

Time: 3 Hrs | Answer All Questions | Max Marks: 100

1.	a	Answer precisely. i) Population mean is 100 and variance is 0. When would this happen? ii) Outlier has greater influence on all these : mean, median and mode. Comment about this statement. iii) Mean is 100. Median is 50. What can you conclude about the distribution? iv) Mean is 100. Standard deviation is 100. Is this possible? Justify your answer. v) Four persons A, B, C and D were born in the years 2000, 2002, 2003, 2009. What is the relationship between the standard deviation of their ages in 2010 compared to the standard deviation in 2016?	10
	b	Assume that the following functions are available. max min from statistics import mean, median, variance. Given a list of lists, each list indicates a sample data, write a function to find the following i) which sample has the highest mean? ii) which sample has the data clustered around the mean? # the parameter x is a list of lists def find_result(x) : # TODO #return (,) # tuple of two lists	6
	c	Assume that the following functions are available. max min from statistics import mean, median, variance. Write a function to the following. Given a list, Find the outliers. Remove them from the list. Return the old mean and the new mean. You may assume that len(x) times 0.5, 0.25 and 0.75 give integer results. def filter_outlier(x) : # TODO	4

2	a	Suppose that the random variable T has the following probability distribution: t | 0 1 2 --- P(T = t) | .5 .3 .2 i. Find P(T <= 0) ii. Find P(T >= 0 and T < 2) iii. Compute the mean of the random variable T.	5
2	b	At a certain airport, 75% of the flights arrive on time. A sample of 10 flights is studied. I) Find the probability that exactly eight of the flights were on time.	5
	c	ii) Find the probability at least eight of the flights are on time? iii) Find the probability at most two flights are delayed?	5
	d	What is the relationship between the mean and the variance in Poisson Distribution? For this distribution, find the following. Assume λ as 3. i) P(X = 2) ii) P(X <= 2) iii) P(X > 2) One hundred students took a test on which the mean score was 73 with a variance of 64. A grade of A was given to all who scored 85 or better. Approximately how many A's were there, assuming scores were normally distributed? What happens if we add 5 marks to each student? What happens if the marks are scaled up by a factor of 1.05? Assume that the cut off limit of 85 is not changed.	5
3	a	A topic of interest in ophthalmology is whether or not spherical refraction differs between the left and right eye on average. In a study to investigate this, refraction was measured on the left and right eye of 17 patients. The differences (right - left) in diopters were $d1,d2,\cdots d17$ and elementary calculations gave $\sum d_i = -3.50$ and $\sum d_i^2 = 19.13$. Provide a 90% confidence interval for the average difference (right - left). Use student T distribution.	5
	b	What is the smallest sample size required to provide a 95% confidence interval for a mean, if it the interval be no longer than 1cm? You may assume that the population is normal with variance $9cm^2$. I) What will be the smallest size if the interval is made 2 cm. ii) Will the smallest size increase or decrease if the confidence level is made 90%?	5

	c	A coin is tossed 100 times. It lands head up 60 times. What is the confidence interval that we get a head 95% time based on this experiment? What would happen to the interval if the coin had landed head up 50 times - would this increase or decrease? You may use the traditional method.	5
	d	How do you distinguish between the confidence interval for the difference of two means and confidence interval with paired data? Given these two samples, find Confidence Intervals at 95% level for difference of two means. X : 15 16 21 17 18 15 19 21 Y : 30 37 39 37 40 39 34 40 $\bar{X}$ = 17.75 $\bar{Y}$ = 37 S_X = 2.4349 S_Y = 3.4641	5
4	a	In a survey in Bangalore, 260 of 500 citizens oppose chopping trees to facilitate metro. Based on the hypothesis testing, can we conclude that the majority are opposed to chopping trees to facilitate metro? What should have been the number of citizens opposing chopping of trees to conclude that the majority oppose chopping of trees? Take P = 0.05.	6
	b	A reading coordinator in a large public school system suspects that poor readers may test lower in IQ than children whose reading is satisfactory. He draws a random sample of 30 fifth grade students who are poor readers. Historically fifth grade students in the school system have had an average IQ of 105. The sample of 30 has XBAR = 101.5 and S(XBAR) = 1.42. Test the appropriate hypothesis at the 5% level	6
	c	What are type 1 and type 2 errors? What is the relationship between type 1 error and α ? What is relationship between power of a test and type 2 error?	6
	d	On what sort of data is Chi-square test applied? What if the chi-square value is large?	2
5	a	Comment In a line or two. i) A parameter (like the mean) of the population is non-variant. Would the confidence interval at some particular level be also non-variant? ii) A small P value in hypothesis testing indicates the probability of H0 being True. iii) two sides 95% confidence interval consists of those values of mean whose P values are greater than 0.025 in a two tailed hypothesis test. {alpha is 5%}. iv) Chi-square tests for homogeneity and independence use two different samples of the same population.	10

		v) if both μ of Null and alternate hypotheses lies on the same side of the critical point, the value of the power test will be less than or equal to 0.5.	
	b	Find mean of x, mean of y, variance of x, variance of y, correlation coefficient. What can you conclude based on the correlation coefficient and the given data set?	6

x	y
8	6.58
8	5.76
8	7.71
8	8.84
8	8.47
8	7.04
8	5.25
19	12.5
8	5.56
8	7.91
8	6.89

5	c	I) What is the unit of correlation coefficient? II) What happens when x and y are interchanged? III) What happens if we add a constant to each value of the variable? IV) What happens if we multiply each variable by a factor?	4