SRN | | | | | | | | | | | | |

PES UNIVERSITY

PES University, Bangalore

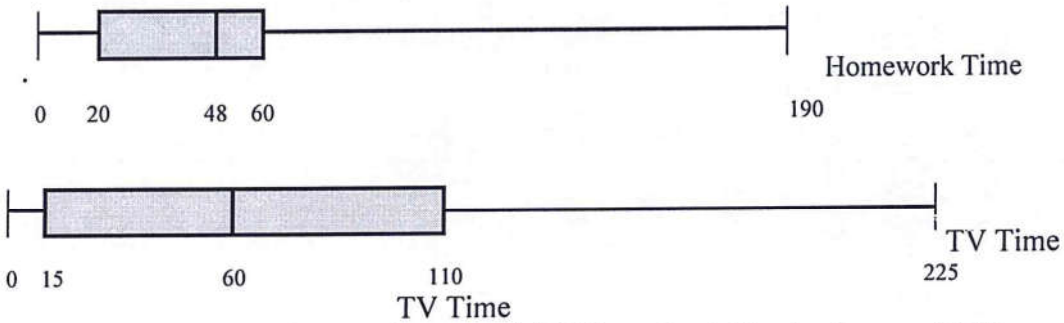
(Established under Karnataka Act No. 16 of 2013)

UE19CS203

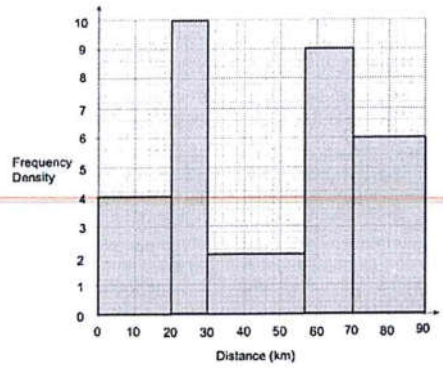
END SEMESTER ASSESSMENT (ESA) B.TECH. III SEMESTER – DEC. 2020

UE19CS203 – STATISTICS FOR DATA SCIENCE

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

1.	a)	i) A sociologist conducts an opinion survey in a major city. Part of the research plan calls for describing and comparing the opinions of four different Dravidian groups: Kannada, Malayalam, Tamil and Telugu. For a total sample of 300, the researcher selects 75 participants from each of the four predetermined subgroups. Name the sampling methods that can be used for random and non-random sampling. ii) A shipment of apples is to be tested for quality. A quality inspector draws a simple random of 40 apples and tests the condition of each. She finds that 6 of them, or 15%, are rotten. She concludes that exactly 15% of the shipment is rotten. However, her supervisor claims that the proportion of rotten apples is close to 15%. Whose conclusion is more statistically appropriate? Justify your answer. iii) Now, a different inspector conducted the same experiment(part ii))but found that 4 apples, or 10%, are rotten. The first inspector claims that he must have done something wrong, since her results showed 15% and not 10%. Is she right? Justify your answer.	5
	b)	i) Define a web scraper. ii) What is imputation? Mention any two techniques that falls under imputation.	5 (2+3)
	c)	The box-and-whisker plots shown below compares homework time per night with TV time per night for the same group of students.  i) What percent of the students watch TV for at least 15 minutes per night? ii) What percent of the students watch TV for more than an hour per night? iii) What percent of the students do homework for more than an hour per night? iv) Which data is more varied? v) Write the distribution of Homework time.	5
	d)	Some cyclists from a local cycling club go out for their usual Sunday ride. There are many different lengths of routes to suit cyclists of all abilities which are shown in the histogram given below. i) Estimate the number of cyclists who rode for 30 kilometers or less. ii) Estimate the number of riders in the interval 57-70.	5 (3+2)

SRN | | | | | | | | | | | |



2.	a)	A pharmaceutical lab states that a drug causes negative side effects in 3 of every 100 patients. To confirm this affirmation, another laboratory chooses 5 people at random who have consumed the drug. What is the probability of the following events? i) None of the five patients experience side effects. ii) At least two experience side effects. iii) What is the average number of patients that the laboratory should expect to experience side effects if they choose 100 patients at random?	6
	b)	Vehicles pass through a junction on a busy road at an average rate of 300 per hour. i) Find the probability that none passes in a given minute. ii) What is the expected number passing in two minutes? iii) Find the probability that this expected number actually pass through in a given two-minute period.	4
	c)	The length of human pregnancies from conception to birth approximates a normal distribution with a mean of 266 days and a standard deviation of 16 days. i) What proportion of all pregnancies will last between 240 and 270 days (roughly between 8 and 9 months)? ii) What length of time marks the shortest 70% of all pregnancies?	5
	d)	The toll on NH-4 is \$1.50 for cars and \$4.50 for trucks. The mean and variance of the number of cars is 15,000 and 250,000; and the mean and variance of the number of trucks is 5,000 and 10,000. What is the mean and standard deviation of the daily toll revenue? Assume that cars and trucks are independent.	5
3	a)	For the following random sample, find the likelihood function and the maximum likelihood estimate of θ. $X_i \sim$ Binomial$(3, \theta)$ and we have observed $(x_1, x_2, x_3) = (3, 2, 2)$.	4
	b)	An article reports that out of 10,500 surgeries, 850 resulted in complications within six months of surgery. A surgeon claims that the rate of complications is less than 8.5%. With what level of confidence can this claim be made?	6
	c)	A telephone company has determined that during non-holidays the number of phone calls that pass through the main branch office each hour follows the normal distribution with mean $\mu = 80000$ and standard deviation $\sigma = 35000$. Suppose that a random sample of 60 non-holiday hours is selected and the sample mean of the incoming phone calls is computed. i) Describe the distribution of X and write the values of the parameter. ii)Find the probability that the sample mean X of the incoming phone calls for these 60 hours is larger than 91970.	4 (2+2)
	d)	The following data presents a confidence interval for a population mean, but some of the numbers are missing. Find the missing numbers for (i), (ii) and (iii).	6

n	mean	sigma	SE of mean	99% Confidence Interval
20	2.39374	(i)	0.52640	((ii), (iii))

SRN | | | | | | | | | | | |

4	a)	Suppose a consumer group suspects that the proportion of households that have two or more cell phones is 30%. A cell phone company has reason to believe that the proportion is not 30%. Before they start a big advertising campaign, they conduct a hypothesis test with a significance level of 10%. Their marketing people survey 150 households with the result that 43 of the households have two or more cell phones. Can we conclude that the consumer group is correct?	5
	b)	The thicknesses of eight pads designed for use in aircraft engine mounts are measured. The results, in mm, are 41.83, 41.01, 42.68, 41.37, 41.83, 40.50, 41.70, and 41.42. Assume that the thicknesses are a sample from an approximately symmetric distribution. The target thickness is 42 mm. Can you conclude that the mean thickness differs from the target value? Compute the appropriate test statistic and find the P-value. Make your conclusion.	6
	c)	Assessments of health outcomes of people working in an environment with high levels of carbon monoxide are presented. Following are the numbers of workers reporting various symptoms, categorized by work shift. The numbers were read from a graph. Shift Morning Evening Night Influenza 16 13 18 Headache 24 33 6 Weakness 11 16 5 Shortness of Breath 7 9 9 Can you conclude that the proportions of workers with the various symptoms differ among the shifts?	7
	d)	For the given null hypothesis, write Type I error as a statement. H_0 : Medicine A cures Disease B	2
5.	a)	A copper smelting process is supposed to reduce the arsenic content of the copper to less than 1000 ppm. Let μ denote the mean arsenic content for copper treated by this process, and assume that the standard deviation of arsenic content is $\sigma = 100$ ppm. The sample mean arsenic content X of 75 copper specimens will be computed, and the null hypothesis H0 : ≥ 1000 will be tested against the alternate H1 : $\mu < 1000$. i). A decision is made to reject H0 if $X \leq 980$. Find the level of this test. ii). Find the power of the test in part (a) if the true mean content is 965 ppm. iii). How large a sample is needed so that a 5% level test has power 0.95 when the true mean content is 965 ppm?	10
	b)	In a study relating the degree of warping, in mm, of a copper plate (y) to temperature in ◦C (x), the following summary statistics were calculated: $n = 40$, $\sum(xi - xbar)^2 = 98{,}775$, $\sum(yi - ybar)^2 = 19.10$, xbar = 26.36, ybar = 0.5188, $\sum (xi - xbar)(yi - ybar) = 826.94$. i). Compute the correlation r between the degree of warping and the temperature. ii). Compute the error sum of squares. iii). Compute the least-squares line for predicting warping from temperature. iv) Predict the warping at a temperature of 40◦C. v). At what temperature will we predict the warping to be 0.5 mm?	10