

OCTOBER 2020: IN SEMESTER ASSESSMENT, B.TECH, III-SEMESTER

TEST – 1

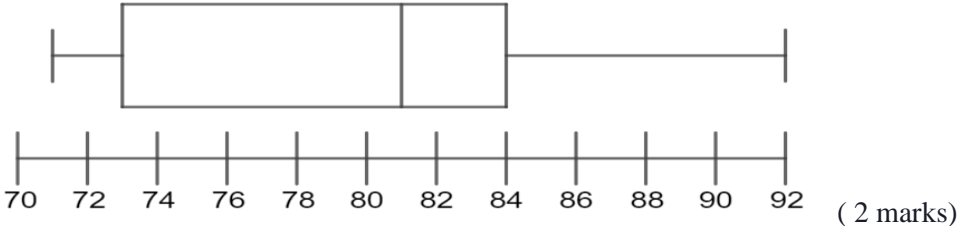
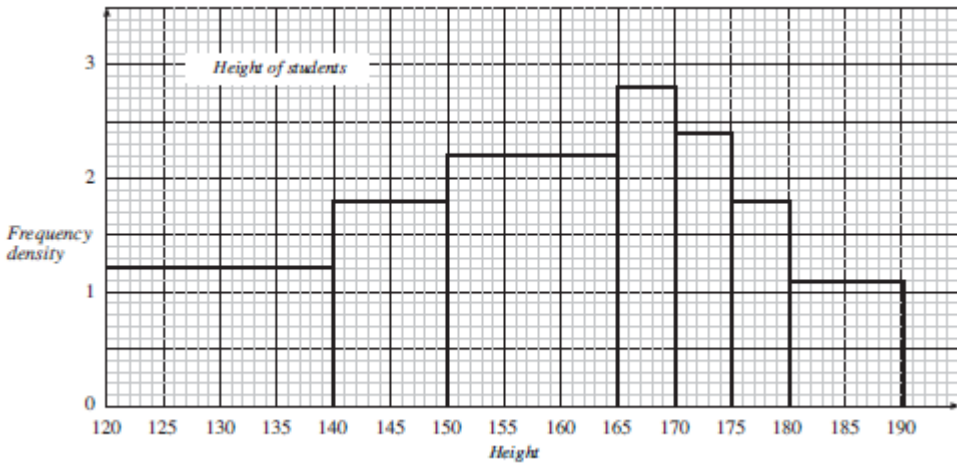
**UE19CS203 – STATISTICS FOR DATA SCIENCE
SCHEME & SOLUTION**

Time: 02 Hrs

Answer All Questions

Max Marks: 60

Q. No.		Marks														
1.	<p>a) Asha is conducting an experiment on training certain breeds of dogs. She wants to know how long, on average, it would take to teach a Labrador to fetch an object. She gets a group of dogs to conduct her experiment. 5 of the dogs are Labradors and 3 of the dogs are Dalmatians. What is the population and sample in this experiment?</p> <p>Solution: Population is all Labradors. Sample is 5 Labradors.</p>	2														
b)	<p>i) A psychologist is studying the sleep patterns of the 3960 students at his university. He decides to start by asking a random sample of 30 students how many hours of sleep they get weekday nights. Students are listed by the neighbourhood they live in. The psychologist randomly selects six neighbourhoods and then randomly selects five students from each one. Name the sampling technique used.</p> <p>ii) A television reporter interviewed travelers stranded at an airport during a snowstorm about the efficiency of air travel in Bangalore. Name the sampling technique used.</p> <p>Solution: i) Cluster Sampling ii) Convenience Sampling</p>	2														
c)	<p>Briefly explain any two treatments that can be given to address missing values.</p> <p>Solution:</p> <ol style="list-style-type: none">1. List-wise deletion or Drop Observation2. Imputation3. Drop the Feature <p>(Brief explanation any two – each 1 mark)</p>	2														
d)	<p>A sample of 100 adult women was taken, and each was asked how many children she had. The results were as follow</p> <table border="1"><tr><td>Children</td><td>0</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr><tr><td>Number of women</td><td>27</td><td>22</td><td>30</td><td>12</td><td>7</td><td>2</td></tr></table> <p>For what proportion of the women was the number of children more than one standard deviation greater than the mean?</p> <p>Solution:</p> <p>Mean = 1.56 (1 mark)</p> <p>Standard deviation $s = 1.3052$ (1 mark)</p> <p>$\bar{X} + s = 1.56 + 1.3052 = 2.8652$ (0.5 mark)</p> <p>(12+7+2) women had more than 2.86 children (0.5 mark)</p> <p>Therefore, proportion is $21/100 = 21\%$. (1 mark)</p>	Children	0	1	2	3	4	5	Number of women	27	22	30	12	7	2	4
Children	0	1	2	3	4	5										
Number of women	27	22	30	12	7	2										

2.	<p>a) A test was conducted for a class of students and the following are the facts about students' scores (Max. score is 100). Lower Quartile=73, Range=21, Largest value=92,Median=81,IQR=11 Construct a Box Plot using the given information and comment about distribution of data.</p> <p>Solution: Upper Quartile=Lower Quartile + IQR=73+11=84; Smallest value=Largest value-Range = 92 – 21 = 71 (1 mark)</p>  <p>(2 marks)</p> <p>Distribution : Left skewed. (1 mark)</p>	4
	<p>b) The results of a survey into the height of students in a college are shown below using a histogram.</p> <p>Find the</p> <ol style="list-style-type: none"> intervals which have approximately equal number of students number of students in the interval 150-165 and total number of students measured.  <p>i) The intervals 170-175 and 180-190 have approximately equal number of students. (1 mark)</p> <p>ii) Number of students in 150-165 = 15*2.2 = 33 (1 mark)</p> <p>iii) Total number of students measured</p> $= 20*1.2 + 10*1.8 + 15*2.2 + 5*2.8 + 5*2.4 + 5*1.8 + 10*1.1$ $= 24 + 18 + 33 + 14 + 12 + 9 + 11$ $= 121.$ <p>(2 marks)</p>	4
c)	<p>Mala is trying to understand why a student would pick computer science branch. 64% of the sample she surveyed said that they picked computer science because of the vast employment opportunities. Her survey comes from 2000 students in 50 different colleges across the nation. Can we infer a parameter from this information? What would it be?</p> <p>Solution: Yes. The statistic is 64% of the students choose computer science branch because of the employment opportunities. (1 mark)</p> <p>We can infer that the parameter is the same. (1 mark)</p>	2

3.	a)	<p>A box contains three \$5 bills, two \$10 bills, five \$15 bills and one \$20bill. Let X be a random variable equal to the value of a single bill drawn at random from the box. Write the probability distribution of X.</p> <p>Solution:</p> <table><tr><td>X:</td><td>5</td><td>10</td><td>15</td><td>20</td></tr><tr><td>P(X=x):</td><td>3/11</td><td>2/11</td><td>5/11</td><td>1/11 (or)</td></tr><tr><td>P(X=x):</td><td>0.273</td><td>0.182</td><td>0.454</td><td>0.091</td></tr></table>	X:	5	10	15	20	P(X=x):	3/11	2/11	5/11	1/11 (or)	P(X=x):	0.273	0.182	0.454	0.091	2
X:	5	10	15	20														
P(X=x):	3/11	2/11	5/11	1/11 (or)														
P(X=x):	0.273	0.182	0.454	0.091														
	b)	<p>If from six to seven in the evening one telephone line in every five is engaged in a conversation: what is the probability that when 10 telephone numbers are chosen at random, only two are in use?</p> <p>Solution:</p> <p>Bin(10, 1/5) (1 mark) p = 1/5 (1 mark) q= 1 – p = 4/5 (1 mark)</p> <p>p(X=2)=10C₂ (1/5)² (4/5)⁸ =0.3020 (1 mark)</p>	4															
	c)	<p>I run a business where my daily revenue has a mean of 1500 and a standard deviation of 400, while my daily costs have a mean of 1000 and a standard deviation of 300. What is the mean and standard deviation of my daily profit, assuming my daily revenue and costs are independent?</p> <p>Solution:</p> <p>Let P be a random variable defines daily profit.</p> <p>P = R – X (1 mark)</p> <p>E[P] = E[R] – E[X] = 500 (1 mark)</p> <p>Var(R – X) = Var(R) + Var(X) = 400² + 300² = 500² (1 mark)</p> <p>σ_{R-X} = 500 (1 mark)</p>	4															
4.	a)	<p>Bacteria in a culture live for an average time of three hours with a standard deviation of 20 minutes. At least what fractions of the bacteria live between two and four hours?</p> <p>Solution:</p> <p>Two and four hours are each one hour away from the mean.</p> <p>One hour corresponds to three standard deviations. (1 mark)</p> <p>So at least 1 – 1/32 = 8/9 =88.89% of the bacteria live between two and four hours. (2 marks)</p>	3															
	b)	<p>Suppose that the weight of pencils is normally distributed with mean 8 grams, and standard deviation 1.5 grams. What proportion of pencils weigh between 10.3 and 14 grams?</p> <p>Solution:</p> <p>P(10.3 < X < 14) = P((10.3 – 8) /1.5 < Z < (14 – 8)/ 1.5)</p> <p>= P(1.53 < Z < 4) ≈ 1 – 0.9370 = 0.0630.</p>	3															
	c)	<p>Generate normal random variates X₁ and X₂ using Box Muller method for the given μ = 15 and the variance =16 and the random numbers R₁ = 0.1543 and R₂ = 0.0724.</p>	4															

		<p>Solution:</p> <p>The random variates X_1 and X_2 are given by</p> $X_1 = \mu + Z_1 * \sigma \quad \text{and} \quad X_2 = \mu + Z_2 * \sigma$ <p>where $Z_1 = \sqrt{-2 \ln(R_1)} * \cos(2 * \pi * R_2)$ and</p> $Z_2 = \sqrt{-2 \ln(R_1)} * \sin(2 * \pi * R_2) \quad (1 \text{ mark})$ $Z_1 = \sqrt{-2 \ln(0.1543)} * \cos(2 * 3.1416 * 0.0724)$ $= 1.933 * \cos(0.4549) = 1.933 * 0.8983 = 1.736$ $Z_2 = \sqrt{-2 \ln(0.1543)} * \sin(2 * 3.1416 * 0.0724) \quad (2 \text{ marks})$ $= 1.933 * \sin(0.4549) = 1.933 * 0.4394 = 0.8494$ $X_1 = \mu + Z_1 * \sigma = 15 + 1.736 * 4 = 21.944 \text{ and } (1 \text{ mark})$ $X_2 = \mu + Z_2 * \sigma = 15 + 0.8494 * 4 = 18.3976$	
5.	a)	<p>Let X_1, X_2, \dots, X_n be an independent random sample from a population with poisson(λ). Find MLE of λ.</p> <p>Solution :</p> $e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}$ <p>The probability mass function is given by $p(X=x) =$</p> <p>The likelihood function is</p> $L(x_1, \dots, x_n; \lambda) = \frac{\lambda^{\sum x_i} e^{-n\lambda}}{\prod_{i=1}^n x_i!} \quad (1 \text{ mark})$ <p>The log likelihood function is</p> $\ln L(x_i; \lambda) = \sum x_i \ln(\lambda) - n\lambda \ln(e) - \ln(x_1!, \dots, x_n!) \quad (1 \text{ mark})$ <p>Taking derivative with respect to λ and set it to 0,</p> $\frac{d}{d\lambda} \ln[L(x_i; \lambda)] = 0 \quad (1 \text{ mark})$ $\sum x_i \left(\frac{1}{\lambda}\right) - n - 0 = 0$ <p>The MLE of λ is $\hat{\lambda} = \bar{X} \quad (1 \text{ mark})$</p>	4
	b)	<p>At a large university, 25% of students are over 21 years old. In a sample of 400 students, calculate the probability that more than 110 of them are over 21 years old.</p> <p>If $X \sim B(n, p)$ and if n is large and/or p is close to $\frac{1}{2}$, then X is approximately $N(n*p, n*p*q)$</p> <p>Here n is large so normal distribution can be used as an approximation to the binomial distribution.</p> <p>mean = $n*p = 400*0.25 = 100$</p> <p>Sample SD $s = \sqrt{n*p*(1-p)} = \sqrt{400(0.25)(1-0.25)} = \sqrt{400(0.25)(1-0.25)} = 8.66$</p>	3

		<p>since $n \cdot p = 400 \cdot 0.25 = 100$ $n \cdot q = 400 \cdot (1 - 0.25) = 300$, both are greater than 10, we use continuity correction factor.</p> <p>$P(X > 110) = P(X > 110.5)$ [using continuity correction factor] $Z = (x - \text{mean}) / \text{sd}$ $Z = (110.5 - 100) / 8.66$ $Z = 1.1124$ $:$ $P(Z > 1.1124) = 1 - 0.8873 = 0.1127$ Therefore, $P(X > 110)$ is approximately 0.1127</p>	
	c)	<p>The record of weights of the female population follows the normal distribution. Its mean and standard deviations are 60 kg and 10 kg respectively. If a researcher considers the records of 50 females, then what would be the mean and standard deviation of the chosen sample?</p> <p>Solution:</p> <p>Mean of the population $\mu = 60$ kg Standard deviation of the population = 10 kg sample size $n = 50$ (formula 1 mark; values each one mark)</p> <p>Mean of the sample is given by: $\mu_{\bar{x}} = \mu = 60$ kg</p> <p>Standard deviation of the sample is given by:</p> <p>$S_{\bar{x}} = \sigma / \sqrt{n}$ $S_{\bar{x}} = 10 / \sqrt{50}$ $S_{\bar{x}} = 1.414 = 1.4$ kg (approx)</p>	3
6	a)	<p>Leakage from underground fuel tanks has been a source of water pollution. In a random sample of 87 gasoline stations, 13 were found to have at least one leaking underground tank. How many stations must be sampled so that a 95% confidence interval specifies the proportion to within ± 0.03?</p> <p>Solution:</p> <p>Given $n = 87$; $X = 13$; $\text{MoE} = 0.03$ $\hat{p} = (X + 2) / (n + 4) = 15 / 91 = 0.1648$ (1 mark) $\text{MoE} = Z_{\alpha/2} * \sqrt{\hat{p} * (1 - \hat{p}) / (n + 4)}$ (1 mark) $0.03 = 1.96 * \sqrt{(0.1648 * 0.8352) / (n + 4)}$ $n + 4 = 587.33$ (1 mark) $n = 583.33$ Therefore, $n = 584$ (1 mark)</p>	4
	b)	<p>City planners wish to estimate the mean lifetime of the most commonly planted trees in urban settings. A sample of 16 recently felled trees yielded mean age 32.7 years with standard deviation 3.1 years. Assuming the lifetimes of all such trees are normally distributed, construct a 99.8% confidence interval for the mean lifetime of all such trees.</p> <p>Solution:</p> <p>Given $n = 16$; sample mean = 32.7 ; $s = 3.1$; $\text{CL} = 99.8\%$ CI is given by $\bar{X} \pm t_{n-1, \alpha/2} * s / \sqrt{n} = 32.7 \pm t_{15, .001} * 3.1 / \sqrt{16}$ (1 mark) $= 32.7 \pm 3.733 * .775$ (1 mark) Therefore, 99.8% CI is 32.7 ± 2.8931 (OR) (29.8069, 35.5931) (1 mark)</p>	3
	c)	Construct a confidence interval for the given data.	3

	<p>Mean of differences is 19.4, Standard Deviation of differences is 2.836273, sample size is 10 and significance level is 0.05.</p> <p>Solution:</p> <p>95% CI is given by $\bar{D} \pm t_{n-1, \alpha/2} * s_D / \sqrt{n}$ (1 mark)</p> $= 19.4 \pm 2.262 * 2.836273 / \sqrt{10}$ $= 19.4 \pm 2.029 \quad (1 \text{ mark})$ <p>Therefore, CI is (17.371 , 21.429) (1 mark)</p>	
--	---	--