SRN													

PES University, Bangalore

(Established under Karnataka Act No. 16 of 2013) **UE17CS203**

END SEMESTER ASSESSMENT (ESA) - B.TECH III SEMESTER – DEC. 2018

UE17CS203 - INTRODUCTION TO DATA SCIENCE

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

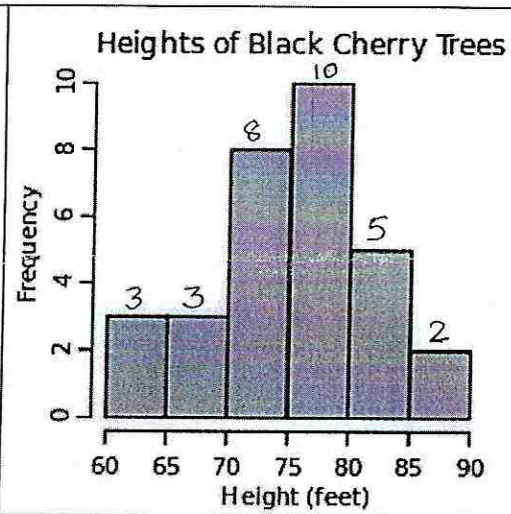
Note:

- **Answer all questions in order and to the point.**
- **IDS Formula sheet must be provided to all the students.**

1	a	I.	A shoe store had a special on women's running shoes and sold the following sizes : 6,10,4,7,8,7,6,5,7,8,7,5,6,7,3,7 Which of the mean, median, mode would be of most use to the shop owner?	6 (2 + 2 + 2)
		II.	Data on income are stored as integers, with 1 standing for the range under Rs. 50k, 2 for Rs. 50k to Rs. 80k and 3 for over Rs. 80k. True or False : "This income data is quantitative". Justify your answer.	
		III.	Suppose we collected purchase data consisting of transaction id, the purchase amount, and the time of day. If we wanted to create a visualization to explore the **purchase behavior,** which of the following plots would likely be helpful? Choose all that apply. [You earn full credit if and only if all right options are picked] A. a bar plot of the amount for each transaction id B. a scatter plot of purchase amount and time of day C. a bar plot with the purchase for each time of day D. a bar plot with total purchase amount aggregated over each hour of the day.	
	b	I.	Provide 2 examples which emphasize on the importance of Cleaning data.	6 (2 + 2 + 2)
		II.	Your Data Science course Professor gave your team a task to estimate the average amount of money spent per student on extracurricular activities in Engineering colleges of Bangalore Region. One of your teammates suggested to randomly select 10 Engineering colleges from Bangalore and all of the students in each of the 10 colleges will be selected to frame the sample. Name the sampling method suggested by your friend.	
		III.	A group of 15 adult females are training to run/walk a 5k as a team. For their first 5k run together the summary statistics were as follows: Median = 36 minutes, Mean = 30 minutes, Standard Deviation = 3 minutes After a month of training each of their individual times decreased by exactly 2 minutes. What is the groups mean and standard deviation after a month of training?	
	c	Consider the following statistics for x, which is infant mortality rate for 200 countries. According to these, which transformation would symmetrize the distribution? Justify your answer.		4 (2 + 2)

Transformation	lower quartile	median	upper quartile
x	13	30	68
$\sqrt{x}$	3.5	5	8
log(x)	1.15	1.5	1.8

SRN												

	d	 For the given histogram of 31 values, answer the following : I. Use the histogram to determine the class that contains Quartile 1 (Q1) value. Explain your answer. II. Use the histogram to determine the class that contains Quartile 3 (Q3) value. Explain your answer. III. We may obtain an estimate of Q1 by computing the midpoint of the class that you found in part I. Similarly, we may obtain an estimate of Q3 by computing the midpoint of the class found in part II. Approximate the interquartile range of scores using the estimated values of Q1 and Q3.	4 (1 + 1 + 2)
2	a	**I.** Suppose X, Y , and Z are random variables that are independent and have the same probability distribution. If $Var(X) = \sigma^2$, then What is $Var(X + Y + Z)$? **II.** Suppose the random variable X can take on values −1, 0, and 1 with chance p^2, $2p(1-p)$ and $(1-p)^2$, respectively, for $0 \le p \le 1$. What is the expected value of X? **III.** A sample of 100 sociology books has a mean cost of Rs. 76.75 and a standard deviation of Rs. 10.42. A. Use this information and Chebyshev's Rule to complete the following statement : "*Approximately ____ of the 100 sociology books cost between Rs. 45.49 and Rs. 108.01*"	6 (2 + 2 + 2)
	b	Does the following sample appear to come from a Normal population? Justify your answer by constructing a Normal Probability Plot: 100, 98, 101, 93, 123	6
	c	If only 5 percent students can secure S grade in IDS paper, find the probability of at most 2 out of 10 students securing S grade in that paper.	4
	d	Answer the following: **I.** A continuous random variable X is distributed by a normal probability distribution. What is $P(X \neq 3)$? **II.** What is the percentage of a normal distribution that is less than the mean? **III.** What value of Z corresponds to Q1 and Q3?	4 (1 + 1+ 2)
3	a	**I.** Assuming the population standard deviation $\sigma = 3$, how large should a sample be to estimate the population mean μ with a margin of error not exceeding 0.5? **II.** Under what conditions does the Central Limit Theorem permits us to assume the sample means are normally distributed? [Must specify 2 conditions] **III.** How is mean squared error of an estimate calculated?	6 (3+2+1)
	b	Suppose 20 donors come to a blood drive. Assume that the blood donors are not related in any way, so that we can consider them independent. The probability that the donor has type-O blood is 0.06, which is constant from donor to donor. Let X = the number of donors that have type-O blood. **I.** For the sample of 100 donors, what is the sampling distribution of the sample proportion? **II.** For a sample of 300 donors, what is the sampling distribution of the sample proportion?	6 (2+2+2)

SRN													

		III. For the sample of 300 donors, what is the probability that the sample proportion is greater than 0.10?	
	c	The following are summary statistics for a data set. Would it be appropriate to use the Student's t distribution to construct a confidence interval from these data? Explain. N = 12, mean = 62.21, median = 50, sd = 41.37, min = 7.5, max = 165, Q1 = 36.5, Q3 = 78	4
	d	Let X1,..., Xn be a random sample from a population with the Poisson(λ) distribution. Find the MLE of λ.	4
4	a	Suppose in a sample of 45 people, the mean height $\bar{x}$ was observed to be 87 inches and σ was found to be 4.5. **I.** Construct a 95% confidence interval for true value of height (μ). **II.** Would you reject the hypothesis H(0): $\mu = 88$ versus H(1): $\mu \neq 88$ on the basis of the observations, when testing at level $\alpha = 0.05$? **III.** Would you reject the hypothesis H(0): $\mu = 90$ versus the alternative H(1): $\mu \neq 90$ on the basis of the observations, when testing at level $\alpha = 0.05$?	6 (2 + 2 + 2)
	b	Bags of a certain brand of Potato chips claim to have a net weight of 14 ounces. The net weights actually vary slightly from bag to bag and are normally distributed with mean μ. A representative of a consumer advocacy group wishes to see if there is any evidence that the mean net weight is less than advertised. For this, the representative randomly selects 40 bags of this brand and determines the net weight of each. He finds the sample mean to be X = 13.82 and the sample standard deviation to be s = 0.24. Use these data to perform an test of hypothesis of **H0:μ >=14 vs H1 : μ<14** at 5% significance level. Answer the following : **I.** Specify Null distribution. **II.** Specify the test statistic value. **III.** Specify the P-value. **IV.** Provide your conclusion with proper justification.	6 (1 + 2 +1+2)
	c	**I.** A random sample of 64 adult wolf in a region of Maharashtra showed the average age to be 5.85 years, with a sample standard deviation of 0.88 years. However, it is thought that the overall population mean age of wolf is 5.45. You would like to test if the sample data indicate that wolf in this region tend to live longer than the average of 5.45 years. What is the appropriate rejection region for this hypothesis test at the 0.05 significance level? **II.** On November 1, 2018 the ABP News-CSDS polling agency reported that a 95% confidence interval for the true proportion of Indian adults in favor of the way Narendra Modi is handling his job as Prime Minister is (0.34, 0.42). Report the margin of error for this confidence interval.	4 (2 + 2)
	d	According to the World Health Organization, 20% of adult Indians experience migraine headaches. Stress is a major contributor to the frequency and intensity of headaches. A massage therapist believes that he has a technique that can reduce the frequency and intensity of migraine headaches. The following hypotheses are used to test the effectiveness of the massage therapist's claim. **H(0):** The true proportion of adults who experience migraine headaches after massage therapy is 0. **H(1):** The true proportion of adults who experience migraine headaches after massage therapy is less than 0.2 Answer the following: **I.** Describe Type 1 error in the context of this situation. **II.** Describe a consequence of Type I error in this situation. **III.** Describe Type II error in the context of this situation. **IV.** Describe a consequence of Type II error in this situation.	4 (1+ 1 + 1+1)

SRN													

5	a	Rachel told Eric that the reason her car insurance is less expensive is that female drivers get in fewer accidents than male drivers. Specifically, she says that male drivers are held responsible in 65% of accidents involving drivers under 23. If Eric does some research of his own and discovers that 46 out of the 85 accidents he investigates involve male drivers. With the help of Chi square goodness-of-fit test, answer whether Eric's data support or refute Rachel's hypothesis? Clearly state the null and alternative hypotheses.	6
	b	Monitoring the yield of a particular chemical reaction at various reaction vessel temperatures produces the results shown in the given table. The equation of the least squares line is : **predicted_yield = 60.26 + 0.1165 * Temp**. Make a residual plot. Does the linear model seem appropriate?	6 (5 + 1)
	c	Test of H0 : $\mu = 4$ verses H1 : μ != 4 at Alpha = 0.05. Assume population standard deviation to be 1.5. Under the alternative μ=5.5, the power was calculated to be 0.85, using sample size of 18. Answer the following: **I.** If the sample size were 25, will the probability of making **type II error** increase, decrease or remain the same? **II.** If the power was calculated for alternative 5 instead of 5.5,would the **power** be less than 0.85, greater than 0.85, or is it impossible to tell from the output? **III.** If the test was made at alpha = 0.10 instead of 0.05,would the **power** be less than 0.85, greater than 0.85, or is it impossible to tell from the output? **IV.** If the population standard deviation is changed to 3 instead of 1.5, will the probability of making **type II** error increase, decrease or remain the same?	4 (1 + 1 +1+ 1)
	d	**I.** Suppose a 95% confidence interval for the slope (BETA_1) of the straight line regression of Y on X is given by -3.5 < BETA_1 < -0.5. True or False : "*A two-sided test of the hypothesis H(0): BETA_1 = -1 would result in rejection of H(0) at the 1% level of significance*". Justify your answer. **II.** A researcher finds that the correlation between the personality traits "greed" and "superciliousness" is -0.40. What percentage of the variation in greed can be explained by the relationship with superciliousness? **III.** True or False : If the correlation coefficient is negative, then below-average values of one variable are associated with below-average values of the other.	4 (2 + 1 + 1)

Temp. (°C)	Yield (%)	Temp. (°C)	Yield (%)
150	77.4	250	88.9
150	76.7	250	89.2
150	78.2	250	89.7
200	84.1	300	94.8
200	84.5	300	94.7
200	83.7	300	95.9