Name:

Enrolment No:

UPES
UNIVERSITY WITH A PURPOSE

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
End Semester Examination, December 2021

Course: Applied Statistical Analysis **Semester: 3rd**
Program: B.Tech CSE (BAO) **Time 03 hrs.**
Course Code: CSBA2009 **Max. Marks: 100**

Instructions: All questions are compulsory

SECTION A

S. No.		Marks	CO
Q 1	For the marks of 25 students studying BAO : 20, 21, 19, 18, 20, 20, 19, 18, 21, 19, 22, 21, 18, 19, 21, 22, 19, 18, 20, 19, 20, 22, 20, 21, 20. Discuss the discrete frequency distribution.	**4**	**CO1**
Q 2	You got a dataset depicting the popularity of two graphic novels given by a critic which contains three variables. 1) Time of survey (in dd-mm-yy format) 2) Rating of 'Marvel' (in range between 0 to 10) 3) Rating of 'DC' (in range between 0 to 10) The data is collected every day since 1970. You need to graphically represent the data in a chart. What will you use? And why?	**4**	**CO2**
Q 3	Suppose we visited a car shop and the following data is provided by the owner: **CAR NO** / **TYPE** 1 / Fast 2 / Fast 3 / Very Fast 4 / Very Slow 5 / Slow 6 / 7 / Fast Discuss the dataset (which type). Also compute the missing value for Car No 6. Explain the methodology you used and why. Furthermore discuss the asymptotic time and space analysis of the method used.	**4**	**CO1**
Q 4	Three athletes A, B and C are participating in the Olympics. A is twice as likely to win as B and B is twice as likely to win as C. What are the probabilities of their winning?	**4**	**CO2**
Q 5	A new flag of Informatics club is to be designed with 5 vertical strips using some or all of the four colours: green, maroon, red and yellow. In how many ways this can be done so that no two adjacent strips have the same colour?	**4**	**CO2**

SECTION B

(Option given in Question 8)

Q 6 — **10** — **CO4**

Today Vivek marked the attendance of students with respect to time which gradually increased as the lecture progressed (positively correlated). He found the following observations:

Time Stamp	No. of Students in class
0	1
1	3
2	4
3	5
4	6
5	9

Vivek happens to a number freak and wants to find an equation for the observed data. He loves the second degree parabolic equation: $Y = a + bX + cX^2$

Help Vivek to formulate the equation.

Q 7 — **10** — **CO2**

During this latest household survey of Dehradun; we came to know the ages of husbands and wives of each family. A sample is provided to you:

Age of Husband	Age of Wife
23	18
27	22
28	23
29	24
30	25
31	26

Compute the correlation coefficient and discuss what the coefficient suggests about the data.

Q 8 — **10** — **CO4**

Table 4.1. Data set for Exercise 2.

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

For the data give in the table find 'Class' given the following information:

Customer ID	Gender	Car Type	Shirt Size	Class
21	M	Sports	Medium	?

	OR In a statistical study relating to the prices (in Rs.) of two shares, X and Y, the following two regression lines were found as 8X – 10Y +70 = 0 20X – 9Y – 65 = 0 The standard deviation of X = 3. Compute i) The values of mean of X and mean of Y, ii) R(X,Y) iii) Standard deviation of Y.		
Q 9	Ashish and Sudhanshu (two best of friends) on phone regularly as they are living in separate cities. Based on Sudhanshu's mood Ashish wants to predict the weather of the city Sudhanshu is currently in. For this purpose, Ashish collected the data below:  If Sudhanshu didn't pickup the phone today; help Ashish to predict the weather. Also if Sudhanshu is Happy today, what is the probability that its Sunny or Rainy?	**10**	**CO4**
SECTION-C (Option given in Question 10 and 11)			
Q 10	If the cost function for a classification method is taken as a tanh function instead of a simple sigmoid function. The tanh function is defined as: $$\tanh(x) = \frac{2}{1+e^{-x}} - 1$$ Compare this new function with the sigmoid function. Find the gradient for each functions. **OR** The variance of a certain dimension article produced by a machine is 7.2 over a long period. A random sample of 20 articles gave a variance 8. Is it justifiable to conclude that variability has increased at 5% level of significance assuming that the measurement of dimension article is normally distributed? For reference:	**20**	**CO3**

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892

Q 11	An urn contains either 4 white and 2 black balls or 2 white and 4 black balls. Two balls are to be drawn from the urn. If less than two white balls are obtained, it will be decided that this urn contains 2 white and 4 black balls. a) Compute the Null and Alternate Hypothesis. b) Compute the critical region. c) Compute both sizes of type-I and type II error. d) Is it a two-tailed or one-tailed test? e) How will you go forward to test the hypothesis? **OR** A big company uses thousands of CFL lights every year. The brand that the company has been using in the past has average life of 1200 hours. A new brand is offered to the company at a price lower than they are paying for the old brand. Consequently, a sample of 100 CFL light of new brand is tested which yields	**20**	**CO3**

an average life of 1220 hours with standard deviation 90 hours. Should the company accept the new brand at 5% level of significance? (Refer table below)

Level of Significance (α)	Two-Tailed Test	One-Tailed Test	
		Right-Tailed Test	Left- Tailed Test
$\alpha = 0.05$ (= 5%)	$\pm z_{\alpha/2} = \pm 1.96$	$z_\alpha = 1.645$	$-z_\alpha = -1.645$
$\alpha = 0.01$ (= 1%)	$\pm z_{\alpha/2} = \pm 2.58$	$z_\alpha = 2.33$	$-z_\alpha = -2.33$