Roll No.: _

Amrita Vishwa Vidyapeetham

Amrita School of Computing, Coimbatore

B.Tech Second Assessment Examinations – Dec 2022

Fifth Semester

Computer Science and Engineering

19CSE304 Foundations of Data Science

Duration: Two hours Maximum: 50Marks

Course Outcomes (COs):

CO	**Course Outcomes**
CO01	Understand the statistical foundations of data science.
CO02	Apply pre-processing techniques over raw data so as to enable further analysis.
CO03	Conduct exploratory data analysis and create insightful visualizations to identify patterns.
CO04	Identify machine learning algorithms for prediction/classification and to derive insights
CO05	Analyse the degree of certainty of predictions using statistical test and models

Answer all questions

Refer Appendix for the relevant Tables of information

1. Given a one-dimensional data set **X = {8.0, 204.0, 17.0, 9.0, 4.0}.** Find the normalized value for **24.0** using the following techniques: [3] [CO01] [BTL 3]
   (a) Decimal scaling on interval [0, 1]. (1mark)
   (b) Min-max normalization on interval [0, 1]. (1mark)
   (c) Z-Score normalization. (1 mark)

2. (a) When is a random sample of size **n** stated to be independent and identically distributed? (1mark) [4] [CO05] [BTL 2]
   (b) State the Central Limit Theorem (1 mark).
   (c) Illustrate by means of a sketch, the outcome of application of Central Limit theorem when the sampling distribution is derived from a (i) Normal distribution, and (ii) Exponential distribution. (2 marks)

3. (a) State the difference between a parameter and a statistic. (1mark)
   (b) How would you justify the use of large random samples in statistical inference? (1mark)
   (c) What are empirical distributions? (1 mark) [3] [CO05] [BTL 2]

4. (a) State the types of errors in statistical hypothesis testing (1mark). [3] [CO05] [BTL 2]
   (b) Under what conditions can a hypothesis test be wrong (2marks)?

5 (a) What do you understand by “p value” (1mark)? [3] [CO05] [BTL 3]
   (b) When would you consider the “p-value” as “highly statistically significant” (1/2mark)?
   (c) What does a “p-value” of .001 mean(1/2 mark)?
   (d) What is the Type I “error probability” that can be observed in hypothesis tests (1mark)?

6. (a) If 3% of electronic units manufactured by a company are defective, find the probability that in a sample of 200 units, less than 2 bulbs are defective. (3 marks) [6] [CO01] [BTL 3]
   (b) Assume that, you usually get 2 phone calls per hour. Calculate the probability that a phone call will come within the next hour. (3 marks)

7. AT&T argues that its rates are such that customers won't see a difference in their phone bills between them and their competitors. They calculate the mean and standard deviation for all their customers at $17.09 and $3.87 (respectively). They then sample 100 customers at random and recalculate a monthly phone bill based on competitor's rates. The results reveal a mean of $17.55.
Use a 0.05 level of significance to verify the claim of AT&T. [10] [CO05] [BTL 4]
(a) State the null and alternative hypotheses (2marks)
(b) What is the Type I error probability in percentage (1mark)?
(c) What are the z values of the rejection region(s)? (2mark)
(d) What is the test statistic? (1mark)
(e) State the outcome of the hypothesis test based on your analysis (4 marks)

8. According to a 2020 demographic report, the average Tamil Nadu household spends INR 90 per day on cereals. A random sample of 30 households in a certain village revealed a mean of INR 84.50. Assume the standard deviation is known to be INR 14.50. The objective is to ascertain, using a 0.05 level of significance, if the average amount spent on cereals per day by households in Tamil Nadu has decreased. Answer the following: [5] [CO05] [BTL 4]
(a) State the Null and Alternative Hypothesis (1 mark)
(b) What is your test statistic? (1mark)
© Evaluate the margin of error or maximum error of the estimate (1 mark)
(d) State the outcome of the hypothesis test based on your analysis (2 marks)

9. A company wants to test the claim that their batteries last more than 40 hours. Using a simple random sample of 15 batteries yielded a mean of 44.9 hours, with a standard deviation of 8.9 hours. This claim has to be tested using a significance level of 0.05. Answer the following:
[7] [CO05] [BTL 4]
(a) State the type of test you would invoke in this case and justify (2 mark).
(b) State the Null and Alternative Hypothesis (1 mark)
(c) What is your test statistic? (1mark)
(d) Evaluate the p value approximately (1 mark).
(e) State the outcome of the hypothesis test based on your analysis (2 marks)

10. Consider the data regarding "*handedness*" of a sample population of American/ Canadian persons. Americans: Right Handedness is 236 and Left Handedness is 19; and in the case of Canadians Right Handedness is 157 and Left Handedness is 16. The objective is to carry out an appropriate test of independence to examine whether these observed frequencies are significantly different from the frequencies expected if *handedness* is unrelated to nationality. Consider sig level as 0.05.
[6] [CO05] [BTL 4]
(a) State the type of test you would invoke in this case and justify(1 mark).
(b) State your Null and Alternate Hypothesis (1 mark).
(c) Evaluate the test statistic (2 marks).
(d) State the outcome of the hypothesis test based on your analysis (2 marks)

*****

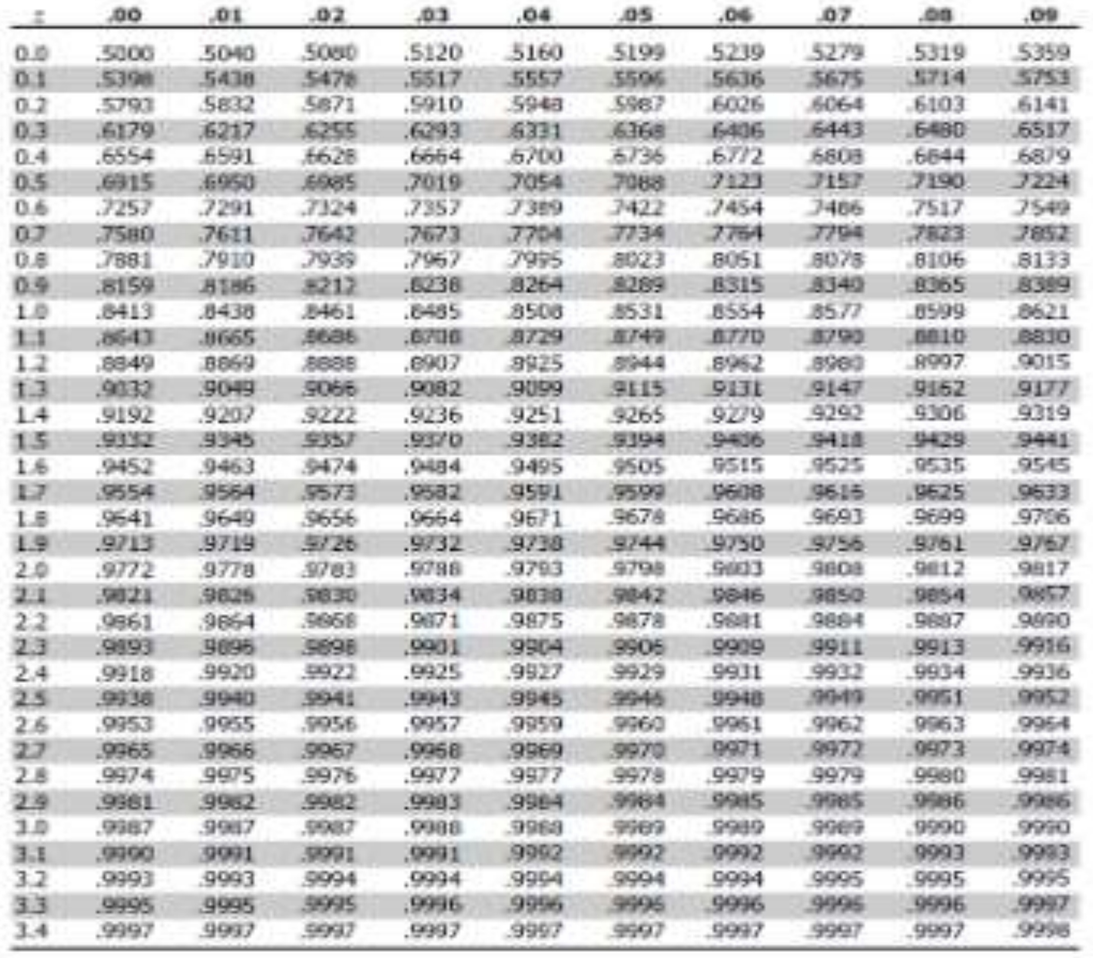
Course Outcome /Bloom's Taxonomy Level (BTL) Mark Distribution Table

CO	Marks	BTL	Marks
CO01	**9**	BTL 1	**-**
CO02	**-**	BTL 2	**10**
CO03	**-**	BTL 3	**12**
CO04	**-**	BTL 4	**28**
CO05	**41**	BTL 5	**-**

Appendix

Relevant Tables for probability Distributions

Z score Table



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

T Distribution Table

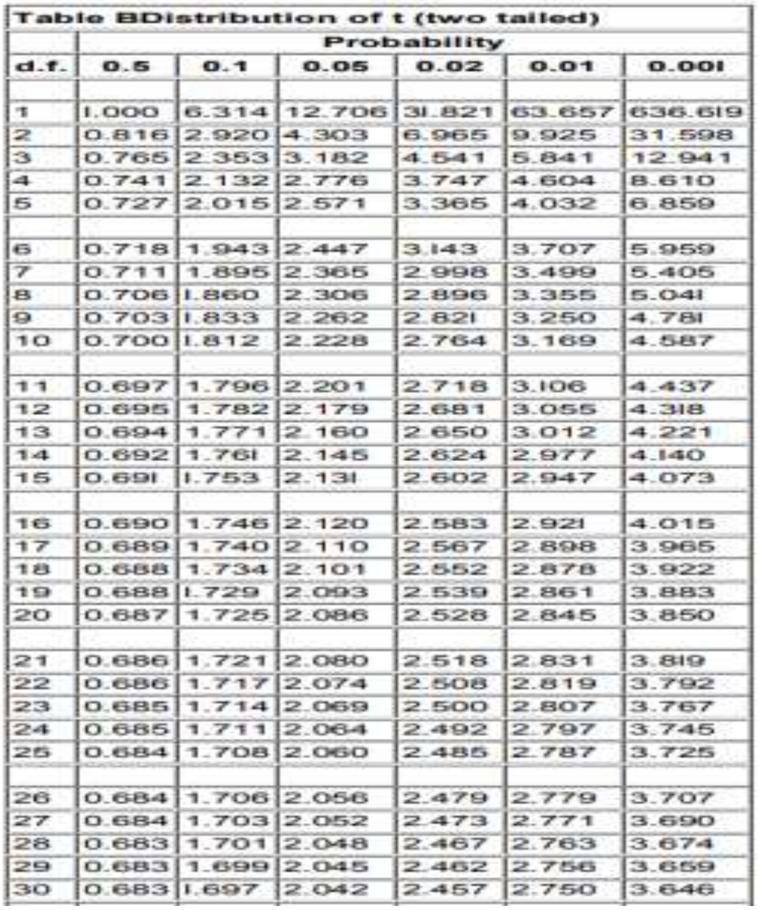
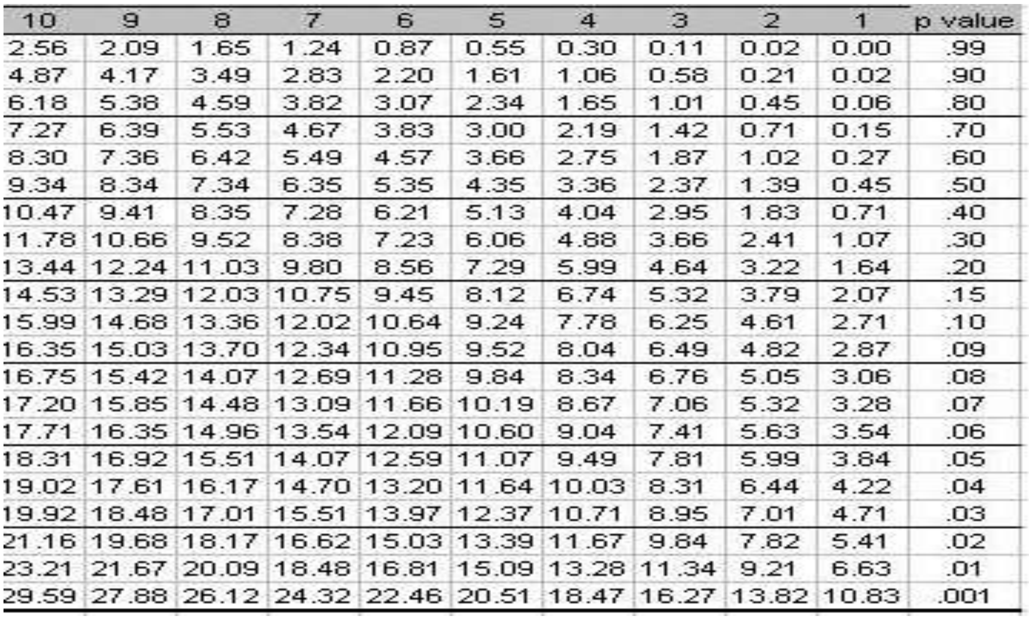


Table BDistribution of t (two tailed)						
	Probability					
d.f.	0.5	0.1	0.05	0.02	0.01	0.00I
1	I.000	6.314	12.706	3I.821	63.657	636.6I9
2	0.816	2.920	4.303	6.965	9.925	31.598
3	0.765	2.353	3.182	4.541	5.841	12.941
4	0.741	2.132	2.776	3.747	4.604	8.610
5	0.727	2.015	2.571	3.365	4.032	6.859
6	0.718	1.943	2.447	3.I43	3.707	5.959
7	0.711	1.895	2.365	2.998	3.499	5.405
8	0.706	I.860	2.306	2.896	3.355	5.04I
9	0.703	I.833	2.262	2.82I	3.250	4.78I
10	0.700	I.812	2.228	2.764	3.169	4.587
11	0.697	1.796	2.201	2.718	3.I06	4.437
12	0.695	1.782	2.179	2.681	3.055	4.3I8
13	0.694	1.771	2.160	2.650	3.012	4.221
14	0.692	1.76I	2.145	2.624	2.977	4.I40
15	0.69I	I.753	2.13I	2.602	2.947	4.073
16	0.690	1.746	2.120	2.583	2.92I	4.015
17	0.689	1.740	2.110	2.567	2.898	3.965
18	0.688	1.734	2.101	2.552	2.878	3.922
19	0.688	I.729	2.093	2.539	2.861	3.883
20	0.687	1.725	2.086	2.528	2.845	3.850
21	0.686	1.721	2.080	2.518	2.831	3.8I9
22	0.686	1.717	2.074	2.508	2.819	3.792
23	0.685	1.714	2.069	2.500	2.807	3.767
24	0.685	1.711	2.064	2.492	2.797	3.745
25	0.684	1.708	2.060	2.485	2.787	3.725
26	0.684	1.706	2.056	2.479	2.779	3.707
27	0.684	1.703	2.052	2.473	2.771	3.690
28	0.683	1.701	2.048	2.467	2.763	3.674
29	0.683	1.699	2.045	2.462	2.756	3.659
30	0.683	I.697	2.042	2.457	2.750	3.646

Chi Square Table



10	9	8	7	6	5	4	3	2	1	p value
2.56	2.09	1.65	1.24	0.87	0.55	0.30	0.11	0.02	0.00	.99
4.87	4.17	3.49	2.83	2.20	1.61	1.06	0.58	0.21	0.02	.90
6.18	5.38	4.59	3.82	3.07	2.34	1.65	1.01	0.45	0.06	.80
7.27	6.39	5.53	4.67	3.83	3.00	2.19	1.42	0.71	0.15	.70
8.30	7.36	6.42	5.49	4.57	3.66	2.75	1.87	1.02	0.27	.60
9.34	8.34	7.34	6.35	5.35	4.35	3.36	2.37	1.39	0.45	.50
10.47	9.41	8.35	7.28	6.21	5.13	4.04	2.95	1.83	0.71	.40
11.78	10.66	9.52	8.38	7.23	6.06	4.88	3.66	2.41	1.07	.30
13.44	12.24	11.03	9.80	8.56	7.29	5.99	4.64	3.22	1.64	.20
14.53	13.29	12.03	10.75	9.45	8.12	6.74	5.32	3.79	2.07	.15
15.99	14.68	13.36	12.02	10.64	9.24	7.78	6.25	4.61	2.71	.10
16.35	15.03	13.70	12.34	10.95	9.52	8.04	6.49	4.82	2.87	.09
16.75	15.42	14.07	12.69	11.28	9.84	8.34	6.76	5.05	3.06	.08
17.20	15.85	14.48	13.09	11.66	10.19	8.67	7.06	5.32	3.28	.07
17.71	16.35	14.96	13.54	12.09	10.60	9.04	7.41	5.63	3.54	.06
18.31	16.92	15.51	14.07	12.59	11.07	9.49	7.81	5.99	3.84	.05
19.02	17.61	16.17	14.70	13.20	11.64	10.03	8.31	6.44	4.22	.04
19.92	18.48	17.01	15.51	13.97	12.37	10.71	8.95	7.01	4.71	.03
21.16	19.68	18.17	16.62	15.03	13.39	11.67	9.84	7.82	5.41	.02
23.21	21.67	20.09	18.48	16.81	15.09	13.28	11.34	9.21	6.63	.01
29.59	27.88	26.12	24.32	22.46	20.51	18.47	16.27	13.82	10.83	.001

==