

Amrita Vishwa Vidyapeetham
Amrita School of Computing, Coimbatore
B.Tech End Examinations – December 2023

Fifth Semester
Computer Science and Engineering
19CSE304 Foundations of Data Science

Duration: Three hours

Maximum: 100 Marks

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.(a) Identify the sampling study in the following cases. Justify your answers.

(i) To determine the average salary of Professors at ABC University, the faculty where divided into the following groups: *Faculty Associate, Assistant Professors, Associate Professors and Professors*. 20 faculty members from each group were selected for the study. [2][CO01][BTL2]

(ii) At a computer facility, every 100th chip is inspected for defects. [2][CO01][BTL2]

(iii) To check the proportion of international students, the admission office takes stock of the number of Asian, European, African and American students. The frequency of each class is about the same. The Office selects 80 people from the population. [2][CO01][BTL2]

(b) A graduate student at the University conducts a research project about communication in India. Se mails a survey to all of the 500 adults that she knows. She asks them to mail back a response to this question: "Do you prefer to use email or snail mail (Department of Posts)?" She gets back 65 responses, with 22 of them indicating a preference for snail mail. She then reported that $22/65 = 33.8\%$ of adults prefer to use snail mail. Identify the parameter, the sample and the statistic in this problem. [2][CO01][BTL2]

(c) Astronomers typically determine the distance to galaxy (a galaxy is a large collection of billions of stars) by measuring the distances to just a few stars within it and taking the mean (average) of these distance measurements. Identify the population, sample, population parameters, and sample statistics. [2][CO01][BTL2]

2. Classify the following attributes as binary, discrete or continuous. In addition, classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio).

(a) Angles as measured in degrees between 0° and 360°. [2][CO01][BTL2]

(b) Bronze, Silver, and Gold medals as awarded at the Olympics. [2][CO01][BTL2]

(c) Number of patients in a hospital. [2][CO01][BTL2]

(d) Military ranks. [2][CO01][BTL2]

(e) Height above sea level. [2][CO01][BTL2]

Note: Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity. Binary/Discrete/Continuous identification - 1 mark, and Qualitative/Quantitative - 0.5 mark

3. For a given application, the burning rate of a solid propellant should be 40 cm/sec. 36 samples of solid propellant are taken and their burning rate is computed to be 41.3 cm/sec. The standard deviation in the burning rate is known to be 3 cm/sec. If the significance level is 0.05, then answer the following questions.

- State the type of test you would invoke in this case with justification. [1][CO05][BTL4]
- State the Null and Alternative Hypotheses. [1][CO05][BTL4]
- What is your population parameter, and the value of the test statistic? [2][CO05][BTL4]
- What is the critical value for your test, based on the significance level? [2][CO05][BTL4]
- Justify the outcome of the hypothesis test based on your analysis. [2][CO05][BTL4]
- For what maximum value of the sample burning rate, you can change the decision that you made based on your hypothesis testing? [2][CO05][BTL4]

4. The specimen of copper wires drawn from a large lot have the following breaking strength (in kg wt): **578, 572, 570, 568, 572, 578, 570, 572, 596, 544**. Test whether the mean breaking strength of the lot may be taken to be **578 kg wt**. Test at 5 per cent level of significance.

- State the type of test you would invoke in this case with proper logic. [2][CO05][BTL4]
- State the null and alternative hypothesis. [2][CO05][BTL4]
- What is your test statistic? [4][CO05][BTL4]
- State the outcome of the hypothesis test based on your analysis. [2][CO05][BTL4]

5. Answer the following questions, based on the results depicted in the Summary table of a regression analysis given below:

OLS Regression Results						
Dep. Variable:	Sales	R-squared:	0.612			
Model:	OLS	Adj. R-squared:	0.610			
Method:	Least Squares	F-statistic:	312.1			
Date:	Wed, 28 Dec 2022	Prob (F-statistic):	1.47e-42			
Time:	09:48:39	Log-Likelihood:	-519.05			
No. Observations:	200	AIC:	1042.			
Df Residuals:	198	BIC:	1049.			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	7.0326	0.458	15.360	0.000	6.130	7.935
TV	0.0475	0.003	17.668	0.000	0.042	0.053
Omnibus:	0.531	Durbin-Watson:	1.935			
Prob(Omnibus):	0.767	Jarque-Bera (JB):	0.669			
Skew:	-0.089	Prob(JB):	0.716			
Kurtosis:	2.779	Cond. No.	338.			

- State the equation of the regression line, and comment on its slope [3][CO04][BTL3]
- Briefly explain the significance of the following parameters along with your comments on the results obtained:-
 - R-square value. [2][CO04][BTL3]
 - Adjusted R-square value. [1][CO04][BTL3]
 - Kurtosis [1][CO04][BTL3]
 - t-values and $P(> |t|)$ obtained [3][CO04][BTL3]

- 6.
- (a) What do you understand by the term correlation coefficient? [1][CO04][BTL2]
 - (b) State the salient properties of correlation coefficient. [2][CO04][BTL2]
 - (c) Does correlation measure causation? Justify. [1][CO04][BTL2]
 - (d) Given: Two variables x and y are measured in standard units. State the regression equation for predicting y based on its dependent variable x , and the regression coefficient r [2][CO04][BTL2]
 - (e) What are residuals? [1][CO04][BTL2]
 - (f) Sketch an illustrative plot of the residuals when (one mark each):- [3][CO04][BTL2]
 - (i) Case1: The residual plot is of a good regression.
 - (ii) Case2: Non-linearity is present in the data.
 - (iii) Case3: Heteroscedasticity is manifest.
7. Virgin coconut oil was considered as a possible medicine to treat cancer. 50 patients were subject to trials to establish the efficacy of this antidote. 30 patients were chosen at random for the treatment group and the balance 20 earmarked as control group. The trial lasted for 45 days: the treatment group was administered the prospective drug 10ml per day while saline solution was given to the control group. On the 46th day, 15 persons from the treatment group showed improvement as against 4 from the control group.
- (a) Under what broad category of experiment does this drug trial fit in? [1][CO05][BTL3]
 - (b) (i) Explain the statistical methodology to validate the efficacy of this experiment. [1][CO05][BTL3]
 - (ii) State the Null and the Alternative clearly. [1][CO05][BTL3]
 - (iii) How do you propose to predict the Statistic Under the Null? [1][CO05][BTL3]
 - (c) State the “*observed statistic*” in the statistical test. [2][CO05][BTL3]
 - (d) What do you understand by the term “*confounding factors*”? [2][CO05][BTL3]
 - (e) State a technique to avoid confounding. [2][CO05][BTL3]
8. The City Corporation is in the process of estimating the median income of approximately one lakh female residents eligible for allowance of the newly introduced monthly family pension. Owing to resource and time constraints, data can be ascertained only for a representative sample of about 10,000 eligible individuals. A methodology is required to be proposed for inference of this parameter (median income) in respect of the population under consideration.
- (a) Explain a suitable methodology to estimate the unknown parameter of median income (3marks). Illustrate by means of a sketch(1mark). [4][CO05][BTL3]
 - (b) State the statistic. [2][CO05][BTL3]
 - (c) How many samples do you propose to draw at a time? Would it be done “with or without replacement”? Justify. [2][CO05][BTL3]
 - (d) How can you obtain an 80% confidence interval? [2][CO05][BTL3]

9. For this cricket world cup, balls were sourced from three different places namely Chennai, Kolkatta and Mumbai. Some of the balls were faulty. Three types of faults were reported. The faults were classified in to three types A, B and C. The data is given in the table below.

	Fault Type		
Source	A	B	C
Chennai	40	28	34
Kolkatta	27	39	32
Mumbai	45	26	29

Conduct a Chi-square test at the 5% level of significance to determine whether fault type is related to the source from which ball was procured. State the hypotheses and infer the result.

10. Suppose we have 2 classifiers, M1 and M2, employed to carry out a 5-fold cross-validation on a dataset. It may seem intuitive to select the model with the lowest error rate; however, these mean error rates are just estimates of error on the true population of future data cases. There can be considerable variance between error rates within any given 5-fold cross-validation experiment.

Although the mean error rates obtained for M1 and M2 may appear different, that difference may not be statistically significant, the difference between the 2 error rates could be just attributed to chance. Use a test of statistical significance and evaluate the two models, M1 and M2 with their Mean Square Error(MSE) tabulated below. Take significance level, $\text{sig} = 0.05$.

Srl No	MSE(M1)	MSE(M2)
1	1	2
2	3	4
3	4	2
4	2	4
5	1	5

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

CO	Marks	BTL	Marks
CO01	20	BTL 1	-
CO02	-	BTL 2	30
CO03	-	BTL 3	30
CO04	30	BTL 4	40
CO05	50	BTL 5	-

Relevant Tables for probability Distributions**Chi Square Table**

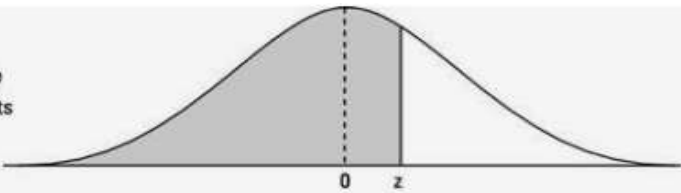
10	9	8	7	6	5	4	3	2	1	p value
2.56	2.09	1.65	1.24	0.87	0.55	0.30	0.11	0.02	0.00	.99
4.87	4.17	3.49	2.83	2.20	1.61	1.06	0.58	0.21	0.02	.90
6.18	5.38	4.59	3.82	3.07	2.34	1.65	1.01	0.45	0.06	.80
7.27	6.39	5.53	4.67	3.83	3.00	2.19	1.42	0.71	0.15	.70
8.30	7.36	6.42	5.49	4.57	3.66	2.75	1.87	1.02	0.27	.60
9.34	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

T Distribution Table

Percent												
	75	90	95	97.5	99	99.5	99.75	99.9	99.95	99.975	99.99	99.995
One-sided α												
	.25	.10	.05	.025	.01	.005	.0025	.001	.0005	.00025	.0001	.00005
Two-sided α												
	.50	.20	.10	.05	.02	.01	.005	.002	.001	.0005	.0002	.0001
df												
1	1.00	3.08	6.31	12.71	31.82	63.66	127.32	318.31	636.62	1273.24	3183.10	6366.20
2	.82	1.89	2.92	4.30	6.96	9.22	14.09	22.33	31.60	44.70	70.70	99.99
3	.76	1.64	2.35	3.18	4.54	5.84	7.45	10.21	12.92	16.33	22.20	28.00
4	.74	1.53	2.13	2.78	3.75	4.60	5.60	7.17	8.61	10.31	13.03	15.54
5	.73	1.48	2.02	2.57	3.37	4.03	4.77	5.89	6.87	7.98	9.68	11.18
6	.72	1.44	1.94	2.45	3.14	3.71	4.32	5.21	5.96	6.79	8.02	9.08
7	.71	1.42	1.90	2.37	3.00	3.50	4.03	4.79	5.41	6.08	7.06	7.88
8	.71	1.40	1.86	2.31	2.90	3.36	3.83	4.50	5.04	5.62	6.44	7.12
9	.70	1.38	1.83	2.26	2.82	3.25	3.69	4.30	4.78	5.29	6.01	6.59
10	.70	1.37	1.81	2.23	2.76	3.17	3.58	4.14	4.59	5.05	5.69	6.21
11	.70	1.36	1.80	2.20	2.72	3.11	3.50	4.03	4.44	4.86	5.45	5.92
12	.70	1.36	1.78	2.18	2.68	3.06	3.43	3.93	4.32	4.72	5.26	5.69
13	.69	1.35	1.77	2.16	2.65	3.01	3.37	3.85	4.22	4.60	5.11	5.51
14	.69	1.35	1.76	2.15	2.63	2.98	3.33	3.79	4.14	4.50	4.99	5.36
15	.69	1.34	1.75	2.13	2.60	2.95	3.29	3.73	4.07	4.42	4.88	5.24
16	.69	1.34	1.75	2.12	2.58	2.92	3.25	3.69	4.02	4.35	4.79	5.13
17	.69	1.33	1.74	2.11	2.57	2.90	3.22	3.65	3.97	4.29	4.71	5.04
18	.69	1.33	1.73	2.10	2.55	2.88	3.20	3.61	3.92	4.23	4.65	4.97
19	.69	1.33	1.73	2.09	2.54	2.86	3.17	3.58	3.88	4.19	4.59	4.90
20	.69	1.33	1.73	2.09	2.53	2.85	3.15	3.55	3.85	4.15	4.54	4.84

Z score Table

Number in the table represents $P(Z \leq z)$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
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