

Amrita Vishwa Vidyapeetham
Amrita School of Computing, Coimbatore
B.Tech Mid Semester Examinations – Odd 2023-24
Fifth Semester
Computer Science and Engineering

19CSE304 Foundations of Data Science

Duration: Two hours

Maximum: 50 Marks

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 tests and models

Answer all questions

1. Suppose an IT company has two stores that sell computers. The company recorded the number of sales each store made each month. The objective is to compare the sales performance of the two stores with box and whisker plots. In the past 12 months, we have the following numbers of sold computers: [10][CO03] [BTL4]

Store 1: 350, 460, 20, 160, 580, 250, 210, 120, 200, 510, 290, 380.

Store 2: 520, 180, 260, 380, 80, 500, 630, 420, 210, 70, 440, 140.

- Find the 5 number summary in respect of both stores(4marks).
- Find the IQR of Store1(1mark).
- Identify the outliers in Store1. Justify your answer(1mark)
- Draw the comparative box and whisker plots side-by-side(4marks).

Key:

- (a) First, sort the data points in ascending order. 20, 120, 160, 200, 210, 250, 290, 350, 380, 460, 510, 580. Q2, the Median in an even data set is: (the sum of the two middle numbers) / 2. The median is $(250 + 290) / 2 = 270$.

Q1: There are six numbers below the median, namely: 20, 120, 160, 200, 210, 250.

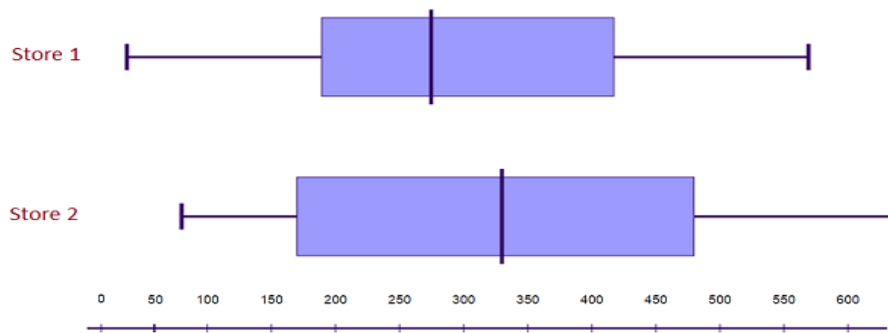
Lower quartile = $(3\text{rd} + 4\text{th data point}) / 2 = (160 + 200) / 2 = 180$

Q3: There are also six numbers above the median, namely: 290, 350, 380, 460, 510, 580. Upper quartile is the median of these six data points. = $(3\text{rd} + 4\text{th data points}) / 2 = 420$.

Finally, the five-number summary for Store 1's sales is 20, 180, 270, 420, 580 (**2marks**)

Similarly the five-number summary for Store 2 is 70, 160, 320, 470, 630 (**2 marks**)

- IQR=240
- No outliers, since the data points are well within the range of 1.5IQR, above/ below Q3 and Q1 respectively
- The comparative double box and whisker plot (2marks each):



2. (a) When is a random sample of size n stated to be independent and identically distributed? (1 mark)
- (b) State the Central Limit Theorem (CLT) (3 marks).
- (c) Illustrate by means of figures, 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)
- (d) State the difference between a parameter and a statistic. (1 mark).
- (e) What are empirical distributions? (1 mark)

[8] [CO01] [BTL2]

Key:

(a) A collection of random variables is independent and identically distributed if each random variable has the same probability distribution as the others and all are mutually independent. Every repetition of the experiment is performed in the same way regardless of the results of all the other repetitions.

(b) The sampling distribution for sample mean is found using the Central Limit Theorem.

NB: Exact definition-3 marks. Accord marks correspondingly.

CENTRAL LIMIT THEOREM

Let $f(x)$ represent a probability density with finite variance σ^2 and mean μ . Also, let \bar{X} be the sample mean for a random sample of size n drawn from this distribution. For large n , the distribution of \bar{X} is approximately normally distributed with mean μ and variance given by σ^2/n .

□

- (c) In both cases it should be a normal distribution only.
- (d) A statistic refers to a sample, as against a parameter is to a population
- (e) The word "empirical" means "observed". Empirical distributions are distributions of observed data, such as data in random samples. They can be visualized by empirical histograms.

3. (a) State the types of errors in statistical hypothesis testing (1 mark). Under what conditions can a hypothesis test be wrong (3 marks)?

- (b) What do you understand by "p value" (1 mark)?
- (c) When would you consider the "p-value" as "highly statistically significant (1 mark).
- (d) What does a "p-value" of .001 mean (2 marks)?

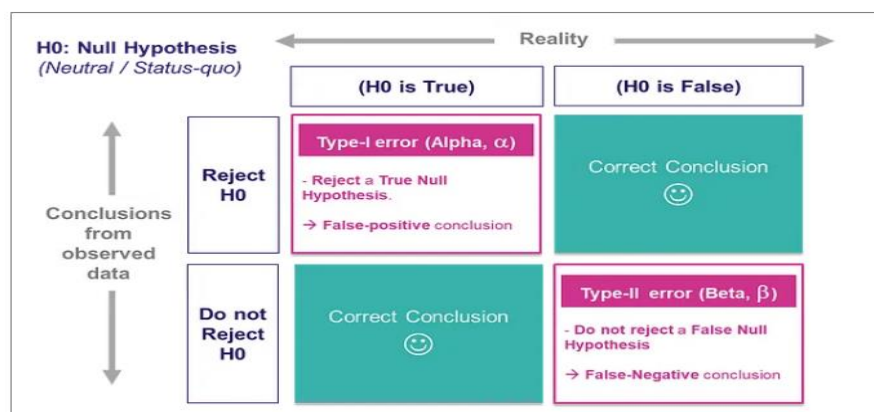
[8] [CO05] [BTL3]

Key:

(a) Types of Errors: Type I and Type II.

A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.

Consider:



(b)

Definition of *P*-value

The *P*-value is the chance,

- under the null hypothesis,
- that the test statistic
- is equal to the value that was observed in the data or is even further in the direction of the alternative.

(c) $p = 0.01$ or the area in the tail is less than 1%

(d) A *p*-value of 0.001 indicates that if the null hypothesis tested were indeed true, there would be a one in 1,000 chance of observing results at least as extreme. This leads the observer to reject the null hypothesis because either a highly rare data result has been observed, or the null hypothesis is incorrect.

4. (a) If 3% of cartons packed with fruits are unfit for consumption, find the probability that in a sample of 200 cartons less than 2 are found spoilt. (3 marks)
- (b) Assume that, you usually get 2 phone calls per hour. Calculate the probability that a phone call will come within the next hour. (2 marks).
- (c) At times, results of Covid-19 tests are incorrect. Let's assume; a diagnostic test has 99% accuracy and 60% of all people have Covid-19. If a patient tests positive, what is the probability that they actually have the disease? (3marks) [8] [CO01] [BTL3]

Key:

(a) The probability of rotten fruits = $3/100 = 0.03$ (p) and Give $n = 200$. Observe that p is small and n is large, hence is a Poisson distribution. (1 mark)

Mean $\lambda = np = 200 \times 0.03 = 6$

$P(X = x)$ is given by the Poisson Distribution Formula as $(e^{-\lambda} \lambda^x)/x!$ (1 mark)

$P(X < 2) = P(X = 0) + P(X = 1)$

$= (e^{-6} 6^0)/0! + (e^{-6} 6^1)/1!$

$= e^{-6} + e^{-6} \times 6$

$= 0.00247 + 0.0148$

$P(X < 2) = 0.01727$

Hence, the probability that less than 2 are unfit is **0.01727. (1mark)**

(b) It is given that, 2 phone calls per hour. So, it would expect that one phone call at every half-an-hour. So, we can take $\lambda = 0.5$ **(1mark)**

It is an exponential distribution

So, the computation is as follows:

$$p(0 \leq X \leq 1) = \sum_{x=0}^1 0.5e^{-0.5x}$$

= 0.393469

Therefore, the probability of arriving the phone calls within the next hour is 0.393469 **(1mark)**

(c)

$$P(\text{covid19}|\text{positive}) = \frac{P(\text{positive}|\text{covid19}) * P(\text{covid19})}{P(\text{positive})}$$

1 mark

$$P(\text{positive} | \text{covid19}) = 0.99$$

$$P(\text{covid19}) = 0.6$$

$$P(\text{positive}) = 0.6 * 0.99 + 0.4 * 0.01 = 0.598 \text{ (1mark)}$$

$$P(\text{covid19}|\text{positive}) = \frac{0.99 * 0.6}{0.598} = 0.993$$

(1marks)

5. (a) The average score on a test is 80 with a standard deviation of 10. With a new teaching curriculum introduced, it is believed that this score will change. A random testing of the scores of 38 students gave the mean as 88. With a 0.05 significance level, is there any evidence to support this claim?(5marks)
- (b) What is the Type I “error probability” of that can be observed in hypothesis tests (3marks)?

[08] [CO05] [BTL4].

Key: This is an example of two-tail hypothesis testing. The z-test will be used as the sample size is greater than 30 and the population standard deviation is given.

$$H_0: \mu = 80, H_1: \mu \neq 80$$

$$\bar{x} = 88, \mu = 80, n = 36, \sigma = 10.$$

$$\alpha = 0.05 / 2 = 0.025$$

The critical value using the normal distribution table is 1.96

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$z = \frac{88 - 80}{\frac{10}{\sqrt{36}}} = 4.8$$

As $4.8 > 1.96$, the null hypothesis is rejected.

Rubrics

Hypothesis - 2 marks

Result-3 marks

(b) Type I “error probability” that can be observed in an hypothesis tests is the “alpha” value chosen (or the cut off for p value). Type I error occurs when the Null hypothesis is actually true, but we fail to accept the Null hypothesis.

An error probability

- The cutoff for the P-value is an error probability.
- If:
 - your **cutoff is 6%**
 - and the **null hypothesis happens to be true**
 - (but you don't know that)
- then there is about a **6% chance** that **your test will reject the null hypothesis**.

6. 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:[8] [CO05] [BTL4]

- State the type of test you would invoke in this case (1 mark).
- State the Null and Alternative Hypothesis (2 mark)
- What is your test statistic? (1mark)
- Evaluate the p value approximately (1 mark)
- State the outcome of the hypothesis test based on your analysis (1 marks).
- How would you justify the use of large random samples in statistical inference? (2marks)

Key:

- State the type of test you would invoke in this case and justify – T –test, since there are only 15 batteries (less than 30).
- State the Null and Alternative Hypothesis

$H_0: \mu = 40$

$H_1: \mu > 40$, where μ is referred to as the mean.

(c) What is your test statistic? The **test statistic is Mean** and the **value of test statistic** is: **$T = (44.9 - 40) / (8.9 / \sqrt{15}) = 2.13$**

(d) p value is **$0.025 < p < 0.05$ (from the t-distribution table). The p-value is approximately 0.02.**

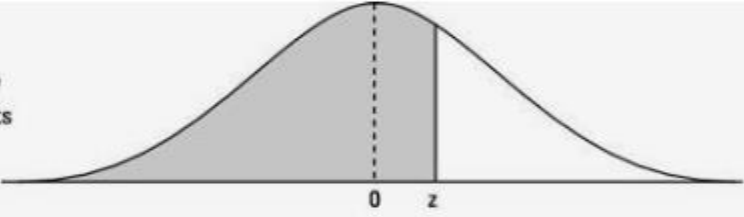
(e) Outcome of the hypothesis test: since **$p < \alpha$** , we **reject the Null**.

(f) Convergence of the Empirical Histogram of the Sample- For a large random sample, the empirical histogram of the sample resembles the histogram of the population, with high probability.

Relevant Tables for probability Distributions

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	.9985	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990

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

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

CO	Marks	BTL	Marks
CO01	16	BTL 1	-
CO02	-	BTL 2	8
CO03	10	BTL 3	16
CO04	-	BTL 4	26
CO05	24	BTL 5	-

CO06	-	BTL 6	-
------	---	-------	---

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

=====