INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

Mid-Autumn Semester Examination 2022-23

Date of Examination: 23.09.2022 Session: FN Duration: 02 hours Full Marks: 50

Subject No. : CS 61061 Subject: Data Analytics Computer Science & Engineering Department

1) Non-programmable calculator may be allowed.
2) Statistical tables may be allowed.

Special instructions:

- Answer to all questions.
- All symbols in the question, if not mentioned explicitly bear their usual meanings.
- You may make reasonable assumptions, if any.

Part-A

1. Multiple options are given against each question below. More than one option may be correct. You have to select all the correct option(s). Choosing a wrong option will not fetch any credit. Otherwise, credit will be given on prorate basis.
No negative marking for selecting any wrong option. [10×1=10]

i. Which of the following category according to NOIR classification does not allow the calculation of median as a measure of central tendency?

(a) Nominal.
(b) Ordinal.
(c) Interval scale.
(d) Ratio scale.

Correct answer: (a)
Explanation:
Only nominal data cannot be ordered and ordering is necessary to calculate the mean. Nominla data cannot be ordered.

ii. Which of the following can be considered to remove outliers in data?

(a) Q1 quartile measure.
(b) Median measure.
(c) IQR (Inter Quartile Range) measure.
(d) Box plot.

Correct answer: (c)
Explanation:
1.5 *IQR can be used to take a decision to remove outliers.

iii. Which of the following statement is true according to the Central Limit theorem?

(a) Population’s mean can be inferred from a sample’s mean for any population.
(b) Population’s variance can be inferred from the sample’s variance for any population.
(c) Population’s mean can be inferred from a sample’s mean for a sample chosen at random.
(d) Population’s variance can be inferred from the sample’s variance for a population provided

that the size of sample is large and the population data is normally distributed.

Correct answer: (d)
Explanation:
Central limit theorem is applicable only when population data follows normal distribution and size of a sample is large enough.

iv. Which of the following probability distribution function(s) is (are) applicable to discrete random variables?

(a) Gaussian distribution
(b) Poisson distribution
(c) Weibull distribution
(d) Chi-square distribution

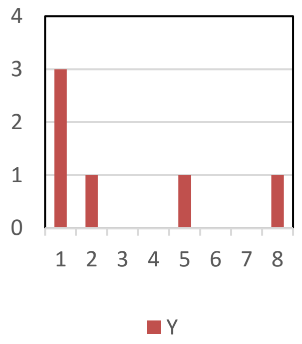
Correct answer: (b)
Explanation:
Poisson distribution is for discrete random variable, whereas others are for continuous random variable.

v. Which of the following is NOT true about a Bernoulli process?

(a) Each trial has two or more outcomes.
(b) Outcomes are mutually exclusive.
(c) Trial are independent to each other.
(d) A trial may be repeated for a number of times.

Correct answer: (a)
Explanation:
A Bernoulli process consists of independent repeated trials, each with dichotomous, that is, binary outcomes.

vi. A distribution of data is shown in the figure given below.



Select the mean and median of the sample from the option given below?

(a) Mean(Y) = 2.0, Median(Y) = 2.5
(b) Mean(Y) = 2.5, Median(Y) = 2.0
(c) Mean(Y) = 3.0, Median(Y) = 1.5
(d) Mean(Y) = 3.5, Median(Y) = 1.0

Correct answer: (c)
Explanation:
Mean = (1×3 + 1×(2+5+8))/6 = 3.0.

Median of the list {1, 1, 1, 2, 5, 8} is (1+2)/2 = 1.5

vii. A probability mass function f(x) for a discrete random variable is given below.

x	**0**	**1**	**2**
f(x)	0.64	0.32	0.04

Select the variance of the data from the option given below?

(a) 0.64
(b) 0.32
(c) 0.20
(d) 0.04

Correct answer: (b)
Explanation:
Mean m = 0.64×0 + 0.32×1 + 0.04×2 = 0.40
Variance = $(0-0.4)^2$ ×0.64+ $(1-0.4)^2$×0.32+$(2-0.4)^2$×0.04 = 0.32

viii. A sample follows normal distribution with mean μ = 0 and variance σ^2 = 1. What is P(X=2)?

(a) 0.1468
(b) 0.1568
(c) 0.1668
(d) 0.1768

Correct answer: (a)
Explanation:
Given mean = 0
Variance = 1

$$f(2) = \frac{1}{(\sqrt{2\pi})} e^{\frac{-1}{2}\frac{2}{1}} = 0.1468.$$

ix. A data distribution in discrete domain satisfies the Poisson distribution. The mean arrival rate is 16 over a fixed period of time, say t = 1. Which of the following is true about the mean and standard deviation of the data?

(a) Mean = 16, Standard deviation = 16
(b) Mean = 16, Standard deviation = 4
(c) Mean = 4, Standard deviation = 16
(d) Mean = 4, Standard deviation = 4

Correct answer: (b)
Explanation:
The following are true for a data which satisfies Poisson's distribution.

$$\mu = \lambda t$$
$$\sigma^2 = \lambda t$$

Here, λ= 16, t = 1

x. In the following Table, Column A lists some sampling distributions, whereas Column B lists the name of sampling distributions. All symbols bear their usual meanings.

Column A		Column B	
(A)	$\frac{\underline{X}-\mu}{\sigma/\sqrt{n}}$	(W)	Normal distribution
(B)	$\frac{\underline{X}-\mu}{S/\sqrt{n}}$	(X)	Chi-squared distribution
(C)	$\frac{{S_1}^2/{\sigma_1}^2}{{S_2}^2/{\sigma_2}^2}$	(Y)	t-distribution
(D)	$\frac{(n-1)S^2}{\sigma^2}$	(Z)	F distribution

Some matchings from Column A and Column B are given below. Select the correct matching?

(a) (A)-(Z), (B)-(X), (C)-(Y), (D)-(Z)
(b) (A)-(X), (B)-(Z), (C)-(W), (D)-(W)
(c) (A)-(Y), (B)-(W), (C)-(X), (D)-(Y)
(d) (A)-(W), (B)-(Y), (C)-(Z), (D)-(X)

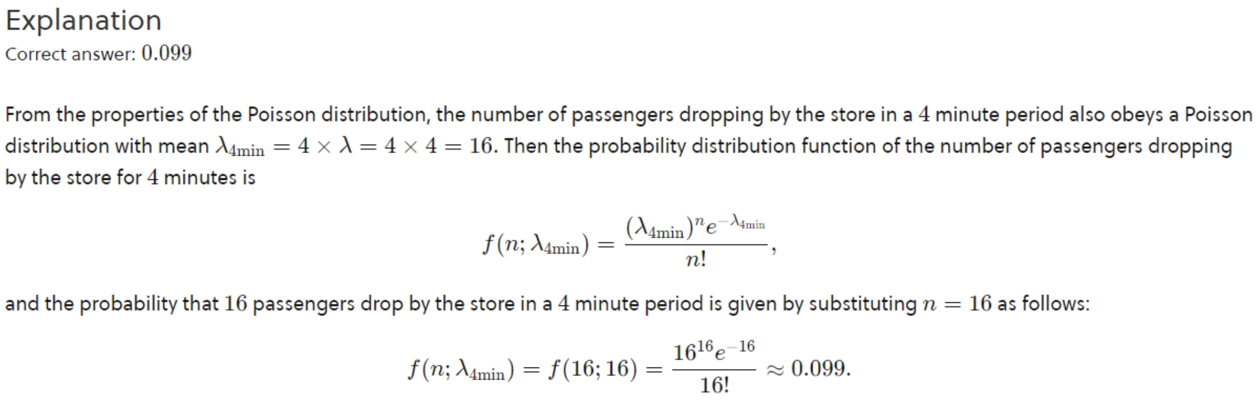
Correct answer: (d)

Explanation:

- χ^2: Describes the distribution of variance.
- t: Describes the distribution of normally distributed random variable standardized by an estimate of the standard deviation.
- F: Describes the distribution of the ratio of two variables.

Part-B

2. Passengers drop by a busy store at an average rate of λ=4 per minute. If the number of passengers dropping by the store obeys a Poisson distribution, what is the approximate probability that16 passengers drop by the store in a particular 4-minute period? [5]



Explanation

Correct answer: 0.099

From the properties of the Poisson distribution, the number of passengers dropping by the store in a 4 minute period also obeys a Poisson distribution with mean $\lambda_{4\text{min}} = 4 \times \lambda = 4 \times 4 = 16$. Then the probability distribution function of the number of passengers dropping by the store for 4 minutes is

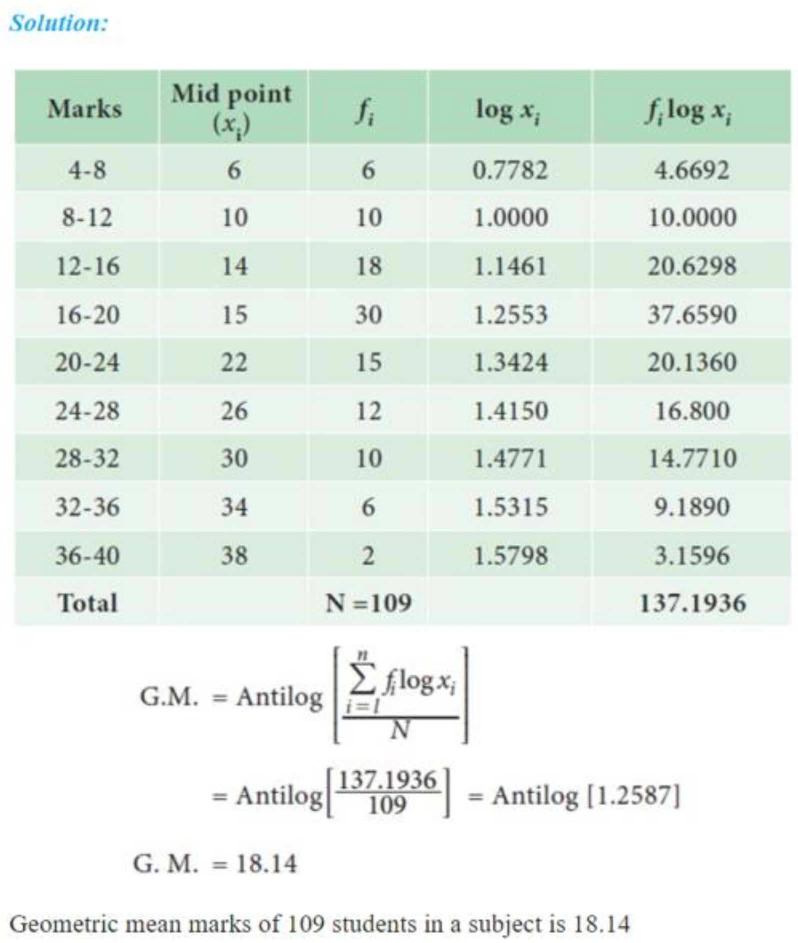
$$f(n; \lambda_{4\text{min}}) = \frac{(\lambda_{4\text{min}})^n e^{-\lambda_{4\text{min}}}}{n!},$$

and the probability that 16 passengers drop by the store in a 4 minute period is given by substituting $n = 16$ as follows:

$$f(n; \lambda_{4\text{min}}) = f(16; 16) = \frac{16^{16} e^{-16}}{16!} \approx 0.099.$$

3. The following is the distribution of marks obtained by 109 students in a class. Find the Geometric Mean (GM). [10]

Marks	4-8	8-12	12-16	16-20	20-24	24-28	28-32	32-36	36-40
No. of students	6	10	18	30	15	12	10	6	2



Solution:

Marks	Mid point (x_i)	f_i	$\log x_i$	$f_i \log x_i$
4-8	6	6	0.7782	4.6692
8-12	10	10	1.0000	10.0000
12-16	14	18	1.1461	20.6298
16-20	15	30	1.2553	37.6590
20-24	22	15	1.3424	20.1360
24-28	26	12	1.4150	16.800
28-32	30	10	1.4771	14.7710
32-36	34	6	1.5315	9.1890
36-40	38	2	1.5798	3.1596
Total		N =109		137.1936

$$\text{G.M.} = \text{Antilog}\left[\frac{\sum_{i=1}^{n} f_i \log x_i}{N}\right]$$

$$= \text{Antilog}\left[\frac{137.1936}{109}\right] = \text{Antilog}\,[1.2587]$$

$$\text{G. M.} = 18.14$$

Geometric mean marks of 109 students in a subject is 18.14

4. The following sample was taken from a normally distributed population.
3, 4, 5, 5, 6, 6, 6, 7, 7, 9, 10, 11, 12 ,12 13, 13, 13, 14, 15

Compute the confidence interval estimation on the population mean. Assume 95% confidence interval. [5]

Solution:

$$n = |S| = 19$$

$$\underline{x} = \frac{3+4+5+5+6+6+6+7+7+9+10+11+12+12+13+13+13+14+15}{19} = 9$$

$$\sigma = \sqrt{\frac{\Sigma(x_i - x)^2}{n-1}} = \sqrt{\frac{260}{18}} = 3.8$$

$$\alpha = confidence\ level = 5\% = 0.05$$

<u>Confidence Interval:</u>

$$\sigma/2 = 0.025 \Rightarrow z\ distribution\ Table = 1.96\ (z_{\alpha/2})$$

$$Formula: \underline{x} \pm z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}$$

$$\underline{x} - z_{\alpha/2}.\frac{\sigma}{\sqrt{n}} = 9 - 1.96 * \frac{3.8}{\sqrt{19}} = 7.29$$

$$\underline{x} + z_{\alpha/2}.\frac{\sigma}{\sqrt{n}} = 9 + 1.96 * \frac{3.8}{\sqrt{19}} = 10.71$$

Ans: [7.29, 10.71]

5. Given a sample drawn randomly which is shown below:

8.08	7.71	7.89	7.72
8.00	7.90	7.77	7.81
8.33	7.67	7.79	7.79
7.94	7.84	8.17	7.87

Test the hypothesis that the population variance is 0.01 at 5% level of confidence.

[10]

Solution:

$$Step\ 1: H_0: \sigma^2 = 0.01$$
$$H_1: \sigma^2 \neq 0.01$$
$$\alpha = 0.05$$

$$Step\ 2: \chi^2 = \frac{(n-1)\, s^2}{\sigma^2}$$
$$Given, \alpha = 0.05\ \&\ with\ degree\ of\ freedom, v = 16 - 1 = 15$$
$$the\ \chi^2\ critical\ value\ is\ 24.996$$
$$Step\ 3: n = 16;\ SS = 0.4761;\ \sigma^2 = 0.01$$
$$\chi^2 = SS / \sigma^2 = 0.4761/0.01 = 47.61$$
$$Step\ 4: Since, \chi^2 = 47.61 > 24.996;\ H_0\ is\ Rejected\ at\ \alpha = 0.05.$$

Conclusion: We conclude that the population variance significantly differs from 0.01.
(ANS)

6. Let the random variable X represent the number of defective parts for a machine when 3 parts are sampled from a production line and tested. The following is the probability distribution of X.

x	0	1	2	3
f(x)	0.51	0.38	0.10	0.01

(a) Find mean μ
(b) Find variance σ^2
(c) Find coefficient of variation, CV

[3+4+3]

Solution:

$$(a)\ \mu = \Sigma x f(x) = 0.61$$
$$(b)\ \sigma^2 = \Sigma(x - \mu)^2 f(x) = 0.4979$$
$$(c)\ Coefficient\ of\ Variation, CV = \sigma / \mu * 100\% = 0.7056/0.61 * 100\% = 115.6755\%$$

(Ans)

---*---