5/41

0207E802

Candidate's Seat No:____

M.Sc. (A.I. & M.L.) Sem.-2 Examination

Statistical Foundation

Time : 3.00 Hours] **July-2025** **[Max.Marks : 100**

SECTION - I

Q. 1 **(20)**

a) A folder on a computer contains 18 code files. Among these files, 6 files on Data Structures, 7 files on Artificial Intelligence, and 5 files on Operating Systems. If three files are randomly selected from the folder, what is the probability that all of them are on Artificial Intelligence? (06)

b) The word "urgent" appears in 25% of emails marked as spam. It appears in only 0.2% of non-spam emails. A user receives emails, 10% of which are spam. What is the probability that an email is spam, given that it contains the word "urgent"? (06)

c) Three compilers — C1, C2, and C3 — are used in a software development environment to compile code. They handle 60%, 20%, and 20% of the total codebase, respectively. It is observed that, 3% of the code compiled by C1 has syntax errors, 6% of the code compiled by C2 has syntax errors, and 5% of the code compiled by C3 has syntax errors. If a code file is randomly selected from the compiled output, what is the probability that it contains a syntax error? (08)

Q. 2 **(15)**

a) Define the following: (10)

i) Discrete random variable

ii) Probability Density Function(pdf)

iii) Cumulative Probability Distribution

iv) Expected value of a random variable

v) Standard normal distribution

b) A data server contains 6 encrypted files and 4 unencrypted files. Two files are selected at random without replacement. Let (05)

(P.T.O)

X denote the number of encrypted files selected. Find the probability distribution of the random variable X.

OR

a) A study on daily coding practice time among junior software developers found that their effective coding time per day (in minutes) follows a uniform distribution ranging from 45 to 75 minutes. (08)

1) What is the probability that a developer codes for 70 minutes or more on a given day?
2) What is the probability that a developer codes between 50 and 65 minutes on a given day?
3) What is the probability that a developer codes for more than 60 minutes on a given day?

b) Difference between normal and standard normal distribution. (07)

Q. 3 **(15)**

a) Suppose the time taken to compile a large software project is normally distributed with a mean of 120 minutes and a standard deviation of 15 minutes. If one compilation process is selected at random, what is the probability that, (05)

i. The project compiles in less than 100 minutes?

ii. The project compiles in 135 to 150 minutes?

b) Let X represent the number of compilation errors, and Y the number of warnings detected by a compiler during automated software testing. (10)

The joint probability mass function (pmf) of X and Y is given in the table below:

X(Error)	Y (Warning)		
	0	1	2
0	0.05	0.10	0.05
1	0.10	0.15	0.05
2	0.05	0.10	0.10
3	0.05	0.05	0.15

Answer the following questions:

a) Construct the marginal pmf for the number of compilation errors (X).

b) Calculate E(X), the expected number of compilation errors.

c) Construct the marginal pmf for the number of warnings (Y).

d) Calculate E(Y), the expected number of warnings.

OR

a) A bug in the code (B) can cause both memory leak (M) and system crash (C). (10)

Both memory leak (M) and heavy server load (L) can cause performance degradation (P) in a software system.

i. Construct a Bayesian network diagram based on the above relationships.

ii. Identify a causal trail in the network.

iii. Identify any common cause trail.

iv. Identify any common effect trail.

b) Explain the applications of Hidden Markov Model in Machine Learning. (05)

SECTION – II

Q. 4 **(20)**

a) When can we say two events are mutually exhaustive? (05)

b) The following data shows the 2024 number of successful project deployments (in units) by 12 software development teams in a tech company: (05)

15	20	14	28	25	50	18	22	28	21	16	19

i. Find the values of the three quartiles (Q_1, Q_2, and Q_3).

ii. Find the interquartile range (IQR).

c) At a large tech institute, the average number of open-source repositories maintained by faculty members is 12, with a standard deviation of 4. The average number of years of programming experience among the same faculty is 18, with a standard deviation of 6. Which of the two data sets — open- (10)

(P.T.O)

source repositories or programming experience — shows greater variability?

Q. 5 **(15)**

a) Answer the following questions regarding Students t-distribution (06)

i. When is the Student's t-test used?

ii. Define degree of freedom

iii. Under what condition does t- distribution resemble the normal distribution

b) The average software engineer attends 5.2 technical webinars per year. A *"tech attendee"* is defined as a software engineer who attends at least one webinar in a 12-month period. A random sample of 40 tech attendees from a reputed tech company revealed that the average number of webinars attended per person was 6.1. The population standard deviation is 2.4 webinars. At the 0.05 level of significance, can it be concluded that this sample represents a significant difference from the national average? (09)

OR

a) Briefly explain the meaning of each of the following terms: (09)

i. Null hypothesis and alternative hypothesis

ii. Critical point

iii. Significance level

b) i) How is the expected frequency of a category calculated for a goodness-of-fit test? (06)

ii) What are the degrees of freedom for such a test?

Q. 6 **(15)**

a) A large dataset of code execution times (in milliseconds) collected from performance testing of a distributed system has a mean of 520 ms and a standard deviation of 65 ms. Using Chebyshev's Theorem, determine at least what percentage of the execution time observations fall within: (08)

i) 2 standard deviations of the mean

ii) 2.5 standard deviations of the mean

iii) 3 standard deviations of the mean

b) Explain Type I and Type II errors with proper examples. (07)

OR

a) In a major tech hub, 12,000 software engineers were laid off last year. Among them, 5,600 were laid off because their companies underwent mergers or relocations, 4,200 were laid off due to project cancellations, and the rest lost their jobs because their roles were automated. (09)

If a software engineer is selected at random from this group, find the probability that they were laid off:

i. Because the company merged or relocated

ii. Due to a project cancellation

iii. Because the role was automated

b) Infosys and TCS have both submitted proposals to develop two independent government AI platforms. The probability that Infosys will win a contract is 0.30, and the same probability applies to TCS. The events are considered independent. (06)

i. What is the probability that both Infosys and TCS will win their respective contracts?

ii. What is the probability that neither Infosys nor TCS will win their contracts?