UNIVERSITY OF JAFFNA

FACULTY OF ENGINEERING

END SEMESTER EXAMINATION– AUGUST, 2025

MC3020- PROBABILITY AND STATISTICS

Date: 20 - 08 -2025 Duration: TWO Hours

Instructions

1. This paper contains **FIVE (5)** questions. Answer **ALL** questions in the answer book provided.
2. Non-programmable scientific calculators are permitted.
3. If you have any doubt as to the interpretation of the wording of a question, make your own decision, but clearly state in the script.
4. This examination accounts for **50%** of module assessment. Total maximum mark attainable is **100**.

Question 1[20 Marks]

1. During a group study session at the University of Jaffna, an engineering student named Kamal challenges his friend Ravi to guess his birth month (excluding the year and day). If Ravi makes a random guess, what is the probability that he guesses correctly?
2. A signal detector correctly identifies 56.25% of the incoming signals as noise-free. When a signal is distorted, the detector identifies it as distorted 94.67% of the time, and when a signal is noise-free, the detector correctly identifies it as noise-free 98.88% of the time. Given that a received signal is noise-free, what is the probability that the detector incorrectly predicted it as distorted? Given that the detector predicted a signal as distorted, what is the probability that the signal is actually distorted? (*Hint:* D = Signal is distorted, N = Signal is noise-free, $\mathrm{Det_D}$ = Detector predicts distorted, $\mathrm{Det_N}$ = Detector predicts noise-free)
3. A tech company is forming a core engineering team by selecting two lead engineers from its current pool of regional engineering managers. Historically, the company has never appointed a female lead engineer for this role. The company has six male regional engineering managers and four female regional engineering managers. Assume all 10 managers are equally qualified, so every possible pair of managers has the same probability of being selected. What is the probability that both selected lead engineers are female?

Question 2[20 Marks]

1. A Structural Health Monitoring Engineer is overseeing a structural health monitoring system for a newly constructed bridge, which has ten critical sensors installed. Each sensor has a 60% chance of functioning correctly during the weekly inspection, independently. The engineer needs to calculate several probabilities to:

   - Decide whether to activate a minimal monitoring setup (supports up to three functional sensors) or the full setup (handles all ten sensors)
   - Prepare for data analysis (requires at least two functional sensors for reliable structural data)
   - Gauge the likelihood of high sensor functionality to prioritize maintenance resources

   (a) Briefly explain why the binomial model is appropriate for this scenario.

   (b) What is the probability that exactly six sensors function correctly?

   (c) What is the probability that at most three sensors function (i.e., the minimal setup is sufficient)?

   (d) What is the probability that data analysis can be performed (i.e., at least two sensors function)?

2. A digital communication system is used to monitor bit errors in a data transmission link. Bit errors occur randomly and independently. Historical data shows that, on average, 2.00 bit errors are detected in any 30-second interval under normal operating conditions.

   (a) What is the probability that no bit errors are detected in a one-minute interval?

   (b) What is the probability of observing at least five but not more than eight bit errors in two minutes of observation?

   (c) Suppose the system has a defect such that each bit error detection has a 15% chance of being a false alarm (i.e., not an actual bit error). If 10 bit errors are detected in a 5-minute interval, what is the probability that exactly 2 of them are false alarms?

   (d) In a dataset of 100 bit error detections, 15 are known to be due to a specific type of impulse noise. If 10 detections are randomly selected for analysis, what is the probability that at least 3 of them are due to this impulse noise?

Question 3[20 Marks]

1. In civil engineering infrastructure assessment, bridge maintenance planning requires careful analysis of structural component reliability. Assume that the time T in years until critical steel reinforcement corrosion occurs in a particular type of concrete bridge deck can be modeled by an exponential distribution with failure rate $\lambda = 0.025$.

(a) Determine the probability that a bridge deck will remain corrosion-free for at least 30 year. This helps municipalities plan inspection intervals for critical infrastructure.

(b) If a new corrosion-resistant concrete mix is used, changing the corrosion rate to $\lambda = 0.018$, would you expect a higher or lower proportion of bridge decks to remain corrosion-free for 30 year? Justify your answer quantitatively.

2. In a study conducted at the Water Resources Engineering Lab at the University of Jaffna, you are analyzing the reliability and performance of water pumps used in municipal distribution systems. These pumps are rigorously tested to assess their operational lifespan under typical working conditions. The lifespans of these pumps follow a normal distribution with a mean of 15,000 hours and a standard deviation of 1,200 hours. The following analysis is designed to evaluate pump performance and make informed decisions about maintenance and replacement schedules.

   (a) Calculate the probability that a randomly selected pump will have a lifespan of less than 13,500 hours.

   (b) To classify pumps into different maintenance categories, determine the lifespan value that corresponds to the 90^{th} percentile of the distribution. This helps identify high-performance pumps for critical applications.

   (c) A quality control standard specifies that pumps with a lifespan of less than 12,000 hours are considered unacceptable for municipal use. Given that a water treatment plant has 150 pumps in operation, estimate how many of these pumps would need replacement based on this lifespan threshold.

   (d) To focus on pumps with optimal performance, determine the range of lifespans within which 92% of the pumps fall. This range should be symmetrically located around the mean lifespan and will help in planning preventive maintenance schedules.

Question 4[20 Marks]

1. As part of their Mechanical Engineering Design project, a team of students at the University of Jaffna has developed a new computer-controlled machining process for manufacturing precision gear shafts. The process uses advanced sensors and actuators to maintain tight tolerances, ensuring high performance in mechanical systems. To evaluate the process, the team measures the diameters (in mm) of ten randomly selected gear shafts and records the following values:

$$23, 28, 34, 39, 31, 37, 38, 30, 32, 28$$

   Assuming the shaft diameters follow a normal distribution:

   (a) Estimate the mean shaft diameter and construct a 98% confidence interval for the mean diameter.

   (b) Additionally, the team seeks to assess the variability in shaft diameters. Construct a 95% confidence interval for the variance of the shaft diameters using the recorded data.

(c) Using the results from part (b), construct a 95% confidence interval for the standard deviation of the shaft diameters, providing valuable insights into the consistency of the manufacturing process.

(d) Is there sufficient evidence that the mean shaft diameter is greater than 30 mm? Use a significance level of $\alpha = 0.01$.

2. An engineering team is evaluating the effectiveness of two different cache replacement algorithms (Algorithm A and Algorithm B) for reducing cache miss rates in a high-performance computing system. They conducted experiments on a simulated processor with 24 benchmark runs: 12 runs using Algorithm A and 12 runs using Algorithm B. The cache miss rates (in misses per thousand instructions) for each run were recorded as follows:

Table 1: Cache miss rates for Algorithms A and B across 12 benchmark runs

Algorithm A:	18	43	28	50	16	32	13	35	38	33	6	7
Algorithm B:	40	54	26	63	21	37	39	23	48	58	28	39

The team wants to determine whether Algorithm A provides a significantly lower cache miss rate than Algorithm B. Assume that the population distributions of the measurements are normal.

(a) Test the hypothesis of equal variances between the two algorithms at a significance level of $\alpha = 0.05$.

(b) Construct a 95% confidence interval for the difference in mean cache miss rates ($\mu_A - \mu_B$), using the appropriate method based on the result from part (a).

(c) Perform a hypothesis test ($\alpha = 0.05$) to determine if Algorithm A's mean cache miss rate is significantly lower than Algorithm B's, accounting for the variance assumption from part (a).

(d) Suppose the team instead uses the same 12 benchmarks and compares both algorithms on each. Perform a paired t-test and construct a 95% confidence interval for the mean difference in cache miss rates.

Question 5[20 Marks]

As an electrical engineer at the Lakvijaya Power Station, you are studying daily power consumption variations in a residential district. Accurate prediction of power consumption is critical for load balancing, grid stability, and energy management. You have collected daily electrical measurements over ten consecutive days from the district's smart grid network. The dataset provided in Table 2 includes four predictor variables: daily mean voltage (V), daily mean current (A), daily power factor (unitless), and ambient temperature (°C). The response variable is daily energy consumption (kWh).

1. Simple Linear Regression: Investigate the relationship between Current and Energy Consumption.

(a) Create a scatter plot with Current on the x-axis and Energy Consumption on the y-axis.

(b) Determine the least squares regression line equation. Interpret the slope coefficient ($\hat{\beta}_1$) in terms of energy consumption response. Discuss the practical meaning of the intercept ($\hat{\beta}_0$) for power distribution applications.

(c) Predict the energy consumption for a current of 24A using the regression equation.

(d) Calculate the correlation coefficient (r) between current and energy consumption. Interpret its strength and direction.

(e) Conduct a hypothesis test at 95% confidence to determine if the regression line is statistically significant. (Hint: Standard error $S_e = 3.4813$)

Table 2: Electrical Data and Energy Consumption Measurements

Voltage (V)	**Current (A)**	**Power Factor**	**Temp. (°C)**	**Energy (kWh)**
220	15	0.95	25	120
225	18	0.93	27	138
230	20	0.90	30	155
235	22	0.88	32	169
240	25	0.85	34	187
245	28	0.82	36	201
250	30	0.80	38	220
255	35	0.78	40	244
260	38	0.75	42	260
265	40	0.72	44	280

2. Multiple Linear Regression: Predict energy consumption using all four electrical factors. The regression output from R is provided on the next page:

(a) Write the complete regression equation for predicting energy consumption.

(b) Interpret each coefficient in the context of energy consumption prediction. Explain how a unit change in each predictor affects energy consumption.

(c) Predict the energy consumption for: Voltage = 232V, Current = 21A, Power Factor = 0.89, Temperature = 31°C.

(d) Interpret the multiple R-squared value (0.9989) and explain what it indicates about the model's effectiveness.

(e) Using the F-test results, determine if the overall regression model is statistically significant at the 5% level. State the null and alternative hypotheses, report the test statistic and p-value, and draw your conclusion.

(f) Conduct a hypothesis test at 95% confidence to determine if Voltage is a significant predictor.

```
Call:
lm(formula = Energy_kWh ~ Voltage_V + Current_A + Power_Factor +
Temp_C, data = electrical_data)

Residuals:
1      2      3      4      5      6      7      8      9     10
0.250 -0.250  1.875  0.125  0.375 -3.375 -0.125 -0.125 -1.875  3.125

Coefficients:
Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  -517.625    663.908  -0.780    0.471
Voltage_V       2.400      2.574   0.932    0.394
Current_A       2.750      1.322   2.081    0.092 .
Power_Factor   75.000    240.985   0.311    0.768
Temp_C         -0.125      3.486  -0.036    0.973
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.387 on 5 degrees of freedom
Multiple R-squared:  0.9989, Adjusted R-squared:  0.998
F-statistic:  1125 on 4 and 5 DF,  p-value: 1.437e-07
```

Important useful formulas

1. If X follows binomial with parameters n, p. Then, the probability mass function is given by
$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x};\ x = 0, 1, 2, \cdots n$$

2. If X follows Poisson with parameters λ. Then, the probability mass function is given by
$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!};\ x = 0, 1, 2, \cdots\cdots$$

3. If X follows exponential distribution with parameter λ. Then, the probability density function is given by
$$f(x) = \lambda e^{-\lambda x};\quad x > 0$$

4. If X follows Hyper-geometric with parameters N, r, n. Then, the probability mass function is given by
$$P(X = x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}};\ x = 0, 1, 2 \cdots r$$

5. $\mu = E(X) = \sum_{\text{for all x}} xP(X = x)$ and $\sigma^2 = Var(X) = E(X^2) - (E(X))^2$
or $Var(X) = E(X-\mu)^2 = \sum_{\text{for all x}} (x-\mu)^2 P(X = x)$

6. Let us consider that a sample space S is divided into two mutually exclusive partitions S_1 and S_2. An event H has occurred, and $P(S_1|H)$ can be written as
$$P(S_1|H) = \frac{P(S_1)P(H|S_1)}{P(S_1)P(H|S_1) + P(S_2)P(H|S_2)}$$

7. If X follows normal distribution with parameter μ and σ then $\left(\frac{X-\mu}{\sigma}\right) = Z$ follows standard normal distribution with $\mu = 0$ and $\sigma = 1$.

8. Sample size calculation based on mean confidence interval : $n = \left(Z_{\frac{\alpha}{2}} * \frac{\sigma}{E}\right)^2$ (where $E-$ margin error)

9. Sample size calculation based on proportion confidence interval : $n = \left(Z^2_{\frac{\alpha}{2}} * \frac{p(1-p)}{E^2}\right)$

10. Sample mean ($\bar{X}$) and standard deviation (s) can be estimated

$$\bar{X} = \frac{\sum X}{n}, \quad s = \sqrt{\frac{\sum \left(X - \bar{X}\right)^2}{n-1}} \quad = \sqrt{\frac{n \sum X^2 \; - \; \left(\sum X\right)^2}{n(n-1)}}$$

11.

Parameter	$(1-\alpha) * 100\%$ **confidence interval**	**Test statistic value**
Mean (μ)	Case 1: $\left(\bar{X} \mp Z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right)$	Case 1: $Z = \frac{\bar{X}-\mu_0}{\left(\frac{\sigma}{\sqrt{n}}\right)}$
	Case 2: $\left(\bar{X} \mp t_{\frac{\alpha}{2},df} * \frac{S}{\sqrt{n}}\right)$	Case 2: $T = \frac{\bar{X}-\mu_0}{\left(\frac{S}{\sqrt{n}}\right)}, df = n-1$
	Case 3: $\left(\bar{X} \mp Z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right)$	Case 3: $Z = \frac{\bar{X}-\mu_0}{\left(\frac{\sigma}{\sqrt{n}}\right)}$
	Case 4: $\left(\bar{X} \mp Z_{\frac{\alpha}{2}} * \frac{S}{\sqrt{n}}\right)$	Case 4: $Z = \frac{\bar{X}-\mu_0}{\left(\frac{S}{\sqrt{n}}\right)}$
Variance (σ)	$\left(\frac{(n-1)*S^2}{\chi^2_U}, \frac{(n-1)*S^2}{\chi^2_L}\right)$	$\chi = \frac{(n-1)s^2}{\sigma_0^2}, df = n-1$

Case 1: when population is normal and σ is known, case 2: when population is normal and σ is unknown, case 3: when population is not normal and σ is known, sample size n is large and case 4: when population is not normal and σ is unknown, sample size n is large.

12.

Parameter	$(1-\alpha) * 100\%$ **confidence interval**	**Test statistic value**
Dependent populations	Case 1: $\left(\bar{D} \mp Z_{\frac{\alpha}{2}} * \frac{\sigma_D}{\sqrt{n}}\right)$	Case 1: $Z = \frac{\bar{D}-\mu_0}{\left(\frac{\sigma_D}{\sqrt{n}}\right)}$
Mean difference ($\mu_1 - \mu_2 = \mu_D$)	Case 2: $\left(\bar{D} \mp t_{\frac{\alpha}{2},df} * \frac{s_D}{\sqrt{n}}\right)$	Case 2: $T = \frac{\bar{D}-\mu_0}{\left(\frac{s_D}{\sqrt{n}}\right)}, df = n-1$
	Case 3: $\left(\bar{D} \mp Z_{\frac{\alpha}{2}} * \frac{\sigma_D}{\sqrt{n}}\right)$	Case 3: $Z = \frac{\bar{D}-\mu_0}{\left(\frac{\sigma_D}{\sqrt{n}}\right)}$
	Case 4: $\left(\bar{D} \mp Z_{\frac{\alpha}{2}} * \frac{s_D}{\sqrt{n}}\right)$	Case 4: $Z = \frac{\bar{D}-\mu_0}{\left(\frac{s_D}{\sqrt{n}}\right)}$

Case 1: when population distribution of the differences is normal and σ_D is known, case 2: when population distribution of the differences is normal and σ_D is unknown, case 3: when population distribution of the differences is not normal and σ_D is known, sample size n is large and case 4: when population distribution of the differences is not normal and σ_D is unknown, sample size n is large.

13. Sample mean ($\bar{D}$) and standard deviation (s_D) for the differences can be estimated

$$\bar{D} = \frac{\sum D}{n}, \quad s_D = \sqrt{\frac{\sum \left(D - \bar{D}\right)^2}{n-1}} \quad = \sqrt{\frac{n \sum D^2 \; - \; \left(\sum D\right)^2}{n(n-1)}}$$

14. Linear regression coefficient estimation formulas,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

15. Standard error (S_e) is given by

$$S_e = \sqrt{\frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{n-2}}$$

16.

Parameter	$(1-\alpha)*100\%$ **confidence interval**	**Test statistic value**
Independent populations	Case 1: $\left(\bar{X}-\bar{Y} \mp Z_{\frac{\alpha}{2}} * \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$	Case 1: $Z = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$
Mean difference $(\mu_1-\mu_2)$	Case 2: $\left(\bar{X}-\bar{Y} \mp t_{\frac{\alpha}{2}} * \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right)$	Case 2: $T = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$ $df = \frac{(A+B)^2}{\frac{A^2}{n_1-1}+\frac{B^2}{n_2-1}}; A = \frac{s_1^2}{n_1},\ B = \frac{s_2^2}{n_2},$
	Case 3: $\left(\bar{X}-\bar{Y} \mp t_{\frac{\alpha}{2}} * s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$ where $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$	Case 3: $T = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ $df = n_1+n_2-2$
	Case 4: $\left(\bar{X}-\bar{Y} \mp Z_{\frac{\alpha}{2}} * \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$	Case 4: $Z = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$
	Case 5: $\left(\bar{X}-\bar{Y} \mp Z_{\frac{\alpha}{2}} * \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right)$	Case 5: $Z = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$
Proportions difference (p_1-p_2)	$\left(\hat{p_1}-\hat{p_2} \mp Z_{\frac{\alpha}{2}} * \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}\right)$	$Z = \frac{\hat{p_1}-\hat{p_2}-(p_1-p_2)}{\sqrt{\bar{p}(1-\bar{p})(\frac{1}{n_1}+\frac{1}{n_2})}}$ $\bar{p} = \frac{X_1+X_2}{n_1+n_2}$

Case 1: When the two independent population distributions are normal and the population variances σ_1^2 and σ_2^2 are known, case 2: When the two independent population distributions are normal and the population variances σ_1^2 and σ_2^2 are unknown and unequal, case 3: When the two independent population distributions are normal and the population variances σ_1^2 and σ_2^2 are unknown but equal, case 4: When the two independent population distributions are not normal and the population variances σ_1^2 and σ_2^2 are known, and the sample size n_1 and n_2 are large and case 5: When the two independent population distributions are not normal and the population variances σ_1^2 and σ_2^2 are unknown, and the sample size n_1 and n_2 are large.

17. Correlation (r) is given by,

$$r = \frac{\sum_{i=1}^{n} X_iY_i - n\bar{X}\bar{Y}}{\sqrt{\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right)}}$$

18. Test statistic value when testing the hypothesis for correlation is given by,

$$T = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}, \quad \text{d.f} = n-2$$

19. Test statistic value when testing the hypothesis for slope coefficient is given by,

$$T = \frac{\hat{\beta}_1 - 0}{\frac{Se}{\sqrt{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}}}, \quad \text{d.f} = n-2$$

20. Test statistic value when testing the hypothesis for intercept coefficient is given by,

$$T = \frac{\hat{\beta}_0 - 0}{Se\sqrt{\frac{1}{n}+\frac{\bar{X}^2}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}}}, \quad \text{d.f} = n-2$$

End of Examination