UNIVERSITY OF JAFFNA

FACULTY OF ENGINEERING

END SEMESTER EXAMINATION– AUGUST, 2024

MC3020- PROBABILITY AND STATISTICS

Date: 22 - 08 -2024 Duration: TWO Hours

Instructions

1. This paper contains **FIVE (5)** questions.
2. Answer **ALL** questions in the answer book provided.
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. A large manufacturing company operates three plants C_1, C_2, and C_3 that produce different portions of the company's total output. Plant C_1, known for its precision and high standards, produces 10% of the total output with only 1% of its products being defective. Plant C_2, the largest plant, contributes 50% of the total production but has a 3% defect rate due to the complexity of its processes. Plant C_3, producing the remaining 40%, has a defect rate of 4%, partly because of the older machinery used in its operations. After production, all items are sent to a central warehouse. Before storage, each item undergoes an additional quality control inspection. This inspection has a 95% probability of correctly identifying a defective item and a 90% probability of correctly identifying a non-defective item (Event D represents the product being defective, and event I represents the product being marked as defective by the inspection process).

   (a) Determine the probability that a randomly selected item from the central warehouse is defective (event D).

   (b) Calculate the overall probability that an item from the central warehouse is marked as defective (event I) after passing through the inspection process.

   (c) Given that a product is marked as defective after inspection, determine the conditional probability that it was produced by plant C_1.

   (d) Evaluate the effectiveness of the inspection process by determining the probability that an item marked as defective by the inspection is actually defective.

2. An E22 batch Engineering student at the University of Jaffna, enrolled in the MC3020 course, is concerned about being late for an 08:00 am lecture class due to a malfunctioning alarm clock. To minimize the risk, the student decides to use three alarm clocks. Each alarm clock has a 92% probability of working correctly. What is the probability that at least one of the alarm clocks will work correctly, ensuring the student wakes up on time?

Question 2[20 marks]

1. The Faculty of Engineering at the University of Jaffna is conducting an evaluation of the attendance of students from the E22 batch in the MC3020 course. The goal is to achieve an attendance rate of 80% among the students. To assess whether this goal is being met, a sample of 28 students from the E22 batch has been selected. These students are representative of the broader E22 cohort. Each of these 28 students is expected to attend the MC3020 course regularly, as part of their academic requirements. The attendance of each student is meticulously recorded over a designated period to determine whether the desired 80% attendance rate is being maintained. We aim to analyze the probabilities associated with the attendance of these 28 sampled students:

   (a) What is the probability that exactly half of the sampled students will attend the MC3020 course.

   (b) What is the probability that 18 or more of the 28 sampled students will attend the MC3020 course consistently, in accordance with the desired attendance rate.

2. As part of the E23 batch orientation program at the Faculty of Engineering, workshops are organized to familiarize incoming students with their English language skills. On average, one workshop experiences a disruption every four days. The faculty is particularly concerned about ensuring a smooth start on the first day of the orientation program.

   (a) What is the probability that a randomly selected workshop will not experience a disruption on the first day of the orientation program?

   (b) Considering the importance of a successful start, what is the probability of having at least one workshop without disruption on the first day of the orientation program? (assuming that there are 5 workshops on the first day)

3. Suppose there has been a recent incident of raggiong (a term used to describe the act of forcefully entering and causing damage or theft) reported at the Faculty of Engineering, University of Jaffna. As a precautionary measure, security personnel are instructed to randomly inspect 16% of all incoming bags for any suspicious items. If any of the inspected bags are found to contain suspicious items, all remaining bags are inspected. Given that two out of the next box of 50 bags are found to contain suspicious items, what is the probability that all the bags will be inspected before use?

Question 3[20 marks]

1. As part of their Design and Prototyping (ID3020) project, a team of students at the University of Jaffna has developed a cutting-edge smart irrigation system designed to optimize water usage in agricultural fields. The system incorporates sensors and actuators to precisely control the amount of water delivered to crops, enhancing crop yield while conserving water resources. To ensure the system's effectiveness, the team conducts a series of tests on randomly selected plots of land. The team measures the soil moisture levels in ten randomly selected plots and records the following percentages:

$$28, 31, 36, 37, 33, 31, 38, 31, 32, 29.$$

   Assuming the soil moisture levels follow a normal distribution

   (a) Estimate the mean soil moisture content per plot and construct a 98% confidence interval for mean soil moisture content.

   (b) Additionally, the team seeks to assess the variability of soil moisture content across the plots. Construct a 95% confidence interval for the variance of soil moisture content per plot using the recorded data.

   (c) Using the results obtained in part (b), constructs a 95% confidence interval for the standard deviation of soil moisture content per plot, providing valuable insights into the consistency of water distribution achieved by their innovative smart irrigation system.

   (d) Is there sufficient evidence that the mean soil moisture content per plot is greater than 30? Use a $\alpha = 0.01$.

2. An engineering team is evaluating the effectiveness of two different insulation materials for reducing heat loss in a specific industrial application. They are testing two different types of insulation materials to determine if there is a significant difference in their thermal resistance. To evaluate this, the team conducts thermal resistance tests on samples coated with each type of insulation. A total of 24 samples are tested, with 12 samples coated using Insulation A and 12 samples coated using Insulation B. After conducting the tests, the thermal resistance in ohms for each sample is measured, and the data are summarized below:

   Insulation A (Thermal Resistance in Ohms):

$$18, 43, 28, 50, 16, 32, 13, 35, 38, 33, 6, 7$$

   Insulation B (Thermal Resistance in Ohms):

$$40, 54, 26, 63, 21, 37, 39, 23, 48, 58, 28, 39$$

   The engineering team wants to determine if Insulation A provides a significant improvement in thermal resistance compared to Insulation B. Assume that the population distributions of the measurements are normal with equal variances.

(a) Construct a 95% confidence interval for the difference in mean thermal resistance $(\mu_1 - \mu_2)$ to assess the size of the difference between the two means.

(b) Test whether the mean thermal resistance of samples coated with Insulation A is significantly greater than the mean thermal resistance of samples coated with Insulation B using a significance level (α) of 0.05.

(c) Suppose instead of using independent samples, the team collects paired data where each sample of Insulation A is compared to a corresponding sample of Insulation B. Redo the analysis using a paired t-test and construct a 95% confidence interval for the mean difference in thermal resistance.

Question 4[20 marks]

1. In civil engineering, drainage systems play a crucial role in urban infrastructure. Assume that the time T in hours until a particular type of drainage pipe becomes clogged can be modeled by an exponential distribution with $\lambda = 0.0004$.

   (a) Find the proportion of pipes that will function without clogging for at least $10,000$ hours.

   (b) If the pipe material is upgraded so that the clogging rate changes to $\lambda = 0.00055$, would you expect a higher or lower proportion of pipes to remain unclogged for at least $10,000$ hours?

2. In a study conducted at the Mechanical Engineering Lab at the University of Jaffna, you are analyzing the reliability and performance of a specific type of bearing used in industrial machinery. These bearings are rigorously tested to assess their lifespan under operating conditions. The lifespans of these bearings follow a normal distribution with a mean of 2000 hours and a standard deviation of 200 hours. The following analysis is designed to evaluate the performance of these bearings and make informed decisions based on the lifespan data:

   (a) Calculate the probability that a randomly selected bearing will have a lifespan of less than 1750 hours.

   (b) To classify bearings into different performance categories and highlights the characteristics of high-performing bearings. Determine the lifespan value that corresponds to the 90th percentile of the distribution.

   (c) A quality control policy specifies that bearings with a lifespan of less than 1650 hours are considered unacceptable and are therefore rejected. Given that a batch of 800 bearings is tested, calculate how many of these bearings would be rejected based on this lifespan threshold.

   (d) To focus on bearings with above-average performance, determine the range of lifespans within which 92% of the bearings fall. This range should be symmetrically located around the mean lifespan. ~~of 125 hours.~~

Question 5[20 marks]

1. Imagine you are a computer engineer working on optimizing the performance of different types of computer processors. You have collected data on the operating frequency (in GHz) and corresponding processing speed (in FLOPS - Floating Point Operations Per Second) for several processors. The dataset you have is as follows:

   Operating Frequency (GHz): 2.0, 2.5, 3.0, 3.5, 4.0, 4.5
   Processing Speed (FLOPS): 50, 60, 75, 90, 110, 130

   You are tasked with analyzing this data to predict the processing speed of a processor based on its operating frequency.

   (a) Calculate the equation of the least squares regression line for predicting processing speed based on operating frequency.

   (b) Interpret the slope and intercept coefficients of the regression line in the context of computer processor performance.

   (c) Use the regression line to predict the processing speed for a processor with an operating frequency of 3.2 GHz.

   (d) Calculate the coefficient of correlation (r) and explain its significance in the context of the relationship between operating frequency and processing speed.

   (e) Perform a hypothesis test to determine whether the regression line is statistically significant at a 95% confidence level. State your hypotheses, the test statistic, and the conclusion clearly.

2. As a civil engineering student working on final year project at the Ariviyal Nagar Weather Station which is located in the Renewal Energy Park of the Engineering Faculty, you are tasked with predicting the soil moisture content at a 30cm depth based on environmental factors such as temperature, wind speed, solar radiation, and overall moisture levels. The datasets you have collected is as follows:

Temperature (°C)	Wind Speed (km/h)	Solar Radiation (W/m²)	Moisture Level (%)	Soil Moisture Content at 30cm Depth (%)
25	10	800	90	14
27	15	850	85	13
30	20	900	60	11
32	25	1350	45	20
34	30	800	70	19
36	35	750	35	17
38	40	700	30	15
40	45	650	25	19

   To utilize this data for predictive analysis and better manage water resources, you need to answer the following questions based on the given R output:

   (a) Determine the equation for predicting soil moisture content at a 30cm depth based on temperature, wind speed, solar radiation, and moisture level.

   (b) Explain the meaning of each coefficient in the context of the problem. Describe how changes in each environmental factor affect the soil moisture content.

(c) Use the regression plane equation to predict the soil moisture content at a 30cm depth for the following conditions: temperature of 26°C, wind speed of 31 km/h, solar radiation of 725 W/m^2, and moisture level of 18%.

(d) Conduct a test to determine the significance of the temperature variable in predicting soil moisture content. State the hypotheses, the test statistic, and the conclusion clearly.

```
Call:
lm(formula = Soil_Moisture_Content_30cm_Depth ~ Temperature +
    Wind_Speed + Solar_Radiation + Moisture_Level)

Residuals:
      1       2       3       4       5       6       7       8
 2.0073 -2.0073 -1.2877  0.4015  0.4292  1.6926 -1.8477  0.6120

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     32.798689  76.570818   0.428   0.6973
Temperature     -2.531816   3.659003  -0.692   0.5387
Wind_Speed       1.628346   1.580284   1.030   0.3786
Solar_Radiation  0.014744   0.006115   2.411   0.0949 .
Moisture_Level   0.160123   0.103913   1.541   0.2210
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.36 on 3 degrees of freedom
Multiple R-squared:  0.7741,	Adjusted R-squared:  0.4729
F-statistic:  2.57 on 4 and 3 DF,  p-value: 0.232
```

Figure 1: R output for question 5

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

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: $Z = \frac{\bar{X}-\mu_0}{(\frac{\sigma}{\sqrt{n}})}$
	Case 2: $\left(\bar{X} \mp t_{\frac{\alpha}{2},df} * \frac{S}{\sqrt{n}}\right)$	Case 2: $T = \frac{\bar{X}-\mu_0}{(\frac{S}{\sqrt{n}})}, df = n-1$
	Case 3: $\left(\bar{X} \mp Z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right)$	Case 3: $Z = \frac{\bar{X}-\mu_0}{(\frac{\sigma}{\sqrt{n}})}$
	Case 4: $\left(\bar{X} \mp Z_{\frac{\alpha}{2}} * \frac{S}{\sqrt{n}}\right)$	Case 4: $Z = \frac{\bar{X}-\mu_0}{(\frac{S}{\sqrt{n}})}$
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 3: when population is not normal and σ is known, 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: $Z = \frac{\bar{D}-\mu_0}{(\frac{\sigma_D}{\sqrt{n}})}$
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: $T = \frac{\bar{D}-\mu_0}{(\frac{s_D}{\sqrt{n}})}, df = n-1$
	Case 3: $\left(\bar{D} \mp Z_{\frac{\alpha}{2}} * \frac{\sigma_D}{\sqrt{n}}\right)$	Case 3: $Z = \frac{\bar{D}-\mu_0}{(\frac{\sigma_D}{\sqrt{n}})}$
	Case 4: $\left(\bar{D} \mp Z_{\frac{\alpha}{2}} * \frac{s_D}{\sqrt{n}}\right)$	Case 4: $Z = \frac{\bar{D}-\mu_0}{(\frac{s_D}{\sqrt{n}})}$

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 known, 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,

$$\hat{\beta_1} = \frac{\sum_{i=1}^n X_iY_i - n\bar{X}\bar{Y}}{\sum_{i=1}^n X_i^2 - n\bar{X}^2} \quad \hat{\beta_0} = \bar{Y} - \hat{\beta_1}\bar{X}$$

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 Mean difference $(\mu_1-\mu_2)$	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 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 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 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 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 1: $Z=\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$ 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: $T=\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ $df=n_1+n_2-2$ Case 4: $Z=\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$ Case 5: $Z=\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$
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 known, 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