National University of Computer and Emerging Sciences

Karachi Campus

Probability and statistics (MT2005)

Final Exam

Total Time (Hrs): 3

Total Marks: 100

Total Questions: 6

Date: May 22nd 2024

Course Instructors

Mr. Muhammad Amjad, Mr.Abdul Basit,
Mr. Nadeem Arif Khan, Ms.Urooj , Ms. Alishba Tariq & Mr. Moheez ur Rehman

_____ _____ _____

Roll No Section Student Signature

Do not write below this line

Attempt all the questions.

CLO1 #: Describe the fundamental concepts in probability & statistics **[5+5+5]**

Q1(a) : A computer assembling company receives 24% of parts from supplier X, 36% of parts from supplier Y, and the remaining 40% of parts from supplier Z. Five percent of parts supplied by X, ten percent of parts supplied by Y, and six percent of parts supplied by Z are defective. If an assembled computer has a defective part in it, what is the probability that this part was received from supplier Z ?

Q1(b): Sometimes stem-and-leaf plots are used to compare two samples. For this purpose, one can put two leaves on the same stem. Consider the samples of round-trip transit times (known as pings) received from two locations (data set Pings).

$Location\ I$:	$0.0156, 0.0396, 0.0355, 0.0480, 0.0419, 0.0335, 0.0543, 0.0350,$ $0.0280, 0.0210, 0.0308, 0.0327, 0.0215, 0.0437, 0.0483\ seconds$
$Location\ II$:	$0.0298, 0.0674, 0.0387, 0.0787, 0.0467, 0.0712, 0.0045, 0.0167,$ $0.0661, 0.0109, 0.0198, 0.0039\ seconds$

Choosing a leaf unit of 0.001, a stem unit of 0.01, and dropping the last digit, construct the two stem-and-leaf plots, one to the left and one to the right of the stem. Looking at these two plots, which location has a more stable connection?

Q1(c): The grade point average (GPA) of 20 college seniors selected at random from the graduating class as follows.

3.2	1.9	2.7	2.4	2.8
2.9	3.8	3.0	2.5	3.3
1.8	2.5	3.7	2.8	2.0
3.2	2.3	2.1	2.5	1.9

i. Compute the mean, median, and mode of the data set. **[1]**

ii. Find the range, variance and standard deviation of the data set. **[1]**

iii. Construct the interval $\bar{x} \pm \sigma$, $\bar{x} \pm 2\,\sigma$ and $\bar{x} \pm 3\,\sigma$. Count the number of observations that fall within each interval and find the percentage of grade point average. **[3]**

CLO 2 #: Analyze the data and produce probabilistic models for different models **[8+4+4+4]**

Q2(a): Each front tire on a particular type of vehicle is supposed to be filled to a pressure of 26*psi*. Suppose the actual air pressure in each tire is a random variable X for the right tire and Y for the left tire, with joint pdf

$$f(x,y) = \begin{cases} k(x^2+y^2) & 20 \le x \le 30, 20 \le y \le 30 \\ 0 & otherwise \end{cases}$$

i. What is the value of k? **[3]**

ii. What is the probability that both tires are under filled? **[3]**

iii. Determine the (marginal) distribution of air pressure in the right tire alone and conditional density $f(y/x)$. **[2]**

Q2(b): The number of computer shutdowns during any month has a Poisson distribution, averaging 0.25 shutdowns per month**.**

i. What is the probability of at least 2 computer shutdowns during the next year? **[2]**

ii. What is the variance of the number of computer shutdowns during the next year? **[2]**

Q2(c): An exciting computer game is released. Sixty percent of players complete all the levels. Thirty percent of them will then buy an advanced version of the game. Among 15 users, what is the expected number of people who will buy the advanced version? What is the probability that at least two people will buy it?

Q2(d): Consider the scenario in which you toss a fair die 12 times. What is the probability that each face value (1-6) will occur exactly twice?

CLO3 #: Apply the rules and algorithms of probability and statistics to their relevant problems. **[15+10]**

Q3(a): Ali wants to brag to his non-Fast colleagues about how smart FAST students are. To give himself credibility, he decides to run a statistical test comparing the IQ scores of FAST students and NUST students. He collects IQ scores from 11 FAST students. The data has a sample mean of 115, with a sample standard deviation of 8. He then collects IQ scores from 11 NUST students. Their scores have a sample mean of 110, with a sample standard deviation of 6? Assume that the two sets of data came from normal populations with equal variances**.**

i. Which test should he run to compare the IQ scores from the two schools? What assumptions will he need to make? What are the null and alternative hypotheses? **[2]**

ii. Run the test with a significance level of $\alpha = 0.05$. Should Ali reject the null hypothesis or not? **(*Critical value: 2.086*)** **[8]**

Estimate the 95% confidence interval for the IQ of FAST students. **[5]**
(*Critical value: 2.086*)

Q3(b): A school mathematics teacher decides to test the effect of using an educational computer package, consisting of geometric designs and illustrations, to teach geometry. Since the package is expensive, the teacher wishes to determine whether using the package will result in an improvement in the pupils understanding of the topic. The teacher randomly assigns pupils to two groups: a control group receiving standard lessons and an experimental group using the new packages. The pupils are selected in pairs of equal mathematical ability, with one form each pair assigned at random to the control group and other to the experimental group. On completion of the topic, the pupils are given a test to measure their understanding. The results, percentage marks are shown in the table:

Pair	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Control	72	82	93	65	76	89	81	58	95	91
Experimental	75	79	84	71	82	91	85	68	90	92

Assuming the percentage marks to be normally distributed, investigate the claim that the educational computer package produces an improvement in pupils understanding of geometry. ***(Critical value: 1.833)***

CLO3 #: Apply the rules and algorithms of probability and statistics to their relevant problems. **[15]**

Q4: A company buys light bulb every year. This year company is considering four brands of light bulbs to choose from. They wants to investigate if the mean life time of the four brands is same.The research department randomly selected few bulbs and tested them, The amount of hours that each bulb in each brand lasts before burning out is listed in the accompanying table:

BRAND 1	23	24	29	26	22	23	25
BRAND 2	19	23	18	24	20	22	19
BRAND 3	23	27	25	26	23	21	27
BRAND 4	26	24	21	29	28	24	28

At a 5% significance level, test the null hypothesis that the mean lifetime of bulbs for each of these four brands is the same. **(*Critical Value : $F_{0.05,\ 3,\ 24} = 3.01$*)**

CLO3#: Apply the rules and algorithms of probability and statistics to their relevant problems. **[15]**

Q5: The following table provides the information on the amount of sugar in gram and calorie count in one serving of a sample of 13 different varieties of cereal:

Sugar(x)	4	15	12	11	8	6	7	2	7	14	20	3	13
Calories(y)	120	200	140	110	120	80	190	100	120	190	190	110	120

(a) Compute (Manually / calculation)$\sum x, \sum y, \sum xy, \sum x^2\ \sum y^2$. **[5]**

(b) Calculate the coefficient of correlation and state the nature of relation. **[2]**

(c) Estimate the linear regression line (y on x). **[2]**

(d) Estimate the calorie when sugar is 9 gram. **[1]**

(e) Test the hypothesis that there is no correlation between x and y, use $\alpha = 0.05$ **[5]**
(*Critical value: 2.201*)

CLO3#: Apply the rules and algorithms of probability and statistics to their relevant problems. **[10]**

Q:6 Choose the Best Answers:

I. Set of populations are called:
a) Samples b) Elements c) Estimates d) Averages

II. Which of the following best represents the center value of a set of data
a) Variance b) Averages c) standard deviation d) proportion

III. When $n < 30$ and the population standard deviation is known, what is appropriate distribution?
a) z b) t c) r d) p

IV. The probability you reject the null hypothesis when in fact the null hypothesis is true is called:
a) Type I error b) a Type II error c) the power d) none of these

V. A intelligent guess about the characteristics of population is called:
a) Mean b) proportion c) hypothesis d) test statistic

VI. Green sea turtles have normally distributed weights, measured in kilograms, with a mean of 134.5 and a variance of 49.0. A particular green sea turtle's weight has a z score of -2.4. What is the weight of this green sea turtle? Round to the nearest whole number
a) 17kg b)151kg c)118kg d) 252kg

VII. Refer to the discrete probability distribution provided in the table below:

$X = x$	0	1	2	3	4
$P(X = x)$	0.040	0.110	0.450	0.230	?

Find the probability that x is equal to 0 or4. Round to 3 decimal places.
a) 0.040 b) 0.210 c) 0.007 d) 1.000

VIII. Which of the following exam scores is better relative to other students enrolled in the course?

- A psychology exam grade of 85; the mean grade for the psychology exam is 92 with a standard deviation of 3.5
- An economics exam grade of 67; the mean grade for the economics exam is 79 with a standard deviation of 8
- A chemistry exam grade of 62; the mean grade for the chemistry exam is 62 with a standard deviation of 5

a) The psychology exam score is relatively better b) The economics exam score is relatively better
c) The chemistry exam score is relatively better d) All of the exam scores are relatively equivalent

IX. Five students take AP Calculus AB one year and AP Calculus BC the next year. Their overall course grades (%) are listed below for both courses. Which of the following statistical procedures would be most appropriate to test the claim that student overall course grades are the same in both courses? Assume that any necessary normality requirements hold.

Student	1	2	3	4	5
AP Cal AB	80%	72.6%	99%	91.3%	68.9%
AP Cal BC	85.5%	71%	93.2%	93%	74.9%

a) Two-tailed two-sample paired/dependent t-test of means b) Two-tailed two-sample indepent t-test of mean
c) Two-tailed two-sample independent z-test of means d) One-tailed two-sample z-test of proportions

X. which of the following serves as a guide in deciding whether reject or accept the null hypothesis
a) Confidence interval b) Decision rule c) Acceptance region d) Interpretation

---------- **Good Luck** ----------

Formula Sheet

Value of Test Statistic	Confidence Interval estimation
$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$; σ known	$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$,
$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$; $v = n - 1$, σ unknown	$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$,
$z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$; σ_1 and σ_2 known	$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$,
$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}}$; $v = n_1 + n_2 - 2$, $\sigma_1 = \sigma_2$ but unknown, $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$	$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$,
$t' = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$; $v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$, $\sigma_1 \neq \sigma_2$ and unknown	$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$,
$t = \frac{\bar{d} - d_0}{s_d/\sqrt{n}}$; $v = n - 1$	$\bar{d} - t_{\alpha/2}\frac{s_d}{\sqrt{n}} < \mu_D < \bar{d} + t_{\alpha/2}\frac{s_d}{\sqrt{n}}$,

Simple linear regression and correlation co-efficients and test statistic for correlation co-efficient

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1\sum_{i=1}^{n} x_i}{n} = \bar{y} - b_1\bar{x}.$$

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \qquad v = n - 2$$

ANOVA Formulae

Sum of squares	Defining formula	Computing formula
Total, SST	$\Sigma(x_i - \bar{x})^2$	$\Sigma x_i^2 - (\Sigma x_i)^2/n$
Treatment, SSTR	$\Sigma n_j(\bar{x}_j - \bar{x})^2$	$\Sigma(T_j^2/n_j) - (\Sigma x_i)^2/n$
Error, SSE	$\Sigma(n_j - 1)s_j^2$	SST − SSTR

Correlation co-efficient and co-variance formulae for joint PMF/PDF

$$\sigma_{XY} = E(XY) - \mu_X\mu_Y. \qquad \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}.$$

Some discrete probability distribution formulae

$$b(x; n, p) = \binom{n}{x} p^x q^{n-x}, \qquad p(x; \lambda t) = \frac{e^{-\lambda t}(\lambda t)^x}{x!},$$

$$f(x_1, x_2, \ldots, x_k; p_1, p_2, \ldots, p_k, n) = \binom{n}{x_1, x_2, \ldots, x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k},$$

with

$$\sum_{i=1}^{k} x_i = n \text{ and } \sum_{i=1}^{k} p_i = 1.$$

Total probability and Baye's Rule formulae

$$P(A) = \sum_{i=1}^{k} P(B_i \cap A) = \sum_{i=1}^{k} P(B_i)P(A|B_i). \qquad P(B_r|A) = \frac{P(B_r \cap A)}{\sum_{i=1}^{k} P(B_i \cap A)} = \frac{P(B_r)P(A|B_r)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)}$$