Biostatistics and R

Time: 2 hours **Maximum Marks: 35**

PART A. Answer any 8 out of 10) (8×1 **Mark)**

A1. A pair of dice are thrown and the two outcomes are summed up. The probability that the sum is greater than 10 is

(a) $\frac{1}{12}$ (b) $\frac{1}{6}$ (c) $\frac{2}{9}$ (d) $\frac{4}{36}$

A2. Which one of the following variables is continuous?

(a) Gender of a person: male or female (b) Choice on a test item: true or false.
(c) Age of a person. (d) Names of colors on the rainbow.

A3. The appropriate plot to visualize the fractional representation of various groups in a population will be a

(a) Histogram (b) Pie Chart (c) Scatter Plot (d) ROC curve

A4. In a Gaussian distribution, the area under the curve between ±1 standard deviation from the mean is approximately equal to

(a) 63% of the total area (b) 95% of the total area
(c) 68% of the total area (d) 50% of the total area

A5. Which of the following changes to a study would result in a narrower confidence interval?

(a) increasing the confidence level, increasing the sample size
(b) decreasing the confidence level, decreasing the sample size
(c) decreasing the confidence level, increasing the sample size.
(d) increasing the confidence level, decreasing the sample size

A6. Which one of the following should not be used to overlay error bars on a line or bar plot of the mean?

(a) The standard deviation of the data
(b) The standard error of the mean
(c) A 95% confidence interval for the mean
(d) The median of the data

A7. In the absence of any exercise, the correlation between daily calorie consumption and body weight is expected to be

(a) zero or near zero
(b) small negative
(c) moderate to large positive
(d) moderate to large negative

A8. In a statistical test, the significance level α represents

(a) false positive
(b) true positive
(c) false negative
(d) true negative

A9. If the distribution of the variable Z is a unit Gaussian, then the distribution of Z^2 is a

(a) Chi-square distribution with one degree of freedom
(b) unit normal distribution
(c) t-distribution with n-1 degrees of freedom
(d) F-distribution with n-1 and m-1 degrees of freedom

A10. Identify the statistical test that does not require the data to be Gaussian?

(a) Welsch t-test
(b) Wilcoxon rank sum test
(c) ANOVA
(d) Chi-square goodness of fit test

PART B (answer any 4 out of 6) (4×3 Marks)

B1. In a class room consiting of 50 girls and 35 boys, it is observed that 25 girls and 19 boys have bicycles. Compute the probability that

(A) a student picked at random has a bicycle given that the student is a boy?

(B) a student picket at random has a bicycle given that the student is not a girl?

B2. If the aritmmetic mean of four numbers (22,31,n,50) is 36, compute their geometric mean.

B3. In a survey on certain population of mothers, 16 percent of them admitted that they have ckewed tobacco during pregnency. If 15 of them are chosen at random, what is the probability that exactly 7 among them would have chewed tobacco?

B4. If the mean number of serious accidents per year in a city block is 5, compute the probability that in the coming year there will be one or no accident in the same block.

B5. For a data set randomly drawn from a Gaussian distribution $N(\mu, \sigma)$, write down the expression for a confidence interval on population mean for a given significance level α. Explain the meaning of this interval.

B6. The weight of certain variety of Mango is known to follow a Gaussian distribution of population mean 160 grams. A random sample of 16 mangos from this population was found to have a sample mean of 172 grams with a standard deviation 35 grams. Compute the statistical significance of this data.

PART C (answer any 3 out of 5) (5×3 Marks)

C1. Compute the Pearson's correlation coefficient for the following two data sets and comment on the level of correlation:

$X = \{10, 20, 30, 40, 50\}$ and $Y = \{13, 24, 33, 48, 53\}$

C2. Scientists working in a lab wish to determine whether the lymphocytes and tumor cells in the biposy tissues of patients with certain cancer differ in their size. The following are the cell diameters (in μm) measured from patients with the disease:

Lymphosites : $\{8.9, 9.3, 4.6, 4.7, 8.9, 4.9, 8.4, 5.9, 6.3\}$

Tumor cells : $\{12.7, 15.2, 15.7, 23.9, 23.3, 17.1, 20.0, 21.0, 19.1\}$

Perform an appropriate t-test to determine whether, on an average, the two cells significantly differ in size.
Let $\alpha = 0.01$ be the level os significance. State your null and alternate hypothesis clearly.

C3. Let X1, X2, X3 and X4 represent the cholesterol level of women under 4 different age groups. Assume that these 4 distributions are Gaussian. We make 7 observations from these 4 groups and tabulate the data:

X1	:	221	213	202	183	185	197	162
X2	:	271	192	189	209	227	236	142
X3	:	262	193	224	201	161	178	265

X4 : 192 253 248 278 232 267 289

Perform an ANOVA for this data to test the null hypothesis that these 4 data sets have equal population mean.

At the significant level of 0.05, do you reject or accept the null hypothesis?

C4. In order to test the effectiveness of a training program, employees of a tech company were make to take a technical test before and after the training program. Their test scored in some scale are listed below:

Subject :	1	2	3	4	5	6	7	8	9	10	11
Score before training :	43	36	43	41	37	37	43	40	45	32	37
Score after training :	44	38	41	39	34	41	39	34	47	30	39

Perform Wilcoxon signed rank test on this data to determine whether training program increases the test score to a significance level of 0.05. State the null and alternate hypothesis clealrly.

C5. Chemists use Ion Sensitive Electrodes to measure ionic concentrations of acquous solutions. In order to calibrate this equipment, the output signal in millivolt was measured for known ion concentrations in units of ppm. The data is reproduced here:

concentration (in ppm) :	0.0	50.0	75.0	100.0	150.0	200.0
signal (in mV) :	1.72	2.11	2.36	2.56	3.05	3.42

Calculate a least square regression line between concentration and signal data.