LS4202:Biostatistics. End-Semester Test. Marks = 50.
Instructors: Dr. Robert J. Chandran and Dr. Dipjyoti Das
Date: May 5, 2019, **Time:** 10:00 AM - 12:30 PM

Section I: Answer all 3 questions. Marks: $3 \times 2 = 6$

1. Over more than 50 years, the number of students failing to appear in the final exams at a university each year was found to be a perfectly Poisson distributed random variable with $\mu = 5$. During last 5 years, out of the1000 students each year, exactly 4 did not turn up each year. Identify the values for:

(a) μ, (b) σ^2, (c) n, (d) $\bar{X}$, and (e) s^2

2. A student is told that the mean weight of a model animal is 120g, but that individuals in a new sample obtained in the lab appear heavier. She is asked to test this at $\alpha = 0.05$.

(a) State the null and alternative hypotheses
(b) Suppose that the mean weight is indeed 120g, but she is asked to repeat the same test 40 times on different samples. How many times do you expect she will draw the wrong conclusion and reject H_0 at $\alpha = 0.05$?

3. A student has been monitoring a physiological parameter for different samples of a mouse strain over several years and determined that the mean of the population is 70.65 with a 95 percent confidence interval of (67.2, 74.1). A new sample yields $\bar{X} = 75.32$.

(a) What is your conclusion concerning the $H_0 : \mu = 70.65$ with $\alpha = 0.05$.
(b) Is this a one-tailed or a two-tailed test? Explain why?

Section II: Answer all 5 questions. Marks: $5 \times 4 = 20$

4. In a binomial experiment, a student obtains 8 successes in 32 trials. The student is surprised as he was expecting close to 50% success rate. Show how you can use *Likelihood* based estimation to test the relevant hypotheses here. Choose your own confidence level.

5. You are given a series of nested hierarchical models m_0, m_1, m_2, . . . each with parameters $k_0 + 1$, $k_1 + 1$, $k_2 + 1$, . . . where $k_0 < k_1 < k_2$, . . . Show how you can do model selection (assuming the principle of parsimony) in the case of the models being General Linear Models and Generalized Linear Models.

6. (a) What is a test statistic? Suppose that a random sample is taken from a normal distribution. If you want to test if the estimated mean of that sample is equal to a given value, which distribution of the sample mean would you use for the following cases:
(i) both the population mean and variance are unknown,
(ii) The population mean is unknown, but the variance is known.

(b) Suppose that a workout session for an olympic athlete is designed to burn an average of 46 KCal. In a random sample of 12 such sessions, it was estimated that the sample mean was 42 KCal with an estimated standard deviation of 11.9. Does it suggest the workout session is configured lower than needed? Explain in detail how you would test this at the 0.05 level of significance.

7. (a) What is the distribution of the variable $\frac{n-1}{\sigma^2}S^2$, where S^2 is the sample-variance of a random sample of size n taken from a normal distribution of standard deviation σ? Use this distribution to solve the problem below.

(b) The following are the cell volumes 10 budding yeast cells: 46.4, 46.1, 45.8, 47.0, 46.1, 45.9, 45.8, 46.9, 45.2, and 46.0 μm^3. Explain in detail how would you find the 95% confidence interval for the variance of all such cells, assuming that cell volumes are taken from a normal distribution.

8. What is the Odds Ratio in logistic regression? Is there any relationship between the odd ratio and the slope of the logistic regression model?

Section III: Answer all 4 questions. Marks: $4 \times 6 = 24$

9. Consider a sample of n ordered-pairs (x_1, y_1), (x_1, y_1), . . . ,(x_n, y_n), where the x_i's and y_i's are independent. Further the x_is are fixed by the experimenter, while y_i's are corresponding response variables. Assuming a linear relationship between x and y, write the following:

Dipjyoti Das

(a) define the regression function for the linear model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, explain the assumptions of the model, and derive the *least squares* estimators for the coefficients of the model.

(b) In the estimate of the variance $s^2 = \frac{1}{(n-2)} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$, why do you need to divide by $(n-2)$?

(c) Suppose that the n ordered pairs appear to represent a power law $Y = aX^b$, can you still use a linear model for inference? How?

10. A severe drought led to extraordinary tree death rates in a forest ecosystem. A student monitored percentage loss of stem hydraulic conductance in 250 trees and recorded whether the tree died in the subsequent year or not. Trees typically tolerate significant loss of conductance but not beyond a certain value. Explain what statistical model you will use test the relationship between loss of conductance and tree death? What are the components, assumptions, and properties of such a model?

11. Design an ANOVA based experiment to test the *efficacy* of a new drug to treat high blood sugar in people. Let the blood sugar values be denoted by the random variable Y. Write down the ANOVA table and explain the components. How does one do statistical inference to test if the drug is effective at all? Now show how can you further test if the drug is equally effective for adult males and females?

12. (a) Males of a certain species of mosquito have lifespans that are strongly skewed to the right with a mean of 8 days and a standard deviation of 6 days. A biologist collects a random sample of 50 of these male mosquitos, and uses them to calculate the sample mean lifespan. We can assume that the mosquitos in each sample are independent. What will be the approximate shape of the sampling distribution of the sample mean lifespan? (only give the mathematical logic).

(b) Consider that three different samples of 500 random numbers are drawn from the following three distributions:
(i) $P(x) = \lambda e^{-\lambda x}$ $(x \geq 0)$
(ii) $P(x) = (\alpha - 1)x^{-\alpha}$ $(x \geq 1)$, with $\alpha = 2.5$
(ii) $P(x) = (\alpha - 1)x^{-\alpha}$ $(x \geq 1)$, with $\alpha = 3.5$

Does Central limit theorem hold true for all of them, *i.e.*, does the sample mean in all of them follow a normal distribution approximately? Give mathematical reason for your answer.

(c) Let X be a random variable, which follows an unknown distribution of mean $\mu = 10$ and standard deviation $\sigma = 4$. A sample of size 100 is taken from this population. Test the hypothesis that the sample mean of these 100 observations is less than 9. (It is given that a unit normal distribution has an area of 0.0062 at the left of $Z = -2.5$).