Assignment 1: Probability

Instructions

Complete the following problem set showing your work. Problems may be worked out "by hand" or in "python" or with the assistance of other analytical software (e.g., Mathematica, MatLab). You may use chatGPT to assist in coding.

Solutions must be type written (e.g., in Jupyter, markdown, or latex). Submit solutions as a PDF to Rebekah Hall (rah11@sfu.ca) by **11:59pm** on the **Saturday** of the corresponding week (see syllabus).

All problems are equally weighted within an assignment. Students in 468 may or may not choose to attempt the challenge question for a bonus 5pts. Students in 795 are required to complete the challenge question.

Problem Set

1. **Infectious Disease Testing**

A new infectious disease has emerged, and a diagnostic test has been developed to identify individuals who are infected. The test is not perfect and can yield both false positives and false negatives. You are given the following information:

- The prevalence of the disease in the population is 5%, meaning that 5% of individuals are infected.
- The sensitivity of the test is 90%, which means it correctly identifies 90% of infected individuals as positive.
- The specificity of the test is 85%, which means it correctly identifies 85% of uninfected individuals as negative.

Part A. Given that an individual tests positive, what is the probability that they are truly infected? Use Bayes' theorem to calculate this quantity which is known as the *positive predictive value* (PPV). Is this a high or low PPV?

Part B. Given that an individual tests negative, what is the probability that they are truly uninfected? Use Bayes' theorem to calculate the negative predictive value (NPV). Is this a high or low NPV?

Part C. Discuss the impact of test sensitivity and specificity on the accuracy of diagnostic tests, especially in the context of infectious diseases.

Part D. Suppose that individuals are tested twice. Assuming that the accuracy of the tests is independent, what is the probability that an individual is infected given two negative test results? Discuss possible pros and cons of a multiple-testing design

2. **Genetic Inheritance and Probability**

In genetics, the principles of probability are often used to understand the inheritance of traits. Consider a specific genetic trait determined by a single gene with two alleles: dominant (D) and recessive (d). In a population, 40% of individuals are homozygous dominant (DD), 30% are heterozygous (Dd), and 30% are homozygous recessive (dd) for this trait.

Part A. What is the probability that a randomly selected individual exhibits the dominant trait?

Part B. Given that two individuals heterozygous for this trait (Dd) mate and have offspring, what is the probability that their child will express the dominant trait)?

Part C. Consider a different co-dominant genetic trait determined by a separate (unlinked) gene. In the focal population, 60% of individuals express the mutant phenotype, 30% express the heterozygous phenotype, and 10% express the ancestral phenotype. What is the probability that a randomly selected individual is heterozygous for the first trait (Dd) and expresses the mutant phenotype (MM) for the second trait?

Part D. Discuss the importance of understanding and applying probabilities in the field of genetics and how they can aid in predicting the likelihood of trait inheritance and genotypic outcomes.

3. **Genetic Inheritance of a Recessive Trait**

In the study of genetics, we often encounter situations where we need to model the probability of specific genetic outcomes. Let's consider a simple genetic trait, where the presence of a recessive allele (r). An individual can inherit the allele from either parent. The probability of inheriting a recessive allele from each parent is represented by p, where p is known to be 0.25.

Part A. Define a random variable that represents the maternally inherited allele. What are the possible outcomes of this random variable, and what are their probabilities? This random variable is distributed according to what distribution?

Part B. Explain how you can model the inheritance of a recessive allele (r) from both parents using a Binomial distribution. What does the r.v. represent in this case? What are the parameters of the Binomial distribution in this context, and what do they represent?

Part C. Calculate the probability that an individual inherits two recessive alleles (rr), one from each from both parents for this trait.

4. **Moment Analysis in Measuring Tree Height in a Forest**

In forest ecology, understanding the height distribution of trees is crucial for assessing forest structure and biomass. Ecologists often use probability distributions to describe the distribution of tree heights. Let's consider a forest where tree heights follow a continuous probability distribution, f(h), where h is the height of a tree in meters.

The probability density function for tree heights in this forest is given by:

$$f(h) = khe^{-0.2h} \quad \text{for } 0 \leq h < \infty$$

where k is a constant that ensures the total probability integrates to 1.

Part A: What is the value of where k?

Part B: Calculate the mean tree height.

Part C: Calculate the variance in tree height distribution.

Part D. Calculate the third moment (skewness) of the tree height distribution. Discuss what the answers to A, B, and C tell you about the tree height distribution.

Part E: Plot the distribution of tree heights.

Part F: Suppose you want to compare the tree height distribution of this forest to another forest. How could you use moments (mean, variance, skewness) to assess and compare the ecological characteristics of these two forests based on tree height data?

5. **Random number generation**

Part A: Modify the Python code to generate a histogram of 500 exponentially distributed random numbers where $\lambda = 1$. Show the resulting plot.

Part B: Use the built in function *np.random.exponential* to generate 500 random numbers from an exponential distribution where $\lambda = \frac{1}{2}$. Show the resulting plot.

Part C: Plot the distributions from part A and part B on the same plot and use the result to compare the two distributions.

6. **Challenge Question: Modelling Disease Incubation Periods**

Understanding the incubation period of a disease is crucial in epidemiology, as it helps assess disease transmission dynamics and develop effective control measures. Let's consider a new infectious disease with a known incubation period, T during which individuals are infected, asymptomatic, but still infectious. The incubation period, denoted as f follows a continuous probability distribution with the following probability density function (PDF):

$$f(t) = 0.2e^{-0.2t} \quad \text{for } t \geq 0$$

Part A. Calculate the cumulative density function (CDF) of the incubation period, F(t). What does the CDF represent in the context of this infectious disease, and how can it be useful for epidemiologists?

Part B. Calculate the probability that an individual exposed to the disease will develop symptoms within the first 5 days ($0 \leq t \leq 5$). Use the CDF to find this probability.

Part C. Calculate the mean and skew of the incubation period using the PDF. What do these movements imply about disease transmission and control strategies?

Part D. Discuss the epidemiological relevance of using probability density and cumulative density functions to model disease incubation periods. How can understanding these functions help in predicting disease spread, estimating outbreak sizes, and planning public health interventions?