Introduction to Statistical Inference

Amirali Soltani Tehrani,
aa.soltanitehrani@gmail.com
Instructor: Mohammad-Reza A. Dehaqani
Deadline:
17 Aban 1402

I. Introduction

In this assignment, we will explore the fundamental concepts of probability theory and how it relates to the field of engineering. Probability is a fundamental concept in engineering that allows us to quantify uncertainties, make informed decisions, and assess risks. By understanding the principles of probability, we can analyze and model complex systems, predict outcomes, and optimize engineering processes. Throughout this assignment, we will dive into various topics, including basic probability theory, probability distributions, and statistical inference. We will learn how to calculate probabilities, work with random variables, and apply probability models to solve engineering problems. By mastering these concepts, we will develop a solid foundation for using probability in our future engineering endeavors.

The problems in this assignment are designed to challenge your understanding and problem-solving skills. They are carefully crafted to simulate real-world scenarios where probability is a critical factor. As you tackle each problem, take time to think through the concepts, apply relevant formulas, and provide clear and concise solutions. Remember, this assignment is not just about finding the right answers but also about understanding the underlying principles. Take the opportunity to discuss your approaches with your peers, as collaboration can often lead to deeper insights and different perspectives. I encourage you to ask questions, seek clarification, and engage in discussions to enhance your understanding of engineering probability. Your active participation will not only benefit you but also contribute to the collective learning experience of our class.

II. Mixed Problems

A. Problem 1

A device is equipped with a sensor that is connected to an alarming system. The sensor has a probability of 0.95 of triggering an alarm when dangerous conditions are present in a given day, and a probability of 0.005 of triggering an alarm when conditions are normal during the day. The probability of dangerous conditions occurring in a day is 0.005. Based on this information, we need to determine:

- What is the probability of a false alarm? In other words, what is the probability that conditions are normal when the alarm system triggers?
- What is the probability of an unidentified critical condition? In other words, what is the probability that conditions are dangerous when the system does not trigger?
- How many false alarms and unidentified critical conditions can be expected to occur during a 10-year period? Additionally, we need to comment on the effectiveness of the alarming system based on this information.

B. Problem 2

Consider two independent and identically distributed random variables, X_1 and X_2, with a common mean value of m and a common variance of σ^2. We need to determine the following:

- Find the mean value and variance of Y_1, where $Y_1 = X_1 + X_2$.
- Find the mean value and variance of Y_2, where $Y_2 = 2X_1$.
- Are the variances of Y_1 and Y_2 the same? If not, provide an intuitive explanation for the difference.
- Find the covariance between Y_1 and Y_2.

C. Problem 3

Excluding leap days, it's possible to assign numbers from 1 to 365 to each day of the year. Let's assume that any day of the year is equally likely for birthdays. Now, let's think about a group of n individuals, with you not being one of them. The sample space Ω represents all possible sequences of n birthdays (one for each person).

- Define the probability function P for Ω
- Let's examine the following events:
  A: "At least one person in the group shares your birthday." B: "At least two people in the group share a birthday." C: "At least three people in the group share a birthday."
  Now, we'll provide a detailed description of the subset of Ω that corresponds to each of these events.
- Determine a precise formula for the probability $P(A)$. What is the minimum value of n that makes $P(A)$ exceed 0.5?
- Provide a rationale for why n must be larger than $\frac{365}{2}$ without performing any calculations.
- Employ R or Python simulations to approximate the smallest value of n for which $P(B)$ exceeds 0.9. Conduct 10,000 trials for these simulations. To ensure the accuracy of the results, repeat the simulation a few times.

You observed minimal variation in the estimated probability of $P(B)$ when using 10,000 trials. To contrast this, attempt the simulation with only 30 trials and validate that the estimated probabilities exhibit greater variability.

- Determine a precise formula for the probability P(B).
- Utilize R or Python simulations to approximate the minimum value of n for which $P(C)$ exceeds 0.5. Employ 10,000 trials for this purpose. You'll notice that two values of n are equally valid outcomes. You have the flexibility to choose either one as your response.
  It's worth mentioning that finding an exact formula for $P(C)$ is significantly more challenging, making simulation a particularly useful approach in this case.
- In practice, the frequency of birthdays varies, with some dates being more common than others. You can access the data by clicking on the initial graph in the following link: http://chmullig.com/2012/06/births-by-day-of-year/. Upon reviewing the data, you'll notice that the author conducted a simulation of the birthday problem using real-life probabilities associated with different birthdates. The results showed that, for a fixed number of individuals, the likelihood of a shared birthday is slightly higher in reality compared to the equal-probability model discussed earlier. What makes this plausible? (Once more, we're seeking a concise response.)

D. Problem 4

These two questions were featured in a column by Martin Gardner in Scientific American in 1959.

- Mr. Jones is the parent of two children, and the older child is a girl. What is the likelihood that both children are girls?
- Mr. Smith is the parent of two children, and at least one of them is a boy. What is the probability that both children are boys?

E. Problem 5

In a city where there are one hundred taxis, one of them is painted blue, while the other 99 are green. During a hit-and-run incident at night, a witness claims to have seen a blue taxi leaving the scene and identifies it as the one involved in the incident. Consequently, the police arrest the blue taxi driver who was on duty that night. The driver asserts his innocence and has sought your legal representation to defend him in court. In order to build a case for reasonable doubt, you decide to conduct a test involving a scientist to assess the witness's ability to differentiate between blue and green taxis under conditions similar to the night of the accident. The collected data indicates that the witness correctly perceives blue cars as blue 99% of the time but misidentifies green cars as blue 2% of the time. Your task is to deliver a brief speech to the jury, aiming to provide them with sufficient doubt regarding your client's guilt. Your speech should be concise and clear, as most jurors may not have a background in this field. Using an illustrative table rather than complex formulas may aid their understanding.

F. Problem 6

nside a drawer, there are four dice: one with four sides (tetrahedron), one with six sides (cube), and two with eight sides (octahedra). Your friend discreetly selects one of these four dice at random. We'll denote the number of sides on the chosen die as S.

- What is the probability mass function (pmf) for S?

Now, your friend proceeds to roll the selected die without revealing the outcome to you. Let's denote the result of this roll as R.

- Utilize Bayes' rule to calculate the probability of S being equal to 4, 6, or 8, given that R equals 3. Which die is the most probable choice when R equals 3? In technical terms, you are determining the pmf of '$SgivenR = 3$.'
- Which die is the most likely choice when R equals 6?
- Which die is the most likely choice when R equals 7? No further calculations are necessary for this case.

G. Problem 7

Consider the following situation: X represents the outcome of rolling a fair 4-sided die, Y represents the outcome of rolling a fair 6-sided die, and Z stands for the average of X and Y.

- Determine the standard deviation for X, Y, and Z.
- Create a comprehensive probability mass function (pmf) and cumulative distribution function (cdf) for Z.
- Participate in a game with the following rules: You win $2X$ dollars if X is greater than Y, and you lose 1 dollar if X is not greater than Y. After engaging in this game for a total of 60 rounds, what is the expected overall gain (positive) or loss (negative)?

H. Problem 8

Raisin Bran cereal boxes have a height of 30 cm. However, due to the settling of contents, the density of raisins within the box varies from the bottom ($h = 0$) to the top ($h = 30$). Let's assume that this density (measured in raisins per cm of height) is described by the function $f(h) = 40 - h$.

- How many raisins can be found in one cereal box?
- Introduce a new random variable, H, representing the height of a randomly selected raisin from within the box. Determine and illustrate the probability density function (pdf) $g(h)$ for H.
- Determine and illustrate the cumulative distribution function (cdf) $G(h)$ for H.
- What is the probability that a randomly chosen raisin is located in the bottom third of the cereal box?

I. Problem 9

Assume X and Y are random variables with the following probabilities: $P(X = 1) = P(X = -1) = \frac{1}{2}$ and $P(Y = 1) = P(Y = -1) = \frac{1}{2}$. Let's denote c as the probability that both $X = 1$ and $Y = 1$.

- Calculate the joint distribution of X and Y, the covariance ($Cov(X,Y)$), and the correlation ($Cor(X,Y)$).
- Determine the values of c for which X and Y are independent. Also, find the values of c for which X and Y are 100% correlated.

J. Problem 10

Use R or Python for this question based on your choice.

- Generate enough number random data from 1) Uniform, 2) Normal, 3) Gamma, 4) Exponential, and 5) Binomal distributions and plot them in different graphs. (ex. with distplot)
- Generate the previous distribution for enough number of iterations, then plot the distribution of their mean values over all iterations. What distribution do you expect this sample mean should be? Justify your answer.
- Load *prob10.csv* file that is provided for you. Get familiar with each column of this dataset then go through the following questions.
  - Use data cleaning approaches on this dataset. Explain every method that you use for this dataset in your report.
  - Describe each column of this dataset. Is there any non-sense data in this dataset?
  - Use plot bar to show the frequency of each car manufacture in a single graph. Which company has the most cars?
  - Use a method to evaluate the dispersion of this dataset. Moreover, calculate the skewness and kurtosis of this dataset. What can these parameters tell you about this dataset?
  - Plot the scatter between engine-size and price value. Are these factors associated with each other?
  - Use *pairplot* for multivariate analysis of some factors in this dataset.
  - Use correlation over all entire numerical records and plot the heatmap for this correlation values.
  - For some categorical columns, plot the boxplot of these variables and calculate percentile, IQR, and whiskers for each of these columns.

K. Problem 11

Which one of the following variables is a random variable?

- Population mean.
- Population size
- Sample size
- Sample mean
- Variance of the sample mean
- The largest value in the sample
- Population variance

L. Problem 12

Use R or Python for this question based on your choice. Write a computer program to estimate the value of pi using a Monte Carlo method. This method involves generating random points within a square and determining how many fall within a quarter of a unit circle. By calculating the ratio of points inside the circle to the total number of points in the square, you can approximate the value of pi. Display the estimated value of pi and compare it to the actual value (math.pi). Also, display the difference between the estimated and actual values in a single plot.

Note: The more random points you generate (larger N), the more accurate your estimate will be. Try the above question using different values of N and plot the difference with an appropriate graph that you know.

M. Problem 13

Use R or Python for this question based on your choice. The Gambler's Ruin is a classical problem in probability theory that explores the concept of random walks and the likelihood of a gambler losing their entire stake over time. In this problem, you will create a computer program to simulate the Gambler's Ruin scenario and analyze the results. Here is the task:

- Create a program that simulates the Gambler's Ruin scenario. The scenario is as follows:
  1) A gambler starts with an initial stake (a certain amount of money).
  2) The gambler repeatedly bets a fixed amount on a game with a win probability of p (e.g., 0.5 for a fair coin toss).
  3) If the gambler wins a bet, their stake increases by the bet amount. If they lose, their stake decreases by the bet amount.
  4) The game continues until the gambler either reaches a desired target amount or loses their entire stake (goes broke).
- Implement the following steps in your program:
  1) Simulate the gambling scenario for a specified number of rounds (e.g., 1,000 rounds).
  2) Track the gambler's stake after each round.
  3) Calculate and display statistics, such as the probability of reaching the target amount before going broke.
  4) Report the mean value for more number of iterations and use an appropriate graph to visualize this metric.
- Extend your program to allow for different initial stakes, bet amounts, win probabilities, and target amounts.
- Visualize the results using charts, such as line plots showing the gambler's stake over time.

III. Submission

For the programming section, each student is required to submit a well-structured, typed PDF report that presents a concise summary of their analysis. The report should include the figures mentioned in the problem description and offer a detailed discussion of each. Please avoid uploading theoritical problem in .jpg format and upload them in a single .pdf file.

For each section of the report, a separate script is expected, which can be written in MATLAB (.m), Python 3 (.py or .py3), or R (.r). Avoid submitting scripts in formats like MATLAB live scripts, Python notebooks, or R Markdown. It is crucial that the submitted code is compatible with the grader's system. Be sure to include all relevant functions and any non-standard libraries used in your code.

The report should be treated as an academic piece of writing, and it should not contain any code snippets or explanations of coding logic. Instead, it should provide the author's insights about the results and demonstrate a strong grasp of the reference article. Academic reports typically maintain a concise and highly formal tone.

Each section of the report should briefly outline the hypothesis being tested. The responsibility for designing and implementing the tests lies with the students, as does explaining the results. Interpretations should be comprehensive without unnecessary verbosity.

The report can be written in either Persian or English, with no preference for either. In Persian reports, use B Nazanin with a font size of 14 for the text body and B Titr with a font size of 18 for titles. English reports should use Times New Roman 12 for the body text and Times New Roman 16 for titles. Sentences should be written in the passive tense. In Persian reports, the correct usage of the zero-width non-joiner is mandatory. In all reports, equations, figures, and tables must be labeled with unique numbers and referenced accordingly. Referring to figures as "the following figure," "the figure above," and similar expressions is considered incorrect.

Every figure in the report should be accompanied by a descriptive caption below it, while tables should have captions above them. Feel free to use footnotes and citations as necessary for clarity and proper attribution.

REFERENCES

[1] Monte Carlo Method https://en.wikipedia.org/wiki/Monte_Carlo_method. Wikipedia.
[2] Pi Number. https://www.britannica.com/science/pi-mathematics/. Britannica.
[3] Gambler's Ruin https://en.wikipedia.org/wiki/Gambler%27s_ruin. Wikipedia.