Statistical Inference

HW Author: Kamand Mesbah *Instructor: Mohammadreza A. Dehaqani*

Spring 2024

Homework 1

- If you have any questions about the homework, don't hesitate to drop an email to the HW Author.
- Feel free to use the class group to ask questions — our TA team will do their best to help out!
- Please consult the course page for important information on submission guidelines and delay policies to ensure your homework is turned in correctly and on time.
- This course aims to equip you with the skills to tackle all problems in this domain and encourages you to engage in independent research. Utilize your learnings to extend beyond the classroom teachings where necessary.

Problem 1: Sample spaces

A hospital administrator categorizes patients with gunshot wounds based on whether they possess insurance (1 for insured, 0 for uninsured) and their condition: good (g), fair (f), or severe (s). This constitutes an experiment.

a: Describe the sample space for this experiment.
b: Define event A as the patient's severe condition. List the possible outcomes in event A.
c: Define event B as the patient being uninsured. List the possible outcomes in event B.
d: Enumerate all possible outcomes if the patient is insured or in a severe condition. $(B^c \cup A)$

Problem 2: Sum of Normal Random Variables

a: Consider two independent normal random variables, X and Y, with means μ_1 and μ_2 respectively, and both sharing the same variance σ^2. Show that their sum, denoted as $T = X + Y$, follows a normal distribution $T \sim \mathcal{N}(\mu_1 + \mu_2, 2\sigma^2)$. Use a convolution integral, and you can use a standardization idea to reduce to the standard Normal case before setting up the integral.

b: Consider a sequence of independent normal random variables, denoted as X_i for $i = 1, \ldots, n$. Each random variable follows a normal distribution with mean μ_i and variance σ_i^2. Show that for any sequence of scalars $a_1, a_2, \ldots, a_n$,

$$\sum_{i=1}^{n} a_i X_i \sim \mathcal{N}\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right).$$

Problem 3: Joint and conditional probability mass functions

Consider the following game: initially, a coin is tossed once ($P(\text{heads}) = q$). If the coin lands tails, a 4-sided die is rolled; otherwise, a 6-sided die is rolled. The amount of money (in dollars \$) you win corresponds to the

outcome of the die roll. Let X be an indicator random variable for the coin toss ($X = 0$ if tails; $X = 1$ if heads), and let Y be the random variable corresponding to the amount of money won.

a: Calculate the joint probability mass function $p_{X,Y}$ as a function of q, the probability of heads on the coin.

b: Compute the conditional probability mass function $p_{X|Y}$, also as a function of q. Given that you win \$2 or less, determine the probability that the initial coin toss was headed, expressed as a function of q.

c: Assume you must pay \$3 each time you play the game. Determine, as a function of q, the average amount of money won or lost per game. At what value of q do you break even?

Problem 4: Conditional probabilities

Koichi's backyard pond contains g goldfish and c catfish. All fish are equally likely to be caught.

a: Suppose Koichi catches a total of k fish (without throwing any back). What is the probability of catching x goldfish?

b: Now, if all k fish are returned to the pond and Koichi starts fishing again, catching a total of m fish, what is the probability that among the m caught fish, exactly 2 goldfish are included from the first catch? (Assume that fish do not learn from experience.)

Problem 5: Variance, and covariance

Let $Y = 5X + \epsilon$, where ϵ follows a standard normal distribution ($\mathcal{N}(0, 1)$) and X follows a uniform distribution on the interval $(-1, 1)$. Assume that X and ϵ are independent.

a: Determine the mean and variance of Y.

b: Calculate the expected value of Y^2.

c: Find the conditional expected value of Y given $X = x$.

d: Determine the expected value of Y^3.

e: Calculate the covariance between ϵ and ϵ^2. Are ϵ and ϵ^2 independent?

f: Prove that for random variables X and Y, the covariance between the linear transformations $n + mX$ and $h + gY$ is equal to $mg\,\mathrm{Cov}(X, Y)$.

g: Find an upper bound for |Cov(X, Y)| using an inequality you learned in class.

Problem 6: Probability distribution

A card is randomly drawn from a standard deck of 52 playing cards. After each draw, the card is replaced, and the deck is reshuffled. This process continues indefinitely. What is the probability that exactly 3 out of the first five cards drawn are red?

Problem 7: Simulation techniques and analytical solutions

For part B, Use R or Python, but R is recommended.

Suppose there are n balls distributed randomly into n cells, where each cell can accommodate multiple balls.

a: Show that the probability of precisely one cell remaining empty is given by:

$$\frac{n(n-1)\binom{n}{2}(n-2)!}{n^n} = \frac{\binom{n}{2}n!}{n^n}$$

b: Define an R function 'sim1(n)' as:

$$\mathrm{sim1}(n) \leftarrow \mathrm{length}(\mathrm{unique}(\mathrm{sample}(1:n, n, \mathrm{replace} = \mathrm{TRUE})))$$

This function simulates the assignment of balls to cells and counts the number of occupied cells. Use 'sim1(n)' to compute the probability of precisely one empty cell for $n = 4, 6, 8, 10$ by simulation. Compare the simulated results with the values calculated using the formula in part (a). You may find the function 'replicate' helpful for this task. (Additionally, you are allowed to utilize Python.)

Problem 8: Cumulative distribution function

Let X be a non-negative random variable with cumulative distribution function (CDF) F. Show that

$$\mathbb{E}[X] = \int_0^\infty (1 - F(t))\, dt.$$

Hint: Consider expressing X as $X = \int_0^1 1_{\{t<X\}}\, dt$, and then assume the possibility of interchanging the order of expectation and integral.

Problem 9: Expected values

Let N represent the population of the state of Canicatti. Suppose that out of these people, a fraction pN support Jansen and $(1-p)N$ support Ezra, where $p \in (0, 1)$. The value of N is known (say $N = 3,000,000$), while p is unknown.

a: Assuming that each person in Canicatti independently decides whether or not to vote on election day, with a probability of 1/2 of voting and 1/2 of not voting, let V_{Jansen} be the number of people who vote for Jansen and V_{Ezra} be the number of people who vote for Ezra. Show that

$$\mathbb{E}[V_{\text{Jansen}}] = \frac{1}{2}pN, \quad \mathbb{E}[V_{\text{Ezra}}] = \frac{1}{2}(1-p)N.$$

Express the standard deviations of V_{Jansen} and V_{Ezra} in terms of p and N. Explain why, when N is large, we expect the fraction of voters who vote for Jansen to be very close to p.

b: Now, suppose there are two types of voters—passive and active. Each passive voter votes on election day with probability 1/4 and abstains with probability 3/4, while each active voter votes with probability 3/4 and abstains with probability 1/4. Let q_J represent the fraction of Jansen supporters who are passive, and $1 - q_J$ represent the fraction of Jansen supporters who are active. Similarly, let q_E represent the fraction of Ezra supporters who are passive, and $1 - q_E$ represent the fraction of active Ezra supporters. Show that

$$\mathbb{E}[V_{\text{Jansen}}] = \frac{1}{4}q_J pN + \frac{3}{4}(1-q_J)pN, \quad \mathbb{E}[V_{\text{Ezra}}] = \frac{1}{4}q_E(1-p)N + \frac{3}{4}(1-q_E)(1-p)N.$$

Express the standard deviations of V_{Jansen} and V_{Ezra} in terms of p, N, q_J, and q_E. Explain why estimating p by $\hat{p}$ using a simple random sample of $n = 1000$ people from Canicatti might not be a good estimate of the fraction of voters who will vote for Jansen.

c: Although q_J and q_E are unknown, suppose that in our simple random sample, we can determine whether each person is passive or active and inquire about their support for Jansen or Ezra. Propose estimators $\hat{V}_{\text{Jansen}}$ and $\hat{V}_{\text{Ezra}}$ for $E[V_{\text{Jansen}}]$ and $E[V_{\text{Ezra}}]$ using this additional information. Demonstrate that

$$\mathbb{E}[\hat{V}_{\text{Jansen}}] = \frac{1}{4}q_J pN + \frac{3}{4}(1-q_J)pN, \quad \mathbb{E}[\hat{V}_{\text{Ezra}}] = \frac{1}{4}q_E(1-p)N + \frac{3}{4}(1-q_E)(1-p)N.$$

Problem 10: Inequalities

a: **Chebyshev Inequality:** According to Chebyshev's inequality, for a given random variable X with an average value of μ and a variance of σ^2, the following holds when $k = 2$:

$$P(|X - \mu| \geq 2\sigma) \leq \frac{1}{4}.$$

This inequality indicates an upper bound on the probability, not the exact value, meaning the actual probability can significantly differ. For the specified distributions below, calculate the precise probability that $|X - \mu| \geq 2\sigma$.

- Let X be a continuous random variable whose probability density function (pdf) is defined as $f(x) = \frac{1}{2}$ within the interval $0 \leq x < 2$, and $f(x) = 0$ for all other values.
- X is a discrete random variable with a probability mass function (pmf) given by $p(1) = p(-1) = \frac{1}{2}$, and 0 everywhere else.

b: **Markov's inequality:** Consider a random variable X for which the fourth moment about the mean, μ, denoted as $E[(X - \mu)^4]$, is defined. Prove that for any positive constant c, the following inequality holds:

$$P(|X - \mu| > c) \leq \frac{E[(X - \mu)^4]}{c^4}.$$

c: **(Extra Point) Jensen's inequality:** Consider two distributions, P and Q, defined on the same measurable space X. Assume there exists a measure μ such that both P and Q are absolutely continuous with respect to μ (a common choice for μ could be $P + Q$). Let $p = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$ represent the Radon-Nikodym derivatives of P and Q with respect to μ, serving as the respective density functions. The Kullback-Leibler divergence from P to Q is given by

$$D_{\mathrm{KL}}(P\|Q) = \int_X p(x) \log\left(\frac{p(x)}{q(x)}\right) d\mu(x).$$

Demonstrate that $D_{\mathrm{KL}}(P\|Q)$ is non-negative, and it equals zero if and only if the distributions P and Q are identical. [Hint: Use Jensen's inequality, acknowledging that a convex function f satisfies $f''(t) > 0$ for almost every t, which indicates strict convexity.]

Problem 11: Different distribution contexts

a: A fire station will be located along a finite-length road A. Where should the station be located to minimize the expected distance from the fire if fires occur at points uniformly chosen on $(0, A)$? In other words, choose a to minimize $E[|X - a|]$, where X is uniformly distributed over $(0, A)$.

b: suppose the road is infinite, stretching from point 0 outward to ∞. Where should the fire station be located if the fire distance from point 0 is exponentially distributed with rate λ? We want to minimize $E[|X-a|]$, where X is now exponential with rate λ.

Problem 12: Bonus question: Defend the innocent

In a city with one hundred taxis, one 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. Consequently, the police arrested the blue taxi driver on duty that night. The driver asserts his innocence and has sought your legal representation to defend him in court. 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 percent of the time but misidentifies green cars as blue 2 percent. Your task is to deliver a brief speech to the jury, to give them 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.
Note: Articulate a compelling defense for the accused using clear and convincing language. This bonus question makes up 10% of your grade and can cover mistakes and shortcomings so write a well-presented argument, emphasize reasonable doubt in your narrative, avoiding complex wording to ensure the jury's comprehension.

Problem 13: Exploratory Data Analysis

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) Binomial distributions and plot them in different graphs. (ex. with distplot)
• Generate the previous distribution for enough iterations, then plot the distribution of their mean values over all iterations. What distribution do you expect this sample to be? Justify your answer.
• Load the prob10.csv file that is provided for you. Get familiar with each column of this dataset, then go through the following questions.
a: Use data cleaning approaches on this dataset. Explain every method that you use for this dataset in your report.
b: Describe each column of this dataset. Is there any non-sense data in this dataset?
c: Use a plot bar to show the frequency of each car manufacturer in a single graph. Which company has the most cars?
d: Use a method to evaluate the dispersion of this dataset. Moreover, the skewness and kurtosis of this dataset are calculated. What can these parameters tell you about this dataset?
e: Plot the scatter between engine size and price value. Are these factors associated with each other?
f: Use pair plots for multivariate analysis of some factors in this dataset.
g: Use correlation overall numerical records and plot the heatmap for these correlation values.
h: For some categorical columns, plot the boxplot of these variables and calculate percentile, IQR, and whiskers for each of these columns.

Problem 14: Monte Carlo Method

Use R or Python for this question based on your choice.
Read about the Monte Carlo method and shortly explain it and it's mathematical concept. Now, 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), your estimate will be more accurate.
Try the above question using different values of N and plot the difference with an appropriate graph you know.

Problem 15: Random Walks and Losing Likelihood

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 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 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.

Good Luck!