CS114. Assignment 1. Discrete Random Variables.

Instructions

- Complete all the problems below and show your work.
- In addition to the LOs listed next to each problem, you will be graded on #MathTools, #CompTools, #Communication, and #Professionalism for the overall quality of your text, math, and code.
- You may upload one Python notebook containing all your work, or a PDF with your text and math and a Python notebook with your code separately. No handwritten submissions.

Problem 1. Coin Spins (#Probability)

One day you overhear how two fellow Minervans discuss a show in which a performer spins a coin many times, resulting in Tails 30 times in a row. You hear them discussing:

Yueh Han: "That outcome would be extremely unlikely with fair coins. They must be using trick coins (maybe double-tailed coins), or the experiment must have been rigged somehow (maybe with magnets)."

Akma: "It's true that the string `TT...T` of length 30 is very unlikely; the chance is $(\frac{1}{2})^{30} \approx 9 \times 10^{-10}$ with fair coins. But any other specific string of $\mathtt{H}$ and $\mathtt{T}$ with length 30 has exactly the same probability! The reason the outcome seems extremely unlikely is that the number of possible outcomes grows exponentially as the number of spins grows, so any outcome would seem extremely unlikely. You could just as well have made the same argument even without looking at the results of their experiment, which means you really don't have evidence against the coins being fair."

Help Yueh Han and Akma resolve their debate.

1. Discuss the conversation of Yueh Han and Akma. With whom do you agree and why?

2. Suppose there are only two models: either the coins are all fair (and spun fairly), or double-tailed coins are being used in which case the probability of Tails is 1. Let p be the prior probability that the coins are fair. Find the posterior probability that the coins are fair, given that they landed Tails in 30 out of 30 trials.

3. For which values of the prior, p, is the posterior probability that the coins are fair greater than 0.5? (What does the prior need to be to make the fair-coin model more probable according to the posterior?)

Problem 2. Medical test (#Probability)

Omer wants to be really certain about a diagnosis so he takes a series of n identical medical tests. He hopes multiple tests will reduce his uncertainty.

Events:

- D: Omer has the disease being tested for.
- T_j: Omer tests positive on the j^{th} test for $j = 1, 2, \ldots, n$.

Let $p = P(D)$ be the prior probability that he has the disease.

1. Assume for this part the test results are conditionally independent given Omer's disease status. Let $a_0 = P(T_j \mid D)$ and $b_0 = P(T_j \mid D^c)$, where a_0 and b_0 don't depend on j.

   Find the posterior probability that Omer has the disease, given that he tests positive on all n of the n tests.

   Hint: Since a_0 does not depend on j (the index of a test result), the algebra simplifies a lot since $P(T_i \mid D)\, P(T_j \mid D) = a_0^2$ for any test indices i and j. The same holds for b_0.

2. Suppose some people have a gene that makes them always test positive on this type of medical test. Let G be the event that Omer has the gene. Assume that $P(G) = \frac{1}{2}$ and that D and G are independent — that is, the gene does not make you more or less susceptible to the disease.

   If Omer has the gene, he'll test positive on all n tests.

   If Omer does not have the gene, then the test results are conditionally independent given his disease status. Let $a_1 = P(T_j \mid D, G^c)$ and $b_1 = P(T_j \mid D^c, G^c)$, where a_1 and b_1 don't depend on j.

   Now, suppose that Omer tests positive on all n tests and find the posterior probability that Omer has the disease.

3. Using the same setup as in part (2.), find the posterior probability that Omer has **the gene** given that he tests positive on all n of the tests.

Problem 3. Proofreading (#Distributions, #ModelSelection)

A book has n typos. Two proofreaders, Sho and Haruna, independently read the book. Sho finds each typo independently with probability p_1, and Haruna with probability p_2. Let X_1 be the number of typos caught by Sho, X_2 be the number caught by Haruna, and X be the number caught by at least one of the two proofreaders.

1. Find the distribution of X. Motivate your answer.

2. Assuming $p_1 = p_2$, find the conditional distribution of X_1 given that $X_1 + X_2 = t$. That is, if we know t typos were found, how many were found by the first proofreader?

   Hint: $X_1 + X_2 \neq X$. Note that we are conditioning on the number of typos found by Sho plus the number found by Haruna. This is not the same as the number of typos found by at least one of the two proofreaders — we are double-counting the typos found by both of them.

3. You write a book with 100,000 words. On average, you make a typo once every 300 words. Sho and Haruna proofread your book. Sho finds 299 typos while Haruna finds 314 typos. Use model selection to decide whether Haruna is better than Sho at finding typos or whether it is just a coincidence that she found more than Sho this time.

   Hints:

   - Follow the frequentist model selection recipe from the Session 12 Study Guide.
   - For the null model, think carefully about which distribution to use for the number of typos that exist in the book.
   - You'll need your answer to part (2.) above.
   - You will need to implement a simulation for this problem. Don't rely on looking up a known distribution of the test statistic.

4. **Stretch goal.** This part is entirely optional — there is no penalty for not attempting it and there is no penalty for getting it wrong if you choose to try it. If you provide a **great, well-written solution** to this problem, you will get an **extra score of Ⓚ** on #ParameterEstimation. If you fall short of a Ⓚ, you can still earn a Ⓐ for a slightly flawed solution but the errors have to be minimal. You cannot score less than Ⓐ on this problem — instead you will get no extra grade if the attempt is insufficient. This is a hard problem and it is possible that nobody will get it right.

   Problem: Use the same setup as in part (3.) above but with the following additional information: of the 299 typos Sho found and the 314 Haruna found, 285 were found by both of them. This means Sho found (299 − 285) = 14 typos that Haruna did not find and Haruna found (314 − 285) = 29 typos that Sho did not find.

   Question: How many typos are left in the book, found by neither Sho nor Haruna?

   Task: Address this question by estimating all of the following unknown variables using Bayesian parameter estimation — p_1 (the probability that Sho finds a typo), p_2 (the probability that Haruna finds a typo), and t (the total number of typos in the book). You can then answer how many of the t typos were not found by either proofreader.

   Why is this hard? We don't know for certain how many typos there are in the book! We only know how many were found by Sho or Haruna (or both) but not how many typos were not found by them. However, with everything we covered in class till now, you have enough tools to estimate the probabilities that the proofreaders find typos and therefore how many of the overall typos they found. Good luck!

Problem 4. Randomized response surveys (#ParameterEstimation)

A researcher wants to estimate the percentage of people in some population who have used illegal drugs by conducting a survey. Concerned that a lot of people would lie when asked a sensitive question like "Have you ever used illegal drugs?", the researcher uses a method known as *randomized response*. A box is filled with slips of paper, each of which says either "I have used illegal drugs" or "I have not used illegal drugs".

- Let p be the proportion of slips of paper that say "I have used illegal drugs". p is chosen by the researcher in advance.

Each participant chooses a random slip of paper from the hat and answers "yes" or "no" to whether the statement on that slip is true. The slip is then returned to the hat. The researcher doesn't know which type of slip the participant had.

- Let y be the probability that a participant will say "yes",
- and d be the probability that a participant has used illegal drugs.

1. Find y in terms of d and p.

2. Given that the researcher is interested in the true proportion of people using illegal drugs, d, what would be the worst possible choice of p that the researcher could make in designing the survey? Explain.

Now consider the following alternative system. Suppose that proportion p of the slips of paper say "I have used illegal drugs", but that now the remaining $(1 - p)$ say "I was born in winter" rather than "I have not used illegal drugs". Assume that 1/4 of people are born in winter, and that a person's season of birth is independent of whether they have used illegal drugs.

3. Find d, in terms of y and p.

4. A randomized response survey like in part (3.) is conducted with 100 participants and $p = \frac{1}{2}$.

   Each participant draws a slip of paper from a box (with replacement) and answers the question on the paper truthfully. At the end of the survey, the responses are counted and there are 20 "yes" responses and 80 "no" responses.

   Use part (3.) to compute the proportion of participants who use drugs. Note that this result is a point estimate without a confidence interval or credible interval.

   Next, write a simulation to compute a Bayesian posterior over d, the fraction of people who are drug users given the information in the problem. Compare your posterior histogram with the value of d computed above.

   Hints:

   - Generate the birth month, drug user state, and which question they get for every participant in the study.
   - Count the "yes" responses and condition on getting 20 of them.
   - Record how many users were drug users.
   - You will have to come up with a reasonable prior for the number of drug users among the 100 participants. Any reasonably motivated choice of prior will be accepted.