Chris Piech
CS 109

Problem Set #4
Oct 27, 2021

Problem Set #4
Due: 1pm on Fri, Nov 5th

For each problem, briefly explain/justify how you obtained your answer. Make sure to describe the distribution and parameter values you used (e.g., Bin(10, 0.3)), where appropriate. Provide a numeric answer for all questions when possible.

Warmup

1. You roll 6 dice. How much more likely is a roll with: [one 1, one 2, one 3, one 4, one 5, one 6] than a roll with six 6s?

2. The joint probability density function of continuous random variables X and Y is given by: $f(X = x, Y = y) = \frac{4y}{x}$ where $0 < y < x < 1$

   a. What is the marginal density function of X?
   b. What is the marginal density function of Y?
   c. What is $E[X]$?

3. Let X_i be the number of weekly visitors to a web site in week i, where $X_i \sim N(2200, 52900)$ for all i. Assume that all X_i are independent of each other. What is the probability that the weekly number of visitors exceeds 2000 in at least 2 of the next 3 weeks?

4. You think your baby might be tired, and you estimate this prior belief to be $P(\text{Tired}) = \frac{3}{4}$. If a baby is tired, the time in minutes until they rub their eyes is distributed as Exp$(\lambda = 3)$. If a baby is not tired, the time in minutes until they rub their eyes is distributed as Exp$(\lambda = 1)$. A baby rubs their eyes after 2 mins. What is your updated belief that they are tired?

5. You are developing medicine that sometimes has a desired effect, and sometimes does not. With FDA approval, you are allowed to test your medicine on 9 patients. You observe that 7 have the desired outcome. Your belief as to the probability of the medicine having an effect before running any experiments was Beta(2, 2).

   a. What is the distribution for your belief of the probability of the medicine being effective after the trial?
   b. Use your distribution from (a) to calculate your confidence that the probability of the drug having effect is greater than 0.5. You may use scipy.stats or an online calculator.

Music Tastes

6. Write a program that reads the data file music.csv and estimates the answers to the following questions. Each row in the csv represents one person and their corresponding ratings of different music types, on a scale of 1-5. For each question write the mathematical formula you used to compute the answer, and include the numeric estimate (see Name2Age for an example). Let R_i be a random variable for the rating a user gives to genre i. You may either use a Frequentest or a Bayesian approach to estimating probabilities from data:

   a. What is $P(R_{\text{Folk}} = 5)$
   b. What is $P(R_{\text{Folk}} = x)$
   c. What is $E[R_{\text{Musical}}]$
   d. What is $P(R_{\text{Folk}} = 5 | R_{\text{Musical}} = 5)$
   e. What is $P(R_{\text{Folk}} = x | R_{\text{Musical}} = 5)$
   f. What is the covariance of R_{Opera} and R_{Punk}?

Biometric Keystrokes

7. Did you know that computers can know who you are not, just by what you write, but also by how you write it? Coursera uses Biometric Keystroke signatures for plagiarism detection. If you can't write a sentence with the same statistical distribution of key press timings as in your previous work, they assume that it is not you who is sitting behind the computer. In this problem we provide you with three files:

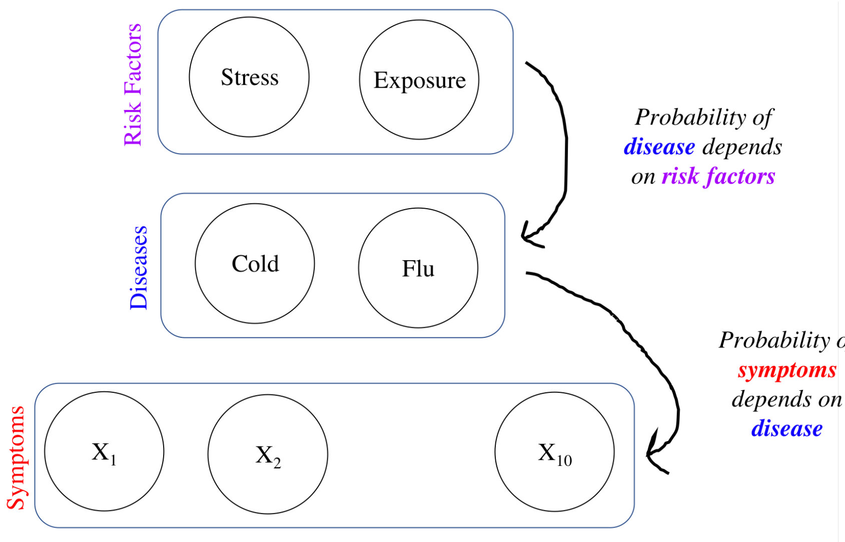
   - personKeyTimingA.txt has keystroke timing information for a user A writing a passage. The first column is the time in milliseconds (***since the start of writing***) when the user hit each key. The second column is the key that the user hit.
   - personKeyTimingB.txt has keystroke timing information for a second user (user B) writing the same passage as the user A. Even though the content of the passage is the same *the timing* of how the second user wrote the passage is different.
   - email.txt has keystroke timing information for an unknown user. We would like to know if the author of the email was user A or user B.

   Let X and Y be random variables for the duration of time, in milliseconds, for users A and B (respectively) to type a key. Assume that each keystroke from a user has a duration that is an independent random variable with the same distribution.

   a. Estimate $E[X]$ and $E[Y]$
   b. Estimate $E[X^2]$ and $E[Y^2]$
   c. Use your answers to part (a) and (b) and approximate X and Y as Normals with mean and variance that match their biometric data. Report both distributions.
   d. Calculate the ratio of the probability that user A wrote the email over the probability that user B wrote the email. You don't need to submit code, but you should include the formula that you attempted to calculate and a few sentence description of how your code works.

WebMD

8. We are writing a WebMd program that is slightly larger than the one we worked through in class. In this program we predict whether a user has a flu ($F = 1$) or cold ($C = 1$) based on knowing any subset of 10 potential binary symptoms (eg headache, sniffles, fatigue, cough, etc) and a subset of binary risk factors (exposure, stress).


Risk Factors
Stress
Exposure
Diseases
Cold
Flu
Symptoms
X_1
X_2
X_{10}
Probability of ***disease*** *depends on* ***risk factors***
Probability of ***symptoms*** *depends on* ***disease***

Write psuedocode that calculates the probability of flu *conditioned on observing* that the patient has had exposure to a sick friend and that they are experiencing symptom 2 (sore throat). In terms of random variables $P(\text{Flu} = 1 \mid \text{Exposure} = 1 \text{ and } X_2 = 1)$:

```
def probFlu(): # P(Flu = 1 | Exposure = 1 and X_2 = 1)
```

We know the prior probability for Stress is 0.5 and Exposure is 0.1.

You are given functions `probCold(s, e)` and `probFlu(s, e)` which return the probability that a patient has a cold or flu, given the state of the risk factors stress (`s`) and exposure (`e`).

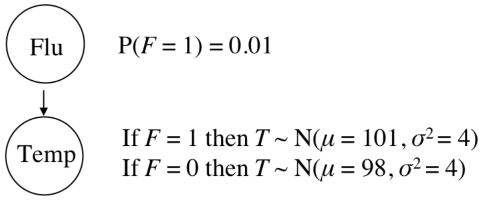
You are given a function `probSymptom(i, f, c)` which returns the probability that the `i`th symptom (X_i) takes on value 1, given the state of cold (`c`) and flu (`f`): $P(X_i = 1|F = f, C = c)$.

a. Write psuedocode that calculates `probFlu()` using **Rejection Sampling**.
b. (Extra Credit) Write pseudocode that calculates `probFlu()` without using sampling.

Cheeky Rejection Sampling

9. In class we observed that Rejection Sampling didn't work well when we asked probability questions that conditioned on rare events. Rejection Sampling is especially problematic if you are conditioning on a continuous variable, where the probability of any exact value is zero.

To get around our inability to condition on continuous random variables we introduce Cheeky Rejection Sampling. Cheeky Rejection Sampling is just like regular Rejection Sampling, except we consider a continuous random variable assignment to “match" a conditioning event if the values **are within 0.1** of each other. We can explore Cheeky Rejection Sampling with this very small Bayesian Network:


Flu
$P(F = 1) = 0.01$
Temp
If $F = 1$ then $T \sim N(\mu = 101, \sigma^2 = 4)$
If $F = 0$ then $T \sim N(\mu = 98, \sigma^2 = 4)$

Flu is a Bernoulli random variable with $p = 0.01$ and Temp is a Normal random variable with μ that changes based on whether the person has a flu.

As an example of the cheeky difference: say we are calculating $P(\text{Flu} = 1|\text{Temp} = 100)$ and we generate a sample with Temp = 100.0001. In the Cheeky algorithm that sample would count as matching the conditioned event. In regular Rejection Sampling it would not since $100.0001 \neq 100$.

a. What is the probability that a person with the flu has a fever greater than 103? You must give a numerical answer for full credit.
b. For this simple model, we can calculate exact probabilities. What is the probability that someone has a flu given they have a temperature of 100?
c. Write pseudocode to calculate the probability that someone has a flu given a fever of 100. Use Cheeky Rejection Sampling with $N = 10000$ samples. Use the functions:

Method	**Description**
`bern(p)`	Returns a random sample from a Bernoulli
`gauss(`μ`, `σ^2`)`	Returns a random sample from a Normal

d. What is the probability that out of 10,000 iterations of Cheeky Rejection Sampling for our model, there are < 20 samples that have a temperature within 0.1 degrees of 100 degrees? Use an approximation to compute a numerical answer.

Learning While Helping

10. You are designing a randomized algorithm that delivers one of two new drugs to patients who come to your clinic—each patient can only receive one of the drugs. Initially you know nothing about the effectiveness of the two drugs. You are simultaneously trying to learn which drug is the best and, at the same time, cure the maximum number of people. To do so we will use the Thompson Sampling Algorithm.

Thompson Sampling Algorithm: For *each* drug we maintain a Beta distribution to represent the drug's probability of being successful. Initially we assume that drug *i* has a probability of success: $\theta_i \sim \text{Beta}(1, 1)$.

When choosing which drug to give to the next patient we **sample** a value from each Beta and select the drug with the largest **sampled** value. We administer the drug, observe if the patient was cured, and update the Beta that represents our belief about the probability of the drug being successful. Repeat for the next patient.

a. Say you try the first drug on 7 patients. It cures 5 patients and has no effect on 2. What is your belief about the drug's probability of success, θ_1? Your answer should be a Beta.

Methods
`V = sampleBeta(a, b)` Returns a real number value in the range [0, 1] with probability defined by a PDF of a Beta with parameters a and b.
`R = giveDrug(i)` Gives drug i to the next patient. Returns a True if the drug was successful in curing the patient or False if it was not. Throws an error if $i \notin \{1, 2\}$.
`I = argmax(list)` Returns the index of the largest value in the list.

b. Write pseudocode to administer either of the two drugs to 100 patients using Thompson's Sampling Algorithm. Use functions from the table above. Your code should execute `giveDrug` 100 times.

c. After running Thomspons' Algorithm 20 times, you end up with the following Beta distributions:

$\theta_1 \sim \text{Beta}(3, 4)$,
$\theta_2 \sim \text{Beta}(13, 4)$

What is the probability that $\theta_2 > \theta_1$? You may provide an approximate answer. Explain briefly how you obtained your answer.