CSE 312: Foundations of Computing II Summer 2022

Problem Set 6 (due Friday, August 5, 11:59pm)

Directions: *For each problem, explain/justify how you obtained your answer, as correct answers without an explanation may receive **no credit**. Moreover, in the event of an incorrect answer, we can still try to give you partial credit based on the explanation you provide. Unless you are asked to, you should leave your answer in terms of factorials, combinations, etc., for instance 26^7 or $26!/7!$ or $26 \cdot \binom{26}{7}$.*

***Submission**: You must upload a **pdf** of your solutions to Gradescope under "Pset 6 [Written]". The use of LaTex is highly recommended, and we have provided a template. (Note that if you want to hand-write your solutions, you'll need to scan them. If we cannot make out your writing, your work may be ungradable, so make sure it is legible.) There is no coding for this homework.*

Instructions as to how to upload your solutions to Gradescope are on the course web page.

*Remember that you must tag your written problems on Gradescope, or you will potentially receive **no credit** as mentioned in the syllabus. Please put each numbered problem on its own page in the pdf (this will make selecting pages easier when you submit), and ensure that your pdfs are oriented correctly (e.g. not upside-down or sideways). As stated above, the coding problem will also be submitted to Gradescope.*

***Collaboration**: This pset must be submitted **individually**. You are welcome and encouraged to discuss approaches with your fellow students, but everyone **must write up their own solutions.** Failure to do so is an instance of academic dishonesty.*

1. Joint Densities (26 points)

Suppose X, Y are jointly continuous rv's with joint density

$$f_{X,Y}(x,y) = \begin{cases} cxy^2 & x > 0, y > 0, x + y < 1 \\ 0 & \text{otherwise} \end{cases}$$

Your answers below should **not** be evaluated. Your answers should usually be in terms of (double) integrals.

(a) [4 Points] Write an expression for c, but do not solve it or compute any integrals.

(b) [4 Points] Write an expression which we can evaluate to find $\Pr(Y \geq X)$. Hint: draw the region of the joint density, and the desired region.

(c) [4 Points] Write an expression which we can evaluate to find the marginal density $f_X(x)$. Specify the value of $f_X(x)$ for all $x \in \mathbb{R}$.

(d) [4 Points] Write an expression for the joint CDF $F_{X,Y}(s,t)$. Specify its value for all $s, t \in \mathbb{R}$.

(e) [6 Points] Are X and Y independent? Justify your answer.

(f) [4 Points] Suppose V, W, X, Y, Z are jointly continuous with joint PDF $f_{V,W,X,Y,Z}(v,w,x,y,z)$. Write an expression for the joint marginal PDF $f_{V,X,Z}(v,x,z)$.

2. Coal Mining (22 points)

A miner is trapped in a mine containing 4 doors, and each door is equally likely to be chosen. The first door leads to a tunnel that will take him to safety after a number of hours which is Poisson with parameter 2. The second door leads to a tunnel that will take him to safety after a number of hours which is Geometric with parameter $\frac{1}{5}$. The third door leads to a tunnel that will take him to safety after a number of hours which is binomial with parameters $n = 100$ and $p = 1/20$. The fourth door leads to a tunnel which brings him back to where he started after 2 hours. Use the law of total expectation to compute the expected number of hours until the miner reaches safety.

3. Statistics Books (30 points)

Alice is going shopping for statistics books for H hours, where H is a random variable, equally likely to be 1, 2 or 3. The number of books B she buys is random and depends on how long she is in the store for. We are told that

$$\text{Pr}(B = b \mid H = h) = \frac{c}{h}, \qquad \text{for } b = 1, \dots, h,$$

for some constant c.

(a) [4 Points] Compute c. You may want to use one of the axioms of probability.

(b) [4 Points] Find the joint distribution of B and H using the chain rule.

(c) [6 Points] Find the marginal distribution of B.

(d) [4 Points] Find the conditional distribution of H given that $B = 1$ (i.e., $Pr(H = h|B = 1)$ for each possible h in 1,2,3). Use the definition of conditional probability and the results from previous parts.

(e) [6 Points] Suppose that we are told that Alice bought either 1 or 2 books. Find the expected number of hours she shopped conditioned on this event. Use the definition of conditional expectation and Bayes Theorem.

(f) [6 Points] The cost of each book is a random variable with mean $3(1 - 0.05(b - 1))$, where b is the number of books bought. So if she buys one book, its expected cost is \$3. If she buys two books the expected cost of each book is \$ $3(1 - 0.05)$ and if she buys three books, the expected cost of each book is \$$3(1 - 0.1)$. What is the expected amount of money Alice spends?
Warning: Be sure to use a formal derivation. Your work should involve the law of total expectation conditioning on the number of books bought, and make use of random variables X_i, where X_i is the amount of money she spends on the ith book she purchases.

4. Improving Models [Real World] (22 points)

(a) [10 Points] Find an analysis that uses probability and statistics tools you're familiar with from this course. By "analysis," we mean any estimate of a "real-world" probability, along with the assumptions that lead to that number. You might want to look at the examples in the final section for what we mean.

We expect most of the answers to this section will be short (2-3 sentences), but you are free to write more if your resource is more complicated.

(a) Provide a link to (or somehow let us access) the analysis you're critiquing.

(b) What is the fundamental claim of the analysis? I.e., what conclusion do they draw at the end of their analysis?

(c) What modelling assumptions do they use? (For example, do they assume some occurrences are independent? Do they assume a set of events all have equal probability? Do they assume they know the probability? Do they use a variable from the zoo?)

(b) [12 Points] Now, see if their modelling assumptions are reasonable or if other ones would lead to a different conclusion. We expect parts a, b, d will be a few sentences each (though you can write more if you have more to say).

 (a) Identify at least one weakness of the modelling assumptions they have made (e.g. a potential dependence on events that are supposed to be independent).

 (b) Now create your own model for the same problem. You might do this by coming up with different probability estimates for the events, or by using a different random variable (e.g. a binomial distribution instead of a Poisson), or by incorporating some outside knowledge about the problem that you think sheds more light. Briefly describe what your model will be, and how it differs from the previous one.

 (c) Under your new model, calculate the probability of the event your source calculated.

 (d) Does the calculation change significantly? If it does, does the conclusion of the analysis change?

(c) [0 Points] The following is a sample analysis to give you an idea of what we are looking for.

i) a) The dataset we use is about the upcoming Chess championship: Carlsen vs. Nepomniachtchi: What Do The Numbers Say?

 b) The analysis is about the upcoming 2021 world chess championship tournament in which current titleholder Magnus Carlsen will play challenger Ian Nepomniachtchi. The tournament works as follows: they play 14 games, and the winner of each game wins a point, whereas a tie is worth 0.5 points. A player needs 7.5 points to win the series. The article concludes that, based on elo-ratings, Magnus has a 72% chance of winning (this assumes they do not tie in the 14-game series).

 c) The articles final analysis assumes that the FIDE rating, which aggregates scores between a players entire match career, is a good predictor of future performance against a specific opponent. They assume that each future match is a direct consequence of current FIDE rating - e.g. that each previous match result is independent of future matches given the FIDE rating of both players.

ii) a) One thing the analysis does not take into account is past player to player match history, which is not necessarily captured by overall FIDE rating, which summarizes a players total past performance against all opponents. That is, a future match between Nepo and Carlsen may not be independent of past matches between those two players given the current FIDE rating of both players, because personal play style factors into match results.

 b) In previous head-to-head matches, Nepo has a 4-1-6 match history against Magnus in the classical setting, meaning that he has won against Magnus 4 times, lost once, and drawn 6 times in classical games. If we take into account only personal history, Nepo has a $\frac{4}{11} \approx 0.36$ chance to win, $\frac{1}{11} \approx 0.1$ chance to lose and $\frac{6}{11} \approx 0.54$ chance to tie.

 c) We will assume that each game has an independent outcome of 0.54 chance to tie, 0.1 chance for a win by Magnus, and a 0.36 chance to win by Nepo.

 d) To compute the probability that Magnus wins the overall 14-game series, we need to break it down into all the possible ways he can win. If he needs 7.5 points that means: he can get at least 8 wins, or he can get 7 wins and at least 1 tie, or 6 wins and at least 3 ties ... and so on. In order to write this out, we will use $B(n, p, x)$ to denote the PDF of a binomial random variable with parameters n, p is equal to x and $B_{\geq}(n, p, x)$ to denote the probability that a binomial random variable with parameters n, p is greater than or equal to x. We will further let p_w, p_t, p_l denote the probabilities that Magnus wins, ties and loses a single game respectively. Using this notation, the probability that Magnus wins is:

$$B_{\geq}(14, p_w, 8) + \Sigma_{i=1}^{7} B(14, p_w, i) \cdot B_{\geq}(14, p_t, 15 - 2i)$$

That is, the probability that he either wins 8 or more games outright, or the probability that he wins exactly some number of games between 1 and 7 and ties at least enough games to total 7.5 point (remember a tie is worth 0.5 points and a win is worth 1 point). Using our computed values of $p_w = 0.1$ and $p_t = 0.54$ from the previous part, we can use a computer to calculate the above equation to get approximately 0.08, while the probability of Nepo winning the championship would be 0.77. If we condition on the fact that they don't tie, the probability that Magnus wins is $\dfrac{0.08}{0.08 + 0.77} \approx 0.09$

The program used to calculate the above probabilities:

```
from scipy.stats import binom

def B(n, p, x):
    return binom.pmf(x, n, p)

def B_g(n, p, x):
    return 1.0 - binom.cdf(x - 1, n, p)

def compute_win_prob(num_games=14, p_w=0.1, p_t=0.54):
    score = B_g(num_games, p_w, num_games // 2 + 1)
    for i in range(1, num_games // 2 + 1):
        score += B(num_games, p_w, i) * \
                B_g(num_games, p_t, num_games + 1 - 2*i)
    return score

magnus_win_prob = compute_win_prob()
nepo_win_prob   = compute_win_prob(p_w=0.36)
tie_prob        = 1.0 - magnus_win_prob - nepo_win_prob
print(magnus_win_prob, nepo_win_prob, tie_prob)
```

The calculation definitely makes the likelihood of a win by Nepo much more likely; essentially it says that Nepo will crush Magnus. In reality, this is unlikely as we are basing our model on **very** little data (11 games is a very small sample size). For example, our model does not take into account the FIDE rating which is the overall stats for a chess player against all opponents. A more balanced and possibly accurate model would take into account both FIDE rating and past history.

(d) [0 Points] Below are some examples of analyses you can use! You are of course free (and encouraged!) to find your own examples outside this list if you have a topic you are passionate about, but if you can't think of anything, you may use any of these as starting points. In many cases, there are already critiques of poor statistical/probability analyses online it's ok to look at these critiques, as long as you tell us if you're using any and still do the new probability calculation independently and put everything in your own words.

- Every year millions of people predict the outcomes of the NCAA men's basketball tournament. It is commonly said that the probability of a perfect bracket is $\frac{1}{2^{63}}$, (since there are 2^{63} ways the 63 games could play out) and therefore no one will ever predict a perfect bracket. Here is a source using that number
- Shortly after the 2020 presidential election, there were many assertions (including by people with PhD's...) that the probability of the election night shift was so low as to be impossible. Page 20 here is one example.

- A video-game streamer named “Dream” did a speedrun of Minecraft where they had incredible luck in a few parts of the game. So lucky, that a speedrunning organization declared that Dream had to be using a modified version of the game, and that the run was therefore invalid. The analysis that lead to the rejection of the run.
- One might be wondering how dangerous lightning really is and if you can die. How Dangerous is Lightning?
- Navigating through the board game of life to make a Monopoly stake and become a real estate tycoon may lead to a visit to prison occasionally. Probability of Going to Jail in Monopoly