CENG 222

Statistical Methods for Computer Engineering

Spring 2023-2024

Sample Solutions for Homework 1

Question 1

a) For the given function to qualify as a probability mass function, we need to ensure $\sum_x P(x) = 1$. Then,

$$\sum_{x=1}^{5} P(x) = N + N/2 + N/3 + N/4 + N/5 = \frac{137N}{60} = 1, \quad N = \boxed{\frac{60}{137} = 0.438}$$

In the following, I will use $\sum_x$ for convenience, the indices always run from 1 to 5.

b) To calculate the expected value,

$$\begin{aligned}
\mathbf{E}(X) &= \sum_x xP(x) \\
&= 1 \times \frac{60}{1 \times 137} + 2 \times \frac{60}{2 \times 137} + 3 \times \frac{60}{3 \times 137} + 4 \times \frac{60}{4 \times 137} + 5 \times \frac{60}{5 \times 137} \\
&= 5 \times \frac{60}{137} = \boxed{\frac{300}{137} = 2.190}
\end{aligned}$$

c) For variance,

$$\begin{aligned}
Var(X) &= \sum_x (x - \mathbf{E}(X))^2 P(x) \\
&= \left(1 - \frac{300}{137}\right)^2 \frac{60}{137} + \left(2 - \frac{300}{137}\right)^2 \frac{30}{137} + \left(3 - \frac{300}{137}\right)^2 \frac{20}{137} + \left(4 - \frac{300}{137}\right)^2 \frac{15}{137} + \left(5 - \frac{300}{137}\right)^2 \frac{12}{137} \\
&= \boxed{1.774}
\end{aligned}$$

d) We can calculate $\mathbf{E}(Y)$ as

$$\begin{aligned}
\mathbf{E}(Y) &= \sum_y yP(y) \\
&= 1 \times \frac{1}{15} + 2 \times \frac{2}{15} + 3 \times \frac{3}{15} + 4 \times \frac{4}{15} + 5 \times \frac{5}{15} \\
&= \boxed{\frac{55}{15} = 3.667}
\end{aligned}$$

$\mathbf{E}(XY)$ is given by

$$\mathbf{E}(XY) = \sum_x \sum_y (x - \mathbf{E}(X))(y - \mathbf{E}(Y))\, P_{X,Y}(x,y) = \sum_x \sum_y (x - \mathbf{E}(X))(y - \mathbf{E}(Y))\, P(x)P(y)$$

since there are 25 terms, we can write a simple code snippet to calculate this value as $\boxed{8.029}$.

We can calculate $E(X)\,E(Y)$ by plugging the numbers we have found as $E(X)\,E(Y) = \dfrac{300}{137} \times \dfrac{55}{15} = \boxed{8.029}$. Thus, the covariance is $\boxed{\mathrm{Cov}(X,Y) = 0}$. This indicates that there are no correlation between the two random variables. This is expected since the joint distribution is given as a separate product of the marginal distributions.

Question 2

a) Each attempt is given as a Bernoulli trial, and we are asked to find the probability of success for an individual attempt, p, so that a successful attempt occurs in 1000 trials with probability 95%. We can write this as

$$\begin{aligned} &\mathbf{P}\{\text{There is a successful attempt in a total of 1000 attempts}\} \\ =&\mathbf{P}\{\text{Not all 1000 attempts fail}\} = 1 - \mathbf{P}\{\text{All 1000 attempts fail}\} \\ =&1 - (1-p)^{1000} = 0.95 \end{aligned}$$

Denoting $q = 1 - p$, we are to solve $1 - q^{1000} = 0.95 \;\rightarrow\; q = (0.05)^{1/1000} = 0.997$. Then the probability of success should be $\boxed{p = 0.003}$.

b) We are given Bernoulli trials. This time, the probability of success is given along with the total number of trials and number of successes. Thus, we will use Negative Binomial distribution, in acquiring which, we will morph the problem into a Binomial distribution problem (*cf.* Example 3.21).

b-i) Write the problem as

$$\begin{aligned} \mathbf{P}\{X > 500\} &= \mathbf{P}\{\text{ more than 500 games are needed to be played in order to get 2 wins }\} \\ &= \mathbf{P}\{\text{ there are less than 2 wins in 500 games }\} \\ &= \mathbf{P}\{Y < 2\} \end{aligned}$$

where X is a Negative Binomial variable, and Y is the Binomial variable corresponding to the number of wins in 500 matches. Then, we need to calculate $\mathbf{P}\{Y < 2\} = \mathbf{P}\{Y \leq 1\} = F(1)$ for $n = 500$ and $p = 3 \times 10^{-3}$. Using MATLAB, we calculate this value as $\boxed{\mathrm{binocdf}(1, 500, 3e-3) = 0.558}$.

b-ii) Same reasoning as above, only the numbers are changed. $\boxed{\mathrm{binocdf}(1, 10000, 1e-4) = 0.736}$.

c) Let X be the number of days you do not feel sick in 366 days, which is a Binomial variable with $p = 0.98$. Applying the Binomial formula is rather cumbersome, thus, we will use Poisson approximation. Since Poisson approximation holds only for small p, we once again transform the problem:

$$\mathbf{P}\{X \geq 360\} = \mathbf{P}\{Y \leq 6\}$$

where Y is a Binomial random variable corresponding to the number of sick days, which admits the probability $q = 0.02$. Then, the Poisson variable is calculated as $\lambda = 366 \times 0.02 = 7.32$. In Table A3, the closest value is provided for $\lambda = 7.5$, which is $\boxed{0.378}$. If you instead used $\lambda = 7.0$, it is also fine, so long as you state it and provide the answer as $\boxed{0.450}$.

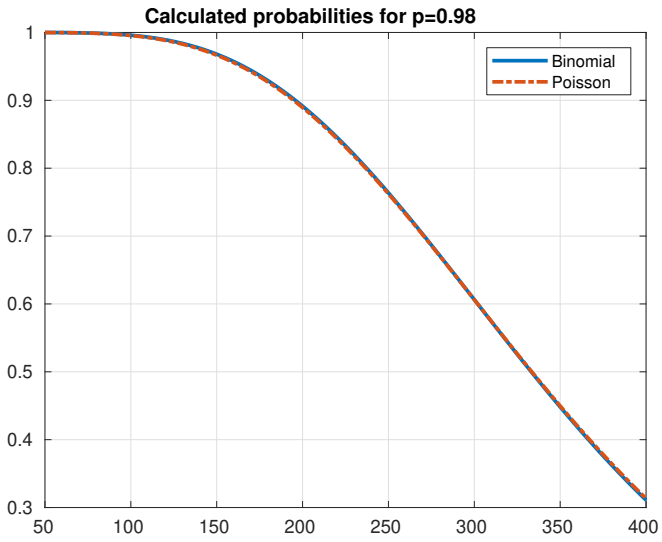
Question 3

a) Using MATLAB, $\boxed{1 - \text{binocdf}(359, 366, 0.98) = 0.401}$ or (less accurately) $\boxed{\text{poisscdf}(6, 7.32) = 0.403}$. If you used $\lambda = 7.5$ in your answer to **Q2c**, this number is higher, because $\lambda = 7.5$ implies a higher probability for sick days, hence underestimating the total number of healthy days. If you used $\lambda = 7.0$, this number is lower because $\lambda = 7.0$ implies a lower probability for sick days. Of course, other explanations are possible; they are all accepted provided they are logical.

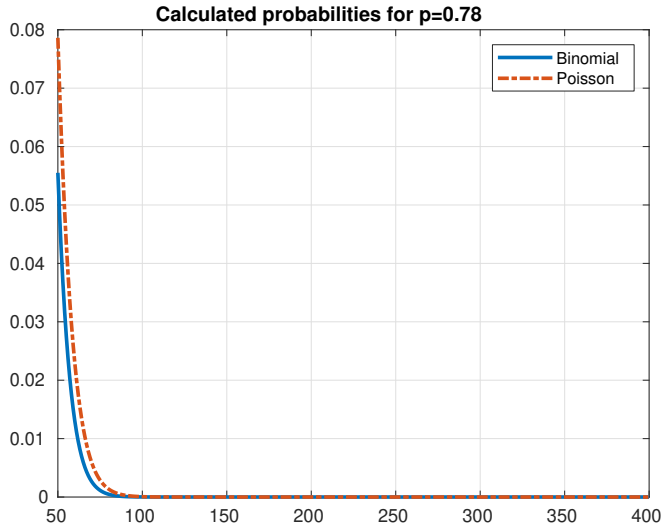
b) Sample code

```
p = 0.98;
ns = 50:400;

% calculate the probabilities using Binomial CDF
B = [];
for n=ns
    % (!) n-7, not n-6
    B(end+1) = 1 - binocdf(n-7, n, p);
end
% calculate the probabilities  using Poisson CDF
P = [];
for n=ns
    P(end+1) = poisscdf(6, n*(1-p));
end

plot(ns, B, 'linewidth', 2);
hold on;
plot(ns, P, ':-', 'linewidth', 2);

% adding some information
title(['Calculated probabilities for p=' sprintf('%.2f',p)]);
legend('Binomial', 'Poisson');
grid on;
```

c) Same code, with $p = 0.78$. You should observe that the relative discrepancy between the Binomial and Poisson-driven probabilities have increased when the probability went from 98% to 78%, which is especially evident for $n < 100$. See Figure 1.



(a)



(b)

Figure 1: Sample plots for (a) Q3b and (b) Q3c.