Question 1

Consider a set of \( n \) people who are members of an online social network. Suppose that each pair of people are linked as “friends” independently with probability \( \frac{1}{2} \). We can think of their relationships as a graph with \( n \) nodes (one for each person), and an undirected edge between each pair that are friends (this type of random graph is called an Erdős-Renyi graph or \( G_{\frac{1}{2}} \)). A clique is a fully connected subset of the graph, or equivalently a subset of people for which all pairs are friends.

(a) A clique of size 2 is simply a pair of nodes that are linked by an edge. Find the expected number of edges as a function of the number of nodes, \( n \). What is the expected number of friend relationships among \( n = 10 \) people?

(b) A clique of size 3 is a triplet of nodes within which all three pairs are linked by an edge. Find the expected number of 3-cliques as a function of the number of nodes, \( n \). What is the expected number of 3-cliques among \( n = 10 \) people?

(c) Larger cliques may occur involving groups of nodes of any size \( k \). Find the expected number of cliques of size \( k \geq 2 \) as a function of the number of nodes, \( n \). What is the expected number of cliques of size \( k \geq 2 \) with \( n = 10 \) people?

(d) Now suppose the probability of each edge is \( p \in (0, 1) \). As a function of the number of nodes \( n \) and \( p \), what is the expected number of cliques of size \( k \geq 2 \)?

Answer 1
Question 2

You have a box with \( n \) ropes, each rope having one red end and one blue end. You choose one red end and one blue end at random from the box, tie them together, and return those ropes to the box. (If you choose a red and a blue end of the same rope, then you will tie one rope into a loop with no free ends. If you choose a red end and a blue end of two different ropes, then you will tie two ropes into a single long rope). This gives a box with \( n - 1 \) untied red and blue ends, and you repeat this process until there are no more untied rope ends. Let \( L(n) \) be the expected number of loops formed.

(a) Consider the cases where \( n = 1 \), \( n = 2 \), or \( n = 3 \). By counting all possible outcomes, determine the expected number of loops formed in each case.

(b) We will now determine a different way to compute \( L(3) \). Suppose you draw the red end of some random rope \( R \). What is the probability that it will be tied to the blue end of \( R \)? What is the probability that it will be tied to the blue end of some different rope? Use these probabilities to find an linear function relating \( L(3) \) to \( L(2) \).

(c) Generalize the argument in part (b) to write \( L(n) \) as an linear function of \( L(n - 1) \), for any \( n \). Use this recursion to express \( L(n) \) directly as a sum of \( n \) terms, corresponding to the \( n \) rope-tying stages. What is \( L(10) \)?

Answer 2
**Question 3**

The time (rounded to the nearest minute) that a TA spends helping an individual student in office hours is geometrically distributed with a mean of 20 minutes, and independent of the time spent with other students. For a geometrically distributed $X$ with *rate parameter* $p$,

$$P(X = x) = (1 - p)^{x-1}p \quad \forall x \geq 1, \quad E[X] = \frac{1}{p}, \quad Var[x] = \frac{1-p}{p^2}.$$  

Suppose there is a homework due tomorrow, and there are 4 people ahead of you in line.

(a) The total time that it takes all 4 students ahead of you to receive help from the TA is a random variable. What is the mean and standard deviation of this total time?

(b) What is the probability that none of the 4 people ahead of you will take more than 30 minutes?

(c) When there is only 1 person left in front of you, you grow very hungry and have a craving for a gyro. It will take you exactly 8 minutes to go to the food truck and return, but the TA is strict and if you are not there when your name is called, your spot will be given to the next person in line. You decide that you will wait 15 more minutes, and if you haven’t been helped by that point, you’ll go get food and (8 minutes later) return. What is the probability that you miss your TA appointment?

(d) Continuing the scenario in part (c), suppose that you waited for 15 minutes, but have not yet been helped by the TA, so you depart for the food truck. Given your knowledge as you leave TA hours, what is the probability that you miss your TA appointment?

**Answer 3**
Question 4

We will now analyze some data collected by observing the famous “Old Faithful” geyser in Yellowstone National Park. We define random variable $S$ to be the time an eruption lasts, and random variable $T$ to be the “waiting time” until the next eruption. These are clearly continuous random variables, but we do not precisely know their true distribution. Instead we have a dataset with $n = 272$ independent observations $(s_i, t_i)$, $i = 1, \ldots, 272$, of the eruption time $s_i$ and subsequent waiting time $t_i$. Use the dataset we provided. See Figure 1 for a plot of this data.

In the following questions, we compute various quantities using the empirical distribution of the data. The empirical distribution of eruption time and waiting time can be represented by a probability mass function $\hat{p}_{ST}(s, t)$ which places probability $\frac{1}{n}$ on each of the $n$ data points, and probability 0 on the continuous range of other $(s, t)$ values. Under this distribution, the expected values of the empirical distribution of $S$ and $T$ then take the following simple form:

$$E[S] = \frac{1}{n} \sum_{i=1}^{n} s_i, \quad E[T] = \frac{1}{n} \sum_{i=1}^{n} t_i.$$  

(a) The variance of random variable $S$ equals $\text{Var}[S] = E[S^2] - E[S]^2$. Give formulas for computing $\text{Var}[S]$ and $\text{Var}[T]$ under the empirical distribution. Use Matlab’s `sum` function to write your own code that computes these variances, and report their values. **Hint:** Various definitions of the “sample variance” can be found in statistics references, and they are not all equivalent to the variance of the empirical distribution.


(c) The empirical cumulative distribution function of $S$ is defined as $\hat{F}_S(s) = P(S \leq s)$, where this probability is under the empirical distribution. Find eruption times $\bar{s}_1, \bar{s}_2, \bar{s}_3$ such that

Figure 1: A scatter plot of the observations of Old Faithful’s eruption time (horizontal axis) and waiting time to the next eruption (vertical axis). Each point is one of the $n = 272$ observations.
\[ F_S(\bar{s}_1) = 0.25, F_S(\bar{s}_2) = 0.50, F_S(\bar{s}_3) = 0.75. \]

Using the cumulative distribution of \( T \), also find waiting times \( \bar{t}_1, \bar{t}_2, \bar{t}_3 \) such that \( F_T(\bar{t}_1) = 0.25, F_T(\bar{t}_2) = 0.50, F_T(\bar{t}_3) = 0.75. \) \textit{Hint:} Use Matlab’s \texttt{sort} function.

Consider two new random variables. Let \( X \) indicate whether the eruption time \( S \) is “short” or “long”: \( X = 0 \) if \( S \leq 3.25 \), and \( X = 1 \) if \( S > 3.25 \). Let \( Y \) indicate whether the waiting time \( T \) is “short” or “long”: \( Y = 0 \) if \( T \leq 68 \), and \( Y = 1 \) if \( T > 68 \).

(d) Using the empirical distribution of \( S \) and \( T \), determine the joint probability mass function \( p_{XY}(X, Y) \). Also determine the marginal probability mass functions \( p_X(x) \) and \( p_Y(y) \).

(e) Are the random variables \( X \) and \( Y \) either exactly or approximately independent? Clearly justify your answer using the probability mass functions from part (d).

\textbf{Answer 4}