1. Sampling without Replacement

Suppose you have $N$ items, $G$ of which are good and $B$ of which are bad ($B$, $G$, and $N$ are positive integers, $B + G = N$). You start to draw items without replacement, and suppose that the first good item appears on draw $X$. Compute the mean and variance of $X$.

Solution:
The expectation is computed with a clever trick: let $X_i$ be the indicator that the $i$th bad item appears before the first good item, for $i = 1, \ldots, B$. Then, $X = 1 + \sum_{i=1}^{B} X_i$, and by linearity of expectation,

$$
E[X] = 1 + B E[X_1] = 1 + \frac{B}{G + 1} = \frac{N + 1}{G + 1}.
$$

Observe that $\text{var} \ X = \text{var}(X - 1)$. Using the same indicators, we compute $E[(X - 1)^2]$.

$$
E[(X - 1)^2] = B E[X_1^2] + B(B - 1) E[X_1 X_2]
$$
$$
= \frac{B}{G + 1} + \frac{2B(B - 1)}{(G + 1)(G + 2)}
$$

Therefore, our answer is

$$
\text{var} \ X = \frac{B}{G + 1} + \frac{2B(B - 1)}{(G + 1)(G + 2)} - \left( \frac{B}{G + 1} \right)^2.
$$

With a little algebra, we can actually simplify the result.

$$
\text{var} \ X = \frac{B(G + 1)(G + 2) + 2B(B - 1)(G + 1) - B^2(G + 2)}{(G + 1)^2(G + 2)}
$$
$$
= \frac{BG(N + 1)}{(G + 1)^2(G + 2)}
$$
2. Poisson Merging

Let $X$ and $Y$ be independent Poisson random variables with means $\lambda$ and $\mu$ respectively. Prove that $X + Y \sim \text{Poisson}(\lambda + \mu)$. (This is known as Poisson merging.) Note that it is not sufficient to use linearity of expectation to say that $X + Y$ has mean $\lambda + \mu$. You are asked to prove that the distribution of $X + Y$ is Poisson.

*Note:* You may need to use the Binomial theorem: $(x + y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i}$. This poisson merging property will be extensively used when we discuss Poisson processes.

**Solution:**

**Intuition:**

The Poisson distribution is used to model rare events, such as the number of customers who enter a store in the next hour. The theoretical justification for this modeling assumption is that the limit of the binomial distribution, as the number of trials $n$ goes to $\infty$ and the probability of success per trial $p$ goes to 0, such that $np \to \lambda > 0$, is the Poisson distribution with mean $\lambda$.

Now, suppose we have two independent streams of rare events (for instance, the number of female customers and male customers entering a store), and we do not care to distinguish between the two types of rare events. We are asked if the combined stream of events can be modeled as a Poisson distribution.

**Mathematical solution:**

For $z \in \mathbb{N}$,

$$
P(X + Y = z) = \sum_{j=0}^{z} \mathbb{P}(X = j, Y = z-j) = \sum_{j=0}^{z} \frac{e^{-\lambda \lambda^j} e^{-\mu \mu^{z-j}}}{j! (z-j)!}$$

$$= \frac{e^{-(\lambda+\mu)}}{z!} \sum_{j=0}^{z} \frac{z!}{j!(z-j)!} \lambda^j \mu^{z-j}$$

$$= \frac{e^{-(\lambda+\mu)}}{z!} \sum_{j=0}^{z} \binom{z}{j} \lambda^j \mu^{z-j} = \frac{e^{-(\lambda+\mu)}(\lambda + \mu)^z}{z!}.$$  

Here is some intuition for why Poisson merging holds. If we are interested in the number of customers entering a store in the next hour, we can discretize the hour into $n$ time intervals, where $n$ is a positive integer. In each time interval, independently of other time intervals, the probability that a female customer enters the store is $\lambda/n$ and the probability that a male customer enters the store is $\mu/n$. Since the two types of customers are assumed to be independent, the probability that a customer, disregarding gender, enters the store is $\lambda/n + \mu/n = \lambda + \mu/n$. As $n \to \infty$, the number of customers who enter the store in the hour is Poisson with mean $\lim_{n \to \infty} n[\lambda/n + \mu/n - \lambda \mu/n^2] = \lambda + \mu$.

We will be able to give a much easier proof of this result after we introduce transforms of random variables.
3. Clustering Coefficient

This problem will explore an important probabilistic concept of clustering that is widely used in machine learning applications today. Consider \( n \) students, where \( n \) is a positive integer. For each pair of students \( i, j \in \{1, \ldots, n\}, i \neq j \), they are friends with probability \( p \), independently of other pairs. We assume that friendship is mutual. We can see that the friendship among the \( n \) students can be represented by an undirected graph \( G \). Let \( N(i) \) be the number of friends of student \( i \) and \( T(i) \) be the number of triangles attached to student \( i \).

We define the clustering coefficient \( C(i) \) for student \( i \) as follows:

\[
C(i) = \frac{T(i)}{\binom{N(i)}{2}}.
\]

Figure 1: Friendship and clustering coefficient.

The clustering coefficient is not defined for the students who have no friends. An example is shown in Figure 1. Student 3 has 4 friends (1, 2, 4, 5) and there are two triangles attached to student 3, i.e., triangle 1-2-3 and triangle 2-3-4. Therefore \( C(3) = \frac{2}{\binom{4}{2}} = \frac{1}{3} \).

Find \( \mathbb{E}[C(i) \mid N(i) \geq 2] \).

**Solution:**

First, we compute \( \mathbb{E}[C(i) \mid N(i) = k] \), for \( k \in \{2, \ldots, n - 1\} \). Suppose that student \( i \) has friends \( f_1, \ldots, f_k \). We can see that \( T(i) \) equals the number of friend pairs among \( \{f_1, \ldots, f_k\} \). Since there are \( \binom{k}{2} \) possible pairs and each pair of students are friends with probability \( p \), independently of other pairs, we know that the expected number of friend pairs among \( \{f_1, \ldots, f_k\} \) is \( \binom{k}{2} p \). Then we have

\[
\mathbb{E}[C(i) \mid N(i) = k] = \frac{\binom{k}{2} p}{\binom{k}{2}} = p.
\]

Since this is true for all \( k \geq 2 \), we have \( \mathbb{E}[C(i) \mid N(i) \geq 2] = p \).