Stat 111 Week 1: Review of Discrete Random Variables and Distributions

1. **Bernoulli Variables**

   The Bernoulli distribution is the simplest probability distribution, and in some ways, a building block for all other discrete distributions.

   a) Define $I_{(A)}$ to be an indicator variable for the event A, meaning $I_{(A)} = 1$ if A occurs and $I_{(A)} = 0$ is A^c occurs. Relate this to the Bernoulli random variable. Explain how an indicator variable represents the *fundamental bridge* between probability and expected value (see Blitzstein 4.4).

   b) Use indicator variables to prove Boole's inequality: $P(A_1 \cup A_2 \cup \ldots \cup A_n) \leq P(A_1) + P(A_2) + \ldots + P(A_n)$. Consider the special case where the events are all independent with the same probability.

   c) Suppose n graduates all throw their caps in the air and then retrieve a cap at random. Find an expression for the probability that none of the students retrieve their own cap (a *derangement*). Find the limit of this probability if the number of caps $n \to \infty$. Hint: For $i = 1, \ldots, n$, let A_i represent the event that person i retrieves their own cap. Then $(A_1 \cup A_2 \cup \ldots \cup A_n)^c$ is the event that nobody ends up with their own cap. See the *Useful Facts* at the end of this document.

   d) A Poisson(λ) variable may be represented as the limit of the sum of n iid Bernoulli(p) variables as $n \to \infty$ and $np \to \lambda$. Explain why, for large n, the count X of graduates who retrieve their own cap is approximately Poisson(1). Compute the exact and approximate probabilities $P(X = 0)$ and $P(X = 1)$ for $n = 6$.

2. **Binomial and Hypergeometric**

   A count X of successes in n trials may be expressed as the sum of Bernoulli variables. With iid Bernoulli trials, the sum follows a Binomial distribution. The Binomial distribution may arise as the limit of a Hypergeometric distribution. The Hypergeometric distribution may arise as a conditional distribution for Binomial counts.

   a) For n independent trials, each with success probability p, the distribution of X is Binomial(n, p). Write out the probability mass function for X. Explain why the sum of two independent Binomial variables is also Binomial, if and only if their probabilities are equal.

   b) If sampling is done without replacement from a population with r successes and $N - r$ failures, then X is a hypergeometric variable. Write out the probability mass function for X. Be careful to designate the appropriate support for X (i.e., what values have non-zero probability?). Could the sum of two Hypergeometric variables also be Hypergeometric?

   c) For a hypergeometric random variable, show that, as $N \to \infty$ with $r/N \to p$, the distribution of X converges to Binomial(n, p). See the useful facts at the end of this document.

   d) I have carried out experimental surveys to determine the effect of wording of a question of the response. Suppose I give out n_1 surveys with wording 1 and n_2 surveys with wording 2. Suppose students answer independently and will agree to either wording with probability p (i.e., the wording does not matter). Let r be the total number of students who agree (to either wording), the distribution of X, the number of students who agree to wording 1 (e.g.) is a Hypergeometric variable. This is the basis for the Fisher Exact Test (Blitzstein 3.9.1).

   e) For the situation in part d, show that the marginal distribution of X is Binomial(n_1, p).

3. **Binomial and Poisson**

For a Poisson process in time, events (e.g., text messages received) occur at a constant rate of λ events per unit time (on average), and the counts of events in non-overlapping time intervals of lengths t_1 and t_2 are independent Poisson random variables with rates λt_1 and λt_2. The probability mass function (pmf) for $X \sim \text{Poisson}(\lambda)$, the count of events in a a unit time interval ($t = 1$), is $P(X = x) = \lambda^x e^{-\lambda}/x!$, for $x = 0, 1, \ldots,$ The Poisson pmf arises as the limit of the Binomial pmf.

a) Find the probability of at least one event occurring in a time interval of length t. As $t \to 0$, show that this probability divided by t converges to λ, meaning the probability behaves like λt for t close to 0. Also consider the probability of exactly 1 event occurring in an interval of length t.

b) Imagine partitioning a unit time interval into n non-overlapping subintervals, each of length $t = 1/n$. Let X be the count of intervals that contain at least one event. Show, in the limit as $n \to \infty$ that X represents the total count of events, and has the Poisson(λ) pmf.

c) Let $N \sim \text{Poisson}(\lambda)$ be the number of scratch-off lottery tickets sold in a day in a particular store. Each ticket has probability p of being a winner, independent of any other ticket outcomes. Let X_1 be the number of winning tickets, and $X_2 = N - X_1$ the number of losing tickets sold in a day. Show that X_1 and X_2 are independent Poisson variables by finding

$$P(X_1 = x_1, X_2 = x_2) = P(N = x_1 + x_2)P(X_1 = x_1 | N = x_1 + x_2)$$

d) If $X_1 \sim \text{Poisson}(\lambda_1)$ is independent of $X_2 \sim \text{Poisson}(\lambda_2)$, show that

$$N = X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$$

4. **Poisson and Negative Binomial**

A Negative Binomial variable arises as the count of failures before a specified number of successes, and as a Poisson variable with a rate parameter generated according to a Gamma distribution.

a) Suppose $X_1 \sim \text{Poisson}(\lambda_1)$ is independent of $X_2 \sim \text{Poisson}(\lambda_2)$, and let $N = X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$. Show $X_1 | N = n \sim \text{Binom}(n, p)$, for $p = \frac{\lambda_1}{\lambda_1 + \lambda_2}$.

b) For a sequence of iid Bernoulli(p) variables, let Y be the number of failures before the rth success. Write out the pmf for Y.

c) Suppose two soccer matches are played simultaneously and that the goals scored in match 1 and in match 2 represent independent Poisson processes with rates λ and 1, respectively. Let Y be the number of goals scored in match 2 at the time of the rth goal in match 1. Explain how Y follows the same distribution as Y in part b (what are the Bernoulli variables and what is p?).

d) The time θ of the rth event in a Poisson process with rate λ is a continuous random variable that follows a Gamma(r, λ) distribution. For the situation in part c, what is the distribution of Y conditional on θ? What is the marginal distribution of Y?

Useful Facts:

$$(1)\qquad P(A_1 \cup \ldots \cup A_n) \;=\; \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \ldots$$

$$(2)\qquad (a+b)^n \;=\; \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

$$(3)\qquad \lim_{n\to\infty} \frac{n!/(n-k)!}{n^k} \;=\; \lim_{n\to\infty} \frac{n(n-1)\ldots(n-k+1)}{n^k} \;=\; 1, \qquad k = 0, 1, \ldots$$

$$(4)\qquad e^x \;=\; \sum_{k=0}^{\infty} \frac{x^k}{k!} \;=\; \lim_{n\to\infty} (1 + x/n)^n$$