Name: _____ PUID _____

STAT 350 Worksheet #6

Previously, we examined discrete random variables by explicitly listing their probability mass functions (PMFs) in tabular form. While this approach works well for small, well-defined cases, it can become cumbersome when dealing with random variables that take on many possible values.
In many situations, patterns emerge in how probabilities are assigned to outcomes. Instead of defining a PMF from scratch, we can use named discrete random variables, which follow standard probability distributions that have been studied extensively. These named distributions provide a structured way to describe random variables with known properties, simplifying calculations and allowing us to apply established theorems and results.

Some advantages of using named distributions include:

- **Compact Representation:** Instead of listing probabilities for each possible outcome, we can describe a random variable using a few key parameters.
- **Generalization:** Many processes share similar probability structures, so named distributions allow us to apply the same mathematical tools across different contexts.
- **Efficient Computation:** Expectation, variance, and probabilities can be computed using well-established formulas rather than recalculating them from first principles.

In this course, we will focus on three fundamental **named discrete distributions**, the **Bernoulli, Binomial,** and **Poisson Distributions.**

However, in this worksheet, you will also get a **preview of additional named discrete distributions**, providing insight into how different random variables arise in various contexts. These will be explored in greater depth in later coursework.

A **Bernoulli random variable** $\boldsymbol{X}$ is the simplest discrete random variable, as it represents the occurrence or non-occurrence of an event in a **single trial**. It takes on only two possible values, typically **0** and **1**, meaning its support is: $\mathbf{Supp(X)} = \{\mathbf{0}, \mathbf{1}\}$.
This setup is useful for modeling scenarios where an event either **happens (1)** or **does not happen (0)**. The probability mass function (PMF) of a Bernoulli random variable with success probability $\boldsymbol{p}$ is:

$$\boldsymbol{P(X = x)} = \begin{cases} \boldsymbol{p}, & \textbf{if } \mathbf{x = 1} \text{ (success)} \\ \mathbf{1 - p}, & \textbf{if } \boldsymbol{x} \textbf{ = 0} \text{ (failure)} \end{cases}$$

and we write:

$$\boldsymbol{X} \sim \mathbf{Bern(p)},$$

where the tilde symbol **(∼)** is used to denote that $\boldsymbol{X}$ "follows a" **Bernoulli distribution** with probability $\boldsymbol{p}$.
You can think of a Bernoulli random variable as an indicator function, where it simply "indicates" whether a specific event occurs. In fact, if $\boldsymbol{A}$ is an event, the random variable:

$$\mathbf{1}_{\boldsymbol{A}} = \begin{cases} \mathbf{1}, & \textbf{if } \boldsymbol{A} \textbf{ occurs} \\ \mathbf{0}, & \textbf{if } \boldsymbol{A'} \textbf{ occurs} \end{cases}$$

follows a Bernoulli distribution with success probability $\boldsymbol{p} = \boldsymbol{P(A)}$.
This makes Bernoulli random variables fundamental in probability, as they serve as building blocks for more complex models, such as the **Binomial distribution**, which describes the total number of successes in multiple independent Bernoulli trials.

A **Binomial random variable** counts the number of successes in $\boldsymbol{n}$ **independent** and **identical Bernoulli trials**, each with success probability $\boldsymbol{p}$. Here $\boldsymbol{n}$ and $\boldsymbol{p}$ are what we call **parameters** of the distribution and once we know the values of the parameters we know everything about the distribution. A Binomial random variable is simply the sum of $\boldsymbol{n}$ **independent Bernoulli random variables**:

$$\boldsymbol{X = X_1 + X_2 + \cdots + X_n}$$

where each $\boldsymbol{X_i} \overset{\text{i.i.d.}}{\sim} \mathbf{Bern}(\boldsymbol{p})$ for $\boldsymbol{i} \in \{\mathbf{1}, \mathbf{2}, \ldots, \boldsymbol{n}\}$**.** Here **i. i. d.** (**independent and identically distributed**) means that:

- Each $\boldsymbol{X_i}$ follows the same **Bernoulli distribution** with success probability $\boldsymbol{p}$.
- The outcomes of different trials do not influence each other.

This property makes the **Binomial distribution** a natural extension of the **Bernoulli,** allowing it to model repeated independent trials of the same process. Companies and researchers leverage the **Binomial distribution** alongside **statistical inference** in **quality control** to test defect rates, in **medical research** to estimate treatment success probabilities, and in **politics** to analyze voter behavior and predict election outcomes, among other applications.
A **Binomial random variable** $\boldsymbol{X}$ has support $\mathbf{Supp(X)} = \{\mathbf{0}, \mathbf{1}, \mathbf{2}, \ldots, \mathbf{n}\}$ and we write $\boldsymbol{X} \sim \mathbf{Binomial(n, p)}$**.**
It has **probability mass function**:

$$\boldsymbol{P(X = x)} = \binom{\boldsymbol{n}}{\boldsymbol{x}} \boldsymbol{p^x (1 - p)^{n-x}}$$

and the **expected value** and **variance** are simply functions of the **parameters**:

- $\boldsymbol{E[X] = np}$
- $\boldsymbol{Var(X) = np(1 - p)}$

Northgate faces severe storms each year, leading to the possibility of widespread power outages. City planners want to estimate how likely it is for a storm to knock out power in many neighborhoods simultaneously, and if it does, how effectively emergency crews can restore service. Use the **Binomial distribution** and the **BINS** criteria to guide your analysis.

Important Note on Independence
In practice, power outages across neighborhoods are often correlated (for example, if a major substation goes down, it can affect multiple neighborhoods at once). For this exercise, however, we **assume** that each neighborhood's power status is **independent** of the others.

1. During a storm, each of the city's 50 neighborhoods independently loses power with probability $\boldsymbol{p = 0.3}$. Let X denote the total number of neighborhoods that lose power during a single storm.

a) Explain why $\boldsymbol{X}$ follows a Binomial distribution and identify its parameters.

b) Compute the probability by hand that there are **exactly 20 neighborhoods** that lose power in a storm under these assumptions. Repeat the calculation using R to compute the probability with the function **dbinom (help(dbinom))** to check your work.

c) Express symbolically the probability that there are **atleast 20 neighborhoods** that lose power and using the Binomial PMF formula, express how you would compute this probability.

d) Directly calculating this probability in **d)** would be tedious. We will learn a technique later to leverage the continuous normal distribution to approximate this probability. However, we can still calculate this using **R**. **Method 1:** Create a list of possible successes from 20 to 50 and apply **dbinom** to the list and sum over the list. **Method 2:** utilize the **pbinom** function to compute the probability. Use both approaches to compute the probability.

e) Given that at least 20 neighborhoods lost power during a particular storm, what is the probability that exactly 25 neighborhoods lost power? Assume the same Binomial model applies to this scenario.

f) For a random storm determine the expected number of neighborhoods that will lose power and determine the standard deviation.

g) For a randomly selected storm, determine the probability that the number of neighborhoods that lose power exceeds two standard deviations above its expected value. Express your answer symbolically in terms of the Binomial PMF and compute the probability using R.

Many real-world phenomena involve **random events occurring over time or space** at an **average rate**. In such cases, the **Poisson distribution** provides a powerful mathematical model for counting the number of events in a **fixed interval of time, space, or volume**. A scenario is **Poisson-distributed** if:

- **Unique Events (no clustering):** Events occur one at a time. The probability of two or more events happening simultaneously in a very small interval is negligible.
- **Independence:** The occurrence of one event does not affect the probability of another occurring. Events happen randomly and are not influenced by previous occurrences. Additionally, the number of events that occur in any interval are **independent** of those that occur in any other non-overlapping interval.
- **Stationarity (Constant Rate):** Events occur at a steady average rate over time or space. The expected number of events in an interval does not change unless external conditions shift the distribution.
- **Proportionality:** The expected number of events is proportional to the size of the interval. Doubling the length of time or space results in twice as many expected occurrences.

Modern MRI scanners are highly sensitive but sometimes suffer from artifacts which are unwanted noise distortions caused by patient movement, hardware issues, or interference. AI powered medical imaging tools detect these artifacts to improve diagnostic accuracy.

In medical imaging, artifacts appear **randomly** across the spatial area of a scan. These artifacts do not overlap, occur independently, and appear at an average rate for a given image size. Because the number of artifacts is counted over a **fixed spatial region** rather than over time, the Poisson distribution provides a natural way to model this process.
To ensure the Poisson model remains valid, we assume that **each MRI scan covers the same fixed spatial area**. If scan sizes varied, the number of artifacts would depend on scan area, and the Poisson rate λ would need to be scaled accordingly. In this case, however, we assume that all scans are uniform in size, meaning that the average number of artifacts per scan remains constant.

2. A hospital's radiology department is analyzing the number of MRI artifacts detected by an AI system per scan. Based on past data, the AI detects an average o**f $\lambda = 3$** artifacts per scan, meaning that each complete image contains an average of 3 randomly occurring noise distortions. The department now wants to analyze data for **10 consecutive MRI scans** to determine the likelihood of different artifact patterns across multiple scans.

a) Let Y denote the total number of artifacts detected by the AI in **10 scans**. Provide a justification why the number of MRI artifacts detected across **10 scans** should follow a Poisson distribution and identify the parameter λ_Y representing the expected number of artifacts across the 10 scans**. Bonus exercise:** prove this mathematically using induction.

b) Compute the probability that there are exactly 20 artifacts detected across the 10 scans. Repeat the calculation using R to compute the probability with the function **dpois (help(dpois))** to check your work.

c) Computer the probability that there are at most 20 artifacts detected across the 10 scans. This calculation is tedious to determine by hand. Use R to calculate this as before: **Method 1:** Create a list of possible successes from 0 to 20 and apply **dpois** to the list and sum over the list. **Method 2:** utilize the **ppois** function to compute the probability.

d) The hospital's radiology department is analyzing two separate collections of MRI scans. **Box A** contains **10 scans**, while **Box B** contains **15 scans**. Compute the probability that **Box A contains at most 20 artifacts** and **Box B contains at most 35 artifacts**.

While we have focused on three **named discrete random variables**, it is important to recognize that there are many others, each designed to model different types of random processes. Not all discrete distributions count the number of successes in a fixed number of trials, nor do they all assume independent events.
For example:

- **Geometric Distribution:** models the **number of trials** until the **first success** occurs. This can model repeated failures before success, such as the number of free throw attempts needed before the first basket.
- **Negative Binomial Distribution:** extends the geometric case by counting the number of trials until the r^{th} **success** occurs.
- **Hypergeometric Distribution**: counts the number of success in a fixed number of trials but differs from the Binomial distribution because it models **dependent trials**. The Hypergeometric distribution applies to cases where **sampling occurs without replacement**, such as drawing defective items from a small batch or selecting colored gummy bears from a finite jar.

A random variable X following a **Hypergeometric distribution** has the following parameters N the total number of items considered, n the number of items sampled (or trials performed), and M the total number of success possible and it has the following probability mass function:

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

with $\binom{N}{n}$ denoting the number of ways to select n items out of N total items, $\binom{M}{x}$ denoting the number of ways to select x successes out of M total successes, and $\binom{N-M}{n-x}$ denoting the number of ways to choose $n-x$ failures out of $N-M$ total failures.

3. Lets consider our gummy bear example:

- $\mathbf{Jar_1}$ contains **30 red**, **10 green**, and **10 blue** gummies.
- $\mathbf{Jar_2}$ contains **20 red** and **40 green** gummies.
- $\mathbf{Jar_3}$ contains **35 yellow** gummies

Using the hypergeometric distribution determine the probability of getting 2 green gummies given you sample from $\mathbf{Jar_1}$ and confirm with how we calculated this in a previous worksheet.