Hypergeometric & Poisson

1 Hypergeometric

1.1 Description & Definition

- Earlier we discussed the binomial distribution in terms of flipping a coin repeatedly, or flipping multiple coins at the same time. Either way it was done such that the flips were independent
- The binomial distirbution can also arise when sampling from a large population where p $(0 < p < 1)$ proportion of the population as one quality, and therefore $1 - p$ of the population doesn't have that quality.
- If the sample is done with replacement (i.e. a sample of n things is done one at a time, with replacement), then the resulting distirbution is binomial
- Additionally, If the population is much larger than the sample size and the proportion is sufficiently large, then a sample will also have a binomial distribution approximately
- In some cases however, these conditions are not met
- That is to say that the sample size is too close to the population size and/or the proportion p of the population that has the quality of interest is not large enough
- In these cases that sample instead takes on a *Hypergeometric* distribution
- Formal Description: Suppose we have a population with N memebers r of which $(r < N)$ have a certain property or quality fo interest. If we take a sample of size n where $n \leq r$ and $n \leq N - r$, and let X be the number of objects in our sample that have the quality of interst, then X will have a Hypergeometric distribution.
- *Note:* the support of X will be $\{0, 1, 2, ..., n\}$
- How can we figure out the PDF of this distribution?
- Consider $p_X(x) = P(X = x)$. The event where $X = x$ will be the event in which we have selected x of the r available objects in the population with the desired property and we will have selected $n - x$ objects that do not have that property
- we can determine the probability of this event the same way we would for any basic event:

$$
\begin{aligned}
P(X = x) &= \frac{\#\text{of ways to select for event} \times \#\text{ of ways to order}}{\#\text{of ways to select in general} \times \#\text{ of ways to order}} \\
&= \frac{\left(\begin{array}{l} \#\text{of ways to select } x \text{ objects with quality} \\ \times \#\text{ of ways to select } n - x \text{ objects without quality} \end{array} \right) \times \#\text{ of ways to order}}{\#\text{of ways to select in general} \times \#\text{ of ways to order}}
\end{aligned}
$$

$$= \frac{(\binom{r}{x} \times \binom{N-r}{n-x}) \times n!}{\binom{N}{n} \times n!}$$
$$= \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$$

- So our PDF will be $p_X = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$
- In general, we say that if a R.V. has the PDF $p_X = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$ for $x = 0, 1, ..., n$ where N, r, n are all positive integers such that $n \leq r < N$ and $n \leq N - r$, then X has a *Hypergeometric* distribution ($X \sim Hypergeo(N, r, n)$)

1.2 Verification, Mean, & Variance

- We will esatablish the following for this PDF
  1. Verify that the PDF is valid (i.e., show that it follows the two rules for discrete R.V. PDFs)
  2. Establish the mean of distribution
  3. Establish the variance of the distribution
- Before we proceed, we will first introduce and prove the following theorem:

  Thoerem 1. *Let n, N_1, and N_2 be positive integers such that $n \leq N_1$ and $n \leq N_2$. Then*
  $$\sum_{k=0}^{n} \binom{N_1}{n}\binom{N_2}{n-k} = \binom{N_1 + N_2}{n}$$

  Proof. Let a be an aribtray real value. Then we see that, by using the binomial expansion it can be shown that
  $$\begin{aligned}
  (1+a)^{N_1+N_2} &= (1+a)^{N_1}(1+a)^{N_2} \\
  \Rightarrow \sum_{i=0}^{N_1+N_2} \binom{N_1+N_2}{i} a^i &= (\sum_{i=k}^{N_1} \binom{N_1}{k} a^k)(\sum_{j=0}^{N_2} \binom{N_2}{j} a^j) \\
  \Rightarrow \sum_{i=0}^{N_1+N_2} \binom{N_1+N_2}{i} a^i &= \sum_{i=k}^{N_1}\sum_{j=0}^{N_2} \binom{N_1}{k}\binom{N_2}{j} a^{k+j}
  \end{aligned}$$

  Because a was selected arbitrarily we see that the coefficients of each powered term must be equal. This implies that
  $$\binom{N_1+N_2}{n} a^n = \left(\sum_{j+k=n; 0 \leq j,k \leq n} \binom{N_1}{k}\binom{N_2}{j} \right) a^n$$

$$
\begin{aligned}
\Rightarrow \binom{N_1+N_2}{n} &= \sum_{j+k=n;0\le j,k\le n} \binom{N_1}{k}\binom{N_2}{j} \\
\Rightarrow \binom{N_1+N_2}{n} &= \sum_{k=0}^{n} \binom{N_1}{k}\binom{N_2}{n-k}
\end{aligned}
$$

$\square$

1. Verify

   a) $\sum_{x\in S} p_X(x) = 1$

   Proof.

$$
\begin{aligned}
\sum_{x\in S} p_X(x) &= \sum_{x=0}^{n} \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} \\
&= \frac{1}{\binom{N}{n}} \sum_{x=0}^{n} \binom{r}{x}\binom{N-r}{n-x} \\
&= \frac{1}{\binom{N}{n}} \binom{r+N-r}{n} \leftarrow \text{From Theorem 1} \\
&= \frac{1}{\binom{N}{n}} \binom{N}{n} \\
&= 1
\end{aligned}
$$

$\square$

   b) $0 \le p_X(x) \le 1 \forall x \in S$

   Proof. We see that $\binom{r}{x}$, $\binom{N-r}{n-x}$, and $\binom{N}{n}$ are all non-negative, thus $0 \le p_X(x) \forall x \in S$, the support of X. Since $0 \le p_X(x) \forall x \in S$ and $\sum_{x\in S} p_X(x) = 1$, we can conclude that $p_X(x) \le 1 \forall x \in S$. Therefore, $0 \le p_X(x) \le 1 \forall x \in S$. $\square$

2. Mean

$$
E[X] = \frac{nr}{N}
$$

Proof.

$$
\begin{aligned}
E[X] &= \sum_{x\in S} x p_X(x) \\
&= \sum_{x=0}^{n} x \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{x=1}^{n} x \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} \leftarrow \text{First term is just 0} \\
&= \sum_{x=1}^{n} x \frac{r!}{x!(r-x)!} \frac{\binom{N-r}{n-x}}{\binom{N}{n}} \\
&= r \sum_{x=1}^{n} \frac{(r-1)!}{(x-1)!(r-x)!} \frac{\binom{N-r}{n-x}}{\binom{N}{n}} \\
&= r \sum_{x-1=0}^{n-1} \frac{\binom{r-1}{x-1}\binom{N-r}{(n-1)-(x-1)}}{\binom{N}{n}} \\
&= r \sum_{y=0}^{n-1} \frac{\binom{r-1}{y}\binom{N-r}{(n-1)-y}}{\binom{N}{n}} \leftarrow \text{Relabel } x-1 \text{ as } y \\
&= r \frac{\binom{(N-r)+(r-1)}{(n-1)}}{\binom{N}{n}} \leftarrow \text{From Theorem 1} \\
&= r \frac{\binom{N-1}{n-1}}{\binom{N}{n}} \\
&= r\left(\frac{(N-1)!}{(n-1)!(N-n)!}\right)/\left(\frac{N!}{n!(N-n)!}\right) \\
&= r \frac{(N-1)!n!(N-n)!}{N!(n-1)!(N-n)!} \\
&= \frac{nr}{N}
\end{aligned}
$$

□

3. Variance

$$V[X] = n\left(\frac{r}{N}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right)$$

Proof. Proof left as Extra Credit

□

2 Poisson

2.1 Description & Definition

- Another distriubtion that is derived from binomial-esque situations is the Poison Distribution
- Suppose that you are hired to investigate the safety at a manufacturing plant.

- Intially, you plan to observe how many accidents happen in the course of a month. We weill let this number be the Random Variable X.
- If we consider the distribution we already have, we see that none of them accurately describe the situation.
  - Geometric/Negative Binomial: While we are counting potentially out until infinity, there is not a point at which we stop counting (in terms of numbers of observations)
  - Hypergeometric: We aren't taking a sample from a population of a known size, so this distirbution doesn't fit
  - Bernoulli/Binomial: We aren't taking a sample of a certain size, so at first it doesn't look like the bernoulli/Binomial family will work either
- What if we take the month that we are observing in and divide it into n sub intervals, and we assume that the probability of an accident occuring during each one of these sub intervals is p and accidents (or lack there of) are independent between intervals. Then the number of accidents WOULD have a binomial distribution.
- There is a problem with this formulation however: what if more than one accident happens in a subinterval?
- To address this issue, we could make the subintervals small enough that only one accdient could possibly occur during an given interval
- We can do this by making n, the number of intervals, very large.
- To guarantee that we can only have a maximum of one accident occur during a given subinterval, we will then consider what happens as we take the limit of the distirbution as n goes to ∞
- So, This would mean that our PDF of X will be $\lim_{n\to\infty} \binom{n}{x} p^x(1-p)^{n-x}$
- To make taking the limit easier, we will note that we can re-write p in terms of n.
- By definition $E[X] = np$
- Let $E[X] = \lambda$.
- This implies that $\lambda/n = p$
- Therefore,

$$\begin{aligned}
\lim_{n\to\infty} \binom{n}{x} p^x(1-p)^{n-x} &= \lim_{n\to\infty} \left(\frac{n!}{x!(n-x)!} (\lambda/n)^x (1-\frac{\lambda}{n})^{n-x} \right) \\
&= \lim_{n\to\infty} \left(\frac{n!}{x!(n-x)!} (\frac{1}{n})^x (\lambda)^x (1-\frac{\lambda}{n})^{-x} (1-\frac{\lambda}{n})^{n} \right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\lambda^x}{x!}\lim_{n\to\infty}\left(\frac{n(n-1)...(n-x+1)}{n^x}(1-\frac{\lambda}{n})^{-x}(1-\frac{\lambda}{n})^n\right) \\
&= \frac{\lambda^x}{x!}\lim_{n\to\infty}\left(\frac{n}{n}\frac{n-1}{n}...\frac{n-x+1}{n}(1-\frac{\lambda}{n})^{-x}(1-\frac{\lambda}{n})^n\right) \\
&= \frac{\lambda^x}{x!}[\lim_{n\to\infty}\frac{n}{n}][\lim_{n\to\infty}\frac{n-1}{n}]...[\lim_{n\to\infty}\frac{n-x+1}{n}] \\
&\quad [\lim_{n\to\infty}(1-\frac{\lambda}{n})^{-x}][\lim_{n\to\infty}(1-\frac{\lambda}{n})^n] \\
&= \frac{\lambda^x}{x!}(1)...(1)(1)e^{-\lambda} \\
&\quad \uparrow \text{ taking limit; one definition of } e^x \text{ is } e^x = \lim_{n\to\infty}(1+\tfrac{x}{n})^n \\
&= \frac{\lambda^x e^{-\lambda}}{x!}
\end{aligned}
$$

- We say that if Random Variable X has a PDF $p_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}$ where $\lambda > 0$, then X is said to have a *Poisson* Distribution ($X \sim Pois(\lambda)$)

2.2 Verification, Mean, & Variance

- We will esatablish the following for this PDF
  1. Verify that the PDF is valid (i.e., show that it follows the two rules for discrete R.V. PDFs)
  2. Establish the meand of distribution
  3. Establish the variance of the distribution

1. Verify
   a) $\sum_{x\in S} p_X(x) = 1$

      Proof.

$$
\begin{aligned}
\sum_{x\in S} p_X(x) &= \sum_{x=0}^{\infty}\frac{\lambda^x e^{-\lambda}}{x!} \\
&= e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} \\
&= e^{-\lambda}(e^{\lambda}) \leftarrow \textstyle\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} \text{ is the taylor expansion of } e^{\lambda} \\
&= 1
\end{aligned}
$$

□

   b) $0 \le p_X(x) \le 1 \forall x \in S$

Proof. We see that $e^{-\lambda}, \lambda^x$, and $\frac{1}{x!}$ are all non-negative, thus $0 \leq p_X(x) \forall x \in S$, the support of X. Since $0 \leq p_X(x) \forall x \in S$ and $\sum_{x \in S} p_X(x) = 1$, we can conclude that $p_X(x) \leq 1 \forall x \in S$. Therefore, $0 \leq p_X(x) \leq 1 \forall x \in S$. □

2. Mean

$$E[X] = \lambda$$

Proof.

$$\begin{aligned}
E[X] &= \sum_{x \in S} x p_X(x) \\
&= \sum_{x=0}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!} \\
&= \sum_{x=1}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!} \leftarrow \text{First term is just 0} \\
&= \sum_{x=1}^{\infty} \frac{\lambda^x e^{-\lambda}}{(x-1)!} \\
&= \lambda \sum_{x=1}^{\infty} \frac{\lambda^{x-1} e^{-\lambda}}{(x-1)!} \\
&= \lambda \sum_{x-1=0}^{\infty} \frac{\lambda^{x-1} e^{-\lambda}}{(x-1)!} \\
&= \lambda \sum_{y=0}^{\infty} \frac{\lambda^y e^{-\lambda}}{(y)!} \leftarrow \text{Relabeling } x-1=y \\
&= \lambda(1) \leftarrow \text{We have already shown that } \textstyle\sum_{y=0}^{\infty} \frac{\lambda^y e^{-\lambda}}{y!} = 1 \\
&= \lambda
\end{aligned}$$

□

3. Variance

$$V[X] = \lambda$$

Proof. Proof Left as Homework □