Hypergeometric and Poisson Distribution

1. Hypergeometric Distribution

Consider a population comprising of $N(\geq 2)$ units out of which $a(\in \{1, 2, \ldots, N-1\})$ are labeled as s (success) and $N-a$ are labeled as f (failure). A sample of size n is drawn from this population drawing one unit at a time. Let

$$X = \text{ number of successes in drawn sample}$$

Case I: Suppose draws are independent and sampling is with replacement (i.e., after each draw the drawn unit is replaced back into the population). Then we have a sequence of n independent Bernoulli trials with probability of success in each trial is $p = \frac{a}{N}$ and, therefore $X \sim \text{Bin}(n, \frac{a}{N})$.

Case II: Suppose sampling is without replacement (i.e., after each draw the drawn unit is not replaced back into the population).

$$\begin{aligned}
&P(\text{obtaining } s \text{ in first draw}) = \frac{a}{N};\\
&P(\{\text{obtaining } s \text{ in second draw}\}) = \frac{a}{N}.\frac{a-1}{N-1} + \frac{N-a}{N}.\frac{a}{N-1} = \frac{a}{N};\\
&P(\{\text{obtaining } s \text{ in third draw}\}) = \frac{a}{N}.\frac{a-1}{N-1}.\frac{a-2}{N-2} + \frac{a}{N}.\frac{N-a}{N-1}.\frac{a-1}{N-2}\\
&+ \frac{N-a}{N}.\frac{a}{N-1}.\frac{a-1}{N-2} + \frac{N-a}{N}.\frac{N-a-1}{N-1}.\frac{a}{N-2} = \frac{a}{N};
\end{aligned}$$

In general, $P(\{\text{obtaining } s \text{ in } k-\text{th draw}\}) = \dfrac{a}{N}$.

Remark 1. *$P($obtaining s in first draw$) = \frac{a}{N}.\frac{a-1}{N-1}$ and*
$P($obtaining s in first and second draw$)P($obtaining s in second draw$) = \frac{a}{N}.\frac{a}{N}$

This implies that the draws are not independent. Therefore, we cannot conclude that $X \sim Bin(n, \frac{a}{N})$.

For $P(\{X = x\}) \neq 0$, we have $0 \leq x \leq n, 0 \leq x \leq a$ and $0 \leq n - x \leq N - a$. Thus $P(\{X = x\}) \neq 0$, $x \in \{\max(0, n - N + a), \ldots, \min(n, a)\}$. Therefore the r.v. X is of discrete type with support $E_X = \{\max(0, n - N + a), \ldots, \min(n, a)\}$ and p.m.f.

$$(1) \quad f_X(x) = P(\{X = x\}) = \begin{cases} \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}}, & \text{if } x \in \{\max(0, n-N+a), \ldots, \min(n, a)\} \\ 0, & \text{otherwise} \end{cases}$$

The random variable X is called a Hypergeometric random variable and it is written as $X \sim \text{Hyp}(a, n, N)$. The probability distribution with the p.m.f. (1) is called a Hypergeometric distribution. Also, we have

$$(2) \qquad \sum_{x=\max(0,n-N+a)}^{\min(n,a)} \binom{a}{x}\binom{N-a}{n-x} = \binom{N}{n}$$

Now, the expectation of $X \sim \mathrm{Hyp}(a, n, N)$ is

$$
\begin{aligned}
E(X) &= \sum_{x \in E_X} x f_X(x) \\
&= \frac{1}{\binom{N}{n}} \sum_{x=\max(0,n-N+a)}^{\min(n,a)} x \binom{a}{x} \binom{N-a}{n-x} \\
&= \frac{a}{\binom{N}{n}} \sum_{x=\max(1,n-N+a)}^{\min(n,a)} \binom{a-1}{x-1} \binom{N-a}{n-x} \\
&= \frac{a}{\binom{N}{n}} \sum_{x=\max(0,n-N+a-1)}^{\min(n-1,a-1)} \binom{a-1}{x} \binom{(N-1)-(a-1)}{(n-1)-x} \\
&= \frac{a\binom{N-1}{n-1}}{\binom{N}{n}} \\
&= \frac{an}{N};
\end{aligned}
$$

$$
\begin{aligned}
E(X^2) &= \sum_{x \in E_X} x^2 f_X(x) \\
&= \frac{1}{\binom{N}{n}} \sum_{x=\max(0,n-N+a)}^{\min(n,a)} x^2 \binom{a}{x} \binom{N-a}{n-x} = \frac{1}{\binom{N}{n}} \sum_{x=\max(1,n-N+a)}^{\min(n,a)} x \frac{a!}{(a-x)!(x-1)!} \binom{N-a}{n-x} \\
&= \frac{1}{\binom{N}{n}} \sum_{x=\max(1,n-N+a)}^{\min(n,a)} (x-1+1) \frac{a!}{(a-x)!(x-1)!} \binom{N-a}{n-x} \\
&= \frac{1}{\binom{N}{n}} \left\{ a \sum_{x=\max(1,n-N+a)}^{\min(n,a)} \binom{a-1}{x-1} \binom{N-a}{n-x} + a(a-1) \sum_{x=\max(2,n-N+a)}^{\min(n,a)} \binom{a-2}{x-2} \binom{N-a}{n-x} \right\} \\
&= \frac{1}{\binom{N}{n}} \left\{ a \sum_{x=\max(0,n-N+a-1)}^{\min(n-1,a-1)} \binom{a-1}{x} \binom{(N-1)-(a-1)}{(n-1)-x} + a(a-1) \right. \\
&\quad \left. \sum_{x=\max(0,n-N+a-2)}^{\min(n-2,a-2)} \binom{a-2}{x} \binom{(N-2)-(a-2)}{(n-2)-x} \right\} \\
&= \frac{a\binom{N-1}{n-1}}{\binom{N}{n}} + \frac{a(a-1)\binom{N-2}{n-2}}{\binom{N}{n}} \\
&= \frac{an}{N} + \frac{a(a-1)n(n-1)}{N(N-1)};
\end{aligned}
$$

$$
Var(X) = E(X^2) - (E(X))^2 = \frac{an}{N} + \frac{a(a-1)n(n-1)}{N(N-1)} - \frac{a^2n^2}{N^2} = n\left(\frac{a}{N}\right)\left(1 - \frac{a}{N}\right)\left(\frac{N-n}{N-1}\right).
$$

Example 2. *An urn contains* 6 *red balls and* 14 *black balls.* 5 *balls are drawn randomly without replacement. What is the probability that exactly* 4 *red balls are drawn?*

Solution: Let us label the drawing of a red as success and the drawing of a black as a failure. Let X be the number of red balls drawn. Then $X \sim \text{Hyp}(6, 5, 20)$. Hence, the required probability is $P(\{X = 4\}) = \frac{\binom{6}{4}\binom{14}{1}}{\binom{20}{5}}$.

2. Poisson Distribution

Suppose some event E is occurring randomly over a period of time. Let X be the number of times the event E has occurred in an unit interval (say $(0, 1]$).

Assumptions:

(1) For each infinitesimal subinterval $(\frac{k-1}{n}, \frac{k}{n}], k = 1, 2, \ldots, n$, the probability that the event E will occur in this subinterval is $\frac{\lambda}{n}$ and the probability that the event E will not occur in this subinterval is $1 - \frac{\lambda}{n}$, where $\lambda > 0$ is a given constant;

(2) chance of two or more occurrences of the event E in each infinitesimal subinterval $(\frac{k-1}{n}, \frac{k}{n}], k = 1, 2, \ldots, n$, is so small that it can be neglected;

(3) if $(\frac{j-1}{n}, \frac{j}{n}]$ and $(\frac{k-1}{n}, \frac{k}{n}] (1 \leq j < k \leq n)$ are disjoint subintervals then the number of times the event E occurs in the interval $(\frac{j-1}{n}, \frac{j}{n}]$ is independent of the number of times the event E occurs in the interval $(\frac{k-1}{n}, \frac{k}{n}]$.

Remark 3. *Such type of events is known as rare events. It means that two such events are extremely unlikely to occur simultaneously or within a very short period of time. Arrivals of jobs, telephone calls, e-mail messages, traffic accidents, network blackouts, virus attacks, errors in software, floods, and earthquakes are examples of rare events.*

Under the above assumptions, in each infinitesimal subinterval $(\frac{k-1}{n}, \frac{k}{n}], k = 1, 2, \ldots, n$, event E can occur only 1 or 0 times and the probability of occurrence of event E in each of these subintervals is the same $(\frac{\lambda}{n})$. If we label the occurrence of event E in any of these subintervals as success and its non-occurrence as failure, then we have a sequence of n independent Bernoulli trials with probability of success in each trial as $p_n = \frac{\lambda}{n}$. Therefore, $X \equiv X_n \sim \text{Bin}(n, p_n)$, where $p_n = \frac{\lambda}{n}$. The p.m.f. of X is given by

$$\begin{aligned} f_n(k) &= \binom{n}{k} p_n^x (1 - p_n)^{n-k}, \text{ if } k = 1, 2, \ldots, n \\ &= \frac{1}{k!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right)(np_n)^k \left(1 - \frac{np_n}{n}\right)^n (1 - p_n)^{-k} \text{ if } k = 1, 2, \ldots, n \end{aligned}$$

Since $np_n = \lambda$ and $p_n \to 0$ as $n \to \infty$, $f_n(k) \to \frac{e^{-\lambda}\lambda^k}{k!}$ as $n \to \infty$.

Definition 4. *A discrete type random variable X is said to follow a Poisson distribution with parameter $\lambda > 0$ (written as $X \sim P(\lambda)$) if its support is $E_X = \{0, 1, 2, \ldots\}$ and its probability mass function is given by*

$$f_X(x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!}, & \text{if } x \in \{0, 1, 2, \ldots\} \\ 0, & \text{otherwise} \end{cases}$$

Remark 5. *From above discussion, it is clear that a Binomial distribution $Bin(n, p)$ with large n and small p can be approximated by a Poisson distribution $P(\lambda)$, where $\lambda = np$.*

Now, the m.g.f. of $X \sim \mathrm{P}(\lambda)$ is

$$\begin{aligned} M_X(t) &= E(e^{tX}) \\ &= \sum_{x \in E_X} e^{tx} f_X(x) \\ &= \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda}\lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} \\ &= e^{\lambda(e^t-1)},\ t \in \mathbb{R} \end{aligned}$$

Therefore,

$$\begin{aligned} &M_X^{(1)}(t) = \lambda e^t e^{\lambda(e^t-1)},\ t \in \mathbb{R}; \\ &M_X^{(2)}(t) = \lambda e^t e^{\lambda(e^t-1)} + (\lambda e^t)^2 e^{\lambda(e^t-1)},\ t \in \mathbb{R}; \\ &E(X) = M_X^{(1)}(0) = \lambda; \\ &E(X^2) = M_X^{(2)}(0) = \lambda + \lambda^2; \\ &\text{and } Var(X) = E(X^2) - (E(X))^2 = \lambda. \end{aligned}$$

Example 6. *Ninety-seven percent of electronic messages are transmitted with no error. What is the probability that out of* 200 *messages, at least* 195 *will be transmitted correctly?*

Solution: Let us label the transmission of messages with no error as success and otherwise as failure. Let X be the number of correctly transmitted messages. Then $X \sim \mathrm{Bin}(200, 0.97)$. Hence the required probability is

$$P(X \geq 195) = 1 - P(X \leq 194) = 1 - \sum_{x=0}^{194} \binom{200}{x} (0.97)^x (0.03)^{200-x}.$$

X cannot be approximated by the Poisson distribution because success probability is too large.

Let Y be the number of failures. Then $Y \sim \mathrm{Bin}(200, 0.03)$. Then Y can be approximated by the Poisson distribution $Z \sim \mathrm{P}(6)$ (since $np = 200 \times 0.03 = 6$). Hence the required probability is

$$P(X \geq 195) = P(Y \leq 5) \approx P(Z \leq 5) = \sum_{x=0}^{5} \frac{e^{-6}6^x}{x!}.$$