Problem 1: Modelling number of goals scored

Sayantan Mondal

Suppose X denote the number of goals scored by home team in premier league. We can assume X is a random variable. Then we have to build the probability distribution to model the probability of number of goals. Since X takes value in $\mathbb{N} = \{0, 1, 2, \cdots\}$, we can consider the geometric progression sequence as possible candidate model, i.e.,

$$S = \{a, ar, ar^2, ar^3, \cdots\}.$$

But we have to be careful and put proper conditions in place and modify S in such a way so that it becomes proper probability distributions.

Solution:-

(i)
Considering the given geometric sequence as the base to create a probability distribution model, we can get a $Geom(1-r)$ distribution i.e, $P(X=x) = (1-r)r^x$. For the given S the necessary conditions are $0 < a, r < 1$ and $a + r = 1$.

(ii)
The mean and variance clearly exist as we know the same for $Geom(1-r)$. We will deduce the same below.

(iii)

$$S_1 = \mathbb{E}[X] = \sum_{x=0}^{\infty} xP(X=x)$$

$$\mathbb{E}[X] = \sum_{x=0}^{\infty} x(1-r)r^x$$

$$S_1 = (1-r)[0 + 1.r + 2.r^2 + \ldots]$$

$$rS_1 = (1-r)[0.r + 1.r^2 + 2.r^3 + \ldots]$$

Subtracting the two equations:-

$$(1-r)S_1 = (1-r)[r + r^2 + r^3 + \ldots]$$

$$S_1 = \mathbb{E}[X] = \frac{r}{1-r}$$

Now we calculate $\mathbb{E}[X^2]$:-

$$S_2 = \mathbb{E}[X^2] = \sum_{x=0}^{\infty} x^2(1-r)r^x$$

$$S_2 = (1-r)[0 + 1.r + 4.r^2 + 9.r^3 + \ldots]$$

$$rS_2 = (1-r)[0 + r^2 + 4r^3 + 9r^4 + \dots]$$

Subtracting the above 2 equations:-

$$(1-r)S_2 = (1-r)[r + 3r^2 + 5r^3 + 7r^4 + \dots]$$

Now let $S_3 = r + 3r^2 + 5r^3 + 7r^4 + \dots$

$$rS_3 = r^2 + 3r^3 + 5r^4 + 7r^5 + \dots$$

Now subtract the above 2 equations:-

$$(1-r)S_3 = r + 2r^2 + 2r^3 + 2r^4 + \dots$$

$$(1-r)S_3 = r + \frac{r^2}{1-r}$$

Now substituting $(1-r)S_3$ into previous equation:-

$$(1-r)S_2 = r + \frac{r^2}{1-r}$$

$$S_2 = \frac{r + r^2}{(1-r)^2}$$

Now $Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

$$Var(X) = \frac{r + r^2}{(1-r)^2} - \frac{r^2}{(1-r)^2}$$

$$Var(X) = \frac{r}{(1-r)^2}$$

(iv)

Clearly the given statistics doesn't seem to follow the Geometric distribution as assuming mean to be true results in variance to be incorrect and vice versa. So if we assume the mean to be true for our model i.e, $\mathbb{E}[X] = 1.5$. So, $\mathbb{E}[X] = \frac{r}{1-r} = 1.5$ gives us $r = 0.6$.

(a)

$$\mathbb{P}(X \geq 1) = 1 - \mathbb{P}(X = 0)$$

$$= 1 - (1-r)$$

$$= r = 0.6$$

(b)

$$\mathbb{P}(1 \leq X < 4) = \sum_{x=1}^{3} \mathbb{P}(X = x)$$

$$= \sum_{x=1}^{3} (1-r)r^x$$

$$= 0.47$$

(v)

Using of the shelf Poisson distribution to model X, runs into the same issue, that the mean and variance is not equal. So likewise we model using mean as the parameter i.e, $\lambda = 1.5$. Therefore $P(X = k) = \frac{e^{-1.5}\lambda^k}{k!}$.

(a)

$$\mathbb{P}(X \geq 1) = 1 - \mathbb{P}(X = 0)$$

$$= 1 - e^{-\lambda}$$

$$= r = 0.77$$

(b)

$$\mathbb{P}(1 \leq X < 4) = \sum_{x=1}^{3} \mathbb{P}(X = x)$$

$$= \sum_{x=1}^{3} \frac{e^{-1.5}\lambda^x}{x!}$$

$$= 0.758$$

(vi)

Poisson would definitely be a better fit for the given model. Because of the two it has the least variance i.e, 1.5. Also the mean, median and mode i.e, the central tendencies are closer to the expected values, so Poisson would definitely be the correct choice for the modelling.

(vii)

For X~$Geom(r)$ we have:-

$$\mathcal{L}(r|x_1, x_2, x_3, \ldots, x_n) = \prod_{i=1}^{n} (1 - r) r^{x_i}$$

$$\mathcal{L}(r|x_1, x_2, x_3, \ldots, x_n) = (1 - r)\left[r^{x_1} . r^{x_2} . r^{x_3} \ldots r^{x_n}\right]$$

$$\mathcal{L}(r|x_1, x_2, x_3, \ldots, x_n) = (1 - r) r^{\sum_{i=1}^{n} x_i}$$

For X~$Poi(\lambda)$ we have:-

$$\mathcal{L}(\lambda|x_1, x_2, x_3, \ldots, x_n) = \prod_{i=1}^{n} \left(\frac{e^{-\lambda}\lambda^{k_i}}{k_i!}\right)$$

$$\mathcal{L}(\lambda|x_1, x_2, x_3, \ldots, x_n) = \frac{e^{-\lambda}\lambda^{k_1}}{k_1!} \cdot \frac{e^{-\lambda}\lambda^{k_2}}{k_2} \cdot \frac{e^{-\lambda}\lambda^{k_3}}{k_3!} \ldots \frac{e^{-\lambda}\lambda^{k_n}}{k_n!}$$

$$\mathcal{L}(\lambda|x_1, x_2, x_3, \ldots, x_n) = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} k_i}}{\prod_{i=1}^{n} k_i!}$$

The assumptions made are that the random variables are independent.