

Binomialfördelning och dess släktingar

När ett försök görs inträffar en händelse A med sannolikhet $p = P(A)$ oberoende av tidigare försök. Antalet försök tills A inträffar för första gången är $\text{ffg}(p)$ -fördelad, dvs

$$P(X = k) = (1 - p)^{k-1}p, \quad k = 1, 2, 3, \dots$$

för $k = 1, 2, 3, \dots$

Sedan tidigare är $E[X] = 1/p$ och man kan visa att $V(X) = (1 - p)/p^2$.

Den omvända frågeställningen är, av n gjorda försök, låt X beskriva antalet gånger A inträffar. Då är X binomialfördelad, det vill säga

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

för $k = 0, 1, \dots, n$.

Det är ofta vettigt att modellera situationen på följande sätt. Låt $U_1, \dots, U_n$ vara oberoende, likafördelade stokastiska variabler, där

$$U_k = \begin{cases} 1 & \text{med sannolikhet } p \\ 0 & \text{med sannolikhet } 1 - p \end{cases}$$

för $k = 1, \dots, n$. Att $U_k = 1$ betyder att A inträffade i försök k och $U_k = 0$ att A inte gjorde det. Notera att

$$E[U_k] = \sum_{i=0}^1 i P(U_k = i) = 0 \cdot (1 - p) + 1 \cdot p = p$$

och

$$E[U_k^2] = \sum_{i=0}^1 i^2 P(U_k = i) = 0^2 \cdot (1 - p) + 1^2 \cdot p = p,$$

och alltså $V(U_k) = p - p^2 = p(1 - p)$.

Med

$$X = \sum_{k=1}^n U_k$$

får vi att X beskriver antalet lyckade försök. Då är X binomialfördelad och

$$E[X] = E\left[\sum_{k=1}^n U_k\right] = \sum_{k=1}^n E[U_k] = np$$

och

$$V(X) = V\left(\sum_{k=1}^n U_k\right) = \{\text{oberoende}\} = \sum_{k=1}^n V(U_k) = np(1 - p)$$

Med modellen

$$X = \sum_{k=1}^n U_k$$

säger Centrala gränsvärdessatsen att för stora n är X approximativt normalfördelad med parametrar $\mu = np$ och $\sigma = \sqrt{np(1 - p)}$. Vi kräver dock för att approximationen skall fungera väl att p inte är för liten eller för stor. Vi formulerar det gemensamma kravet på n och p som

$$np(1 - p) = V(X) \geq 10.$$

Betrakta nu följande situation. Vi gör n försök i tiden, tex. under ett år. Om vi gör försöken tätare och tätare men samtidigt låter sannolikheten p för att något skall inträffa vid varje tidpunkt (försök) minska, så borde vi i gräns få en modell som räknar antalet inträffade händelser under året där händelser kan inträffa närsomhelst under vårt tidsintervall.

Formellt, låt X vara binomialfördelad

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

och låt $n \rightarrow \infty$ och $p \rightarrow 0$ så att $np = E[X]$ är konstant. Med, $\mu = np$ får vi att

$$\begin{aligned} P(X = k) &= \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{(n-k)!k!} \left(\frac{\mu}{n}\right)^k \left(1 - \frac{\mu}{n}\right)^{n-k} \\ &= \frac{n!}{(n-k)!n^k} \frac{1}{k!} \mu^k \left(1 - \frac{\mu}{n}\right)^n \frac{1}{(1 - \mu/n)^k}. \end{aligned}$$

Om vi låter n växa får vi

$$\lim_{\substack{n \rightarrow \infty \\ p \rightarrow 0}} P(X = k) = \lim_{\substack{n \rightarrow \infty \\ p \rightarrow 0}} \frac{n!}{(n-k)!n^k} \frac{1}{k!} \mu^k \left(1 - \frac{\mu}{n}\right)^n \frac{1}{(1 - \mu/n)^k} = \frac{\mu^k}{k!} e^{-\mu}.$$

Detta är en giltig sannolikhetsfunktion på $\{0, 1, \dots\}$

$$\sum_{k=0}^{\infty} \frac{\mu^k}{k!} e^{-\mu} = e^{-\mu} \underbrace{\sum_{k=0}^{\infty} \frac{\mu^k}{k!}}_{=e^{\mu}} = e^{-\mu} e^{\mu} = 1$$

för alla värden på $\mu > 0$. Vi gör följande definition

X är Poissonfördelad med parameter $\mu > 0$ om

$$P(X = k) = \frac{\mu^k}{k!} e^{-\mu}$$

för $k = 0, 1, \dots$. Modellsituationen: antal händelser som inträffar under ett (tids-)intervall där händelser inträffar oberoende av varandra och med konstant intensitet. Ur härledningen ovan följer även att

$$E[X] = np = \mu \quad \text{och} \quad V(X) = \mu.$$

För två oberoende Poissonfördelade stokastiska variabler, X är $\text{Po}(\mu_x)$ och Y är $\text{Po}(\mu_y)$, har man att

$$X + Y \text{ är } \text{Po}(\mu_x + \mu_y).$$

Bevis:

$$\begin{aligned} P(X + Y = n) &= \sum_k P(X = k, Y = n - k) = \{\text{ober.}\} = \sum_{k=0}^n P(X = k) P(Y = n - k) \\ &= \sum_{k=0}^n \frac{\mu_x^k}{k!} e^{-\mu_x} \frac{\mu_y^{n-k}}{(n-k)!} e^{-\mu_y} = \frac{1}{n!} e^{-(\mu_x + \mu_y)} \sum_{k=0}^n \frac{n!}{k!(n-k)!} \mu_x^k \mu_y^{n-k} \\ &= \frac{1}{n!} e^{-(\mu_x + \mu_y)} (\mu_x + \mu_y)^n = \frac{\mu^n}{n!} e^{-\mu} \end{aligned}$$

en Poissonfördelning med parameter $\mu = \mu_x + \mu_y$.

Tag nu $\mu > 0$. Låt, för ett stort n , $\mu_n = \mu/n$, och $X_1, \dots, X_n$ vara oberoende och Poissonfördelade med parameter μ_n . Enligt det ovanstående är

$$X = \sum_{k=1}^n X_k$$

Poissonfördelad med parameter μ , men enligt Centrala gränsvärdessatsen är X approximativt normalfördelad med parametrar m och $\sigma = \sqrt{\mu}$. Slutsatsen är att Poissonfördelningen kan approximeras med Normalfördelningen. För att approximationen skall fungera bra kräver vi att μ_n inte är alltför liten, dvs variablerna X_k skall ha hyfsade sannolikheter att inte bara vara nollor. Vi formulerar kravet i termer av μ . Om $\mu > 15$ kan vi approximera Poissonfördelningen med normalfördelningen.

Konstruktionen vi gjorde av Poissonfördelningen som gränsvärde till Binomialfördelningen medger approximation av Binomial med Poisson om n är stor och p är liten. Det viktiga är värdena på p och vi kräver $p < 0.10$ för en hyfsad approximation.

Hypergeometrisk fördelning

Exempel: Opinionsundersökning (ändlig population): I en population om N individer finns det s sympatisörer, dvs. $p = s/N$ är andelen sympatisörer. Välj ut n på måfå utan återläggning och låt X beskriva antalet sympatisörer. Då är

$$P(X = k) = \frac{\binom{s}{k} \binom{N-s}{n-k}}{\binom{N}{n}} = \frac{\binom{Np}{k} \binom{N(1-p)}{n-k}}{\binom{N}{n}}, \quad \max(n - (N - s), 0) \leq k \leq \min(n, s)$$

(Notera: om urvalet sker med återläggning är X binomialfördelad, X är $\text{Bin}(n, p)$.)

$$E[X] = np \quad V(X) = np(1-p) \frac{N-n}{N-1}$$

Om populationen är stor i förhållande till urvalet kan den hypergeometrisk fördelningen approximeras av binomialfördelningen. Tumregel: $n/N < 0.10$.