Binomialfördelning och dess släktingar

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

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

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

för $k = 0, 1, \ldots, n$.

Vi skall nu bevisa påståendet ovan på ett alternativt sätt. Det är ofta vettigt att modellera situationen på följande sätt. Låt $U_1, \ldots, U_n$ vara oberoende, likafördelade stokastiska variabler, där

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

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

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

och

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

och alltså $\mathrm{V}(U_k) = p - p^2 = p(1-p)$.

Med

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

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

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

och

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

Fördelningen för X kan nu bestämmas med t.ex. dess sannolikhetsgenererande funktion.

$$m_X(s) = \mathsf{E}\left[s^X\right] = \sum_k k \mathsf{P}(X = k) = \mathsf{P}(X = 0) + \mathsf{P}(X = 1)\, s + \cdots + \mathsf{P}(X = n)\, s^n$$

men

$$m_X(s) = \mathsf{E}\left[s^X\right] = \mathsf{E}\left[s^{\sum_{i=1}^{n} U_i}\right] = \mathsf{E}\left[s^{U_1}\right] \cdots \mathsf{E}\left[s^{U_n}\right] = \left[\mathsf{E}\left[s^U\right]\right]^n.$$

Nu är

$$\mathsf{E}\left[s^U\right] = s^0 \underbrace{(1-p)}_{=q} + s^1 p = q + ps$$

så

$$m_X(s) = (q + ps)^n = \sum_{k=0}^{n} \binom{n}{k} (ps)^k q^{n-k},$$

det vill säga $\mathsf{P}(X=k) = \binom{n}{k} p^k q^{n-k}$.

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

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

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

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

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

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

Om vi låter n växa får vi

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

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

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

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

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

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

för $k = 0, 1, \ldots,$. 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

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

En Poissonfördelad stokastisk variabel har

$$g_X(s) = \mathsf{E}\left[s^X\right] = \sum_{k=0}^{\infty} s^k \frac{\mu^k}{k!} \mathrm{e}^{-\mu} = \mathrm{e}^{-\mu} \cdot \mathrm{e}^{\mu s} = \mathrm{e}^{\mu(s-1)}.$$

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

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

Bevis:

$$g_{X+Y}(s) = g_X(s)g_Y(s) = \mathrm{e}^{\mu_x(s-1)}\mathrm{e}^{\mu_y(s-1)} = \mathrm{e}^{(\mu_x+\mu_y)(s-1)} = \mathrm{e}^{\mu(s-1)}$$

den sannolikhetsgenererande funktionen för en Poissonfördelning med parameter $\mu = \mu_x + \mu_y$.

Tag nu $\mu > 0$. Låt, för ett stort n, $\mu_n = \mu/n$, och $X_1, \ldots, X_n$ 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

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

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

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

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