

Binomialfördelning och dess släktingar (forts)

Genomgående modellsituation: Betrakta oberoende försök där vid varje försök det finns en sannolikhet p att en händelse A inträffar.

Låt X beteckna antalet gånger man får göra försöket tills man ser att A inträffar för första gången.

Då är X för första gången-fördelad, skrivet X är $\text{ffg}(p)$, om

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

för $k = 1, 2, 3, \dots$

$$\mathbf{E}[X] = \frac{1}{p} \quad \mathbf{V}(X) = \frac{1-p}{p^2}.$$

Den omvända frågeställningen är, om man gör n försök, låt X vara antalet gånger av dessa som A inträffar. Då får vi att X är binomialfördelad, skrivet X är $\text{Bin}(n, p)$, dvs.

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

för $k = 0, 1, \dots, n$.

Bevis i repris: Betrakta följande sekvensen av n försök där man noterat att A inträffat k gånger.

$$\underbrace{AA \cdots A}_{k \text{ st}} \underbrace{A^* \cdots A^*}_{n-k \text{ st}}$$

Varje sekvens av k stycken A och $n - k$ stycken A^* har pga. oberoendet sannolikhet $p^k(1 - p)^{n-k}$. När $X = k$ skall vi summera över antalet sådana sekvenser, men detta är $\binom{n}{k}$, och således är

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

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

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

och

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

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

Med

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

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

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

och

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

pga oberoendet mellan $U_1, \dots, U_n$.

Med modellen

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

säger Centrala gränsvärdessatsen att för stora n är X approximativt normalfördelad med parametrar $m = np$ 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

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

Vi kan även approximera den hypergeometrisk fördelningen (som uppstår vid dragning utan återläggning) med binomialfördelningen (med återläggning) om populationen man väljer ifrån är stor, N , i förhållande till antalet element man väljer ut, n , $n \ll N$.

Tänk 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 tidintervall.

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, $m = np$ får vi att

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

Om vi låter n växa får vi

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

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

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

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

X är Poissonfördelad med parameter $m > 0$ om

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

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

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

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

$$X + Y \text{ är } \text{Po}(m_x + m_y).$$

Bevis:

$$\begin{aligned} P(X + Y = n) &= \sum_k P(X = k, Y = n - k) = \sum_{k=0}^n P(X = k) P(Y = n - k) \\ &= \sum_{k=0}^n \frac{m_x^k}{k!} e^{-m_x} \frac{m_y^{n-k}}{(n-k)!} e^{-m_y} = \frac{1}{n!} e^{-(m_x+m_y)} \underbrace{\sum_{k=0}^n \frac{n!}{k!(n-k)!} m_x^k m_y^{n-k}}_{=(m_x+m_y)^n} \\ &= \frac{(m_x + m_y)^n}{n!} e^{-(m_x+m_y)} = \frac{m^n}{n!} e^{-m}. \end{aligned}$$

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

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

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

Sammanfattning approximationer:

