Binomialfördelningen (forts)

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

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.

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

för $k=0,1,\ldots,n$.

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

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

och

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

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

Med

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

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

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

och

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

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

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

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

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

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

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

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

Om vi låter n växa får vi

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

Detta är en giltig sannolikhetsfunktion på $\{0,1,\ldots\}$. Vi gör följande definition:

Definition: X är Poissonfördelad med parameter $m>0$ om

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

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

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

Sats. Låt X och Y vara två oberoende $\mathrm{Po}(m_x)$- och $\mathrm{Po}(m_y)$-fördelade stokastiska variabler. Då är

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

Bevis: Se boken.

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

Omvänd problemställning: 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 $\mathrm{ffg}(p)$, om

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

för $k=1,2,3,\ldots$.

$$\mathsf{E}[X]=\frac{1}{p} \qquad \mathrm{V}(X)=\frac{1-p}{p^2}.$$

Punktskattningar

Stickprovsundersökningar — observationer på delar av en population för att få kunskap om helheten.

Exempel: Opinionsundersökning. Av $n = 2854$ intervjuade personer sade sig $x = 1085$ sympatisera med socialdemokraterna. En skattning av p, andelen socialdemokrater i populationen, är $p^* = x/n = 1085/2854 = 0.380$ dvs andelen av de observerade.

Exempel: brinntid för ljus beskrivs av X med $m = \mathsf{E}[X]$ och $\sigma^2 = \mathrm{V}(X)$. Vad är m och σ? Med observationer skattas m och σ^2 med

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

respektive

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left(\left(\sum_{i=1}^{n} x_i^2\right) - n\overline{x}^2\right).$$

Definition: Ett (slumpmässigt) *stickprov* (av storlek n) är en serie av observationer $x_1, \ldots, x_n$ av oberoende stokastiska variabler $X_1, \ldots, X_n$.

Oftast har $X_1, \ldots, X_n$ samma fördelning.

En punktskattning av en parameter θ är en funktion av ett stickprov som skall ge oss information om θ:

$$\theta^* = \theta^*(x_1, \ldots, x_n)$$

Skattningar modelleras med skattningsvariabeln (stickprovsvariabeln)

$$\theta^* = \theta^*(X_1, \ldots, X_n).$$