Förel 8 15 11 23

7 Binomial- och Poissonfördelningen. Normalapproximation

Vi avslutade förra föreläsningen med centrala gränsvärdessatsen som sade att för en oändlig följd av oberoende, likafördelade s.v. $X_1, \dots X_n, \dots$ med $E(X_i) = \mu$ och $V(X_i) = \sigma$, så gäller det att

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \in AsN(\mu, \sigma/\sqrt{n}) \quad \text{då } n \to \infty.$$

Från sambandet mellan allmänna och standardiserade normalfördelade s.v. så får vi approximationen

$$P(a < \overline{X}_n \leq b) \approx \Phi\left(\frac{b-\mu}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{a-\mu}{\sigma/\sqrt{n}}\right)$$

som alltså är användbar även om X_i:na inte är normalfördelade. Vi kommer att undersöka ytterligare några approximationer under dagens föreläsning. Vi repeterar först binomialfördelningen. Betrakta ett försök där en händelse A inträffar med sannolikhet p. Om vi upprepar försöket n gånger och låter X vara antalet gånger som A inträffar så gäller $X \in Bin(n,p)$, dvs

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{för } k = 0, 1, \dots, n.$$

Vi kan också komma fram till binomialfördelningen på följande sätt. Till vart och ett av de n försöken kan vi koppla en Bernoullifördelad s.v. I_i som är ett om A inträffar (sker med sannolikhet p) och noll annars. Antalet gånger som A inträffar på n försök kan då skrivas $X = I_1 + I_2 + \dots + I_n \in Bin(n,p)$. Notera att de s.v. I_i är oberoende och likafördelade. Vi använder detta skrivsätt för att bestämma väntevärde och varians av binomialfördelade s.v.

$$E(X) = E(I_1 + \dots + I_n) = nE(I_1) = n(p \cdot 1 + (1-p)\cdot 0) = np$$

$$V(X) = V(I_1 + \dots + I_n) \underset{I_i \text{ oberoende}}{=} nV(I_1) = n(E(I_1^2) - \underbrace{(E(I_1))^2}_{p^2}) = n(p \cdot 1^2 + (1-p)\cdot 0^2 - p^2) = np(1-p)$$

Sats: Låt $X_1 \in Bin(n_1, p)$ och $X_2 \in Bin(n_2, p)$ vara oberoende. Då är $X_1 + X_2 \in Bin(n,p)$

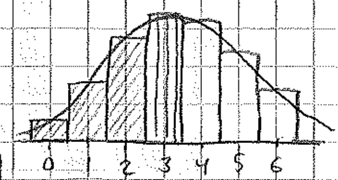
Satsen är naturlig om vi tänker på X_1 och X_2 som summor av oberoende $Be(p)$-fördelade s.v. Från representationen $X = I_1 + \dots + I_n$ följer också att vi kan använda centrala gränsvärdessatsen på $X \in Bin(n,p)$.

Sats: Om $X \in Bin(n,p)$ och $np(1-p) \geq 10$, så gäller approximativt $X \in N(np, \sqrt{np(1-p)})$. Därmed gäller

$$P(a < X \leq b) \approx \Phi\left(\frac{b-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-np}{\sqrt{np(1-p)}}\right)$$

Vi approximerar här en diskret s.v. (för vilken $P(X \leq k)$ är skilt från $P(X < k)$ för alla $k \in \{0, 1, \dots, n\}$) med en kontinuerlig s.v. (för vilken $P(X \leq k) = P(X < k)$ för alla $k \in \{0, 1, \dots, n\}$). Approximationen blir därför bättre om använder så kallade halvkorrektion dvs

$$P(X < k) = \Phi\left(\frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) \qquad P(X \leq k) = \Phi\left(\frac{k + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$$



vilket ger $P(a < X \leq b) \approx \Phi\left(\frac{b+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right)$

Med halvkorrektion blir approximationen god för $np(1-p) \geq 3$.

Ex: Vi kastar en tärning 60 gånger och låter X beteckna antalet gånger vi får en sexa. $X \in Bin(60, 1/6)$. Vi vill bestämma $P(X > 13)$. Det exakta antalet är $\sum_{k=14}^{60} \binom{60}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{60-k} = 0{,}115$. Med normalapproximation får vi

$$P(X > 13) = 1 - \Phi\left(\frac{13 - 60 \cdot 1/6}{\sqrt{60 \cdot \frac{1}{6} \cdot \frac{5}{6}}}\right) = 0{,}149$$

och med halvkorrektion

$$P(X > 13) = 1 - \Phi\left(\frac{13 + \frac{1}{2} - 60 \cdot \frac{1}{6}}{\sqrt{60 \cdot \frac{1}{6} \cdot \frac{5}{6}}}\right) = 0{,}113$$

Notera att $np(1-p) = 8{,}33$ här.

Vi har tidigare sett att när n är stort och p litet så är $X \in Bin(n,p)$ approximativt $Po(np)$-fördelat, dvs

$$p_X(k) \approx \frac{\mu^k}{k!}e^{-\mu}, \quad k=0,1,2,\ldots \quad \text{för } \mu = np.$$

Approximationen visar sig vara god för $p<0,1$ (faktiskt oavsett n). Vi har visat att för $Y \in Po(\mu)$, så är $E(Y)=\mu$. För att bestämma $V(Y)$ beräknar vi $E(Y(Y-1))$:

$$E(Y(Y-1)) = \sum_{k=0}^{\infty} k(k-1)\frac{\mu^k}{k!}e^{-\mu} = \mu^2\sum_{k=2}^{\infty} \frac{\mu^{k-2}}{(k-2)!}e^{-\mu} = \mu^2 e^{-\mu}\underbrace{\sum_{l=0}^{\infty} \frac{\mu^l}{l!}}_{e^{\mu}} = \mu^2$$

Termerna med $k=0$ och $k=1$ är noll $\qquad l=k-2$

så $V(Y)=E(Y^2)-(E(Y))^2 = E(Y(Y-1))+E(Y)-(E(Y))^2 = \mu^2+\mu-\mu^2=\mu$.

När vi undersökte summor av oberoende kontinuerliga s.v. så härledde vi faltningsformeln $f_{X+Y}(z)=\int_{-\infty}^{\infty} f_X(x)f_Y(z-x)\,dx$. För diskreta s.v. gäller, på samma sätt, att $p_{X+Y}(k)=\sum_{i=0}^{k} p_X(i)p_Y(k-i)$.

<u>Sats:</u> Låt $X\in Po(\mu_1)$ och $Y\in Po(\mu_2)$ vara oberoende. Då är $X+Y\in Po(\mu_1+\mu_2)$

<u>Bevis:</u>
$$P(X+Y=k)=\sum_{i=0}^{k} P(X=i)P(Y=k-i) = \sum_{i=0}^{k} \frac{\mu_1^i}{i!}e^{-\mu_1}\frac{\mu_2^{k-i}}{(k-i)!}e^{-\mu_2}$$
$$= e^{-(\mu_1+\mu_2)}\frac{1}{k!}\underbrace{\sum_{i=0}^{\infty}\frac{k!}{i!(k-i)!}\mu_1^i\mu_2^{k-i}}_{(\mu_1+\mu_2)^k} = \frac{(\mu_1+\mu_2)^k}{k!}e^{-(\mu_1+\mu_2)}$$

Additionssatsen implicerar att om μ är ett heltal, så kan $X\in Po(\mu)$ skrivas $X=Z_1+Z_2+\ldots+Z_\mu$, där $Z_i\in Po(1)$. Eftersom $E(Z_i)=V(Z_i)=1$, så följer det av centrala gränsvärdessatsen att X är approximativt $N(\mu,\sqrt{\mu})$-fördelad för stora μ. Approximationen fungerar även när μ inte är ett heltal och är relativt god för $\mu>15$. Precis som vid normalapproximation av binomialfördelade s.v. förbättras approximationen med halvkorrektion.

<u>Ex:</u> Sannolikheten för att en etta byts mot en nolla i en dataöverföring av en bit är $p=10^{-7}$. Antalet fel X vid överföring av $8\cdot10^6$ bitar är $Bin(8\cdot10^6,10^{-7})$-fördelat, dvs approximativt $Po(0,8)$-fördelat. Sannolikheten för minst tre fel $P(X\geq3)=1-P(X\leq2)$ är enligt tabellen för $Po(0,8)$ 4,74%. Med miniräknaren får vi samtidigt $P(X\geq3)=1-\text{binocdf}(2,8\cdot10^6,10^{-7})=4,74\%$.

Vi avslutar med att undersöka den hypergeometriska fördelningen. Vi har tidigare studerat urnmodeller. Antag att en urna innehåller v vita och s svarta kulor. Låt X beteckna antalet vita kulor som fås vid dragning av n kulor utan återläggning

(med återläggning $\Rightarrow X\in Bin(n,\frac{v}{s+v})$) $\quad P(X=k)=\binom{v}{k}\binom{s}{n-k}\Big/\binom{s+v}{n}, \quad k=0,1,\ldots,n$

Vi ändrar beteckningarna något och låter nu N vara totala antalet kulor ($N=s+v$) och p proportionen av vita kulor. Då får vi, för $k=0,1,\ldots,Np$, $n-k=0,1,\ldots,N(1-p)$ att

$$P(X=k)=\binom{Np}{k}\binom{N(1-p)}{n-k}\Big/\binom{N}{n}$$

Vi säger att X är en hypergeometrisk s.v. ($X\in Hyp(N,n,p)$). Precis som för binomialfördelade s.v. så kan vi skriva $X=I_1+\ldots+I_n$ där $I_i\in Be(p)$, eftersom det gäller att vid var och en av de n dragningarna så är sannolikheten för vit kula p. I motsats till fallet med binomialfördelade s.v. så är I_i:na inte oberoende, eftersom dragningen sker utan återläggning. Kovariansen mellan I_i och I_j är inte noll och i boken visas att för $X\in Hyp(N,n,p)$ så gäller

$$E(X)=np \qquad V(X)=\underbrace{\frac{N-n}{N-1}}\,np(1-p)$$

Korrektionsfaktor d_n^2 för ändliga populationer

Om $n \ll N$ så spelar inte återläggningen så stor roll och $d_n^2\approx1$. Om $X\in Hyp(N,n,p)$ och $n/N\leq0,1$ så är X approximativt $Bin(n,p)$-fördelad. Även normalapproximation kan användas.