Föreläsning 8

Thomas Önskog

22/11 2017

Binomialfördelningen

På förra föreläsningen visade vi **centrala gränsvärdessatsen** som säger att en oändlig följd $X_1, X_2, ...$ av oberoende, likafördelade s.v. med väntevärde μ och standardavvikelse σ uppfyller

$$\frac{1}{n}\sum_{i=1}^{n} X_i \in \mathrm{AsN}(\mu, \sigma/\sqrt{n}), \qquad \text{dvs} \qquad \mathbb{P}\Big(a < \frac{1}{n}\sum_{i=1}^{n} X_i \leq b\Big) \approx \Phi\Big(\frac{b-\mu}{\sigma/\sqrt{n}}\Big) - \Phi\Big(\frac{a-\mu}{\sigma/\sqrt{n}}\Big),$$

vilket alltså gäller även om de s.v. X_i inte är normalfördelade. Vi kommer att undersöka ytterligare några approximationer idag, men börjar med att repetera **binomialfördelningen**.

Betrakta ett försök där en händelse A inträffar med sannolikhet p. Om försöket upprepas n gånger och vi låter X beteckna antalet gånger som händelsen A inträffar, så är $X \in \mathrm{Bin}(n,p)$. Vi kan också härleda 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 och noll annars. Antalet gånger som A inträffar på n försök kan då skrivas

$$X = I_1 + \cdots + I_n \in \mathrm{Bin}(n,p)$$

Notera att de s.v. I_i är oberoende och likafördelade med väntevärde

$$E(I_i) = p \cdot 1 + (1-p)\cdot 0 = p,$$

och varians

$$V(I_i) = E(I_i^2) - (E(I_i)^2) = (p\cdot 1^2 + (1-p)\cdot 0^2) - p^2 = p - p^2 = p(1-p).$$

Väntevärdet och variansen för binomialfördelade s.v. blir därmed

$$\underline{E(X)} = E(I_1 + \cdots + I_n) = E(I_1) + \cdots + E(I_n) = \underline{np},$$

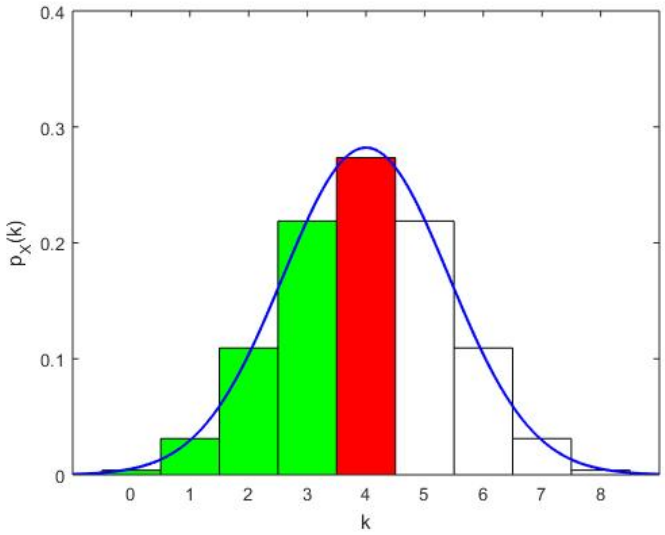
$$\underline{V(X)} = V(I_1 + \cdots + I_n) = [I_1, ..., I_n \text{ är oberoende}] = V(I_1) + \cdots + V(I_n) = \underline{np(1-p)}.$$

Genom att betrakta binomialfördelade s.v. som summor av oberoende $\mathrm{Be}(p)$-fördelade s.v., så är det uppenbart att om $X_1 \in \mathrm{Bin}(n_1,p)$ och om $X_2 \in \mathrm{Bin}(n_2,p)$, så är $X_1 + X_2 \in \mathrm{Bin}(n_1+n_2,p)$. Med detta synsätt följer också att vi kan använda centrala gränsvärdessatsen för $X \in \mathrm{Bin}(n,p)$.

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

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

Vi approximerar här en diskret s.v. (för vilken $\mathbb{P}(X \leq k)$ är skilt från $\mathbb{P}(X < k)$ för alla $k = 0, 1, ..., n$) med en kontinuerlig s.v. (för vilken $\mathbb{P}(X \leq k) = \mathbb{P}(X < k)$ för alla $k = 0, 1, ..., n$). Approximationen förbättras om vi använder så kallad **halvkorrektion**.



Figuren ovan visar sannolikhetsfunktionen för en s.v. $X \in \text{Bin}(8, 1/2)$. Sannolikheten $\mathbb{P}(X < 4)$ motsvarar arean av de gröna staplarna i figuren. Denna area är nästan lika stor som arean under normalfördelningens täthetsfunktion på intervallet $(-\infty, 3.5)$. På samma sätt är $\mathbb{P}(X \leq 4)$ lika med summan av arean av de gröna staplarna och den röda stapeln i figuren, dvs ungefär lika med arean under normalfördelningens täthetsfunktion på intervallet $(-\infty, 4.5)$. Med halvkorrektion blir approximationen i satsen ovan god för $np(1-p) \geq 3$.

Exempel. *Vi kastar en tärning 60 gånger och låter X beteckna antalet gånger som vi får en sexa. Vi vill bestämma $\mathbb{P}(X > 13)$. Eftersom $X \in Bin(60, 1/6)$, så är den sökta sannolikheten*

$$\sum_{k=14}^{60} \binom{60}{k} \Big(\frac{1}{6}\Big)^k \Big(\frac{5}{6}\Big)^{60-k} = 0.115.$$

Med normalapproximation får vi (notera att $np(1-p) = 8.33$ i detta exempel)

$$\mathbb{P}(X > 13) = 1 - \mathbb{P}(X \leq 13) \approx 1 - \Phi\Big(\frac{13 - 60(1/6)}{\sqrt{60(1/6)(5/6)}}\Big) = 0.149,$$

och med halvkorrektion

$$\mathbb{P}(X > 13) = 1 - \mathbb{P}(X \leq 13) \approx 1 - \Phi\Big(\frac{13 + 1/2 - 60(1/6)}{\sqrt{60(1/6)(5/6)}}\Big) = 0.113.$$

Poissonfördelningen

Vi har tidigare sett att när n är stort och p litet, så är $X \in \text{Bin}(n,p)$ approximativt Poissonfördelat med parameter $\mu = np$. Approximationen är god för $p < 0.1$ (faktiskt oavsett n). Vi har också visat att för $Y \in \text{Po}(\mu)$, så gäller $E(Y) = \mu$. Vi ska nu bestämma $V(Y)$, för $Y \in \text{Po}(\mu)$, och börjar med att bestämma $E(Y(Y-1))$.

$$\begin{aligned} E(Y(Y-1)) &= \sum_{k=0}^{\infty} k(k-1)\frac{\mu^k}{k!}e^{-\mu} = [\text{Första två termerna är lika med noll}] \\ &= \sum_{k=2}^{\infty} \frac{\mu^k}{(k-2)!}e^{-\mu} = \mu^2 e^{-\mu}\sum_{k=2}^{\infty}\frac{\mu^{k-2}}{(k-2)!} = \mu^2 e^{-\mu}\underbrace{\sum_{i=0}^{\infty}\frac{\mu^i}{i!}}_{=e^{\mu}} = \mu^2. \end{aligned}$$

Vi får sedan

$$\underline{V(Y)} = E(Y^2) - (E(Y))^2 = E(Y(Y-1)) + E(Y) - (E(Y))^2 = \mu^2 + \mu - \mu^2 = \underline{\mu}.$$

Sats. *Låt $X_1 \in Po(\mu_1)$ och $X_2 \in Po(\mu_2)$ vara oberoende s.v. Då gäller $\underline{X_1 + X_2 \in Po(\mu_1 + \mu_2)}$.*

Bevis. Vi utvecklar sannolikhetsfunktionen för $X_1 + X_2$.

$$\begin{aligned} p_{X_1+X_2}(k) &= \mathbb{P}(X_1 + X_2 = k) = \sum_{i=0}^{k}\mathbb{P}(X_1 = i)\mathbb{P}(X_2 = k-i) \\ &= \sum_{i=0}^{k}\frac{\mu_1^i}{i!}e^{-\mu_1}\frac{\mu_2^{k-i}}{(k-i)!}e^{-\mu_2} = e^{-(\mu_1+\mu_2)}\frac{1}{k!}\underbrace{\sum_{i=0}^{k}\frac{k!}{(k-i)!i!}\mu_1^i\mu_2^{k-i}}_{=(\mu_1+\mu_2)^k} \\ &= \frac{(\mu_1+\mu_2)^k}{k!}e^{-(\mu_1+\mu_2)}. \end{aligned}$$

□

Additionssatsen implicerar att om μ är ett heltal, så kan $X \in \text{Po}(\mu)$ skrivas som

$$X = Z_1 + ... + Z_\mu, \qquad \text{där } Z_i \text{ är oberoende Po(1)-fördelade s.v.}$$

Eftersom $E(Z_i) = V(Z_i) = 1$, så följer det av centrala gränsvärdessatsen att X är approximativt $\text{N}(\mu, \sqrt{\mu})$-fördelad för stora μ. Approximationen fungerar även när μ inte är ett heltal och är god för $\mu \geq 15$. Precis som vid normalapproximation av binomialfördelade s.v. så förbättras approximationen med halvkorrektion.

Hypergeometriska fördelningen

Vi avslutar med att undersöka den **hypergeometriska fördelningen** och återvänder därför till urnmodellerna som vi studerade på föreläsning 2. Antag att en urna innehåller s svarta och v vita kulor. Låt X beteckna antalet vita kulor som erhålls vid dragning av n kulor. Vid

dragning med återläggning, så gäller $X \in \text{Bin}(n, v/(s+v))$. Vid dragning utan återläggning vet vi sedan tidigare att

$$\mathbb{P}(X = k) = \binom{v}{k}\binom{s}{n-k}\Big/\binom{s+v}{n}.$$

Om vi nu betecknar totala antalet kulor $s+v$ med N och proportionen av vita kulor $v/(s+v)$ med p, så får vi, för $0 \leq k \leq Np$ och $0 \leq n-k \leq N(1-p)$, att

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

Vi säger då att X är en **hypergeometrisk** s.v. Detta betecknas $X \in \text{Hyp}(N, n, p)$. Precis som för binomialfördelade s.v., så kan vi skriva $X = I_1 \cdots + I_n$, där $I_i \in \text{Be}(p)$ för $i = 1, ..., n$, eftersom vid var och en av de n dragningarna så är sannolikheten att få en vit kula p. I motsats till det binomialfördelade fallet, så är I_i:na här beroende, eftersom dragningen sker utan återläggning. Vad vi drar i omgång j påverkas av vad vi har dragit i omgångarna $1, 2, ..., j-1$. Kovariansen är alltså nollskild och i boken visas att för $X \in \text{Hyp}(N, n, p)$, så gäller

$$\underline{E(X) = np} \qquad \text{och} \qquad \underline{V(X) = \frac{N-n}{N-1}np(1-p)}.$$

Termen $(N-n)/(N-1)$ kallas korrektionsfaktor för ändliga populationer. Om $n \ll N$, så är korrektionsfaktorn nära ett och återläggningen spelar inte så stor roll för resultatet av dragningen. Mer specifikt gäller att om $n/N \leq 0.1$, så är $X \in \text{Hyp}(N, n, p)$ approximativt $\text{Bin}(n, p)$-fördelad. Om dessutom $np(1-p) \geq 10$, så är $X \in \text{Hyp}(N, n, p)$ approximativt $\text{N}(np, \sqrt{\frac{N-n}{N-1}np(1-p)})$-fördelad. Notera att centrala gränsvärdessatsen här gäller trots att de s.v. I_i inte är oberoende.