Föreläsning 7

Thomas Önskog

15/11 2017

Normalfördelningen

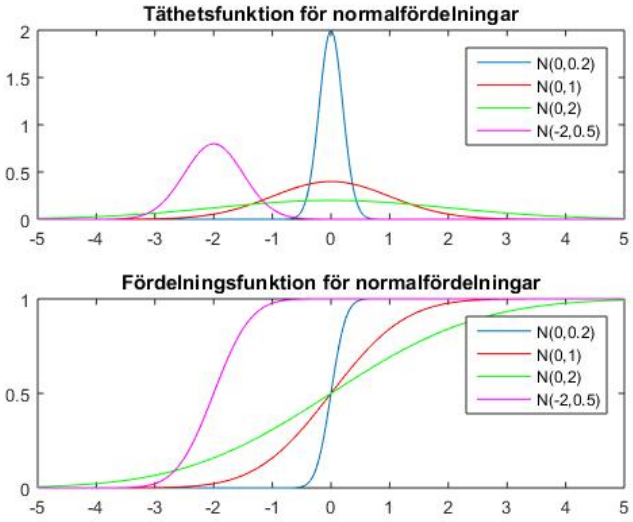
På förra föreläsningen visade vi **stora talens lag** som säger att det aritmetiska medelvärdet av n oberoende s.v. med samma väntevärde och samma varians konvergerar mot väntevärdet av de s.v. Målet med dagens föreläsning är **centrala gränsvärdessatsen** som säger att fördelningen för det aritmetiska medelvärdet kommer att bli alltmer lik en normalfördelning när n ökar. Vi börjar med att studera normalfördelningens egenskaper i detalj och vet sedan tidigare att täthetsfunktionen ges av

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}, \qquad \text{där } \mu \in \mathbb{R} \text{ och } \sigma > 0 \text{ är parametrar.}$$

Vi betecknar detta $X \in N(\mu, \sigma)$. Fördelningsfunktionen är

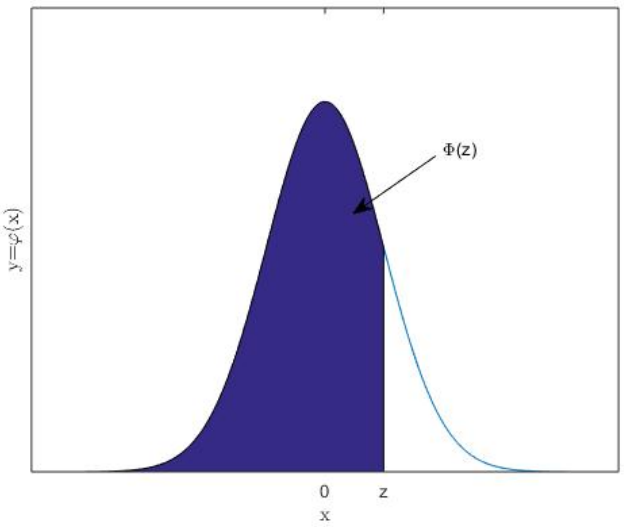
$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2} dy,$$

men denna integral kan inte förenklas ytterligare. Vi har tidigare noterat att μ är ett lägesmått och σ är ett spridningsmått för normalfördelningen.



Definition. *En s.v. Z sägs vara* ***standardiserat normalfördelad*** *om $Z \in N(0,1)$. Vi inför egna beteckningar för täthets- och fördelningsfunktionen för standardiserat normalfördelade s.v.*

$$\underline{\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}} \qquad \textit{respektive} \qquad \underline{\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy.}$$



Att täthetsfunktionen φ uppfyller $\int_{-\infty}^{\infty} \varphi(z)dz = 1$ visas i boken. Senare under denna föreläsning visar vi att med hjälp av tabellvärden för Φ kan vi beräkna fördelningsfunktionens värden för alla normalfördelade s.v. Vi härleder nu några egenskaper hos φ och Φ.

$$\begin{aligned}
\underline{\varphi(z)} &= \underline{\varphi(-z)} \qquad \text{(följer direkt från uttrycket för } \varphi\text{)}. \\
\underline{\Phi(-z)} &= \mathbb{P}(Z \le -z) = \int_{-\infty}^{-z} \varphi(y)dy = [x = -y] = -\int_{\infty}^{z} \varphi(-x)dx = \int_{z}^{\infty} \varphi(x)dx \\
&= \mathbb{P}(Z > z) = \underline{1 - \Phi(z)}.
\end{aligned}$$

Ovanstående samband förklarar varför vi bara behöver tabeller med positiva värden på z (se Tabell 1 i kursens tabellsamling). Vidare gäller

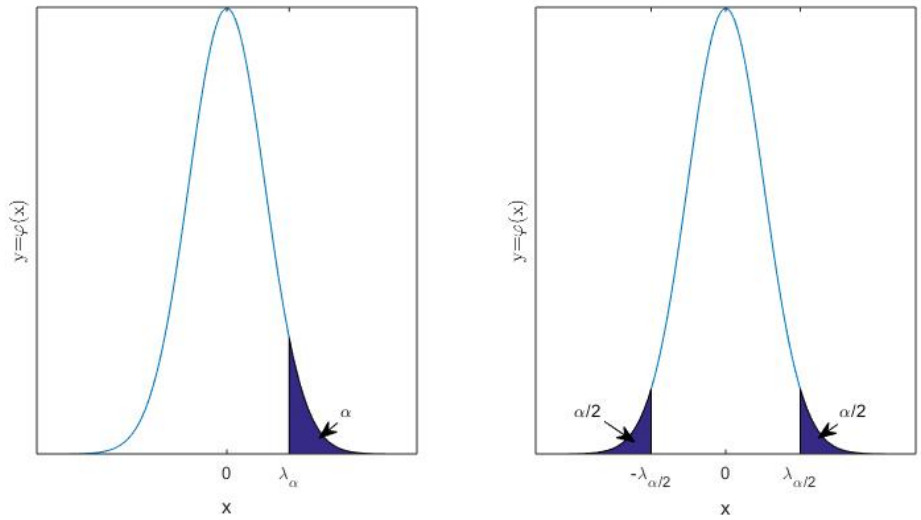
$$\begin{aligned}
\underline{E(Z)} &= \underline{0} \qquad \text{(följer av symmetriskäl eftersom } \varphi(z) = \varphi(-z)\text{)}. \\
\underline{V(Z)} &= E(Z^2) - \underbrace{(E(Z))^2}_{=0} = \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \text{[Partiell integration]} \\
&= \underbrace{\Big[z\Big(-\frac{1}{\sqrt{2\pi}} e^{-z^2/2}\Big)\Big]_{-\infty}^{\infty}}_{=0} - \int_{-\infty}^{\infty} -\frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \underline{1}.
\end{aligned}$$

Inom statistikteorin behöver vi ofta bestämma intervall sådana att en s.v. $Z \in \mathrm{N}(0,1)$ ligger i intervallet med en viss förutbestämd sannolikhet. Intervallen i fråga kan vara symmetriska (på formen $(-a,a)$) eller ensidiga (på formen $(-\infty,b)$ eller $(-b,\infty)$). Sådana intervall kan härledas både med hjälp av fördelningsfunktionen och med hjälp av kvantiler. Kvantiler för standardiserat normalfördelade s.v. används så ofta att de har fått en egen notation.

Definition. *α-**kvantilen** för en standardiserat normalfördelad s.v. $Z \in N(0,1)$ betecknas λ_α och uppfyller*

$$\underline{\mathbb{P}(Z > \lambda_\alpha) = \alpha},$$

eller, ekvivalent, $\Phi(\lambda_\alpha) = 1-\alpha$. Notera att $\lambda_{1-\alpha} = -\lambda_\alpha$.



Exempel. *Bestäm talet $a > 0$ för vilket $\mathbb{P}(-a \le Z \le a) = 0.95$. Av symmetriskäl följer att $\mathbb{P}(Z > a) = (1-0.95)/2 = 0.025$, så $\Phi(a) = \mathbb{P}(Z \le a) = 0.975$. Från tabellen för Φ utläses att $a = 1.96$. Alternativt, så kan vi använda tabellvärden för λ_α och iakttagelsen att $\mathbb{P}(Z > a) = 0.025$ för att dra slutsatsen att $a = \lambda_{0.025} = 1.96$.*

Exempel. *Bestäm talet $b > 0$ för vilket $\mathbb{P}(-b \le Z) = 0.99$. Av symmetriskäl gäller $\mathbb{P}(Z \le b) = 0.99$, dvs $\Phi(b) = 0.99$. Från tabellen för Φ utläses att $b = 2.33$. Vi kan också notera att $\mathbb{P}(Z > b) = 0.01$, så $b = \lambda_{0.01} = 2.33$.*

Vi kan definiera $\mathrm{N}(\mu,\sigma)$-fördelade s.v. med hjälp av standardiserat normalfördelade s.v. genom att säga att $X \in \mathrm{N}(\mu,\sigma)$ för $\mu \in \mathbb{R}$ och $\sigma > 0$ om

$$\underline{Z = \frac{X-\mu}{\sigma} \text{ är } \mathrm{N}(0,1)\text{-fördelad.}}$$

Det gäller då att $X = \mu + \sigma Z$. Eftersom $E(Z) = 0$ och $V(Z) = 1$, så följer det direkt att

$$\underline{E(X)} = E(\mu+\sigma Z) = \mu + \sigma E(Z) = \underline{\mu} \qquad \text{och} \qquad \underline{V(X)} = V(\mu+\sigma Z) = \sigma^2 V(Z) = \underline{\sigma^2}.$$

Vidare gäller

$$F_X(x) = \mathbb{P}(X \le x) = \mathbb{P}(\mu+\sigma Z \le x) = \mathbb{P}\Big(Z \le \frac{x-\mu}{\sigma}\Big) = \Phi\Big(\frac{x-\mu}{\sigma}\Big).$$

Derivering ger sedan direkt

$$f_X(x) = \frac{1}{\sigma}\varphi\Big(\frac{x-\mu}{\sigma}\Big),$$

vilket överensstämmer med formlerna för F_X och f_X som vi tidigare tagit upp. Vi kan alltså använda tabellvärdena för Φ för att bestämma fördelningsfunktionen för $X \in \mathrm{N}(\mu, \sigma)$.

Exempel. *Ett företag tillverkar tvåtonsbalkar. Vikten X hos en enskild balk är $N(2000, 2)$-fördelad. Sannolikheten att en balk väger över* 2003 *kg är*

$$\begin{aligned}\mathbb{P}(X > 2003) &= 1 - \mathbb{P}(X \le 2003) = 1 - F_X(2003) = 1 - \Phi\Big(\frac{2003-2000}{2}\Big)\\ &= 1 - \Phi(1.5) = 6.7\%.\end{aligned}$$

Svaret fås på en TI-miniräknare som 1-normcdf(-1E99,2003,2000,2).
Vi vill härnäst bestämma den vikt som bara en procent av balkarna överskrider, dvs $x_{0.01}$. Per definition, så gäller

$$0.01 = \mathbb{P}(X > x_{0.01}) = \mathbb{P}\Big(\underbrace{\frac{X-\mu}{\sigma}}_{\in N(0,1)} > \frac{x_{0.01}-\mu}{\sigma}\Big)$$

vilket implicerar

$$\frac{x_{0.01}-\mu}{\sigma} = \lambda_{0.01} \qquad \Rightarrow \qquad x_{0.01} = \mu + \sigma\lambda_{0.01} = 2004.7 \text{ kg}.$$

Centrala gränsvärdessatsen

Vi noterar först att linjärkombinationer av normalfördelade s.v. är normalfördelade.

Sats. *Låt $a_0, a_1, ..., a_n$ vara reella konstanter och låt $X_1, ..., X_n$ vara oberoende $N(\mu_i, \sigma_i)$-fördelade s.v. Då gäller*

$$\underline{a_0 + \sum_{i=1}^{n} a_i X_i \in N\Big(a_0 + \sum_{i=1}^{n} a_i\mu_i, \sqrt{\sum_{i=1}^{n} a_i^2\sigma_i^2}\Big).}$$

Satsens utsaga om väntevärdet och standardavvikelsen följer direkt från satser på förra föreläsningen. Det nya i ovanstående sats är att vi i fallet med normalfördelade s.v. X_i också vet fördelningen för linjärkombinationen. Från satsen följer direkt att om alla X_i är $\mathrm{N}(\mu, \sigma)$-fördelade (dvs har samma väntevärde och samma varians), så är $\overline{X} \in \mathrm{N}(\mu, \sigma/n)$ (inses genom att sätta $a_0 = 0$ och $a_i = 1/n$, för $i = 1, ..., n$, i satsen ovan). **Centrala gränsvärdessatsen**, sannolikhetsteorins viktigaste resultat, säger att även om de ingående s.v. X_i inte i sig är normalfördelade, så kommer $\overline{X}$ att vara approximativt normalfördelade för stora n.

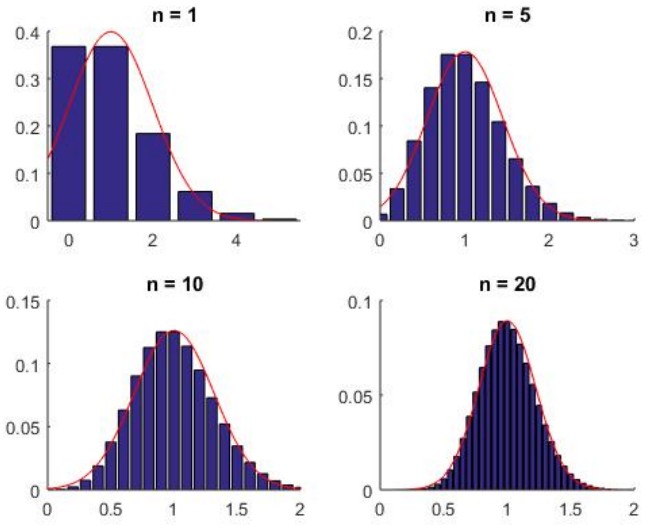
Sats. *Låt $X_1, X_2, ...$ vara en oändlig följd av (nästan) oberoende och (nästan) likafördelade s.v. med väntevärde μ och standardavvikelse σ. Sätt $Y_n = X_1 + \cdots + X_n$. Då gäller, för alla $a < b$, att*

$$\underline{\mathbb{P}\Big(a < \frac{Y_n - n\mu}{\sqrt{n}\sigma} \le b\Big) \to \Phi(b) - \Phi(a), \qquad \text{då } n \to \infty.}$$

För stora n är alltså $(\sum_{i=1}^{n} X_i - n\mu)/\sqrt{n}\sigma$ approximativt $\mathrm{N}(0,1)$-fördelat och $\sum_{i=1}^{n} X_i$ är approximativt $\mathrm{N}(n\mu, \sqrt{n}\sigma)$-fördelat. Vidare är $\overline{X}$ approximativt $\mathrm{N}(\mu, \sigma/\sqrt{n})$-fördelat. För tillräckligt stora n gäller således

$$\mathbb{P}(a < \overline{X} \leq b) = F_{\overline{X}}(b) - F_{\overline{X}}(a) \approx \Phi\Big(\frac{b-\mu}{\sigma/\sqrt{n}}\Big) - \Phi\Big(\frac{a-\mu}{\sigma/\sqrt{n}}\Big).$$

Vi säger att $\overline{X}$ är **asymptotiskt normalfördelat** med väntevärde μ och standardavvikelse $\sigma/\sqrt{n}$. Detta betecknas $\overline{X} \in \mathrm{AsN}(\mu, \sigma/\sqrt{n})$.



Figuren visar fördelningen för medelvärdet av 1, 5, 10 och 20 Po(1)-fördelade s.v. Som synes blir medelvärdet alltmer normalfördelat. Det finns ingen generell regel för hur stort n måste vara för att $\overline{X}$ ska kunna antas vara normalfördelat. Om X_i är normalfördelade, så fungerar alla n, men om fördelningarna för X_i är skeva kan n behöva vara i storleksordningen tiotal, eller i värsta fall hundratal (se laboration 2).