TAMS79: Föreläsning 6
Normalfördelningen

Johan Thim (johan.thim@liu.se)

10 november 2018

Normalfördelning

Definition. Låt $\mu \in \mathbf{R}$ och $\sigma > 0$. Om X är en stokastisk variabel med täthetsfunktion

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbf{R},$$

så kallar vi X för normalfördelad med parametrarna μ och σ. Vi skriver $X \sim \mathrm{N}(\mu, \sigma)$.

Om $\mu = 0$ och $\sigma = 1$ kallar vi X för **standardiserad**, och i det fallet betecknar vi täthetsfunktionen med

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \quad x \in \mathbf{R}.$$

Fördelningsfunktionen för en normalfördelad variabel ges av

$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right) du, \quad x \in \mathbf{R},$$

och även här döper vi speciellt den standardiserade fördelningsfunktionen till

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{u^2}{2}\right) du, \quad x \in \mathbf{R}.$$

Är det någon som kommer ihåg vad som hände när man försökte integrera e^{x^2} i envariabelanalysen? Man kan visa att det inte går att uttrycka den primitiva funktionen i elementära funktioner, utan man definierar helt enkelt en *ny* funktion utifrån den bestämda integralen (slå upp *erf*-funktionen i något matematiskt uppslagsverk). Vad detta innebär för oss är att vi kommer att använda tabell för att beräkna numeriska värden för uttryck som innehåller funktionen Φ. Är då detta en täthetsfunktion? Beviset är så roligt att jag inte kan låta bli. Vi visar att

$$\int_{-\infty}^{\infty} e^{-x^2/2}\, dx = \sqrt{2\pi}.$$

Sättet vi kommer åt det hela är genom att titta på kvadraten och tolka denna som en itererad dubbelintegral:

$$\left(\int_{-\infty}^{\infty} e^{-x^2/2}\, dx\right)^2 = \left(\int_{-\infty}^{\infty} e^{-x^2/2}\, dx\right)\left(\int_{-\infty}^{\infty} e^{-y^2/2}\, dy\right) = \iint_{\mathbf{R}^2} e^{-(x^2+y^2)/2}\, dx\, dy$$
$$= \int_0^{\infty}\int_0^{2\pi} re^{-r^2/2}\, d\theta\, dr = 2\pi\left[e^{-r^2/2}\right]_0^{\infty} = 2\pi.$$

Standardisering av variabel

Om X är en stokastisk variabel med $E(X) = \mu$ och $V(X) = \sigma^2$, så är $Z = (X - \mu)/\sigma$ en stokastisk variabel med $E(Z) = 0$ och $V(Z) = 1$. Vi kallar Z för **standardiserad**.

Beviset för att $E(Z) = 0$ följer direkt från linjäriteten hos väntevärdet. Variansen kan ses genom följande argument:

$$V((X-\mu)/\sigma) = E((X-\mu)^2/\sigma^2) - 0^2 = \frac{1}{\sigma^2}\left(E(X^2) - 2\mu E(X) + \mu^2\right)$$
$$= \frac{1}{\sigma^2}\left(E(X^2) - E(X)^2\right) = \frac{\sigma^2}{\sigma^2} = 1.$$

Standardiserad normalfördelning

Sats. Om $X \sim \mathrm{N}(0,1)$ så är $E(X) = 0$ och $V(X) = 1$.

Eftersom $e^{-x^2/2}$ går mot noll väldigt snabbt, så kommer

$$\int_{-\infty}^{\infty} |x\varphi(x)|\, dx < \infty.$$

Alltså måste

$$E(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x\varphi(x)\, dx = 0$$

eftersom $x \mapsto x\varphi(x)$ är en udda funktion. På liknande sätt ser vi att

$$E(X^2) = \int_{-\infty}^{\infty} x^2\varphi(x)\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x\left(xe^{-x^2/2}\right) dx$$
$$= \frac{1}{\sqrt{2\pi}} \left(\left[x(-e^{-x^2/2})\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-x^2/2}\, dx \right) = 0 + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\, dx = 1,$$

där vi partialintegrerat och utnyttjat att $\varphi(x)$ är en täthetsfunktion så den sista integralen blir ett.

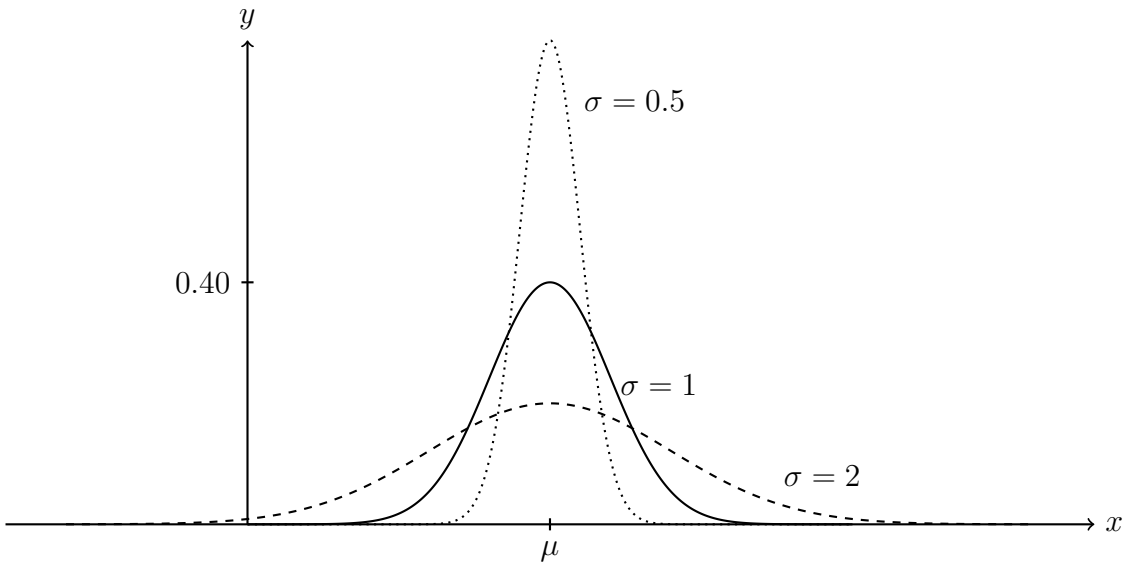
Ett korrolarie av satsen ovan (gör ett variabelbyte i integralerna) är följande.

$X \sim \mathbf{N}(\mu, \sigma)$

Om $X \sim \mathrm{N}(\mu, \sigma)$ så är $E(X) = \mu$ och $V(X) = \sigma^2$.

Standardavvikelse eller varians?

I kursboken (Blom et al) används beteckningen $X \sim N(\mu, \sigma)$, så precis som i vårt fall ovan är den andra parametern är alltså standardavvikelsen σ, inte variansen σ^2. Varför ta upp detta? I mycket av litteraturen så används variansen som andra parameter. Var försiktig när ni slår upp saker eller använder färdiga formler!



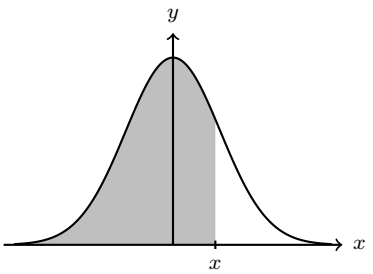
Bruk av tabell för $\Phi(x)$

Låt $X \sim \mathrm{N}(0,1)$. Då gäller

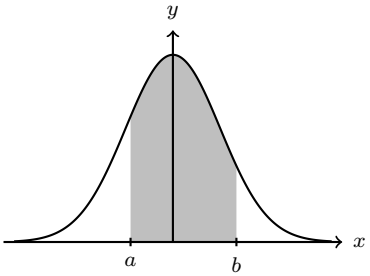
(i) $P(X \leq x) = \Phi(x)$ för alla $x \in \mathbf{R}$;

(ii) $P(a \leq X \leq b) = \Phi(b) - \Phi(a)$ för alla $a, b \in \mathbf{R}$ med $a \leq b$;

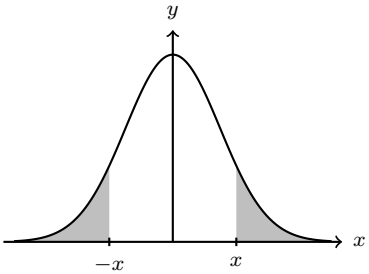
(iii) $\Phi(-x) = 1 - \Phi(x)$ för alla $x \in \mathbf{R}$.



$\Phi(x)$



$\Phi(b) - \Phi(a)$



$\Phi(-x) = 1 - \Phi(x)$

Exempel

Låt $X \sim \mathrm{N}(0,1)$. Bestäm $P(X \leq 1)$, $P(X < 1)$, $P(X \leq -1)$, samt $P(0 < X \leq 1)$.

Direkt ur tabell, $P(X \leq 1) = \Phi(1) \approx 0.8413$. Eftersom X är kontinuerlig kvittar det om olikheterna är strikta eller inte, så $P(X < 1) = P(X \leq 1) = \Phi(1)$ igen. Vidare har vi

$$P(X \leq -1) = \Phi(-1) = 1 - \Phi(1) = 0.1587$$

och $P(0 < X \leq 1) = \Phi(1) - \Phi(0) = 0.8413 - 0.5 = 0.3413$.

Standardisering av normalfördelning

Sats. $X \sim \mathrm{N}(\mu, \sigma) \quad \Leftrightarrow \quad Z = \dfrac{X-\mu}{\sigma} \sim \mathrm{N}(0,1).$

Bevis. Antag att $Z \sim N(0,1)$. Eftersom

$$F_X(x) = P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = P\left(Z \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

och $\Phi'(x) = \varphi(x)$ då φ är kontinuerlig, så följer det att

$$f_X(x) = F_X'(x) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbf{R},$$

vilket är precis hur vi definierat $N(\mu, \sigma)$ tidigare.
Omvänt, om $X \sim N(\mu, \sigma)$ så är

$$F_Z(z) = P(Z \le z) = P(\mu + \sigma Z < \mu + \sigma z) = F_X(\mu + \sigma z),$$

så

$$f_Z(z) = f_X(\mu + \sigma z)\sigma = \varphi(z),$$

dvs $Z \sim N(0,1)$. □

Exempel

Låt $X \sim \mathrm{N}(1,2)$. Bestäm $P(X \le 1)$, $P(X \le -1)$, $P(0 < X \le 1)$, samt $P(|X-2| < 3)$.

(i) $P(X \le 1) = P\left(\dfrac{X-1}{2} \le \dfrac{1-1}{2}\right) = \Phi(0) = \dfrac{1}{2}.$

(ii) $P(X \le -1) = P\left(\dfrac{X-1}{2} \le \dfrac{-1-1}{2}\right) = \Phi(-1) = 1 - \Phi(1) \approx 0.1587.$

(iii)

$$\begin{aligned} P(0 < X \le 1) &= P\left(\frac{0-1}{2} < \frac{X-1}{2} \le \frac{1-1}{2}\right) = \Phi(0) - \Phi(-1/2) \\ &= 1/2 - (1 - \Phi(1/2)) = -1/2 + \Phi(1/2) \approx 0.1915. \end{aligned}$$

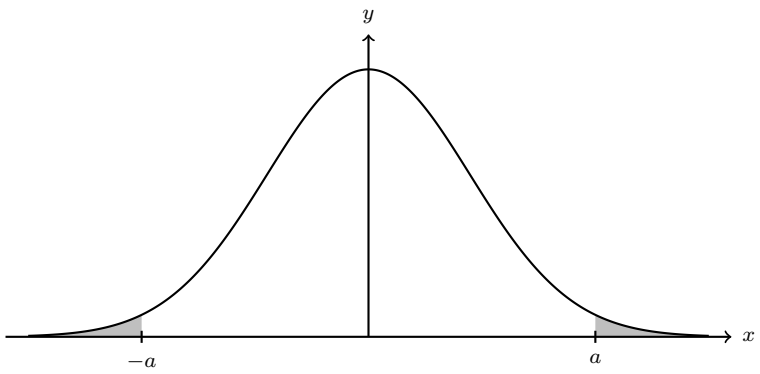
(iv)

$$\begin{aligned} P(|X-2| < 3) &= P\left(-3 < X - 2 < 3\right) = P\left(\frac{-1-1}{2} < \frac{X-1}{2} \le \frac{5-1}{2}\right) \\ &= \Phi(2) - \Phi(-1) = \Phi(2) - 1 + \Phi(1) \approx 0.8186. \end{aligned}$$

Exempel

Låt $X \sim \mathrm{N}(0,1)$. Hitta ett tal a så att $P(|X| > a) = 0.05$.

Situationen ser ut som i bilden nedan. De skuggade områdena utgör tillsammans 5% av sannolikhetsmassan, och på grund av symmetri måste det vara 2.5% i varje ”svans”.



Om vi söker talet a, och vill använda funktionen $\Phi(x) = P(X \leq x)$, måste vi söka det tal som ger $\Phi(a) = 0.975$ (dvs de 2.5% i vänstra svansen tillsammans med de 95% som ligger i den stora kroppen). Detta gör vi genom att helt enkelt leta efter talet 0.975 i tabellen över $\Phi(x)$ värden. Där finner vi att $a = 1.96$ uppfyller kravet att $P(X \leq a) = 0.975$.

6.1 Linjärkombinationer av normalfördelade variabler

Det är inte på något sätt uppenbart att summan av två likafördelade variabler har samma fördelning. Oftast är det inte ens sant. Men, just normalfördelningen har precis denna trevliga egenskap!

Summa av normalfördelade variabler

Sats. Låt $X_1 \sim \mathrm{N}(\mu_1, \sigma_1)$ och $X_2 \sim \mathrm{N}(\mu_2, \sigma_2)$ vara oberoende och låt $a, b \in \mathbf{R}$. Då gäller att

$$aX_1 + bX_2 \sim \mathrm{N}(a\mu_1 + b\mu_2, \sqrt{a^2\sigma_1^2 + b^2\sigma_2^2}).$$

Beviset är inte trivialt. Att $E(aX_1 + bX_2) = a\mu_1 + b\mu_2$ och $V(aX_2 + bX_2) = a^2\sigma_1^2 + b^2\sigma_2^2$ är enkelt att se. Detta gäller oavsett vad variablerna har för slags fördelning (enbart egenskaper för väntevärde och varians). Det faktum att summan blir normalfördelad kräver ett djupare argument; se boken (man använder faltningssatsen).
Satsen ovan generaliserar direkt till flera variabler. Vi har följande användbara specialfall.

Summor och medelvärde

Sats. Låt $X_1, X_2, \ldots, X_n$ vara oberoende och $X_k \sim \mathrm{N}(\mu, \sigma)$ för $k = 1, 2, \ldots, n$. Då gäller följande:

$$X := \sum_{k=1}^{n} X_k \sim \mathrm{N}(n\mu, \sigma\sqrt{n}) \qquad \text{och} \qquad \overline{X} := \frac{1}{n}\sum_{k=1}^{n} X_k \sim \mathrm{N}(\mu, \sigma/\sqrt{n}).$$

Den sista likheten är intressant, då det innebär att ju fler ”likadana” variabler vi tar med i ett medelvärde, desto *mindre* blir variansen. Till exempel får vi alltså säkrare resultat ju fler mätningar vi gör (något som känns intuitivt korrekt). Det är dock mycket viktigt att variablerna är *oberoende*. Annars gäller inte satsen! Vi bildar aldrig heller några skillnader mellan varianser, utan det som gör att variansen minskar med antalet termer är faktorn $1/n$ i medelvärdet:

$$V(\overline{X}) = V\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) = \frac{1}{n^2}V\left(\sum_{k=1}^{n} X_k\right) = \frac{1}{n^2}\sum_{k=1}^{n} V(X_k) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n},$$

eftersom variablerna är oberoende och $V(X_k) = \sigma^2$ för alla k.

Exempel

Låt $X \sim \mathrm{N}(10, 3)$ och $Y \sim \mathrm{N}(21, 6)$ vara oberoende. Vad är sannolikheten att Y är mer än dubbelt så stor som X?

Lösning: Låt $W = Y - 2X \sim \mathrm{N}(21 - 20, \sqrt{6^2 + 4 \cdot 3^2}) = \mathrm{N}(1, \sqrt{72})$. Vi söker alltså

$$\begin{aligned} P(Y > 2X) &= P(Y - 2X > 0) = P(W > 0) = 1 - P(W \leq 0) = 1 - \Phi\left(\frac{0-1}{\sqrt{72}}\right) \\ &= 1 - (1 - \Phi(0.12)) = 0.5478. \end{aligned}$$

Exempel

Låt T vara livslängden för en viss sorts lysrör (enhet: månader) och $T \sim \mathrm{N}(20, \sqrt{8})$.

(i) Vad är sannolikheten att man klarar 43 månader om man har två (oberoende) lysrör och byter direkt det första går sönder?

(ii) Hur många lysrör måste man skaffa för att medellivslängden ska vara mer än 19 månader med sannolikhet 95%?

Lösning:

(i) Vi har två lysrör, T_1 och T_2. Vi söker sannolikheten att $T_1 + T_2 \geq 43$. Satsen ovan visar att $T_1 + T_2 \sim \mathrm{N}(40, 4)$, så

$$\begin{aligned} P(T_1 + T_2 \geq 43) &= 1 - P(T_1 + T_2 < 43) = 1 - P\left(\frac{T_1 + T_2 - 40}{\sqrt{16}} < \frac{43 - 40}{\sqrt{16}}\right) \\ &= 1 - \Phi(3/\sqrt{16}) \approx 1 - \Phi(0.75) \approx 0.2266. \end{aligned}$$

(ii) Låt $\overline{T}$ vara medelvärdet av n stycken lysrör. Det följer att $\overline{T} \sim \mathrm{N}(20, \sqrt{8}/\sqrt{n})$. Vi vill att

$$\begin{aligned} 0.95 = P(\overline{T} > 19) &= P\left(\frac{\overline{T} - 20}{\sqrt{8/n}} > \frac{19 - 20}{\sqrt{8/n}}\right) = 1 - P(Z \leq -1/\sqrt{8/n}) \\ &= 1 - \Phi(-1/\sqrt{8/n}) = 1 - \Big(1 - \Phi(1/\sqrt{8/n})\Big) = \Phi(\sqrt{n/8}), \end{aligned}$$

där $Z = \dfrac{\overline{T} - 20}{\sqrt{8/n}} \sim \mathrm{N}(0, 1)$. Ur tabell finner vi då att $\sqrt{n/8} = 1.645$, eller ekvivalent, att $n \approx 21.6$. Således behövs åtminstone 22 stycken lysrör.