TAMS79: Föreläsning 3
Kontinuerliga stokastiska variabler

Johan Thim (johan.thim@liu.se)

10 november 2018

Vi kommer nu att utveckla teori för kontinuerliga stokastiska variabler som motsvarar den vi tog fram i det diskreta fallet förra gången. Åtminstone i de fall där det finns en så kallad täthetsfunktion. Så vi börjar med det.

3.1 Kontinuerliga stokastiska variabler

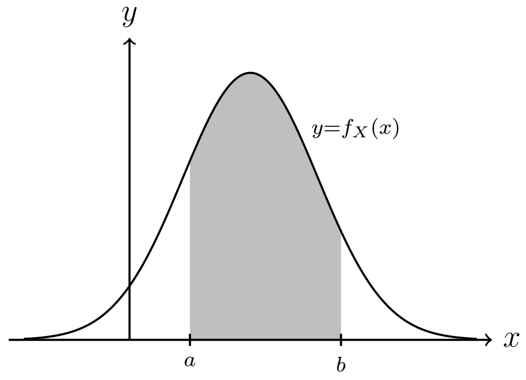
Täthetsfunktion

Definition. Om det finns en icke-negativ *integrerbar* funktion f_X så att

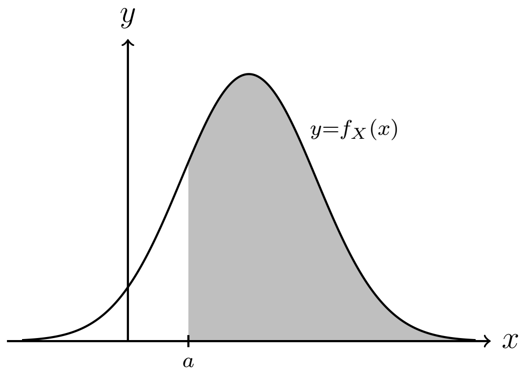
$$P(a < X < b) = \int_a^b f_X(x)\,dx$$

för alla intervall $(a, b) \subset \mathbf{R}$, kallar vi f_X för variabelns **täthetsfunktion**.

Exempel:



Skuggad area: $P(a \leq X \leq b)$.



Skuggad area: $P(X > a) = \int_a^\infty f_X(x)\,dx$.

Egenskaper hos täthetsfunktionen

(i) $f_X(x) \geq 0$ för alla $x \in \mathbf{R}$.

(ii) $\displaystyle\int_{-\infty}^{\infty} f_X(x)\,dx = 1$.

(iii) $f_X(x)$ anger hur mycket sannolikhetsmassa det finns per längdenhet i punkten x.

Definitionen kanske ser oskyldig ut, men här finns det både hundar och ugglor begravda i mossen (som säkert ligger i Danmark). Problemet ligger i integralbegreppet och hur generella händelser vi vill tillåta. I grundanalysen introducerar man Riemann-integralen, men tyvärr räcker den inte riktigt till för allt. Betrakta följande funktion: $f(x) = 0$ om $x < 0$, $x > 1$, eller om x är rationell (dvs ett bråk p/q av heltal p och q). I övriga punkter är $f(x) = 1$ (dvs på alla irrationella punkter i intervallet $[0, 1]$).

Man kan tänka sig den stokastiska variabeln X som indikerar om ett slumptal mellan noll och ett är irrationellt eller inte. Kanske skulle täthetsfunktionen då ges av $f(x)$ ovan, men är detta verkligen en täthetsfunktion? Den är icke-negativ, så den biten är OK. Men har den "area" ett??

Av nödvändighet kommer alla undertrappor till $f(x)$ på $[0, 1]$ att vara identiskt lika med noll, och på samma sätt är alla övertrappor identiskt lika med ett. Vi kan alltså aldrig approximera funktionen med över- och undertrappor. Således är $f(x)$ *inte* Riemann-integrerbar. Så hur löser man detta? Med ett nytt integralbegrepp (Lebesgueintegralen) smidigt nog, där det visar sig att integralen av f mycket riktigt blir ett.

Lebesgueintegralen konstrueras på ett annorlunda sätt i jämförelse med Riemannintegralen. Istället för att bara stycka upp definitionsmängden (dvs x-axeln) i finare och finare likadana bitar och försöka approximera integralen med arean av rektanglar (över- och undertrappor), så styckar vi istället upp värdemängden. Genom att approximera funktionen med så kallade *enkla funktioner* – funktioner som är konstant på ett ändligt antal mätbara mängder och lika med noll annars – så kan man komma åt betydligt fler funktioner. Mätbarheten här blir i en sådan här kurs med avseende på det sannolikhetsmått man är intresserad av, så sannolikheten kommer in på ett väldigt naturligt sätt. Detta ligger dock utanför ramarna för denna kurs. Men termen *integrerbar* i definitionen syftar på denna "nya" typ av integral.

För *snälla* funktioner (funktioner som till exempel bara har uppräkneligt många diskontinuiteter) så sammanfaller de båda integralbegreppen. Vi kommer alltså inte att fundera så mycket mer på detta.

Strikt olikhet eller inte?

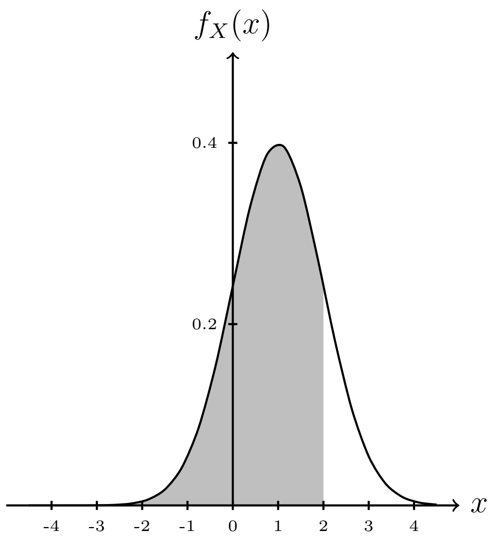
Om X är en kontinuerlig variabel, så är $P(X < x) = P(X \le x)$. Detta följer från att integralen inte gör någon skillnad på om ändpunkten är med eller ej. Vi kan till och med definiera om funktionen i uppräkneligt många punkter (även mer, men det kräver lite måttteori för att definiera) utan att ändra sannolikheten. Detta gäller dock *absolut inte* i det diskreta fallet.

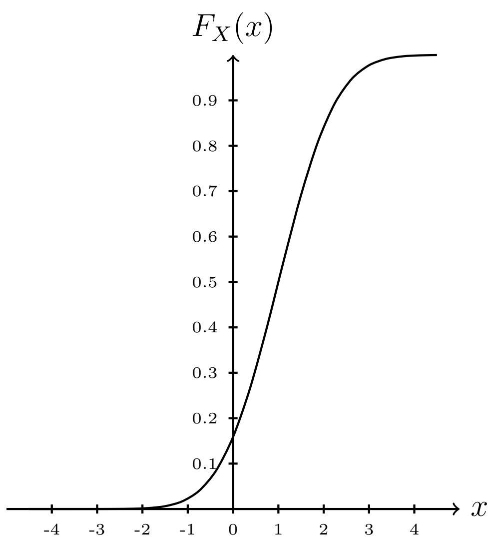
Vi definierar **fördelningsfunktionen** $F_X(x)$ på samma sätt som i det diskreta fallet, och finner att

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} f_X(t)\,dt, \qquad x \in \mathbf{R}.$$

Fördelningsfunktionen uppfyller (i)–(iii) från det diskreta fallet, och i alla punkter där $f_X(x)$ är kontinuerlig gäller dessutom att $F_X'(x) = f_X(x)$. Det sista är i princip analysens huvudsats. Man kan fundera över hur pass diskontinuerlig f_X skulle kunna vara, men som exemplet ovan visar finns det inte så mycket begränsningar på det. I denna kurs kommer dock de flesta kontinuerliga fördelningar ha täthetsfunktioner som är kontinuerliga för det mesta.

Exempel på hur en täthetsfunktion och motsvarande fördelningsfunktion kan se ut:





Täthet: Hur ”sannolikhetsmassan” är fördelad. Skuggad area är $P(X \leq 2) = F_X(2)$.

Fördelningsfunktionen är växande och gränsvärderna mot $\pm\infty$ verkar stämma!

Exempel

Låt $f_1(x) = x^2 + bx$ och $f_2(x) = c(x^3 + x)$ båda för $x \in [0, 2]$. Om det går, bestäm konstanterna b och c så dessa blir täthetsfunktioner och beräkna sannolikheten att respektive variabel är ≤ 1.

Lösning: Vi börjar med f_1:

$$1 = \int_0^2 \left(x^2 + bx\right) dx = \frac{8}{3} + 2b \quad \Rightarrow \quad b = -5/6.$$

Men om b är negativ kommer $f_1(x)$ att vara negativ för x nära noll (x^2 termen går mot noll snabbare än x). Detta kan alltså *inte* vara en täthetsfunktion. Vi testar f_2:

$$1 = c \int_0^2 \left(x^3 + x\right) dx = 6c \quad \Rightarrow \quad c = 1/6.$$

Det är även klart att $f_2(x) \geq 0$ för alla $x \in [0, 2]$. Med $c = 1/6$ är alltså f_2 en täthetsfunktion. Den eftersökta sannolikheten kan beräknas enligt

$$P(X \leq 1) = \int_0^1 \frac{1}{6} \left(x^3 + x\right) dx = \frac{1}{8}.$$

3.2 Vanliga kontinuerliga fördelningar

Analogt med diskreta variabler definieras de kontinuerliga ofta från sina respektive täthetsfunktioner. Vi definierar några av de vanligaste. Det finns många andra fördelningar som ofta används, men dessa är de vi kommer att använda mest. Se boken för fler exempel (Gammafördelning, Weibullfördelning, χ^2-fördelning, t-fördelning mfl.)

Normalfördelning

Variabeln X kallas **normalfördelad** med parametrarna μ och σ, $X \sim \mathrm{N}(\mu, \sigma)$, om

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

Vi kommer att studera normalfördelningen i mer detalj senare (mycket mer detalj...). Det är antagligen den viktigaste fördelningen ni kommer att stöta på. Fler vanliga fördelningar följer.

Likformig fördelning

Variabeln X kallas **likformigt** fördelad (eller **rektangel**-), $X \sim \mathrm{U}(a, b)$ eller $X \sim \mathrm{Re}(a, b)$, om

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b, \\ 0, & \text{övriga } x \end{cases}$$

Exempel

Ubbe häller upp Whisky i sitt glas. Vätskenivån är likformigt fördelad mellan två och fem fingrar. Vad är sannolikheten att Ubbe häller upp mindre än 3.2 fingrar?

Lösning: Låt $X \sim \mathrm{Re}(2, 5)$ vara vätskenivån. Vi söker $P(X < 3.2)$:

$$P(X < 3.2) = \int_2^{3.2} \frac{1}{5-2}\,dx = \frac{1}{3}(3.2 - 2) = 0.4.$$

Exponentialfördelning

Variabeln X kallas **exponentialfördelad** med parametern $\lambda > 0$, $X \sim \mathrm{Exp}(\lambda)$, om

$$f_X(x) = \begin{cases} \lambda \exp(-\lambda x), & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Parametern λ tolkas ibland som *intensiteten*.

Exempel

Låt X vara väntetiden i en telefonkö (minuter). Av någon anledning har det visat sig att X har en täthetsfunktion $f_X(x) = c\,e^{-0.05x}$ för $x \geq 0$, där c är en konstant.

(i) Bestäm c så att f_X blir en täthetsfunktion.

(ii) Vad är sannolikheten att få vänta i mer än 50 minuter vid ett samtal?

(iii) Om man ringer 10 olika (oberoende) samtal, vad är sannolikheten att högst ett av dessa har en väntetid på över 50 minuter?

Lösning:

(i) $1 = \int_{-\infty}^{\infty} f_X(x)\,dx = c\int_0^{\infty} e^{-0.05x}\,dx = \frac{c}{-0.05}\left[e^{-0.05x}\right]_0^{\infty} = \frac{c}{20}$, så $c = 1/20$.

(ii) $P(X > 50) = \int_{50}^{\infty} f_X(x)\,dx = \frac{1}{20}\left[\frac{e^{-0.05x}}{-0.05}\right] = e^{-5/2} \approx 0.082.$

(iii) Varje samtal har sannolikheten $e^{-5/2}$ att ha mer än 50 minuters väntetid. Antalet Y av tio stycken samtal som har mer än 50 minuters väntetid blir alltså Binomialfördelad med $n = 10$ och $p = e^{-5/2}$. Vi erhåller

$$\begin{aligned} P(Y \leq 1) &= \sum_{k=0}^{1} \binom{10}{k} (e^{-5/2})^k (1 - e^{-5/2})^{10-k} \\ &= (1 - e^{-5/2})^{10} + 10e^{-5/2}(1 - 10^{-5/2})^9 \approx 0.804. \end{aligned}$$

Exempel

En komponent (som inte åldras) antas ha en livslängd T som är Exp(1/100)-fördelad (enhet: dagar).

(i) Vad är sannolikheten att komponenten går sönder innan 80 dagar?

(ii) Givet att komponenten överlevt 80 dagar, vad är sannolikheten att den klarar 100 dagar?

Lösning:

(i) $P(T \leq 80) = \int_{-\infty}^{80} f_X(x)\,dx = \frac{1}{100}\int_0^{80} e^{-x/100}\,dx = \frac{1}{100}\left[\frac{e^{-x/100}}{-1/100}\right]_0^{80} = 1 - e^{-4/5}$. Sannolikheten blir alltså ca 55.1%.

(ii) Här använder vi definitionen av betingad sannolikhet och erhåller

$$\begin{aligned} P\big(T \geq 100 \mid T \geq 80\big) &= \frac{P(\{T \geq 100\} \cap \{T \geq 80\})}{P(T \geq 80)} = \frac{P(T \geq 100)}{P(T \geq 80)} \\ &= \frac{\frac{1}{100}\int_{100}^{\infty} e^{-x/100}\,dx}{\frac{1}{100}\int_{80}^{\infty} e^{-x/100}\,dx} = \frac{e^{-100/100}}{e^{-80/100}} = e^{-1/5} \approx 0.8187. \end{aligned}$$

Detta är ett exempel på en *betingad fördelning*. Vi återkommer till detta. Observera även att denna sannolikhet är densamma som

$$P(T \geq 20) = \frac{1}{100}\int_{20}^{\infty} e^{-x/100}\,dx = e^{-1/5}.$$

Detta gäller generellt för exponentialfördelningen. Sannolikheten att komponenten klarar 20 dagar är oberoende av hur länge den levt tidigare. Kanske inte alltid rimligt för komponenter?

3.3 Funktioner av stokastiska variabler

Vad händer om vi har en funktion av en stokastisk variabel, säg att $Y = g(X)$, där vi känner till fördelningen för X och hur funktionen g ser ut? Vi belyser med ett par exempel.

Exempel

Låt $X \sim \text{Exp}(1)$ och definiera $Z = 5X + 2$. Vad blir $f_Z(z)$?

Lösning: Fördelningsfunktionen för Z kan beräknas genom

$$F_Z(z) = P(Z \leq z) = P(5X + 2 \leq z) = P\left(X \leq \frac{z-2}{5}\right) = F_X\left(\frac{z-2}{5}\right).$$

Vidare får vi då

$$f_Z(z) = F_Z'(z) = F_X'\left(\frac{z-2}{5}\right)\frac{1}{5} = f_X\left(\frac{z-2}{5}\right)\frac{1}{5} = \begin{cases} \frac{1}{5}e^{-(z-2)/5}, & z \geq 2, \\ 0, & z < 2. \end{cases}$$

Exempel

Om $X \sim \text{Exp}(\lambda)$, vad får $Y = e^X$ för täthetsfunktion?

Lösning: Vad blir f_Y? Vi ser att $Y > 0$ från definitionen så $f_Y(y) = 0$ för $y \leq 0$. Vi ställer upp $F_Y(y)$ för $y > 0$:

$$F_Y(y) = P(Y \leq y) = P(e^X \leq y) = P(X \leq \log y) = F_X(\log y), \quad y > 0.$$

Vi deriverar fram $f_Y(y) = F_Y'(y) = \frac{1}{y} \cdot f_X(\log y)$. Vi vet att $f_X(x) = \lambda e^{-\lambda x}$ för $x > 0$ och om $x \leq 0$ blir $f_X(x) = 0$. Eftersom $\log y < 0$ då $0 < y < 1$ får vi två fall:

$$f_Y(y) = \begin{cases} \frac{\lambda}{y}e^{-\lambda \log y}, & y > 1, \\ 0, & y \leq 1, \end{cases} = \begin{cases} \lambda y^{-1-\lambda}, & y > 1, \\ 0, & y \leq 1. \end{cases}$$

Exempel

Om $X \sim \text{Re}(-1, 2)$, vad får $Y = X^2$ för täthetsfunktion?

Lösning: Om $y < 0$ så måste $F_Y(y) = P(Y \leq y) = 0$ ($Y = X^2$ kommer aldrig att vara negativ). Låt oss anta att $y \geq 0$. Vi beräknar

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= P(X \leq \sqrt{y}) - P(X \leq -\sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}). \end{aligned}$$

Vi antar att f_Y är kontinuerlig och deriverar fram ett uttryck:

$$f_Y(y) = F_Y'(y) = \begin{cases} \frac{1}{2\sqrt{y}}\big(f_X(\sqrt{y}) + f_X(-\sqrt{y})\big), & y > 0, \\ 0, & y \leq 0. \end{cases}$$

Vidare vet vi att $f_X(x) = 1/3$ om $-1 \leq x \leq 2$ och $f_X(x) = 0$ annars, så

$$f_Y(y) = \begin{cases} 0, & y < 0, \\ \frac{1}{3\sqrt{y}}, & 0 \leq y < 1, \\ \frac{1}{6\sqrt{y}}, & 1 \leq y < 4, \\ 0, & y \geq 4. \end{cases}$$