Föreläsning 3: Stokastiska variabler (forts)

Johan Thim (johan.thim@liu.se)

January 9, 2022

1 Väntevärde

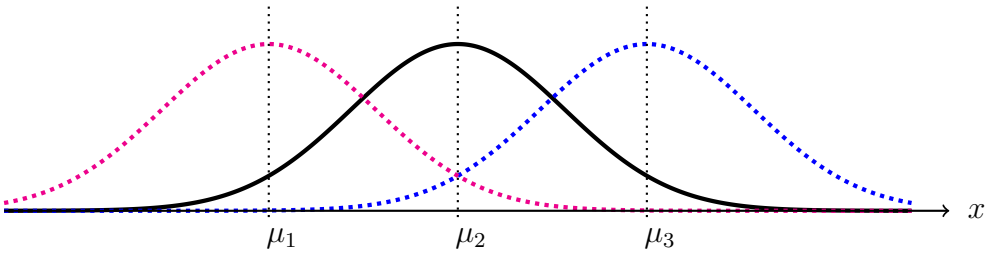
Väntevärde

Definition. Väntevärdet $E(X)$ av en stokastisk variabel X definieras som

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\,dx \qquad \text{respektive} \qquad E(X) = \sum_k k p_X(k)$$

för kontinuerliga (om täthetsfunktion finns) och diskreta variabler.

Andra vanliga beteckningar: μ eller μ_X. Väntevärdet är ett lägesmått som anger var sannolikhetsmassan har sin tyngdpunkt (jämför med mekanikens beräkningar av tyngdpunkt). Om fördelningen är symmetrisk blir det i mitten, men är fördelningen skev blir det annorlunda. I figuren nedan ser vi tre täthetsfunktioner som har samma form men olika väntevärden. De är helt enkelt translationer av samma funktion i detta fall.



Exempel

Vid tillämpningar tolkas ofta väntevärdet som just det *förväntade värdet* för en stokastisk variabel.

(i) Om X är koncentrationen i en flaska salpetersyra så är $E(X)$ den koncentration vi förväntar oss när vid tar ned flaskan från hyllan.

(ii) Om X är antalet kast med en tärning innan vi får en 6:a för första gången så är $E(X)$ det förväntade antalet kast innan vi ser den första 6:an.

Observera att väntevärdet är ett reellt tal, så det kan mycket väl vara så att en variabel (då oftast en diskret sådan) inte *kan* anta sitt väntevärde.

Exempel

Kasta en 4-sidig tärning och låt X vara utfallet $1, 2, 3$ eller 4. Beräkna $E(X)$.

Lösning. Vi antar att tärningen är ärlig så $p_X(k) = 1/4$ för $k = 1, 2, 3, 4$. Då blir

$$E(X) = \sum_{k=1}^{4} k p_X(k) = \frac{1+2+3+4}{4} = \frac{5}{2}.$$

Det förväntade resultatet är alltså 2.5. Knappast ett resultat vi förväntar oss vid ett enskilt kast!

Vad är *egentligen* ett väntevärde?

Vi har gjort en definition av begreppet väntevärde ovan, så det är den som gäller. Men åtminstone följande tolkningar eller alternativa definitioner finns.

(i) Ett sannolikhetsviktat medelvärde av de värden X kan anta.

(ii) Integralen av X med avseende på sannolikhetsmåttet: $\int_\Omega X(\omega)\, dP(\omega)$ (vad nu detta betyder).

(iii) Masscentrum för sannolikhetsfördelningen.

(iv) Det värde X hamnar på i snitt vid väldigt många upprepningar.

Vad punkt (ii) betyder kräver mer analys än vi har tillgång till.

1.1 Funktioner och väntevärden

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? Om g är snäll så ger detta upphov till en ny stokastisk variabel och ibland kan man explicit härleda fördelning för denna (vi återkommer till detta senare), men faktum är att vi kan hantera funktioner av stokastiska variabler på ett smidigare sätt om vi endast behöver väntevärdet.

Följande sats (ofta känd som *the law of the unconscious statistician*) visar hur detta fungerar, men resultatet behöver egentligen lite diskussion (samt ett bevis).

Väntevärde och funktioner av stokastiska variabler

Sats. Låt $Y = g(X)$ där g är snäll. Då gäller

$$E(Y) = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx \qquad \text{och} \qquad E(Y) = \sum_k g(k) p_X(k)$$

för Y kontinuerlig (med täthetsfunktion) respektive diskret.

Exempel

Låt X vara utfallet vid ett tärningskast med en symmetrisk 4-sidig tärning. Beräkna $E(X^2)$.

Lösning. Vi söker $E(X^2)$, så $g(t) = t^2$ är funktionen som transformerar $Y = g(X)$. Enligt satsen ovan blir då

$$E(X^2) = \sum_{k=1}^{4} k^2 p_X(k) = \frac{1^2 + 2^2 + 3^2 + 4^2}{4} = \frac{15}{2}.$$

Notera speciellt att $E(X^2) \neq E(X)^2 = (5/2)^2$.

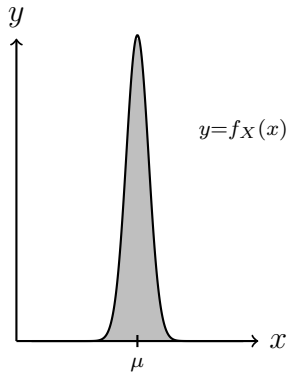
2 Varians och standardavvikelse

Varians och standardavvikelse

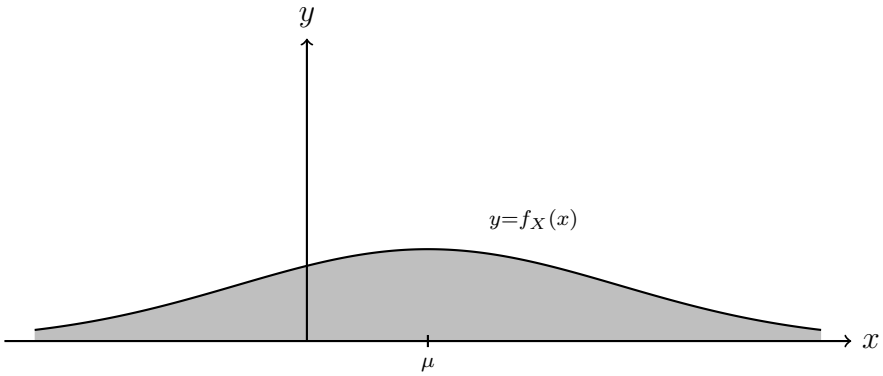
Definition. Låt X vara en stokastisk variabel med $|E(X)| < \infty$. **Variansen** $V(X)$ definieras som $V(X) = E((X - E(X))^2)$. **Standardavvikelsen** $D(X)$ definieras som $D(X) = \sqrt{V(X)}$.

Andra vanliga beteckningar för standardavvikelsen: σ, σ_X, $\sigma(X)$.

Variansen är ett spridningsmått. Stor varians (eller standardavvikelse) betyder att sannolikhetsfördelning har stor spridning. Många värden är troliga. Liten varians betyder att fördelningen är centrerad, hög sannolikhet att hamna kring en viss punkt; se figuren nedan.



Litet $D(X) = \sigma$.



Stort $D(X) = \sigma$.

Steiners sats

Sats. $V(X) = E(X^2) - E(X)^2$.

Exempel

Låt X_1 anta värdena $\{-1, 1\}$ med $p_{X_1}(-1) = p_{X_1}(1) = 1/2$ och låt X_2 anta värdena $\{-10, 10\}$ med $p_{X_2}(-10) = p_{X_2}(10) = 1/2$. Beräkna $V(X_1)$ och $V(X_2)$.

Lösning: Det är klart att $E(X_1) = E(X_2) = 0$ (varför?), så

$$V(X_1) = E(X_1^2) - E(X_1)^2 = \frac{1}{2}(-1)^2 + \frac{1}{2}1^2 - 0 = 1$$

och

$$V(X_2) = E(X_2^2) - E(X_2)^2 = \frac{1}{2}(-10)^2 + \frac{1}{2}(10)^2 - 0 = 50 + 50 = 100.$$

Tydligt att X_2 har mycket större varians även om fördelningarna kan tyckas se snarlika ut, men sannolikheten är mycket mer utspridd för X_2.

3 Räknelagar

För väntevärdet gäller bland annat följande regler.

Linjäritet och oberoende produkt

Låt X och Y vara stokastiska variabler. Då gäller

(i) $E(aX + b) = aE(X) + b$ för alla $a, b \in \mathbf{R}$;

(ii) $E(aX + bY) = aE(X) + bE(Y)$ för alla $a, b \in \mathbf{R}$;

(iii) Om X och Y är oberoende gäller $E(XY) = E(X)E(Y)$.

Vi visar dessa räkneregler genom att sätta in allt i definitionen och sedan utnyttja att både integralen och summan är linjära operationer (så vi kan dela upp över plus-tecknet och bryta ut konstanter).

För variansen kan vi visa istället visa följande. Här blir beviset lite bökigare på grund av att vi inte längre arbetar med en så kallad linjär operator, men tanken är snarlik (sätt in i definitionen och se vad som händer).

Låt X och Y vara stokastiska variabler. Då gäller

(i) $V(aX + b) = a^2V(X)$ för alla $a, b \in \mathbf{R}$;

(ii) $V(aX \pm bY) = a^2V(X) + b^2V(Y) + 2ab(E(XY) - E(X)E(Y))$ för alla $a, b \in \mathbf{R}$;

(iii) Om X och Y är oberoende gäller $V(aX \pm bY) = a^2V(X) + b^2V(Y)$.

Varianser adderas alltid!

Observera att det *alltid* blir ett plustecken mellan varianserna för linjär-kombinationer:

$$V(aX \pm bY) = a^2V(X) + b^2V(Y).$$

Vi kommer *aldrig* att bilda skillnader mellan varianser!

Det följer att standardavvikelsen för en linjärkombination $aX + bY$ av två oberoende stokastiska variabler ges av $\sigma_{aX+bY} = \sqrt{a^2\sigma_X^2 + b^2\sigma_Y^2}$.

4 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.)

Likformig fördelning

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

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

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

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

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 $x \ge 0$, 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

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

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

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

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

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

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

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?

5 Median

Ett annat lägesmått än väntevärdet är medianen.

Median

Definition. En **median** för en stokastisk variabel X är ett tal $m \in \mathbf{R}$ så att

$$P(X \leq m) = P(X \geq m) = \frac{1}{2}.$$

Observera att medianen inte behöver vara entydig!

Medianen för en Exponentialfördelning

Låt $X \sim \text{Exp}(\lambda)$. Beräkna medianen och väntevärdet för X.

Lösning: Vi räknar ut fördelningsfunktionen för X. Om $x > 0$,

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt = \int_0^x \lambda e^{-\lambda t}\,dt = 1 - e^{-\lambda x}.$$

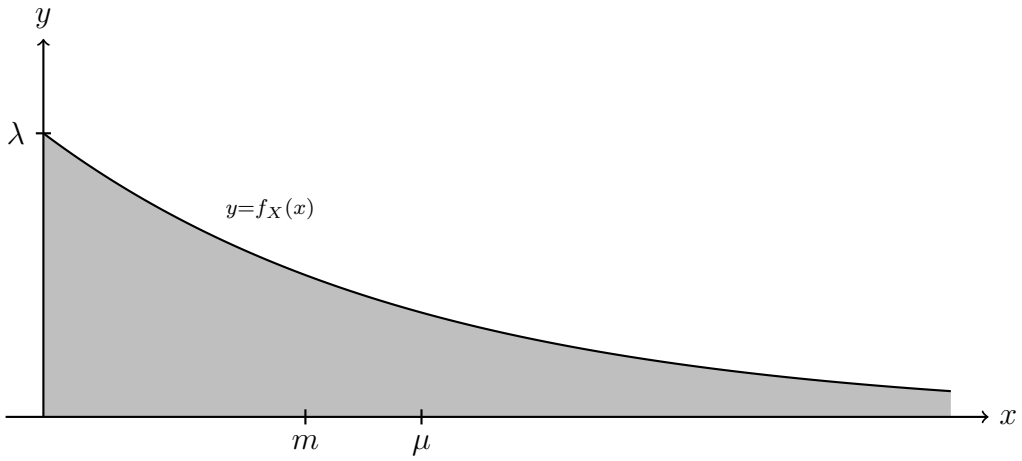
Medianen finner vi ur ekvationen $F_X(m) = 1/2$, dvs

$$1 - e^{-\lambda m} = \frac{1}{2} \quad \Leftrightarrow \quad e^{-\lambda m} = \frac{1}{2} \quad \Leftrightarrow \quad m = \frac{\ln 2}{\lambda}.$$

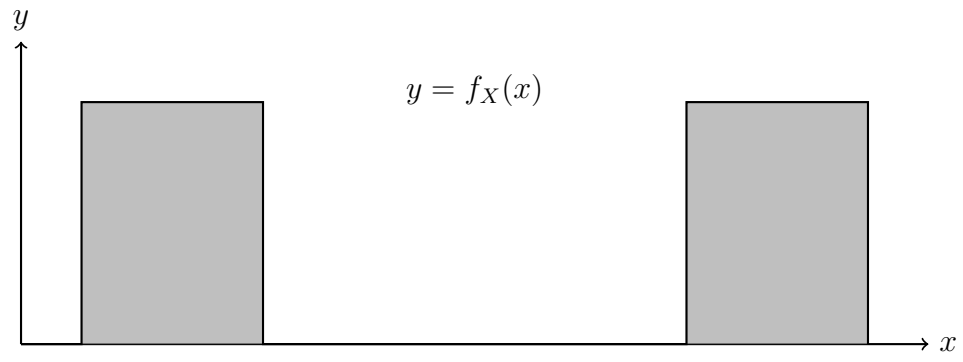
Jämför detta med väntevärdet för X:

$$E(X) = \int_0^\infty x\lambda e^{-\lambda x}\,dx = \left[-xe^{-\lambda x}\right]_0^\infty + \int_0^\infty e^{-\lambda x}\,dx = 0 - 0 + \left[-\frac{e^{-\lambda x}}{\lambda}\right]_0^\infty = \frac{1}{\lambda}.$$

Medianen och väntevärdet behöver alltså *inte* vara samma sak!



Ett annat exempel, där medianen *inte* är entydigt definierad:



Var är medianen??

6 Normalfördelning

Normalfördelning

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

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

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

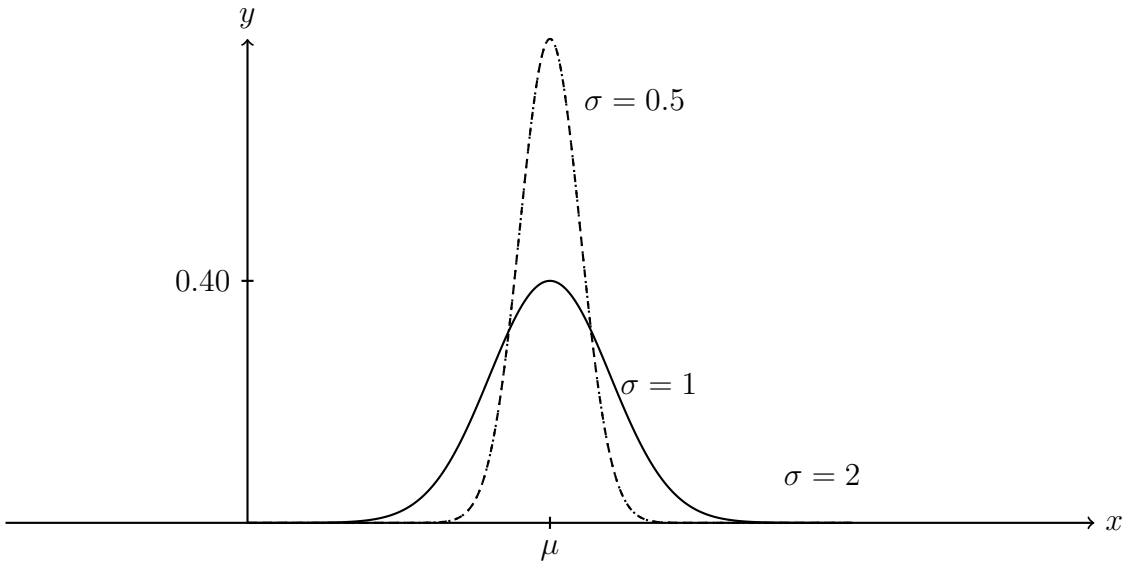
$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right) du, \quad x \in \mathbf{R},$$

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.

Vi kommer visa att $E(X) = \mu$ och $V(X) = \sigma^2$ om $X \sim N(\mu, \sigma)$.



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!

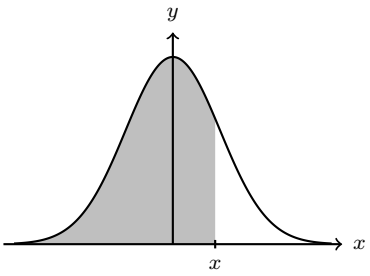
Bruk av tabell för $\Phi(x)$

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

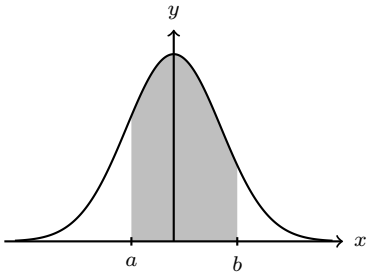
(i) $P(X \leq x) = \Phi(x)$ för alla $x \in \mathbf{R}$;

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

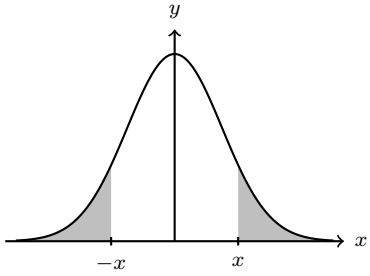
(iii) $\Phi(-x) = 1 - \Phi(x)$ för alla $x \in \mathbf{R}$.



$\Phi(x)$



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



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

Exempel

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

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

7 ($\star\star$)Väntevärde för geometrisk fördelning

Följande avsnitt kräver att man arbetat en del med potensserier (kapitel 10 i envariabelboken).

Exempel

Låt X anta värdena $0, 1, 2, \ldots$ med sannolikheterna $p_X(k) = 2^{-k-1}$. Beräkna $E(X)$ och $V(X)$.

Lösning: Till att börja med kan vi kontrollera att p_X verkligen är en sannolikhetsfunktion. Klart att $p_X(k) \geq 0$, och summan nedan är geometrisk så

$$\sum_{k=0}^{\infty} p_X(k) = \sum_{k=0}^{\infty} 2^{-k-1} = 2^{-1} \sum_{k=0}^{\infty} 2^{-k} = 2^{-1} \cdot \frac{1}{1-1/2} = 1.$$

Vi beräknar väntevärdet. Låt $q \in]0,1[$ så $p_X(k) = (1-q)q^k$ med $q = 1/2$. Vi kan beräkna summan av kq^k genom följande manöver:

$$\sum_{k=0}^{\infty} kq^k = \sum_{k=1}^{\infty} kq^k = q \sum_{k=1}^{\infty} kq^{k-1} = q \sum_{k=0}^{\infty} \frac{d}{dq} q^k = q \frac{d}{dq} \sum_{k=0}^{\infty} q^k = q \frac{d}{dq} \frac{1}{1-q} = \frac{q}{(1-q)^2}.$$

Alltså blir

$$E(X) = (1-q) \sum_{k=0}^{\infty} kq^k = \frac{q}{1-q} \qquad \text{och} \qquad E(X) = 1 \text{ om } q = \frac{1}{2}.$$

För att beräkna $E(X^2)$ kikar vi på motsvarande kalkyl för andraderivatan:

$$q^2 \sum_{k=0}^{\infty} \frac{d^2}{dq^2} q^k = q^2 \sum_{k=0}^{\infty} (k^2 - k) q^{k-2} = \sum_{k=0}^{\infty} (k^2 - k) q^k = \sum_{k=0}^{\infty} k^2 q^k - \frac{q}{(1-q)^2}.$$

Således,

$$\sum_{k=0}^{\infty} k^2 q^k = \frac{q}{(1-q)^2} + q^2 \frac{d^2}{dq^2} \sum_{k=0}^{\infty} q^k = \frac{q}{(1-q)^2} + \frac{2q^2}{(1-q)^3},$$

vilket medför att

$$E(X^2) = (1-q) \sum_{k=0}^{\infty} k^2 q^k = \frac{q}{1-q} + \frac{2q^2}{(1-q)^2}$$

så

$$V(X) = E(X^2) - E(X)^2 = \frac{q}{1-q} + \frac{q^2}{(1-q)^2} = \frac{q}{(1-q)^2}.$$

och med $q = 1/2$ får vi $V(X) = 2$. Den observante läsaren kanske känner igen sannolikhetsfunktionen vi arbetar med då det är den geometriska fördelningen $X \sim \text{Geo}(1-q)$. Så vad vi visat ovan är följande:

Geometrisk fördelning

Sats. Om $X \sim \text{Geo}(p)$ så är $E(X) = \dfrac{1-p}{p}$ och $V(X) = \dfrac{1-p}{p^2}$.