www.math-stockholm.se/cirkel

27 januari 2022

KTH VETENSKAP OCH KONST

STOCKHOLMS MATEMATIKCENTRUM

STOCKHOLMS UNIVERSITET

Hej!

- Siobhán (Sha-von) Correnty
- Matteintresserad
  - Lärare
  - Master i matematik (Lunds universitet)
  - Doktorand i tillämpad matematik (KTH)

Välkomna!

Idag:

- Introducera de funktioner som beskriver slumpstrukturen hos diskreta och kontinuerliga slumpvariabler
    - Sannolikhetsfunktionen
    - Fördelningsfunktionen
- Dessa ligger till grund för nästan all analys av slumpvariabler
- Behövas för att kunna definiera **väntevärde** och **varians**
- Introducera den kanske viktigaste fördelningen av alla, nämligen **normalfördelningen**

Diskreta slumpvariabler och sannolikhetsfunktioner

Definition

En slumpvariabel X är *diskret* om den endast kan anta ändligt eller uppräkneligt oändligt antal värden $x_1, x_2, \ldots$

Notera: uppräkneligt oändligt är t.ex. $0, 1, 2, 3, 4, \ldots$ men inte alla siffror mellan 0 och 1

Definition

Sannolikhetsfunktionen, p_X, för en diskret slumpvariabel X definieras av

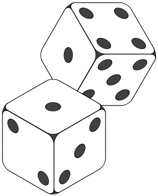
$$p_X(x) := P(X \in \{x\}) = P(X \text{ antar värdet } x), \quad x = x_1, x_2, \ldots$$

Vi tar ett exempel . . .

Exempel: Diskreta slumpvariabler och sannolikhetsfunktioner (1)

Exempel: Slår två tärningar

- X är sannolikhetsfunktionen som är 1 om vi får par (samma tal på båda tärningarna) och 0 annars
- Denna slumpvariabel är diskret då den endast kan anta värdena 0 och 1



Figur: Slår två tärningar

Exempel: Diskreta slumpvariabler och sannolikhetsfunktioner (2)

- Sannolikheten att X ska bli 1 (vi får par) är

$$\begin{aligned} p_X(1) &= P(u \in \Omega : X(u) = 1) \\ &= P(u \in \{(1,1),(1,2),\dots(6,6)\} : X(u) = 1) \\ &= P(u \in \{(1,1),(2,2),\dots,(5,5),(6,6)\}) \\ &= 1/6 \end{aligned}$$

eftersom:

  - Vi har 6 **möjliga utfall** som ger värdet 1 av **totalt** 36 stycken
  - Alla utfall är **lika troliga**

Exempel: Diskreta slumpvariabler och sannolikhetsfunktioner (3)

- Sannolikheten att X ska bli 0 (inte par) är

$$X(u) = 0 \iff X(u) \neq 1$$

- Vi beräknar

$$p_X(0) = P(u \in \{(1,2),(1,3),(1,4)\ldots,(6,3),(6,4),(6,5)\})$$

- Vi har

$$p_X(0) = P(X = 0) = 1 - P(X = 1) = 1 - 1/6 = 5/6$$

Diskreta slumpvariabler

Sats 5.1.4

För en diskret slumpvariabel X gäller

- $0 \leqslant p_X(k) \leqslant 1$ för alla k
- $\sum_k p_X(k) = 1$
- $P(a \leqslant X \leqslant b) = \sum_{k: a \leqslant k \leqslant b} p_X(k)$
- $P(X \leqslant a) = \sum_{k: k \leqslant a} p_X(k)$
- $P(X > a) = \sum_{k: k > a} p_X(k) = 1 - \sum_{k: k \leqslant a} p_X(k) = 1 - P(X \leqslant a)$

Notera: dessa egenskaper följer direkt från Kolmogorovs axiom

Exempel 5.1.5 (1)

- Tre personer korrekturläser en text i vilken det finns exakt **ett** stavfel
- För var och en av dessa personer är det 10 procents chans att de **missar** detta stavfel
- Låt X vara det **antal** bland de tre personerna som missar stavfelet
- Vilka observationsvärden k kan X ha? Vad är $p_X(k)$ för dessa observationsvärden?

Exempel 5.1.5 (2)

- Eftersom 3 personer läser texten kan anting inga, $1, 2$ eller 3 personer **missa** stavfelet
- Vi börjar med att härleda $p_X(k)$ för de enklaste värdena, nämligen 0 och 3

Att exakt 0 eller 3 personer missar stavfelet

- Eftersom det är 10 procents sannolikhet att en person missar stavfelet är sannolikheten att alla missar stavfelet

$$p_X(3) = 0.1^3$$

- Eftersom det är 90 procents sannolikhet att en enskild person hittar stavfelet är sannolikheten att ingen missar stavfelet

$$p_X(0) = 0.9^3$$

Exempel 5.1.5 (3)

Att exakt en person missar stavfelet

- Sannolikheten att den *första* personen missar stavfelet är 0.1, och därefter är sannolikheten att den andre hittar det 0.9, och att den tredje hittar det också 0.9
- Sannolikheten för att den första missar stavfelet och de andra hittar det $0.1 \cdot 0.9^2$
- Sammanlagt får vi att sannolikheten för att *precis* en person missar är

$$p_X(1) = 3 \cdot 0.1 \cdot 0.9^2$$

eftersom den första, andre eller tredje kan missa det

Exempel 5.1.5 (4)

Att två personer missar stavfelet

- Slutligen kan vi använda egenskaperna för sannolikhetsfunktioner för att härleda att

$$p_X(2) = 1 - p_X(0) - p_X(1) - p_X(3) = 1 - 0.9^3 - 3 \cdot 0.1 \cdot 0.9^2 - 0.1^3$$

- Och vi är klar

Binomialfördelning

- Sannolikhetsfördelning som i Exempel 5.1.5 är väldigt vanligt förekommande och kallas en *binomialfördelning*
- De uppkommer i situationer som den ovan, dvs
  - Vi utför ett experiment ett visst antal gånger, n, och experimentet
    - Lyckas med någon sannolikhet p (missa stavfelet)
    - Misslyckas med någon sannolikhet $1 - p$ (hitta stavfelet)
  - Vi undrar vad sannolikheten är att exakt k av de n utförda experimenten lyckas

Fördelningsfunktioner (1)

- Ofta är man intresserad av att beräkna uttryck som

$$P(a < X < b), P(X \leqslant a), \text{ och } P(X > a)$$

- Vill beräkna sannolikheten att summan av fem tärningsslag är större än 25 är man intresserad av att beräkna $P(X > 25)$ för någon lämplig slumpvariabel X
- För en **diskret slumpvariabel** kan vi använda sannolikhetsfunktionen
- För en **kontinuerlig slumpvariabel** är ofta sannolikheten att få *exakt* ett specifikt värde k helt obefintlig, och vi kan därför inte prata om en sannolikhetsfunktion, $P(X = k)$, på ett meningsfullt sätt

Fördelningsfunktioner (2)

Definition

Fördelningsfunktionen $F_X(t)$ för en slumpvariabel X definieras av

$$F_X(t) := P(X \leqslant t), \quad -\infty < t < \infty$$

Notera att för en **diskret slumpvariabel** gäller $F_X(t) = \sum_{x \leqslant t} p_X(x)$

Kontinuerlig slumpvariabel

Definition

En slumpvariabel X sägs vara *kontinuerlig* om det finns en funktion $f_X(t)$ så att det för "alla" mängder A gäller att

$$P(X \in A) = \int_A f_X(t)dt$$

och funktionen f_X kallas för slumpvariabelns *täthetsfunktion*

Notera: definitionen ovan inte är helt entydig

- Vi kan alltid ändra funktionens värde i någon enstaka punkt utan att det påverkar integralen till höger
- Det är underförstått att vi alltid väljer "den mest kontinuerliga" täthetsfunktionen

Kontinuerlig slumpvariabel

Definition

En slumpvariabel X sägs vara *kontinuerlig* om det finns en funktion $f_X(t)$ så att det för "alla" mängder A gäller att

$$P(X \in A) = \int_A f_X(t)dt$$

och funktionen f_X kallas för slumpvariabelns *täthetsfunktion*

Notera: definitionen ovan fungerar dock bra för mängder A som kan beskrivas som **unioner av öppna intervall**, vilket är tillräckligt för våra ändamål

Täthetsfunktionen och fördelningsfunktionen

Sats 5.2.3

För en kontinuerlig slumpvariabel X med täthetsfunktion f_X och fördelningsfunktion F_X gäller

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

och omvänt gäller det att

$$f_X(x) = F_X'(x) = \lim_{h \to 0} \frac{F_X(x+h) - F_X(x)}{h}$$

för de punkter där $f_X(x)$ är kontinuerlig

Notera: Bevis i kompendiet

Täthetsfunktionen och fördelningsfunktionen

Notera även att:

$$\begin{aligned}\lim_{t\to\infty}\int_{-\infty}^{t} f_X(t)dt &= \lim_{t\to\infty} F_X(t)\\ &= \lim_{t\to\infty} P\left(u\in\Omega : X(u)\in(-\infty, t]\right)\\ &= 1\end{aligned}$$

Exempel 5.2.4 (1)

- Låt $a > 0$ vara ett reellt tal och låt X vara en slumpvariabel med täthetsfunktion
$$f_X(t) = ae^{-at}$$
för $t > 0$ och $f_X(t) = 0$ för $t \leqslant 0$
- En sådan slumpvariabel sägs vara *exponentialfördelad*

- Exponentialfördelningar uppkommer i situationer då:
  - Man mäter tiden fram tills att något inträffar
  - Denna sak har samma sannolikhet att inträffa i varje givet ögonblick

Exempel 5.2.4 (1)

- Låt $a > 0$ vara ett reellt tal och låt X vara en slumpvariabel med täthetsfunktion
$$f_X(t) = ae^{-at}$$
för $t > 0$ och $f_X(t) = 0$ för $t \leqslant 0$
- En sådan slumpvariabel sägs vara *exponentialfördelad*

- Exempel:
  - Tiden till att nästa person kommer in genom dörren på ett kafé (under öppettider och någon relativt fix tid på dagen, säg 10:00 - 10:30) **exponentialfördelad** för någon parameter a
  - Hur lång tid det tar innan en radioaktiv partikel sönderfaller är också **exponentialfördelad** för någon parameter a

Exempel 5.2.4 (2)

- För en exponentialfördelad slumpvariabel kan vi räkna ut att fördelningsfunktionen ges av

$$F_X(x) = \int_0^x ae^{-at} dt = [-e^{-at}]_0^x = 1 - e^{-ax}$$

- $\lim_{x \to \infty} F_X(x) = 1$, som är ett villkor för att F_X ska vara en riktig fördelningsfunktion

Exempel 5.2.4 (3)

- Låt X vara en slumpvariabel som beskriver tiden det tar innan nästa person går in genom dörren på ett kafé
- Täthetsfunktionen för X ges av

$$f_X(t) = 2e^{-2t}$$

där t mäter tiden i timmar (och $a = 2$)

- Sannolikheten att nästa person har kommit in på kaféet **inom en halvtimme**

$$F_X(1/2) = 1 - e^{-2\cdot(1/2)} \approx 0.63$$

- Sannolikheten att nästa person har kommit in **inom en timme**

$$F_X(1) = 1 - e^{-2} \approx 0.86$$

Väntevärdet

Definition

Väntevärdet för en slumpvariabel X betecknas med $E(X)$, μ_X, eller bara μ om det inte kan förväxlas med andra väntevärden. För en diskret slumpvariabel definieras väntevärdet genom summan

$$E(X) := \sum_k k \cdot p_X(k),$$

och för en kontinuerlig slumpvariabel genom integralen

$$E(X) := \int_{-\infty}^{\infty} x \cdot f_X(x) dx.$$

Väntevärdet är den stokastiska versionen av **medelvärde**; man beräknar summan respektive integralen av alla möjliga utfall **viktat** med hur sannolikt det är att slumpvariabeln antar **varje utfall**.

Väntevärdet

Definition

Väntevärdet för en slumpvariabel X betecknas med $E(X)$, μ_X, eller bara μ om det inte kan förväxlas med andra väntevärden. För en diskret slumpvariabel definieras väntevärdet genom summan

$$E(X) := \sum_k k \cdot p_X(k),$$

och för en kontinuerlig slumpvariabel genom integralen

$$E(X) := \int_{-\infty}^{\infty} x \cdot f_X(x)dx.$$

Notera att definitionen bara gäller om summan eller integralen ovan är **väldefinierad**. Om integralen eller summan ovan är oändlig säger vi att X **saknar väntevärde**.

Väntevärdet

Definition

Väntevärdet för en slumpvariabel X betecknas med $E(X)$, μ_X, eller bara μ om det inte kan förväxlas med andra väntevärden. För en diskret slumpvariabel definieras väntevärdet genom summan

$$E(X) := \sum_k k \cdot p_X(k),$$

och för en kontinuerlig slumpvariabel genom integralen

$$E(X) := \int_{-\infty}^{\infty} x \cdot f_X(x) dx.$$

Notera även att väntevärdet för en slumpvariabel är ett **fixt reellt tal**, till skillnad från själva slumpvariabeln som är en **funktion på utfallsrummet** och kan **anta olika värden**.

Exempel 5.3.2

- Betrakta återigen exemplet med att vi slår en tärning
- Låt X vara slumpvariabeln som ger oss talet som kommer upp när vi slagit tärningen
- Eftersom alla utfall är lika sannolika, det vill säga

$$p_X(j) = P(X = j) = \frac{1}{6}$$

för $j = 1, 2, \ldots, 6$, ges väntevärdet av

$$E(X) = \sum_{k=1}^{6} k p_X(k) = \sum_{k=1}^{6} \frac{k}{6} = \frac{1+2+3+4+5+6}{6}$$
$$= \frac{21}{6} = 3.5$$

Exempel 5.3.3

- Betrakta återigen exemplet där X är en exponentialfördelad slumpvariabel
- X har täthetsfunktion

$$f_X(t) = ae^{-at}$$

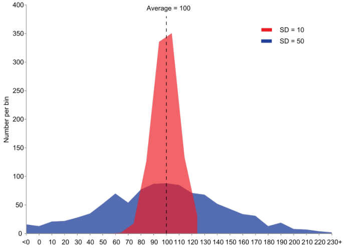
för något $a > 0$ om $t > 0$ och $f_X(t) = 0$ om $t \leqslant 0$

- Väntevärdet för X ges då av

$$E(X) = \int_0^\infty tae^{-at}dt = \left[-te^{-at} - \frac{e^{-at}}{a}\right]_0^\infty = \frac{1}{a}$$

Varians (1)

- Väntevärdet beskriver det förväntade utfallet av en slumpvariabel
- Detta är dock bara *en* siffra, och ger inte all relevant information om en slumpvariabel.
- En annan kvantitet som ofta är av intresse är *variansen*, vilket är ett mått på hur mycket spridning slumpvariabeln har



Figur: Två fördelningar med samma medelvärde, olika varians

Varians (2)

Exempel

- Betrakta två grupper med sex personer i varje
  - I den ena gruppen är personerna $100, 200, 120, 180, 190$ och 110 cm långa
  - I den andra gruppen är alla personerna 150 cm långa
- Båda grupperna har samma medellängd
- I den andra gruppen är fördelningen av längder uppenbarligen mycket mer (helt) koncentrerad kring medelvärdet
- Variansen kvantifierar den förväntade avvikelsen från medelvärdet för en slumpvariabel

Varians (3)

Definition

Variansen $V(X)$ för en slumpvariabel X med väntevärde μ definieras som $V(X) := E((X-\mu)^2)$ om detta uttryck är ändligt. Variansen betecknas ofta med σ^2 eller σ_μ^2.

- Om X med 100 procents sannolikhet antar sitt medelvärde blir variansen 0
- Om X med stor sannolikhet hamnar väldigt långt ifrån sitt medelvärde blir variansen stor

Varians (3)

Definition

Variansen $V(X)$ för en slumpvariabel X med väntevärde μ definieras som $V(X) := E((X-\mu)^2)$ om detta uttryck är ändligt. Variansen betecknas ofta med σ^2 eller σ_μ^2.

Notera: det inte kan ske någon kancellering av termer i beräkningen av väntevärdet i formeln då alla termer är positiva eftersom vi betraktar **kvadraten av avvikelsen från väntevärdet**, $(X-\mu)^2$.

Varians (4)

Exempel 5.4.2

- Har beräknat väntevärdet för ett tärningsslag: 3.5
- Eftersom $p_X(k) = 1/6$ för $k = 1, \ldots, 6$ ges variansen av

$$\begin{aligned} E((X-3.5)^2) &= \sum_{k=1}^{6} (k-3.5)^2 p_X(k) \\ &= \frac{1}{6} \sum_{k=1}^{6} (k-3.5)^2 \\ &= \frac{1}{6} \frac{35}{2} \\ &= \frac{35}{12} \end{aligned}$$

Varians (5)

Sats 5.4.3

För en slumpvariabel X med väntevärde μ gäller

$$V(X) = E(X^2) - \mu^2 = E(X^2) - E(X)^2.$$

Bevis (kontinuerlig slumpvariabel)

$$\begin{aligned}
V(X) &= \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)dx = \int_{-\infty}^{\infty} (x^2 - 2x\mu + \mu^2) f_X(x)dx \\
&= \int_{-\infty}^{\infty} x^2 f_X(x)dx - 2\mu \int_{-\infty}^{\infty} x f_X(x)dx + \mu^2 \int_{-\infty}^{\infty} f_X(x)dx \\
&= E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2
\end{aligned}$$

eftersom $\int_{-\infty}^{\infty} f_X(x)dx = 1$

Standardavvikelsen

Definition

Standardavvikelsen $D(X)$ för en slumpvariabel X definieras som

$$D(X) := \sqrt{V(X)}.$$

Standardavvikelsen betecknas ofta med σ eller σ_X.

Notera: Man behöver dock räkna ut variansen för att kunna räkna ut standardavvikelsen.

Standardavvikelsen

Definition

Standardavvikelsen $D(X)$ för en slumpvariabel X definieras som

$$D(X) := \sqrt{V(X)}.$$

Standardavvikelsen betecknas ofta med σ eller σ_X.

- Standardavvikelsen, precis som variansen och väntevärdet är ett **reellt tal**
- Slumpvariabeln är en **funktion** och kan **anta olika värden** vid skilda observationer

Standardavvikelsen

Definition

Standardavvikelsen $D(X)$ för en slumpvariabel X definieras som

$$D(X) := \sqrt{V(X)}.$$

Standardavvikelsen betecknas ofta med σ eller σ_X.

- Standardavvikelsen har samma **enhet** som slumpvariabeln själv
- Om slumpvariabeln ger längden på fotbollsspelare i **meter**
  - Väntevärdet och standardavvikelsen också vara uttryck i **meter**
  - Variansen ger ett uttryck i **kvadratmeter** eftersom variansen är väntevärdet av avvikelsen från väntevärdet **i kvadrat** (se definitionen)

Normalfördelning (1)

Definition

En kontinuerlig slumpvariabel X sägs vara *normalfördelad* med parametrar μ och σ^2 om täthetsfunktionen ges av

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty.$$

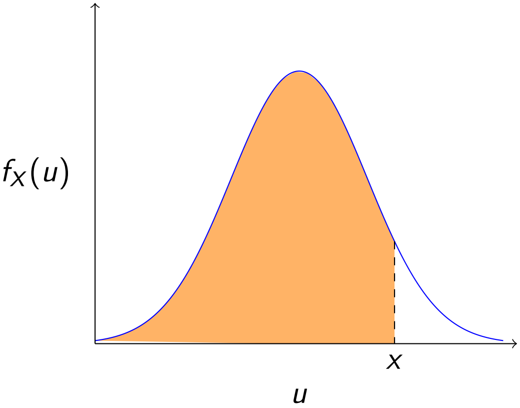
För sådana X används ofta beteckningen $X \sim N(\mu, \sigma^2)$

- Väntevärde μ
- Varians σ^2
- Standardavvikelse σ

Normalfördelning (2)

- *Normalfördelningen* är kanske viktigaste fördelningen av alla
- Oerhört viktig sats som kallas *centrala gränsvärdessatsen* (inte i kursen)
- Normalfördelningar uppkommer hela tiden, i alla möjliga olika sammanhang
- Det enda som krävs är att man betraktar en kumulativ effekt av något som sker tillräckligt ofta

Normalfördelning (3)



Figur: Bilden visar en normalfördelning, och arean av det orangefärgade området ger värdet av fördelningsfunktionen evaluerad i punkten x