TAMS79: Föreläsning 2
Stokastiska variabler

Johan Thim*

6 november 2018

2.1 Endimensionella stokastiska variabler

För att kunna precisera vad för slags funktion (för det är en funktion) en stokastisk variabel är, behöver vi diskutera öppna mängder på den reella axeln $\mathbf{R}$.

Definition. Den minsta (minst antal element) σ-algebran på $\mathbf{R}$ som innehåller *alla* öppna intervall betecknar vi med $\mathscr{B}$. Denna algebra brukar kallas för **Borel-σ-algebran** på $\mathbf{R}$.

Algebran $\mathscr{B}$ innehåller alltså alla mängder av typen $(a, b) \subset \mathbf{R}$, $(-\infty, c) \cup (d, \infty) \subset \mathbf{R}$, komplement av sådana mängder, samt alla uppräkneliga unioner av mängder av föregående typ. Detta är ganska tekniskt, och inget vi kommer att arbeta med direkt. Men för att få en korrekt definition behövs begreppet.

Stokastisk variabel

Definition. En **stokastisk variabel** är en reellvärd funktion definierad på ett utfallsrum Ω. Funktionen X avbildar alltså olika utfall på reella tal; $X : \Omega \to \mathbf{R}$.
Mer precist så kräver vi att $X^{-1}(B) \in \mathcal{F}$ för alla $B \in \mathscr{B}$. Mängden $X^{-1}(B)$ definieras som $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$ och kallas för **urbilden** av B. Mängden består alltså av alla $\omega \in \Omega$ som avbildas in i B.
Bilden av en delmängd A av Ω betecknas med $X(A)$, och

$$X(A) = \{x \in \mathbf{R} : X(\omega) = x \text{ för något } \omega \in A\}.$$

Mängden $X(A)$ är alltså värdemängden för X på mängden A.
Om $X(\Omega)$ är ändlig, eller bara har uppräkneligt många värden, så kallar vi X för en **diskret stokastisk variabel**. Annars kallar vi X för **kontinuerlig**.

*johan.thim@liu.se

Uttrycket $X(\omega)$ är alltså det siffervärde vi sätter på ett visst utfall $\omega \in \Omega$. Varför kravet att urbilden $X^{-1}(B)$ skall tillhöra de tillåtna händelserna? Det faller sig ganska naturligt, då $X^{-1}(B)$ är precis de utfall i Ω som avbildas in i mängden B. Således vill vi gärna att denna samling utfall verkligen utgör en händelse, annars kan vi inte prata om någon sannolikhet för denna samling utfall.
Termen variabel är egentligen lite olycklig då våra stokastiska variabler är funktioner, men det är en gammal tradition som lever kvar. Ett par exempel kan vara på sin plats.

Exempel

(i) Kasta en tärning. Utfallsrummet $\Omega = \{$⚀, ⚁, ⚂, ⚃, ⚄, ⚅$\}$. Låt X vara antalet ögon vid ett tärningskast. X antar värderna $1, 2, 3, 4, 5, 6$, så X är diskret.

(ii) Handla tacosås på måfå. $\Omega = \{\text{Het}, \text{Medel}, \text{Mild}\}$. Låt $X = 1$ om såsen är mild, $X = 5$ om såsen är medel och $X = 10$ om såsen är het. X är diskret.

(iii) Låt X vara livslängden för ett kylskåp. Då kan X (teoretiskt) anta alla värden i intervallet $[0, \infty[$, så X är kontinuerlig.

(iv) Låt Ω bestå av alla möjlig färger på gräset. Låt $X = \lambda$ vara motsvarande våglängd för färgen (kontinuerlig variabel). Låt $Y = 1$ om färgen är grön och $Y = 0$ annars (diskret variabel).

Definitionen ovan är av ganska teknisk karaktär, så vad måste man ta med sig för att kunna tillgodogöra sig resten av denna kurs?

Vad måste jag förstå av all matematiska?

- En stokastisk variabel är en *snäll* funktion från Ω till $\mathbf{R}$.
- Utfallsrummet Ω kan vara abstrakt, e.g., $\Omega = \{\text{Krona}, \ \text{Klave}\}$.
- Det är mängden $X(\Omega)$ som består av siffror.
- Ibland finns en naturlig koppling mellan Ω och $X(\Omega)$, säg om vi kastar en tärning och räknar antalet ögon vi får.
- En händelse är en *snäll* delmängd av Ω.
- Om A är en händelse så är $X(A)$ *värdemängden* för funktionen X med A som definitionsmängd. Speciellt så är $X(\Omega)$ alla möjliga värden vi kan få från variabeln X.
- Urbilden $X^{-1}(B)$ av en delmängd $B \subset \mathbf{R}$ består av *alla* utfall $\omega \in \Omega$ så att siffran $X(\omega)$ ligger i mängden B.

2.2 Diskreta stokastiska variabler

Om $X(\Omega)$ är ändlig eller uppräkneligt oändlig så kallade vi X för diskret. En sådan variabel kan vi karaktärisera med en så kallad **sannolikhetsfunktion**.

Sannolikhetsfunktion

Definition. Sannolikhetsfunktionen $p_X \colon X(\Omega) \to [0,1]$ för en diskret stokastisk variabel definieras av $p_X(k) = P(X = k)$ för alla $k \in X(\Omega)$.

Den vanligaste situationen vi stöter på är att utfallsrummet är numrerat med heltal på något sätt så att p_X är en funktion definierad för (en delmängd av) heltal (när det finns en naturlig koppling mellan Ω och $X(\Omega)$). Ibland är vi slarviga och tänker oss att $p_X(k) = 0$ för siffror k som ej är möjliga ($p_X(-1) = 0$ om X är antal ögon vi ett tärningskast till exempel).
Vissa egenskaper gäller för alla alla sannolikhetsfunktioner:

Egenskaper hos sannolikhetsfunktionen

(i) $p_X(k) \geq 0$ för alla $k \in X(\Omega)$.

(ii) $\displaystyle\sum_{k \in X(\Omega)} p_X(k) = 1$.

(iii) Om $A \subset X(\Omega)$ så är $P(X \in A) = \displaystyle\sum_{k \in A} p_X(k)$.

En sannolikhetsfunktion är alltså *aldrig* negativ, om vi summerar över alla möjliga värden (alla $k \in X(\Omega)$) så måste summan bli ett, och om vi är ute efter sannolikheten att få vissa värden på X så summerar vi sannolikheten för vart och ett av dessa värden!

Fördelningsfunktion

Definition. Fördelningsfunktionen $F_X(x)$ för en stokastisk variabel X definieras av för alla $x \in \mathbf{R}$ av sambandet $F_X(x) = P(X \leq x)$.

Det följer från definitionen att följande påståenden gäller.

Egenskaper hos fördelningsfunktionen

(i) $F_X(x) \to \begin{cases} 0, & x \to -\infty, \\ 1, & x \to +\infty. \end{cases}$

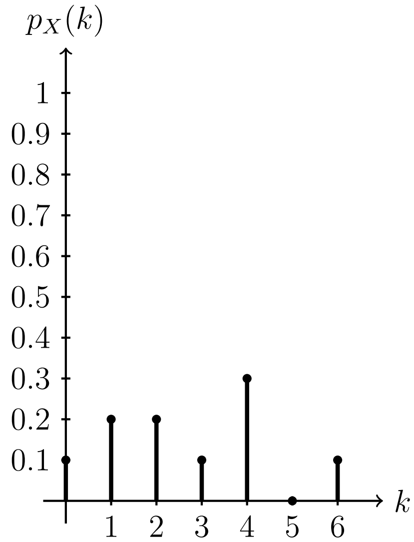
(ii) $F_X(x)$ är icke-avtagande och högerkontinuerlig.

(iii) $F_X(x) = \displaystyle\sum_{\{k \in X(\Omega) : k \leq x\}} p_X(k)$.

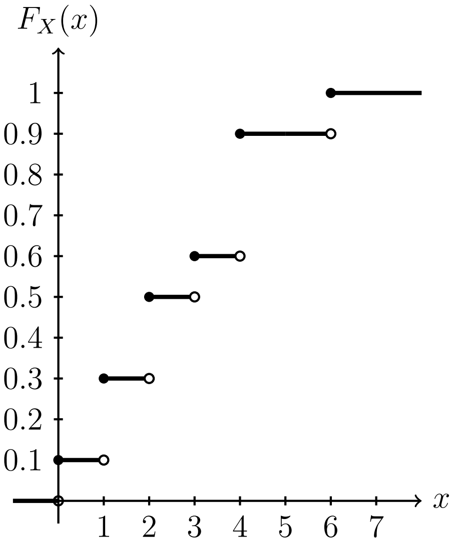
(iv) $P(X > x) = 1 - F_X(x)$.

(v) $F_X(k) - F_X(k-1) = p_X(k)$ för $k \in X(\Omega)$.

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



Sannolikhetsfunktion $p_X(k) = P(X = k)$.



Fördelningsfunktion $F_X(x) = P(X \leq x)$.

2.3 Vanliga diskreta fördelningar

Det räcker med sannolikhetsfunktionen för att karakterisera en diskret variabel, så vi sammanfattar några av de vanligaste fallen. Vi antar genomgående att $0 < p < 1$. En av de enklaste fördelningarna vi stöter på är Bernoullifördelningen, eller 2-punkts fördelningen.

Tvåpunktsfördelning (Bernoullifördelning)

Den stokastiska variabeln X kan anta två värden: a och b. Vi kallar X för **tvåpunktsfördelad**, $X \sim \mathrm{Be}(p)$, om $p_X(a) = p$ och $p_X(b) = 1 - p$.

Exempel

Slantsingling med osymetriskt mynt. Låt $X = -1$ vid krona och $X = 1$ vi klave. Till exempel kan vi ha $p_X(-1) = P(X = -1) = 0.4$ och $p_X(1) = P(X = 1) = 0.6$.

Vad händer om vi betraktar summan av oberoende Bernoullivariabler?

Exempel

En händelse har 30% sannolikhet. Vi upprepar försöket 10 gånger, oberoende av varandra. Vad blir sannolikheten att:

1. Händelsen inträffar exakt k gånger;
2. Händelsen inträffar högst 1 gång;
3. Händelsen inträffar minst 2 gånger.

Lösning: Låt X vara antalet gånger händelsen inträffar. Då är $X = 0, 1, \ldots, 10$ möjliga värden.

1. Enligt exemplet från föregående föreläsning (inbrottstjuven) måste

$$P(X = k) = \binom{10}{k} 0.3^k \cdot 0.7^{10-k}, \qquad k = 0, 1, \ldots, 10.$$

   Observera att $P(X = k) = p_X(k)$, så uttrycket ovan är sannolikhetsfunktionen för X.

2. $P(X \leq 1) = \sum_{k=0}^{1} p_X(k) = \binom{10}{0} 0.3^0 \cdot 0.7^{10} + \binom{10}{1} 0.3^1 \cdot 0.7^9 \approx 0.149.$

3. $P(X \geq 2) = 1 - P(X < 2) = 1 - P(X \leq 1) \approx 0.851.$

Vi säger att X är binomialfördelad med parametrarna $n = 10$ och $p = 0.3$.

Binomialfördelning

Vi kallar X binomialfördelad med parametrarna n och p om X har sannolikhetsfunktionen

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n,$$

och vi skriver $X \sim \text{Bin}(n, p)$.

Antag att vi har en större mängd med N element där $N = v + s$ består av två olika sorters element. Låt $p = v/N$ vara andelen v-märkta element. Om vi på måfå plockar ut n stycken element från N, hur många är v-märkta? Svaret kommer i form av den Hypergeometriska fördelningen.

Hypergeometrisk fördelning

Vi skriver $X \sim \text{Hyp}(N, n, p)$ om

$$p_X(k) = \frac{\binom{Np}{k}\binom{N(1-p)}{n-k}}{\binom{N}{n}}, \quad k = 0, 1, 2, \ldots, Np.$$

Vi kallar X för **Hypergeometriskt** fördelad.

Varför blir det så? Det är bara multiplikationsprincipen *in action*. Vi väljer k stycken av de Np v-märkta kulorna och $n - k$ stycken av de $N(1-p)$ s-märkta kulorna (vilket ger de gynsamma utfallen). Totalt sett väljer vi n stycken kulor från de N som finns (det totala antalet). Vi räknar allt utan ordning och använder den klassiska definitionen av sannolikhet.

Exempel

Ur en grupp bestående av 100 studenter väljer vi 20 på måfå. Den stora gruppen består till 60% av kvinnor. Vad är sannolikheten att vi har precis elva kvinnor i den mindre gruppen?

Lösning: Låt X vara antalet kvinnor i den mindre mängden. Det följer att $X \sim \text{Hyp}(100, 20, 0.6)$, så sannolikheten vi söker kan beräknas enligt

$$P(X = 11) = p_X(11) = \frac{\binom{60}{11}\binom{40}{9}}{\binom{100}{20}} \approx 0.175.$$

Likformig (rektangel-) fördelning

Om X antar ändligt många värden, säg $X \in E = \{1, 2, \ldots, m\}$, och vi definierar $p_X(k) = \dfrac{1}{m}$ för varje $k = 1, 2, \ldots, m$, så kallar vi X för **likformigt** fördelad (på mängden E).

Exempel

Låt $\Omega = \{0, 1, 2, \ldots, 9\}$ och låt X vara likformigt fördelad på Ω. Vidare, låt $Y(x) = 0$ om x är jämnt delbart med 3, annars är $Y = 1$. Bestäm p_X och p_Y.

Lösning: Vi har $p_X(k) = 1/10$ om $k = 0, 1, 2, \ldots, 9$ och $p_X(k) = 0$ annars. Vi låter mängden $A = \{0, 3, 6, 9\}$ bestå av de tal som är delbara med 3 och $B = \{1, 2, 4, 5, 7, 8\}$ de som inte är delbara med tre. Klassiska definitionen på sannolikhet ger $P(A) = 4/10$ och $P(B) = 6/10$. Variabeln Y blir Bernoullifördelad med $p = 2/5$ (med $a = 0$ och $b = 1$).

För-första-gången-fördelning

Vi skriver $X \sim \text{Ffg}(p)$ om $p_X(k) = (1 - p)^{k-1}p, \quad k = 1, 2, \ldots$. Vi kallar X för **För-första-gången**-fördelad.

Ett slumpförsök har två olika utfall, säg A och B, med sannolikheterna p respektive $1 - p$. Vi upprepar försöker oberoende tills dess att händelsen A inträffar för första gången. Antalet försök X till och med att A inträffar för första gången är Ffg(p)-fördelad. Om $X = k$ innebär det att A inträffade för första gången vid den k:te upprepningen, och att i de $k - 1$ första försöken inträffade B. Alltså måste $P(X = k) = P(B)^{k-1}P(A) = (1 - p)^{k-1}p$ eftersom försöken är oberoende.

Exempel

Låt oss kasta en 6-sidig tärning tills dess att vi för första gången får en 1:a eller 3:a. Låt X vara antalet kast. Händelsen att få en 1:a eller 3:a vid ett kast är $p = 2/6 = 1/3$ (gynsamma/möjliga). Då blir alltså $X \sim \text{Ffg}(1/3)$-fördelad. Vad är sannolikheten att det tar fyra eller fler kast innan vi får en 1:a eller 3:a för första gången?

Lösning: Som bekant är $X \sim \text{Ffg}(1/3)$, så

$$P(X \geq 4) = 1 - \sum_{k=1}^{3} P(X = k) = 1 - \left(\frac{1}{3} + \frac{2}{3} \cdot \frac{1}{3} + \left(\frac{2}{3}\right)^2 \cdot \frac{1}{3}\right) = \frac{8}{27}.$$

Alternativt,

$$P(X \geq 4) = \sum_{k=4}^{\infty} P(X = k) = \frac{1}{3} \sum_{k=4}^{\infty} \left(\frac{2}{3}\right)^{k-1} = \frac{2^3}{3^4} \sum_{k=0}^{\infty} \left(\frac{2}{3}\right)^{k} = \frac{2^3}{3^4} \cdot \frac{1}{1 - 2/3} = \frac{8}{27}.$$

En nära besläktad fördelning är den geometriska. Vi kan tänka oss att vi räknar antalet misslyckade försök innan en händelse inträffar för första gången.

Geometrisk fördelning

Vi skriver $X \sim \text{Geo}(p)$ om $p_X(k) = (1-p)^k p, \quad k = 0, 1, 2, \ldots$. Vi kallar X för **Geometriskt** fördelad.

En annan diskret fördelning vi kommer att stöta på framöver är Poissonfördelningen.

Poissonfördelning

Vi skriver $X \sim \text{Po}(\mu)$ om $p_X(k) = \dfrac{\mu^k}{k!} e^{-\mu}, \quad k = 0, 1, 2, \ldots$ och $\mu > 0$. Vi kallar X för **Poisson**-fördelad.

Försök visa att detta faktiskt är en sannolikhetsfunktion (vad måste gälla?).
För vissa fördelningar och parametervärden har vi tabeller av sannolikheter att tillgå, speciellt för Poisson- och Binomialfördelning. Studera formelsamlingen! Se till att ni lär er känna igen och skilja de olika fördelningarna åt. De flesta kommer dyka upp i andra sammanhang och andra kurser senare i utbildningen.