SF1901: Sannolikhetslära och statistik
Föreläsning 3.
Stokastiska variabler, diskreta och kontinuerliga

Jan Grandell & Timo Koski

25.01.2016

KTH Matematik

Stokastiska variabler: inledning

I nästan alla situationer som vi betraktar, kommer resultaten av slumpförsöken att vara tal, kontinerliga mätvärden eller antal. Det är praktiskt att anpassa beteckningarna till detta.

Definition

En stokastisk variabel s.v. (eller en slumpvariabel) X är en funktion från Ω till reella linjen.

Lite löst kommer vi att uppfatta X som en beteckning för resultatet av ett slumpförsök.

KTH Matematik

Lärandemål: X st.v.

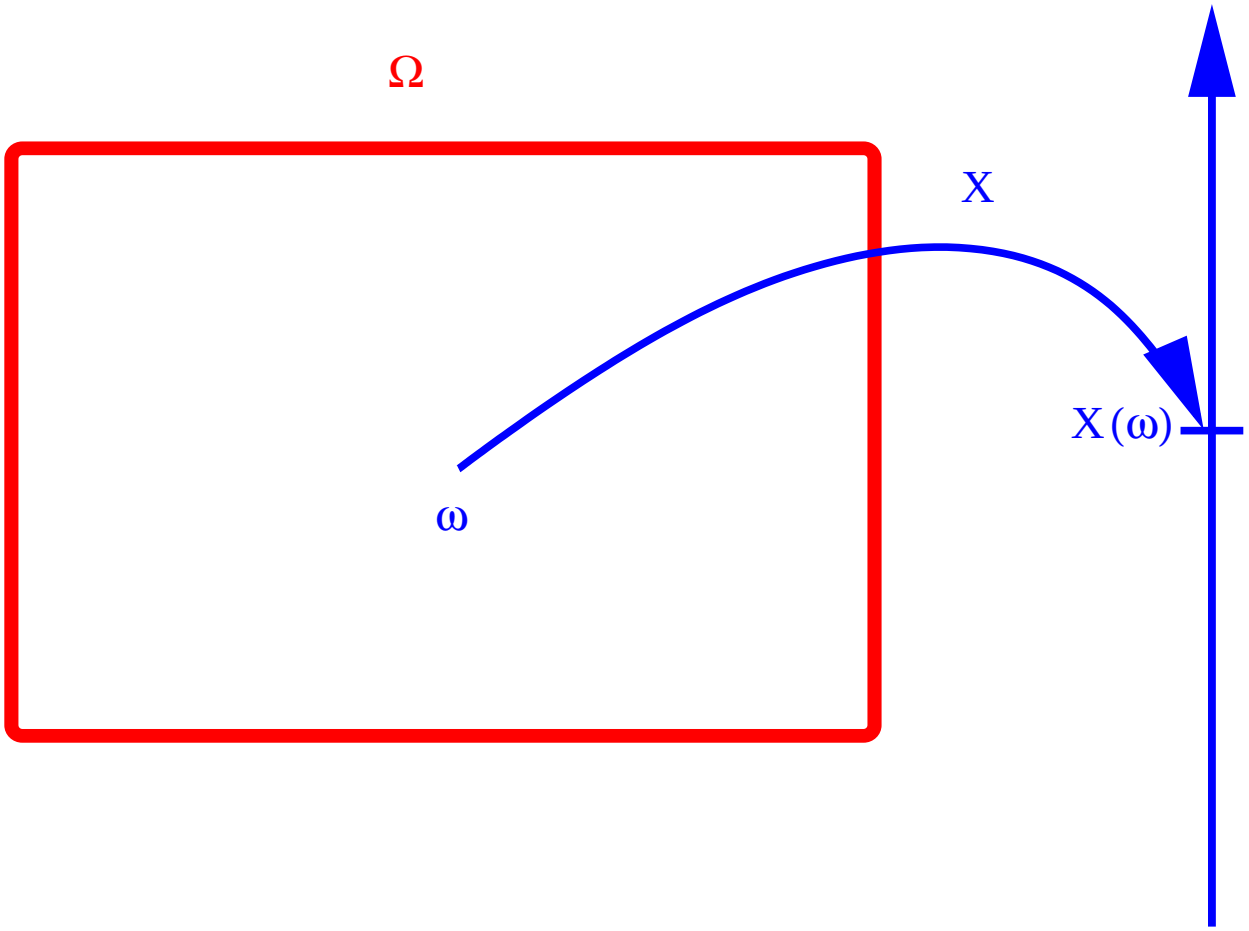
- Fördelningsfunktion $F_X(x)$;

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a)$$

- diskreta st.v:er, $F_X(x) = \sum_{j \leq [x]} p_X(j)$,
  $p_X(k) = P(X = k)$,sannolikhetsfunktion.
- kontinuerliga st.ver $F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$. Funktionen $f_X(x)$ kallas *täthetsfunktionen* för X. $f_X(x) = F_X'(x)$.
- binomialfördelning, poissonfördelning, exponentialfördelning,

KTH VETENSKAP OCH KONST
KTH Matematik

Stokastiska variabler



Stokastiska variabler: inledning

Låt X vara en stokastisk variabel. Det mest allmänna sättet att beskriva X, dvs. hur X varierar, är att ange dess fördelningsfunktion.

Definition

Fördelningsfunktionen $F_X(x)$ till en s.v. X definieras av

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

KTH VETENSKAP OCH KONST
KTH Matematik

Stokastiska variabler: inledning

En fördelningsfunktion $F_X(x)$ har följande egenskaper:

1) $F_X(x)$ är icke-avtagande[1] ;
2) $F_X(x) \to 1$ då $x \to \infty$;
3) $F_X(x) \to 0$ då $x \to -\infty$;
4) $F_X(x)$ är högerkontinuerlig.

[1] d.v.s. $x_1 \leq x_2 \Rightarrow F_X(x_1) \leq F_X(x_2)$

KTH Matematik

Beräkning av sannolikheter

Låt X vara en stokastisk variabel.

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

Då fås

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a)$$

Sats

$$P(a < X \leq b) = F_X(b) - F_X(a)$$

KTH Matematik

Stokastiska variabler: **Diskret stokastisk variabel**

Det är lämpligt att skilja på fallen då vår stokastiska variabel representerar kontinuerliga mätvärden eller antal.
Vi ska nu betrakta fallet med antal.

Definition

En s.v. X säges vara diskret om den kan anta ett ändligt eller uppräkneligt oändligt antal olika värden.

Det viktiga är att de möjliga värdena ligger i en ändlig eller högst uppräknelig mängd. Oftast tar en diskret s.v. icke-negativa heltalsvärden "räknar ett antal". Vi kommer att förutsätta detta, om vi inte explicit säger något annat.

KTH Matematik

Definition

För en diskret s.v. definieras sannolikhetsfunktionen $p_X(k)$ *av*

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

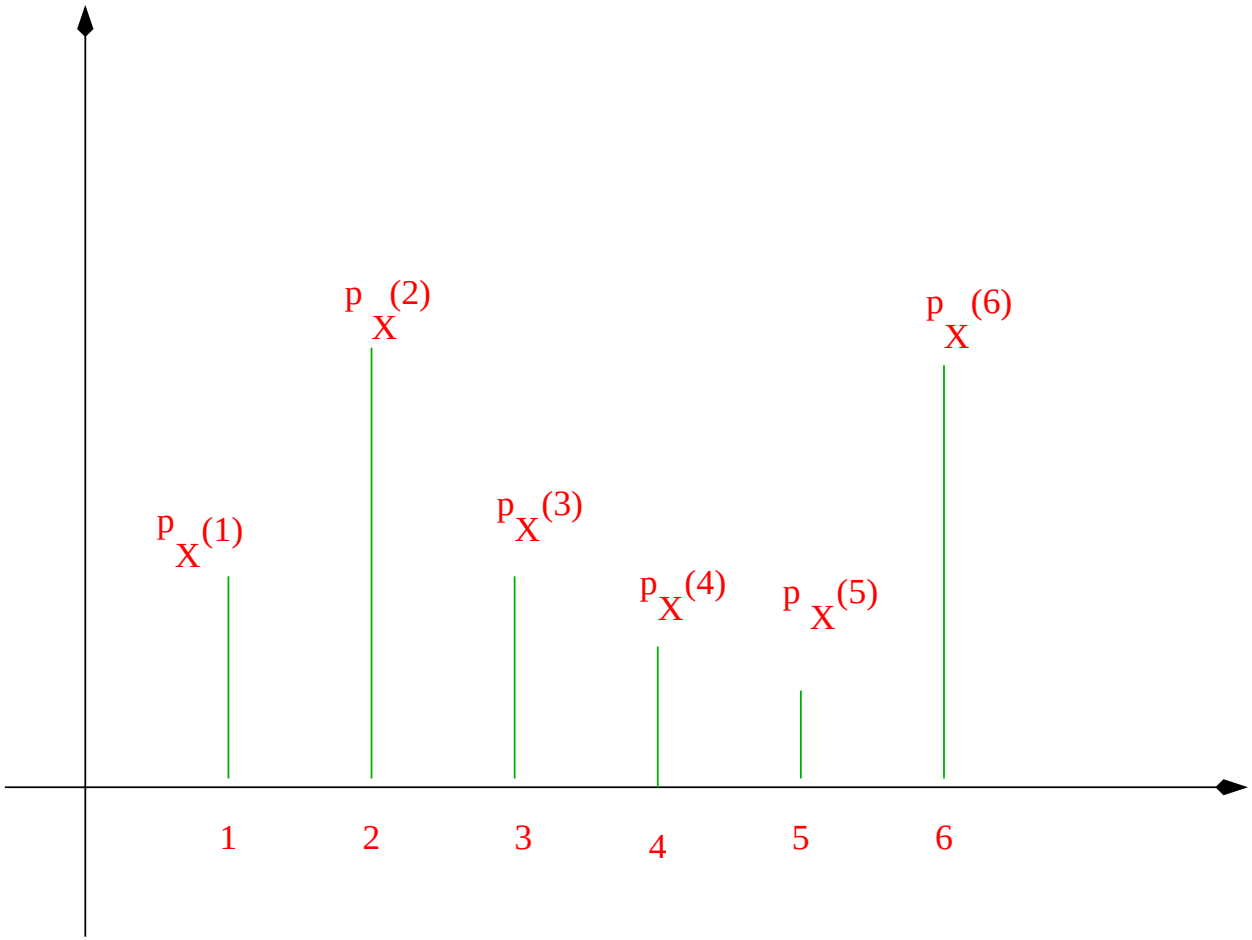
Om X beskriver ett tärningskast gäller således

$$p_X(k) = \begin{cases} \frac{1}{6} & \text{för } k = 1, 2, 3, 4, 5, 6 \\ 0 & \text{för övriga värden på } k. \end{cases}$$

Gör vi nu slumpförsöket att på måfå dra en av 6 lappar med talen 1, 2, 3, 4, 5 eller 6, så får vi samma s.v. som i tärningskasten.

KTH VETENSKAP OCH KONST
KTH Matematik

En sannolikhetsfunktion



Diskreta stokastiska variabler

Relationen mellan sannolikhetsfunktionen och fördelningsfunktionen för en diskret stokastisk variabel fås av sambanden

$$F_X(x) = \sum_{j \leq [x]} p_X(j), \quad \text{där } [x] \text{ betyder heltalsdelen av } x,$$

och

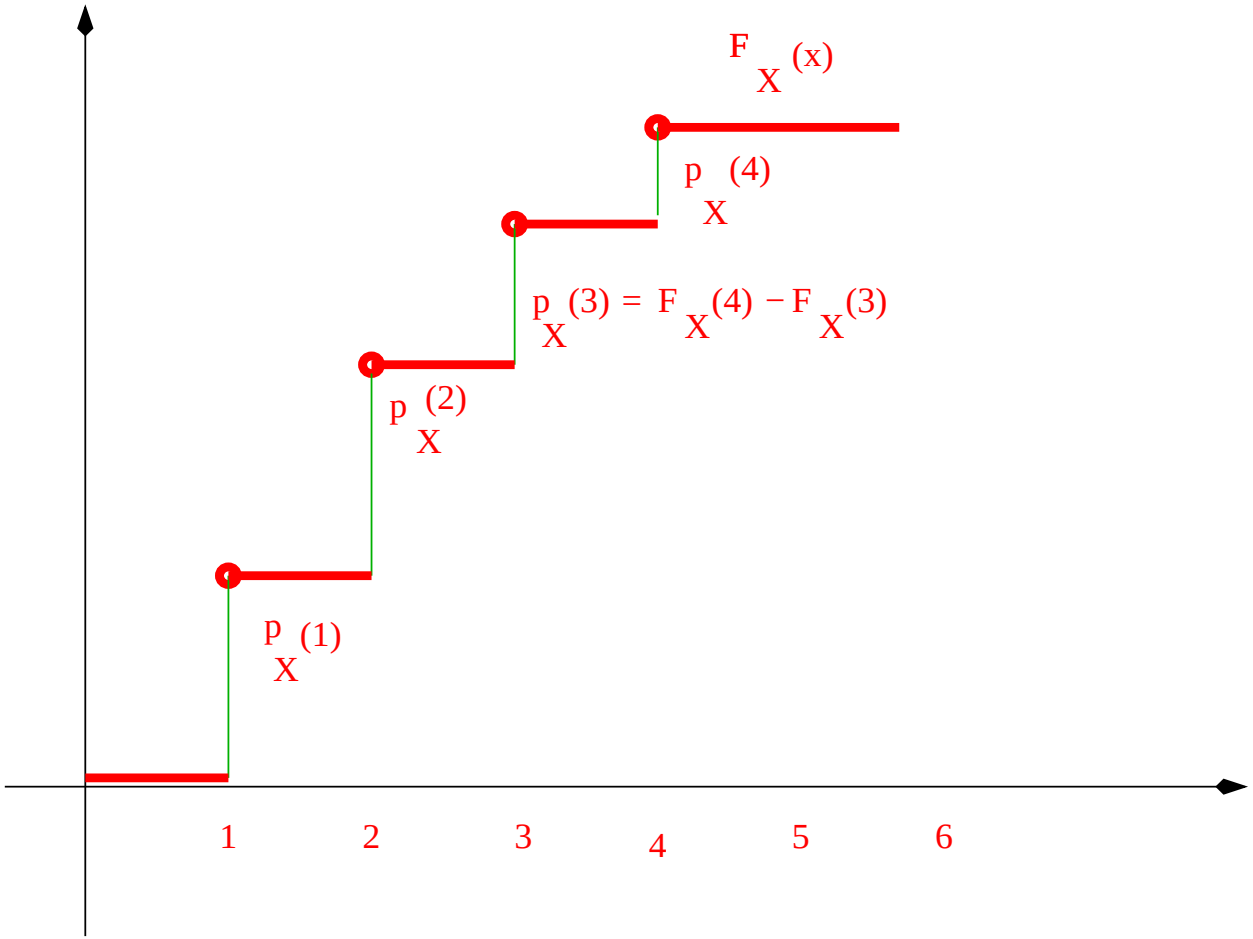
$$p_X(k) = F_X(k) - F_X(k-1).$$

Det följer av detta att

$$p_X(k) \geq 0 \quad \text{och} \quad \sum_0^\infty p_X(k) = 1.$$

KTH Matematik

En fördelningsfunktion



obs Ej samma sannolikheter som i föregående figur om sannolikhetsfunktion

Benfords lag

Benfords lag beskriver hur olika siffror är statistiskt fördelade som förstasiffror i olika material och ges av sannolikhetsfunktionen

$$p_X(k) = P(X = k) = \log_{10}\left(1 + \frac{1}{k}\right), \quad k = 1, \ldots, 9.$$

En formell check (av att $p_X(k)$ är en sannolikhetsfunktion):

- $p_X(k) \geq 0$, ty $1 + \frac{1}{k} \geq 1$.

KTH Matematik

Benfords lag

$$\sum_{k=1}^{9} p_X(k) = \sum_{k=1}^{9} \log_{10}\left(1+\frac{1}{k}\right)$$

$$= \sum_{k=1}^{9} \log_{10}\left(\frac{k+1}{k}\right) = \sum_{k=1}^{9} \left[\log_{10}(k+1) - \log_{10}(k)\right] =$$

$$= \log_{10}(2) - \log_{10}(1) + \log_{10}(3) - \log_{10}(2)$$

$$+ \log_{10}(4) - \log_{10}(3) + \log_{10}(5) - \log_{10}(4)$$

$$+ \log_{10}(6) - \log_{10}(5) + \log_{10}(7) - \log_{10}(6)$$

$$+ \log_{10}(8) - \log_{10}(7) + \log_{10}(9) - \log_{10}(8)$$

$$+ \log_{10}(10) - \log_{10}(9) = -\log_{10}(1) + \log_{10}(10)$$

$$= 0 + 1 = 1.$$

KTH Matematik

Benfords lag

Benfords lag ger fördelningen för olika siffror som förstasiffror. Lagen säger till exempel att siffran 1 bör vara frstasiffra i 30,1% av fallen, siffran 2 i 17,6% av fallen och siffran 9 i 4,6% av fallen i en mycket stor datamängd. Benfords lag visar sig tillämplig inom många skilda områden. Exempel är lagen tillämplig vid ekonomisk redovisning, prislistor, antal röster vid omröstningar mellan ett stort antal alternativ, samt folkmängd i städer. Om en stor datamängd inte följer Benfords lag är det en indikation på att siffrorna kan vara påhittade eller manipulerade.

KTH Matematik

Benfords lag

Till exempel de tal som förekommer i en inkomstdeklaration, borde mer eller mindre följa Benfords lag. Om talen inte uppför sig enligt Benfords ordination kan det vara något skumt i görningen. Antingen har någon manipulerat siffrorna med flit, till exempel en skattesmitare, eller så är det något annat som har gått snett.
För få tal som börjar med en etta räcker som en varningssignal. Då bör myndigheterna eller företaget gå igenom siffrorna en extra gång. I Sverige använder till exempel företaget Trelleborg AB Benfords lag i sin internrevision fr att avslöja bedrägerier.

KTH Matematik

Beräkning av sannolikheter

Sats

$$P(a < X \leq b) = F_X(b) - F_X(a)$$

Om X är diskret, så fås, om a och b är positiva heltal, att

$$P(a < X \leq b) = F_X(b) - F_X(a) = \sum_{j \leq b} p_X(j) - \sum_{j \leq a} p_X(j)$$

$$= \sum_{j=a+1}^{b} p_X(j).$$

KTH Matematik

Beräkning av sannolikheter

Om X är diskret, och a och b är positiva heltal, gäller det att

$$P(a \leq X \leq b) = P(a < X \leq b) + P(X = a)$$

$$= F_X(b) - F_X(a) + P(X = a) = \sum_{j=a+1}^{b} p_X(j) + p_X(a)$$

$$= \sum_{j=a}^{b} p_X(j).$$

obs! Om X är diskret, och $p_X(a) > 0$, så gäller alltså enligt ovananförda att

$$P(a < X \leq b) \neq P(a \leq X \leq b)$$

Beräkning av sannolikheter

X är diskret, och b är ett positivt heltal,

$$P(X \geq b) = 1 - P(X < b) = 1 - P(X \leq b - 1)$$

$$= 1 - F_X(b-1) = 1 - \sum_{j \leq b-1} p_X(j)$$

$$P(X > b) = 1 - P(X \leq b) = 1 - F_X(b)$$

KTH Matematik

Beräkning av sannolikheter: allmänt

A en delmängd av den reella axeln, $P(X \in A)$ = sannolikheten för att X antar ett värde i A.

- Om X diskret:

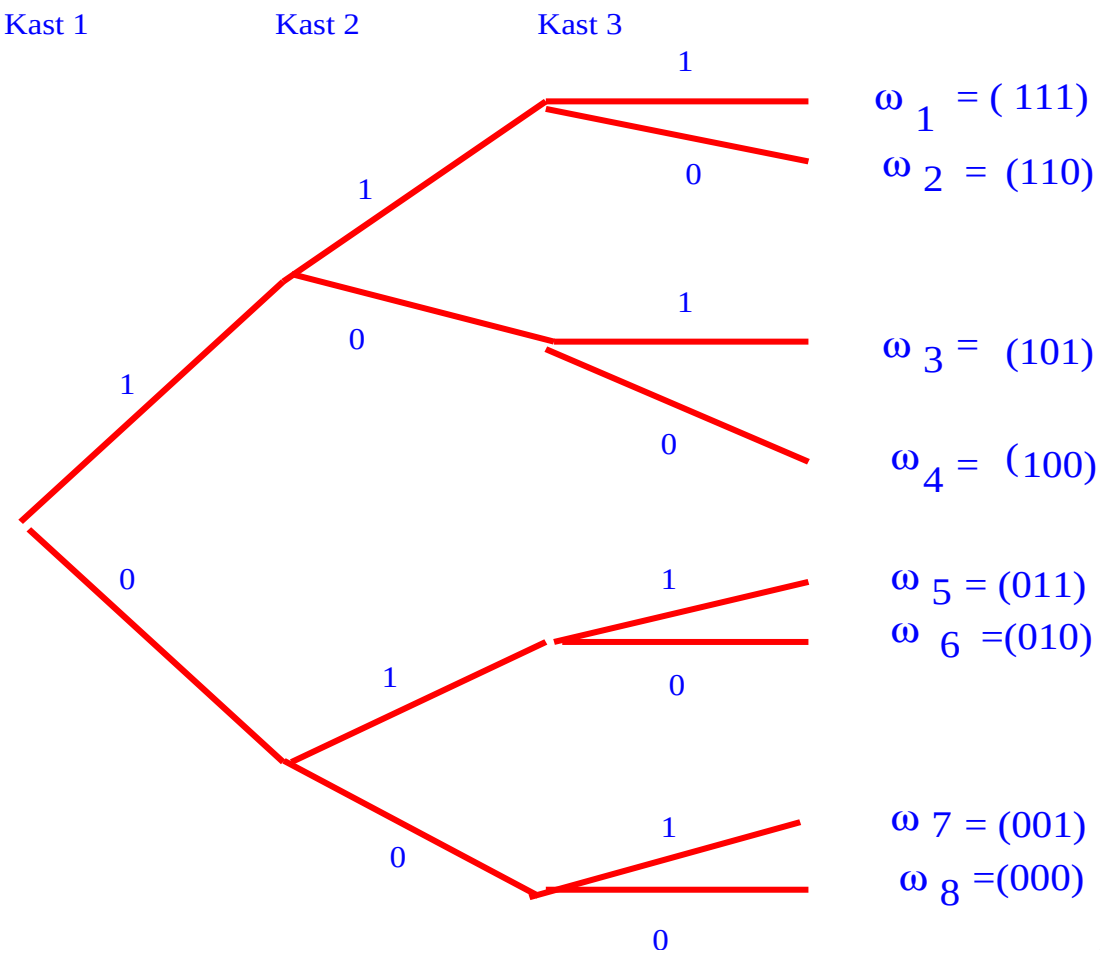
$$P(X \in A) = \sum_{j \in A} p_X(j)$$

Kast av ett häftstift (= Thumbtack på (amerikansk) engelska)

Slumpexperiment: Kast av ett häftstift. Om det landar på spetsen som i bilden ovan, säger vi att en 'etta' (1) inträffar. Vi säger att en nolla (0) inträffar om häftstiftet landar på hatten.

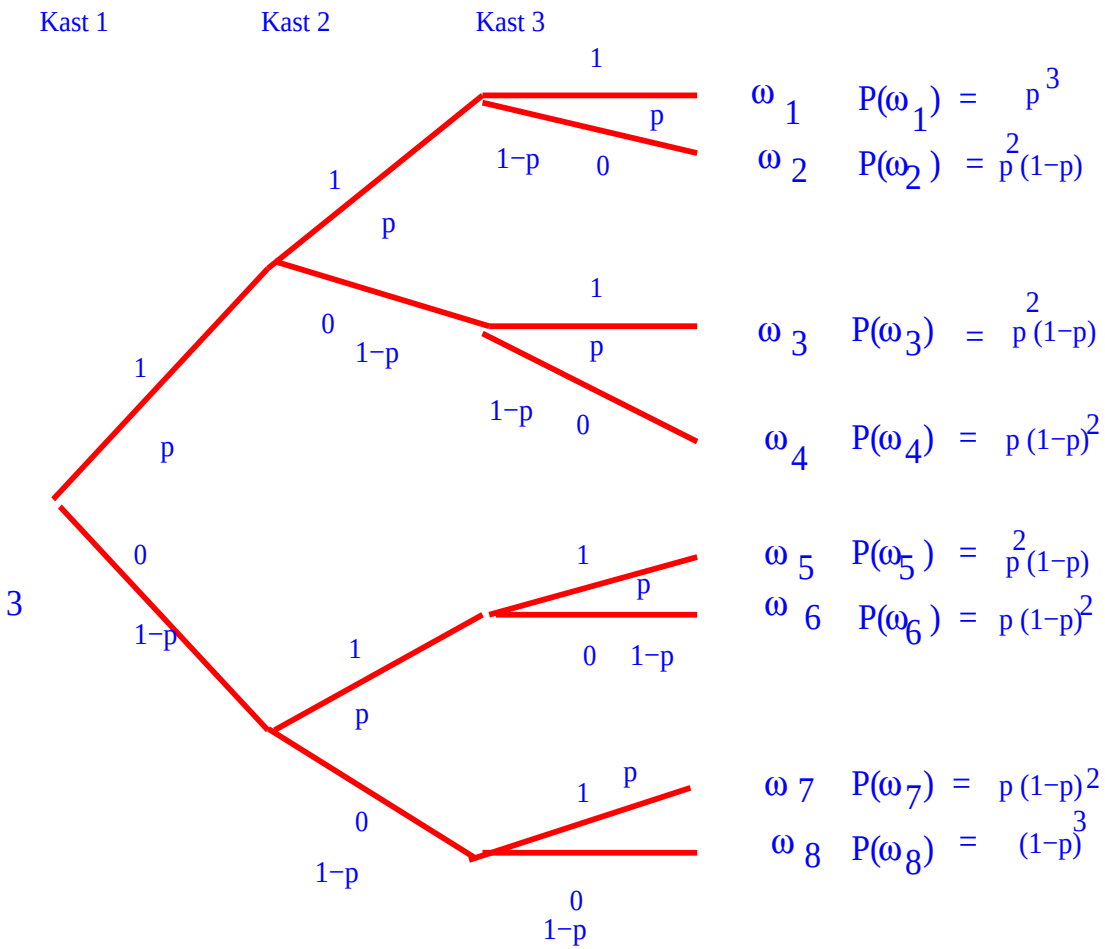
KTH Matematik

Tre kast av ett häftstift: utfallsrummet som ett träddiagram



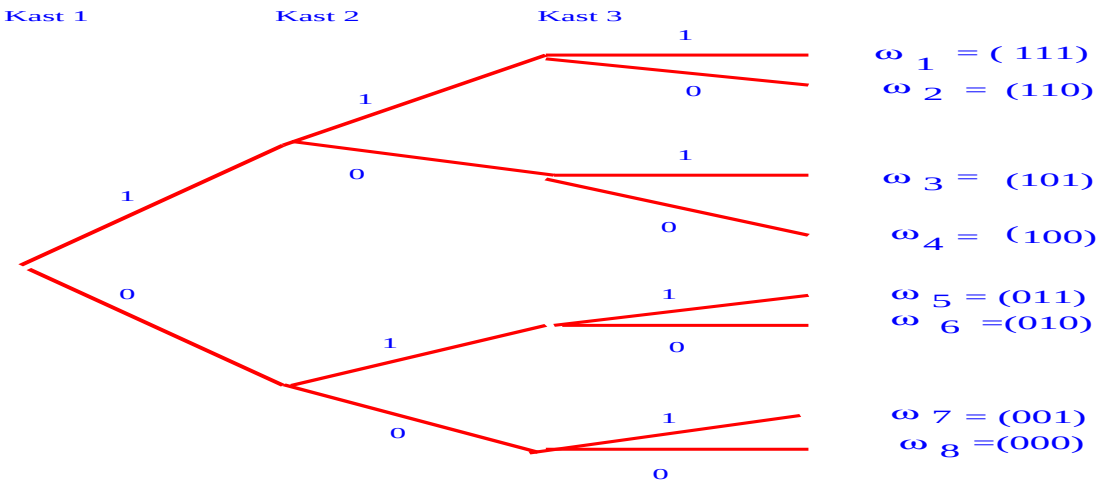
Antalet utfall$= 2^3$

Tre oberoende kast av ett häftstift: sannolikheter



$$P(\text{en nolla } (0)) = 1 - p, P(\text{en etta } (1)) = p.$$

Tre oberoende kast av ett häftstift: en stokastisk variabel



X = 'antalet ettor i tre oberoende kast av ett häftstift'.

$X(\omega_1) = 3, X(\omega_2) = X(\omega_3) = X(\omega_5) = 2,$

$X(\omega_4) = X(\omega_6) = X(\omega_7) = 1,\ X(\omega_8) = 0.$

KTH VETENSKAP OCH KONST
KTH Matematik

Tre oberoende kast av ett häftstift: sannolikhetsfunktion för antalet ettor

$$X(\omega_1) = 3, X(\omega_2) = X(\omega_3) = X(\omega_5) = 2,$$
$$X(\omega_4) = X(\omega_6) = X(\omega_7) = 1,\ X(\omega_8) = 0.$$

$$P(X = 3) = P(\omega_1) = p^3,$$

$$P(X = 2) = P(\omega_2) + P(\omega_3) + P(\omega_5) = 3p^2(1-p)$$

$$P(X = 1) = P(\omega_4) + P(\omega_6) + P(\omega_7) = 3p(1-p)^2$$

$$P(X = 0) = P(\omega_8) = (1-p)^3$$

KTH Matematik

Tre oberoende kast av ett häftstift: sannolikhetsfunktion för antalet ettor

Vi omskriver dessa med hjälp av binomialkoefficienterna:

$$P(X=3) = p^3 = \binom{3}{3} p^3 (1-p)^0$$

$$P(X=2) = 3p^2(1-p) = \binom{3}{2} p^2 (1-p)$$

$$P(X=1) = 3p(1-p)^2 = \binom{3}{1} p(1-p)^2$$

$$P(X=0) = \binom{3}{0} (1-p)^3$$

eller med en enda formel

$$P(X=k) = \binom{3}{k} p^k (1-p)^{3-k}, \quad , k = 0, 1, 2, 3.$$

Denna är sannolikhetsfunktionen för binomialfördelningen med parametrarna 3 och p.

KTH Matematik

Diskreta st. v:er: Binomialfördelningen

Definition

En diskret s.v. X säges vara binomialfördelad med parametrarna n och p, $\mathrm{Bin}(n,p)$*-fördelad, om*

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

Vi skriver detta med $X \in \mathrm{Bin}(n,p)$.

KTH VETENSKAP OCH KONST
KTH Matematik

Diskreta st. v:er: Binomialfördelningen

De generella villkoren för detta:

- *n oberoende upprepningar av ett försök.*
- *varje försök har två utfall, 0 och 1.*
- *sannolikheten för lyckat försök (=1) är densamma = p vid varje försök.*

X = antalet lyckade försök.

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

Vi skriver detta med $X \in \text{Bin}(n, p)$.

Diskreta st. v:er: Poissonfördelningen

Ofta när det är rimligt att anta att en s.v. X är Bin(n,p)-fördelad, så är det även rimligt att anta att p är liten och att n är stor. Låt oss anta att $p = \mu/n$, där n är "stor" men μ är "lagom". Då gäller

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n(n-1)\dots(n-k+1)}{k!} \left(\frac{\mu}{n}\right)^k \left(1-\frac{\mu}{n}\right)^{n-k}$$

$$= \frac{\mu^k}{k!} \underbrace{\left(1-\frac{\mu}{n}\right)^n}_{\approx e^{-\mu}} \underbrace{\frac{n(n-1)\dots(n-k+1)}{n^k}}_{\approx 1} \underbrace{\left(1-\frac{\mu}{n}\right)^{-k}}_{\approx 1} \approx \frac{\mu^k}{k!} e^{-\mu}.$$

Diskreta st. v:er: Poissonfördelningen

Definition

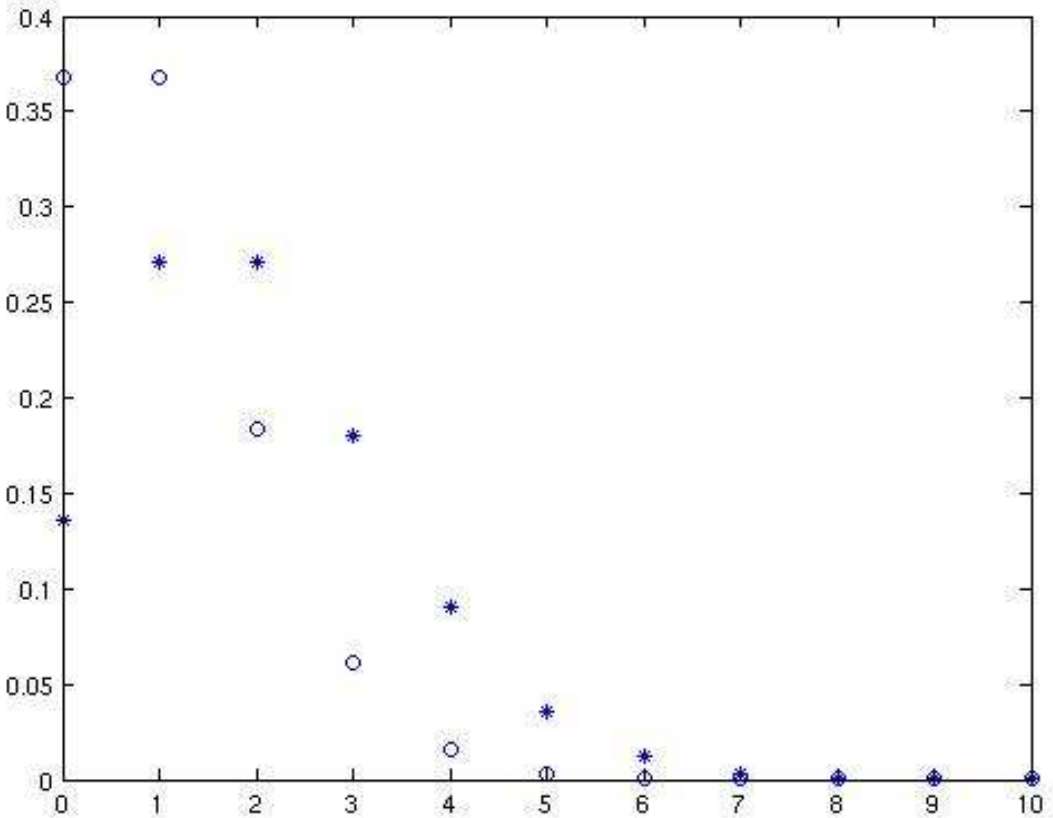
En diskret s.v. X säges vara Poissonfördelad med parameter μ, $\mathrm{Po}(\mu)$-fördelad, om

$$p_X(k) = \frac{\mu^k}{k!} e^{-\mu}, \text{ för } k = 0, 1, 2 \ldots .$$

Vi skriver detta med $X \in \mathrm{Po}(\mu)$.

KTH VETENSKAP OCH KONST
KTH Matematik

Sannolikhetsfunktionerna för Po(2) och Po(1)



$* \leftrightarrow \mathrm{Po}(2)$, $\circ \leftrightarrow \mathrm{Po}(1)$

KTH VETENSKAP OCH KONST
KTH Matematik

En urnmodell: Dragning utan återläggning

I en urna finns kulor av två slag: v vita och s svarta. Drag n kulor ur urnan slumpmässigt och så att en kula som dragits inte stoppas tillbaka. dvs dragning utan återläggning.
Sätt $X =$ "Man får k vita kulor i urvalet".
Välj Ω: Alla uppsättningar om n kulor utan hänsyn till ordning.
Då fås:

$$m = \binom{v+s}{n} \quad \text{och} \quad g = \binom{v}{k}\binom{s}{n-k}$$

och således

$$P(X = k) = \frac{\binom{v}{k}\binom{s}{n-k}}{\binom{v+s}{n}}.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Hypergeometrisk fördelning, Hyp(N, n, p)

Definition

Om den st.v. X har sannolikhetsfunktionen

$$p_X(k) = \frac{\binom{v}{k}\binom{s}{n-k}}{\binom{v+s}{n}}.$$

där k har värden sådana att $0 \leq k \leq v$, $0 \leq n-k \leq s$, så säges X vara hypergeometriskt fördelad.

Vi sätter $N = v + s$ och $p = v/(v+s)$ och skriver $X \in \text{Hyp}(N, n, p)$.

KTH VETENSKAP OCH KONST
KTH Matematik

Kontinuerlig stokastisk variabel

Här kan vi tyvärr inte ge definitionen i termer av den stokastiska variabeln själv. Det räcker inte att säga att X kan ta ett överuppräkneligt antal värden. Vi får därför ge definitionen i termer av fördelningsfunktionen, som ju är den allmännaste beskrivningen av en s.v.

Definition

En s.v. X säges vara kontinuerlig om dess fördelningsfunktion har framställningen

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

för någon funktion $f_X(x)$. Funktionen $f_X(x)$ kallas täthetsfunktionen för X.

Omvänt gäller att $f_X(x) = F_X'(x)$.

KTH VETENSKAP OCH KONST
KTH Matematik

Kontinuerlig stokastisk variabel

Täthetsfunktionen kan inte direkt tolkas som en sannolikhet, men vi har, för små värden på h,

$$P(x < X \leq x + h) = F_X(x + h) - F_X(x) = \int_x^{x+h} f_X(t)\,dt \approx h\, f_X(x).$$

Beräkning av sannolikheter

Låt X vara en kontinuerlig stokastisk variabel. $a < b$ är reella tal,

$$P(a < X \leq b) = F_X(b) - F_X(a) = \int_{-\infty}^{b} f_X(t)\,dt - \int_{-\infty}^{a} f_X(t)\,dt$$

$$= \int_{a}^{b} f_X(t)\,dt$$

Beräkning av sannolikheter

Låt X vara en kontinuerlig stokastisk variabel. $a < b$ är reella tal

$$P(a \leq X \leq b) = F_X(b) - F_X(a) + P(X = a)$$

Men för $h > 0$

$$P(X = a) = \lim_{h \to 0} P(a - h < X \leq a + h) = \lim_{h \to 0} (F_X(a + h) - F_X(a - h)) =$$

ty $F_X(x)$ är kontinuerlig (och därmed högerkontinuerlig).

KTH Matematik

Beräkning av sannolikheter

Med andra ord, om X är en kontinuerlig stokastisk variabel och $a < b$ är reella tal,

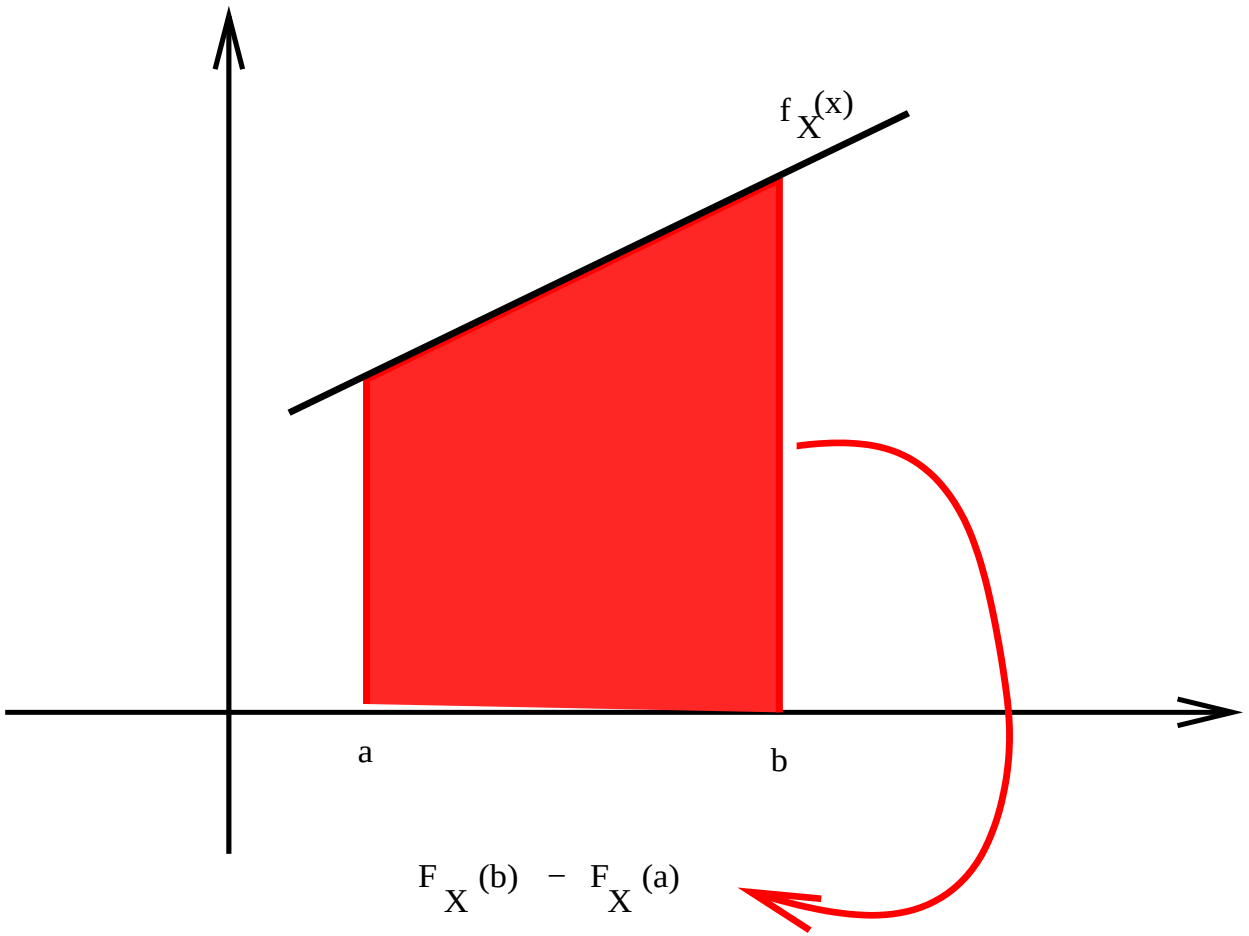
Sats

$$P(a \leq X \leq b) = P(a < X \leq b) = \int_a^b f_X(t)\,dt.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Beräkning av sannolikheter: arean under kruvan

$$P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(t)\,dt.$$



Beräkning av sannolikheter

Om X är en kontinuerlig stokastisk variabel och b är ett reellt tal,

$$P(X > b) = P(X \geq b) = 1 - P(X < b) = 1 - P(X \leq b)$$

$$= 1 - F_X(b) = \int_{-\infty}^{\infty} f_X(t)\,dt - \int_{\infty}^{b} f_X(t)\,dt$$

$$= \int_{b}^{\infty} f_X(t)\,dt.$$

dvs.

$$P(X > b) = P(X \geq b) = \int_{b}^{\infty} f_X(t)\,dt.$$

Likformig fördelning $U(a, b)$

$X \in U(a, b)$

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

$$F_X(x) = \begin{cases} 0 & \text{för } x \leq a, \\ \frac{x-a}{b-a} & \text{för } a \leq x \leq b, \\ 1 & \text{för } x \geq b. \end{cases}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Exponentialfördelningen $\mathrm{Exp}(\lambda)$

$X \in \mathrm{Exp}(\lambda)$

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

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & \text{för } x \geq 0, \\ 0 & \text{för } x < 0. \end{cases}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Exponentialfördelningen Exp(λ)

Denna fördelning är viktig i väntetidsproblem. För att inse detta så tar vi ett enkelt exempel:
Antag att n personer går förbi en affär per tidsenhet. Låt var och en av dessa gå in i affären oberoende av varandra och med sannolikheten p. Låt X vara tiden tills första kunden kommer. $X > x$ betyder att ingen kund kommit efter x tidsenheter.
$P(X > x) = (1-p)^{nx}$ ty nx personer har gått förbi.
Låt oss anta precis som då vi "härledde" Poissonfördelningen, att $p = \mu/n$, där n är "stor" men μ är "lagom". Då gäller

$$P(X > x) = (1-p)^{nx} = (1 - \frac{\mu}{n})^{nx} \approx e^{-\mu x}.$$

Detta ger att $F_X(x) = 1 - P(X > x) \approx 1 - e^{-\mu x}$, dvs X är approximativt Exp(μ). Observera att väntevärdet (ännu ej definierat, men det kommer) är $1/\mu$!

Kvantil, median

Ett par begrepp:

Definition

Lösningen till ekvationen $1 - F_X(x) = \alpha$ *kallas* α*-kvantilen till* X *och betecknas med* x_α.

$x_{0.5}$ kallas för *medianen* och är således det värde som överskrides med samma sannolikhet som det underskrides.

KTH VETENSKAP OCH KONST
KTH Matematik