

Föreläsning 3

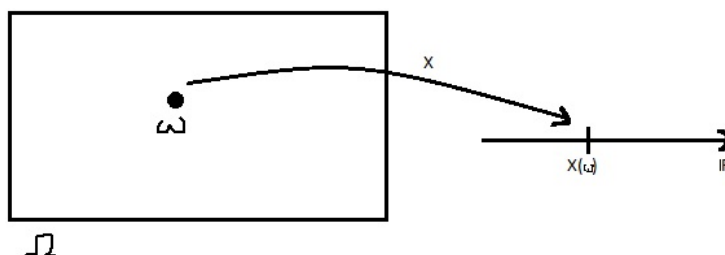
Thomas Önskog

1/11 2017

Stokastiska variabler

Vi inleder dagens föreläsning med en definition.

Definition. En *stokastisk variabel* (även kallad slumpvariabel) X är en funktion från Ω till den reella tallinjen $\mathbb{R}$.



Av utrymmesskäl kommer vi härnäst att använda beteckningen s.v. för stokastiska variabler. En s.v. kopplar alltså ett tal $X(\omega)$ till varje utfall ω av ett slutförsök. Talet $X(\omega)$ skrivs oftast kortfattat som X och detta är anledningen till att X benämns stokastisk variabel och inte stokastisk funktion. Stokastiska variabler kan vara diskreta eller kontinuerliga.

Exempel. För en enkätundersökning där godtyckligt valda personer får ange sin födelsemånad kan vi definiera en diskret s.v. X där varje utfall "januari", "februari", "mars", osv. motsvarar ett värde 1, 2, 3, osv. på X . Om vi mäter livslängden hos glödlampor kan vi definiera en kontinuerlig s.v. X för livslängden som antar värden på intervallet $(0, \infty)$.

I det andra exemplet ovan är utfallen ω redan numeriska värden och då är det naturligt att definiera X som $X(\omega) = \omega$, varvid X också antar numeriska värden, såsom önskat.

Definition. En s.v. X sägs vara **diskret** om den antar ändligt eller uppräkneligt oändligt antal olika värden. Värdeområdet kan skrivas $k_1, k_2, \dots$. För en diskret s.v. definieras **sannolikhetsfunktionen** p_X av

$$p_X(k) = \mathbb{P}(X = k), \quad \text{för } k = k_1, k_2, \dots$$

Sannolikhetsfunktionen beskriver hur sannolika de olika värdena på X är och kan åskådliggöras med ett stolpdigram med en stolpe för varje möjligt utfall. Oftast antar diskreta s.v. heltalsvärden och sannolikhetsfunktionen har då följande egenskaper

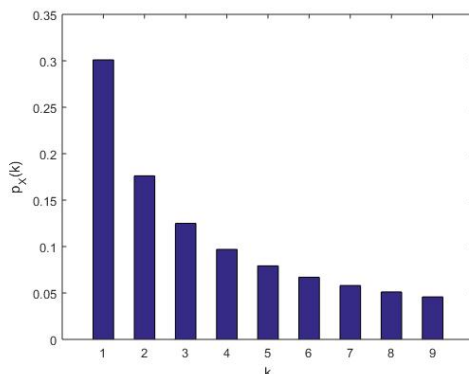
$$0 \leq p_X(k) \leq 1, \quad \text{för alla } k, \quad \sum_{k=-\infty}^{\infty} p_X(k) = 1,$$

$$\mathbb{P}(a \leq X \leq b) = \sum_{k=a}^b p_X(k), \quad \mathbb{P}(X \leq a) = \sum_{k=-\infty}^a p_X(k) := F_X(a),$$

Funktionen F_X kallas fördelningsfunktionen för X och kommer att studeras ytterligare på föreläsning 4.

Exempel. Benfords lag beskriver hur förstasiffrorna i många numeriska datamängder (exempelvis städers invånarantal, belopp i årsredovisning, fysikaliska konstanter) är fördelade. Lagen gäller särskilt om värdena varierar över många storleksordningar och kan användas för att testa om det är troligt att siffrorna i en datamängd har blivit manipulerade. Låt den s.v. X beteckna förstasiffran för ett tal i en datamängd som följer Benfords lag. Sannolikhetsfunktionen för X är då

$$p_X(k) = \log_{10} \left(\frac{k+1}{k} \right) = \log_{10}(k+1) - \log_{10}(k), \quad \text{för } k = 1, 2, \dots, 9.$$



Notera att $p_X(k) \geq 0$, ty $\log_{10}$ är en växande funktion, så att $\log_{10}(k+1) \geq \log_{10}(k)$ och att

$$\sum_{k=1}^9 p_X(k) = \sum_{k=1}^9 (\log_{10}(k+1) - \log_{10}(k)) = \log_{10}(10) - \log_{10}(1) = 1,$$

där vi i den andra likheten använt att summan är en teleskopsumma där alla termer utom första och sista termen tar ut varandra.

En kontinuerlig s.v. kan anta alla värden i ett intervall på $\mathbb{R}$. Utfallen ligger oändligt tätt, så inget enskilt utfall har positiv sannolikhet. Vi kan därför inte definiera sannolikhetsfunktionen som för diskreta s.v.

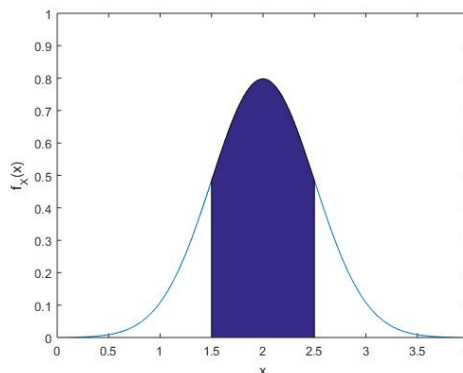
Definition. En s.v. X sägs vara **kontinuerlig** om det existerar en icke-negativ funktion f_X så att

$$\mathbb{P}(X \in A) = \int_A f_X(x) dx \quad \text{för alla mängder } A \in \mathbb{R}.$$

Funktionen f_X kallas för **täthetsfunktionen** för X . Notera att $\int_{-\infty}^{\infty} f_X(x) dx = 1$.

Sannolikheten för en händelse ges för en kontinuerlig s.v. av integration av täthetsfunktionen istället för summation av sannolikhetsfunktionen. Täthetsfunktionen kan inte direkt tolkas som en sannolikhet, men det gäller att

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x) dx,$$



så sannolikheten att $X \in [a, b]$ är lika med arean under kurvan $y = f_X(x)$ på intervallet $[a, b]$. Det spelar ingen roll om vi använder $<$ eller $\leq$ i ovanstående ekvation eftersom $\mathbb{P}(X = c) = 0$ för alla $c \in \mathbb{R}$ om X är en kontinuerlig s.v.

Diskreta sannolikhetsfördelningar

Vi ska nu undersöka några **diskreta sannolikhetsfördelningar**, dvs ofta förekommande klasser av sannolikhetsfunktioner.

Definition. En s.v. X sägs vara **likformigt fördelad** på $\{1, 2, \dots, m\}$ om $p_X(k) = 1/m$, för alla $k \in \{1, 2, \dots, m\}$.

Exempel. Antalet prickar vid ett träningskast är ett exempel på en s.v. som är likformigt fördelad på $\{1, 2, \dots, 6\}$.

Definition. En s.v. X sägs vara **tvåpunktsfördelad** om $p_X(a) = p$ och $p_X(b) = 1 - p$, där $p \in (0, 1)$ och $a, b \in \mathbb{R}$. Om $a = 1$ och $b = 0$, så sägs X vara **Bernoullifördelad**. Detta betecknas $X \in Be(p)$.

Exempel. Utfallet vid ett experiment som lyckas ($X = 1$) med sannolikhet p och annars misslyckas ($X = 0$) kan beskrivas med en Bernoullifördelad s.v.

Definition. En s.v. X sägs vara **för-första-gången-fördelad** om $p_X(k) = (1 - p)^{k-1}p$, där $p \in (0, 1)$ och $k = 1, 2, \dots$. Detta betecknas $X \in ff(p)$.

Exempel. Om ett experiment som lyckas med sannolikhet p upprepas ett obegränsat antal gånger, så är den gång i ordningen då det lyckas för första gången ff-fördelad.

Vi härleder nu den viktiga binomialfördelningen genom att ytterligare undersöka exemplet med ett försök som lyckas med sannolikhet p . Vi ska bestämma sannolikheten att om försöket upprepas n gånger, så lyckas det k av dessa gånger. Sannolikheten att försöket först lyckas k gånger och sedan misslyckas resterande $n - k$ gånger är $p^k(1 - p)^{n-k}$. Detta är även

sannolikheten för varje annan sekvens av n försök som innehåller exakt k lyckade försök. Totalt finns $\binom{n}{k}$ sådana sekvenser (vi kan välja de k omgångar då försöket lyckas på $\binom{n}{k}$ olika sätt, jfr. urnmodellen på föreläsning 2). Den totala sannolikheten för k lyckade försök är således $\binom{n}{k}p^k(1-p)^{n-k}$.

Definition. En s.v. X sägs vara **binomialfördelad** om

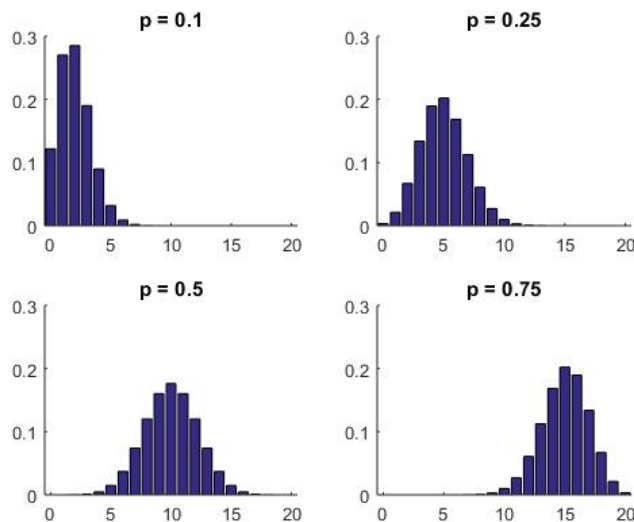
$$p_X(k) = \binom{n}{k}p^k(1-p)^{n-k},$$

där $p \in (0, 1)$ och $k = 0, 1, \dots, n$. Detta betecknas $X \in \text{Bin}(n, p)$.

Binomialsatsen säger att $(x + y)^n = \sum_{k=0}^n \binom{n}{k}x^k y^{n-k}$, så med $x = p$ och $y = 1 - p$ fås

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

som önskat. Bilden nedan visar sannolikhetsfunktionen $p_X(k)$ för fyra olika binomialfördelningar med $n = 20$, nämligen för $p = 0.1$, $p = 0.25$, $p = 0.5$ och $p = 0.75$. Som synes sker en förskjutning mot större värden på k när p ökas.



Exempel. Låt X vara en s.v. som betecknar antalet sexor som fås vid 20 tärningskast. Då är $X \in \text{Bin}(20, 1/6)$. Sannolikheten att få minst 8 sexor vid 20 tärningskast är

$$\mathbb{P}(X \geq 8) = \sum_{k=8}^{20} \binom{20}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{20-k} \approx 1.1\%.$$

Vi undersöker nu fallet då sannolikheten p att experimentet lyckas är mycket liten. Vi låter samtidigt n vara stort, så att $\mu = np$ har storleksordning ett. Då gäller

$$\begin{aligned} p_X(k) &= \binom{n}{k} p^k (1-p)^{n-k} = \frac{n(n-1)\cdots(n-k+1)}{k!} \left(\frac{\mu}{n}\right)^k \left(1 - \frac{\mu}{n}\right)^{n-k} \\ &= \underbrace{\frac{\mu^k}{k!} \left(1 - \frac{\mu}{n}\right)^n}_{\approx e^{-\mu}} \underbrace{\frac{n(n-1)\cdots(n-k+1)}{n^k}}_{\approx 1 \text{ om } k \ll n} \underbrace{\left(1 - \frac{\mu}{n}\right)^{-k}}_{\approx 1 \text{ om } k \ll n} = \frac{\mu^k}{k!} e^{-\mu}. \end{aligned}$$

Definition. En s.v. X sägs vara **Poissonfördelad** om

$$p_X(k) = \frac{\mu^k}{k!} e^{-\mu},$$

där $\mu > 0$ och $k = 0, 1, \dots$. Detta betecknas $X \in Po(\mu)$.

Poissonfördelningen kan användas för att beskriva en situation då ett försök med liten sannolikhet att lyckas upprepas många gånger. För $n > 10$ och $p < 1/10$ så är approximationen $\text{Bin}(n, p) \approx \text{Po}(np)$ god. Detta illustreras i figuren nedan.

