Föreläsning 1, Kösystem

Här följer en kort sammanfattning av det viktigaste i Föreläsning 1.

Kolla kursens hemsida minst en gång per vecka. Övningar kommer att läggas ut där, skriv ut dem och ha med på övningstillfället. Också laborationshandledningarna kommer att finnas på hemsidan.

Föreläsningar

Kursen innehåller sju föreläsningar. Första veckan är det två föreläsningar, därefter maximalt en per vecka.

Övningar

Övningarna börjar redan vecka 1. Skriv ut övningsmaterialet på hemsidan och ha med till övningen. Både problem och fullständiga lösningar till dem kommer att finnas på hemsidan.

Laborationer

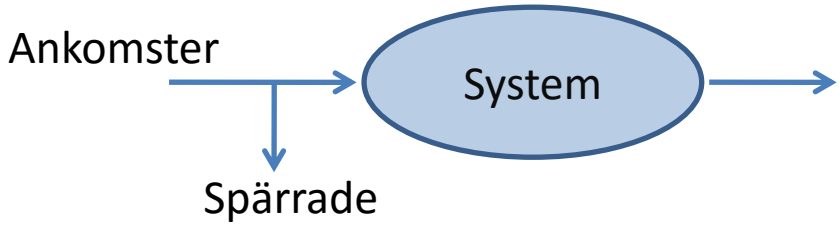
Det är två obligatoriska laborationer. Man anmäler sig till laborationer på kursens hemsida, se under länken ”Anmälan” till vänster på kursens hemsida.

Vad ska vi studera i denna kurs?

Vi ska titta på betjäningssystem av olika slag. Till ett betjäningssystem kommer kunder som får betjäning av betjänare. Om det inte finns några lediga betjänare så kan kunden ibland vänta i en kö. En kund kan vara en människa av kött och blod (som i en butik) men i våra tillämpningar är kunden ofta något abstrakt som till exempel en transaktion i en databas, ett mobilsamtal som ska betjänas av ett mobilnät eller http-paket som kommer till en webbserver.

Vi är naturligtvis intresserade av att kunderna ska få en tillräckligt bra service av betjäningssystemet. En kund ska inte bli så fördröjd att den blir missnöjd eller att det uppstår tekniska problem, kanske genom time-outer. Dock är de resurser som finns för att bygga systemet alltid begränsade, så det gäller att kunna avgöra om resurserna räcker. Om man dessutom kan optimera sitt betjäningssytem så kanske mindre resurser räcker för att ge en lika bra service som i ett ej optimerat system.

Grundproblemet vi ska studera i denna kurs ser ut så här:



Kunder kommer till ett betjäningssystem av något slag, i fortsättningen bara kallat system för korthets skull. Om systemet är fullt så kanske en kund spärras. I livsmedelsbutiker i det gamla DDR så fick inte fler kunder komma in i butiken om korgarna eller vagnarna var slut vid ingången. Om man

försöker ringa när alla frekvenser och tidluckor är upptagna i en cell i ett mobilnät blir man spärrad. Om en webbserver har startar maximalt antalet trådar så får nya kunder bara meddelandet ”server busy”. Om kunden får komma in i systemet så blir den så småningom färdig och kommer ut igen. Då är det naturligtvis intressant hur lång tid det tar innan kunden är färdig. Vi antar att inga nya kunder skapas inne i systemet. Vi ska huvudsakligen behandla frågor i kursen:

1. Vad är sannolikheten att en kund spärras?
2. Hur lång tid tillbringar en kund i systemet?
3. Hur många kunder kan systemet betjäna per tidsenhet och fortfarande uppfylla kraven på god kvalité?

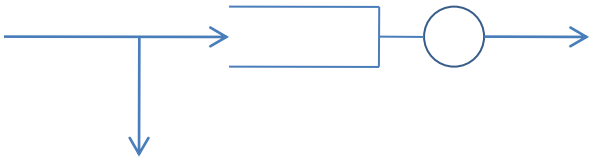
Ofta så är ankomsterna till ett system slumpmässiga. Vi vet inte exakt när någon vill kolla på en webbsida eller när någon vill använda sig av en molntjänst som en serverhall tillhandahåller. Vi vet inte heller hur lång tid det tar att betjäna en kund. Ibland kastar den som surfar in på webbsidan bara en snabb blick på den och försvinner sedan, ibland så stannar surfaren kvar, klickar på många länkar, lägger varor i korgen etc. Därför måste man använda modeller hämtade från sannolikhetsläran för att svara på frågorna ovan.

Några exempel på kösystem

Här följer några enkla exempel på betjäningssystem och en modell för dem. Avsnittet visar också hur vi ska rita kösystem med cirklar, pilar och köutrymmen.

Webbserver

Till en webbserver kommer begäran om att hämta sidor. Dessa betjänas av servern som skickar tillbaka ett antal filer som kan innehålla text och bilder. En modell som enligt mätningar ger bra resultat är så här enkel:



Den runda cirkeln är betjänaren, i detta fall processorn i servern. Kunder (i det här fallet begäran om att få se en webbsida) kommer till systemet. Om betjänaren inte är ledig så kan de lagras i köutrymmet och få vänta på att bli betjänade. Om köutrymmet är fullt så avvisas de.

Mobilsystem

En basstation i ett GSM-nät har ett antal frekvenser. Om det inte finns några lediga frekvenser när någon vill ringa så spärras det nya samtalet. Man kan använda följande kösystem för att beskriva detta:

Här är varje radiokanal en betjänare så det finns många betjänare. Kunderna är abonnenter som vill ringa ett telefonsamtal. Det finns inget köutrymme, så när man avvisas så får man inte vänta.

Charkuteridisk

En charkuteridisk med två biträden kan man modellera på följande sätt:

Kunderna avvisas inte utan alla som vill får vänta.

En stor serverpark för molntjänster

Modeller där flera köer är sammankopplade är viktiga för att undersöka kapaciteten hos molntjänster. Man kan även optimera tjänsterna med hjälp av kömodeller.

Andra tillämpningar

Kömodeller används inom många områden. Till exempel studeras ofta lagerproblem, tillverkningsprocesser, bagagesystem och vägtrafik med kömodeller.

Kömodeller i denna kurs

Vi ska studera enskilda kösystem som består av ett köutrymme och ett antal betjänare. Köutrymmet kan vara oändligt (charkuteridisken), ändligt (webbservern) eller inte finnas alls (GSM-nätet). Det kan finnas en eller fler betjänare. Vi ska också studera könät som är system av flera ihopkopplade köer (routern). Eftersom både ankomster och betjäningstider i allmänhet är slummässiga så behöver vi använda sannolikhetsteori.

I kursen kommer vi att studera ***köteori*** som är teorin för enskilda köer. Vi kommer också att studera ***könätsteori*** som behandlar ihopkopplade köer. Dock måste man göra en hel del antaganden om de statistiska fördelningarna för ankomster och betjäningstider i köteori och könätsteori som ibland kan vara orealistiska. Därför ska vi också studera ***simulering***, som är en teknik som är mycket allmängiltig, men som också har sina begränsningar.

Vilka statistiska egenskaper som ankomster och betjäningstider har måste man i allmänhet mäta. Vi behandlar dock inte mätningar i denna kurs.

Några beteckningar

Vi definierar några storheter som hör ihop med frågorna ovan:

Sannolikheten att en kund ***spärras*** eller ***avvisas:*** *P(*spärr*)*.

Tiden i systemet för en kund som inte har spärrats, *T*. Detta är i allmänhet en stokastisk variabel och vi är oftast intresserade av att beräkna medelvärde och varians för *T*. Kallas ofta för **svarstid**.

Genomströmningen, det vill säga hur många kunder per tidsenhet som blir färdigbetjänade i systemet, Λ.

Ankomstintensiteten till systemet, det vill säga hur många kunder per tidsenhet som kommer till systemet (både de som avvisas och som får komma in i systemet), λ.

Den effektiva ankomstintensiteten till systemet, det vill säga hur många som per tidsenhet får komma in i systemet, λ_{eff}.

I allmänhet så bildar ankomsterna till ett system en stokastisk process och betjäningstiderna är också slumpmässiga. Därför måste vi använda sannolikhetsteori och teorin för stokastiska processer för att studera kösystem.

Littles sats

Littles sats är ett mycket användbart och enkelt samband. Om vi har ett system av något slag som befinner sig i jämvikt där inga kunder vare sig skapas eller förstörs inne i systemet så gäller:

$$E(N) = E(T) \cdot \lambda_{eff}$$

Detta samband är mycket generellt. Vi ska använda det för att studera kösystem, men det gäller också för andra system.

Repetition av sannolikhetslära

Här följer en repetition av de viktigaste begreppen i sannolikhetsläran. Vi övar på dessa begrepp på de första övningarna. Om du vill ha en mer fyllig version så kolla i läroböckerna i matematisk statistik och i läroboken där ett kapitel ger en snabböversikt.

Diskreta stokastiska variabler

En ***stokastisk variabel*** är ett slumpmässigt talvärde som man får genom att göra ett försök (t ex kasta tärning) eller en observation (t ex iaktta hur många kunder som det finns i ett kösystem). ***Utfallsrummet*** är alla värden som en stokastisk variabel kan anta.

En ***diskret stokastisk variabel*** antar heltalsvärden. Exempel på diskreta stokastiska variabler i denna kurs är antalet kunder i ett kösystem. Det betyder att värdena som de antar för det mesta kommer att tillhöra mängden $\{0,1,2 \dots\}$. Låt N vara en diskret stokastisk variabel. Vi kommer att kalla $P(N = k) =$ sannolikheten att har värdet k för variabelns ***frekvensfunktion***. Ett trivialt exempel är att om N är antalet ögon vid kast med en tärning, då är

$$P(N = k) = \begin{cases} {}^{1}/_{6}, k \in \{1,2, \dots ,6\} \\ 0 \, för \, övrigt \end{cases}$$

Medelvärdet för en diskret stokastisk variabel är

$$E(N) = \sum kP(N = k)$$

Summan tas för alla k för vilka $P(N = k) \neq 0$.

Variansen för en diskret stokastisk variabel är

$$V(N) = E((N - E(N))^2)$$

Om värdet av den stokastiska variabeln ofta ligger långt från medelvärdet så blir variansen stor, annars blir den liten. Ju större varians desto längre från medelvärdet tenderar N att vara.

Man kan enkelt visa följande formler:

$$E(N + M) = E(N) + E(M)$$

$E(aN) = aE(N)$

$V(N) = E(N^2) - E^2(N)$

Kontinuerliga stokastiska variabler

Kontinuerliga stokastiska variabler antar reella värden. I denna kurs är de ofta tiden mellan två händelser vilket innebär att den oftast är ett positivt tal. Om X är en kontinuerlig stokastisk variabel så kallar vi

$F_X(t) = P(X \leq t)$

variabelns ***fördelningsfunktion***.

Några enkla egenskaper för fördelningsfunktionen:

$\lim_{t\to\infty} F_X(t) = 1$ eftersom $P(X < \infty) = 1$.

$F_X(t)$ är en växande funktion

$0 \leq F_X(t) \leq 1$ eftersom frekvensfunktionen är en sannolikhet.

Frekvensfunktionen för en kontinuerlig stokastisk variabel definieras som

$f_X(t) = \frac{d}{dt} F_X(t)$

Medelvärdet definieras som

$E(X) = \int_{-\infty}^{\infty} t f_X(t) dt$

Variansen är precis som för diskreta stokastiska variabler

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

Man kan också visa att

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

Oberoende stokastiska variabler

Intuitivt så förstår vi vad som menas med att två stokastiska variabler är oberoende av varandra. Om den ena har ett visst värde så påverkar det inte värdet på den andra variabeln. Vi uttrycker det matematiskt på följande sätt:

$P(N = i, M = j) = P(N = i) \cdot P(M = j)$ för diskreta stokastiska variabler

$P(X \leq t, Y \leq u) = P(X \leq t) \cdot P(y \leq u)$ för kontinuerliga stokastiska variabler

Ju fler variabler i en modell som är oberoende av varandra, desto lättare blir det i allmänhet att göra beräkningar. Längre fram kommer vi att se några trevliga konsekvenser av oberoende.

Transformer

Transformer är ett hjälpmedel som ibland underlättar beräkningar och härledningar. För diskreta stokastiska variabler använder man ***z-transformen*** som definieras som

$$P_N(z) = \sum_{\forall k} z^k P(N = k)$$

Denna trasform har egenskaperna

$P_N(z) \to 1$ då $z \to 1$

$\frac{d}{dz} P_N(z) \to E(N)$då $z \to 1$

$\frac{d^2}{dz^2} P_N(z) \to E(N^2) - E(N)$ då $z \to 1$

Man kan också visa att om M och N är oberoende diskreta stokastiska variabler och $A = M + N$ så gäller

$$P_A(z) = P_M(z) \cdot P_N(z)$$

För kontinuerliga stokastiska variabler använder man i stället ***Laplacetransformen*** som definieras på följande sätt:

$$F_X^*(s) = \int_0^\infty e^{-st} f_X(t)dt$$

Man kan visa följande samband

$F_X^*(s) \to 1$ då $s \to 0$

$\frac{d}{ds} F_X^*(s) \to -E(X)$ då $s \to 0$

$\frac{d^2}{ds^2} F_X^*(s) \to E(X^2)$ då $s \to 0$

Om X och Y är oberoende och $U = X + Y$ så gäller

$$F_U^*(s) = F_X^*(s) \cdot F_Y^*(s)$$

Betingade stokastiska variabler

Antag att A och B är två händelser. Då inför vi beteckningen

$P(A|B) =$ sannolikheten att A har inträffat om vi vet att B inträffat

Man kallar detta en ***betingad sannolikhet***. Några exempel med tärningskast visar hur det fungerar. Låt N vara antalet ögon som man får när man kastar. Då gäller:

$$P(N = 1|N \leq 3) = \frac{1}{3}$$

$$P(N = 1|N > 3) = 0$$

$$P(N < 3|N = 2) = 1$$

$$P(N \leq 3 | N \geq 2) = \frac{2}{5}$$

Man kan visa att

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

$P(A,B)$ är sannolikheten att både A och B inträffar.

Antag att vi delar in allt som kan hända (utfallsrummet) i ett antal händelser $B_1, B_2, \ldots$ som är ömsesidigt uteslutande (det vill säga bara en av dem kan inträffa) och att de fyller ut hela utfallsrummet. Om då A är en annan händelse så gäller ***satsen om total sannolikhet***:

$$P(A) = \sum_{\forall k} P(A|B_k)P(B_k)$$

Om N är en diskret stokastisk variabel och X en kontinuerlig så kan man visa följande:

$$P(N = k) = \int_{-\infty}^{\infty} P(N = k | X = t) f_X(t) dt$$

Integralen tas oftast bara från 0 till ∞ i denna kurs eftersom vi nästan alltid har positiva stokastiska variabler.

$$f_X(t) = \sum_{\forall k} f_X(t|N = k)P(N = k)$$

Man kan visa att dessa två formler följer från satsen om total sannolikhet.