UMEÅ UNIVERSITET
Institutionen för matematik
och matematisk statistik
Anton Grafström

Laboration

Laboration 2
Statistik för datavetare

Uppgifterna i den här laborationen ska lösas med hjälp av MATLAB.

Redovisning: En laborationsrapport ska lämnas till David Källberg eller Jonas Bygren, eller i labblådan utanför institutionen, senast 2/6 2009 kl. 15.00.

OBS! Laborationsrapporten ska skrivas enligt reglerna i dokumentet "Instruktion för rapportskrivning" som återfinns på kurshemsidan. Gå gärna igenom den färdiga rapporten och se till så att allt som ska finnas med också finns med då du/ni annars kommer att få tillbaka rapporten som ofullständig.

Vi kommer att bedöma rapporterna och sätta O, K eller G på dessa. O står för ofullständig vilket innebär att du/ni får komplettera med det som saknas och lämna in på nytt. K innebär att vi har lite kommentarer eller frågor på rapporten och att vi vill diskutera dessa med er innan rapporten godkänns. G betyder att rapporten är godkänd.

Laborationsrapporten ska hämtas ut senast 1 månad efter inlämningsdatum. Rapporter med K som inte hämtats ut inom en månad får betyg O, vilket innebär att den måste göras om. Vid en andra inlämning, efter komplettering, måste den först inlämnade och rättade rapporten bifogas.

Uppgift 1
Du jobbar på ett visst företag som tillverkar mobiltelefoner och du är chef för avdelningen där man utvecklar mjukvara till telefonerna. Nu ska en ny modell tas fram och du måste (med stor sannolikhet) kunna hålla den deadline som du lovar. Detta beror på att varje extra dag som leveransen av mobilen försenas efter deadline kostar flera miljoner. Du kan inte heller välja att lägga deadline så långt fram att du nästan helt säkert klarar deadline eftersom då kommer de konkurrerande företagen att hinna komma ut med motsvarande modell före er och ni förlorar en massa pengar. Du bestämmer dig för att det är rimligt att lägga deadline så att ni med sannolikhet 0.95 kommer att klara av att hålla den. Vid deadline får det maximalt vara två upptäckta fel kvar i systemet.

För att snabba upp processen kommer ni som vanligt att återanvända mycket kod från andra modeller och får på så vis ett operativ i version 0.01 redan första dagen. Det är några miljoner rader kod i programmet och givetvis uppstår ett antal fel när man gör så här. Du har data sedan tidigare som visar att antal fel $X(1)$ i version 0.01 kan komma från följande fördelning

$$X(1) \sim Po(70),$$

dvs. Poissonfördelning med väntevärde 70. Därefter börjar arbetet med att få bort felen. För att förenkla kommer vi att arbeta med diskret tid $t = 1, 2, ..., 100$. Vid $t = 1$ har ni alltså $X(1)$ fel. Vid ett tidssteg händer följande

$$X(t) = X(t-1) - Y(t) + Z(t) + W(t)$$

Där $Y(t)$ är antalet fel som ni har hunnit åtgärda $Y(t) \sim Bin(X(t-1), p)$ där $p = 0.5$. Det innebär att för varje fel som finns vid $t-1$ är sannolikheten 0.5 att ni hunnit åtgärda det vid tid t.

Vi har också $Z(t) \sim Bin(Y(t), q)$ där $q = 0.1$. Det innebär att för varje fel som ni åtgärdar så leder det till ett nytt fel med sannolikhet 0.1.

Den sista termen $W(t)$ är antalet andra nya fel som upptäcks vid tid t. Tidigare studier visar att $W(t) \sim Po(\lambda(t))$ där $\lambda(t) = 3 \cdot (100 - t)/100$. Det innebär att antal nya fel som upptäcks vid tid t har en Poissonfördelning med väntevärde $\lambda(t)$. Parametern $\lambda(t)$ är konstruerad så att det förväntade antalet nya fel som upptäckts minskar med tiden.

Din uppgift är nu att simulera $X_i(t)$, $t = 1, 2, ..., 100$, för $i = 1, 2, ..., N$ där $N = 1000$, dvs simulera $X(t)$ 1000 ggr. Deadline väljs som den första tiden

(d) då 95% av alla $X_i(d)$ samtidigt har 2 eller färre fel.

En sådan simulering ger en uppskattning av bästa tiden för deadline. Upprepa simuleringen ett antal gånger och presentera medelvärdet för deadline samt standardavvikelsen för deadline. Prova också att förändra parametrarna p, q och $\lambda(t)$ och förklara vad som händer. Tips: Använd plot och rita upp $X_i(t)$:na för att se vad som händer. I rapporten ska du/ni förklara hur ni gått till väga för att lösa problemet.

OBS! Koden för simuleringen ska bifogas som bilaga till rapporten.

Uppgift 2
SUNET är en gemensam organisation för universitet, vars mål är att genom samverkan mellan högskolorna i Sverige främja datakommunikation som är till nytta för högskolan, i första hand genom att tillhandahålla möjligheter till datakommunikation till/från/mellan universitet och högskolor (och ytterligare organisationer som tillhör samma intressesfär) nationellt och internationellt, i andra hand genom att stödja och samarbeta med organisationer som har som målsättning att erbjuda datakommunikation åt hela samhället (För mer information om SUNET se www.sunet.se). Filerna Data1.txt och Data2.txt innehåller lite statistik för trafiken mellan enheter i nätverket.

2.1. Data1.txt innehåller information om genomsnittlig trafik per månad för 70 "interfaces", för mars och april. Finns det tillräckliga bevis för att genomsnittstrafiken var högre i mars? Vilka antaganden måste vara uppfyllda för testet du gör? Är de uppfyllda? Använd signifikansnivån 0.05. Försök förklara resultatet.

2.2. Data2.txt innehåller genomsnittlig trafik per dag för tre "interfaces" (2 mellan servrar i Stockholm och en mellan Umeå, Luleå och Sundsvall) i april.

a) Finns det tillräckliga bevis för att genomsnittstrafiken för Umeå är högre än 210 000kb/s? Vilka antaganden ska vara uppfyllda för testet? Är de uppfyllda? Använd signifikansnivån 0.05.

b) Finns det tillräckliga bevis för att påstå att varianserna mellan de två servrarna i Stockholm skiljer sig åt? Vilka antaganden måste vara uppfyllda för testet? Är de uppfyllda? Använd signifikansnivån 0.05.

Lycka till!