TAIU06/9GMA05 DATORÖVNING 1

Datorövningen behandlar simulering av observationer från diskreta och kontinuerliga fördelningar, illustration av skattningars osäkerhet, transformation av data för att få bättre anpassning till normalfördelning.

Innan du startar Matlab bör du ladda ned filen `lab1_uppg3.m` från kurshemsidan.

Du kan kontrollera att filen finns i din katalog med kommandot `ls` (skrives i terminalfönstret).

När du arbetar i Matlab bör du skriva allt i m-filer så att du lätt kan gå tillbaka och ändra eller återupprepa processen senare om du skulle vilja.

Redovisning

Datorövningen ska göras, individuellt eller i par om två.

Spara alla filer med kod så att du är beredd att visa den eller testköra under redovisningstillfället. Spara även, eller skriv ut alla plottar.

Uppgift 1.
Simulering av observationer från en diskret fördelning

Vi har ofta utnyttjat tärningskast för att illustrera olika räknelagar för sannolikheter. Du ska nu med hjälp av Matlab simulera sådana kast.

Starta Matlab och öppna en m-fil. Namnge filen till *uppg1.m* och spara den på lämpligt ställe.

Skriv följande kommandon i m-filen och evaulera.

```
n=600; % n stycken kast

% Simulera en tärning
kast = randi(6,n,1);
```

Skriv `help randi` i kommandofönstret för att se hur `randi` fungerar. Se speciellt exemplet i hjälpavsnittet.

Det kan nu vara intressant att titta på frekvenserna f_i för de olika utfallen. Skriv `sum(kast==i)` för $i = 1, \ldots, 6$ i kommandofönstret (skriv alltså, t ex `sum(kast==2)` etc). Vilka frekvenser får du?

Vi kan även plotta frekvenserna med hjälp av ett histogram. Börja med att "kommentera bort" simuleringen ovan genom att skriva ett % före den, dvs. så vi får `% kast = randi(6,n,1)` (om vi inte gör det så kommer vi att simulera nya kast).

Fyll på i m-filen och evaluera.

```
% Histogram
figure
hist(kast,[1:6])
ylabel('antal kast')
xlabel('antal ögon på tärningen')
```

Beräkna stickprovsmedelvärdet och stickprovsstandardavvikelsen för de 600 kasten. Jämför medelvärdet med det teoretiska väntevärdet för ett tärningskast. Skriv i m-filen.

```
% Stickprovsmedelvärde
medel = mean(kast)
% Stickprovsstandardavvikelse
stdev = std(kast)
```

Fyll i resultaten på resultatbladet som finns sist i det här häftet och redovisa uppgiften så snart din handledare är ledig.

Glöm inte att spara din m-fil så att du kan gå tillbaka till din programkod.

Uppgift 2.
Punktskattning vid rektangelfördelning

Rensa i kommandofönstret och minnet genom att skriva `clear` och `clc` i kommandofönstret. Öppna en ny m-fil och namnge filen till *uppg2.m*.

Vi ska nu skapa 100 observationer från $\mathrm{Re}(0, a)$ (välj själv a-värde). Skriv i m-filen.

```
N = 100;
n = 20;
X = a*rand(N,n); % där a är det värdet som du har valt
```

Vi har härlett momentskattningen $\hat{a}_1 = 2\bar{x}$ och den korrigerade ML-skattningen $\hat{a}_2 = ((n+1)/n) \cdot \max x_i$. Du ska nu jämföra hur dessa båda skattningar fungerar för n=20.

Ovan har vi nu en matris X som är 100×20 och som vi tolkar som 100 stycken stickprov med 20 observationer i varje stickprov. Vi kan nu beräkna momentskattningen och maximum-likelihood-skattningen för de 100 stickproven. Alltså, vi får 100 skattningar från varje metod. Skriv i m-filen och evaluera.

```
ahat1 = 2*mean(X,2);
ahat2 = ((n+1)/n)*max(X,[],2);
```

Skriv `help mean` och `help max` i kommandofönstret för att se hur funktionerna fungerar. Vi vill nu plotta de olika skattningarna. Skriv i m-filen men evaluera **inte** än.

```
figure
plot(ahat1,'b*');
hold on
plot(ahat2,'ro');
legend('Skattning med moment-metoden','Korrigerad MLE-skattning',
                                        'Location','NorthEastOutside')
```

Markera de nya raderna. Högerklicka med musen på det markerade stycket och välj **"Evaluate Selection"** (eller tryck F9). Om du kör en hel evaluering så kommer nya observationer att simuleras vilket också är ok.

Varför ser skattningarna ut som de gör?

Beräkna också approximativa värden på väntevärdena $E(\widehat{A}_i)$ och standardavvikelserna $D(\widehat{A}_i)$, $i = 1, 2$, dvs. beräkna medelvärdet och stickprovsstandardavvikelsen för respektive skattning.

Spara m-filen, fyll i resultaten på resultatbladet och redovisa.

Uppgift 3.
Jämförelse av volatilitet

NASDAQ-börsen brukar beskrivas som en marknad för spännande tillväxtföretag. Stora vinster betyder ofta också större risker, vilket i sin tur är kopplat till större volatilitet jämfört med t ex New York-börsen (NYSE) som anses vara ganska stabil. Vi ska studera detta med hjälp av ett datamaterial som innehåller

värdeförändringar under en slumpmässigt vald vecka för 30 aktier från NASDAQ och 30 från NYSE. Absolutbeloppet av förändringen anses vara ett rimligt mått på volatiliteten under en period då den genomsnittliga förändringen ligger nära noll. En ökning på 10% anges i datamaterialet som 0.1 etc.

Öppna m-filen **uppg3.m** som du kopierade från kursbiblioteket. Börja med att skapa absolutbeloppet av förändringarna. Fyll på i m-filen.

```
x = abs(kurs_NASD);
y = abs(kurs_NYSE);
```

Ta fram beskrivande statistik för x och y genom att använda `normfit` och `histfit`.

```
[muhatX sX] = normfit(x)
figure
subplot(1,2,1); histfit(x); title('NASD')

[muhatY sY] = normfit(y)
subplot(1,2,2); histfit(y); title('NYSE')
```

Det är ganska tydligt att en normalfördelningsmodell inte passar så bra. Det finns en tydlig skevhet med lång högersvans och den anpassade normalfördelningskurvan har sannolikhetsmassa på den negativa axeln medan våra mätvärden är ≥ 0. I sådana fall kan logaritmisk transformation vara bra, men man får se upp så att man inte skapar en lång vänstersvans. Skriv i m-filen.

```
% Transformera data
u = log(x + 0.01);
v = log(y + 0.01);
```

Addition av 0.01 hjälper oss att undvika långa vänstersvansar. Här är valet av värdet 0.01 inte självklart.

Ta fram beskrivande statistik för u och v på samma sätt som för x och y. Kommentera bort eller markera och "Evaluate Selection" om du inte vill få upp samma figurer flera gånger. För att stänga flera figurer samtidigt kan du använda `close all`.

```
figure
subplot(1,2,1); histfit(u); title('log-NASD')
subplot(1,2,2); histfit(v); title('log-NYSE')

[muhatU,sU,muU_CI,sigmaU_CI] = normfit(u)
[muhatV,sV,muV_CI,sigmaV_CI] = normfit(v)
```

Nu är skevheten i stort sett borta. Taggigheten hos histogrammen hänger delvis ihop med hur Matlab bildar klasser för histogrammet.

Spara m-filen och fyll i redovisningsbladet. Se till att spara filen till nästa laboration också!

REDOVISNINGSBLAD

Fyll i namn och personnummer med en bläckpenna.

1)

.....

2)

.....

Du ska även kunna visa upp analyserna för de olika uppgifterna som utskrift eller i datorn.

Uppgift 1

$f_1 = \ldots\ldots$ $f_2 = \ldots\ldots$ $f_3 = \ldots\ldots$ $f_4 = \ldots\ldots$ $f_5 = \ldots\ldots$ $f_6 = \ldots\ldots$

Medelvärdet $\bar{x} = \ldots\ldots$; väntevärdet $E(X_i) = \ldots\ldots$;

stickprovsstandardavvikelse $s = \ldots\ldots$ OK .....

Uppgift 2

Sant värde $a = \ldots\ldots$

Approximativa värden på $E(\widehat{A}_1)$: ..... på $D(\widehat{A}_1)$: .....

på $E(\widehat{A}_2)$: ..... på $D(\widehat{A}_2)$: ..... OK .....

Uppgift 3

Skattade parametrar $\hat{\mu}_1 = \ldots\ldots$, $s_1 = \ldots\ldots$ och

$\hat{\mu}_2 = \ldots\ldots$, $s_2 = \ldots\ldots$

Då alla uppgifterna är godkända ska du skriva upp dig på närvarolistan.

Appendix.
Punktskattning vid rektangelfördelning

Låt $x_1, x_2, \ldots, x_n$ vara observationer av oberoende stokastiska variabler som är $\mathrm{Re}(0, a)$ (har likformig fördelning över intervallet $(0, a)$). Då har vi täthetsfunktionen

$$f(x) = \begin{cases} \dfrac{1}{a} & \text{om } 0 \le x \le a \\ 0 & \text{f.ö.} \end{cases}$$

Momentmetoden: Vi har $E(X_i) = \frac{a}{2}$, se formelsamlingen. När man använder momentmetoden för att hitta en skattning så löser man ekvationen $\mu(\widehat{a}) = \bar{x}$, där funktionen $\mu(\widehat{a})$ är väntevärdet för fördelningen som variablerna X_i kommer ifrån. Det vill säga, $E(X_i) = \mu(a)$. Ekvationen $\hat{a}/2 = \bar{x}$ ger momentskattningen $\hat{a}_1 = 2\bar{x}$.

Då är $E(\widehat{A}_1) = E(2\bar{X}) = 2E(\bar{X}) = 2E(X_i) = 2 \cdot \frac{a}{2} = a$ vilket innebär att $\widehat{A}_1$ är väntevärdesriktig.

Vidare blir $V(\widehat{A}_1) = V(2\bar{X}) = 4V(X_i)/n = 4 \cdot \frac{a^2}{12} \cdot \frac{1}{n} = \frac{a^2}{3n}$. Alltså är $D(\widehat{A}_1) = a/\sqrt{3n}$.

ML-metoden: ML-metoden står för *maximum likelihood*-metoden, och den går ut på att välja det värde på parametern (a i vårt fall) som gör de observerade värdena $(x_1, x_2, \ldots, x_n)$ mest troliga som observationer från den givna fördelningen ($\mathrm{Re}(0, a)$ i vårt fall). Detta åstadkommer vi genom att maximera den så kallade likelihoodfunktionen $L(a)$ som definieras enligt följande:

$$L(a) = \prod_{i=1}^{n} f_{X_i}(x_i\,;\, a). \tag{1}$$

Funktionerna f_{X_i} är täthetsfunktiorna för respektive X_i, och dessa beror på den okända parametern a. Värdena x_i är fixa (observationerna) och varierar ej när vi utför maximeringen. Anledningen till att maximering av denna produkt ger den situation som gör de observerade värdena mest troliga kan ses i följande argument. Antag för ett ögonblick att vi har en diskret fördelning, där Y_i är diskreta s.v., $i = 1, 2, \ldots, n$ med sannolikhetsfunktioner $p_{Y_i}(k\,;\, a)$ där a är en okänd parameter. Vi har

$$\begin{aligned} P(Y_1 = k_1, Y_2 = k_2, \ldots, Y_n = k_n) &= \{\text{oberoende variabler}\} = \prod_{i=1}^{n} P(Y_i = k_i) \\ &= \prod_{i=1}^{n} p_{Y_i}(k_i\,;\, a). \end{aligned}$$

Om vi ersätter sannolikhetsfunktionerna med täthetsfunktioner erhåller vi i högerledet samma uttryck som i (1). Oftast har variablerna (som i vårt fall) samma täthetsfunktion, vilket underlättar arbetet.

För rektangelfördelningen får vi likelihoodfunktionen

$$L(a) = \begin{cases} \left(\dfrac{1}{a}\right)^n & \text{om } 0 \le x_i \le a \text{ för i=1,..., n} \\ 0 & \text{f.ö.} \end{cases}$$

Tänk på att a ska variera och att vi söker det a som gör $L(a)$ så stor som möjligt. Vi ser att $L(a) \neq 0$ bara om $a \ge x_i$ för alla i, vilket innebär att $L(a) \neq 0$ bara om $a \ge \max x_i$. Vi har alltså

$$L(a) = \begin{cases} \dfrac{1}{a^n} & \text{om } a \ge \max x_i \\ 0 & \text{f.ö.} \end{cases}$$

Det följer att $\hat{a} = \max x_i$ ger störst värde på $L(a)$.

Vi ska nu undersöka väntevärdesriktigheten. Låt $Y = \max X_i$. Vi behöver täthetsfunktionen för Y, men bestämmer först fördelningsfunktionen. Låt $0 \leq y \leq a$. Vi har då

$$F_Y(y) = P(Y \leq y) = P(\max X_i \leq y) = P(X_1 \leq y, \ldots, X_n \leq y) = \left(\frac{y}{a}\right)^n$$

eftersom de s.v. $X_1, \ldots, X_n$ är oberoende.

Detta ger täthetsfunktionen

$$f_Y(y) = F_Y'(y) = n \cdot \frac{y^{n-1}}{a^n} \quad \text{för } 0 < y < a$$

och

$$E(\widehat{A}) = E(Y) = \int_0^a y f(y) dy = \ldots = \frac{n}{n+1} \cdot a,$$

vilket innebär att $\widehat{A}$ inte är väntevärdesriktig.

Vi får den korrigerade ML-skattningen $\hat{a}_2 = \frac{n+1}{n} \cdot \hat{a} = \frac{n+1}{n} \cdot \max x_i$ med

$$E(\widehat{A}_2) = \frac{n+1}{n} E(Y) = a$$

$$V(\widehat{A}_2) = \left(\frac{n+1}{n}\right)^2 V(Y)$$

vilket efter en del räknande ger

$$V(\widehat{A}_2) = \frac{a^2}{n(n+2)}$$

och

$$D(\widehat{A}_2) = a/\sqrt{n(n+2)}.$$

Vi ser att $\widehat{A}_2$ är mycket effektivare än $\widehat{A}_1$, vilket också kommer att bekräftas av simuleringarna i uppgift 2.