DATORÖVNING 2
MATEMATISK STATISTIK FÖR F OCH FYSIKER;
FMSF45 & MASB03

Syfte:

Syftet med denna datorövning är att du skall:

- få förståelse för diskreta, bivaria och betingade fördelningar.
- bli bekant med summor av stokastiska variabler.
- få förståelse för hur och när centrala gränsvärdessatsen kan användas.

Specialrutiner finns att hämta på kursens hemsida:
www.maths.lth.se/matstat/kurser/fmsf45masb03/f/

1 Bakgrund

Laborationen består av två delar. Först studerar vi hur en bivariat diskret fördelning kan konstrueras från enklare del komponenter och undersöker de resulterande marginal och betingade fördelningar. Därefter undersöker vi summor av stokastiska variabler och centrala gränsvärdessatsen.

1.1 En modell för skördeutfall

I första delen av laborationen kommer vi att studera en enkel modell för skördeutfall. Frågan är hur stor skörd man kan förvänta sig om man planterar n st frö. För att modellera skördeutfallet kan vi dela upp problemet i två steg

1. Först konstruerar vi en modell för antalet av de planterade fröna som gror
2. Därefter funderar vi på hur stort skördeutfallet (antal nya frö) blir om precis k st frö gror.

Den resulterande modellen består nu av en fördelning för antalet frö $p_X(k)$ och en betingad fördelning för skördeutfallet, $p_{Y|X=k}(l|k)$. Bayessats och satsen om total sannolikhet ger oss nu den gemensam fördelningen för antalet frö som gror och skördeutfallet samt marginalfördelningen för skördeutfallet.

$$p_{X,Y}(k,l) = p_X(k) \cdot p_{Y|X=k}(l) \quad \text{och} \quad p_Y(l) = \sum_k p_X(k) \cdot p_{Y|X=k}(l)$$

Utöver dessa fördelningar är även den betingade fördelningen för X givet Y intressant, $p_{X|Y=l}(k|l)$. Om vi enbart observerar den totala skörden y så kan denna fördelning användas för att säga något om hur många frö som faktiskt grott.

1.2 Förberedelseuppgifter

1. Förvissa dig om att du förstår vad en sannolikhetsfunktion är.
2. **Mozquizto 1:** Vi planterar 7 frö med grobarhet 75%. Ange fördelningen för antalet frön som kommer att gro (om de gror oberoende av varandra) samt fördelningens väntevärde och varians.
3. **Mozquizto 2:** Om $X_i \in \mathsf{Po}(\mu_i)$ och oberoende vilken fördelning har då summan $Y = \sum_{i=1}^{n} X_i$?

4. Förvissa dig om att du förstår hur total sannolikhet fungerar för väntevärde, d.v.s. hur man kan beräkna $\mathsf{E}(Y) = \mathsf{E}(\mathsf{E}(Y \mid X))$

5. **Mozquizto 3:** Förvissa dig om att du förstår vad Centrala gränsvärdessatsen innebär och när den kan användas.

6. Vi beräknar medelvärdet $\overline{X}$ av de oberoende s.v. $X_i \in \mathsf{Po}(3)$, $i = 1, \ldots, n$ (samma väntevärde för alla X_i). Ange väntevärde och varians för $\overline{X}$. Vilken fördelning får $\overline{X}$ (approximativt) när n är stort? Ungefär hur stort måste n vara för att approximationen ska bli bra?

2 Modell för skördeutfall

2.1 Diskret variabel: Antal frö som gror

Vi vill simulera antalet frön som kommer att gro bland de sju planterade fröna. Det kan vi göra på två sätt. Det mest rättframma är att simulera 7 frön och räkna antalet som gror. Funktionen `rand(1,n)` ger en radvektor med n rektangelfördelade slumptal, U, mellan 0 och 1. För att sannolikheten att ett frö kommer att gro skall bli p kan vi helt enkelt se efter om $U \leq p$. I så fall kommer fröet att gro. Om $U > p$ så kommer det inte att gro. För att få reda på antalet frön som kommer att gro bland de 7 summerar vi den resulterande 0/1-variabeln:

```
fx >> n = 7;
      p = 0.75;
      U = rand(1,n)   % 1 rad och n kolumner med observation från R(0,1)
      U<=p            % 0 = gror inte, 1 = gror
      X = sum(U<=p)   % antal frön som gror
```

Uppgift: Jämför resultatet av `U=rand(1,n)` och `U<=p` och förvissa dig om att du förstår vad som hände.

För att illustrera vad som händer så kan vi också plotta slumptalen och den sannolikhet som vi jämför med.

```
fx >> figure(1)
      stem(U)         % 7 st R(0,1) slumptal
      refline(0, p)   % sannoliheten vi vill jämföra U med.
```

Mozquizto 4: Hur många frön grodde?

Ett smidigare sätt är att utnyttja att vi vet att antalet frön som kommer att gro är $\mathsf{Bin}(7, 0.75)$-fördelat. Då kan vi simulera X direkt med hjälp av MATLABs färdiga rutiner:

```
fx >> help binornd
      X = binornd(n,p)
```

Uppgift: Gör om simuleringen några gånger. Hur många frön brukar gro?

Antalet frön som kommer att gro varierar uppenbarligen från gång till gång. För att se hur vanligt det är med olika antal frön som kommer att gro simulerar vi $N = 100$ planterings tillfällen och ritar ett stolpdiagram (vi har ju en diskret variabel).

```
fx >> N = 100;                % antal fröpåsar
      X = binornd(n,p,N,1) % X = antal groende frön i var och en av Nx1 påsar
      antal = hist(X,0:n)  % använd hist för att räkna antalet gånger vi får 0,1,...,n
      antal(4)             % antal X==3 (4:e siffra i vektorn 0,1,2,3,...)
      sum(X==3)            % jfr med antalet X som är lika med 3
      figure(2)
      bar(0:n,antal)       % stolpdiagram
      xlabel('antal frön som gror')
      ylabel('antal tillfällen')
```

Uppgift: Var det någon av planterings tillfällen som inte hade några groende frön alls?

Uppgift: Hur många av planterings tillfällen gav 5 groende frön? Hur många gav högst 2 groende frön?

Vi vill nu jämföra våra 100 påsar med den teoretiska sannolikhetsfunktionen. För att göra det måste vi skala om y-axeln till andelar

```
fx >> bar(0:n,[antal/N; binopdf(0:n,n,p)]') % rita två uppsättningar staplar
      xlabel('antal frön som gror')
      ylabel('andel påsar')
```

För att hålla ordning på vilken färg som är vilken adderar vi också en `legend` med förklarande text

```
fx >> legend('simulering','exakt','Location','NorthWest')
```

Mozquizto 5: Hur stämmer andelen av de simulerade planterings tillfällena som hade precis 5 groende frön eller högst 2 groende frön med motsvarande sannolikheter? (Jämför med resultatet från `binopdf(5,n,p)` och `binocdf(2,n,p)`)

Mozquizto 6: Experimentera med att ändra grobarheten från $p = 0.75$ och antalet frön från $n = 7$. Hur ändrar sig fördelningen när n eller p minskar eller ökar?

2.2 Centrala gränsvärdessatsen för binomialfördelning

Om $np(1-p) > 10$ kan binomialfördelningen approximeras med en normalfördelning. Vi kan jämföra fördelningsfunktionerna och se hur bra det blir:

```
fx >> n = 7;
      p = 0.75;
      np = n*p              % väntevärde
      npq = np*(1-p)        % varians
```

```
x = linspace(np-4*sqrt(npq),np+4*sqrt(npq));  % mu +/- 4 sigma

figure(3)
stairs(0:n,binocdf(0:n,n,p))      % Stegfunktion p.g.a diskret s.v.
hold on
plot(x,normcdf(x,np,sqrt(npq)))   % men den här är kontinuerlig
hold off
```

Uppgift: Pröva med lite olika värden på n och p. Testa både när det går bra att normalapproximera och när det inte går.

Mozquizto 7: Beräkna sannolikheten att högst 2 frön gror, både exakt och med normalapproximation.

2.3 Simulering med hjälp av betingad fördelning: Skördeutfall

Vi tänker oss nu att varje frö som gror ger upphov till ett Poissonfördelat antal nya frön, i medeltal 10 frön per groende ursprungligt frö. Frön som inte gror ger naturligtvis inga nya frön. Vi är intresserade av fördelningen för det totala antalet nya frön som erhålls om vi planterar 7 frön med 75 % grobarhet.
Sedan tidigare har vi att $X =$ ”antal frön som gror” $\in \mathsf{Bin}\,(7, 0.75)$. Om exakt $X = k$ frön grodde blir $Y =$ ”antal nya frön” en summa över antalet frö från k st oberoende plantor. D.v.s. Summan av k stycken oberoende $\mathsf{Po}\,(10)$-fördelade variabler, en för varje groende frö:

$$Y = \sum_{i=1}^{k} Z_i \quad \text{där} \quad Z_i \in \mathsf{Po}\,(10)$$

Från förberedelseuppgifterna har vi då att fördelningen för $Y \mid X = k \in \mathsf{Po}\,(10 \cdot k)$ där $k = 0, \ldots, 7$. Fördelningen för Y ges då av (Satsen om Total Sannolikhet)

$$p_Y(l) = \sum_{k=0}^{7} p_{Y|X=k}(l) \cdot p_X(k) = \sum_{k=0}^{7} \frac{(10 \cdot k)^l}{l!} \cdot \mathrm{e}^{-10 \cdot k} \cdot \binom{7}{k} \cdot 0.75^k \cdot 0.25^{7-k}$$

För att ta reda på hur denna fördelning ser ut studerar vi först det enklare fallet med enbart $n = 2$ planterade frön. Först illustrerar vi sannolikheten att 0, 1 eller 2 frö gror

```
fx >> figure(4)
      subplot(211)
      bar(0:2, binopdf(0:2,2,.75))
      title('Antal frö som gror')
      ylabel('p(x)')
```

Därefter illustrerar vi de tre olika varianterna av betingade fördelningar: $\mathsf{Po}\,(0)$, $\mathsf{Po}\,(10)$, $\mathsf{Po}\,(20)$.

```
fx >> mu = 10;
      x = 0:4*mu;

      figure(4)
      subplot(234)
      bar(x, poisspdf(x, 0*mu))
```

```
title('Skörd om 0 frö gror')
ylabel('p(y|x=0)')

subplot(235)
bar(x, poisspdf(x, 1*mu))
title('Skörd om 1 frö gror')
ylabel('p(y|x=1)')

subplot(236)
bar(x, poisspdf(x, 2*mu))
title('Skörd om 2 frö gror')
ylabel('p(y|x=2)')
```

Mozquizto 8: Hur ändrar sig den betingade fördelningen för Y givet X när antalet groende frön ändrar sig?

Uppgift: Tänk efter hur fördelningen för Y borde se ut, när vi viktat ihop dessa 3 fördelningar med vikter enligt binomialfördelningen för antalet groende frön.

Även om vi inte vill räkna ut sannolikhetsfunktionen för Y så är det ganska enkelt att låta Matlab göra det:

```
fx >> pY = poisspdf(x,0*mu)*binopdf(0,2,0.75);      % fallet X=0
      pY = pY + poisspdf(x,1*mu)*binopdf(1,2,0.75); % fallet X=1
      pY = pY + poisspdf(x,2*mu)*binopdf(2,2,0.75); % fallet X=2
      figure(5)
      bar(x,pY)
      xlabel('antal nya frön')
```

Uppgift: Ser fördelningen ut som du hade väntat dig?

För det allmänna fallet kan vi använda en `for`-sats för att beräkna summan över k:

```
fx >> n=7; p=0.75; mu=10;
      y = 0:100;
      pY = zeros(size(y));   % Fyll först p_Y(y) med nollor.
      for k=0:n              % Uppdatera p_Y(y) för varje k
      pY=pY+poisspdf(y,mu*k)*binopdf(k,n,p);
      end
      figure(6)
      bar(y,pY)
      xlabel('antal nya frön')
```

Funktionen `harvest.m` (som finns på kurshemsidan) ritar upp sannolikhetsfunktionen för Y där $Y \mid X = x \in \mathsf{Po}\left(\mu \cdot x\right)$ och $X \in \mathsf{Bin}\left(n, p\right)$ för valfria värden på n, p och μ.

```
fx >> help harvest
      harvest(7, 0.75, 10)
```

Mozquizto 9: Experimentera med olika värden på n, p och μ. Vad händer om antalet planterade frö, n, minskar eller ökar? Om grobarheten, p, minskar eller ökar? Om medelantalet nya frön per frö som gror, μ, minskar eller ökar?

Uppgift: Vad händer om grobarheten är 100 %?

Uppgift: Kan du få fördelningen att se normalfördelad ut?

2.4 Bivariat fördelning och betingad fördelning

Funktionen `harvest2D.m` illustrerar den gemensamma fördelningen för antalet frön som gror, X, och antalet nya frö som skördas, Y.

$$p_{X,Y}(k,l) = p_{Y|X=k}(l) \cdot p_X(k), \quad k = 0,1,\ldots,n;\ l = 0,1,2,3,\ldots$$

```
fx >> figure(1)
      harvest2D(7, 0.75, 10)
      figure(2)
      harvest(7, 0.75, 10)
```

Mozquizto 10: Experimentera med olika värden på n, p och μ. Hur hänger den bivariata sannolikhetsfunktionen, $p_{X,Y}(k,l)$, ihop med marginal sannolikheten, $p_Y(l)$?

Givet den gemensamma sannolikhetsfunktionen kan vi också räkna ut den betingade fördelningen för hur många frön som grott givet att vi vet skördeutfallet

$$p_{X|Y=l}(k) = \frac{p_{X,Y}(k,l)}{p_Y(l)} = \frac{p_{Y|X=k}(l) \cdot p_X(k)}{p_Y(l)}, \quad k = 0,1,\ldots,n$$

Använd funktionen `harvestCond.m` för att illustrera den betingade sannolikhetsfunktionen om vi har skördat $y = 25$ frö

```
fx >> figure(1)
      harvestCond(7, 0.75, 10, 25)
      figure(2)
      harvest2D(7, 0.75, 10)
```

Mozquizto 11: Experimentera med olika värden på n, p, μ och observerat skördeutfall y. Hur hänger den bivariata sannolikhetsfunktionen, $p_{X,Y}(k,l)$, ihop med den betingade sannolikheten, $p_{X|Y=l}(k)$?

Mozquizto 12: Vad är det troligaste antalet frö som grott (högst betingad sannolikhet) givet att man skördade 50 frön?

3 Centrala gränsvärdessatsen

Vi skall nu titta lite närmare på Centrala Gränsvärdessatsen (CGS). Vi börjar med en liten simulering från en känd fördelning, två slumpmässiga obervationer x_1, x_2 från $X \in \mathsf{Po}(\mu)$ där $\mu = 3$. Vi ska sedan beräkna medelvärdet $\bar{x}$ och se hur nära väntevärdet μ det hamnar.

```
fx >> mu = 3;                    % det sanna my-värdet
      x = poissrnd(mu,2,1)       % en 2x1-matris med Po(my)-slumptal
      xmedel = mean(x)           % medelvärdet
```

Uppgift: Gör om simuleringen och medelvärdesberäkningen några gånger. Verkar medelvärdet variera mindre än de enskilda observationerna? Borde den det? I så fall, hur mycket mindre?

Låt oss göra om simuleringarna ett stort antal gånger så vi får bättre uppfattning om hur medelvärdet beter sig:

```
fx >> mu = 3;
      n = 2;                 % antal termer i medelvärdet
      M = 1000;              % antal simuleringar
      x = poissrnd(mu,n,M)   % n x M-matris. x1 i första raden, xn i sista.
      xmedel = mean(x)       % M st medelvärden
      subplot(2,1,1)
      hist(x(1,:),0:15)      % histogram över de Mst x1-värdena
      subplot(2,1,2)
      hist(xmedel,0:0.5:15) % histogram över de Mst x-medelvärdena
```

Uppgift: Experimentera med lite olika värden på n och se vad som händer med medelvärdet. Du kan behöva ändra klassbredden i det undre histogrammet för att se något, t.ex. `0:0.1:15`, som ger klasser från 0 till 15 med bredd 0.1.

Mozquizto 13: Jämför variationen hos de enskilda observationerna i den övre figuren och variationen för skattningarna i den undre figuren. Hur ändrar sig variationen hos observationerna när vi ändrar n?

Enligt centrala gränsvärdessatsen vet vi att $\sum X_i$ och därmed också medelvärdet $\bar{X}$ blir normalfördelat om vi summerar tillräckligt många variabler; **oavsett** vilken fördelning X_i har. Funktionen `cgsgui.m` illustrerar hur summor och medelvärde av ett antal olika standard fördelningar ser ut när n blir stort.

Uppgift: Experimentera med `cgsgui` för $\mathsf{Po}(3)$-fördelningen. Hur ser approximationen ut för olika n (och μ)?

Uppgift: För små n kan normalfördelningen uppenbarligen bli negativ. Blir $\bar{X}_n$ någonsin negativ?

Mozquizto 14: Undersök hur några andra fördelningar (t.ex. binomial-, exponential- och rektangefördelningarna) beter sig. Hur stort måste n vara för att normalapproximationen ska bli bra? Skiljer det sig mellan fördelningarna?