DATORÖVNING 1
MATEMATISK STATISTIK FÖR E— FMSF20

Syfte:

Syftet med dagens laborationen är att du skall:

- träna på olika sätt att illustrera och beskriva ett datamaterial
- få förståelse för begreppen fördelningsfunktion, täthets- och sannolikhetsfunktion samt sambandet mellan stickprov och population
- träna på att beräkna sannolikheter i Matlab
- bli bekant med inversmetoden för att generera slumptal

1 Förberedelseuppgifter

1. Förvissa dig om att du förstår vad täthetsfunktion och fördelningsfunktion är och hur de förhåller sig till varandra.
2. Vad är en kvantil?
3. Ta reda på hur man använder inversmetoden för att transformera slumptal till en önskad fördelning. (Kap. 8.4).

2 Relativa frekvenser och fördelningar

I denna del av laborationen ska vi använda ett datamaterial av kroppstemperatur hos 130 st friska 18–40 åringar[1]. Data finns i filen `kroppsTemp.mat` som kan hämtas från kurshemsidan.

2.1 Kroppstemperatur

Börja med att läsa in och titta på data

```
>> load kroppsTemp.mat
>> whos T
```

Data består av två kolumner där den första är temperaturer (i C) och den andra indikerar om personen var man (=1) eller kvinna (=2).
En god regel, när man står inför ett nytt datamaterial, är att rita upp det på några olika sätt.
Vi börjar med att göra ett histogram:

```
>> histogram(T(:,1)) %eller hist(T(:,1))
```

och en plot av data

```
>> figure(2)                % Ritar i ett nytt fönster
>> plot(T(:,1),'.')
```

och relatera det till histogrammet.

[1] Mackowiak, Wasserman, Levine. (1992) *A critical appraisal of 98.6 degrees F, the upper limit of the normal body temperature*

Mozquizto 1: Jämför histogrammet med ploten. Hur syns egenskaperna hos data i histogrammet, och tvärtom?

I `figure(2)` ser vi att data är grupperat i män och kvinnor. För att jämföra data mellan flera grupper kan man göra två histogram

```
>> figure(3)
>> subplot(211)
>> histogram( T(T(:,2)==1,1) )
>> subplot(212)
>> histogram( T(T(:,2)==2,1) )
```

där vi använder en indikator `T(:,2)==1` för att plocka ut all data som relaterar till män eller kvinnor. Ett annat alternativ för att jämföra data är att använda `boxplot`

```
>> figure(4)
>> boxplot(T(:,1), T(:,2))
```

Mozquizto 2: Jämför `figure(3)` och `figure(4)`. Är det någon skillnad mellan män och kvinnor? Hur stor verkar skillnaden vara?

Eftersom skillnaden inte är särskilt stor fortsätter vi att analysera alla data på en gång.
Ett annat sätt är att rita de sorterade data, med ordningsnumret på y-axeln:

```
>> figure(5)
>> plot(sort(T(:,1)),1:size(T,1),'.')
```

Uppgift: Jämför denna plot med `figure(1)` och `figure(2)`. Hur hänger de ihop med varandra?

Uppgift: Välj ut några datapunkter i `figure(2)` och försök hitta dem i `figure(1)` och `figure(5)`.

I `figure(5)` kan vi t.ex. avläsa hur många av observationerna som är mindre än eller lika med ett visst tal.

Uppgift: Välj $x = 37$ och försök avgöra i figuren (det går att zooma) hur många av värdena som är mindre än eller lika med 37.

När antalet observationer i stickprovet ökar kan vi tolka kvoten som sannolikheten att få ett värde mindre än eller lika med x. Kvoten kan beräknas så här:

```
>> ratio = mean(T(:,1)<=37)
```

Uppgift: Stämmer det med din uppskattning från figuren?

För att förstå hur `T(:,1)<=37` fungerar så jämför vi det med ursprungsdata:

```
>> T(:,1)
>> T(:,1)<=37
```

Vad är det som händer?

Mozquizto 3: Pröva med några andra värden på x. Hur borde andelen ändra sig när x ökar respektive minskar? Jämför med figuren.

Den omvända proceduren, hitta det värde x som motsvarar en given sannolikhet, dvs en given kvantil, är ofta viktigare. Vi återkommer till det lite senare.
Vi kan naturligtvis låta datorn välja ett stort antal värden att undersöka och sedan försöka få en överblick. Eftersom vi har ett ändligt antal observationer så blir antalet, eller andelen, observationer som än mindre än eller lika med ett visst x-värde en stegfunktion som vi kan rita upp:

```
>> figure(6)
>> stairs(sort(T(:,1)),(1:size(T(:,1),1))/size(T(:,1),1))
>> grid on
```

Den visar hur värdena är fördelade och denna typ av figur kallas *empirisk fördelningsfunktion* (empirical distribution function[2]. För ett värde på x-axeln, t.ex. 37, hittar vi, på y-axeln, andelen värden som är mindre än eller lika värdet på x-axeln.

Uppgift: Kolla att andelen värden som är mindre än eller lika med 37 stämmer med det du fick fram tidigare.

2.2 Kommer data från en standardfördelning?

Histogramet i `figure(1)` och den empiriska fördelningsfunktionen i `figure(6)` kan jämföras med täthet och fördelningsfunktioner för standardfördelningar för att undersöka om någon sådan passar som modell för data.

Uppgift: Påminner histogrammet och fördelningsfunktionen om någon vanliga modell för stokastiska variabler?

Mozquizto 4: Innan vi kan jämföra data med standard fördelningar behöver vi beräkna medelvärde och stickprovsstandardavvikelse

```
>> mu = mean(T(:,1))
>> sigma = std(T(:,1))
```

Vi kan nu rita ut histogrammet och tätheten i samma figur.

```
>> figure(1)
>> histogram(T(:,1), 'Normalization', 'pdf') %normalisera histogramet till area=1
>> x = linspace(35.5,38.5,1e2); %skappa en x-vector
>> hold on
>> plot(x, normpdf(x,mu,sigma)) %plotta tätheten (pdf) i samma figur
>> hold off
```

[2]Fördelningsfunktioner kallas ofta *cumulative distribution functions*.

Ett alternativ är att använda funktionen `histfit`

```
>> histfit(T(:,1))
```

som också tillåter jämförelser med andra täthetsfunktioner (t.ex. exponential)

```
>> histfit(T(:,1),[],'exp')
```

`help histfit` ger en lista på möjliga fördelningar att jämföra med.
På samma sätt kan vi jämföra de teoretiska och emperiska fördelningsfunktionerna.

```
>> figure(6)
>> stairs(sort(T(:,1)),(1:size(T(:,1),1))/size(T(:,1),1),'-')
>> grid on
>> hold on
>> plot(x, normcdf(x,mu,sigma)) %plotta F(x) (cdf) i samma figur
>> hold off
```

Jämför gärna med någon annan fördelning, t.ex. exponential (`exppdf`, `expcdf`).

Mozquizto 5: Vilken fördelning verkar data komma från?

2.3 Större stickprov — Fördelningsfunktionen för en slumpvariabel

En intressant fråga är hur storleken på datamaterialet påverkar hur bra anpassningarna kan förväntas bli. För att undersöka effekten av antalet observationer vill vi nu studera ett större datamaterial, t.ex. 2000 observationer från samma fördelning som tidigare. Eftersom vi bara har 130 observationer simulerar vi ny data och ritar dem i en ny figur:

```
>> data = normrnd(mu, sigma, 1, 2000);
>> figure(7)
>> histogram(data, 'Normalization', 'pdf') %normalisera histogramet till area=1
>> hold on
>> plot(x, normpdf(x,mu,sigma)) %plotta tätheten (pdf) i samma figur
>> hold on

>> figure(8)
>> stairs(sort(data),(1:length(data))/length(data),'.-')
>> hold on
>> plot(x, normcdf(x,mu,sigma)) %plotta F(x) (cdf) i samma figur
>> hold off
>> grid on
```

Pröva även att rita histogram och empirisk fördelningsfunktion för 50, 100 och 500 observationer.

Uppgift: Hur förändras histogram och empirisk fördelningsfunktion när antalet observationer ökar? Hur bra stämmer dessa med sina teoretiska motsvarigheterna?

Uppgift: Vad blir nu andelen värden som är mindre än eller lika med 37?

Eftersom resultatet närmare sig en normalfördelning kan vi beräkna andelen värden som är ≤ 37 som $\mathsf{P}(X \leq 37) = F_X(37)$ där $X \in \mathsf{N}\left(\mu, \sigma\right)$. Använd funktionen `normcdf` för att beräkna $\mathsf{P}(X \leq 37)$. Funktionen `normspec` kan användas för att illustrera sannolikhets beräkningar i en normalfördelad täthet:

```
>> normspec([-Inf 37], mu, sigma)
```

Mozquizto 6: Hur stämmer sannolikhets beräkningen med tidigare beräkningar av andelen värden som är mindre än eller lika med 37 (eller andra värden på x)?

2.4 Kvantiler

Begreppet *kvantil* är viktigt. Kvantilen kan definieras på olika sätt men vi (och många andra) använder följande definition: kvantilen är det tal x_α som uppfyller

$$\mathsf{P}(X \leq x_\alpha) = 1 - \alpha \tag{1}$$

där α är ett tal mellan 0 och 1 (vanliga val är: 0.05, 0.01, 0.001).

Mozquizto 7: Läs av kvantilen $x_{0.05}$ där $\alpha = 0.05$ ur dina figurer, med hjälp av definitionen (1). Både som skattningar i de två empiriska fördelningsfunktionerna och exakt i den teoretiska.

Mozquizto 8: Jämför med det exakta värdet, som kan fås med `norminv(1-0.05, mu, sigma)`.

Uppgift: Hur mycket skiljer sig skattningen av $\mathsf{P}(X \leq 37)$ och $x_{0.05}$ baserat på 50, 100, 500 eller 2000 observationer?

Hur datasetets storlek påverkar osäkerheten i uppskattningarna kommer vi tillbaka till under hela resten av kursen.

2.5 Andra fördelningar

Vi ska nu rita upp några andra normalfördelningar, $\mathsf{N}\left(\mu, \sigma\right)$, och se hur de ändrar sig när vi ändrar på parametrarna μ och σ.

```
>> close all                    % stäng alla gamla figurer
>> x = linspace(0,10,1000);     % Genererar 1000 tal jämnt utspridda
                                % mellan 0 och 10.
>> figure(1)
>> plot(x,normpdf(x,2,0.5))     % N(2, 0.5)
>> hold on                      % Lås plotten, övriga ritas i samma bild.
>> plot(x,normpdf(x,7,0.5))     % N(7, 0.5)
>> plot(x,normpdf(x,5,2))       % N(5, 2)
>> plot(x,normpdf(x,5,0.2))     % N(5, 0.2)
>> hold off                     % Lås upp plotten
>> xlabel('x')
>> title('Täthetsfunktioner, f(x)')

>> figure(2)
>> plot(x,normcdf(x,2,0.5))
>> hold on
```

```
>> plot(x,normcdf(x,7,0.5))
>> plot(x,normcdf(x,5,2))
>> plot(x,normcdf(x,5,0.2))
>> hold off
>> xlabel('x')
>> title('Fördelningsfunktioner, F(x)')
```

Uppgift: Vad händer med fördelningen när μ och σ ändras? Vad representerar μ och σ i fördelningen?

Uppgift: Fördelningsfunktionen är ju integralen av täthetsfunktionen. Relatera dem till varandra i figuren. Hur ändrar sig, t.ex. fördelningsfunktionen när x ligger nära μ jämfört med när x ligger långt från μ? Hur ser täthetsfuktionen ut då (stor eller liten?)

Mozquizto 9: Experimentera med andra värden på μ och σ och se vad som händer. Du kan behöva ändra `x` för att för att få plats i figuren (tips: det allra mesta av en normalfördelning ryms inom $\mu \pm 4\sigma$).

3 Inversmetoden

Ett sätt att få fram slumptal från olika fördelningar är att använda inversmetoden. Då genereras först slumptal från en $\mathsf{U}(0,1)$-fördelning. Dessa stoppas sedan in i inversen till den önskade fördelningens fördelningsfunktion, dvs i
I Matlab kan man få en vektor med n slumptal från en $\mathsf{U}(0,1)$-fördelning genom

```
>> u = rand(n, 1)
```

Uppgift: Gör en vektor u med $n = 1000$ slumptal från denna fördelning. Gör ett histogram och övertyga dig om att det verkar vara rätt fördelning.

Nu vill vi göra om dessa slumptal till att komma från $Exp(\lambda)$-fördelning. Fördelningsfunktionen för en sådan fördelning är som bekant $F_X(x) = 1 - e^{-\lambda x},\ x \geq 0$.

Mozquizto 10: Beräkna fördelningsfunktionens invers och använd den för att transformera slumptalen i vektorn `u` till att komma från en exponentialfördelning med $\lambda = 3$. (dvs lös ut x som funktion av u i $u = 1 - e^{-\lambda x}$.

Uppgift: Gör ett histogram över de exponentialfördelade slumptalen. Ser det rimligt ut? Gör även en empirisk fördelningsfunktion och se så att det verkar vara rätt fördelning (Återanvänd funktioner från avsnitt 2.2).