LABORATION 1
MATEMATISK STATISTIK FÖR B, K, N, BME OCH KEMISTER;
FMSF70 & MASBO2

Syfte:

Syftet med dagens laborationen är att du skall:

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

Kurskompendium:

Olbjer kap 1–3

1 Förberedelseuppgifter

Till laborationens start har du med dig lösningar till uppgifterna nedan. Godkända uppgifter är ett krav för att bli godkänd på laborationen.

1. Antag att du gör n upprepade mätningar av en variabel X och får då ett datamaterial $x_1, \ldots, x_n$.
   (a) Vilka mått kan man använda för att sammanfatta materialet numeriskt?
   (b) Hur kan man beskriva materialet grafiskt?
2. Slumpvariabler brukar delas in i två olika kategorier, diskreta respektive kontinuerliga slumpvariabler.
   (a) Vad karaktäriserar de två olika typerna av variabler?
   (b) Hur beskriver man variationen (fördelningen) för respektive variabel?
3. Hur definieras fördelningsfunktionen, $F(x)$, för en slumpvariabel? Hur kan den beräknas då sannolikhetsfunktion, respektive täthetsfunktion, är given?
4. Skissera täthetsfunktionen för normalfördelningen $\mathsf{N}\left(\mu, \sigma^2\right)$. Vad är tolkningen av μ och σ?

Läs igenom hela handledningen före laborationstillfället. Notera vilka resultat och figurer du behöver till den muntliga redovisningen.

2 Luftföroreningar Hornsgatan

Den 1 Januari 2010 infördes ett dubbdäcksförbud på Hornsgatan i Stockholm. Målsättningen med förbudet var att minska halten av små luftburna partiklar, $PM_{2.5}$ (dvs partiklar mindre än 2.5 μm). Under första delen av laborationen kommer du att titta på mätningar av luftburna partiklar under vintern 2008/2009 (oktober t.o.m. mars) och 2010/2011.

Titta på data

Börja med att ladda in data som ligger i filen `Hornsgatan.mat`.

```
>> load Hornsgatan.mat
>> whos           % visar vilka variabler som finns
```

Variabeln `PM25_2010` innehåller uppmätt dygnsmedelvärde av $PM_{2.5}$ under vintern 2010 (`PM25_2008` mätningar från 2008). Titta på datamaterialet. Det är alltid viktigt att plotta större datamaterial för att kunna upptäcka eventuella konstigheter.

```
>> plot(PM25_2010)         % sammanbinder mätningarna
>> plot(PM25_2010,'*')
>> xlabel('PM 2.5 - Vintern 2010')
```

Finns det några konstigheter i data eller tycks mätsituationen varit under kontroll hela tiden?
Titta också på antalet observationer under de två vintrarna

```
>> length(PM25_2008)
>> length(PM25_2010)
```

Är det samma antal observationer båda åren?
Gör ett histogram (`hist`) över $PM_{2.5}$-halterna

```
>> help hist
>> hist(PM25_2010)
>> hist(PM25_2010, 25)      % fler staplar
```

Beskrivande statistik

Använd funktionerna `mean`, `std`, `min` och `max` för att beräkna medelvärde, standardavvikelse, minsta och största värdet för `PM25_2010`. Jämnför med motsvarande värden för 2008 (`PM25_2008`).
Årsmedelskoncentrationen av $PM_{2.5}$ får inte överstiga 25 μg/m^3. Använd den relativa frekvensen för att uppskatta sannolikheten att dygnsmedelvärdet under vintern 2010 översteg 25 μg/m^3 med

```
>> sum(PM25_2010>=25) %antal observationer >= 25
>> sum(PM25_2010>=25)/length(PM25_2010) %sanolikheten
```

Jämför med den relativa frekvensen för 2008. En framtida målsättning är att årsmedelvärdet inte ska överstiga 8.5 μg/m^3. Uppskatta även sannolikheterna att dygnsmedel under vintern 2010 och 2008 översteg den framtida målsättningen.
För att kunna uttala som om mer udda händelser, prediktera framtida värden, eller mer rigoröst jämnföra mätningar mellan 2008 och 2010 är det ofta lämpligt att anpassa en fördelning till data.

Fördelningar — Simulerad data

För att kontrollera om ett stort datamaterial kan vara normalfördelat används ofta normalfördelningspapper. Nufördtiden är det namnet knappast passande längre eftersom ingen använder papper när det finns datorer. Ordet normalkvantilplot är nog bättre.
Prova först hur ett datamaterial med samma storlek som `PM25_2010` ser ut om det verkligen är normalfördelat, genom att använda simulerade värden. Kommandot `normrnd` simulerar observationer från en normalfördelning, se `help normrnd`. Här simuleras en normalfördelning med väntevärde och standardavvikelse som är medelvärde respektive stickprovsstandardavvikelse av $PM_{2.5}$ mätningarna.

```
>> n = length(PM25_2010);
>> mu = mean(PM25_2010);
>> sigma = std(PM25_2010);
>> X = normrnd(mu, sigma, n, 1);
```

- Rita upp täthetsfunktionen för de simuleradevärdena, X, i intervallet (-5, 25). Rita också upp fördelningsfunktionen för X. (Med kommandot `subplot` kan du få båda graferna i samma figur.) Täthetsfunktion och fördelningsfunktion för normalfördelningen fås genom `normpdf` respektive `normcdf` (använd `help` funktionen eller se stencilen om "Användbara Matlabkommandon").

```
>> x = linspace(-5,25,200)  % x-variabel för täthets- och fördelningsfunktionen
>> subplot(2,1,1)
>> plot(x, normpdf(x,mu,sigma))
>> subplot(2,1,2)
>> plot(x, normcdf(x,mu,sigma))
```

- Använd `normcdf` eller `normspec` för att beräkna sannolikheten att X är
  - större än 25
  - i intervallet (8.5, 25)
  - mindre än 0

  Kontrollera att du förstår vad sannolikhetsberäkningarna innebär grafiskt i täthets- respektive fördelningsfunktionen.

- Vi kan också undersöka hur histogram och trappstegsdiagram (eller kumulativa histogrammet, eller empiriska fördelningsfunktionen) för den simulerade datan stämmer med normalfördelningens täthets- och fördelningsfunktion.

```
>> subplot(2,1,1)
>> histfit(X) %histogram med normaltäthet
>> subplot(2,1,2)
>> cdfplot(X) %trappstegsdiagram
>> hold on
>> plot(x,normcdf(x,mu,sigma),'r') %fördelningsfunktionen
>> hold off
```

- Undersök hur histogrammet och det kumulativa histogrammet ändras nära antalet simuleringar ändras, pröva tex 10, 1 000 och 10 000 simuleringar.

Anmärkning: Givet ett stickprov $x_1, \ldots, x_n$ skattas fördelningsfunktionen med den *empiriska fördelningsfunktionen* $F_n(x)$ som definieras

$$F_n(x) = \begin{cases} 0 & x < x_{(1)} \\ \frac{i}{n} & x_{(i)} \leq x < x_{(i+1)} \\ 1 & x_{(n)} \leq x \end{cases}$$

där $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ betecknar det ordnade stickprovet.

Normalfördelning?

Pröva nu om `PM25_2010` kan vara normalfördelat. En bra ansatts är att göra en normalfördelningsplot för data (`PM25_2010`)

```
>> subplot(2,1,1)
>> normplot(PM25_2010)
```

och/eller undersöka hur en normalfördelning passar till histogrammet för `PM25_2010`.

```
>> subplot(2,1,2)
>> histfit(PM25_2010)
```

Ofta kan man genom att **transformera sina data**, dvs genom att bilda en funktion av dem, få ett datamaterial som verkar vara bättre anpassat till en normalfördelning (eller annan standardfördelning). Vanliga transformationer är $\log(x)$, $\sqrt{x}$ och x^{-1}. Vilken av transformationerna passar bäst för $PM_{2.5}$-mätningarn?
Funktionen `histfit` kan användas för att jämföra histogrammet för data (eller transformerad data) med andra standardfördelningar

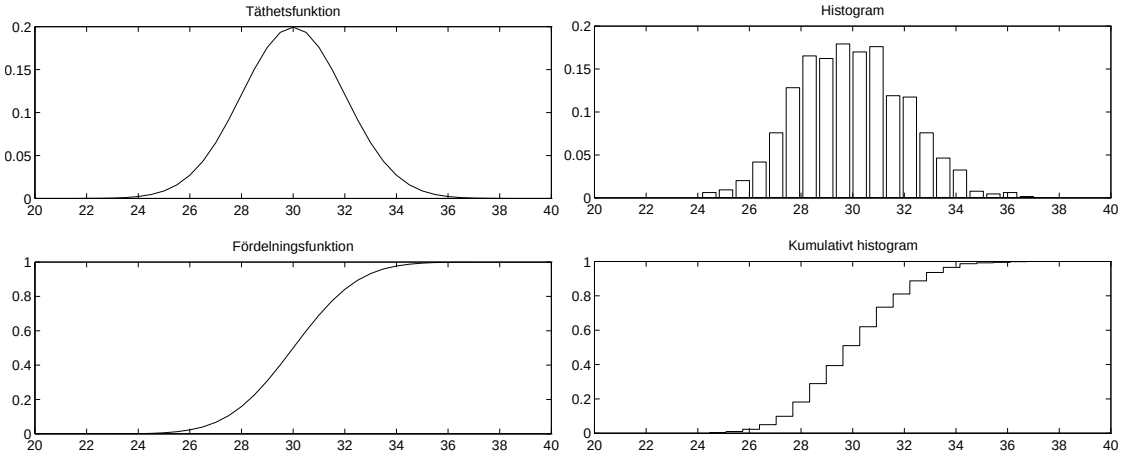
```
>> histfit(PM25_2010,[],'exponential')
```

`help histfit` ger en lista på möjliga fördelningar att pröva. Fundera på sambandet mellan transformationen $\log(x)$ och log-normalfördelningen.

3 Muntlig Redovisning

Diskutera följande frågor med laborationshandledaren

1. Hur kan man avgöra om data är normalfördelad?
2. Vilken av transformationerna gör $PM_{2.5}$-mätningarna approximativt normalfördelade?
3. Vad är en uppskattning av $\mathsf{P}(PM_{2.5} \geq 8.5)$ för 2010 och 2008?
4. Hur påverkar antalet simulerade värden histogrammet och det kumulativa histogrammet?
5. Föklara, med hjälp av figuren nedan, vad beräkningarna av $\mathsf{P}(X < 26)$ och $\mathsf{P}(29 < X < 31)$ innebär grafiskt.



Population Stickprov

Om du har tid över börja gärna med Projekt 1.