Förberedelseuppgifter

1.
    a. ML-skattning av b = $\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i^2}{2}\right)\right)}$
        i. Följde föreläsningens steg
            1. l(theta)
            2. log(l(theta))
            3. Derivatan av log(l(theta)) = 0
    b. MK-skattning av b = $\bar{x} \cdot \sqrt{\frac{\pi}{2}}$
        i. Följde föreläsningens steg
            1. Använde miniräknare för att hitta integral resultat, det var komplicerat.

2. Approximativt konfidensintervall för parametern b:
    a. $\sqrt{\frac{2}{\pi} \cdot \frac{1}{n} \cdot \frac{4-\pi}{2} \cdot MK - skattning}$
    b. $\sqrt{\frac{2}{\pi} \cdot \frac{1}{n} \cdot \frac{4-\pi}{2} \cdot (\bar{x} \cdot \sqrt{\frac{\pi}{2}})}$
        i. $D[{\sigma^2}^*] = \sqrt{V[\sigma^2]} = \sqrt{V[\bar{X}\sqrt{\frac{2}{\pi}}]} = \sqrt{\frac{2}{\pi} \times \frac{1}{n} \times V[X]} = \sqrt{\frac{2}{\pi} \times \frac{1}{n} \times \frac{4-\pi}{2}\sigma^2}$
        ii. σ = b = $\bar{x} \cdot \sqrt{\frac{\pi}{2}}$

3. Simple linear regression is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line. Linear Regression is the process of finding a line that best fits the data points available on the plot, so that we can use it to predict output values for inputs that are not present in the data set we have, with the belief that those outputs would fall on the line.

    `Regress` is used in MatLab as a function for linear regression

1. Simulering av konfidensintervall

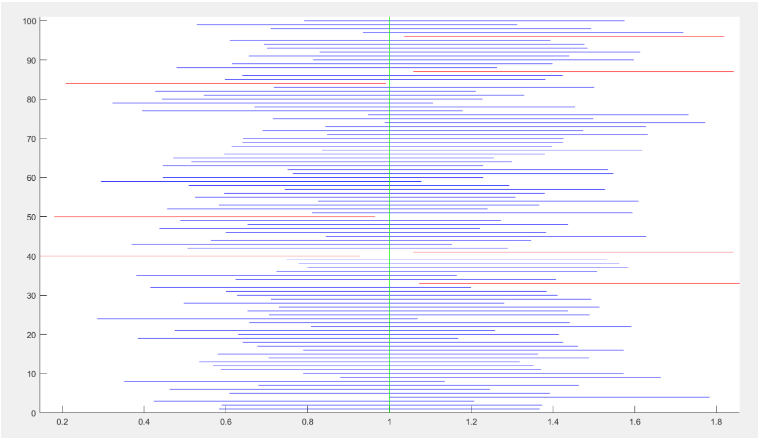
Försök:

1. 8 röda
2. 1 röd
3. 4 röda
4. 9 röda
5. 1 röd
6. 5 röda
7. 3 röda
8. 9 röda

9. 13 röda
10. 8 röda

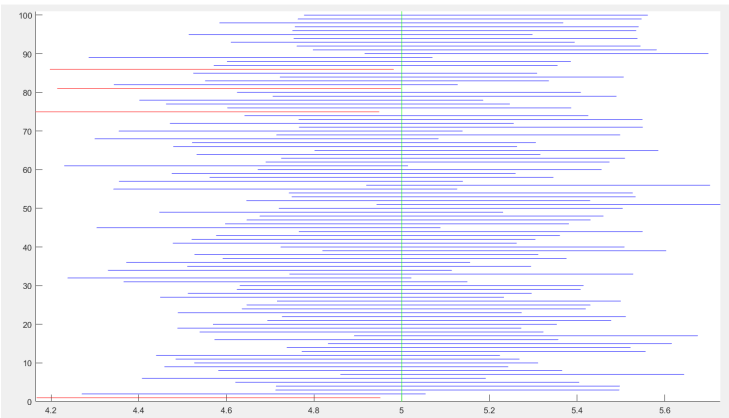
6.1 average - borde ligga vid 5
Vertikal: mu
    Grön = mu, som är väntevärdet
Horisontella: konfidensintervall
    Blå = inom mu, borde vara 95% sannolikhet att det sker
    Röd = innehåller ej mu, borde var 5% sannolikhet att det sker

1.1 Vad som sker om man varierar:
Mu = 1



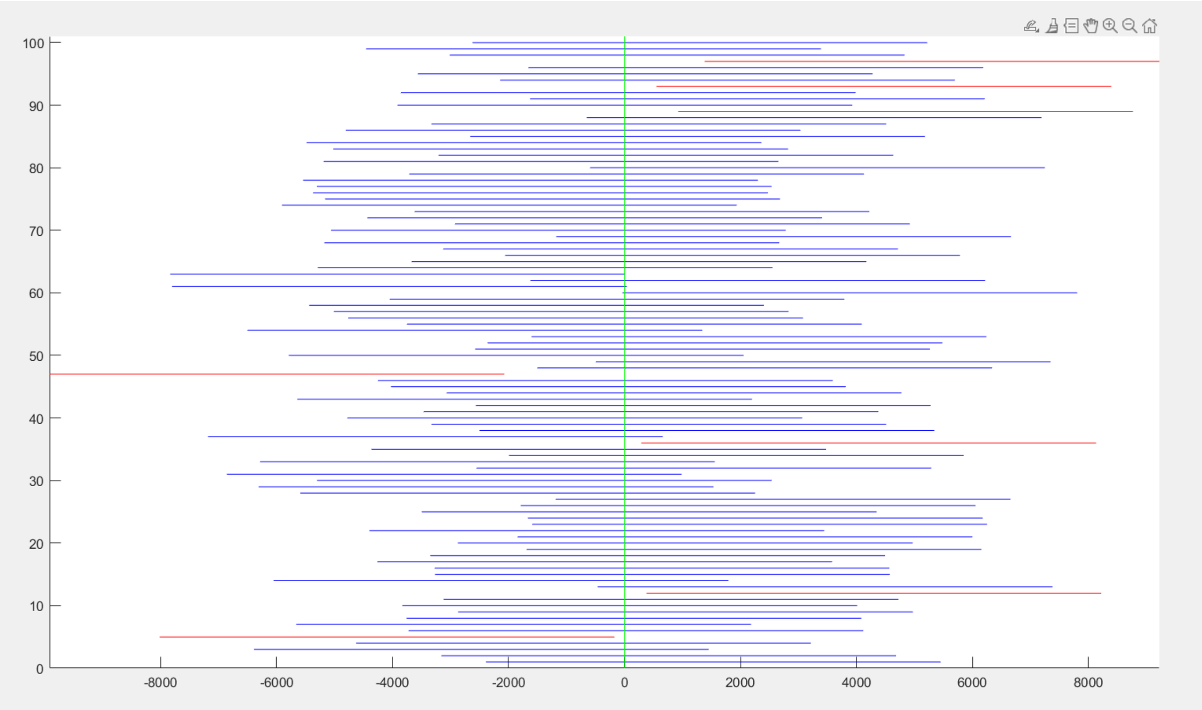
7 röda, verkar normalt - bara väntevärdet som ändras
6 röda, verkar normalt

Mu = 5



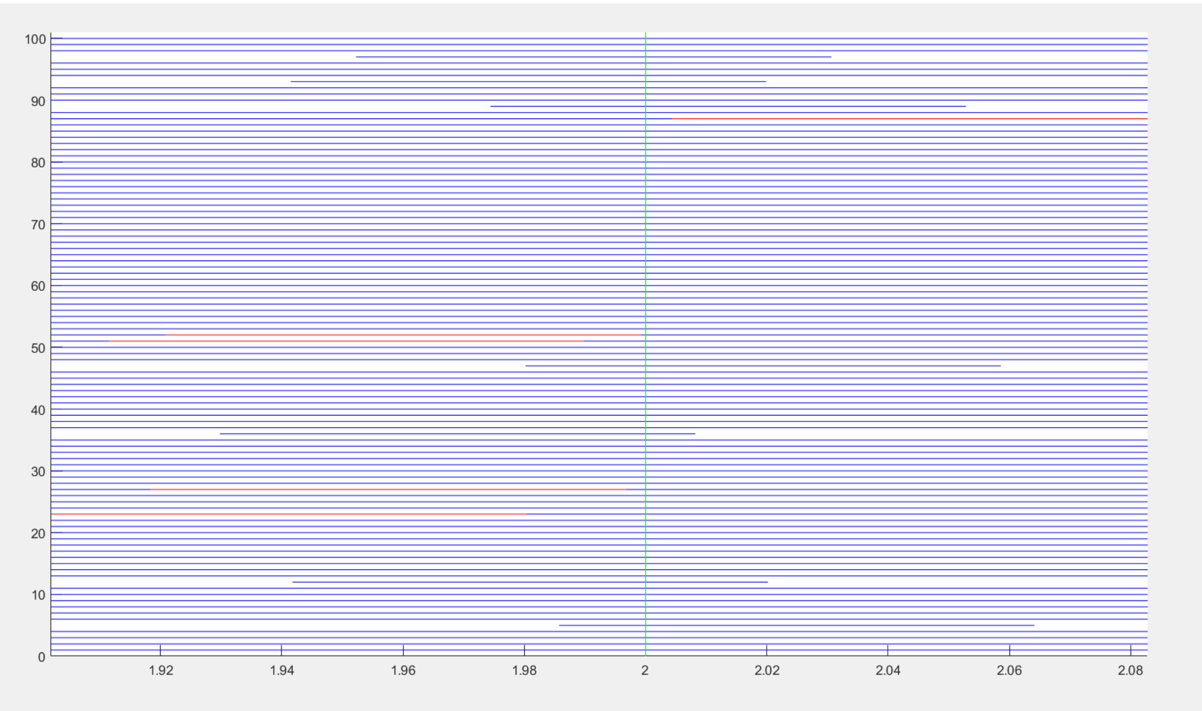
4 röda - samma som innan, bara väntevärdet som ändras

Sigma = 10000

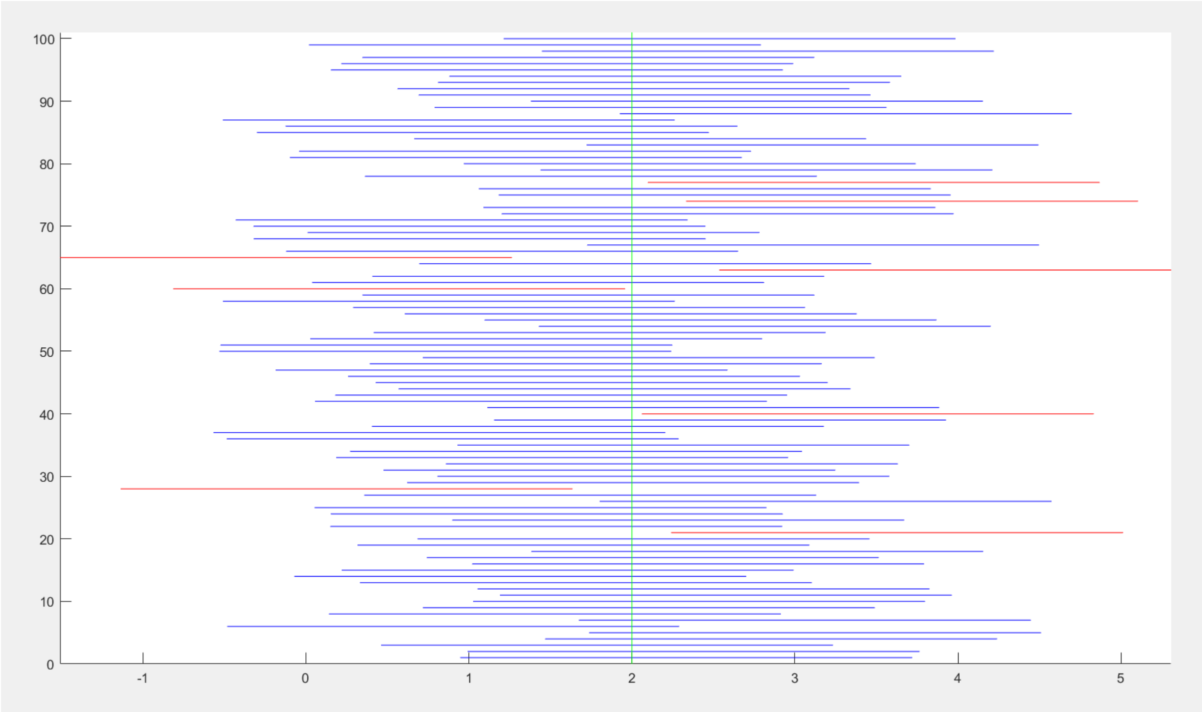


Intervallet ökar till med mer avvikelse (mot (-10,000, 10,000)

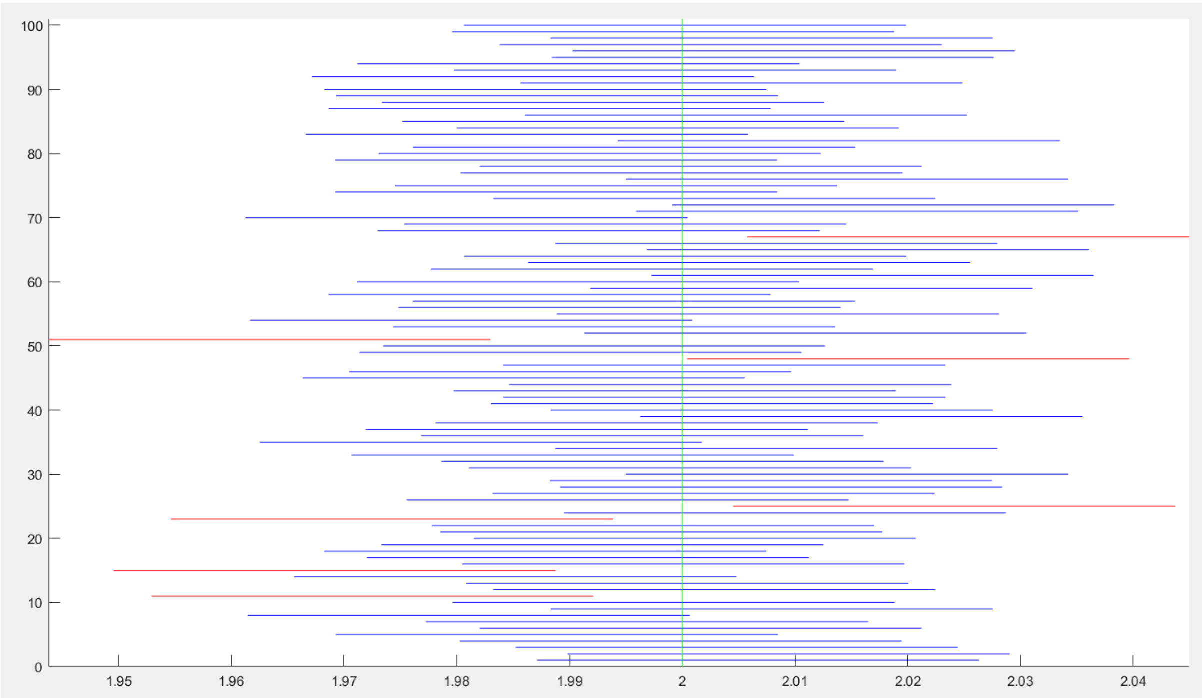
1.2 Sigma = 0.1 -> samma men (1.9, 2.2) - mindre avvikelse



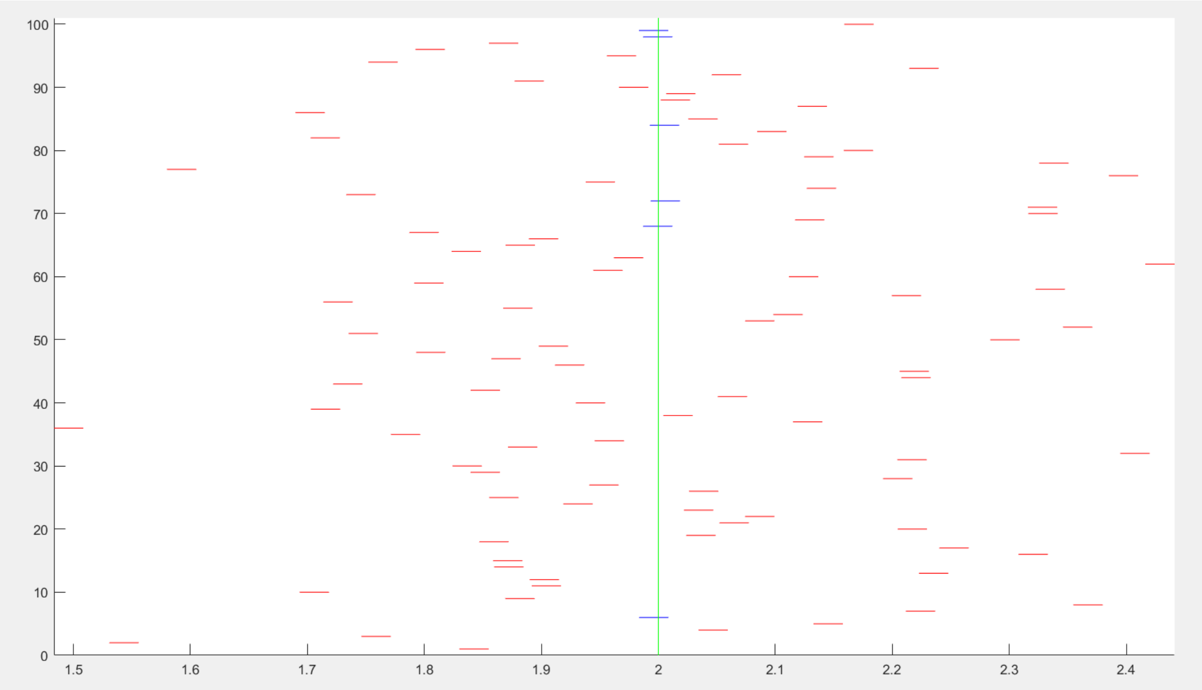
1.3 n = 2 -> konfidensintervall blir sämre (no confidence) antal mätningar är för få för att dra bra slutsats.



n = 10,000 -> bättre intervall närmare mu

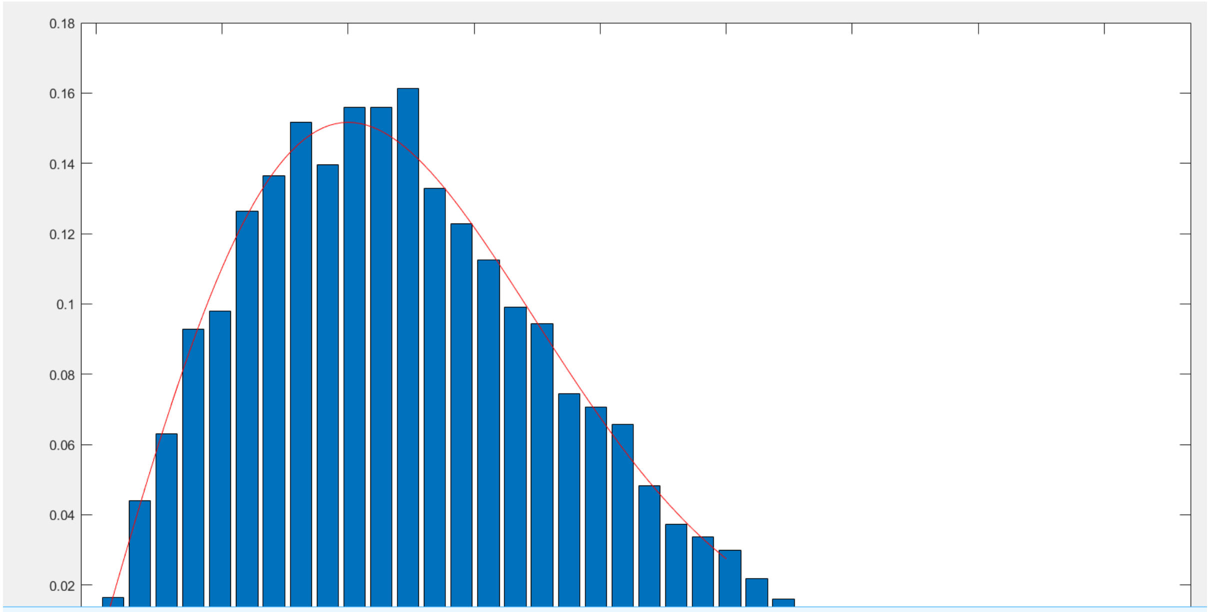


1.4 Alfa = 0.95 -> 95% chans att konfidensintervallet ligger utanför mu



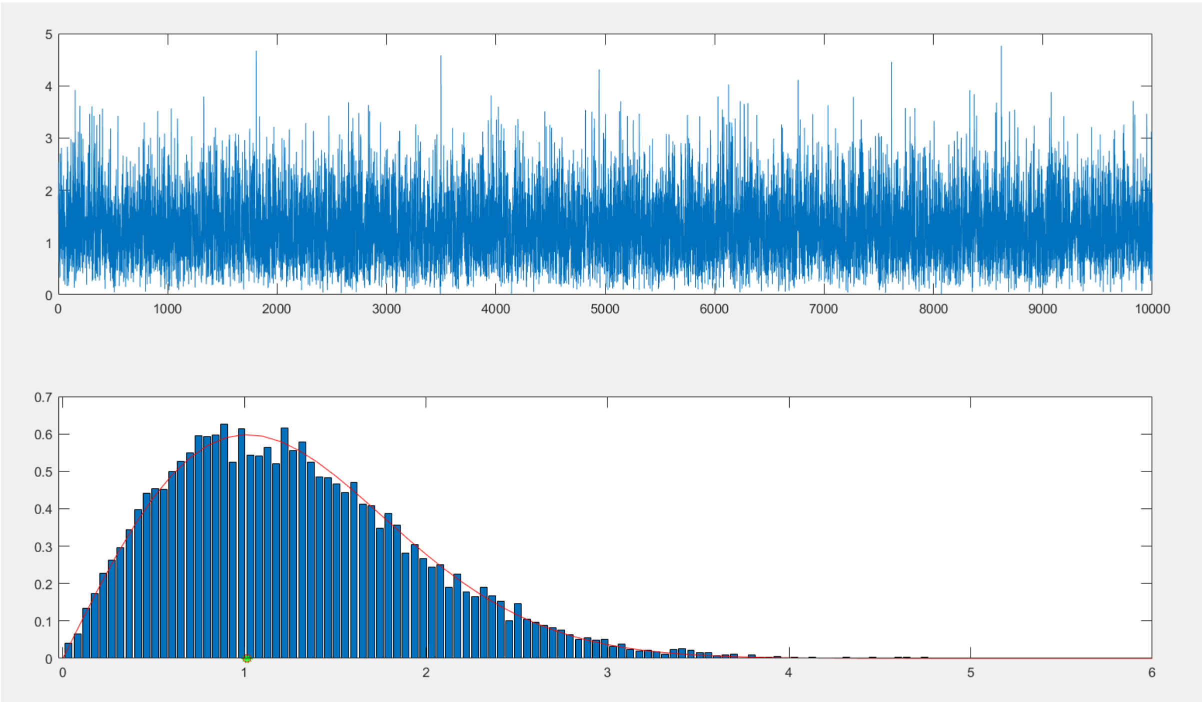
2. Maximum likelihood skattning och minsta kvadrat skattning

Skattningarna ser bra ut, rayleigh fördelningen följs av täthetsfunktionen



3. Konfidensintervall för Rayleigh Fördelning

Täthetsfunktionen passar bra - följer den röda Rayleigh linjen



Name	Value
alpha	0.0500
d	0.0053
lower_bound	1.0038
my_est	1.0141
n	10000
upper_bound	1.0245
y	*1x10000 double*

4. Jämförelse av fördelning hos olika populationer

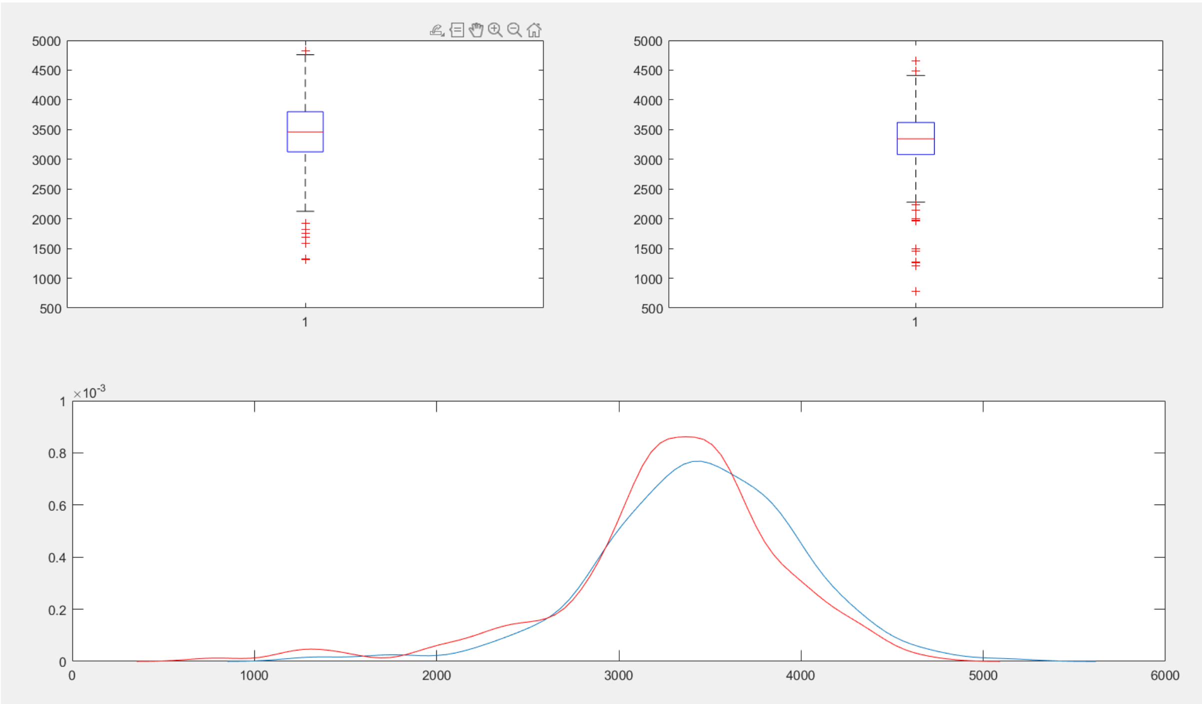
1:a boxen är icke rökare, 2:a är rökare

Röda linjen i blåa boxen är median, blåa boxen 50%

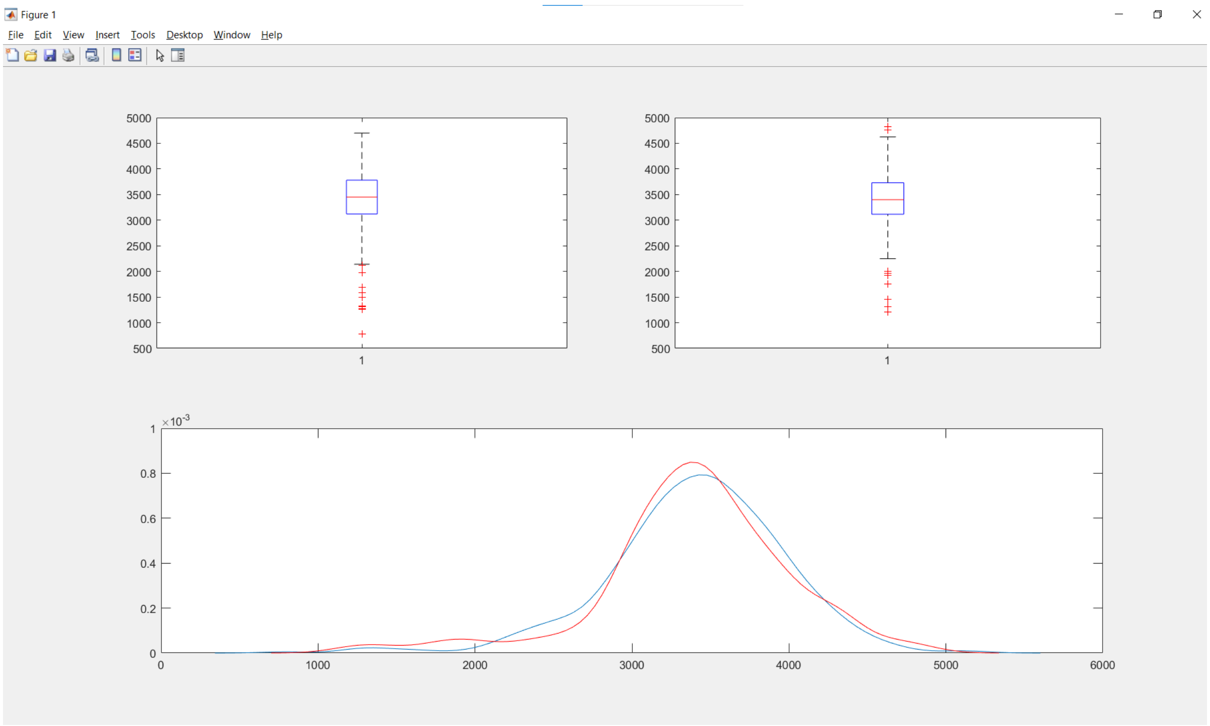
Andra röda linjer är outliers

Grafen nere är en visuell representation av barnens vikt, där röda är rökande mamma och blåa är icke-rökande mamma

Vi ser att rökande mammor fick oftare barn med lägre vikt, men de fick även barn med vanlig och högre vikt fast med en mindre median och barn som väger mer.



Precis som för grafen med rökande mamma, har vi grafen som plottades för mammor som dricker alkohol vs mammor som inte dricker alkohol under graviditeten.
Vi gjorde samma sak, fast ändrade linjen från 20 -> 26 enligt birth.txt

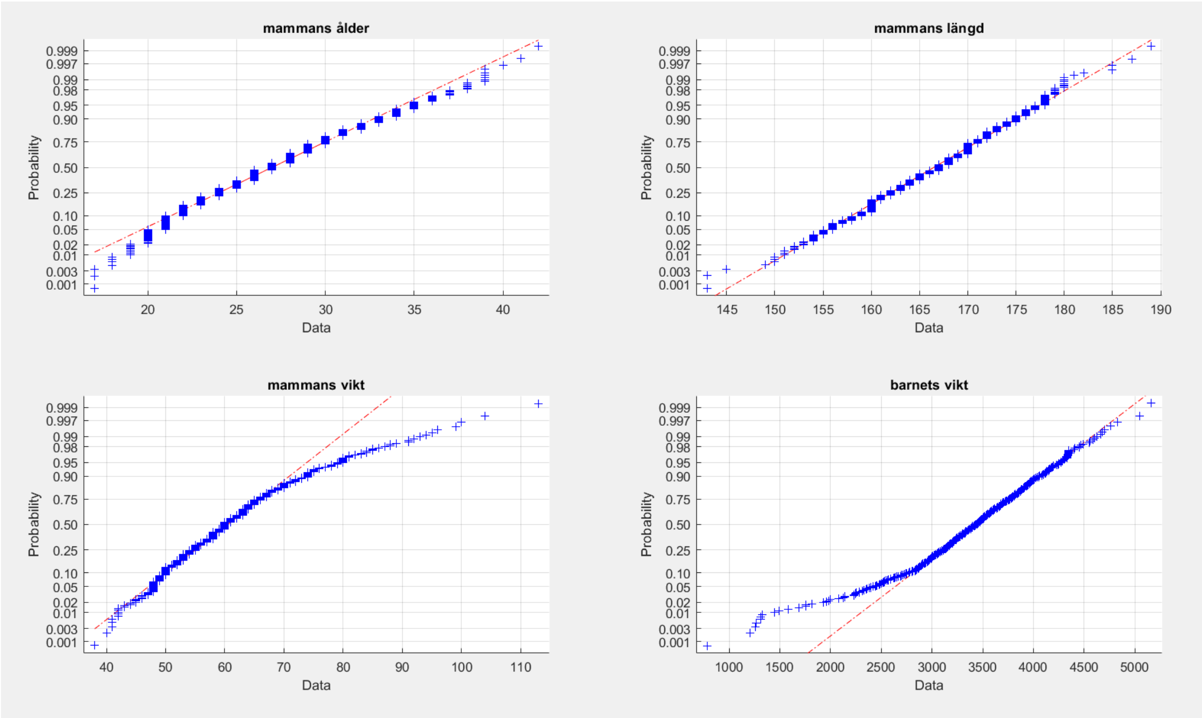


Ungefär samma mönster som rökare, mammor som dricker alkohol under graviditeten har en större chans att få barn som ligger under viktstandarden.

5. Test av normalitet

```
load birth.dat
% Load datan som vi är intresserade över
alder = birth(:, 4);
hojd = birth(:, 16);
viktMamma = birth(:, 15);
viktBarn = birth(:, 3);

% Plotta jämförelse med Normalfördelning

subplot(2,2,1)
normplot(alder), title('mammans ålder')
subplot(2,2,2)
normplot(hojd), title("mammans längd")
subplot(2,2,3)
normplot(viktMamma), title("mammans vikt")
subplot(2,2,4)
normplot(viktBarn), title("barnets vikt")


alder = birth(:, 4);
hojd = birth(:, 16);
viktMamma = birth(:, 15);
viktBarn = birth(:, 3);
procent = 0.05;
% Kollar JB-test
alderJB = jbtest(alder, procent);
hojdJB = jbtest(hojd, procent);
viktMammaJB = jbtest(viktMamma, procent);
viktBarnJB = jbtest(viktBarn, procent);
fprintf('\nMammans ålder är Normalfördelad'), if alderJB==1, fprintf(", är falsk"), end
fprintf('\nMammans längd är Normalfördelad'), if hojdJB==1, fprintf(", är falsk"), end
fprintf('\nMammans vikt är Normalfördelad'), if viktMammaJB==1, fprintf(", är falsk"), end
fprintf('\nBarnets vikt är Normalfördelad'), if viktBarnJB==1, fprintf(", är falsk"), end
```

Visuellt:



Vi tycker att mammans ålder och längd verkar följa normal distributionen (den röda linjen är normal distributionen!) bra medan mammans och barnens vikt inte följer normal distributionen.

JB-test med 5% siknifikansnivå via jbtest:
Vårt resultat tyder på att endast mammans längd är normalfördelad:
Mammans ålder är Normalfördelad, är falsk
Mammans längd är Normalfördelad
Mammans vikt är Normalfördelad, är falsk
Barnets vikt är Normalfördelad, är falsk

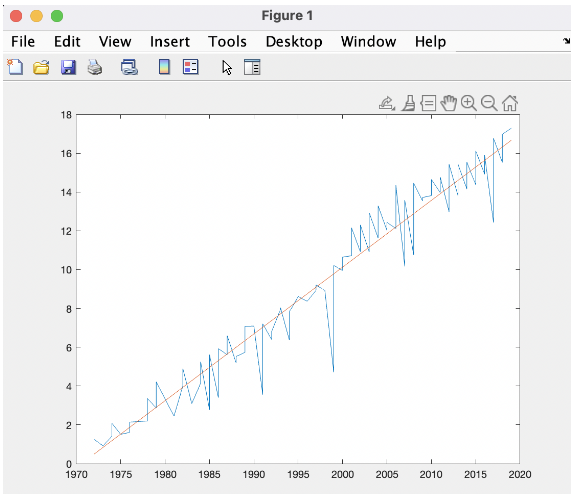
6. Enkel linjär regression

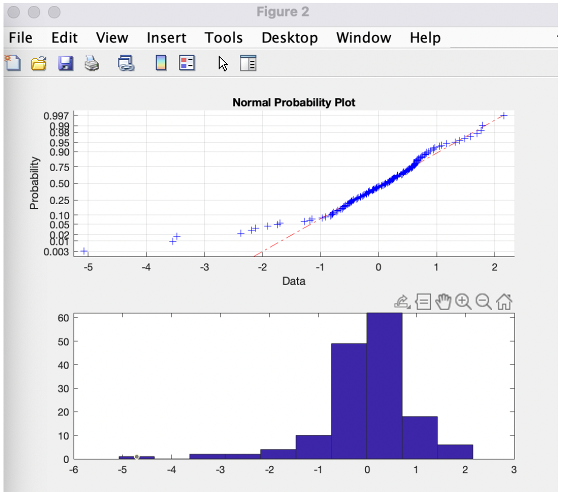
```
load moore.dat

x = moore(:, 1);
y = moore(:, 2);
%Transistorernas värde från expontentiellt till länjärt
w = log(y);

% matris med 1or till vänster och datavärden till höger
X = [ones(length(x),1), x];

[B_circumflex,bint,r,rint,STATS] = regress(w, X);
w_circumflex = X*B_circumflex;
figure
plot(x, w)
hold on
plot(x, w_circumflex)

%Kollar om skillnaden mellan datan (w) och modellens värde
(w_circumflex) (moore's lag) är normalfördelat figure
diff = w-w_circumflex;
subplot(2,1,1), normplot(diff), subplot(2,1,2), hist(diff)

% Få R^2
R2 = STATS(1);
% Funktion för antal transistorer/ytenhet för något år enligt
modellen, använder ekvation för att hitta
```

% $w_i = log(y_i) = \beta_0 + \beta_1 x_i$

```
antal2025 = exp(B_circumflex(1) + B_circumflex(2)*2025);
fprintf('2025 antal: %d \n', antal2025)
fprintf('R^2: %d \n', R2)
```

```
% Funktion för antal transistorer/ytenhet för något år enligt modellen
antal2025 = exp(B_circumflex(1) + B_circumflex(2)*2025);
fprintf('2025 antal: %d \n', antal2025)
fprintf('R^2: %d \n', R2)
2025 antal: 1.359867e+08
R^2: 9.586177e-01
```





Vilken fördelning ser de ut att komma från?
Följer normalfördelningen, “vänsterlutande” där nere
R^2: 9.586177e-01
från 1972 till 2019, vad är då din prediktion för antalet transistorer år 2025?
1.359867e+08 rätt
2021 då? Kolla om det stämmer (dvs om ni hittar själva utfallet från 2021).
3.433789e+07 rätt

7. Multipel linjär regression

```
load birth.dat

%hur mammans längd påverkar barnens vikt, scatter plot
%enkel linjär modell, liten positiv relation
langdM = birth(:,16);   % Mammans längd
viktB = birth(:,3);    % Barns vikt

X = [ones(length(langdM),1), langdM];
B_circumflex = regress(viktB, X);
viktB_circumflex = X*B_circumflex;  % Skattning av barnets vikt

%plottar koden ovan
figure
scatter(langdM, viktB)
hold on
plot(langdM, viktB_circumflex)

%multipel linjär regressionsmodell med alla andra variabler
%innehåller mammans vikt, rökvanor, alkoholvanor i relation till
barnens vikt
viktM = birth(:,15);      % Mammans vikt
rokM = birth(:,20)==3;    % Om mamman röker
drickM = birth(:,26)==2;   % Om mamman dricker

viktB2 = birth(:, 3);      % Barns vikt

X2 = [ones(size(viktM)), viktM, rokM, drickM];
[B,BINT,R] = regress(viktB2, X2);
viktB2_circumflex = X2*B;    % Skattning av barnets vikt

fprintf('%d \n %d \n %d \n %d \n', B, BINT(:,1), BINT(:,2))
% Residualer verklig - regressvärde
figure
normplot(R)
```

Medelvärdet

B: 2.762359e+03 //hur mycket påverkan alla variabler har på barnets vikt
Mammans vikt: 1.130653e+01
Mammans rökvanor: -1.556048e+02
Mammans alkvanor: -1.824818e+01

Konfidensintervall 5%
B: 2.505642e+03
Mammans vikt: 7.187224e+00
Mammans rökvanor: -2.444365e+02
Mammans alkvanor: -1.157097e+02

Konfidensintervall 95%
B: 3.019076e+03
Mammans vikt: 1.542583e+01
Mammans rökvanor: -6.677311e+01
Mammans alkvanor: 7.921330e+01

Vikt påverkar barnen lite positivt
Rökvanor påverkar barnen negativt
Alkvanor påverkar otydligt barnen