Christoffer Mølck

Eks nr: 187952

Std nr: 202009347

Opgave 1

I en glasproduktion inspiceres de færdige glas for fejl. Ved inspektionen findes der erfaringsmæssigt 6.2 glas med fejl pr. time.

a. Hvilken sandsynlighedsfordeling er hensigtsmæssig at anvende til beskrivelse af antal glas med fejl pr. time?
Opskriv det generelle udtryk for denne fordelings sandsynlighedsfunktion (tæthedsfunktion).

Da vi har et gøre med nået pr tid eller pr enhed ville det være en god ide at pro poissons fordeling.

$$p(y) = \frac{\lambda^y}{y!} e^{-\lambda}$$

b. Bestem fordelingens middelværdi, varians og standardafvigelse.

```
u = 6.2 %Middelværdig også ligmed lamda
```

```
u = 6.2000
```

```
varians = u
```

```
varians = 6.2000
```

```
stdafvigelse = sqrt(varians)
```

```
stdafvigelse = 2.4900
```

```

```

c. Bestem sandsynligheden for netop 4 glas med fejl.

```
lambda = 6.2
```

```
lambda = 6.2000
```

```
poisspdf(4, lambda) %den ikke komuleret chance
```

```
ans = 0.1249
```

d. Bestem sandsynligheden for mindst 5 glas med fejl.

```
1 - poisscdf(4, lambda) %chancen må være 1 - den samlet sandsynlighed for 1 - 4.
```

```
ans = 0.7408
```

```
% det ville sige alt over 5 uden grænse ville være
% inkluderet i den chance. Den burde være høj da der generelt
% produceres 6.2 fejl i timen.
%det passer meget godt.
```

e. Bestem sandsynligheden for mindst 5 og højst 9 glas med fejl.

```
%Nu skal vi egentlig gøre det samme men fratrække det overstøe så vi ikke
%får uendelig høj.
(1 - poisscdf(4, lambda)) - (1 - poisscdf(10, lambda))
```

```
ans = 0.6894
```

```
%jeg tænker poisscdf(10) da 9 skal inkluderes.
```

f. Bestem sandsynligheden for netop 10 glas med fejl i et tilfældigt valgt interval på 2 timer.

jeg tænker at vis det er over 2 timer så må mængden af fejl være fordoblet, så lamda = 12.4 fejl.

```
lambda_2t = 6.2 * 2
```

```
lambda_2t = 12.4000
```

```
poisspdf(10, lambda_2t) %NETTOP 10 fejl, så chancen for at få lige precis den mængde fejl er ikke så stor.
```

```
ans = 0.0975
```

a. Beregn sandsynlighederne for følgende:

Propelbladet er intakt, $P(\text{In})$
Propelbladet er defekt, $P(\text{D})$
Propelbladet er fremstillet på maskine A, $P(\text{A})$
Propelbladet er fremstillet på maskine B, $P(\text{B})$
Propelbladet er fremstillet på maskine C, $P(\text{C})$

```
Totalt = 50 + 45 + 45 %Total mængde produceret på alle maskiner
```

```
Totalt = 140
```

```
In_tot = 39 + 41 + 36 %Mængde produceret som er OK på alle maskiner
```

```
In_tot = 116
```

```
D_tot = 11 + 4 + 9 %Mængde produceret som er er defejte på alle maskiner
```

```
D_tot = 24
```

```
P_in = In_tot/Totalt %Chance for at delen er OK
```

```
P_in = 0.8286
```

```
P_D = D_tot/Totalt %Chance for delen er defekt
```

```
P_D = 0.1714
```

```
%Eller
P_D = 1 - P_in %Chance for delen er defekt
```

```
P_D = 0.1714
```

```
P_A = 50 / Totalt %Mængden produceret på maskine A
```

```
P_A = 0.3571
```

```
P_B = 45 / Totalt %Mængden produceret på maskine B
```

```
P_B = 0.3214
```

```
P_C = 45 / Totalt %Mængden produceret på maskine C
```

```
P_C = 0.3214
```

```
P_A + P_B + P_C %Skal give 1, sanity check
```

```
ans = 1
```

```

```

Antag, at der er uafhængighed mellem kvaliteten af de fremstillede propelblade, og hvilken maskine propelbladene er fremstillet på.

b. Udfyld nedenstående tabel med forventede antal propelblade pr. døgn. Nødvendige mellemregninger skal fremgå.

Tabel over forventede antal propelblade pr. døgn:

Vis det er over samme tid, så ændre total mængden sig vel ikke? Vis der er uafhængighed imellem kvaliteten og hvilke maskine det bliver produceret på så må det betyde at det er tilfældigt hvor de defekkte propeller bliver produceret og hvor de gode bliver produceret.

	A	B	C	Alt
In	= 116/3	= 116/3	= 116/3	116
D	= 24/3	= 24/3	= 24/3	24
Tot	50	45	45	50 + 45 + 45 = 140

Det skal nu undersøges ved hjælp af en hypotesetest om kvaliteten af de fremstillede propelblade er uafhængig af hvilken maskine propelbladene er fremstillet på. Ved hypotesetesten anvendes et signifikansniveau på 10%.

c. Opstil nulhypotese og alternativ hypotese for hypotesetesten.

Der skal undersøges om 2 variabler er uafhængige så en chi i anden test bruges.

Til denne opsættes hypostesen:

H0 = Kvaliteten af det produceret propelblad er afhængig af maskinen

Ha = Kavliteten af det produceret propelblad er IKKE afhængig af maskinen

```
Tabl = [[39, 11];[41, 4];[36, 9]]
```

```
Tabl = 3×2
    39    11
    41     4
    36     9
```

d. Opstil en formel for teststørrelsen (teststatistikken), og angiv hvilken fordeling den følger.

```
displayFormula("chi_0^2 = (Sigma_i)^k*(Sigma_j)^k*((O_ij - E_ij)^2/E_ij)")
```

$$\chi_0^2 = \Sigma_i^k \Sigma_j^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Da det det er antal fejl over en vis mængde tid, er det formentlig en poission fordeling.

e. Bestem den kritiske værdi og angiv det kritiske område for testen, når der vælges et signifikansniveau på 10%.

f. Beregn teststørrelses (teststatistikkens) værdi. Mellemregninger skal fremgå.

```
forv = KontingensTabel(Tabl) %Først lave en kontigens tabel til testen.
```

```
     Data navn          In | D
    ____________    ____________

    "Maskine A"     41.429    8.5714
    "Maskine B"     37.286    7.7143
    "Maskine C"     37.286    7.7143
forv = 3×2
   41.4286    8.5714
   37.2857    7.7143
   37.2857    7.7143
```

```
Tabl = [[39, 11] ;[41, 4];[36, 9]]'
```

```
Tabl = 2×3
    39    41    36
    11     4     9
```

```
forv = [forv(1,:); forv(2,:); forv(3,:)]'
```

```
forv = 2×3
   41.4286   37.2857   37.2857
    8.5714    7.7143    7.7143
```

```
chi2normal_Test_8_uafh(Tabl, forv, 10)
```

```
Formel for teststørrelsen
```

$$\chi_0^2 = \Sigma_i^k \Sigma_j^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

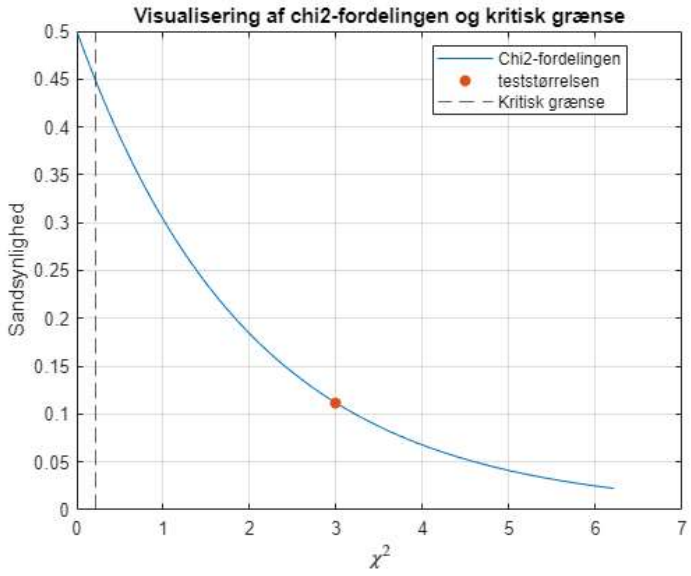
```
-----------------------------------------------------------------
Antal frihedsgrader
```

$$\mathrm{dfs} = (r-1)\,(c-1)$$

```
-----------------------------------------------------------------
Kritiske grænse - MATLAB kommando
```

$$\chi_\alpha^2 = \mathrm{chi2inv}\,\alpha\,\mathrm{dfs}$$

```
-----------------------------------------------------------------
    Signifikansniveau    Frihedsgrader    Kritisk grænse    Teststørrelse
    _________________    _____________    ______________    _____________

          "%90"                2              0.21072          2.9888
```



g. Konkluder på hypotesetesten.

Det kan hurtigt se at testørelsen > kritiske grænse**.** Derfor må vi forkaste H0 og sige at Ha er sand, det ville sige at kavliteten af det produceret propelblad er IKKE afhængig af maskinen og derfor godt kan producere på alle maskiner lige meget uden at være bekymret for at nogle af dem laver flere fejl en andre**.**

h. Bestem p-værdien.

```
p = 1 - chi2cdf(2.988, 2)
```

```
p = 0.2245
```

```
data = importdata("Christoffer\StatEksamen\Data_M4STI1_2023F.xlsx", "H:I")
```

```
data = struct with fields:
        data: [14×7 double]
    textdata: {4×9 cell}
```

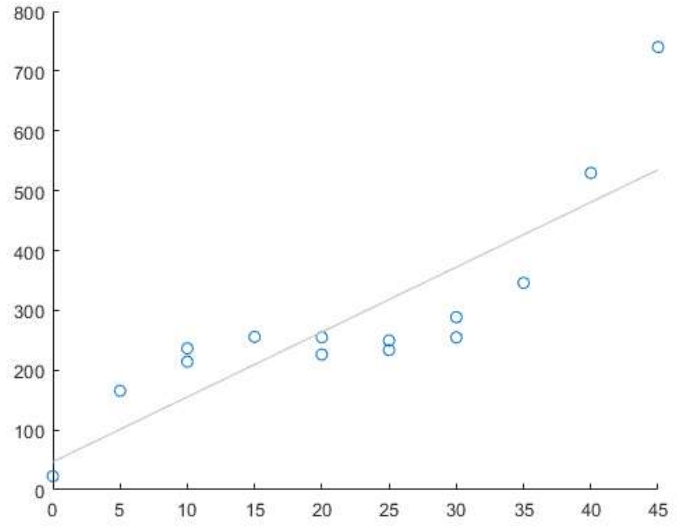
```
data = data.data(:,6:7);
```

```
data = 14×2
         0   22.7000
    5.0000  165.5000
   10.0000  236.5000
   10.0000  214.3000
   15.0000  255.6000
   20.0000  226.1000
   20.0000  255.0000
   25.0000  249.7000
   25.0000  233.6000
   30.0000  254.8000
```

```
data_x = data(:,1);
data_y = data(:,2); %Nået data import sjov.
```

a. Lav en lineær regression med y som funktion af x og skriv regressionsligningen op.

```
%Jeg laver lige et scatter plot først får at få overblik over data
figure(1)
scatter(data_x, data_y)
lsline
```




```
%Ja det gør det jo ret tydligt at linær nok ikke ligefrem er den beste
%model til denne data.

mdl = fitlm(data_x, data_y)
```

```
mdl =
Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate      SE       tStat       pValue
                   ________    ______    _______    __________

    (Intercept)      46.73     49.296    0.94795       0.36186
    x1               10.851    1.9283     5.6275    0.00011118


Number of observations: 14, Error degrees of freedom: 12
Root Mean Squared Error: 92.2
R-squared: 0.725,  Adjusted R-Squared: 0.702
```

```
syms x
y(x) = 46.73 + 10.851 * x %ville være ligningen
```

```
y(x) =
```

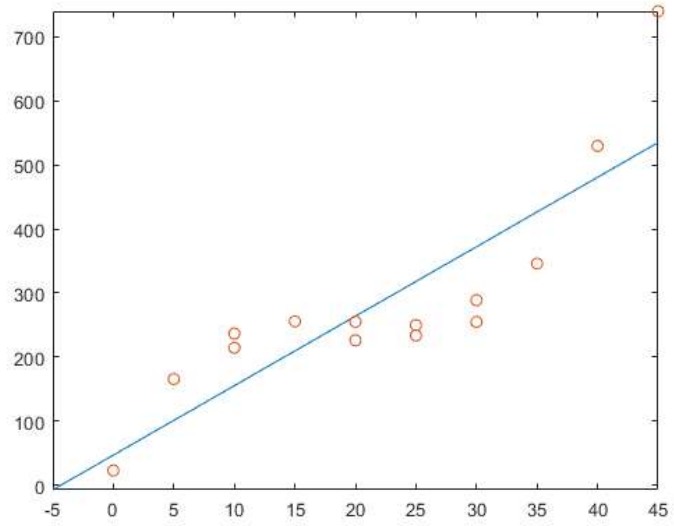
$$\frac{10851\,x}{1000}+\frac{4673}{100}$$

b. Vurder ud fra regressionsanalysens statistikker (f.eks. R^2, F og p-værdier), om modellen beskriver observationerne godt.

Ud fra r^2 adjusted kan vi se at den ligger på en 0.7 hvilket ikke er fantastisk og betyder at vores model nok ikke passer særlig godt til dataen.

c. Lav et plot, der viser data og regressionsligningen. Diskutér sammenhængen mellem regressor- og responsvariablen.

```
fplot(y(x))
hold on
scatter(data_x, data_y)
hold off
```



plottet er heldigvis identisk med sanity plottet jeg lavet først i ogave a. Det burde den da også helst være men godt nok lige at tjække at alt stemmer overens.

Sammenhængende imellem regressor og responsvariabel ses som at være positiv da vi ser en positiv trænd i både fit linjen men også i scatter dataen.

d. Lav en polynomiel regression, der udtrykker y som et tredjegradspolynomium af x:

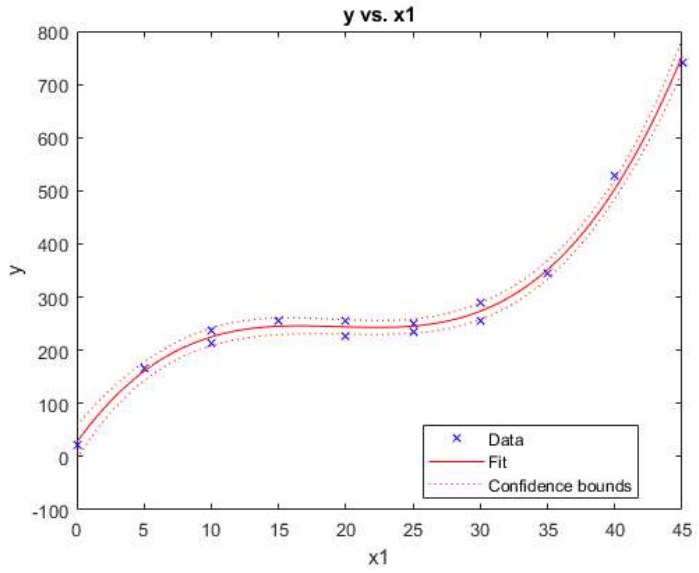
$$y = b_0 + b_1 x + b_2 x^2 + b_3 x^3$$

hvor b_0, b_1, b_2 og b_3 er konstanter.
Skriv regressionsligningen op.

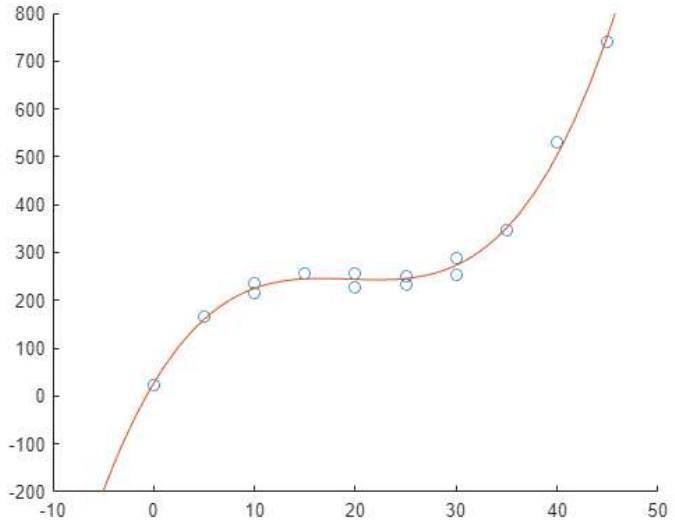
```
polyfit = fitlm(data_x, data_y, 'y ~ x1 + x1^2 + x1^3')
```

```
polyfit =
Linear regression model:
    y ~ 1 + x1 + x1^2 + x1^3

Estimated Coefficients:
                   Estimate       SE         tStat       pValue
                   ________    _________    _______    _________

    (Intercept)      27.325       14.127     1.9343     0.081848
    x1               35.085       2.6749     13.116    1.2604e-07
    x1^2            -1.8455      0.13924    -13.254     1.141e-07
    x1^3           0.031651    0.0020149     15.708    2.2424e-08


Number of observations: 14, Error degrees of freedom: 10
Root Mean Squared Error: 15.7
```

```
plot(polyfit)
```




```
scatter(data_x, data_y)
y(x) = 27.325 + 35.085 * x + -1.8455 * x^2 + 0.031651 * x^3;
hold on
fplot(y(x))
hold off %Sanity check, det passer ganske fint
```



```
y(x) = 27.325 + 35.085 * x + -1.8455 * x^2 + 0.031651 * x^3 % ville være vores polynomie
```

y(x) =

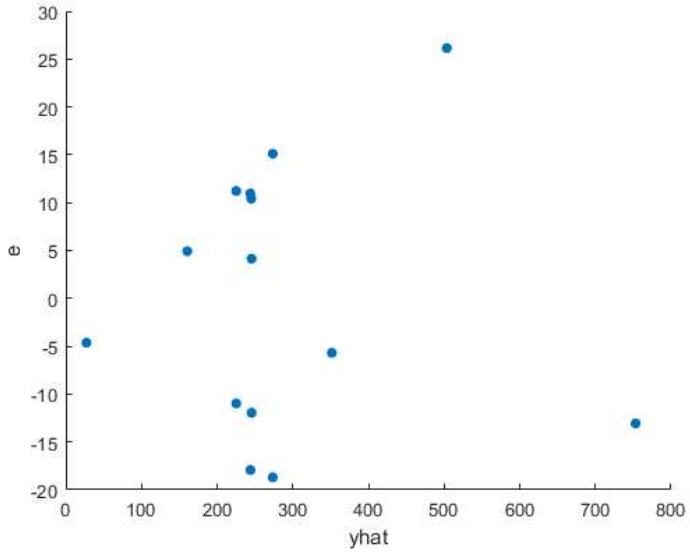
$$\frac{2280694908894457\,x^3}{72057594037927936} - \frac{3691\,x^2}{2000} + \frac{7017\,x}{200} + \frac{1093}{40}$$

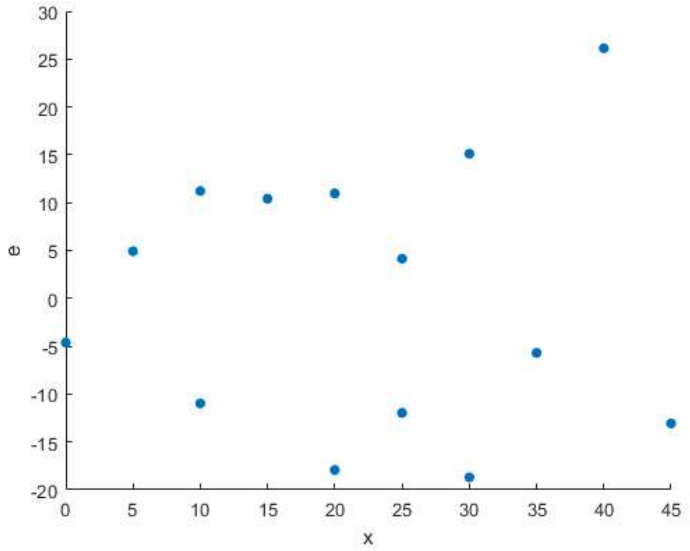
e. Er den polynomielle regressionsmodel bedre end den lineære model? Begrund dit svar.

Ja markant bedre. R^2 korrigeret er helt oppe på 0.991 hvilket er en rigtig godt passende model.

f. Lav et residualplot, der viser studentiserede residualer mod $\hat{y}$. Diskutér resultatet.

```
ris = STAT.Residual(polyfit, [data_x, data_y])
```






```
ris = struct with fields:
         data: [14×9 table]
            k: 1
            n: 14
    lev_limit: 0.2857
    rst_limit: 3
```

Jammen det ser sådan set ganske findt ud. Risidualerne ligger tilfældigt hvilket er som ønsket og viser at poly modellen passer godt med dataen. Der er ingen trends.

g. Undersøg om der er unormale observationer i datasættet og angiv eventuelle unormale observationer med deres type (outliers, løftestangs-punkter eller indflydelsespunkter).

```
ris.data
```

ans = 14×9 table

	x	y	lev	rst	levrage	outlier	yhat	yci	
1	0	22.7000	0.8091	-0.6545	1	0	27.3255	-4.1508	58.8018
2	5	165.5000	0.2511	0.3464	0	0	160.5699	143.0333	178.1066
3	10	236.5000	0.2247	0.7966	0	0	225.2771	208.6878	241.8665
4	10	214.3000	0.2247	-0.7780	0	0	225.2771	208.6878	241.8665
5	15	255.6000	0.2019	0.7245	0	0	245.1850	229.4604	260.9096
6	20	226.1000	0.1479	-1.2750	0	0	244.0314	230.5720	257.4908
7	20	255	0.1479	0.7393	0	0	244.0314	230.5720	257.4908
8	25	249.7000	0.1445	0.2719	0	0	245.5543	232.2530	258.8556
9	25	233.6000	0.1445	-0.8086	0	0	245.5543	232.2530	258.8556
10	30	254.8000	0.1981	-1.3895	0	0	273.4916	257.9169	289.0663
11	30	288.6000	0.1981	1.0836	0	0	273.4916	257.9169	289.0663
12	35	345.9000	0.2337	-0.3954	0	0	351.5812	334.6644	368.4979
13	40	529.7000	0.2734	2.3548	0	0	503.5609	485.2644	521.8573

	x	y	lev	rst	levrage	outlier	yhat	yci	
14	45	740.1000	0.8002	-2.1851	1	0	753.1686	721.8651	784.4721

```
ris.data.levrage(1)
```

```
ans = 1
```

```
ris.data.levrage(14)
```

```
ans = 1
```

Det ses at vi har 2 løftestangs punkter (levrage) ved x = 0 og x = 45. Ud fra risidual plottet virker de ikke til at være gale nok til at være til bekymring og kan ignoreres uden problemer. Nok grundet af de ligger ved polynomiets kraftigste hældning som får y værdigen til at ændre sig kraftigt på meget få x.

h. Beregn den forventede værdi af y, når $x = 27$. Beregn et interval for værdier af y, hvor 95 % af målinger med $x = 27$ må forventes at ligge indenfor.

```
vpa(y(27), 5)
```

```
ans = 252.24
```

```
coefCI(polyfit, 0.05)
```

```
ans = 4×2
   -4.1508   58.8018
   29.1250   41.0453
   -2.1557   -1.5353
    0.0272    0.0361
```

```
y_h(x) = 58.8018 + 41.0453 * x + -1.5353 * x^2 + 0.0361 * x^3
```

y_h(x) =

$$\frac{361\,x^3}{10000} - \frac{15353\,x^2}{10000} + \frac{180519138462217\,x}{4398046511104} + \frac{4137808821386163}{70368744177664}$$

```
y_l(x) = -4.1508 + 29.1250 * x + -2.1557 * x^2 + 0.0272 * x^3
```

y_l(x) =

$$\frac{17\,x^3}{625} - \frac{21557\,x^2}{10000} + \frac{233\,x}{8} - \frac{10377}{2500}$$

```
vpa(y_h(27), 5) - vpa(y(27), 5)
```

```
ans = 506.11036700004478916525840759277
```

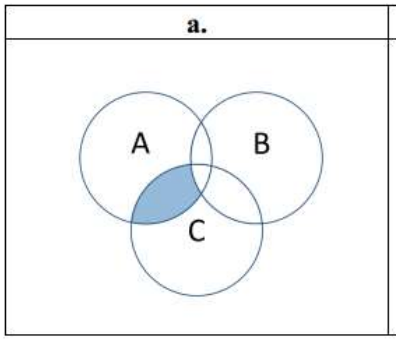
```
vpa(y(27), 5) - vpa(y_l(27), 5)
```

```
ans = 506.14063299997360445559024810791
```

```
%Det ville sige at 252.42 +- 506.1 cirka. hvilket ikke lyder helt
%rigtigt...
```

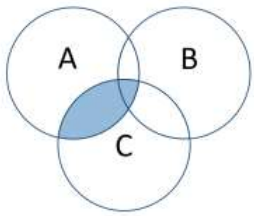
Beskriv på tilsvarende måde det blå område for hver af nedenstående tre Venn diagrammer med et eller flere af følgende fem udtryk:

a.



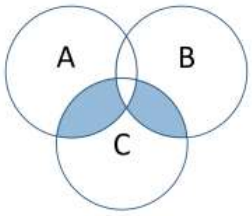
2) $(A \cap C) \cap B^c$

b.



1) $A \cap C$

c.



4) $(A \cap C \cap B^c) \cup (B \cap C \cap A^c)$