Phd-kursus i Basal Statistik, Opgaver til 2. uge

Opgave 1: Sædkvalitet

Filen `oeko.txt` på hjemmesiden indeholder datamateriale til belysning af forskellen i sædkvalitet mellem SAS-ansatte og mænd, der lever økologisk (i det følgende ofte blot omtalt som økologer). Variablene er (i den nævnte rækkefølge):

- `sas_ansat`: ansat i SAS (ja/nej)
- `abstid`: abstinenstid (1: kort, 2: medium, 3: lang)
  (et mål for længden af seksuel afholdenhed)
- `konc`: sædkoncentrationen (mill/ml)

Formålet med opgaven er at undersøge, om der er forskel på de to populationsgruppers sædkoncentrationsniveau.

Vi indlæser data (i form af `txt`-filen direkte fra hjemmesiden), og foretager samtidig en logaritmetransformation, fordi det viser sig, at vi senere kan få brug for dette:

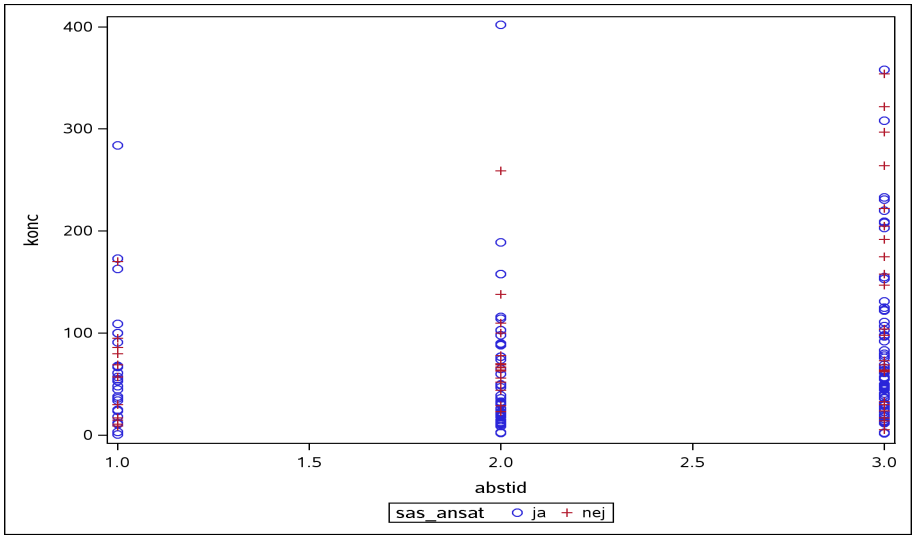
```
FILENAME navn URL "http://biostat.ku.dk/~lts/basal/data/oeko.txt";

data oeko;
infile navn firstobs=2;
input sas_ansat $ abstid konc;

lkonc=log10(konc);

/* variablen gruppe er beskrevet i spørgsmål 1 og 4a */
saskode=(sas_ansat='ja');
gruppe=10*saskode+abstid;

run;
```

1. **Lav en passende illustration af data.**

   En umiddelbar optegning af sædkoncentration mod abstinenstid, med farveangivelse for grupperne kan udføres ved at skrive som nedenfor:

```
proc sgplot data=oeko;
    scatter x=abstid y=konc / group=sas_ansat;
run;
```

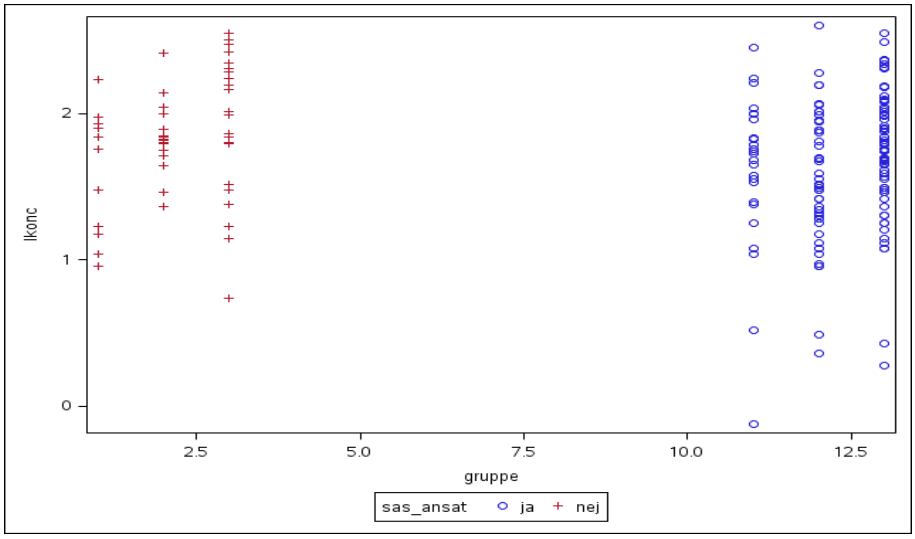
hvorved vi får figuren



Heraf ses, at sædkoncentration næppe er normalfordelt inden for hver gruppe, og vi laver derfor en logaritmetransformation (her er brugt `log10`). Samtidig skiller vi observationerne ud i 6 grupper, så vi bedre kan se, hvad der er hvad, og hertil benytter vi den variabel, vi kaldte `gruppe` og som blev dannet i forbindelse med indlæsningen vha de to viste sætninger.

```
proc sgplot data=oeko;
    scatter x=abstid y=lkonc / group=gruppe;
run;
```

Variablen `gruppe` har (som det ses nedenfor) værdierne 1,2,3 (for økologerne, abstinenstid 1,2 og 3) og 11,12,13 (for de SAS-ansatte, abstinenstid 1,2 og 3), og figuren ser nu således ud:



På denne skala ser både normalfordelingsantagelse og varianshomogenitet rimelig fornuftig ud, omend ikke perfekt.

2. **Vi skal nu kvantificere niveauet af sædkoncentration for de to grupper af mænd og sammenligne disse niveauer, i første omgang uden at tage hensyn til abstinenstiden. Overvej, om der skal logaritmetransformeres, når I svarer på nedenstående spørgsmål:**

   (a) *Giv et estimat for nivauet af sædkoncentrationen for hver af de to grupper af mænd. Husk et 95% konfidensinterval.*

   Som estimat for niveauet vil vi jo umiddelbart anvende gennemsnittet, men da fordelingen er skæv, vil det nok være mere passende at benytte medianen, eller at transformere til logaritmisk skala. Vi gør det hele på en gang nedenfor:

```
proc means N mean median stddev stderr clm data=oeko;
class sas_ansat;
var konc lkonc;
run;
```

   hvorved vi får

```
The MEANS Procedure

              N
sas_ansat   Obs  Variable    N          Mean        Median       Std Dev
-------------------------------------------------------------------------
ja          135  konc      135    71.2845185    48.0000000    70.8536049
                 lkonc     135     1.6486399     1.6812412     0.4715158

nej          53  konc       53   100.9547170    69.0000000    86.9397173
                 lkonc      53     1.8345764     1.8388491     0.4197763
-------------------------------------------------------------------------

              N                                    Lower 95%     Upper 95%
sas_ansat   Obs  Variable     Std Error    CL for Mean    CL for Mean
-------------------------------------------------------------------------
ja          135  konc         6.0981074     59.2235247     83.3455123
                 lkonc        0.0405816      1.5683766      1.7289033

nej          53  konc        11.9420886     76.9911630    124.9182709
                 lkonc        0.0576607      1.7188718      1.9502811
-------------------------------------------------------------------------
```

De estimerede middelværdier (med tilhørende 95% konfidensinterval) ses at være

Data	Gruppe	gennemsnit	SEM	95% konfidensinterval (CI)
utransformeret	SAS	71.28	6.10	(59.22, 83.35)
	Øko	100.95	11.94	(76.99, 124.92)
log10-transformeret	SAS	1.649	0.0406	(1.568, 1.729)
	Øko	1.835	0.0577	(1.719, 1.950)
tilbagetransformeret	SAS	44.56	-	(36.98, 53.58)
	Øko	68.39	-	(52.36, 89.13)

Tilbagetransofrmerede "gennemsnit" = kaldes geometriske gennemsnit

Tilbagelogaritmering:
log(a)-log(b)=log(a/b)
Derfor efter tilbagelogaritmering har vi en ratio

Til sammenligning kan det anføres, at medianerne i de to grupper er hhv. 48 (SAS-ansatte) og 69 (økologer), hvilket ses at passe en del bedre med de estimater, der fremkommer ved at tilbagetransforme gennemsnittene på logaritmisk skala, i forhold til de gennemsnit, der er lavet direkte på den utransformerede skala.

Diff = 0.18
10^(-0.18)=0.65 —> a/b=0.65 —> a=0.65b

(b) *Sammenlign de to estimater og de to tilhørende konfidensintervaller fundet ovenfor, og giv en intuitiv vudering af, hvorvidt der er forskel på de to grupper eller ej.*

I R
10^c(a, b)

På såvel de utransformerede som de logaritmetransformerede gennemsnit ses, at økologerne har et højere niveau af sædkoncentrationen end de SAS-ansatte. Der er nogen overlap mellem de tilhørende konfidensgrænser, men ikke ret meget for de logaritme-

Hvis differesen ved ANOVAen (korrigeret gor abstinenstid) er den samme som t-testen (som tog alle i en samlet gruppe) er ens, er der ingen confounding

transformeredes vedkommende. Vi vil derfor nok forvente, at der er en faktisk forskel, men det ser vi på nedenfor.

(c) *Foretag nu en sammenligning af de to grupper, og kvantificer forskellen i sædkoncentration for grupperne, igen med 95% konfidensinterval.*

Når vi skal sammenligne de to grupper uden hensyntagen til abstinenstiden, drejer det sig blot om et T-test. Antagelserne er bedst på log-skala, så det er den, vi benytter:

```
proc ttest data=oeko;
    class sas_ansat;
    var lkonc;
run;
```

```
The TTEST Procedure

Variable:  lkonc

sas_ansat       N        Mean    Std Dev    Std Err    Minimum    Maximum
ja            135      1.6486     0.4715     0.0406    -0.1249     2.6042
nej            53      1.8346     0.4198     0.0577     0.7404     2.5490
Diff (1-2)            -0.1859     0.4576     0.0742

sas_ansat     Method               Mean       95% CL Mean         Std Dev
ja                                1.6486      1.5684   1.7289      0.4715
nej                               1.8346      1.7189   1.9503      0.4198
Diff (1-2)    Pooled             -0.1859     -0.3323  -0.0396      0.4576
Diff (1-2)    Satterthwaite      -0.1859     -0.3257  -0.0461

Method           Variances        DF    t Value    Pr > |t|
Pooled           Equal           186      -2.51      0.0131
Satterthwaite    Unequal      106.17      -2.64      0.0096

             Equality of Variances

Method      Num DF    Den DF    F Value    Pr > F
Folded F       134        52       1.26    0.3414
```

Det ses, at SAS-ansatte har en signifikant lavere sædkoncentration end økologer (P=0.013). Forskellen på logaritmisk-skala er -0.186, svarende til at de SAS-ansattes sædkoncentration udgør i gennemsnit kun udgør $10^{-0.186} = 0.65$, dbvs. 65% af økologernes koncentrationsniveau.

95% sikkerhedsintervallet for denne forskel er angivet af SAS til $(-0.332, -0.040)$, og når dette tilbagetransformeres, fås:

$$(10^{-0.332}, 10^{-0.040}) = (47\%;\ 91\%)$$

(d) *Er der signifikant forskel på de to gruppers sædkoncentration?*

Ja, på et sædvanligt 5% signifikansniveau er der forskel, da P=0.013.

3. **Lav nu en underopdeling af de to grupper, baseret på længden af abstinenstiden og udregn passende summary statistics for de nu i alt 6 grupper.**

Vi ser igen på nogle *summary statistics*, men for at minimere outputtet, tager vi denne gang kun gennemsnit og median (og antallet, som man altid bør have med):

```
proc means N mean median data=oeko;
class sas_ansat abstid;
var konc lkonc;
run;
```

Dette giver en del output

sas_ansat	abstid	N Obs	Variable	N	Mean	Median
ja	1	25	konc	25	67.0420000	53.0000000
			lkonc	25	1.6045615	1.7242759
	2	47	konc	47	60.1229787	33.0000000
			lkonc	47	1.5679402	1.5185139
	3	63	konc	63	81.2949206	55.0000000
			lkonc	63	1.7263360	1.7403627
nej	1	12	konc	12	59.0083333	63.0000000
			lkonc	12	1.6136191	1.7973620
	2	16	konc	16	80.5000000	66.5000000
			lkonc	16	1.8345952	1.8228094
	3	25	konc	25	134.1800000	98.0000000
			lkonc	25	1.9406240	1.9912261

(a) *Ser det ud som om abstinenstiden har indflydelse på sædkoncentrationen?*
Og i givet fald, ser denne indflydelse så ens ud i grupperne?

Mændene med den lange abstinenstid ses at have en noget højere sædkoncentration end dem med kort eller mellem abstinenstid. Forskellen på de to korte abstinenstider er lidt mere uklar, men noget kunne tyde på, at effekten af abstinenstid er mere udtalt for økologerne end for de SAS ansatte (altså at der *kunne være* en interaktion).

(b) *Ser det ud som om fordelingen af abstinenstider er den samme i de to grupper?*

Dette spørgsmål vedrører slet ikke sædkoncentrationen, men udelukkende de to potentielle forklarende variable. Hvis de disse to forklarende variabel har relation til hinanden, altså hvis abstinenstiden i norgen grad afhænger af om man er SAS-ansat eller økolog, så kan abstinenstiden virke som en confounder for sammenligningen mellem de to grupper af mænd, således at vores estimat fra spørgsmål 2c bliver et misvisende udtryk for effekten af at leve økologisk.

Vi vil lave en simpel tabel, så som:

```
Table of sas_ansat by abstid

sas_ansat     abstid

Frequency|
Row Pct  |       1|       2|       3|  Total
---------+--------+--------+--------+
ja       |     25 |     47 |     63 |    135
         |  18.52 |  34.81 |  46.67 |
---------+--------+--------+--------+
nej      |     12 |     16 |     25 |     53
         |  22.64 |  30.19 |  47.17 |
---------+--------+--------+--------+
Total          37       63       88      188
```

Denne er lavet ved at skrive

```
proc freq data=oeko;
table sas_ansat*abstid / nocol nopercent;
run;
```

I denne tabel ses antallene af mænd i hver af de 6 grupper, samt rækkeprocenterne, dvs. fordelingen af abstinenstider for hver af de to grupper mønd (SAS-ansatte og økologer). Der synes ikke at være nogen særlig forskel på disse fordelinger (man kunne lave et χ^2-test for dette, det lærer I i næste uge).

4. **Benyt en variansanalysemodel til at besvare følgende:**

   (a) *Find et estimat for forskellen i sædkoncentration mellem de to populationer af mænd, for fastholdt værdi af abstinenstid.*

   Hvis abstinenstiden har en effekt på sædkoncentrationen (som det ser ud til, at den har) og hvis den også var relateret til SAS-ansat ja/nej (som det *ikke* ser ud til, at den er), så ville estimatet fra spørgsmål 2c som nævnt ikke være en rimelig sammenligning af de to grupper af mænd.

   I så fald ville vi hellere sammenligne SAS-ansatte med økologer, under forudsætning af samme abstinenstid, og det er præcis hvad en (additiv) tosidet variansanalysemodel gør.

   Den additive model kan skrives som :

$$Y_{sai} = \mu + \alpha_s + \beta_a + \varepsilon_{sai}$$

   hvor indices betyder **s:** SAS-ansat ja/nej, **a:** abstinenstid og **i:** individ.

   SAS-koden til dette ses nedenfor. Denne indeholder tillige dannelse af passende modelkontrol, ved hjælp af `ods`-systemet:

```
ods graphics on;
proc glm plots=DiagnosticsPanel data=oeko;
  class sas_ansat abstid;
model lkonc=sas_ansat abstid / solution clparm;
run;
ods graphics off;
```

Outputtet bliver nu (lettere beskåret):

```
The GLM Procedure

    Class Level Information

Class          Levels    Values
sas_ansat           2    ja nej
abstid              3    1 2 3

Number of observations    188
Dependent Variable: lkonc
                                  Sum of
Source                     DF       Squares   Mean Square  F Value  Pr > F
Model                       3    2.57196501    0.85732167     4.18  0.0068
Error                     184   37.69867193    0.20488409
Corrected Total           187   40.27063694

R-Square     Coeff Var      Root MSE    lkonc Mean
0.063867      26.60939      0.452641      1.701058

Source                     DF     Type I SS   Mean Square  F Value  Pr > F
sas_ansat                   1    1.31577345    1.31577345     6.42  0.0121
abstid                      2    1.25619157    0.62809578     3.07  0.0490

Source                     DF   Type III SS   Mean Square  F Value  Pr > F
sas_ansat                   1    1.32546710    1.32546710     6.47  0.0118
abstid                      2    1.25619157    0.62809578     3.07  0.0490

                                        Standard
Parameter             Estimate             Error    t Value    Pr > |t|
Intercept          1.921020681 B      0.07138491      26.91      <.0001
sas_ansat ja      -0.186905638 B      0.07348387      -2.54      0.0118
sas_ansat nej      0.000000000 B       .               .         .
abstid    1       -0.187233951 B      0.08873769      -2.11      0.0362
abstid    2       -0.145921066 B      0.07473457      -1.95      0.0524
abstid    3        0.000000000 B       .               .         .

Parameter           95% Confidence Limits
Intercept          1.780182489  2.061858873
sas_ansat ja      -0.331884951 -0.041926325
sas_ansat nej       .            .
abstid    1       -0.362308129 -0.012159772
abstid    2       -0.293367934  0.001525802
abstid    3         .            .
```

Begge kovariater ses at være signifikante, abstinenstiden dog kun lige akkurat (P=4.9%). For `abstid` ses de to laveste abstinens-

tider at ligge nogenlunde på samme niveau, mens sædkoncentrationen er højere for mænd med lang abstinenstid, ganske som vi konkluderede ud fra gennemsnittene ovenfor. Endvidere ses, at SAS-ansatte har en signifikant lavere sædkoncentration end økologer *med samme abstinenstid*. Forskellen på logaritmisk-skala er -0.187, svarende til at de SAS-ansattes sædkoncentration udgør i gennemsnit kun udgør $10^{-0.187} = 65\%$ af økologernes koncentrationsniveau.

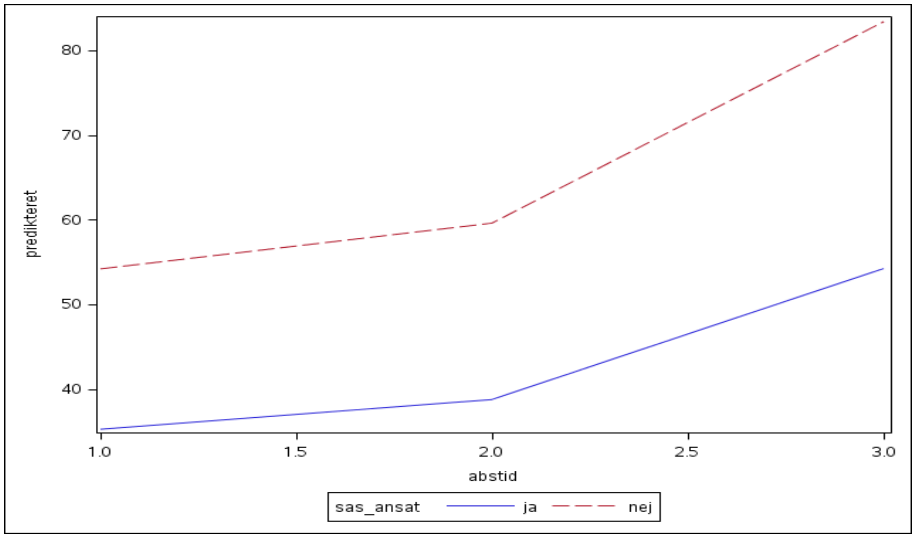
95% sikkerhedsintervallet for denne forskel er angivet af SAS til $(-0.332, -0.042)$, og når dette tilbagetransformeres, fås:

$$(10^{-0.332}, 10^{-0.042}) = (47\%;\ 91\%)$$

altså (med den valgte nøjagtighed) fuldstændig det samme som det ukorrigerede estimat fra spørgsmål 2c.

De tilhørende predikterede værdier for sædkoncentrationen ses i figuren nedenfor. De er dannet ved at tilføje en output-sætning, med efterfølgende fiksfakserier (sortering samt udvælgelse af netop 1 person fra hver af de 6 grupper), der skal sørge for at få tegningen til at se pæn ud....

```
proc glm plots=DiagnosticsPanel data=oeko;
  class sas_ansat abstid;
model lkonc=sas_ansat abstid / solution clparm;
output out=ny p=predikt;
run;

data ny;
set ny;
predikteret=10**predikt;
run;
proc sort data=ny;
by sas_ansat abstid;
run;

proc sgplot data=ny; where nr in (13,11,1,148,146,147);
    series y=predikteret x=abstid / group=sas_ansat;
run;
```



Bemærk, at de predikterede værdier er tilbagetransformeret til den oprindelige skala, og på denne skala er der *ikke* additivitet. Når effekterne er additive på logaritmisk skala, er de multiplkative på den oprindelige skala.

Modelkontrollen.

- *Varianshomogenitet?*

  Vi kan checke antagelsen om ens varians i alle 6 grupper ved at bruge Levenes test fra en ensidet variansanalyse (one-way ANOVA), der sammenligner alle disse 6 grupper under et. Hertil skal vi bruge den variabel, vi kaldte `gruppe` og som blev dannet i forbindelse med indlæsningen, og tidligere benyttet til en figur.

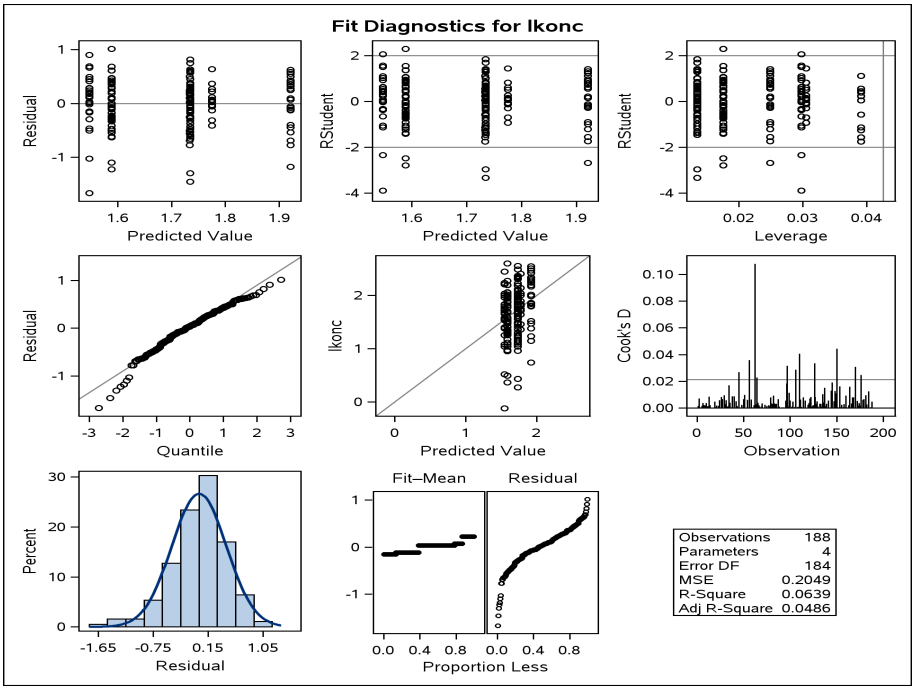
```
proc glm data=oeko;
  class gruppe;
model lkonc=gruppe;
means gruppe / hovtest=levene;
run;
```

og vi får så outputtet

```
The GLM Procedure

        Class Level Information
Class         Levels    Values
gruppe             6    1 2 3 11 12 13

Number of Observations Read         188
Number of Observations Used         188

Dependent Variable: lkonc
                                 Sum of
Source                DF        Squares    Mean Square   F Value   Pr > F
Model                  5     2.91775922     0.58355184      2.84   0.0169
Error                182    37.35287772     0.20523559
Corrected Total      187    40.27063694

R-Square     Coeff Var      Root MSE    lkonc Mean
0.072454      26.63221      0.453029      1.701058

Source                DF    Type III SS    Mean Square   F Value  Pr > F
gruppe                 5     2.91775922     0.58355184      2.84  0.0169

       Levene's Test for Homogeneity of lkonc Variance
        ANOVA of Squared Deviations from Group Means

                      Sum of        Mean
Source        DF     Squares      Square    F Value    Pr > F
gruppe         5      0.5646      0.1129       0.86    0.5067
Error        182     23.7958      0.1307

The GLM Procedure
Level of            ------------lkonc------------
gruppe        N             Mean          Std Dev
1            12       1.61361913       0.41933915
2            16       1.83459515       0.24822103
3            25       1.94062399       0.47598617
11           25       1.60456154       0.55156819
12           47       1.56794018       0.45641776
13           63       1.72633595       0.44309791
```

Antagelsen om ens varianser ser altså rimelig ud, idet Levenes test accepteres (P=51%)

Vi kunne også se på figuren i øverste venstre hjørne nedenfor. Det er et plot af residualer mod predikterede=forventede værdier, og det bør ikke udvise nogen form for struktur (bortset fra, at det jo ligner 6 søjler, da der kun er 6 forskellige predikterede værdier i denne model). Vi ser ingen tendens til trompetfacon eller anden form for struktur.



Normalfordelingsantagelsen?

Tegnes histogrammer eller residual-plots (se midti og nederst i venstre kolonne i figuren ovenfor) vil man opdage at logaritmetransformationen har bevirket en skævhed til 'den anden side', så normalfordelingsantagelsen er tvivlsom.

En bedre overensstemmelse kan opnås efter en kubikrodstranformation ($f(konc) = konc^{1/3}$). De overordnede konklusioner ændres dog ikke. Til gengæld kan parametrene i den nye model ikke direkte fortolkes (forskellene kan ikke kvantificeres på en enkel måde), så vi foretrækker at fortsætte på logaritmisk skala og glæde os over det rimeligt store datamateriale, der nedsætter behovet for en perfekt normalfordeling, så længe vi afholder os fra at lave normalområder.

(b) *Sammenhold ovenstående estimat med det tilsvarende fra spørgsmål 2a og kommenter.*

Denne sammenligning er allerede kommenteret ovenfor. Der er ikke nævneværdig confounding at spore.

(c) *Er der evidens for, at abstininstiden har en forskellig effekt på sædkoncentrationen i de to populationer?*

En model, der tillader effekten af abstinenstid at afhænge af SAS-ansat ja/nej, er en model med et interaktionsled (vekselvirkningsled):

$$Y_{sai} = \mu + \alpha_s + \beta_a + \gamma_{sa} + \varepsilon_{sai}$$

Koden bliver derfor nu udbygget til

```
proc glm data=oeko;
  class sas_ansat abstid;
model lkonc=sas_ansat abstid sas_ansat*abstid / solution;
run;
```

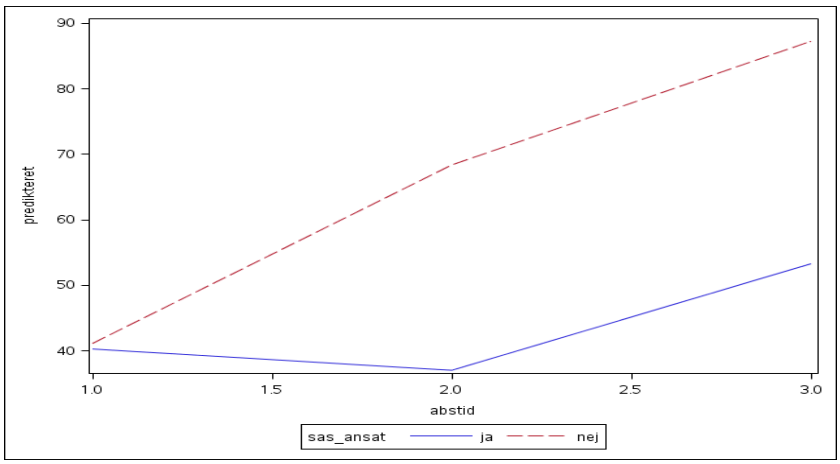
som resulterer i nedenstående output:

```
The GLM Procedure

    Class Level Information
 Class          Levels    Values
sas_ansat           2    ja nej
abstid              3    1 2 3

Number of observations    188

Dependent Variable: lkonc
                                    Sum of
Source                     DF       Squares   Mean Square  F Value  Pr > F

Model                       5    2.91775922    0.58355184     2.84  0.0169
Error                     182   37.35287772    0.20523559
Corrected Total           187   40.27063694

R-Square     Coeff Var      Root MSE    lkonc Mean
0.072454      26.63221      0.453029      1.701058

Source                     DF     Type I SS   Mean Square  F Value  Pr > F

sas_ansat                   1    1.31577345    1.31577345     6.41  0.0122
abstid                      2    1.25619157    0.62809578     3.06  0.0493
sas_ansat*abstid            2    0.34579420    0.17289710     0.84  0.4323

Source                     DF   Type III SS   Mean Square  F Value  Pr > F

sas_ansat                   1    0.91298917    0.91298917     4.45  0.0363
abstid                      2    1.25068088    0.62534044     3.05  0.0499
sas_ansat*abstid            2    0.34579420    0.17289710     0.84  0.4323
```

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		1.940623988	B	0.09060587	21.42	<.0001
sas_ansat	ja	-0.214288035	B	0.10708469	-2.00	0.0469
sas_ansat	nej	0.000000000	B	.	.	.
abstid	1	-0.327004862	B	0.15909868	-2.06	0.0413
abstid	2	-0.106028838	B	0.14504016	-0.73	0.4657
abstid	3	0.000000000	B	.	.	.
sas_ansat*abstid	ja 1	0.205230452	B	0.19177988	1.07	0.2860
sas_ansat*abstid	ja 2	-0.052366937	B	0.16929581	-0.31	0.7574
sas_ansat*abstid	ja 3	0.000000000	B	.	.	.
sas_ansat*abstid	nej 1	0.000000000	B	.	.	.
sas_ansat*abstid	nej 2	0.000000000	B	.	.	.
sas_ansat*abstid	nej 3	0.000000000	B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

Hypotesen om ingen vekselvirkning kan accepteres med $P = 43.2\%$. De predikterede forløb ses i figuren nedenfor, og vi bemærker, at de ser meget anderledes ud end de tilsvarende for den additive model. Hvis vi bare skulle udtale os udfra figuren, ville vi således gætte på, at der *var* interaktion mellem SAS-ansat og abstinenstid, men figuren viser jo heller ikke usikkerhederne, og disse er altså så store, at vi ikke kan påstå at have evidens for en interaktion.



Til gengæld kan vi selvfølgelig heller ikke afvise, at der *kunne* være en interaktion, som vi bare ikke finder på grund af et for lille datamateriale.

Opgave 2: Space shuttle

Et studie involverer de 26 astronauter, der deltog på de første 8 rejser med space shuttle (Bungo et.al., 1985). På frivillig basis konsumerede 17 af disse astronauter store mængder af salt og væske inden landingen, i et forsøg på at imødegå 'space deconditioning' (salt=1). De 9 øvrige indtog intet specielt (salt=0). Tabellen nedenfor viser pulsen (slag pr. minut) før og efter flyvningen for hver af de 26 astronauter.

Countermeasure taken			Countermeasure not taken		
Pre	Post	Change	Pre	Post	Change
71	61	−10	61	61	0
65	59	−6	59	66	7
52	47	−5	52	61	9
68	65	−3	54	68	14
69	69	0	53	77	24
49	50	1	78	103	25
49	51	2	52	77	25
57	60	3	54	80	26
51	57	6	52	79	27
55	64	9			
58	67	9			
57	69	12			
59	72	13			
53	69	16			
53	72	19			
53	75	22			
48	77	29			
Mean 56.88	63.76	6.88	57.22	74.67	17.44
SD 7.30	8.86	10.70	8.44	13.01	10.11

Filen `"space.txt"` fra hjemmesiden ser således ud:

```
salt pre post
1 71 61
1 65 59
1 52 47
.......
.......
.......
0 52 77
0 54 80
0 52 79
```

Data indlæses derfor i 3 kolonner, som f.eks. kaldes `salt`, `pre` og `post`, ligesom det står i overskriften. Der er således i alt tale om 26 observationer, idet de to grupper lægges 'i forlængelse af hinanden' (kun oplysninger fra 1 person på hver linie!).

Indlæsningen (til det midlertidige `WORK`-datasæt `space`), definition af to nye variable, `dif` og `snit`, samt print af datamaterialet, kunne se ud som nedenfor, hvis data forinden var anbragt i filen `space.txt` i folderen `C:\Basal`:

```
data space;
infile 'C:\Basal\space.txt' firstobs=2;
input salt pre post;

dif=post-pre;
snit=(pre+post)/2;
run;

proc print data=space;
run;
```

Man kunne selvfølgelig også indlæse filen direkte fra hjemmesiden, ligesom vi gjorde det i forrige opgave.

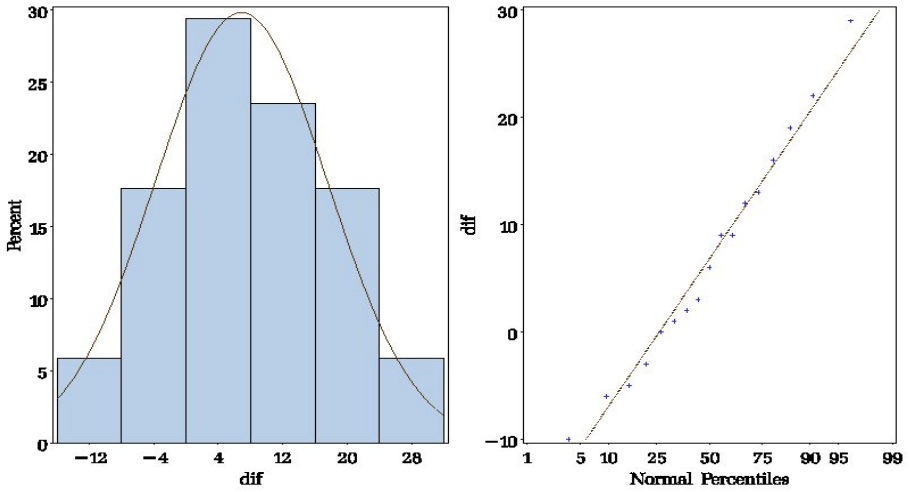
1. *Sammenlign før- og efter-målingerne i 'salt'-gruppen, og husk samtidig at lave passende illustrationer.*

   Vi skal sammenligne før- og efter-målingerne i salt-gruppen. Selv om vi således i første omgang kun bliver bedt om at se på salt-gruppen, er det ligeså let at foretage sammenligne for begge grupper på en gang, ved at benytte `by salt;`, når den relevante analyse foretages. Man skal dog bare huske at sortere først. I nedenstående plots er dog vist "filtreringsversionen", hvor plottet kun udføres for salt-gruppen, idet vi skriver `where salt=1;`.

   Hvis vi skal foretage et parametrisk test (og det foretrækker vi, da det giver et konfidensinterval), bliver der tale om et parret t-test. Forudsætningen for dette er rimelig normalitet for differenserne `dif=post-pre`, som er udregnet ovenfor.

Et histogram og et fraktildiagram kan fås ved at skrive:

```
proc univariate data=space; where salt=1;
var dif;
histogram / height=3 normal(mu=EST sigma=EST);
probplot / height=3 normal(mu=EST sigma=EST l=33);
run;
```

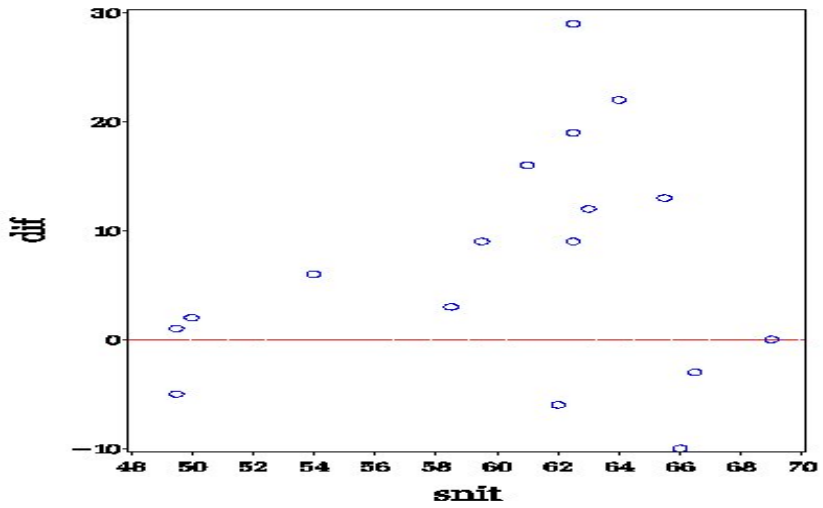


Vi ser her ingen særlige tegn på afvigelse fra normalfordelingen (men det betyder ikke, at vi har stor tiltro til denne antagelse, da der er tale om et ganske lille materiale).

En anden vigtig forudsætning er, at differenserne er 'lige store' over hele skalaen, altså at der ikke er nogen sammenhæng mellem differenser og niveau. Dette undersøges ved et Bland-Altman plot, som simpelthen er et scatterplot af differenser mod gennemsnit, her udført med proceduren `gplot` (den nyere procedure `sgplot` kan gøre noget tilsvarende, men vi har endnu ikke så stor erfaring med detaljerne endnu, f.eks. som her at indlægge en rød vandret linie i 0 med stiplet linie `vref=0 lv=33 cv=red`):

```
proc gplot data=space; where salt=1;
  plot dif*snit
  / vref=0 lv=33 cv=red haxis=axis1 vaxis=axis2 frame;
```

```
axis1 value=(H=2) minor=NONE label=(H=3);
axis2 value=(H=2) minor=NONE label=(A=90 R=0 H=3);
symbol v=circle i=none c=BLUE h=2 l=1 w=2;
run;
```



Da dette heller ikke viser udprægede tegn på sammenhæng (eller gør det??), vil vi fortsætte med et parret t-test. Vi udfører t-testet for begge grupper på en gang ved at skrive

```
proc sort data=space;  by salt;
run;

proc ttest data=space; by salt;
    paired pre*post;
    run;
```

eller

```
proc sort data=space;  by salt;
run;

proc ttest data=space; by salt;
    var dif;
    run;
```

Vi finder resultatet (her er kun den nederste del vist, svarende til salt-gruppen)

```
salt=1

The TTEST Procedure
Difference:  pre - post

  N        Mean     Std Dev     Std Err     Minimum     Maximum
 17     -6.8824     10.6998      2.5951    -29.0000     10.0000

    Mean       95% CL Mean        Std Dev      95% CL Std Dev
 -6.8824    -12.3837  -1.3810     10.6998      7.9689  16.2843

    DF    t Value    Pr > |t|
    16      -2.65      0.0174
```

P-værdien for test af identiske middelværdier for puls før og efter flyvningen ses at være 0.017, hvilket er signifikant på et 5% niveau og altså viser, at der formentlig sker en pulsstigning.

Hvis vi føler os usikre på normalfordelingsantagelsen, kunne vi i stedet udføre et non-parametrisk test (Wilcoxon signed-rank test), se kode og output nedenfor. Herved finder vi en P-værdi på 0.024, som understøtter konklusionen fra t-testet. Vi kunne også lave et test for normalfordelingen, men det giver ikke rigtig nogen mening på sådan et lille datamateriale.

Koden til den nonparametriske analyse er

```
proc univariate data=space; by salt;
var dif;
run;
```

og output er (igen kun for salt-gruppen):

```
salt=1

The UNIVARIATE Procedure
           Tests for Location: Mu0=0
```

```
Test              -Statistic-     -----p Value------

Student's t     t   2.65208     Pr > |t|     0.0174
Sign            M         4     Pr >= |M|    0.0768
Signed Rank     S        43     Pr >= |S|    0.0241
```

2. *Huskede du at give en kvantificering af effekten af flyvning på pulsen i 'salt'-gruppen? Med konfidensinterval!*

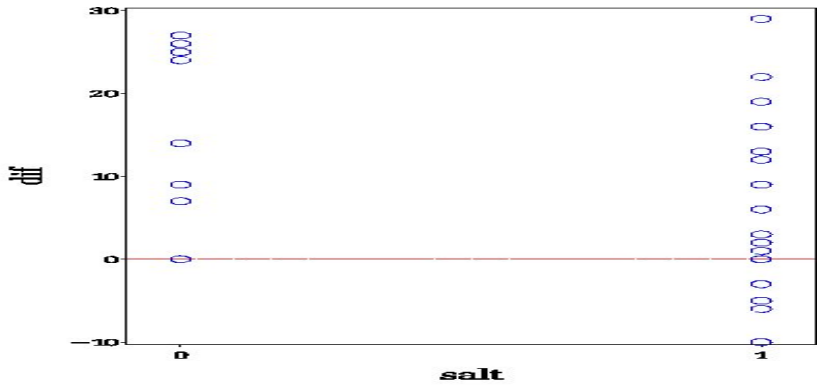
   Udregning af et konfidensinterval for middelværdien af pulsstigningen fås automatisk ved at udføre t-testet. Det ses under betegnelsen `95% CL Mean`, dvs. (-12.3837,-1.3810).

   Udfra ovenstående resultater kvantificeres stigningen i puls altså til 6.88 (med en standard error på 2.60), dvs. med 95% konfidensintervallet (1.38,12.38), altså ganske bredt. Testet gav signifikans på et 5% niveau, svarende til, at 0 ikke er inkluderet i konfidensintervallet. Vi er altså noget usikre på, hvor stor denne pulsstigning er, men den er næppe af afgørende betydning.

3. *Sammenlign effekten af flyvning i de to grupper. Hvilken konklusion opnås for effekten af saltindtagelse? Husk konfidensinterval!*

   Vi skal nu se på en sammenligning af differenserne i de to grupper.

   I kontrolgruppen har vi kun 9 personer, hvilket simpelthen er for lidt selv til grafiske illustrationer af fordelingen. Vi tillader os derfor (i hvert fald til en start) at gå ud fra, at differenserne `post-pre` er 'ligeså' normalfordelte i denne gruppe som i 'salt'-gruppen. En illustration af differenserne i de to grupper gøres bedst ved et scatterplot, da der er så få observationer:



På trods af den ikke så pæne fordeling i kontrolgruppen, fortsætter vi alligevel med at basere en sammenligning af de to gruppers differenser på et **uparret** t-test:

```
proc ttest data=space;
class salt;
var dif pre;
run;
```

Så får vi

```
The TTEST Procedure

Variable:  dif

salt           N        Mean     Std Dev     Std Err     Minimum     Maximum
0              9     17.4444     10.1132      3.3711           0     27.0000
1             17      6.8824     10.6998      2.5951    -10.0000     29.0000
Diff (1-2)            10.5621     10.5079      4.3317

salt          Method                Mean       95% CL Mean        Std Dev
0                                 17.4444       9.6707  25.2182     10.1132
1                                  6.8824       1.3810  12.3837     10.6998
Diff (1-2)    Pooled              10.5621       1.6219  19.5023     10.5079
Diff (1-2)    Satterthwaite       10.5621       1.5967  19.5275

Method           Variances        DF    t Value    Pr > |t|
Pooled           Equal            24       2.44      0.0225
Satterthwaite    Unequal       17.26       2.48      0.0236

              Equality of Variances

Method      Num DF    Den DF    F Value    Pr > F
Folded F        16         8       1.12    0.9123
```

Vi ser, at P-værdien for sammenligning af middelværdierne for differenserne i de to grupper er 0.0225, svarende til, at de er signifikant forskellige på et 5% niveau. Det betyder, at de astronauter, der ikke traf nogen foranstaltninger havde en mere udtalt pulsøgning end de, der gjorde noget. Denne øgede stigning er estimeret til 10.56, med et 95% konfidensinterval på (1.62,19.50). Ikke særligt overbevisende, men alligevel ...

Det tilsvarende non-parametriske test fås (for differenser og før-målinger på en gang, se forklaring under spørgsmål 5), ved at skrive

```
proc npar1way wilcoxon data=space;
class salt;
var pre dif;
exact hl;
run;
```

og for differenserne finder vi outputtet

```
The NPAR1WAY Procedure
            Wilcoxon Scores (Rank Sums) for Variable dif
                    Classified by Variable salt

                    Sum of      Expected       Std Dev          Mean
salt       N        Scores      Under H0      Under H0         Score
-----------------------------------------------------------------------
0          9        161.50        121.50     18.534936     17.944444
1         17        189.50        229.50     18.534936     11.147059
                 Average scores were used for ties.

   Wilcoxon Two-Sample Test

Statistic             161.5000

Normal Approximation
Z                       2.1311
One-Sided Pr >  Z       0.0165
Two-Sided Pr > |Z|      0.0331

t Approximation
One-Sided Pr >  Z       0.0215
Two-Sided Pr > |Z|      0.0431

Z includes a continuity correction of 0.5.

     Kruskal-Wallis Test

Chi-Square              4.6573
DF                           1
Pr > Chi-Square         0.0309



                        Hodges-Lehmann Estimation

                        Location Shift    12.0000

                                                    Interval        Asymptotic
Type                 95% Confidence Limits          Midpoint    Standard Error
Asymptotic (Moses)      1.0000     22.0000           11.5000            5.3572
Exact                   2.0000     21.0000           11.5000
```

Vi finder altså også en signifikans i det non-parametriske test (P-værdien er ca. 3-4%), og et konfidensinterval på (2,21), altså ikke langt fra det tilsvarende parametriske.

4. To astronauter deltog i to forskellige flyvninger og optræder altså i datamaterialet to gange. Spiller det nogen rolle?

   Vi ved ikke hvilke par af observationer, der stammer fra samme astronauter, så helt konkret kan vi ikke stille noget op med vores viden. Men hvis vi havde kunnet identificere dem, ville det nok være klogest kun at benytte første flyvetur for disse. Hvis pulsøgningen er meget personspecifik skaber det nemlig problemer for antagelsen om uafhængighed mellem observationerne, at der er flere målinger for hver person.

   Herudover kunne man tænke sig

   - at det er nogle selekterede personer, der tager afsted flere gange
   - at personer, der allerede har været afsted en gang, er blevet varigt ændret, så de anden gang adskiller sig fra de øvrige

   Den konkrete betydning for analyseresultaterne er svær at sige ret meget om. Det afhænger f.eks. af om personerne er med i samme gruppe begge gange:

   - Hvis de er med i samme gruppe, bliver variationen indenfor grupper for lille, og dermed kan man lettere finde en (måske ikke-eksisterende) forskel på de to grupper (type 1 fejl).
   - Hvis de er med i hver sin gruppe, bliver grupperne for ens, og vi får dermed sværere ved at se en evt. forskel (type 2 fejl).

5. *Kommenter frivilligheden i opdelingen i de to grupper og hvordan dette kunne tænkes at påvirke fortolkningen af resultaterne.*

   Frivilligheden i gruppeopdelingen kan tænkes at skabe problemer, som kan gå begge veje

   - Måske er det de overforsigtige/velovervejede, der tager deres forholdsregler, og hvis disse samtidig er i fysisk bedst form, kan de tænkes i forvejen at ville opleve en mindre pulsstigning
   - eller måske er det dem med en kendt risiko for pulsstigning, der vælger at tage forholdsregler, og så er det sandsynligt, at forskellen på de to grupper bliver mindre udtalt.

   For at få en valid sammenligning, burde grupperne have været randomiseret.

En lille indsigt i en evt. skævvridning kan fås ved at sammenligne `pre`-værdierne i de 2 grupper. Bemærk, at et t-test nu vil kræve normalitet af disse `pre`-målinger i hver gruppe og ikke som tidligere kun af differenserne. Vi finder

Mann-Whitney (Kruskal-Wallis) test: P=0.94
T-test, med ens varianser: P=0.92
T-test, med forskellige varianser: P=0.92

Der er altså ikke her nogen indikation af selektion.