Phd-kursus i Basal Statistik, Opgaver til 1. uge

Opgave 1: Wright

For 17 patienter er der målt **peak expiratory flow rate** på to forskellige måder, dels ved at anvende det **traditionelle** *Wright peak flow meter*, og dels med det **nye** såkaldte *mini Wright flow meter* (Bland and Altman, 1986). Med begge apparater er der foretaget dobbeltbestemmelser, således at der i alt foreligger 4 observationer for hver person.

1. **Kør indlæsningsprogrammet, og overvej, hvad hver enkelt linie refererer til, så I næste gang kan udnytte en modificeret udgave af denne kode.**

   Vi benytter den i opgaveteksten angivne indlæsningskode, idet vi dog erstatter det intetsigende datasæt-navn `a1` med det mere sigende `wright`:

```
data wright;
infile "http://staff.pubhealth.ku.dk/~lts/basal/data/wright.txt"
        URL firstobs=2;
input wright1 wright2 mini1 mini2;
run;
```

   Kommentarer til de enkelte linier:

   Linie1: angiver navnet på det datasæt, man vil danne

   Linie 2-3: angiver navnet på den fil, man vil importere, her en fil fra nettet, og derfor skrives efterfølgende `URL`.
   Option `firstobs=2` angiver, at man først skal starte indlæsningen fra linie 2, idet der står variablenavne i første linie.

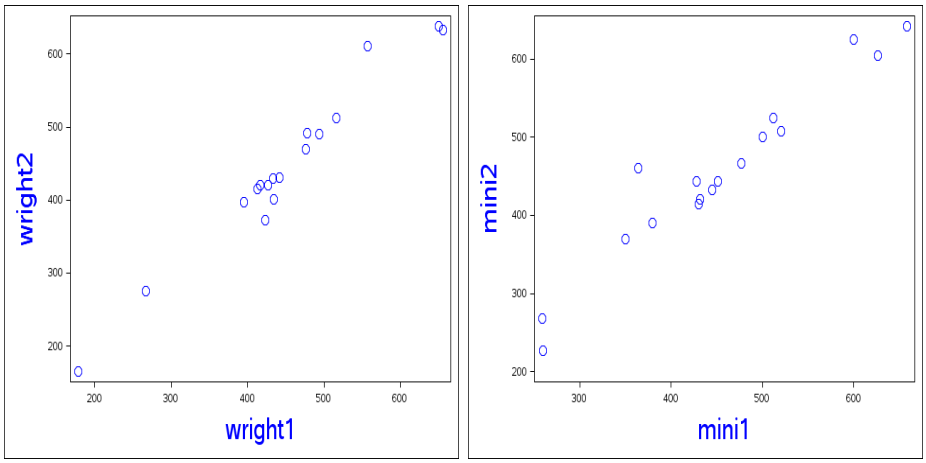
   Linie 4: Angiver de variablenavne, man ønsker at anvende for de enkelte kolonner i datafilen.

   Vores datasæt ligger nu i `WORK`-folderen under navnet `wright` (er altså *ikke* permanent - ellers skulle man have skrevet `sasuser.wright` i stedet for blot `wright`), og det består af 17 observationer og fire variable, nemlig `wright1`, `wright2`, `mini1` og `mini2`.

Til en start kan vi se på et plot af dobbeltbestemmelser mod hinanden, for hver af de to målemetoder:

```
proc sgplot data=wright;
    scatter x=wright1 y=wright2;
run;

proc sgplot data=wright;
    scatter x=mini1 y=mini2;
run;
```

Figurerne nedenfor er dog lavet ved hjælp af en kode, der er lidt udbygget for at forbedre grafernes udseende, se løsningsprogrammet.



Det ses, at observationerne fordeler sig rimeligt omkring en linie, og hvis man kigger nærmere efter, er denne linie næsten identitetslinien, og dermed vil vi umiddelbart sige, at dobbeltbestemmelserne stemmer rimeligt godt overens.

De efterfølgende spørgsmål skal lede igennem forskellige betragtninger vedrørende vurdering af hver af målemetoderne samt sammenligning af de to målemetoder. Det endelige formål er at kvantificere overensstemmelsen mellem de to målemetoder (hhv. Wright og Mini Wright).

2. *Vurder (for hver af de to målemetoder for sig) om differensen mellem dobbeltmålinger afhænger af niveauet af lungefunktionen. En god metode til dette er det såkaldte Bland-Altman plot (scatter plot af differenser*

mod gennemsnit).

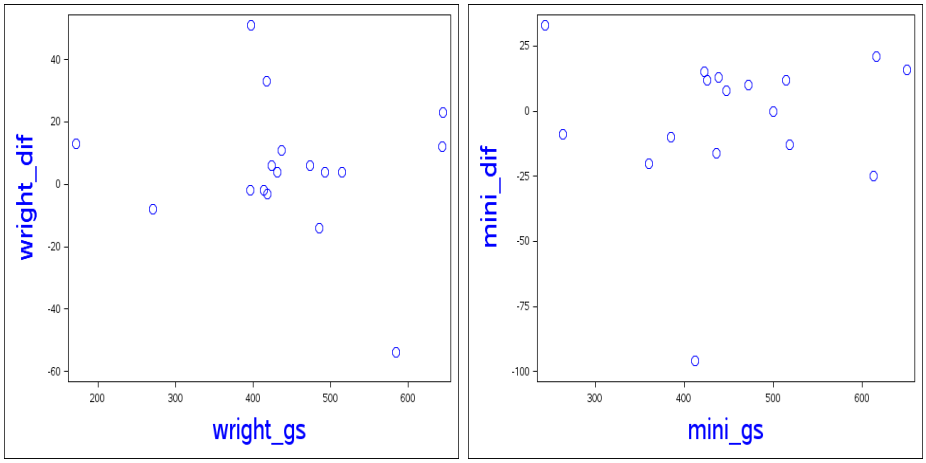
Vi har nu brug for at ændre i datasættet `wright`, fordi vi skal danne nogle nye variable. Nedenfor udregner vi differenser mellem dobbeltbestemmelser for hver af metoderne, gennemsnit af selvsamme dobbeltbestemmelser, samt (til brug i spm. 6) differenser mellem gennemsnit af de to metoder (`dif`) samt gennemsnittet af disse gennemsnit (`gnsnit`, som altså blot er gennemsnittet af alle fire målinger). Vi kalder det nye datasæt `wright1` og gemmer det som permanent datasæt i folderen `sasuser`:

```
data sasuser.wright1;
set wright;

wright_dif=wright1-wright2;
wright_gs=(wright1+wright2)/2;
mini_dif=mini1-mini2;
mini_gs=(mini1+mini2)/2;

dif=wright_gs-mini_gs;
gnsnit=(wright_gs+mini_gs)/2;
run;
```

Vi laver herefter (for hver af målemetoderne for sig) et plot af differenserne mod gennemsnittet

```
proc sgplot data=sasuser.wright1;
    scatter x=wright_gs y=wright_dif;
run;

proc sgplot data=sasuser.wright1;
    scatter x=mini_gs y=mini_dif;
run;
```

hvorved vi får figurerne



Disse figurer går under betegnelsen 'Bland-Altman plots', efter artiklen Bland&Altman(1986). Vi ser af disse plots, at differenserne generelt ligger i et bånd omkring 0 af nogenlunde lige stor bredde hele vejen, omend det lille antal observationer ikke tillader alt for kategoriske konklusioner.

Der synes at være en enkelt perons, hvor der er meget stor forskel mellem de to observationer af `mini`. Vi vil komme tilbage til dette senere.

3. **Reproducerbarhed:** *Udregn og fortolk limits of agreement (normalområde for differenser), igen separat for hver af metoderne, uden at transformere. Vurder rimeligheden af de nødvendige antagelser.*

   Limits of agreement er normalområder for differenserne, så vi skal finde gennemsnit og spredning for disse, ved hjælp af `proc means`. Vi gør dette for alle tre differenser på en gang ved at skrive

```
proc means data=sasuser.wright1;
var wright_dif mini_dif dif;
run;
```

   hvorved vi får

```
The MEANS Procedure

Variable      N           Mean        Std Dev        Minimum        Maximum
-------------------------------------------------------------------------
```

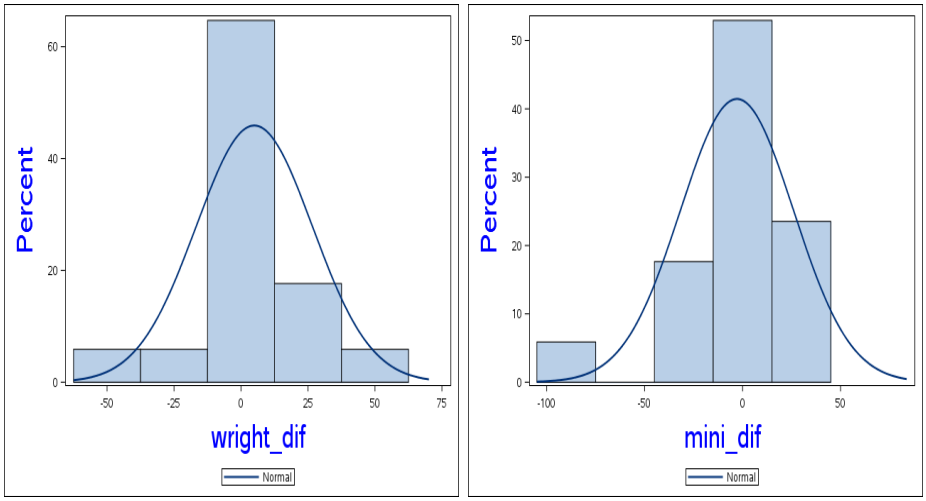
```
wright_dif   17      4.9411765     21.7240379    -54.0000000     51.0000000
mini_dif     17     -2.8823529     28.8723102    -96.0000000     33.0000000
dif          17     -6.0294118     33.2041369    -92.0000000     51.5000000
-------------------------------------------------------------------------------
```

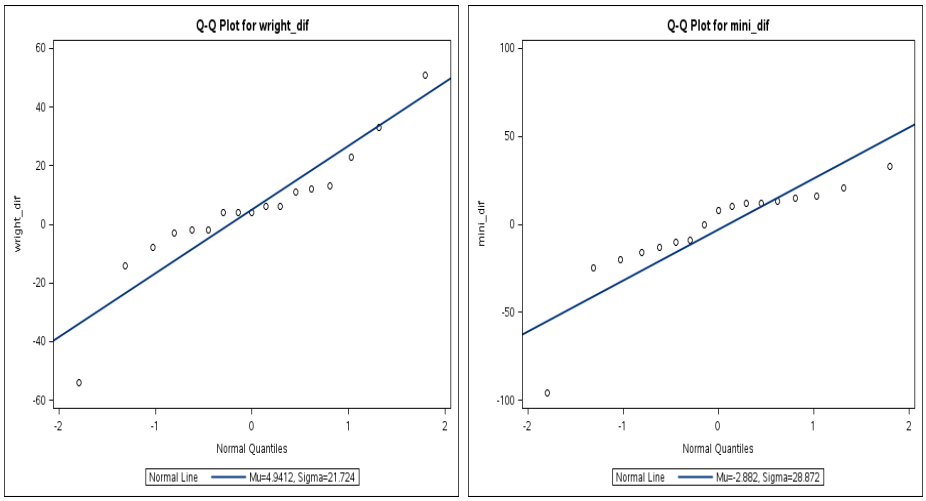
Vi går ud fra, at de 17 personer ikke er familiemæssigt relateret, og at de 17 differenser derfor er uafhængige. For at anvende ovenstående spredninger til at udregne normalområder, skal vi yderligere sikre os, at differenserne er rimeligt normalfordelte og nogenlunde af samme størrelsesorden uanset niveau. Det sidste var netop hvad vi vurderede i spørgsmålet ovenfor, så tilbage står antagelsen om normalitet. Nedenfor ses histogrammer for hhv. `wright_dif` og `mini_dif` og vi ser, at der er nogen afvigelse fra en normalfordeling. Usikkerheden i vurderingen er imidlertid stor med så få observationer.

Endvidere er vist et fraktildiagram, kun for `wright_dif`, da vi alligevel må erkende, at vi ikke med nogen rimelighed kan vurdere normalfordelingstilpasningen med så få observationer.

Figurerne kan dannes ved at skrive (kun vist for wright-apparaturet):

```
proc sgplot data=sasuser.wright1;
    histogram wright_dif;
    density wright_dif;
run;

proc univariate normal data=sasuser.wright1;
    var wright_dif;
    probplot wright_dif / normal(mu=est sigma=est);
run;
```





Tests for normalitet (som man normalt ikke får meget relevant information fra) giver faktisk her en afvisning for Mini Wright, på trods af det sparsomme materiale:

For Wright:

```
The UNIVARIATE Procedure
Variable:  wright_dif

                  Tests for Normality

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

Shapiro-Wilk          W     0.899039    Pr < W      0.0655
```

```
Kolmogorov-Smirnov    D     0.180881    Pr > D      0.1429
Cramer-von Mises      W-Sq  0.135555    Pr > W-Sq   0.0346
Anderson-Darling      A-Sq  0.780059    Pr > A-Sq   0.0358
```

og for Mini Wright:

```
The UNIVARIATE Procedure
Variable:  mini_dif

                  Tests for Normality

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

Shapiro-Wilk          W     0.791298    Pr < W      0.0015
Kolmogorov-Smirnov    D     0.176293    Pr > D     >0.1500
Cramer-von Mises      W-Sq  0.158384    Pr > W-Sq   0.0175
Anderson-Darling      A-Sq  1.082043    Pr > A-Sq   0.0058
```

Denne afvisning - ligesom i øvrigt det lave antal observationer - gør, at vi skal tage de nedenfor udregnede grænser med et stort forbehold. Vi finder limits of agreement til

Wright: $-4.94 \pm 2 \times 21.72 = (-48.38, 38.50)$
Mini Wright: $2.88 \pm 2 \times 28.87 = (-54.86, 60.62)$

Betydningen af limits of agreement er, at differenserne mellem dobbeltbestemmelser med 95% sandsynlighed vil ligge indenfor disse grænser, dvs. de udtrykker troværdigheden af en enkelt måling med hver af apparaterne.

Da datamaterialet er så lille, kunne vi også have valgt at bruge en passende t-fraktil til at udregne disse normalområder, det ville i så fald være med 16 frihedsgrader, altså 2.12.

Teknisk note:
Man kunne ligeledes overveje, om man skulle kræve, at differenserne havde middelværdi 0 og dermed estimere spredningen ved $\frac{1}{17}\sum_{p=1}^{17} \mathrm{dif}_p^2$ i stedet for $\frac{1}{16}\sum_{p=1}^{17}(\mathrm{dif}_p - \bar{\mathrm{dif}})^2$

Herved ville vi få symmetriske normalområder (limits of agreement):

Wright: $0 \pm 2 \times 21.65 = (-43.30, 43.30)$
Mini Wright: $0 \pm 2 \times 28.16 = (-56.32, 56.32)$

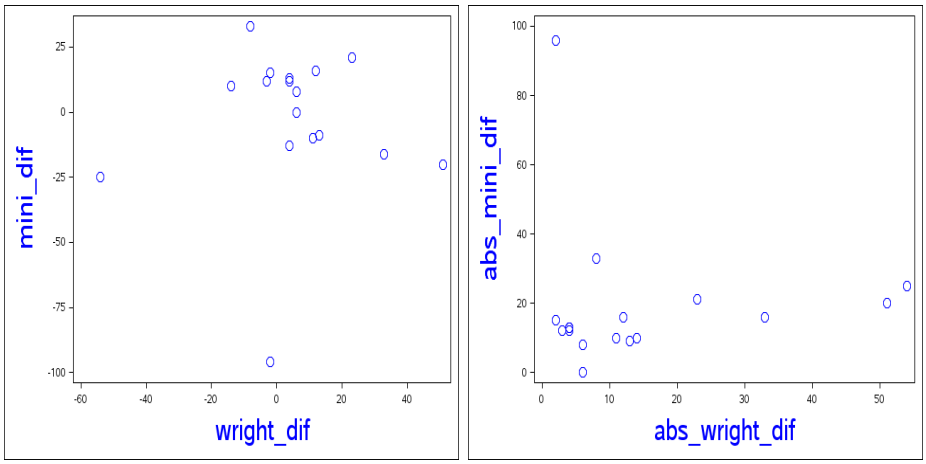
4. *Hvilken af metoderne har den bedste reproducerbarhed?*

   Baseret på de udregnede limits of agreement ovenfor, ser det ud som om Wright metoden har en noget bedre reproducerbarhed end Mini Wright, idet dens limits of agreement er smallest.

   Man kunne godt lave et egentligt test for dette, men det fører alt for vidt her...

5. *Tegn et scatter plot af de to sæt differenser (differenser mellem dobbeltbestemmelser for hver af de to metoder), og vurder på baggrund af dette, om der er nogen personer, der ser ud til at være mere ustabile at måle på end andre.*

   Den venstre af figurerne nedenfor viser de to sæt differenser (med fortegn) plottet mod hinanden, medens den højre figur plotter de tilsvarende numeriske (absolutte) differenser, dannet ved f.eks.
   `abs_wright_dif=abs(wright_dif);`
   Hvis fortegnet på differensen skønnes at være vigtigt (hvis der f.eks. ses en generel stigning fra første til anden måling) bør venstre figur benyttes, ellers er højre lettere at se på.



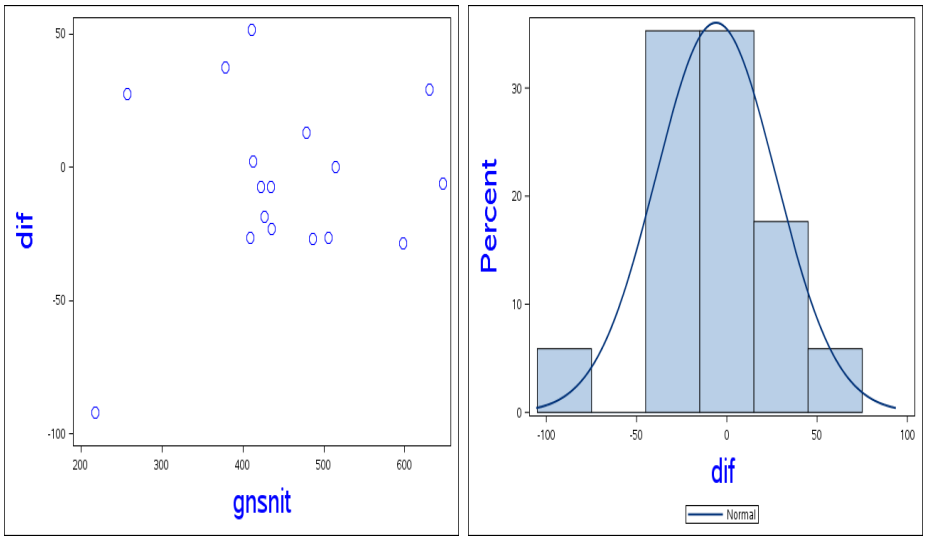
   Vi skal vurdere om der er enkelte personer, der har store differenser mellem dobbeltbestemmelserne for *begge* målemetoder, og dette ses ikke umiddelbart at være tilfældet. Vi har (som tidligere bemærket) en enkelt person med en stor diskrepans mellem de to målinger for Mini Wright, men denne person har pænt overensstemmende målinger for

Wright apparaturet. Sådanne personer, der er 'svære at måle på' ses i andre sammenhænge, såsom vurdering af leverstørrelse, hvor overvægtige personer er sværere at vurdere.

6. **Overensstemmelse:** *Sammenlign nu gennemsnittene af dobbeltbestemmelserne for de to metoder, dvs. tegn igen Bland-Altman plot og udregn limits of agreement, denne gang for sammenligning af de to målemetoder. Kommenter den kliniske anvendelighed af disse grænser.*

   Vi arbejder nu videre med de to gennemsnit, ovenfor kaldet `wright_gs` hhv. `mini_gs`. Igen skal vi se på et plot af differenser (`dif`) mod gennemsnit (`gnsnit`) samt vurdere rimeligheden af normalfordelingsantagelsen, inden vi går over til at udregne normalområder for differenserne.

   De relevante tegninger er



   og ved hjælp af `proc means` finder vi de størrelser, vi skal bruge til at udregne normalområder

```
The MEANS Procedure

Variable     N            Mean         Std Dev         Minimum         Maximum
-------------------------------------------------------------------------------
```

```
dif          17      -6.0294118      33.2041369     -92.0000000      51.5000000
-------------------------------------------------------------------------------
```

Selv om det (igen) er lidt vovet på så få observationer, udregner vi nu limits of agreement til

Wright vs. Mini Wright: $-6.03 \pm 2 \times 33.20 = (-72.43, 60.37)$

Når vi anvender disse grænser i praksis, skal vi huske på, at de er udregnet på baggrund af gennemsnit af to dobbeltbestemmelser. Hvis dette ikke er sædvanlig klinisk praksis, dvs. hvis man i praksis kun foretager en enkelt måling, så vil disse grænser være *for snævre*!

7. *Er der systematisk forskel på de to målemetoder? Kvantificer!*

Vi interesserer os her for middelværdierne af de to målemetoder, nærmere betegnet om disse afviger signifikant fra hinanden. Igen er der tale om parrede observationer (gennemsnittene `wright_gs` hhv `mini_gs`), så vi ser enten på differenserne `dif` og tester om disse har middelværdi 0 eller foretager et parret t-test (`paired wright_gs*mini_gs;` i `proc ttest`). Forudsætningen for dette er rimelig normalitet for differenserne, hvilket vi allerede checkede ovenfor.

```
The TTEST Procedure
Variable:  dif

  N        Mean     Std Dev     Std Err     Minimum     Maximum
 17     -6.0294     33.2041      8.0532    -92.0000     51.5000

    Mean       95% CL Mean        Std Dev       95% CL Std Dev
 -6.0294    -23.1014  11.0426     33.2041     24.7294  50.5343

    DF    t Value    Pr > |t|
    16      -0.75      0.4649
```

Vi ser altså, at T-testet giver teststørrelsen t=-0.75, svarende til P=0.46, og altså ingen signifikant forskel på de to målemetoder. En tilsvarende konklusion opnås fra et nonparametrisk test.

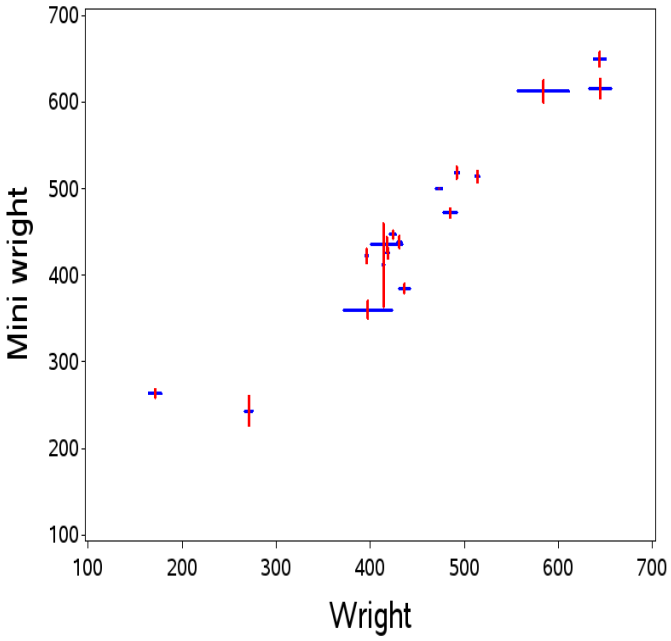
Hermed kan vi imidlertid ikke være sikre på, at der ingen forskel er, så vi kvantificerer den sandsynlige forskel ved et konfidensinterval for forskellen mellem middelværdier, som her aflæses fra outputtet til at være $(-23.10, 11.04)$

Vi kan altså ikke udelukke at forskellen på middelværdierne kan være op til ca. 10 'den ene vej' eller lidt over 20 'den anden vej'.

8. *Hvis en forskel på 75 l/min skønnes at have klinisk betydning, kan vi så erstatte Wright med det nye mini Wright?*

   Her skal vi vurdere om der hyppigt forekommer forskelle på 75 l/min, når man måler to gange på samme person med hvert af de to forskellige apparater. Ud fra limits of agreement ser vi, at 75 l/min ligger udenfor det, der 'normalt' forekommer, dvs. det, der forekommer i 95% af tilfældene. Det vil således være relativt sjældent, at vi blot ved et tilfælde ser klinisk betydelige afvigelser mellem de to målemetoder, igen **forudsat at vi til daglig virkelig benytter gennemsnit af dobbeltbestemmelser!**

Sluttelig skal vi se en figur, der forsøger at medtage alle observationer på en gang:



For hver person råder vi over 4 observationer, 2 med hver målemetode. Disse 4 er opsat som et kors, idet dobbeltbestemmelser foretaget med samme målemetode er forbundet med et liniestykke.

Kodningen af denne figur er ikke helt let (se løsningsprogrammet), men figuren er illustrativ, fordi den både viser overensstemmelsen mellem de to typer af måleappartur (ligger krydsene cirka på identitetslinien?) samt reproducerbarheden for hver af metoderne (hhv. længden af de blå og røde streger). Det, vi så i spørgsmål 3, var, at det traditionelle apparatur (`wright`) var en anelse bedre end det nye (`mini`), hvilket her svarer til, at de blå streger er en anelse kortere end de røde. Vi kan af tegningen se, at dette hovedsagelig skyldes en enkelt person, hvor der var stor uoverensstemmelse mellem de to `mini`-målinger.

Reference:

Bland, J.M. and Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet,* **i**, 307-310.

Opgave 2: Sundby

Vi betragter nu et lille uddrag af det såkaldte **Sundby95**-materiale, der er en stor undersøgelse af københavnernes sundhed.

Det totale datasæt ligger på hjemmesiden som præfabrikeret SAS-datasæt, men der ligger også et lille udpluk af informationerne i tekstfilen `Sundby_lille`, indeholdende variablene (i den nævnte rækkefølge)

kon: Personens køn (1: mand, 2:kvinde)

v75: Personens vægt, i *kg*

v76: Personens højde, i *cm*

v17: Fysisk aftivitet i fritid (kategorier 1-4, lave tal betyder mest aktiv)

v24af: Antal drukkede genstande sidste weekend

1. *Indlæs data ved at benytte en modifikation af det indlæsningsprogram, I brugte til den forrige opgave. Brug passende valgte navne til variablene.*

   Vi indlæser uddraget af Sundby-datasættet på måde, der er en anelse anderledes end i den forrige opgave (denne bid ligger også på hjemmesidens link til indlæsning):

```
FILENAME navn URL "http://staff.pubhealth.ku.dk/~lts/basal/data/sundby_lille.txt";

data sundby;
infile navn firstobs=2;
input kon v75 v76 v17 v24af;

if kon<0 then delete;
if kon=2 then gender='female';
if kon=1 then gender='male';
vaegt=v75;
hoejde=v76/100;
bmi=vaegt/hoejde**2;
run;
```

   Bemærk, at vi her samtidig har foretaget et par ændringer i det *temporære* (midlertidige) datasæt `sundby`, nemlig:

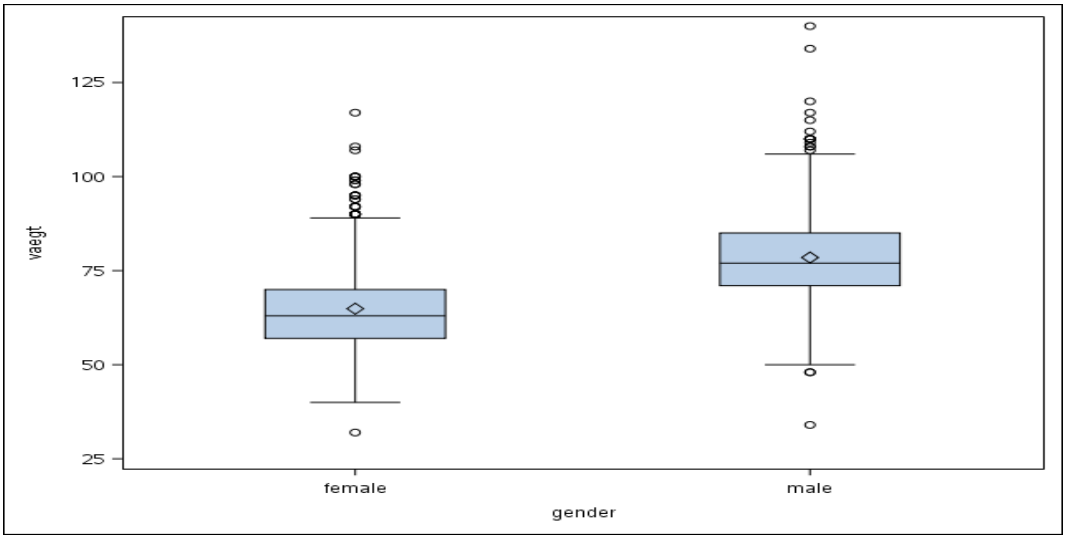
   - Sletning af observationer med ukendt køn

- Definition af en mere forståelig angivelse af kønnet
- Mere intuitive betegnelser for højde- og vægt-variable, samt omkodning af højde fra *cm* til *m*
- Definition af en ny variabel, `bmi`, se nedenfor under spm. 6.

2. *Lav en illustration af vægtfordelingen (*`v75`*) for mænd hhv. kvinder (brug f.eks. Box plots eller histogrammer), og beskriv også fordelingen i tal, dvs. gennemsnit, median, spredning mv.*

   Box plottene kan laves med koden

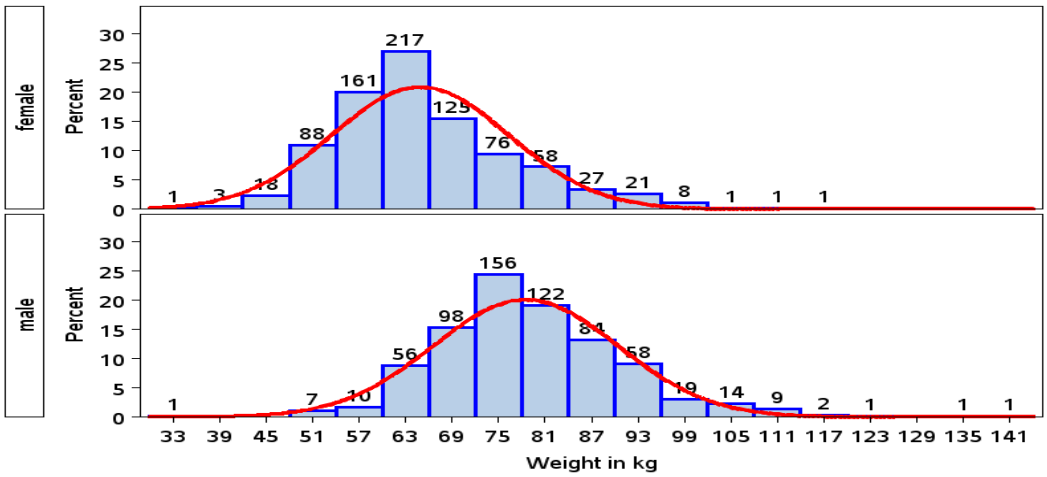
```
proc sgplot data=sundby;
    vbox vaegt / category=gender;
    run;
```

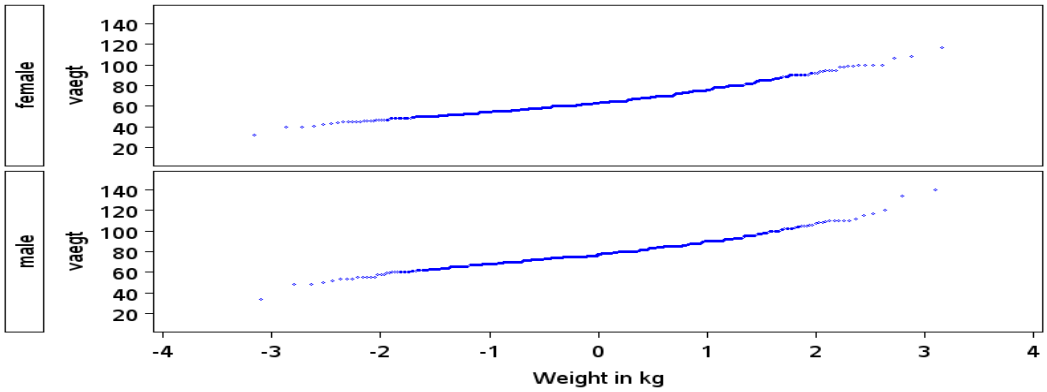


   De synes at vise en vis skævhed i fordelingerne, så lad os se nærmere på histogrammer for at vurdere tilpasningen til normalfordelingen (som skal bruges i næste spørgsmål).

   I forbindelse med histogrammerne udregner vi også 2.5% og 97.5% fraktilerne, idet vi benytter proceduren `univariate` og herunder anvender sætningen `histogram`, her med en masse lir:

```
proc univariate normal gout=plotud data=sundby;
  class gender;
  var vaegt;
  histogram vaegt / normal(color=red w=3)
                    height=3
                    haxis=axis1 barlabel=count;
   axis1 label=(a=90 r=0 'Weight in kg');
   output out=regn pctlpre=P_ pctlpts=2.5,97.5;
run;
```





Fordelingen i tal, opdelt efter køn:

```
proc means n mean median stddev min max data=sundby;
    class gender;
    var vaegt hoejde bmi;
run;
```

```
The MEANS Procedure

            N
gender     Obs  Variable     N          Mean        Median       Std Dev
-------------------------------------------------------------------------
female     827  vaegt      806    64.8959057    63.0000000    11.4403653
                hoejde     805     1.6742484     1.6800000     0.0694423
                bmi        788    23.1517030    22.4190897     3.9116406

male       647  vaegt      639    78.4882629    77.0000000    11.9084584
                hoejde     632     1.7990348     1.8000000     0.0779310
                bmi        628    24.2283560    23.8087970     3.2047620
-------------------------------------------------------------------------

            N
gender     Obs  Variable       Minimum         Maximum
-----------------------------------------------------
female     827  vaegt       32.0000000     117.0000000
                hoejde       1.4800000       1.9800000
                bmi         13.6699560      37.3702422

male       647  vaegt       34.0000000     140.0000000
                hoejde       1.5500000       2.0000000
                bmi         11.6275093      39.6353547
-----------------------------------------------------
```

Ikke overraskende kan vi konstatere, at mændene vejer mere end kvinderne. En del af årsagen hertil kunne jo være, at de også er højere (og det er de jo, hvilket også klart ses af såvel boxplot som summary statistics)

3. *Kommenter fundene fra forrige spørgsmålet med henblik på at konstruere normalområder for vægten, for hvert køn for sig. Bestem et sådant normalområde, både med og uden brug af normalfordelingsantagelsen. Hvordan passer de sammen?*

   Udfra gennemsnit og spredning udregner vi normalområder under antagelse af, at vi har at gøre med normalfordelinger for hvert køn. Her er udregningerne foretaget i R, som (bl.a.) fungerer som regnemaskine:

```
> 64.896+c(-2,2)*11.44
[1] 42.016 87.776
> 78.488+c(-2,2)*11.908
[1]  54.672 102.304
```

   Normalområderne for vægten er således:

   Kvinder: (42.0, 87.8) kg

   Mænd: (54.7, 102.3) kg

   Sammenligner vi med $2\frac{1}{2}\%$ og $97\frac{1}{2}\%$ fraktilerne, som blev udregnet og lagt i datasættet `regn` (nedenfor et `proc print data=regn;`):

Obs	gender	P_2_5	P_97_5
1	female	47	92
2	male	58	106

   ser vi, at de normalfordelingsbaserede intervaller skyder lidt for lavt, fordi de ikke tager hensyn til den lille skævhed, der er i fordelingerne.

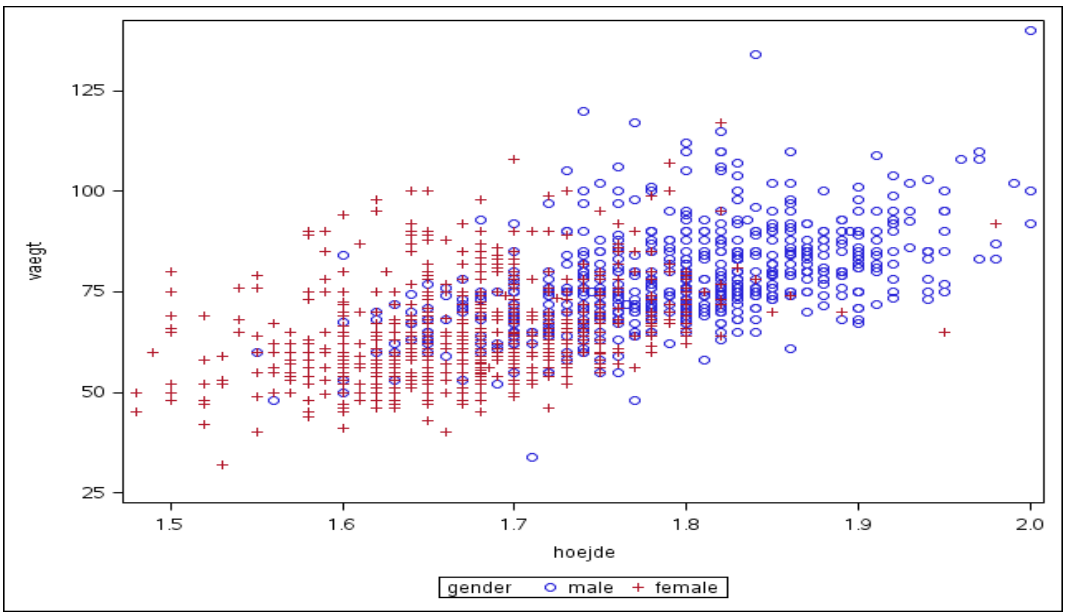
4. *Det ser klart ud til, at mændene vejer mere end kvinderne. Hvordan kunne man forklare det?*

   Det har vi allerede berørt ovenfor: Det kan skyldes, at mændene er højere.

5. *Lav et scatterplot af vægt overfor højde (*`v76`*), med forskellige symboler for mænd og kvinder. Ser det ud som om vægtforskellen kan forklares ved at mænd generelt er højere end kvinder?*

   Scatterplottet kræver farver for at man skal kunne se forskel på kønnene. Vi benytter koden:

```
proc sgplot data=sundby;
    scatter x=hoejde y=vaegt / group=gender;
    run;
```



   Umiddelbart ser det på plottet ud som om højden kan forklare en god del af forskellen på vægten på mænd og kvinder.

   Et velkendt højde-korrigeret mål for vægt er *body mass index (BMI)*, der defineres som

$$\text{BMI} = \frac{\text{vægt i kg}}{\text{højde i meter, kvadreret}}$$

6. *Definer den nye variabel* `bmi` *ved at indføje en sætning inden det første* `run;`*. Check om den er blevet rigtigt defineret, f.eks. ved at udregne gennemsnit mv.*

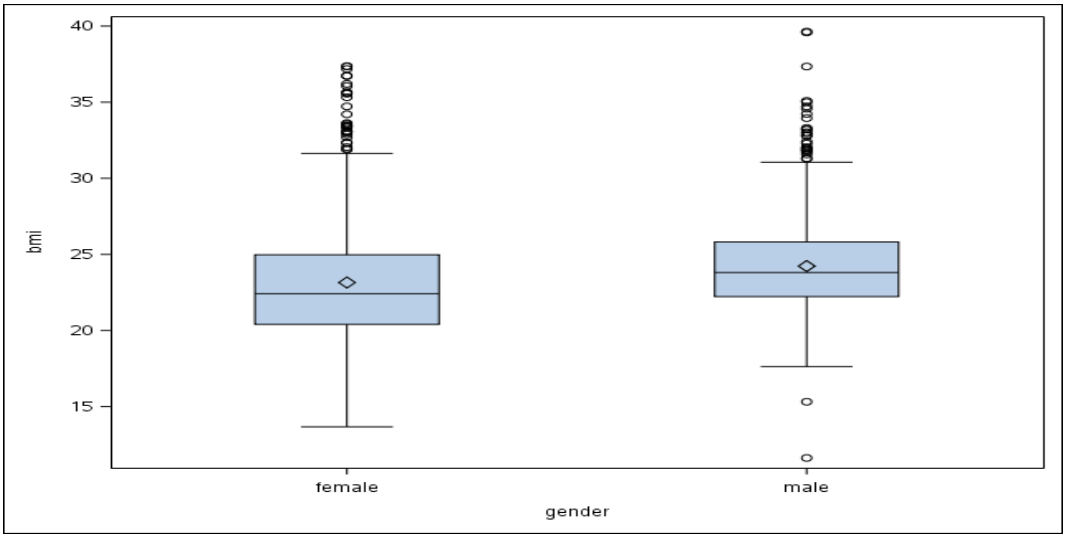
   For at definere denne variabel, skal man op i indlæsningsdelen af SAS-programmet, dvs. **inden det første** `run` og tilføje linien

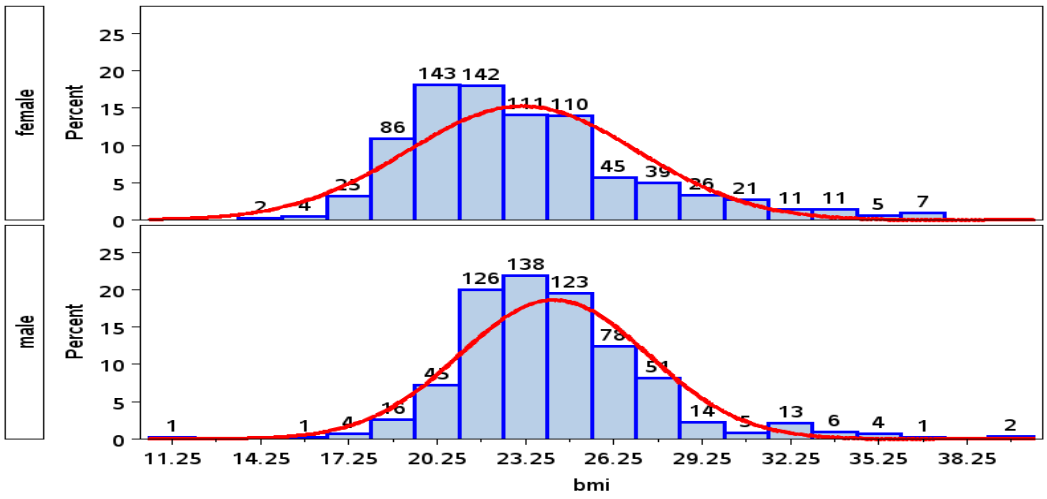
   ```
   bmi=vaegt/hoejde**2;
   ```

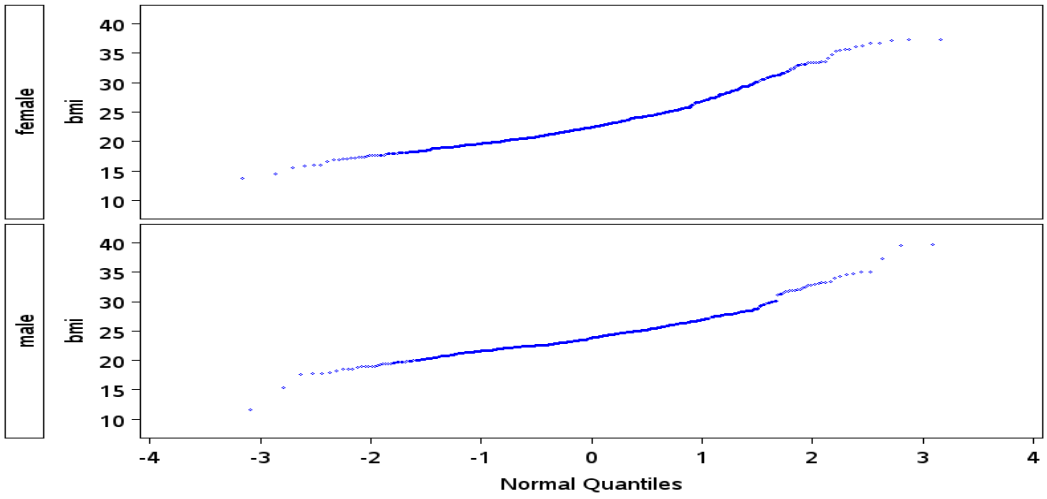
   hvilket vi faktisk allerede havde gjort helt fra starten.

7. *Bestem nu et normalområde for* `bmi` *ved hjælp af en normalfordelings-antagelse.*

   Vi tager lige et kig på figurerne svarende til vægten ovenfor, nu blot for body mass index, `bmi`:







Igen ser vi, at normalfordelingen næppe er helt rimelig, da der ses en skævhed i fordelingen. Vi fortsætter alligevel, men sørger for at huske, at vi ikke kan stole helt på udregninger, der baserer sig på normalfordelingsantagelsen. Det ville nok være en bedre ide at foretage udregningerne på de logaritmetransformerede værdier og så tilbagetransformere endepunkterne....

Vi benytter de udregnede størrelser ovenfor (fra `proc means`) og regner videre i R:

```
> 23.1517+c(-2,2)*3.9116
[1] 15.3285 30.9749
> 24.2284+c(-2,2)*3.2048
[1] 17.8188 30.6380
```

Normalområderne for body mass index er således:

Kvinder: (15.3,31.0)

Mænd: (17.8,30.6)

- *Hvordan passer det med den faktiske fordeling?*

  De direkte udregnede fraktiler for body mass index ses nedenfor:

Obs	gender	P_2_5	P_97_5
1	female	17.6254	33.3333
2	male	19.0364	32.7880

  og vi ser, at de normalfordelingsbaserede grænser her ligger lidt lavere end de, der er baseret på fraktilerne.

- *Hvor mange procent falder udenfor?*

  Her skal man lige lave lidt ekstra programmering, så det er et lidt svært spørgsmål. Vi udregner to nye variable, `over` og `under`, hvorefter vi summerer dem med `proc means`. Bemærk, at man pga. manglende værdier af `bmi` (som anses for at være minus uendelig) er nødt til at kræve `bmi>0`, da disse manglende værdier ellers ville tælle med som liggende under normalområdet.

```
data check;
set sundby;

if gender="female" then over=(bmi>31);
if gender="female" then under=(bmi<15.3 and bmi>0);
```

```
if gender="male" then over=(bmi>30.6);
if gender="male" then under=(bmi<17.8 and bmi>0);
run;

proc means n sum mean data=check;
    class gender;
    var under over;
run;
```

som giver outputtet

```
The MEANS Procedure

             N
gender     Obs    Variable       N            Sum            Mean
-----------------------------------------------------------------
female     827    under        827      2.0000000       0.0024184
                  over         827     42.0000000       0.0507860

male       647    under        647      5.0000000       0.0077280
                  over         647     29.0000000       0.0448223
-----------------------------------------------------------------
```

Vi finder altså ret få (under 1%), der ligger under normalområdet (og altså ville blive vurderet til at være for tynde), men lidt for mange, der ligger over (4-5%).

Man kan evt. gå videre med denne opgave og se på forskellige forklarende variable for BMI.

Måske er BMI højt for

- fysisk inaktive personer (`v17`)
- personer, der drikker meget (`v24af`)