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

Opgave 1: Sundby

Vi betragter 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, f.eks ved at indlæse direkte fra hjemmesiden. Bagefter kan I evt. omdøbe variablene, så de er til at huske.*

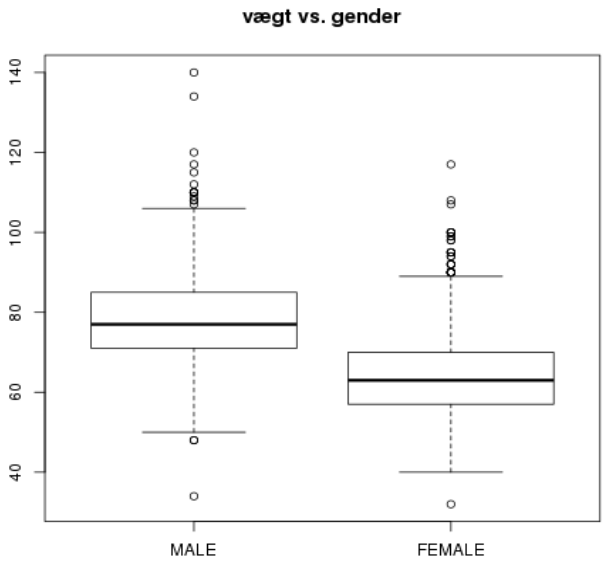
   Vi indlæser uddraget af Sundby-datasættet, og erstatter samtidig de intetsigende variabelnavne med nogle mere sigende, simpelthen ved at udskifte kolonner i data framen `su`. Vi udelader endvidere variablesn `v24af`, som vi ikke skal bruge i denne opgave:

```
su <- read.table("sundby_lille.txt", header=T)

su$gender <- factor(su$kon, levels=1:2, labels=c("MALE", "FEMALE"))

su$vaegt=su$v75;
su$hoejde=su$v76/100;

su$v75 = NULL
su$v76 = NULL
su$v17 = NULL
su$v24af = NULL
```

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

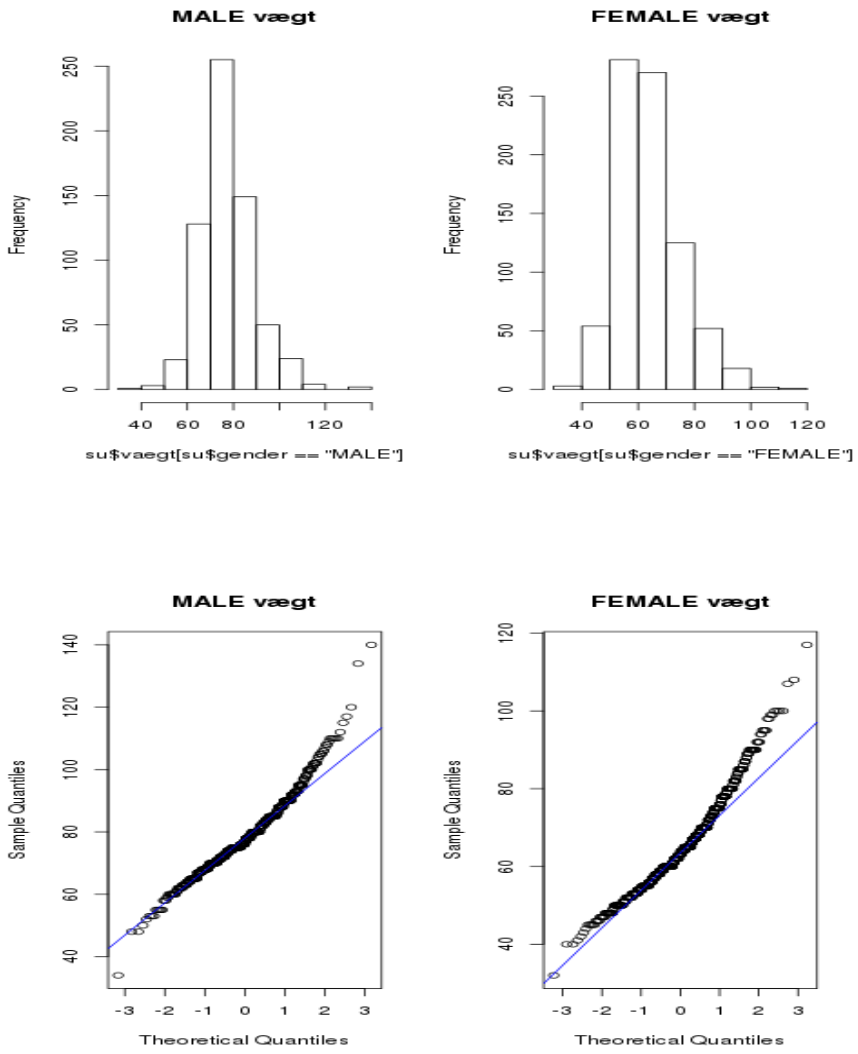
```
boxplot(su$vaegt~su$gender)
```



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

```
par(mfrow=c(1,2))
hist(su$vaegt[su$gender=="MALE"])
hist(su$vaegt[su$gender=="FEMALE"])

par(mfrow=c(1,2))
qqnorm(su$vaegt[su$gender=="MALE"])
qqline(su$vaegt[su$gender=="MALE"],col="blue")
```

```
qqnorm(su$vaegt[su$gender=="FEMALE"])
qqline(su$vaegt[su$gender=="FEMALE"],col="blue")
```



Fordelingen i tal, opdelt efter køn:

```
> tapply(su$vaegt, su$gender, summary)
$MALE
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  34.00   71.00   77.00   78.49   85.00  140.00       8
```

```
$FEMALE
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   32.0    57.0    63.0    64.9    70.0   117.0      21

> tapply(su$hoejde, su$gender, summary)
$MALE
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  1.550   1.750   1.800   1.799   1.850   2.000      15

$FEMALE
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  1.480   1.630   1.680   1.674   1.720   1.980      22
```

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 summary statistics).

3. *Kommenter fundene fra forrige spørgsmål 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 (vist ovenfor) udregner vi normalområder under antagelse af, at vi har at gøre med normalfordelinger for hvert køn.

```
> mean(su$vaegt[su$gender=="MALE"],na.rm=T)+
+   qt(0.975,826)*c(-1,0,1)*sd(su$vaegt[su$gender=="MALE"],na.rm=T)
[1]  55.11386  78.48826 101.86266

> mean(su$vaegt[su$gender=="FEMALE"],na.rm=T)+
+   qt(0.975,826)*c(-1,0,1)*sd(su$vaegt[su$gender=="FEMALE"],na.rm=T)
[1] 42.44030 64.89591 87.35151
```

Normalområderne for vægten er således:

Kvinder: (42.4, 87.4) kg

Mænd: (55.1, 101.9) kg

Sammenligner vi med $2\frac{1}{2}\%$ og $97\frac{1}{2}\%$ fraktilerne, ser vi, at de normalfordelingsbaserede intervaller skyder lidt for lavt, fordi de ikke tager hensyn til den lille skævhed, der er i fordelingerne (mere om dette senere).

```
> pvec=c(0.025,0.975)
> quantile(su$vaegt[su$gender=="MALE"],pvec,na.rm=T)
  2.5%  97.5%
 58.95 106.00

> quantile(su$vaegt[su$gender=="FEMALE"],pvec,na.rm=T)
 2.5% 97.5%
47.00 91.75
```

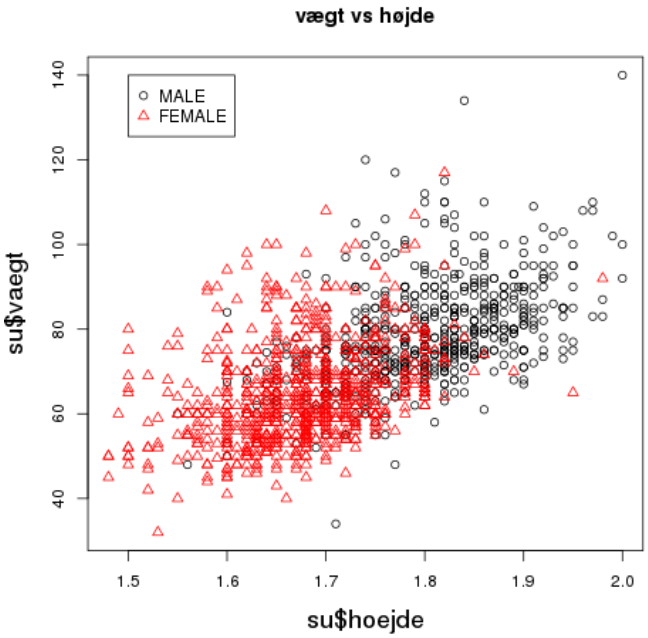
4. *Det ser klart ud til, at mændene vejer mere end kvinderne. Hvilke grunde kunne der være til det?*

   Det har vi allerede berørt ovenfor: Det kan skyldes, at mændene er højere - men det kan selvfølgelig også skyldes kraftigere knoglebygning, større muskler osv.

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:

```
plot(su$hoejde,su$vaegt,pch=as.numeric(su$gender),col=as.numeric(su$gender),
main="vægt vs højde", cex.lab=1.5)
kode=unique(su$gender)
legend(x=1.5, y=140, legend=kode, pch=as.numeric(kode),col=as.numeric(kode))
```



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

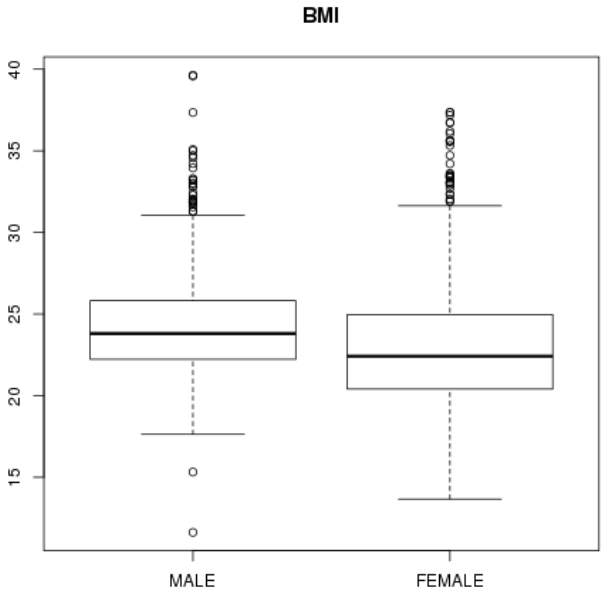
$$\text{BMI} = \frac{\text{vægt i kg}}{\text{højde i meter, kvadreret}}$$

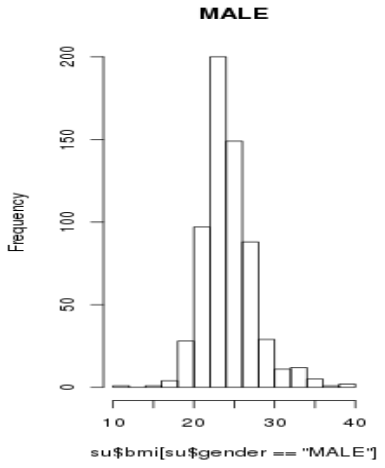
6. *Definer den nye variabel* `bmi` *ved at indføje en sætning inden det første* `run;`. *Check om den er blevet rigtigt defineret, f.eks. ved at udregne gennemsnit mv.*

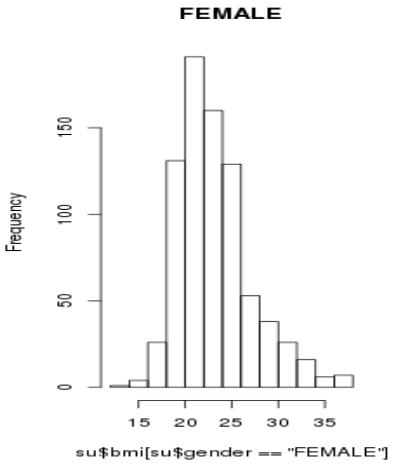
   Dette gør vi ved at skrive `su$bmi=su$vaegt/su$hoejde^2`.

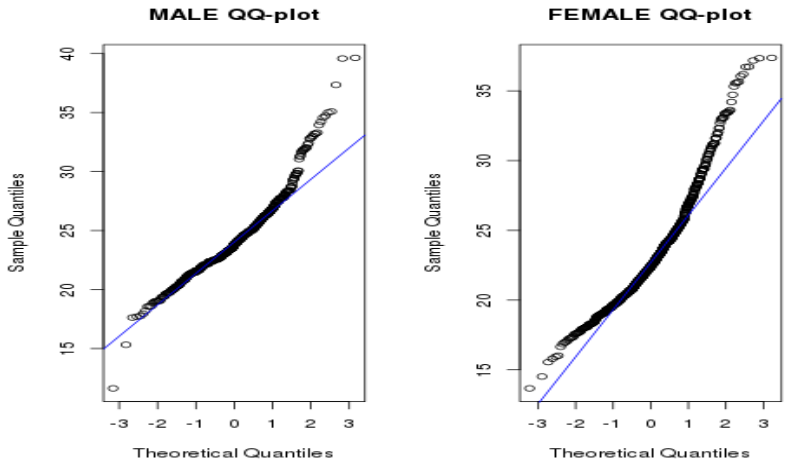
7. *Bestem nu et normalområde for* `bmi` *ved hjælp af en normalfordelingsantagelse.*

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


BMI
40
35
30
25
20
15
MALE
FEMALE


MALE
Frequency
200
150
100
50
0
10
20
30
40
su$bmi[su$gender == "MALE"]


FEMALE
Frequency
150
100
50
0
15
20
25
30
35
su$bmi[su$gender == "FEMALE"]



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....

Alligevel udregner vi de relevante størrelser:

```
> mean(su$bmi[su$gender=="MALE"],na.rm=T)
   +qt(0.975,826)*c(-1,0,1)*sd(su$bmi[su$gender=="MALE"],na.rm=T)
[1] 17.93792 24.22836 30.51879

> mean(su$bmi[su$gender=="FEMALE"],na.rm=T)
    +qt(0.975,646)*c(-1,0,1)*sd(su$bmi[su$gender=="FEMALE"],na.rm=T)
[1] 15.47064 23.15170 30.83277
```

og normalområderne for body mass index er således:

Kvinder: (15.5, 30.8)

Mænd: (17.9, 30.5)

- *Hvordan passer det med den faktiske fordeling (fraktiler=percentiler)?*

  De direkte udregnede fraktiler for body mass index ses nedenfor:

```
> pvec=c(0.025,0.975)
> quantile(su$bmi[su$gender=="MALE"],pvec,na.rm=T)
    2.5%    97.5%
19.04356 32.53110
> quantile(su$bmi[su$gender=="FEMALE"],pvec,na.rm=T)
    2.5%    97.5%
17.62908 33.16937
```

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

- *Hvor mange procent falder udenfor det normalfordelingsbaserede område?*

  Her skal man lige lave lidt ekstra programmering ud fra grænserne for normalområderne, som vi udregnede ovenfor, og som vi derfor nedenfor giver navne.

```
upper.male=mean(su$bmi[su$gender=="MALE"],na.rm=T)
    +qt(0.975,826)*sd(su$bmi[su$gender=="MALE"],na.rm=T)
lower.male=mean(su$bmi[su$gender=="MALE"],na.rm=T)
    -qt(0.975,826)*sd(su$bmi[su$gender=="MALE"],na.rm=T)

upper.female=mean(su$bmi[su$gender=="FEMALE"],na.rm=T)
    +qt(0.975,646)*sd(su$bmi[su$gender=="FEMALE"],na.rm=T)
lower.female=mean(su$bmi[su$gender=="FEMALE"],na.rm=T)
    -qt(0.975,646)*sd(su$bmi[su$gender=="FEMALE"],na.rm=T)
```

Herefter kan vi bruge dem til at tælle hvor mange der falder hhv ovenfor og nedenfor disse grænser:

```
> sum(su$bmi[su$gender=="FEMALE"]>upper.female,na.rm=T)
[1] 43
> sum(su$bmi[su$gender=="MALE"]>upper.male,na.rm=T)
```

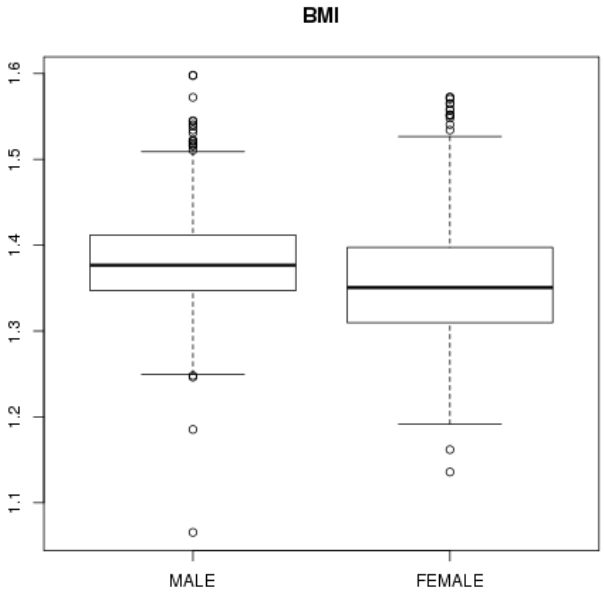
```
[1] 29
> sum(su$bmi[su$gender=="FEMALE"]<lower.female,na.rm=T)
[1] 2
> sum(su$bmi[su$gender=="MALE"]<lower.male,na.rm=T)
[1] 5
```

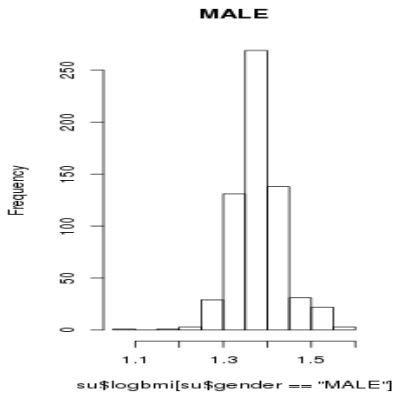
Vi finder altså ret få (under 1%), der ligger under normalområdet (og altså ville blive vurderet til at være for tynde), men lidt for mange, der ligger over (4-5%). Det afspejler (igen) den skæve fordeling af `bmi`.

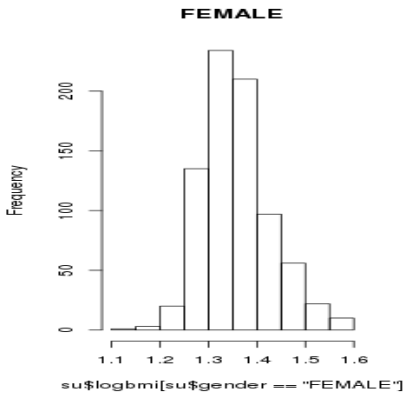
8. *Hvordan ser normalområderne ud, hvis de er baseret på en normalfordelingsantagelse for logaritmen til* `bmi`*?*

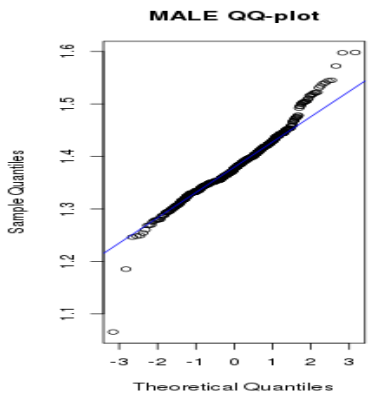
   Vi bemærkede ovenfor en vis skævhed i (bl.a) fordelingen af `bmi`, og en deraf følgende diskrepans mellem normalområder udregnet dels fra gennemsnit og spredning og dels fra de empiriske fraktiler. Denne skævhed forsøger vi nu at transformere os ud af, idet vi tilføjer variablen `logbmi=log10(bmi)`

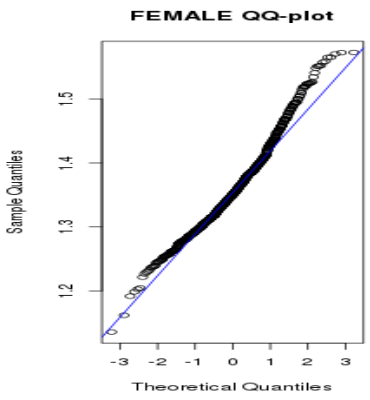
   Vi kan nu gentage figurerne fra tidligere, blot med denne transformerede variabel, og finder:


BMI
1.1
1.2
1.3
1.4
1.5
1.6
MALE
FEMALE


MALE
Frequency
0
50
100
150
200
250
1.1
1.3
1.5
su$logbmi[su$gender == "MALE"]


FEMALE
Frequency
0
50
100
150
200
1.1
1.2
1.3
1.4
1.5
1.6
su$logbmi[su$gender == "FEMALE"]


MALE QQ-plot
Sample Quantiles
1.1
1.2
1.3
1.4
1.5
1.6
-3
-2
-1
0
1
2
3
Theoretical Quantiles


FEMALE QQ-plot
Sample Quantiles
1.2
1.3
1.4
1.5
-3
-2
-1
0
1
2
3
Theoretical Quantiles

Disse figurer synes at vise en noget bedre normalfordelingstilpasning end den utransformerede variabel, og vi fortsætter derfor med at udregne normalområder for denne, som vi med det samme transformerer tilbage til oprindelig skala ved at benytte grænserne som 10'er potenser:

```
> 10^(mean(su$logbmi[su$gender=="MALE"],na.rm=T)
   +qt(0.975,826)*c(-1,0,1)*sd(su$logbmi[su$gender=="MALE"],na.rm=T))
[1] 18.68054 24.02743 30.90475
> 10^(mean(su$logbmi[su$gender=="FEMALE"],na.rm=T)
   +qt(0.975,646)*c(-1,0,1)*sd(su$logbmi[su$gender=="FEMALE"],na.rm=T))
[1] 16.66686 22.84657 31.31759
```

Normalområderne for body mass index er således:

Kvinder: (16.7,31.3)

Mænd: (18.7,30.9)

hvilket er en smule bedre i overensstemmelse med områderne, som udregnes fra de empiriske fraktiler.

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`)

men så skal man bruge analyseteknikker, vi endnu ikke har set på.

Opgave 1: Wright

For 17 patienter er der målt **peak expiratory flow rate** (maksimal udåndingshastighed, i l/min) 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.**

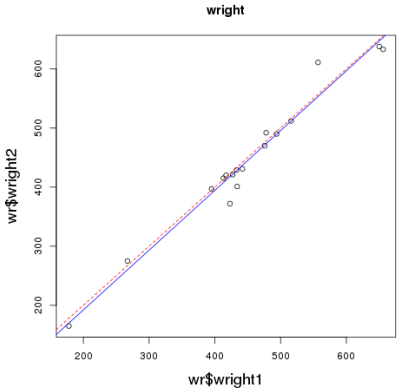
   Der er ikke tale om noget egentligt indlæsningsprogram, men blot koden

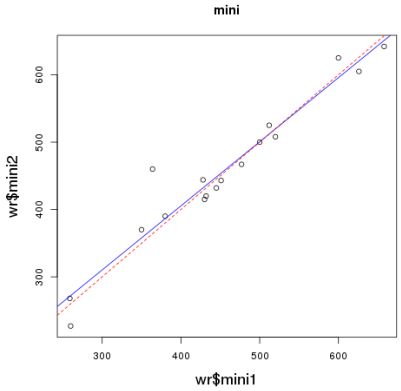
   ```
   wr <- read.table("wright.txt", header=T)
   ```

   Her har vi benyttet navnet `wr` til vores dataframe, som nu består af 17 observationer og 4 variable, nemlig `wright1`, `wright2`, `mini1` og `mini2` (fordi vi har benyttet `header=T`).

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

   ```
   plot(wr$wright1, wr$wright2, main="wright")
   abline(lm(wr$wright2~wr$wright1), col="blue")
   abline(a=0, b=1, col="red",lty=2)

   plot(wr$mini1,wr$mini2, main="mini")
   abline(lm(wr$mini2~wr$mini1),col="blue")
   abline(a=0, b=1, col="red",lty=2)
   ```





Det ses, at observationerne fordeler sig rimeligt omkring en linie (den blå, fuldt optrukket), og hvis man kigger nærmere efter, er denne linie næsten identitetslinien (den røde skraverede), 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).*

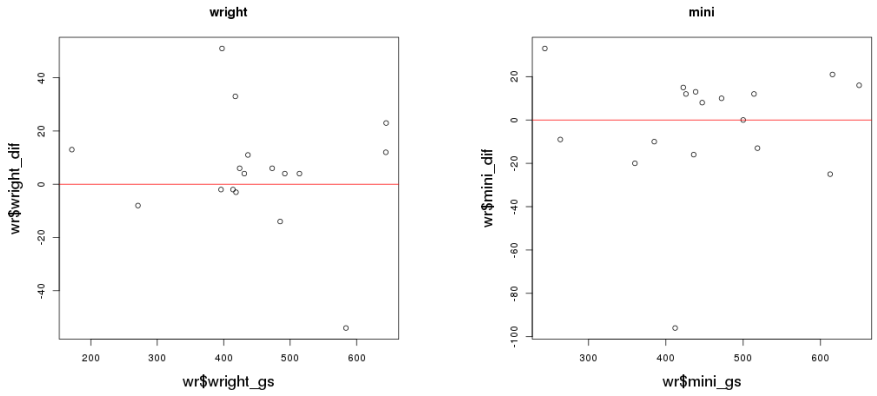
   Vi har nu brug for at tilføje nogle variable i vores data frame `wr`, nemlig differenser mellem dobbeltbestemmelser for hver af metoderne samt gennemsnit af selvsamme dobbeltbestemmelser:

```
wr$wright_dif=wr$wright1-wr$wright2
wr$wright_gs=(wr$wright1+wr$wright2)/2
wr$mini_dif=wr$mini1-wr$mini2
wr$mini_gs=(wr$mini1+wr$mini2)/2
```

   Vi laver herefter (for hver af målemetoderne for sig) et plot af differenserne mod gennemsnittet, med en vandret linie i 0:

```
plot(wr$wright_gs,wr$wright_dif, main="wright")
abline(a=0, b=0, col="red")

plot(wr$mini_gs,wr$mini_dif, main="mini")
abline(a=0, b=0, col="red")
```

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 person, 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.*

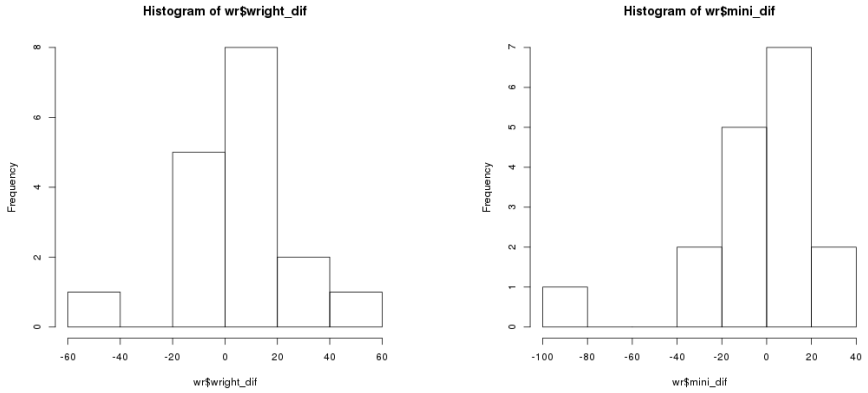
Limits of agreement er normalområder for differenserne, så vi skal finde gennemsnit og spredning for disse. Det drejer sig om 5. og 7. kolonne i data framen, og derfor skriver vi som nedenfor:

```
sapply(wr[,c(5,7)],mean)

sapply(wr[,c(5,7)],sd)
```

hvorved vi får

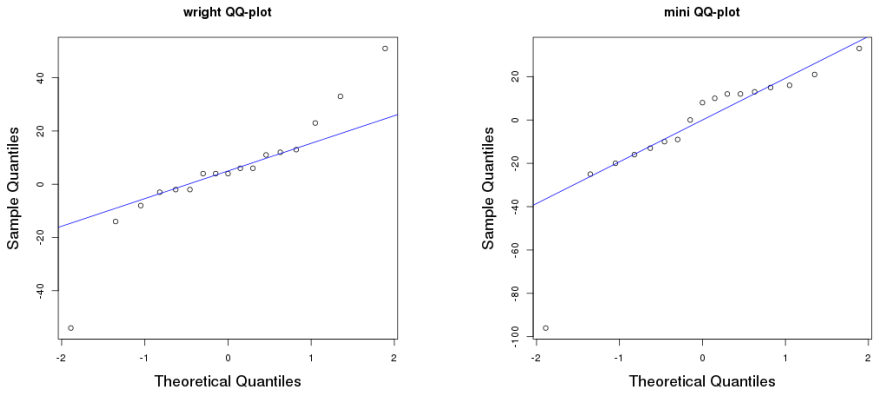
```
> sapply(wr[,c(5,7)],mean)
wright_dif   mini_dif
  4.941176  -2.882353

> sapply(wr[,c(5,7)],sd)
wright_dif   mini_dif
  21.72404   28.87231
```

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, selv om vi må erkende, at vi ikke med nogen rimelighed kan vurdere normalfordelingstilpasningen med så få observationer.

Derfor bliver limits-of-agreement også usikre, hvilket vi vil komme tilbage til nedenfor i en teknisk note.



Histogram og fraktildiagram kan dannes ved at skrive (kun vist for wright-apparaturet):

```
hist(wr$wright_dif)

qqnorm(wr$wright_dif, main="wright QQ-plot")
qqline(wr$wright_dif,col="blue")
```

Tests for normalitet (som man normalt ikke får meget relevant information fra, fordi de plejer at blive godkendt, når man har få observationer og forkastet, når man har mange...) giver faktisk her en afvisning for Mini Wright, på trods af det sparsomme materiale. Det skyldes nok hovedsagelig den (numerisk) meget store differens på -96.

For Wright:

```
> shapiro.test(wr$wright_dif)

Shapiro-Wilk normality test

data:  wr$wright_dif
W = 0.89904, p-value = 0.06546
```

og for Mini Wright:

```
> shapiro.test(wr$mini_dif)

Shapiro-Wilk normality test

data:  wr$mini_dif
W = 0.7913, p-value = 0.001545
```

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 ved at skrive

```
> mean(wr$wright_dif)+c(-2,0,2)*sd(wr$wright_dif)
[1] -38.506899   4.941176  48.389252
> mean(wr$mini_dif)+c(-2,0,2)*sd(wr$mini_dif)
[1] -60.626973  -2.882353  54.862267
```

og finder altså:

Wright: $4.94 \pm 2 \times 21.72 = (-38.51, 48.39)$
Mini Wright: $-2.88 \pm 2 \times 28.87 = (-60.63, 54.86)$

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, og ville dermed lede til lidt bredere grænser:

```
> mean(wr$wright_dif)+qt(0.975,16)*c(-1,0,1)*sd(wr$wright_dif)
[1] -41.111727   4.941176  50.994080
> mean(wr$mini_dif)+qt(0.975,16)*c(-1,0,1)*sd(wr$mini_dif)
[1] -64.088916  -2.882353  58.324210
```

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

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

Wright: $0 \pm 2 \times 21.65 = \pm 43.30$
Mini Wright: $0 \pm 2 \times 28.16 = \pm 56.32$

Teknisk note 2: Usikkerhed på grænserne:
På grund af det lille materiale, vil selve grænserne være ret usikkert bestemt, nemlig med en standard error på ca. $\sqrt{\frac{3}{n}} \times s$, hvilket her giver

- Wright:
  - nedre grænse: $-48.38 \pm 2 \times \sqrt{\frac{3}{17}} \times 21.72 = (-66.6, -30.1)$
  - øvre grænse: $38.50 \pm 2 \times \sqrt{\frac{3}{17}} \times 21.72 = (20.3, 56.7)$
- Mini Wright:
  - nedre grænse: $-54.86 \pm 2 \times \sqrt{\frac{3}{17}} \times 28.87 = (-79.1, -30.6)$
  - øvre grænse: $60.62 \pm 2 \times \sqrt{\frac{3}{17}} \times 28.87 = (36.4, 84.9)$

Det ses af ovenstående, at grænserne er frygteligt dårligt bestemt ud fra så få observationer, så når man vil udtale sig om limits-of-agreement (eller normalområder generelt) bør man have langt flere observationer.

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

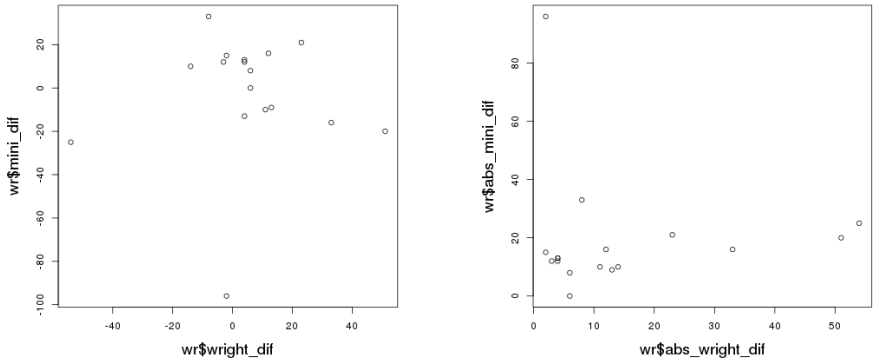
   Baseret på de udregnede limits of agreement ovenfor, ser det ud som om Wright metoden har en noget bedre reproducerbarhed end Mini Wright, idet dens limits of agreement er smallest.

   Man kunne godt lave et egentligt test for dette, men det fører alt for vidt her...specielt når vi lige har indset, *hvor* usikre, disse grænser er.

5. *Tegn et scatter plot af de numeriske differenser mellem dobbeltbestemmelser, med de to metoder på hver sin akse, og vurder på baggrund af*

dette, om der er nogle 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 at definere `wr$abs_wright_dif=abs(wr$wright_dif)` og tilsvarende for `mini`. 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.



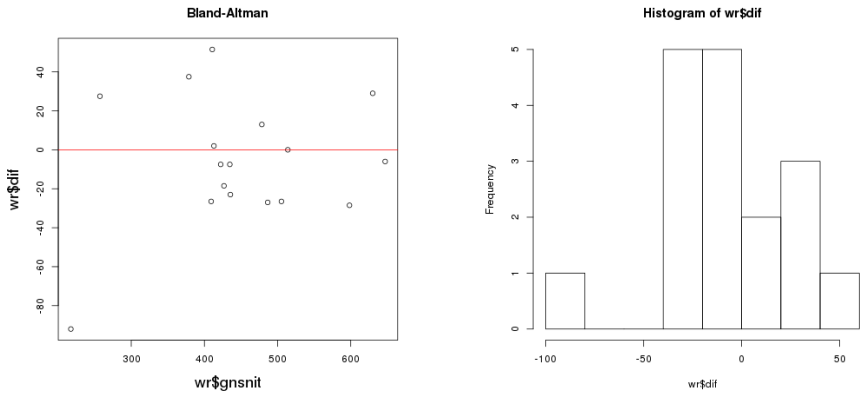
Vi skal vurdere om der er enkelte personer, der har store differenser mellem dobbeltbestemmelserne for *begge* målemetoder, og dette ses ikke umiddelbart at være tilfældet. Vi har (som tidligere bemærket) en enkelt person med en stor diskrepans mellem de to målinger for Mini Wright, men denne person har pænt overensstemmende målinger for Wright apparaturet, så man kan ikke sige, at vedkommende er svær at måle på. Sådanne personer, der er 'svære at måle på' ses i andre sammenhænge, såsom vurdering af leverstørrelse, hvor overvægtige personer er sværere at vurdere.

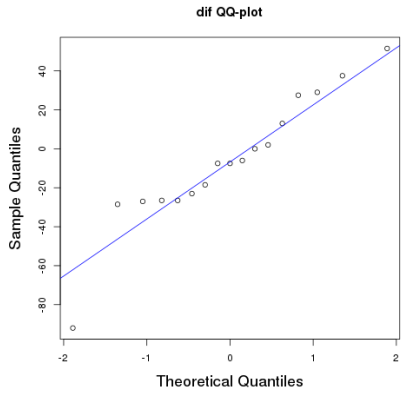
6. **Overensstemmelse:** *Sammenlign nu gennemsnittene af dobbeltbestemmelserne for de to metoder, dvs. tegn igen Bland-Altman plot og udregn limits of agreement, denne gang for sammenligning af de to målemetoder. Kommenter den kliniske anvendelighed af disse grænser.*

   Vi arbejder nu videre med de to gennemsnit, ovenfor kaldet `wr$wright_gs` hhv. `wr$mini_gs`. Igen skal vi se på et plot af differenser

(`wr$dif=wr$wright_gs-wr$mini_gs`) mod gennemsnit (`wr$gnsnit=(wr$wright_gs+wr$mini_gs)/2`) samt vurdere rimeligheden af normalfordelingsantagelsen, inden vi går over til at udregne normalområder for differenserne.

De relevante tegninger er ganske som før





og med kommandoen `mean(wr$dif)+qt(0.975,16)*c(-1,0,1)*sd(wr$dif)` finder vi normalområder

```
> mean(wr$dif)+qt(0.975,16)*c(-1,0,1)*sd(wr$dif)
[1] -76.419037  -6.029412  64.360214
```

Selv om det (igen) er lidt vovet på så få observationer, finder vi altså limits of agreement til (-76.4, 63.4)

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 kan enten foretage et parret T-test:

```
> t.test(wr$wright_gs,wr$mini_gs,paired=T)

Paired t-test

data:  wr$wright_gs and wr$mini_gs
t = -0.7487, df = 16, p-value = 0.4649
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -23.10140  11.04258
sample estimates:
mean of the differences
              -6.029412
```

   eller et test for middelværdi 0 for differenserne `dif`:

```
> t.test(wr$dif)

One Sample t-test

data:  wr$dif
t = -0.7487, df = 16, p-value = 0.4649
alternative hypothesis: true mean is not equal to 0
```

```
95 percent confidence interval:
 -23.10140  11.04258
sample estimates:
mean of x
-6.029412
```

Forudsætningen for dette er rimelig normalitet for differenserne, hvilket vi allerede checkede ovenfor.

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, i form af et Wilcoxon signed rank test på differenserne `dif`:

```
> wilcox.test(wr$wright_gs,wr$mini_gs,paired=T)

Wilcoxon signed rank test with continuity correction

data:  wr$wright_gs and wr$mini_gs
V = 59, p-value = 0.6602
alternative hypothesis: true location shift is not equal to 0

Warning messages:
1: In wilcox.test.default(wr$wright_gs, wr$mini_gs, paired = T) :
  cannot compute exact p-value with ties
2: In wilcox.test.default(wr$wright_gs, wr$mini_gs, paired = T) :
  cannot compute exact p-value with zeroes
```

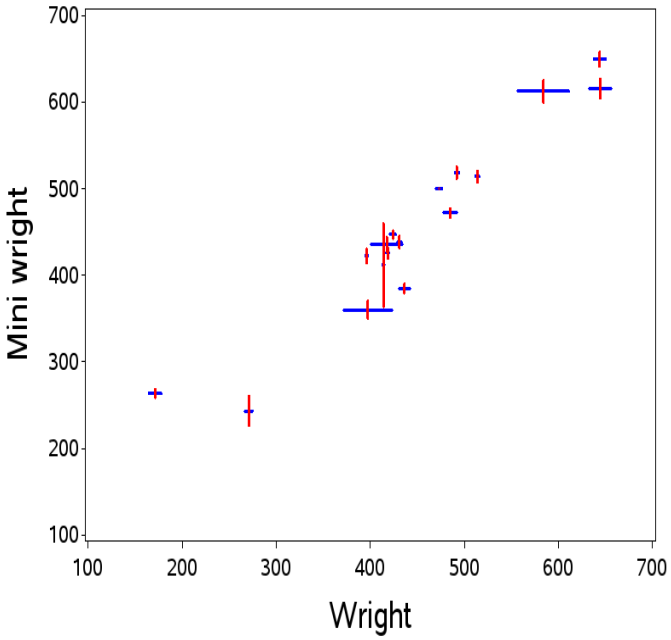
Hermed kan vi imidlertid ikke være *sikre* på, at der ingen forskel er, så vi kvantificerer den sandsynlige forskel ved et konfidensinterval for forskellen mellem middelværdier, som aflæses fra T-test outputtet til at være $(-23.10, 11.04)$

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 (og er foretaget i SAS), 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 (som allerede kommenteret på tidligere).

Reference:

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