

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 aktivitet i fritid (kategorier 1-4, lave tal betyder mest aktiv)

v24af: Antal drukke genstande sidste weekend

1. Indlæs data, f.eks. ved at benytte nedenstående:

```
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;
run;
```

(men lav først om på gåseøjnene, da de bliver forkerte, når man henter dem fra en .pdf-fil).

Overvej, hvad hver enkelt linie refererer til, så I en anden gang kan udnytte en modificeret udgave af denne kode.

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

Som et højdekorrigeret mål for vægt anvendes ofte det såkaldte **BMI** (**B**ody **M**ass **I**ndex), defineret 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.
7. Bestem nu et normalområde for `bmi` ved hjælp af en normalfordelingsantagelse.
 - Hvordan passer det med den faktiske fordeling (fraktiler=percentiler)?
 - Hvor mange procent falder udenfor det normalfordelingsbaserede område?
8. Hvordan ser normalområderne ud, hvis de er baseret på en normalfordelingsantagelse for logaritmen til `bmi`?

Opgave 2: Sammenligning af målemetoder

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, sådan som det fremgår af tabellen nedenfor.

Data: Alle størrelser har enheden l/min

<i>person</i> <i>nr.</i>	<i>Wright</i>		<i>mini Wright</i>	
	X_{1p1}	X_{1p2}	X_{2p1}	X_{2p2}
1	494	490	512	525
2	395	397	430	415
3	516	512	520	508
4	434	401	428	444
5	476	470	500	500
6	557	611	600	625
7	413	415	364	460
8	442	431	380	390
9	650	638	658	642
10	433	429	445	432
11	417	420	432	420
12	656	633	626	605
13	267	275	260	227
14	478	492	477	467
15	178	165	259	268
16	423	372	350	370
17	427	421	451	443

Reference: J.M. Bland & D. Altman: Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 1986, p.307-310

Data kan hentes som tekstfil (`wright.txt`) fra hjemmesiden, og indlæses f.eks. med nedenstående programstump:

```
data opg2;
infile "http://staff.pubhealth.ku.dk/~lts/basal/data/wright.txt"
       URL firstobs=2;
input wright1 wright2 mini1 mini2;
run;
```

1. Kør ovenstående indlæsningsprogram
(men husk at lave om på gåseøjnene).

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).
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.
4. Hvilken af metoderne har den bedste reproducerbarhed?
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.
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.
7. Er der systematisk forskel på de to målemetoder?
Kvantificer!
8. Hvis en forskel på 75 l/min skønnes at have klinisk betydning, kan vi så erstatte Wright med det nye mini Wright?