Faculty of Health Sciences

Basal Statistik

Begreber. Parrede sammenligninger, i SAS

Lene Theil Skovgaard

3. februar 2020

Indhold

- Planlægning af undersøgelse, protokol
- Grafik, Basale begreber
- Parrede sammenligninger
- Limits of agreement
- Appendix med kodning

Home pages:
`http://publicifsv.sund.ku.dk/~sr/BasicStatistics`
E-mail: `ltsk@sund.ku.dk`

$*$: Siden er lidt teknisk

Ide, Problemstilling

- Har "folk" et tilstrækkeligt højt niveau af vitamin D?
- Og hvis ikke, kan vi så gøre noget ved det?
  – eller i hvert fald forstå hvorfor...

Vi ser her på:
studie af kvinder fra 4 lande:
Danmark, Polen, Finland og Irland

Udvælgelse af personer?

- Hvem? Inklusionskriterier kontra repræsentativitet.
- Hvor mange? Dimensionering.
- Design?

Planlægning af undersøgelse

- Formulering af de(t) centrale spøgsmål
  - Er folk generelt oppe på det anbefalede niveau på 25 nmol/l?
  - Er der forskel på landene?
  - I givet fald, hvorfor?
- Hvilke oplysninger skal registreres?
  Formodede forklarende variable = kovariater
  - spisevaner
  - sol eksponering
  - fedme
  - rygning
  - alkohol

Skriv en protokol

Dette er en vigtig del af processen!!

- Man får tænkt sig om på forhånd
- Der bliver udarbejdet information til brug for kolleger mv.
- Det tjener som "ekstra hukommelse" - man glemmer en del hvis dataindsamling eller andet trækker ud
- Det er en nødvendig del af dokumentation i forbindelse med f.eks. etisk komite, ansøgning om midler, anmeldelse af trial etc.
- I forbindelse med den statistiske analyse dokumenterer det, hvad der var den oprindelige strategi og hvad der bør betegnes som tilfældige fund

Eksempel på data

Obs	country	category	vitd	age	bmi	sunexp	vitdintake
1	Ireland	Woman	37.6	71.153	26.391	Sometimes in sun	5.430
2	Ireland	Woman	53.0	70.233	20.540	Sometimes in sun	9.257
3	Ireland	Woman	66.7	70.301	23.500	Sometimes in sun	30.040
4	Ireland	Woman	62.7	70.203	20.800	Avoid sun	3.005
5	Ireland	Woman	89.1	69.932	21.800	Avoid sun	4.068
6	Ireland	Woman	24.3	70.652	36.000	Prefer sun	3.443
38	Ireland	Woman	26.2	71.518	26.950	Sometimes in sun	2.158
39	Ireland	Woman	43.7	70.326	25.723	Prefer sun	3.205
40	Ireland	Woman	35.2	70.638	21.107	Sometimes in sun	7.753
41	Ireland	Woman	17.0	72.049	30.978	Prefer sun	2.906

Det oprindelige datasæt i tekstformat, samt SAS-programmet til indlæsning fremgår af appendix bagest i disse slides (s. 78-80).

Datastruktur, terminologi

- Rækkerne kaldes **observationer** (typisk 1 pr. person)
- Søjlerne kaldes **variable** (en bestemt type oplysning). De kan være
  - Kvantitative variable (Numeriske variable) , dvs. tal, som man kan regne på
    - Vitamin D koncentration (`vitd`, i nmol/l)
    - Alder (`age`)
    - Body mass index (`bmi`)
  - Kategoriske variable (Class-variable, *factors*), som kun kan antage nogle få bestemte værdier, her repræsenteret ved tekst (string)
    - Personens hjemland (`country`)
    - Personens solvaner (`sunexp`)

Anbefalet rækkefølge af aktiviteter

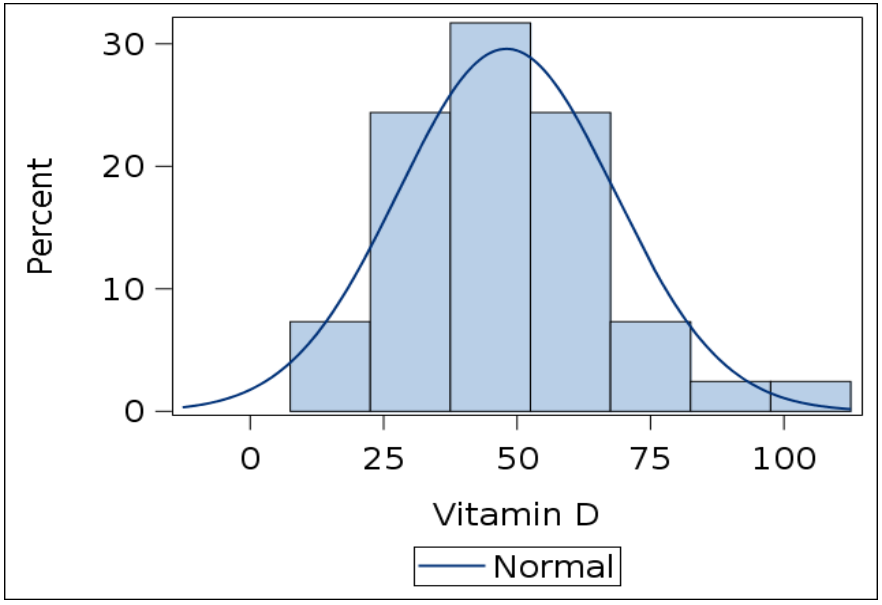
1. **Tænk** (forhåbentlig allerede på protokolstadiet)
2. **Tegn**
   - Histogram
   - Boxplot (typisk for at sammenligne grupper)
   - Scatter plot
3. **Regn**
   - Tabeller
   - Summary statistics
4. Lav analyser
   - Model
   - Estimation
   - Test

Histogram for Irske kvinder

overlejret med fittet normalfordeling

```
proc sgplot data=irlwomen;
    histogram vitd;
    density vitd;
run;
```

God til vurdering
af fordelingen



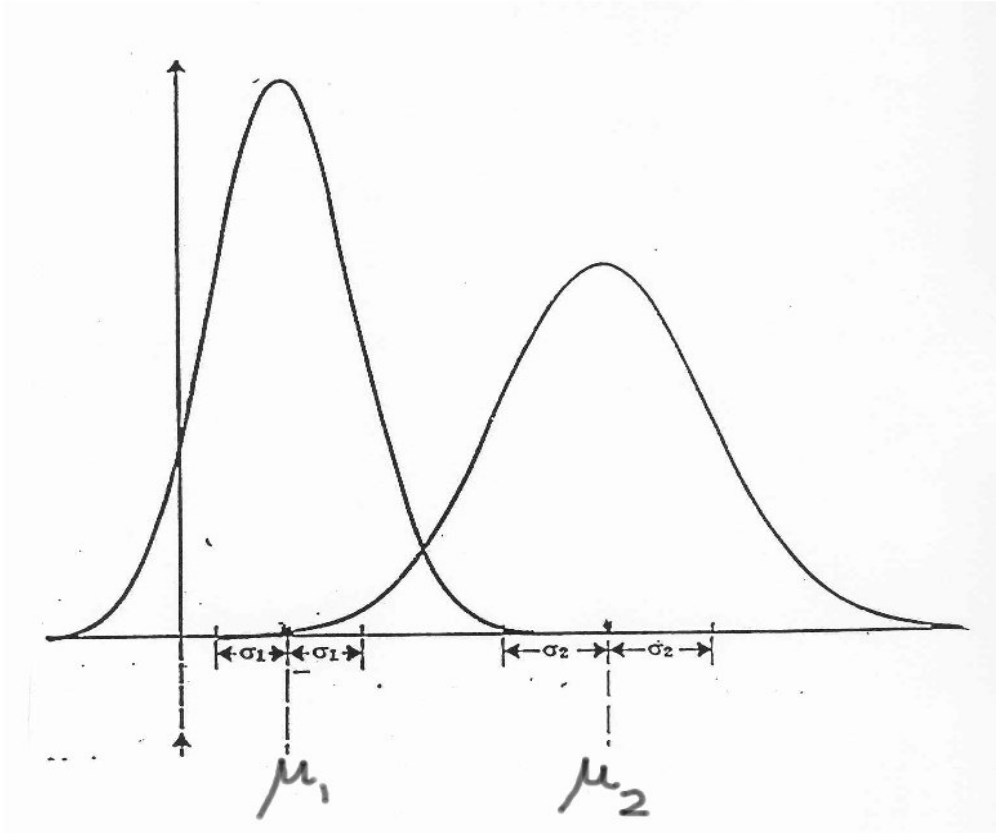
Normalfordelinger

som ikke er nær så vigtige, som nogle af jer sikkert tror!

Middelværdi = mean,
ofte benævnt μ, δ el.lign.

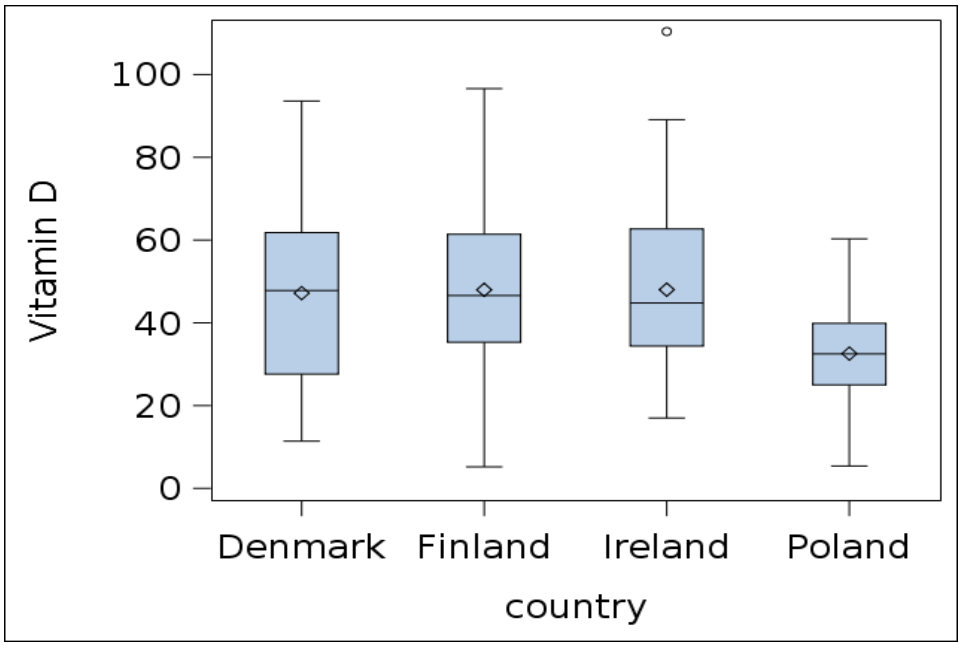
Spredning, ofte benævnt σ
(eller s, når den udregnes):

N(μ,σ^2)



Box-plot for alle kvinder

```
proc sgplot data=women;
    vbox vitd / category=country;
run;
```



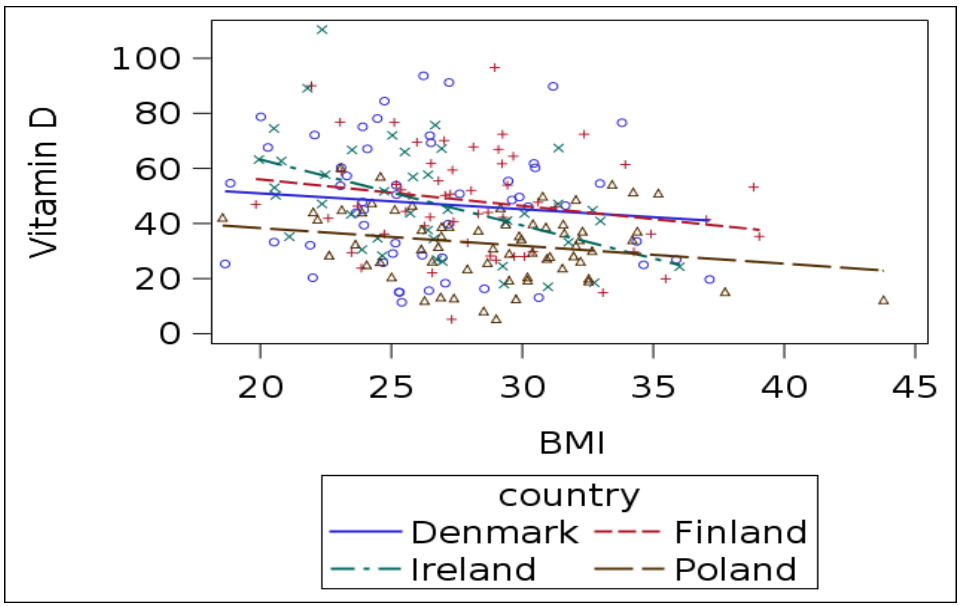
God til sammenligninger

- Box: 25% - 75% fraktil
- Streg: Median
- ◇: Gennemsnit
- Whiskers: definitionsafhængig

...noget med variansanalyse

Scatter-plot af Vitamin D niveau mod BMI

```
proc sgplot data=women;
    reg X=bmi Y=vitd / group=country;
run;
```



Er der en afhængighed af BMI? Måske lineær?

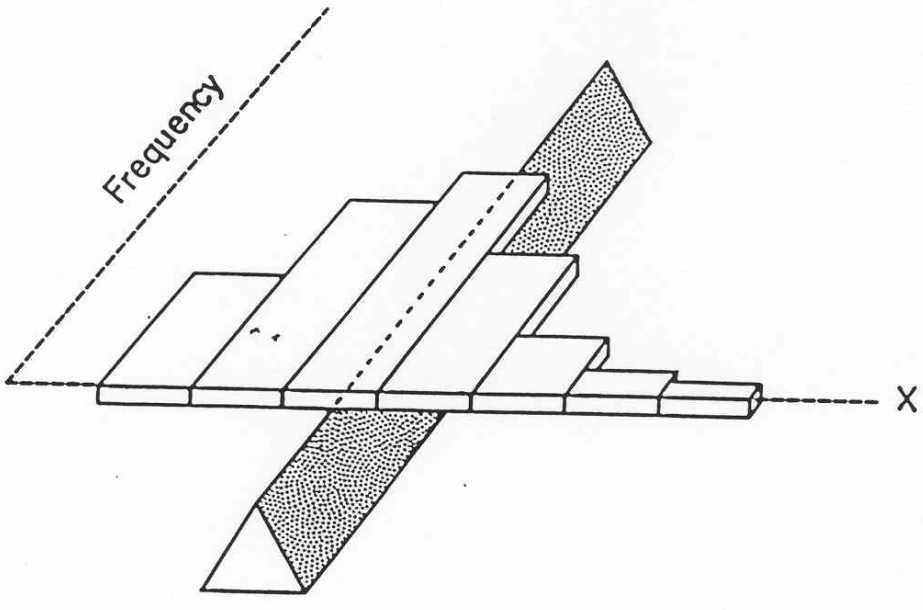
... noget med regressionsanalyse

Regn: Summary statistics, I

Observationer $y_1, \ldots, y_n$

- Location, centrum
  - Gennemsnit: $\bar{y} = \frac{1}{n}(y_1 + \cdots + y_n)$
  - Median: midterste observation, efter størrelsesorden
- I symmetriske fordelinger vil gennemsnit og median være ens (pånær tilfældigheder, naturligvis)
- I skæve fordelinger vil de *ikke* være ens:
  Typisk er der hale mod de høje værdier,
  så *gennemsnittet er større end medianen.*

Gennemsnit=tyngdepunkt



- kan opfattes som ligevægtspunkt
- påvirkes kraftigt af yderlige observationer

Eksempel:
Indlæggelsestider:
5,5,5,7,10,16,106 dage
Gennemsnit: 154/7=22 dage.
Repræsentativt for hvad?

På den anden side, hvis omkostninger er proportionale med indlæggelsestiden, så **er** det måske gennemsnittet, der er interessant for hospitalsledelsen.

Regn: Summary statistics, II

Observationer $y_1, \ldots, y_n$

Variation

- Varians: $s^2 = \frac{1}{n-1}\Sigma(y_i - \bar{y})^2$

  Spredning = Standardafvigelse= Standard Deviation = $\sqrt{\text{varians}} = s = \text{SD}$

- Fraktiler
  Medianen er 50% fraktilen, men der er en fraktil svarende til alle procenter, se næste side

Fraktiler for vitamin D

Sorter data med den mindste først, tæl:

5% fraktil: 5% er mindre end dette, 95% er større

25% fraktil: 25% er mindre, 75% er større
kaldes også **nedre kvartil** eller Q1

50% fraktil: 50% er mindre, 50% er større
Midterste observation, kaldes også **median**

75% fraktil: 75% er mindre, 25% er større
kaldes også **øvre kvartil** eller Q3

$2\frac{1}{2}\%$ og $97\frac{1}{2}\%$ er vigtige,
fordi 95% af observationerne ligger imellem disse

Summary statistics for vitamin D

```
proc means mean median Q1 Q3 stddev
           data=vitamind maxdec=2;
  class country;
  var vitd;
run;
```

giver outputtet

Analysis Variable : vitd

country	N Obs	Mean	Median	Lower Quartile	Upper Quartile	Std Dev
Denmark	53	47.17	47.80	27.60	61.80	22.78
Finland	54	47.99	46.60	35.30	61.40	18.72
Ireland	41	48.01	44.80	34.40	62.70	20.22
Poland	65	32.56	32.50	25.00	39.90	12.46

Summary statistics for vitamin D

Bemærk:

- Median og gennemsnit er nogenlunde ens,
  svarende til rimelig symmetri i Boxplottene s. 11
- Dette kan *ikke* bruges til at påstå, at der er tale om en Normalfordeling
- Polen ligger lavere end de øvrige lande, undtagen måske for Q1. Bemærk, at også spredningen er lavere for Polen.

Traditionel fortolkning af spredningen s

Hovedparten af observationerne ligger inden for $\bar{y} \pm ca.2 \times s$
dvs. *sandsynligheden for at en tilfældig udtrukket person fra populationen har en værdi i dette interval* er *stor...*

For Vitamin D blandt irske kvinder finder vi
reference område = normalområde

$$48.0 \pm 2 \times 20.2 = (7.6, 88.4)$$

Hvis data er normalfordelt, vil dette interval indeholde ca. 95% af fremtidige observationer. **Hvis ikke, tja....**

Noter: "Ca. 2-tallet" er i virkeligheden $(1 + \frac{1}{n}) t_{97.5\%}(n-1)$,
og usikkerheden på grænserne (st.err.) er ca. $\sqrt{\frac{3}{n}} s \approx 5.46$

Er folk oppe på det anbefalede niveau på 25 nmol/l?

Vi har lige fundet normalområdet til (7.6, 88.4), hvilket fortæller os, at en del personer må forventes at have værdier under 25 - *hvis* vi har at gøre med en normalfordeling.

Med definitionen `lav=(vitd<25);` kan vi også bare tælle...

```
proc freq data=irlwomen;
    table lav;
run;
```

hvorved vi får en lille tabel

lav	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	36	87.80	36	87.80
1	5	12.20	41	100.00

som altså viser, at

mere end 12% af de irske kvinder har et for lavt niveau.

Normalområde / Referenceområde

Område, der omslutter de centrale 95% af observationerne:

- nedre grænse: $2\frac{1}{2}\%$ fraktil
- øvre grænse: $97\frac{1}{2}\%$ fraktil

  (For irske kvinder fås 18 hhv 89.1 nmol/l, se kode s. 84)

Hvis fordelingen kan beskrives ved en normalfordeling $\mathrm{N}(\mu,\sigma^2)$, kan de sande fraktiler udtrykkes som

$2\frac{1}{2}\%$ fraktil: $\mu - 1.96\sigma \approx \bar{y} - 1.96s \approx \bar{y} - 2s$

$97\frac{1}{2}\%$ fraktil: $\mu + 1.96\sigma \approx \bar{y} + 1.96s \approx \bar{y} + 2s$

og normalområdet udregnes derfor som

$$\bar{y} \pm ca.2 \times s \approx (\bar{y} - 2 \times s, \quad \bar{y} + 2 \times s)$$

Praktisk konstruktion af referenceområde

- Store datasæt:
  Brug fraktiler

- Mellemstore datasæt:
  Brug en rimelig fordelingsantagelse,
  typisk normalfordelingen,
  evt. efter transformation

  Her er normalfordelingsantagelsen vigtig!

  Mellemstort...: Med 100 observationer er usikkerheden på grænserne ca. 20%

- Små datasæt:
  Lad være med det!!

Hvad er en rimelig fordelingsantagelse?

Her: passer normalfordelingen nogenlunde?

- Gode argumenter
  - Tegn histogram, er det symmetrisk?
  - Fraktildiagram, er det lineært? (kræver tilvænning)
- Svagere indicier
  - Er gennemsnit og median tæt på hinanden?
  - Er fraktilerne (f.eks. 25% og 75%) symmetriske omkring medianen?

Bemærk:

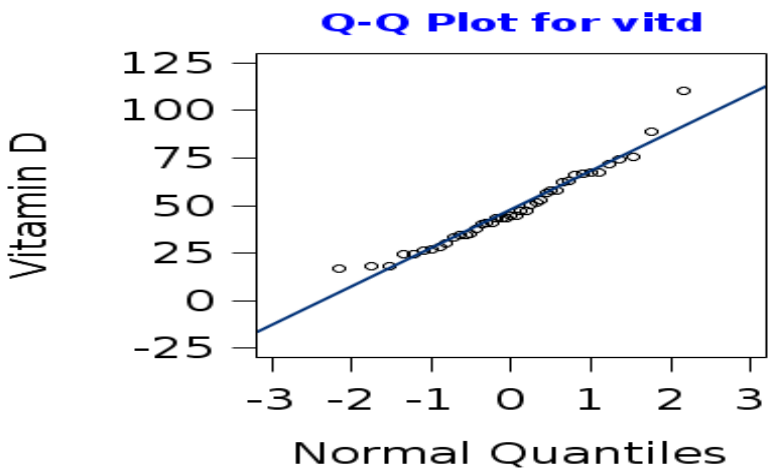
Et stort antal observationer sikrer ikke,
at der er tale om en normalfordeling.
– og et lille materiale *kan* sagtens være et sample fra en Normalfordeling – vi kan bare ikke afgøre det....

Fraktildiagram, "qqplot"

til check af Normalfordelingen:

Sammenlign observationerne (Y-aksen) med teoretiske fraktiler, baseret på en fittet normalfordeling, her normeret (X-aksen),

```
proc univariate data=irlwomen;
  var vitd;
  qqplot / normal(mu=EST sigma=EST color=red);
run;
```



Dette bør se lineært ud, hvis der er tale om en normalfordeling.

Hvorfor normalfordelingen?

- Det er ofte en rimelig approksimation
  - Evt. efter transformation
    med logaritme, kvadratrod, invers,...
- **Central grænseværdisætning:**
  - Sum (eller gennemsnit) af et stort antal variable får en fordeling, der efterhånden kommer til at *ligne* en normalfordeling
    (sum af normalfordelinger er igen en normalfordeling).
- Rimelig let at arbejde med, fordi standard programmel er udviklet for normalfordelingen.

men som regel er antagelsen ikke specielt vigtig!
Undtagelsen er konstruktion af referenceområder

Eksempler på pæne normalfordelinger

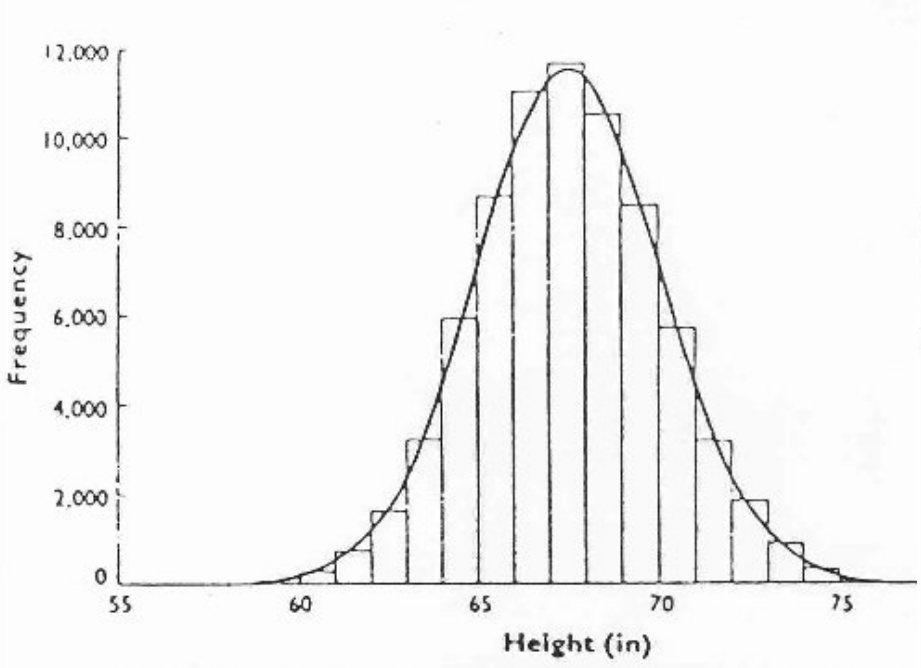


Fig. 2.10. A distribution of heights of young adult males, with an approximating normal distribution (Martin, 1949, Table 17 (Grad

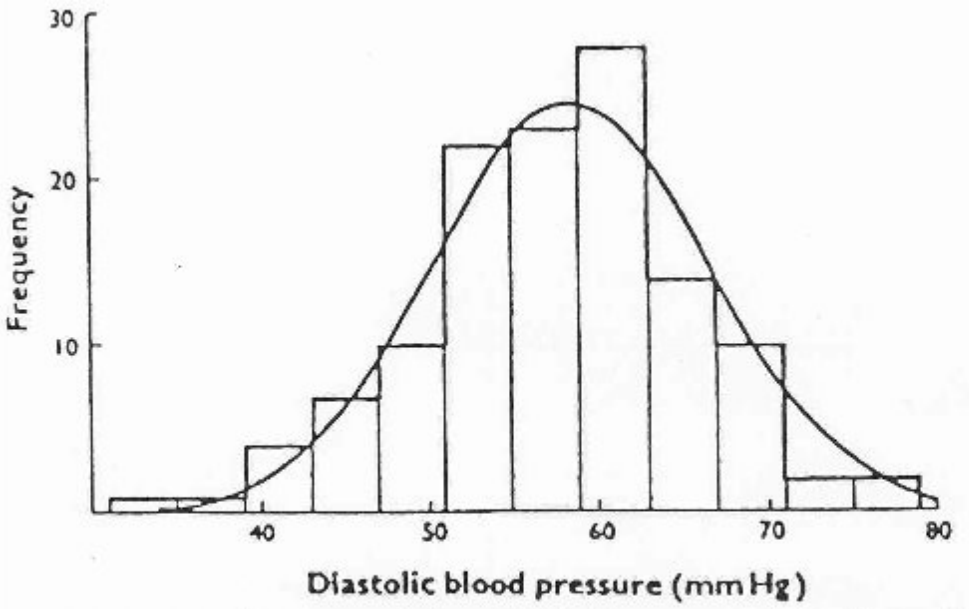


Fig. 2.11. A distribution of diastolic blood pressures of schoolboys with an approximating normal distribution (Rose, 1962, Table 1).

Typisk afvigelse fra normalfordelingen

som regel når der er tale om ret lave værdier, f.eks. hormonmålinger (eller immunoglobulin):

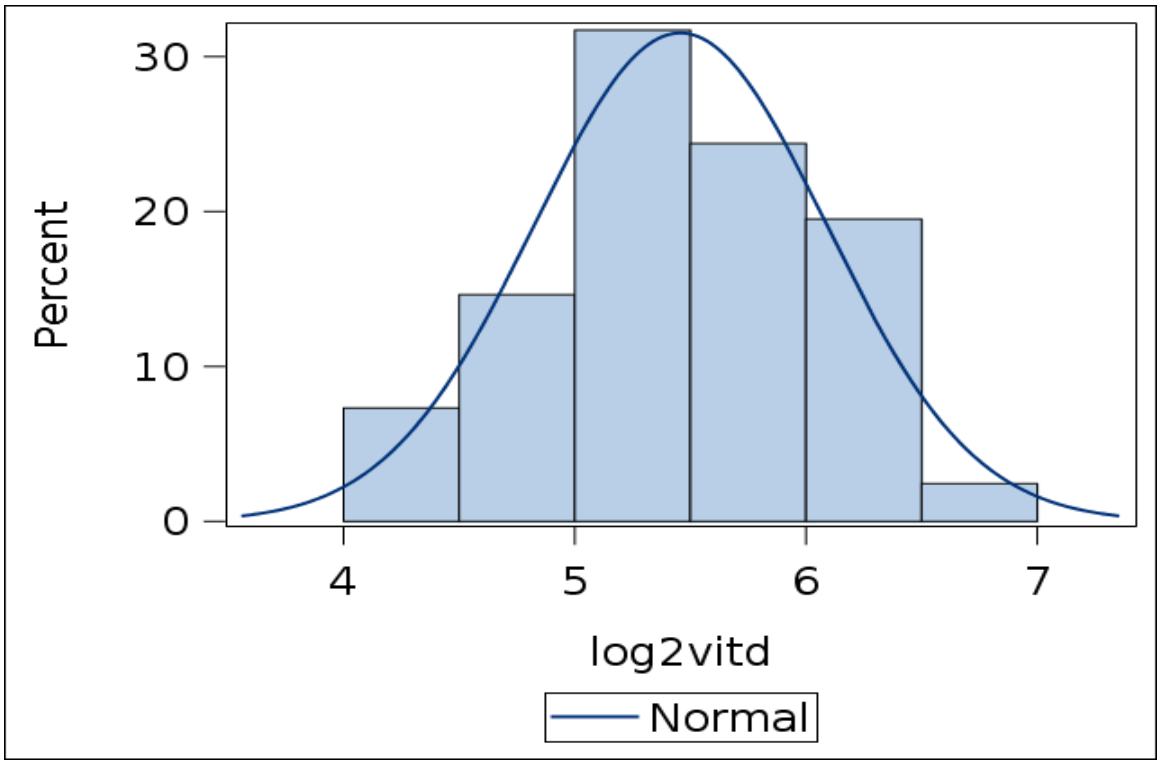
- Histogrammet er skævt, med en hale mod de høje værdier
- Gennemsnittet er en del større end medianen

Løsning: Transformer med en logaritme

- ligegyldig hvilken: naturlig, 10-tals, 2-tals
  Tilføj linien: `log2vitd=log2(vitd);` (se s. 80)
- bare man transformerer tilbage med den samme anti-logaritme, *når man har regnet færdigt*... dvs. $\exp(\text{noget})$, 10^{noget} eller 2^{noget}

Histogram for logaritmerede værdier af vitamin D

her 2-tals logaritmen



igen med fittet overlejret normalfordeling

Her har vi - måske - *lidt* bedre symmetri

Referenceområde, baseret på logaritmer

kode som s. 17 (uden `maxdec=2`):

```
       Analysis Variable : log2vitd

 N            Mean          Median         Std Dev
-----------------------------------------------------
41       5.4564121       5.4854268       0.6327180
-----------------------------------------------------
```

For logaritmen til Vitamin D for irske kvinder finder vi
$5.456 \pm 2 \times 0.633 = (4.19, 6.72)$

Dette interval skal tilbagetransformeres med anti-logaritmen:
$(2^{4.19}, 2^{6.72}) = (18.3, 105.6)$ nmol/l

Sammenlign med (7.6, 88.4) uden transformation, eller de empiriske fraktiler (18.0, 89.1)

Skæve fordelinger: Immunoglobulin

Summary statistics for 298 personer:

Analysis Variable : igm

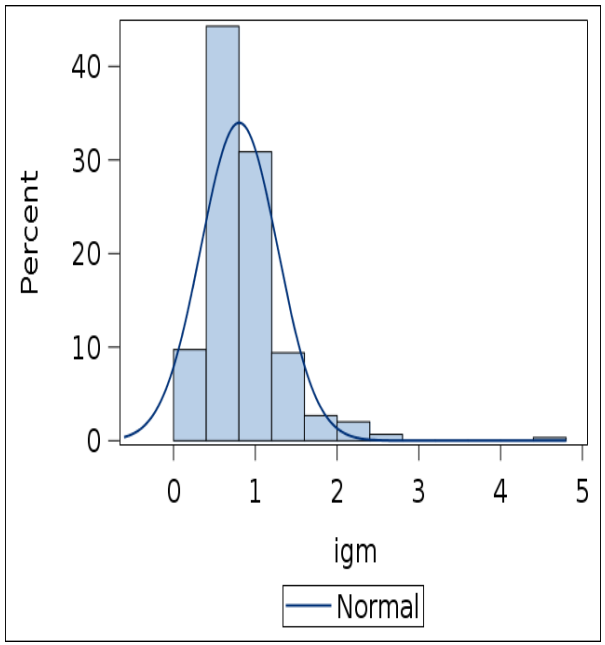
N	Mean	Median	Minimum	Lower Quartile	Upper Quartile	Maximum	Std Dev
298	0.803	0.700	0.100	0.500	1.000	4.500	0.469

Bemærk:

- Gennemsnittet er noget højere end medianen, så vi har nok en hale med høje værdier (jvf. s. 13)
- Maximum (og Q3) viser, at det specielt er de øverste 25% af fordelingen, der er trukket op mod de høje værdier.
- Spredningen er stor i forhold til gennemsnittet - den kan faktisk overhovedet ikke fortolkes!

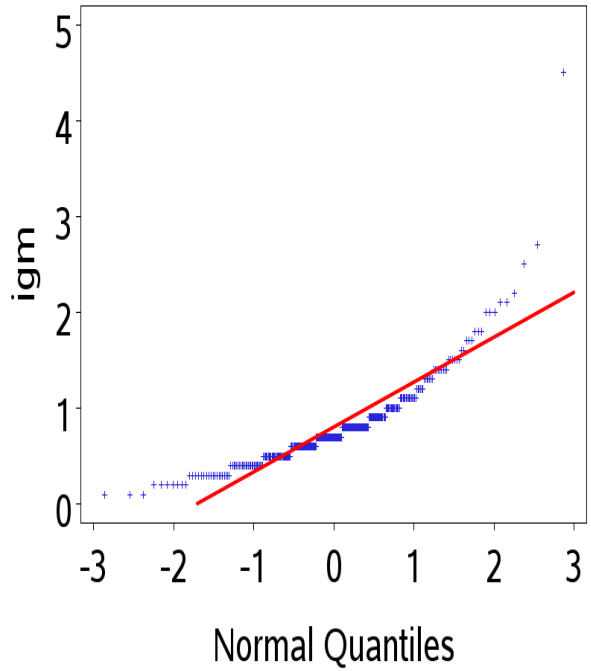
Immunoglobulin, fortsat

Histogram



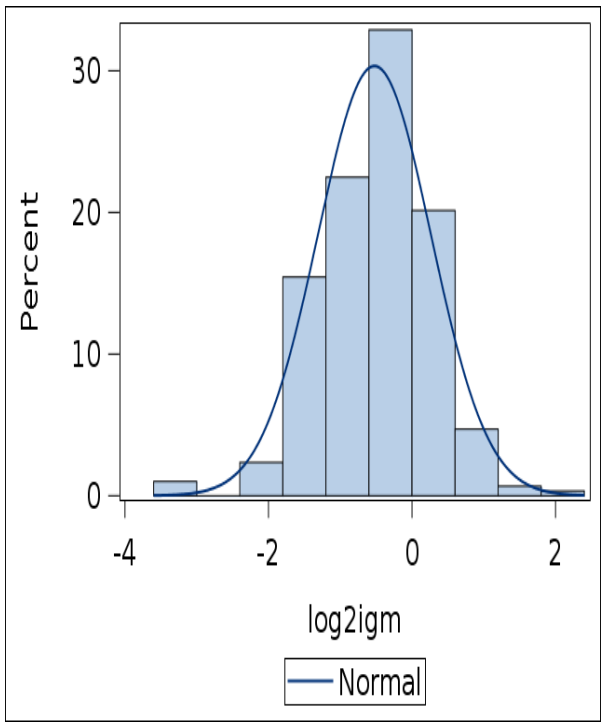
Tydeligt ikke-normalfordelt
(bemærk observationer
langt ude til højre)

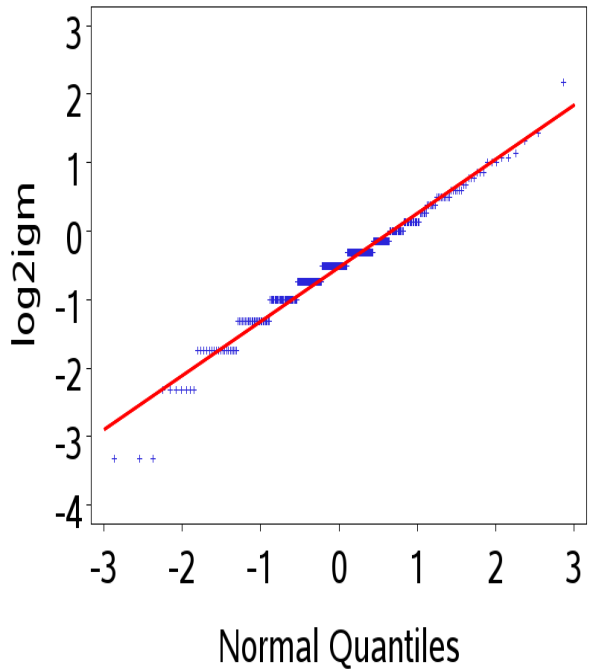
Fraktildiagram



Tydeligt ikke-lineært
(hængekøjefacon)

Immunoglobulin, log2-transformeret





Næsten lineært

Væsentlig bedre normalfordelingstilpasning

Referenceområde for immunoglobulin

Urimelige værdier er i rød *kursiv*:
Ufortolkelige værdier er bare i *kursiv*

Data	gennemsnit (median)	spredning	Referenceområde
utransformeret	*0.803*	*0.469*	*(-0.135, 1.741)*
log2-transformeret	*-0.524*	*0.789*	*(-2.102, 1.054)*
tilbagetransformeret	(0.695)	-	(0.233, 2.076)
empiriske fraktiler	(0.700)	-	(0.2, 2.0)

Sådan foregår tilbagetransfomationen:
Referenceområde for logaritmer: (-2.102, 1.054)
Tilbagetransformeret: $(2^{-2.102}, 2^{1.054}) = (0.23, 2.08)$

Lad være med at tilbagetransformere spredningerne!

Vigtigheden af normalfordelingen

afhænger af *formålet* med undersøgelsen

- vigtig
  - ved beskrivelser
  - *specielt* ved konstruktion af referenceområder
- ikke så vigtig
  - ved sammenligninger, vurdering af effekter
    hvor det kun er residualerne, der antages normalfordelte, og
    hvor antallet af observationer kan redde situationen
- ikke på nogen måde påkrævet for kovariater!
  — som I ikke ved så meget om endnu...

Parrede sammenligninger

Vi skal se på en situation, hvor vi ønsker at sammenligne to fordelinger/situationer, men hvor observationer fra den ene situation
"er parret med" observationer fra den anden fordeling.

Eksempler:

- Målinger på samme person før og efter en behandling
- Sammenligning af to grupper/behandlinger, hvor individerne er individuelt matchet på f.eks. køn, alder, bopæl etc.
- To målemetoder, der benyttes på samme person/dyr/blodprøve

Formålet med undersøgelsen

kan være flere forskellige:

- Vurdering af effekten af en behandling
  (således at man har målinger både før og efter behandling).
  Sædvanligvis vil man dog her også have en kontrolgruppe,
  hvis det er muligt (5. uges emne)
- Sammenligning af to behandlinger, hvor man ved hjælp af matchning eller cross-over har sørget for parrede observationer for behandlingerne
- Vurdering af, om to målemetoder/apparaturer måler det samme, eller rettere:
  kvantificering af, hvor stor diskrepans, der ses imellem dem

Sammenligning af målemetoder

To metoder til bestemmelse af slagvolumen:

- MF: bestemt ved Doppler ekkokardiografi
- SV: bestemt ved cross-sectional ekkokardiografi

Ubrugelig tabel:

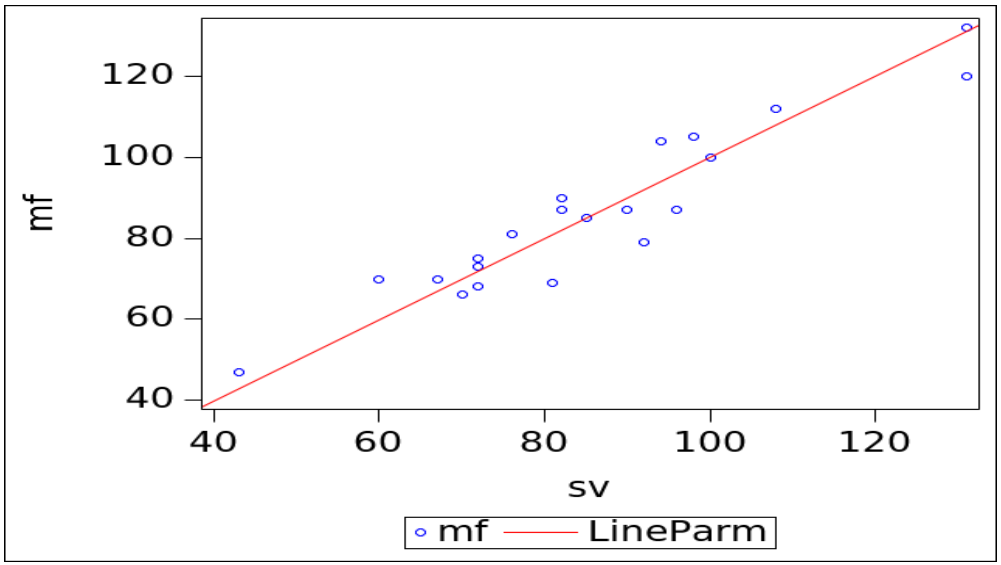
person	MF	SV
1	47	43
2	66	70
3	68	72
4	69	81
5	70	60
.	.	.
.	.	.
.	.	.
.	.	.
17	104	94
18	105	98
19	112	108
20	120	131
21	132	131
gennemsnit	86.05	85.81
SD	20.32	21.19
SEM	4.43	4.62

Måler de to målemetoder "det samme"?

Kode for indlæsning s. 86

Scatter plot af MF vs. SV

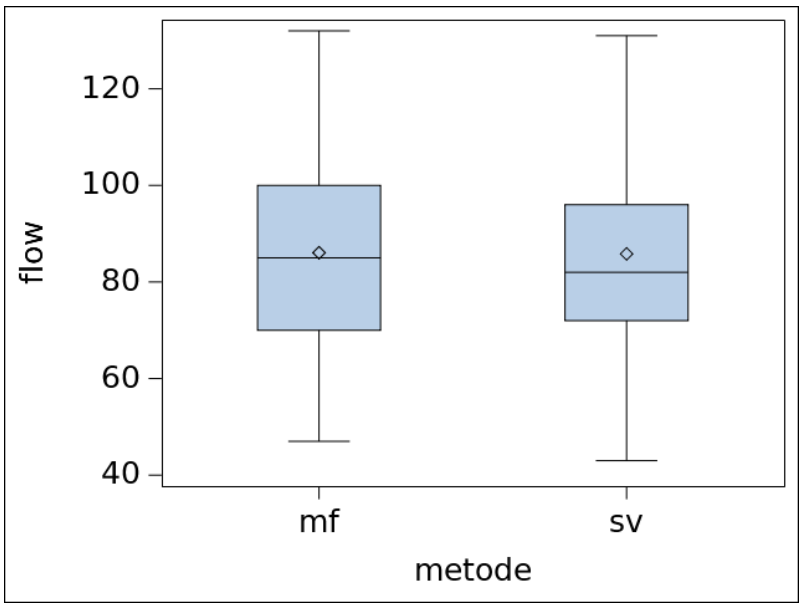
med indlagt identitetslinie



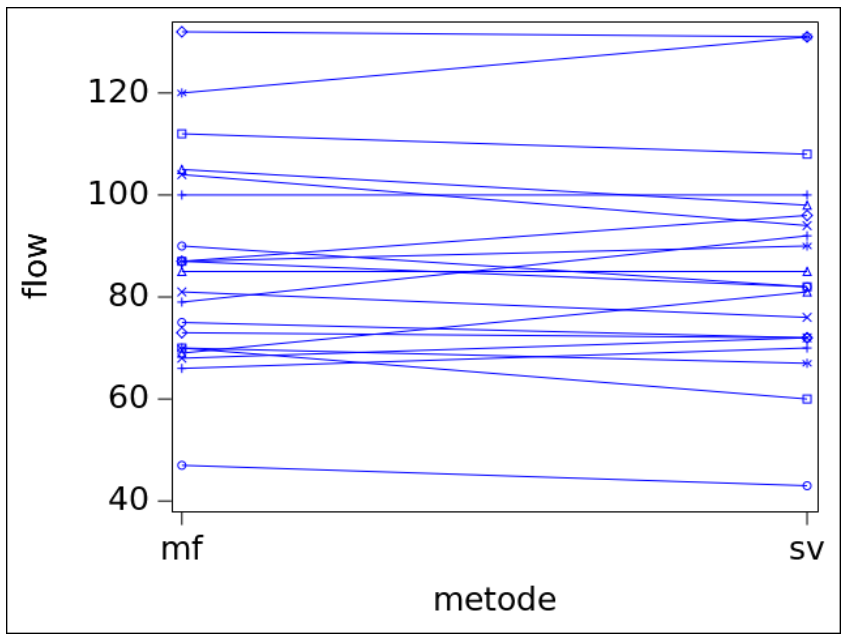
Er der en pæn lineær sammenhæng mellem de to målemetoder?
Er de måske endda rimeligt ens?
(kode s. 87)

Man skal kunne se parringen!

Forkert tegning



Rigtig tegning



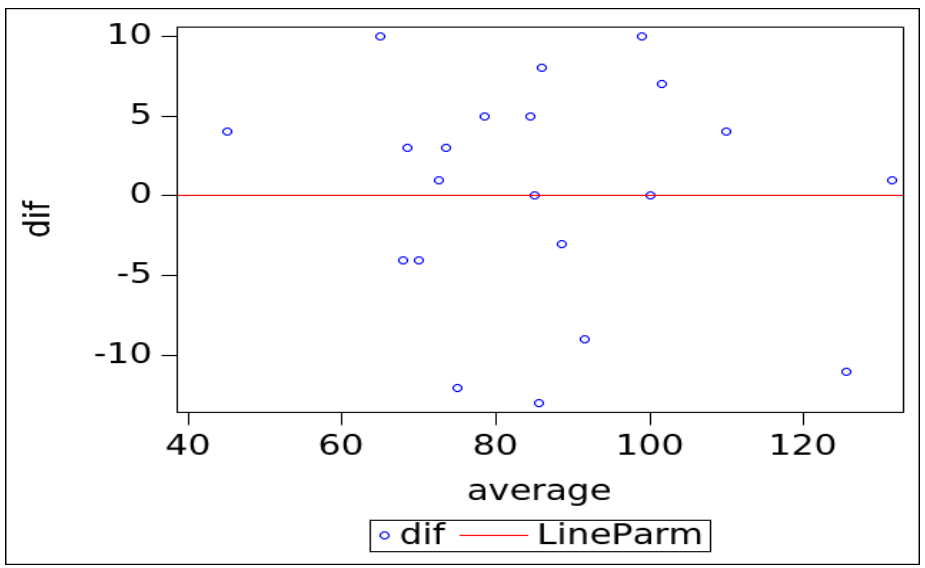
Kode s. 88 (kræver omstrukturering af data til *langt* format)

Analyse af parrede data

- Personen er **sin egen kontrol**
  Det giver **stor styrke** til at opdage evt. forskelle.
- Se på individuelle differenser
  – men på hvilken skala?
  - Er differensernes størrelse nogenlunde uafhængig af niveauet?
  - Eller er der snarere tale om relative (procentuelle) forskelle:
    I så fald skal der tages differenser på en logaritmisk skala.
- Undersøg om differenserne har middelværdi 0:
  parret T-test

Bland-Altman plot

Scatter plot af differenser `dif=mf-sv` mod gennemsnit `average=(mf+sv)/2` for den enkelte person (kode s. 87):



Ligger differenserne omkring 0?
Er der ca. den samme fordeling for alle gennemsnit?

Statistisk model for parrede data

X_i: flowmålingen MF for den i'te person
Y_i: flowmålingen SV for den i'te person

Differenser $D_i = X_i - Y_i$ $(i = 1, \cdots, 21)$
uafhængige, normalfordelte
med middelværdi δ og spredning σ

Bemærk:

- Kun antagelser om differenser er nødvendige
  – fordi det er et parret design
- Intet krav om fordeling af *selve flowmålingerne*!
  – *kun* af differenserne.

Inferens, Statistisk analyse

Med udgangspunkt i indsamlede data, hvad kan vi så sige om den sandsynlighedsmekanisme (model, f.eks. de ukendte parametre), der har frembragt disse data?

- Estimation:
  Når vi ser disse 21 differenser, hvad kan vi så sige om de to ukendte parametre, δ og σ?
- Test:
  Ser der ud til at være systematisk forskel på de to metoder, dvs. er $\delta = 0$?
- Prædiktion:
  Hvor store forskelle kan vi forvente i praksis?

Gangen i en statistisk analyse

- **Modelkontrol**: Er forudsætningerne opfyldt?
  Burde komme først, men kommer af praktiske grunde efter estimationen.

- **Estimation**:
  Hvilke parameterværdier passer bedst med observationerne?
  Og hvor sikkert er de bestemt?

- **Modelreduktion** (test af hypoteser):
  Er simplere beskrivelser tilladelige?
  Passer en simplere model næsten lige så godt?

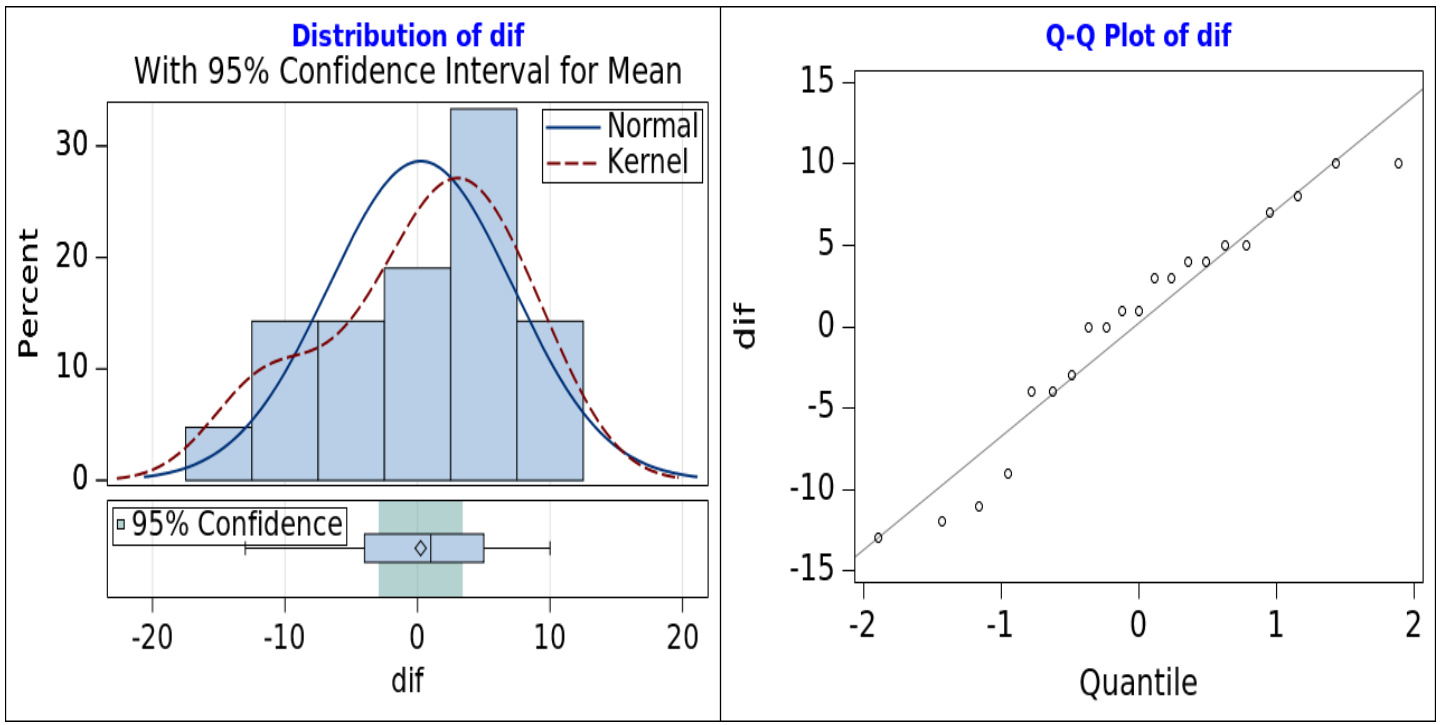
Antagelser for den parrede sammenligning

Differenserne $D_i = X_i - Y_i, \quad i = 1, \ldots, 21$:

- er **uafhængige:**
  personerne har ikke noget med hinanden at gøre
- har **samme spredning** (varians):
  vurderes ved det såkaldte Bland-Altman plot af differenser mod gennemsnit (se s. 41)
- er **normalfordelte:**
  vurderes grafisk eller numerisk
  - histogram og fraktildiagram, hmm....kun 21 observationer
  - formelt test? nix...
  - somme tider vigtig, andre gange ikke

Fordeling af differenser

Figurerne fremkommer ved at udføre T-test (s. 59ff)



Passer normalfordelingen nogenlunde?

*Estimation

Differenserne $D_i = \text{MF}_i - \text{SV}_i \sim N(\delta, \sigma^2)$ er 21 uafhængige observationer fra en normalfordeling

Maximum likelihood princippet giver her, at

parametrene δ og σ estimeres ved henholdsvis gennemsnittet $\bar{D}$ og den tidligere omtalte spredning s (s. 15).

Vi markerer sædvanligvis estimater ved at sætte en ˆ (hat) over, altså $\hat{\delta} = \bar{D}$ (udtales *delta-hat*)
Her finder vi $\hat{\delta} = 0.238$, men

Estimater skal angives med tilhørende usikkerheder!
– gerne i form af et konfidensinterval

Usikkerhed på et estimat

Hvad betyder det?

F.eks. (som her) usikkerhed på et gennemsnit, som estimat for (skøn over) en ukendt middelværdi.

Vi kan tænke på gentagelser af undersøgelsen:

- Hver gang får vi et nyt estimat (for middelværdien)
- Vi kan studere fordelingen af sådanne estimater
- Spredningen i denne fordeling angiver usikkerheden. Den kaldes som regel standard error of the estimate

Hvis det er en middelværdi, kaldes den
standard error of the mean

Altså:

Spredning på gennemsnittet kaldes

Standard error (of the mean), SEM

Hermed angives usikkerheden på gennemsnittet, og der gælder

$$\text{SEM} = \frac{\text{SD}}{\sqrt{n}}$$

SEM bliver således mindre, når n bliver større

Den bruges til at konstruere konfidensintervaller,
som kommer nu...

Konfidensinterval = Sikkerhedsinterval

Interval, der *"fanger"* den ukendte parameter
(her middelværdien δ)
med stor (typisk 95%) sandsynlighed.

Hvor kan vi tro på at den faktiske middelværdi δ ligger?

(Approksimativt) 95% konfidensinterval for middelværdi

$$\bar{D} \pm 2 \times \text{SEM}$$

Eksakt for normalfordelingen, bortset fra "ca. 2",
("mere præcist 2-tal" er $t_{97.5\%}(20) = 2.086$)

Vi siger, at intervallet har dækningsgrad 95%,
på engelsk *coverage*

Konfidensinterval for δ = forskel MF-SV

95% konfidensinterval: $\bar{D} \pm$ 'ca. 2' $\times$ SEM,
eller med et "mere præcist 2-tal": $t_{97.5\%}(20) = 2.086$

men vi behøver ikke håndregning, vi bruger i stedet SAS:

```
proc means N mean stderr clm;
 var dif;
run;
```

og får outputtet

```
   Analysis Variable : dif

                                          Lower 95%       Upper 95%
 N            Mean       Std Error     CL for Mean     CL for Mean
-------------------------------------------------------------------
21       0.2380952       1.5195625      -2.9316567       3.4078471
-------------------------------------------------------------------
```

Konfidensintervallet ses at være (-2.93, 3.41)

Fortolkning af konfidensinterval (= sikkerhedsinterval)

Konfidensinterval for middelværdien af forskellen δ mellem MF og SV blev estimeret til

$$(-2.93, 3.41)$$

Det betyder:

- Der kan ikke påvises nogen systematisk forskel (bias) mellem de to typer målinger
- Vi kan dog heller ikke afvise, at der *kan* være forskel
- En evt. bias vil med stor sikkerhed (her 95%) være mindre end. ca. $3 - 3.5$ (til hver side)

Test af hypotese (ofte kaldet nulhypotese)

Kan vi nøjes med en simplere model?
Kunne en eller flere parametre i en model være en kendt værdi (ofte 0, deraf navnet)?

Modelreduktion: Model $\longrightarrow$ (nul)hypotese (H_0).
Kan den forenklede model *tænkes* at være den rigtige?
Eksempelvis:

- Er der systematisk forskel på de to målemetoder? ($\delta = 0$)
- Har mænd og kvinder samme middelværdi af blodtrykket? ($\mu_1 = \mu_2$ eller $\mu_1 - \mu_2 = 0$)
- Er blodtrykket uafhængig af alderen? (hældning $\beta = 0$)
- Er der samme sandsynlighed for farveblindhed hos piger og drenge? ($p_1 = p_2$ eller $p_1 - p_2 = 0$)

Test af hypotese, II

Ofte ønsker vi at forkaste hypotesen, fordi vi så har fundet en effekt, f.eks. en forskel på to grupper.

Andre gange ønsker man at vise, at der ingen forskel er, og så skal man bruge konfidensintervaller i stedet for!

Teststørrelse: En størrelse, der måler
diskrepans mellem observation og hypotese

- Stor diskrepans: Forkast hypotesen, fordi den passer dårligt sammen med data. Men hvor stor?
- Undersøg om teststørrelsen (diskrepansen) er
  værre / mere ekstrem
  end hvad der kan forventes ved tilfældighedernes spil.

Teststørrelsens fordeling

Teststørrelsen måler diskrepansen mellem observationerne og hypotesen.

Selv når hypotesen H_0 er *fuldstændig sand*, vil vi aldrig opnå fuldstændig overensstemmelse mellem model og observationer (f.eks. ikke nøjagtigt samme gennemsnit for MF og SV).

- Hvor store vil afvigelserne typisk være, når H_0 er sand?
- Hvilke værdier af teststørrelsen vil vi typisk få, og med hvilken hyppighed (sandsynlighed)?
- Det kaldes *fordelingen* af teststørrelsen, og den kan beregnes, så vi ved, hvad der er *normalt* og hvad der er unormalt/ekstremt

Test af "ingen bias" mellem MF og SV

dvs. test af nulhypotesen $H_0 : \delta = 0$

Vi benytter et T-test på differenserne, og dette fremkommer som brøken

$$\frac{\text{estimat - hypotesev\ae rdi}}{\text{standard error for estimat}}$$

og det viser sig, at denne (under H_0) er T-fordelt (Student-fordelt) med et antal
frihedsgrader, som er antallet af observationer minus 1

- Lille (numerisk) t: God tilpasning
- Stor (numerisk) t: Dårlig tilpasning

T-test for MF vs. SV, fortsat

Her finder vi teststørrelsen:

$$t = \frac{\hat{\delta} - 0}{\text{SEM}} = \frac{0.24 - 0}{1.52} = 0.158 \sim t(20)$$

Der er 20 frihedsgrader, fordi der er 21 observationer og kun 1 fælles middelværdi.

Passer denne værdi (0.158) godt med en t-fordeling med 20 frihedsgrader?

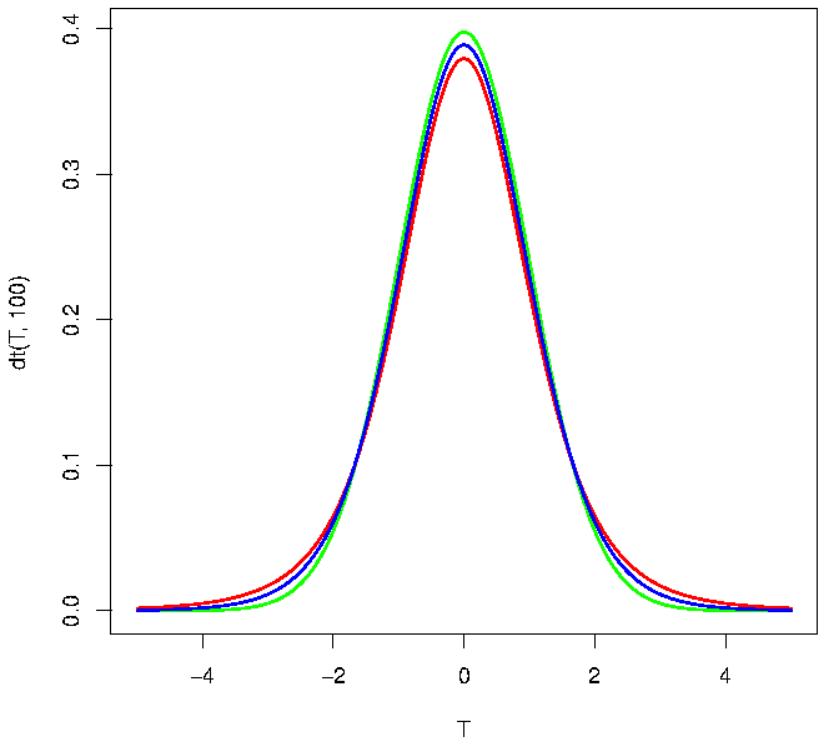
Ja, den ligger ret centralt i fordelingen, og vi kan derfor ikke se noget galt med vores hypotese (se næste side).

∗Teknisk note

t-fordelingen (Student fordelingen)

har en parameter df, der kaldes *antallet af frihedsgrader* (her: 5, 10, 100).

- Mange frihedsgrader: Fordelingen ligner normalfordeling
- Få frihedsgrader: Tungere haler.



Parret T-test i praksis

Vi vil teste ens middelværdi for MV og SV, med differenser `dif`
Der er flere alternativer til at gøre dette:

1. ```
   proc ttest data=mf_sv;
     paired mf*sv;
   run;
   ```
2. ```
   proc ttest data=mf_sv;
     var dif;
   run;
   ```
3. Summary statistics for differenserne:

   ```
   proc means N mean std stderr t probt data=mf_sv;
    var dif;
   run;
   ```

Typisk output fra parret T-test

Her fra det første alternativ fra forrige side:

```
The TTEST Procedure
Difference:  mf - sv

  N          Mean       Std Dev      Std Err      Minimum      Maximum
 21        0.2381        6.9635       1.5196     -13.0000      10.0000

    Mean          95% CL Mean          Std Dev       95% CL Std Dev
  0.2381       -2.9317    3.4078        6.9635       5.3275  10.0558

    DF     t Value     Pr > |t|
    20        0.16        0.8771
```

Estimeret differens: 0.2381 (1.5196)

P-værdi: P=0.88

Fortolkning af P-værdi

P-værdien er sandsynlighed for "dette eller værre", altså
større diskrepans end den observerede,
under nulhypotesen (dvs. hvis nulhypotesen er sand).

Hvis der kun er en ganske lille sandsynlighed for at få noget, der er værre end det vi har, så må det være ret slemt, og vi må forkaste.

Her finder vi $P = 0.88$, altså stor sandsynlighed for at få noget, der er værre end det vi har, så nulhypotesen ser rimelig ud: vi kan ikke forkaste.

Signifikans

Hvis P-værdien er under 0.05, siger man at
testet er signifikant på 5% niveau. Man forkaster hypotesen.

Signifikansniveauet α vælges sædvanligvis til 5% ($\alpha = 0.05$), men der er tale om et arbitrært valg.

Man bør derfor angive selve P-værdien, og allervigtigst:

Angiv estimat med konfidensinterval!

Her blev det udregnet til (-2.93, 3.41), – så vi kunne med det samme have set, at 0 *var* en rimelig værdi for middelværdien

Test vs. konfidensinterval

Der er **ækvivalens**, i den forstand, at:

- Hvis konfidensintervallet (sikkerhedsintervallet) indeholder 0, er testet ikke signifikant
- Hvis konfidensintervallet (sikkerhedsintervallet) *ikke* indeholder 0, er testet signifikant

Her fik vi konfidensintervallet (-2.93, 3.41), med tilhørende P-værdi P=0.88, så vi kan ikke forkaste hypotesen om middelværdi 0 for differenserne.

Men var det alt, hvad vi gerne ville vide?
Nej, vi vil gerne vide, hvor store forskellene typisk er....

Limits-of-agreement

Hvor store afvigelser vil man typisk se mellem de to metoder for individuelle personer (enkelt individer)

Limits of agreement er en speciel betegnelse for normalområdet for differenser, dvs.
gennemsnit $\pm$ *2 spredninger*, for differenserne

$$\bar{D} \pm \text{'ca. 2'} \times \text{SD} = 0.24 \pm 2 \times 6.96 = (-13.68, 14.16)$$

Disse grænser er vigtige for at afgøre om to målemetoder kan erstatte hinanden. Det er nemlig ikke nok,
at der ikke er nogen systematisk forskel!!

Og her er normalfordelingen vigtig!

Repetition: De to slags spredninger

SD: Spredningen i populationen

SEM: Standard error of the mean

$$SEM = SD(\bar{x}) = \frac{SD(x)}{\sqrt{n}}$$

(eller mere generelt blot standard error)

SD bruges til beskrivelser

SEM bruges til sammenligninger

SD til beskrivelser ("*Tabel 1*")

Variable	Antal	Gennemsnit $\bar{X}$	Spredning SD
Alder	100	45	10
Immunoglobulin	100	0.80	0.47

Her tænker man:

- Patienterne er nok ca. 25-65 år
- og har immunoglobulinværdier på ca. ....
  UPS: De kan være negative!!
  (så der burde være transformeret, eller....)

Her er normalfordelingen vigtig!
fordi vi udtaler os om enkeltobservationer

Eksempel fra litteraturen

Malhotra, Welch, Rosenbaum & Poiesz:

American Journal of Clinical Oncology • Volume 39, Number 3, June 2016 *Efficacy and Toxicity of Docetaxel*

TABLE 1. Demographic and Clinical Characteristics of Metastatic Prostate Cancer Patients (n = 41) Treated With 3 Dosing Regimens of Docetaxel

Characteristics	Total Group	q1w (n = 12)	q2w (n = 14)	q3w (n = 15)
Age (mean [SD]) (y)	69.7 (8.9)	72.3 (7.7)	70.2 (10.0)	67.2 (8.5)
Site of metastasis (n [%])				
Bone	36 (87.8)	11 (91.7)	12 (85.7)	13 (86.7)
Lymph node	10 (24.4)	3 (25)	4 (28.6)	3 (20)
Liver*	3 (7.3)	0	3 (21.4)	0
Lung	2 (4.9)	0	1 (7.1)	1 (6.7)
Bladder/ureter	5 (12.2)	1 (8.3)	3 (21.4)	1 (6.7)
PSA (mean [SD])				
Start	321.9 (744.4)	378.8 (458.8)	271.6 (812.1)	323.3 (894.6)
Nadir	94.6 (191.9)	141.9 (201.7)	68.1 (192.2)	81.3 (190.0)
Total dose (mean [SD]) (mg/m^2)	705.9 (479.0)	601.9 (286.9)	841.1 (649.5)	663.0 (411.7)
Total # cycles (mean [SD])*	14.8 (11.1)	19.0 (11.6)	17.1 (13.4)	9.3 (5.2)
Response (n [%])				
Yes	27 (66)	7 (58)	10 (71)	10 (67)
No	14 (34)	5 (42)	4 (29)	5 (33)
Follow-up status (n [%])				
Dead	31 (76)	11 (92)	10 (71)	10 (67)
Lost to follow up	1 (2)	0	0	1 (6)
Alive	9 (22)	1 (8)	4 (29)	4 (27)

The median age for the total group was 71.5 years with a range of 49.9 to 88.5 years.
Median PSA at start was 67.4 with a range of 0.4 to 3528.
Median PSA at nadir was 20.9 with a range of 0 to 733.9
Median total dose of docetaxel was 540 mg/m^2, with a range of 100 to 1980 mg/m^2.
*$P < 0.05$; remainder of comparisons were not significantly different from one another across treatment groups.
PSA indicates prostate-specific antigen; q1w, weekly regimens of docetaxel; q2w, 2 weekly regimens of docetaxel; q3w, 3 weekly regimens of docetaxel.

Hvad bør man så gøre?

Hvis fordelingen er **tydeligt skæv**
eller på anden måde afviger tydeligt fra normalfordelingen, bør man ikke engang angive gennemsnit og spredning, men snarere:

- fraktiler:
  - median
  - inter-quartile range, IQR:
    intervallet mellem 25% og 75% fraktil

For **helt små materialer** angives evt.

- median og range

..og så laver man ikke statistik, men kasuistik

SEM til sammenligninger ("*Tabel 2*")

	Gruppe 1 (n=35)		Gruppe 2 (n=65)	
Variable	Gennemsnit $\bar{X}_1$	SEM_1	Gennemsnit $\bar{X}_2$	SEM_2
Alder	43	1.7	46	1.2
Immunoglobulin	0.63	0.08	0.89	0.06

Her tænker man:

- De to grupper ser ens ud rent aldersmæssigt
- men har måske nok forskellige niveauer af immunoglobulin

Her er normalfordelingen ikke så vigtig,
fordi det er gennemsnit, vi udtaler os om

Central grænseværdisætning

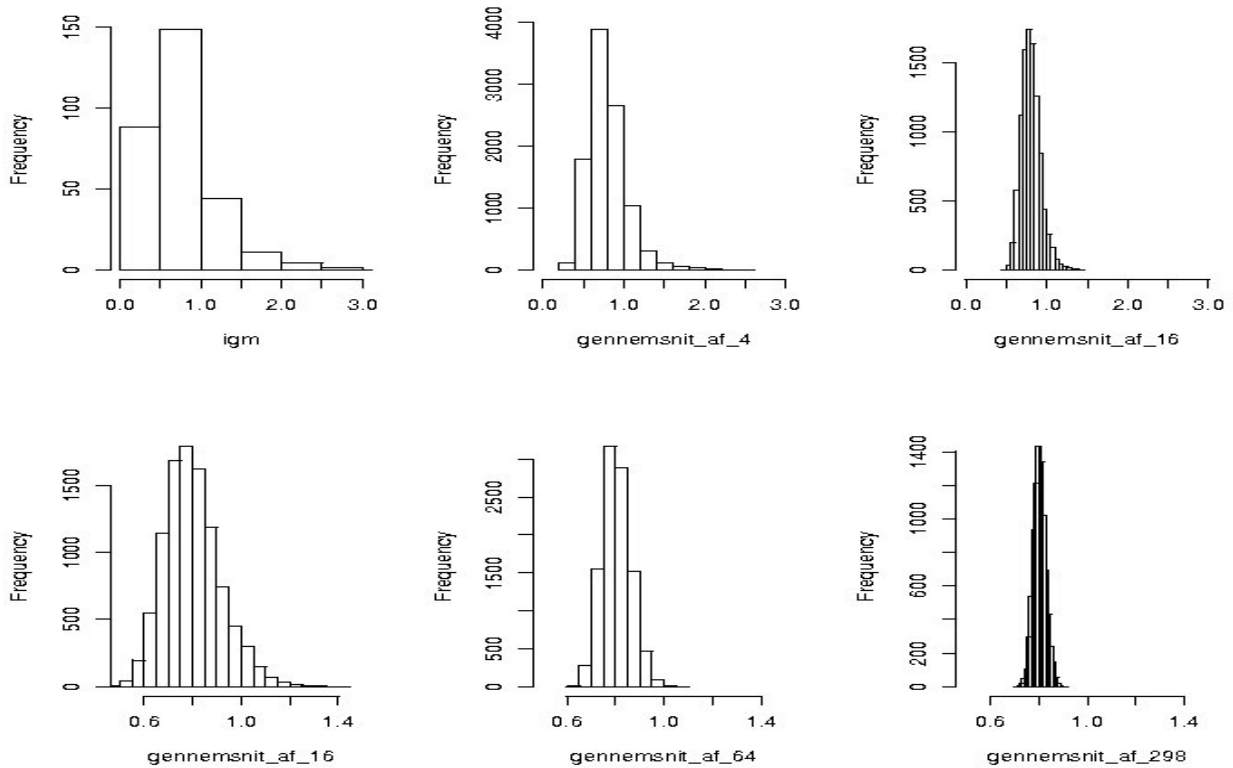
Fordelingen af et gennemsnit er pænere end fordelingen af de individuelle observationer (mere normalfordelt)

Jo flere observationer, der indgår i gennemsnittet

- des mere normalfordelt ser det ud
- des mindre spredning har dets fordeling (standard error of the mean, SEM), dvs. jo mere præcist fanger vi den sande middelværdi

Når man har mange observationer, gør det altså ikke så meget med fordelingsantagelsen - så længe man ser på gennemsnit!

Gennemsnit af flere og flere - immunoglobulin



Øverste linie:

Oprindelig fordeling, samt gennemsnit af 4 og 16

Nederste linie (i ny skala): gennemsnit af 16, 64 og 298

Hvis man *ikke* har mange observationer

- kan man *ikke* kontrollere normalfordelingsantagelsen
- og man bliver *ikke* reddet af *den centrale grænseværdisætning*

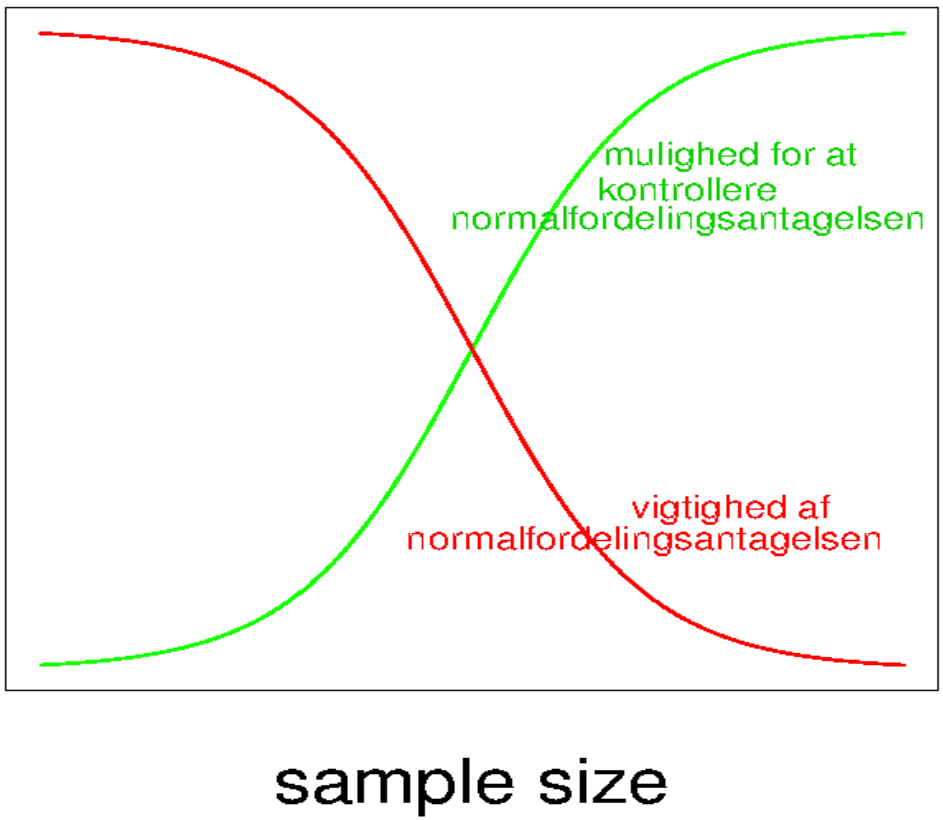
Man kan sige, at

- Muligheden for at kontrollere (forkaste) normalfordelingsantagelsen *vokser* med antallet af observationer
- Vigtigheden af normalfordelingsantagelsen *falder* med antallet af observationer

Så ved små studier kan man blive nødt til at benytte non-parametriske metoder

Kontrol af normalfordeling

Lidt af et dilemma:



Non-parametriske metoder - tests

Tests, der *ikke* bygger på en normalfordelingsantagelse

– men de er *ikke* forudsætningsfri

Ulemper

- tab af efficiens (sædvanligvis lille)
- uklar problemformulering
  - manglende model, og dermed ingen fortolkelige parametre
- **ofte ingen estimater!** – og ingen konfidensintervaller
- kan kun anvendes i simple problemstillinger
  – med mindre man har godt med computerkraft

Nonparametrisk one-sample test

af middelværdi 0 (parret two-sample test)

- Sign test, fortegnstest
  - udnytter kun observationernes fortegn, ikke deres størrelse
  - ikke særligt stærkt
  - invariant ved transformation
- Wilcoxon signed rank test
  - udnytter observationernes fortegn, kombineret med rangordenen af de numeriske værdier
  - stærkere end sign-testet
  - kræver at man kan tale om 'store' og 'små' forskelle
  - kan påvirkes af transformation

Men vi får hverken estimat, konfidensinterval
eller limits of agreement...

Nonparametriske parrede tests

I SAS kan disse kun foretages som tests af middelværdi 0 på differenserne:

```
proc univariate data=mf_sv;
  var dif;
run;
```

```
            Tests for Location: Mu0=0

   Test           -Statistic-    -----p Value------
    Student's t    t  0.156687    Pr > |t|    0.8771
   Sign           M       2.5    Pr >= |M|   0.3593
   Signed Rank    S         8    Pr >= |S|   0.7603
```

Disse giver kun en P-værdi, og
hverken *estimat*, *konfidensinterval* eller *limits of agreement*...

Forskellige programmer benytter lidt forskellige teststørrelser!
(og benytter approksimationer, som regel for $n > 25$)

APPENDIX

Programbidder svarende til diverse slides:

- Indlæsning af vitamin D datasæt, s. 78-80
- Tegninger vedrørende vitamin D, s. 81-82
- Udregning af summary statistics og fraktildiagram, s. 83-85
- Indlæsning af MF-SV data, s. 37
- Tegninger vedrørende MF-SV, s. 87-88
- Parret T-test, s. 90
- Non-parametrisk test, s. 91

Det oprindelige Vitamin D datasæt

De første 5 linier (4 observationer):

```
country;category;vitd;age;bmi;sunexp;vitdintake
1;1;22.400;11.888;19.254;2;7.188
1;1;37.000;12.441;17.567;3;1.186
1;1;12.900;13.025;17.700;3;1.480
1;1;13.600;13.501;16.953;3;1.612
```

- Kode til indlæsning ses på de næste sider
- Sværere indlæsning end normalt pga format-sætninger.
- Datasættet `vitamind.txt` ligger på hjemmesiden

SAS-kode til indlæsning

Slide 6

```
PROC FORMAT;
    VALUE categoryf
        1 = "Girl"
        2 = "Woman";
    VALUE sunf
        1 = "Avoid sun"
        2 = "Sometimes in sun"
        3 = "Prefer sun";
    VALUE countryf
        1 = "Denmark"
        2 = "Finland"
        4 = "Ireland"
        6 = "Poland";
RUN;
```

SAS-kode til indlæsning, II

```
Data vitamind;
INFILE
  "http://publicifsv.sund.ku.dk/~lts/basal/data/VitaminD.txt"
  URL DLM=";" FIRSTOBS=2;
INPUT country category vitd age bmi sunexp vitdintake;
    FORMAT category categoryf. ;
    FORMAT country countryf. ;
    FORMAT sunexp sunf. ;
RUN;

DATA irlwomen;
SET vitamind; WHERE country=4 and category=2;

log2vitd=log2(vitd);
run;

proc print data=irlwomen;
var country category vitd age bmi sunexp vitdintake;
run;
```

Printet ses s. 6

SAS-kodning af histogram og Box-plot

Slides 9 og 11

```
proc sgplot data=irlwomen;
    histogram vitd;
    density vitd;
run;

proc sgplot data=women;
    vbox vitd / category=country;
run;
```

eller med "gammeldags kodning":

```
proc univariate normal data=irlwomen;
  var vitd;
  histogram / cfill=yellow height=3 normal;
run;

proc boxplot data=women;
  plot vitd*country /
     boxstyle=schematic cboxfill=yellow height=4;
run;
```

SAS-kodning af Scatter plot med linier

Slide 12

```
proc sgplot data=women;
    reg X=bmi Y=vitd / group=country;
run;
```

eller med "gammeldags" kodning:

(som man kan modificere på forskellig vis, men som tegner linierne helt ud til kanten)

```
proc gplot normal data=women;
  plot vitd*bmi=country
  / 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);
symbol1 v=circle i=rl c=black h=2 l=1 w=2;
symbol2 v=star i=rl c=blue h=2 l=1 w=2;
symbol3 v=triangle i=rl c=red h=2 l=1 w=2;
symbol4 v=dot i=rl c=green h=2 l=1 w=2;
run;
```

Udregning af summary statistics

Slide 17

```
proc means mean median Q1 Q3 stddev
           data=vitamind maxdec=2;
  class country;
  var vitd;
run;
```

Udregning af specielle fraktiler

Slide 21

```
proc univariate data=irlwomen;
    var vitd;
    output out=regn pctlpre=P_ pctlpts=2.5,97.5;
run;

proc print data=regn;
run;
```

som giver outputtet

```
Obs    P_2_5    P_97_5
 1       18      89.1
```


Fraktildiagram

Slide 24

```
proc univariate data=irlwomen;
  var vitd;
  qqplot / height=4 square
           normal(mu=EST sigma=EST color=red w=3 l=1);
run;
```

Datafilen vedr. MF og SV

Slide 37

Data-filen `'mf_sv.txt'`,
(beliggende på hjemmesiden)
er en tekstfil med 2 kolonner a 21 linier, en for hver person, med variabelnavne i første linie.

Vi indlæser og definerer derefter to nye variable:

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

infile navn firstobs=2;
input mf sv;

/* definition af nye variable */
dif=mf-sv;
average=(mf+sv)/2;
run;
```

Scatter plot og Bland-Altman plot

Slides 38 og 41

```
/* Scatter plot */
proc sgplot data=mf_sv;
    scatter X=sv Y=mf / markerattrs=(color=blue);
  lineparm x=40 y=40 slope=1 / lineattrs=(color=red);
run;

/* Bland-Altman plot */
proc sgplot data=mf_sv;
  scatter X=average Y=dif / markerattrs=(color=blue);
  lineparm x=40 y=0 slope=0 / lineattrs=(color=red);
run;
```

Box-plot og Spaghetti-plot

Slide 39

```
data lang;
set mf_sv;

flow=mf;  metode='mf';  output;
flow=sv;  metode='sv';  output;
run;

/* Forkert Boxplot */
proc sgplot data=lang;
    vbox flow / category=metode;
run;

/* Korrekt Spaghetti-plot */
proc sgplot data=lang;
    series X=metode Y=flow / group=person;
run;
```

Estimat med konfidensgrænser

Slide 51

```
proc means N mean stderr clm;
 var dif;
run;
```

Parret T-test

af MV vs. SV, med differenser `dif`

Slide 59 og 60

```
proc ttest data=mf_sv;
  paired mf*sv;
run;
```

eller

```
proc ttest data=mf_sv;
  var dif;
run;
```

Parret non-parametrisk test

af MV vs. SV, med differenser `dif`

Slide 76

```
proc univariate data=mf_sv;
  var dif;
run;
```