Basale statistiske begreber

Parrede sammenligninger

Susanne Rosthøj
Section of Biostatistics
sro@sund.ku.dk

Indhold

- Deskriptiv statistik og illustration af data
- Vurdering af normalfordeling
- Normalområder (referenceområder)
- Konfidensintervaller
- Sammenligning af to målemetoder

Vitamin D

Problemstilling

Videnskabeligt spørgsmål

- Har 'folk' et tilstrækkeligt højt niveau af vitamin D?
- Hvis ikke, kan vi så forstå hvorfor?
  - og kan vi gøre noget ved det?

Vi ser på et studie af ældre kvinder fra DK, EI, PL og SF

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?

- **Outcome:** serum Vitamin D
- **Kovariater / forklarende variable:** land, spisevaner, soleksponering, fedme, rygning, alkohol, ...

Skriv en protokol!

- Man får tænkt sig om på forhånd
- Det er en nødvendig del af dokumentation (etisk komite, ansøgning om midler, anmeldelse af trial, ...)
- Det tjener som “ekstra hukommelse”
- I forbindelse med den statistiske analyse dokumenterer det, hvad der var den oprindelige strategi og hvad der bør betegnes som tilfældige fund

Data

Studie af ældre kvinder fra DK, PO, FI og SF.

```
  country vitd    age    bmi           sunexp vitdintake
1      DK 28.5 73.005 26.177        Avoid sun      1.182
2      DK 32.9 72.929 25.155        Avoid sun      1.052
3      DK 60.3 72.597 23.081        Avoid sun     18.710
4      DK 50.5 71.512 25.200        Avoid sun      8.417
5      DK 15.0 71.847 25.310        Avoid sun      1.273
6      DK 47.8 73.764 23.904 Sometimes in sun      7.135
```

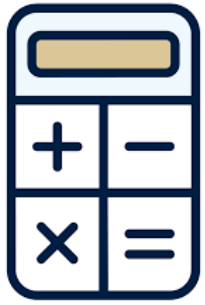
Indlæsning af data: R / SAS / SPSS

Rækkerne kaldes **observationer** (typisk 1 pr. person)

Søjlerne kaldes **variable**

Typer af variable

- **Kvantitative** (numeriske) variable
  - Vitamin D koncentration (`vitd`, nmol/l)
  - Alder (`age`)
  - Body Mass Index (`bmi`)



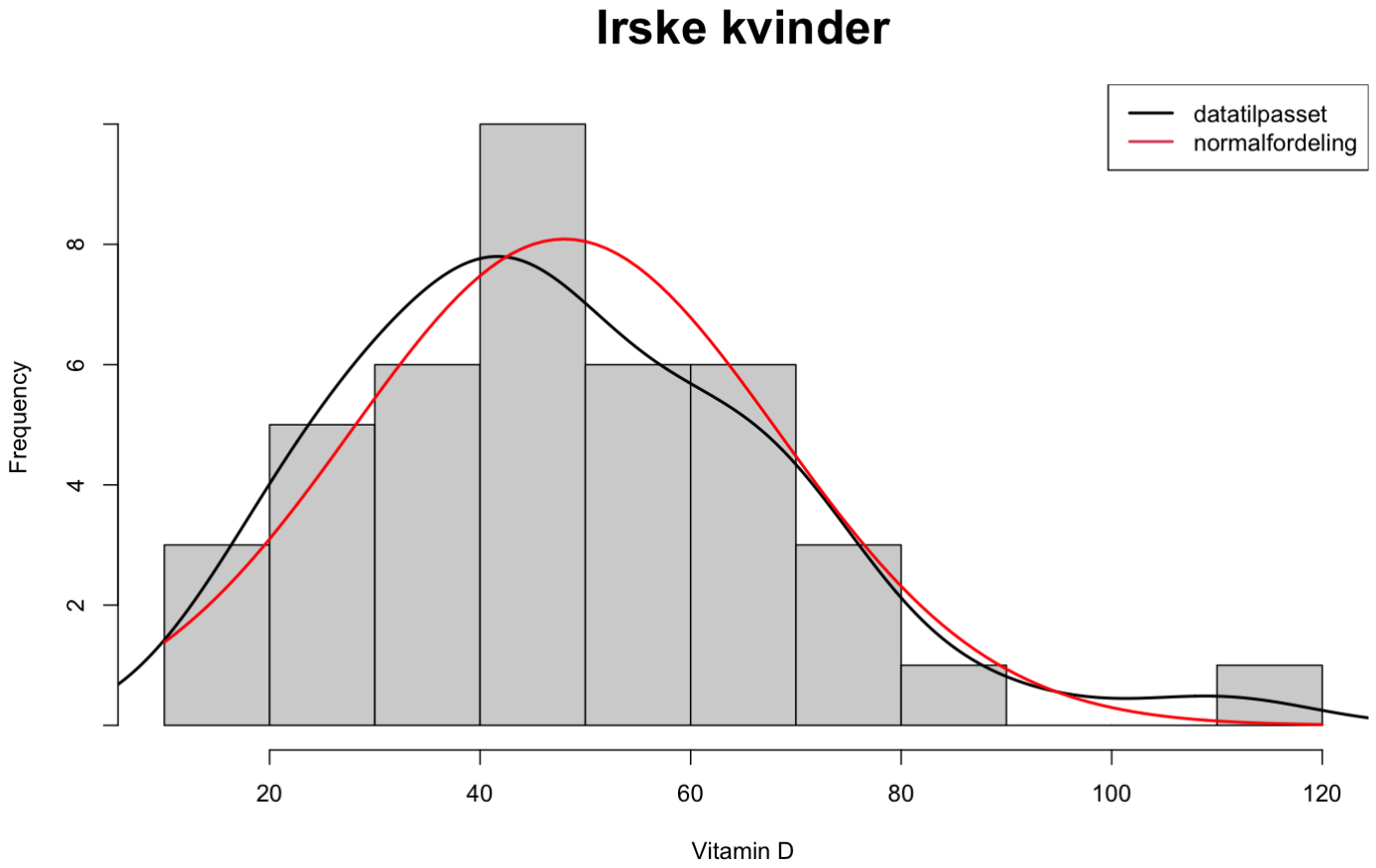
- **Kategoriske** (kvalitative) variable
  (class, factor, nominal)
  - Personens hjemland (`country`)
  - Personens solvaner (`sunexp`, **ordinal**)

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

Tegn

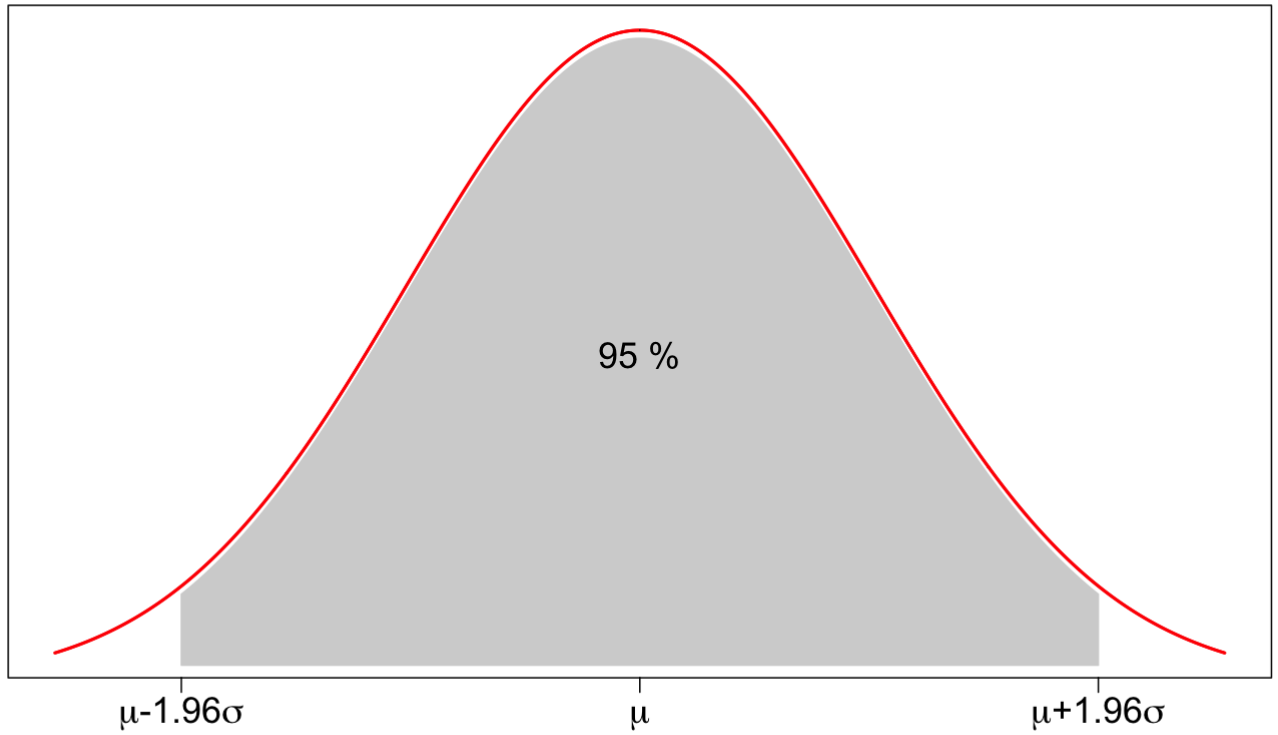
Tegn I: Histogram



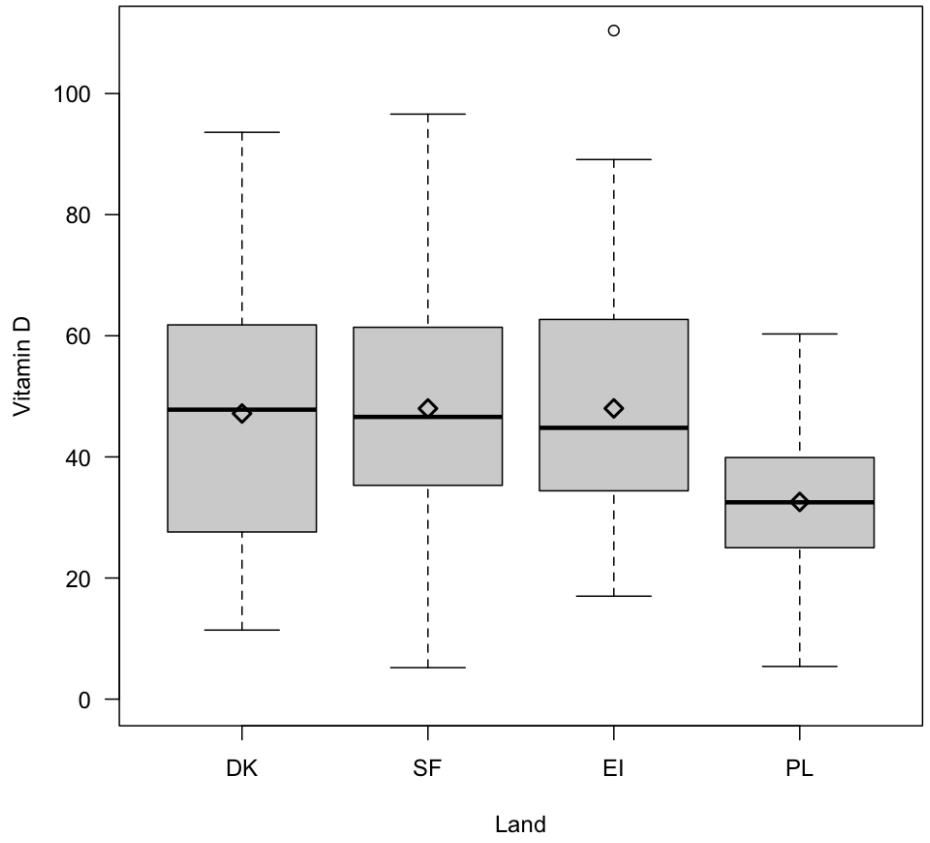
R / SAS / SPSS

Normalfordelingen $\mathcal{N}(\mu, \sigma^2)$

Middelværdi μ, spredning σ



Tegn II: boxplot



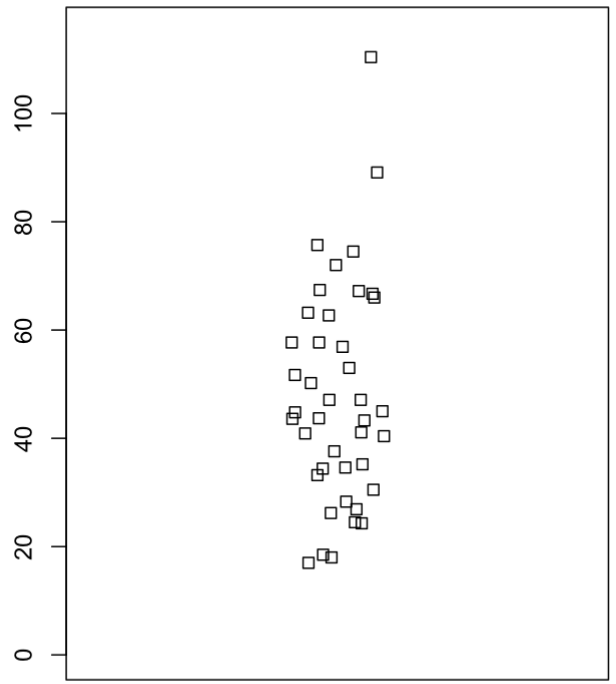
God til sammenligninger

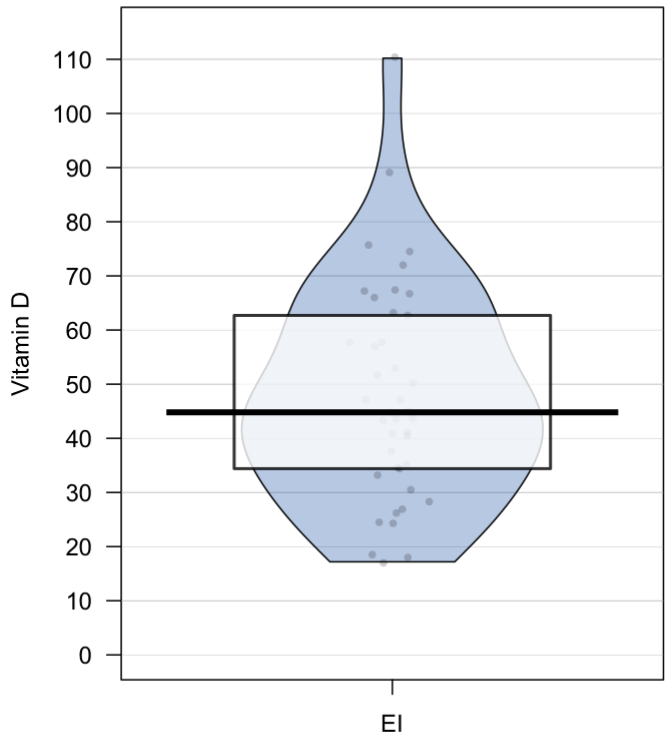
R / SAS / SPSS

Mere om boxplots

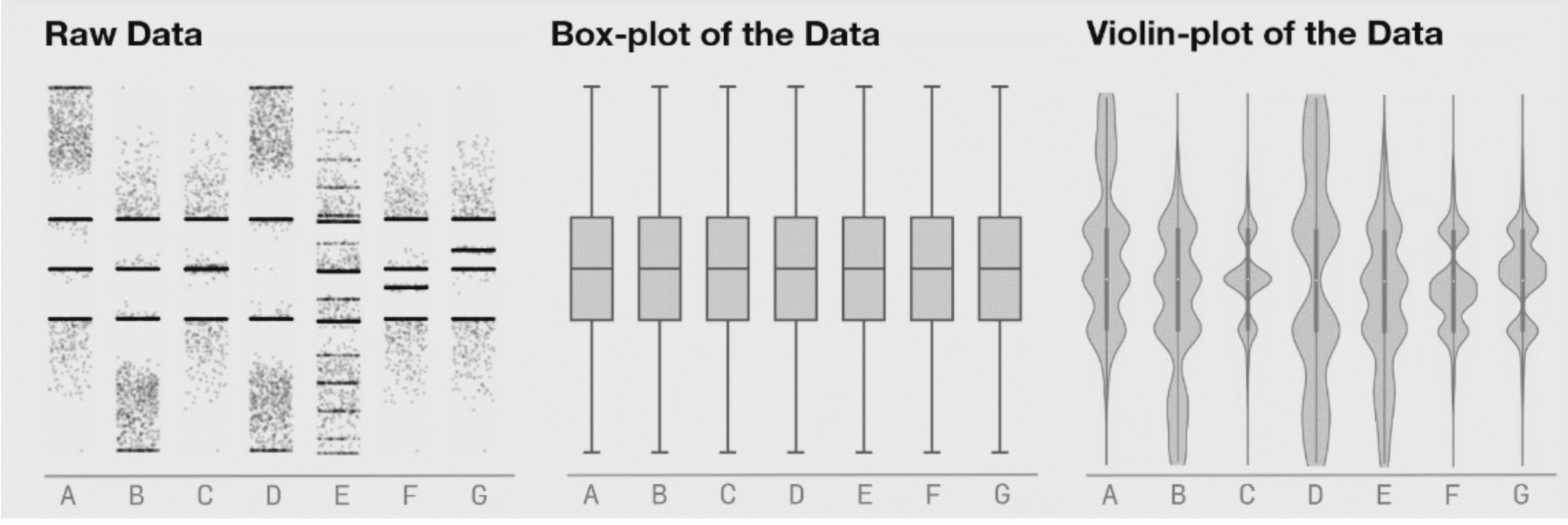
Tegn III

Stripchart og violinplot





Er det tilstrækkeligt at tegne boxplot?



Link til animation

Regn

Regn I: Location / centrum

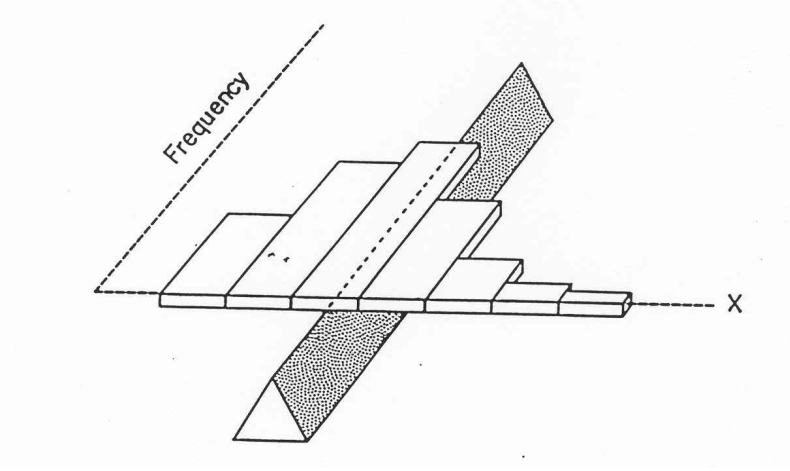
Observationer $Y_1, \ldots, Y_n$.

- **Gennemsnit:** $\bar{Y} = \frac{1}{n}(Y_1 + \cdots + Y_n)$
- **Median:** midterste observation

I *symmetriske* fordelinger vil gennemsnit og median være ens (pånær tilfældigheder)

I *skæve* fordelinger vil de ikke være ens:
Typisk er der hale mod de høje værdier (og 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

Median 7 dage. Gennemsnit 22 dage.

Repræsentativt for hvad?

Regn II: Variation

- Varians: $s^2 = \frac{1}{n-1}\sum(Y_i - \bar{Y})^2$
  - **Spredning** = Standardafvigelse = Standard Deviation = $\sqrt{}$varians = s = SD.

- Fraktiler:
  - **Interquartiles** Q1 (25%) og Q3 (75%) (InterQuartile Range Q3-Q1 (IQR))

Summary statistics for Vitamin D

```
       mean median       SD     Q1   Q3  min   max
DK 47.16604   47.8 22.78292 27.600 61.8 11.4  93.6
SF 47.99074   46.6 18.72471 35.525 60.9  5.2  96.6
EI 48.00732   44.8 20.22212 34.400 62.7 17.0 110.4
PL 32.56154   32.5 12.46448 25.000 39.9  5.4  60.3
```

Median og gennemsnit er nogenlunde ens (~ symmetri i boxplot).

*Det betyder **ikke** at der er tale om en normalfordeling*

R / SAS / SPSS

Referenceområder

Referenceområde baseret på fraktiler

Reference- eller **normal** område:

De centrale 95% af observationerne

- Nedre grænse: 2.5% fraktil
- Øvre grænse: 97.5% fraktil

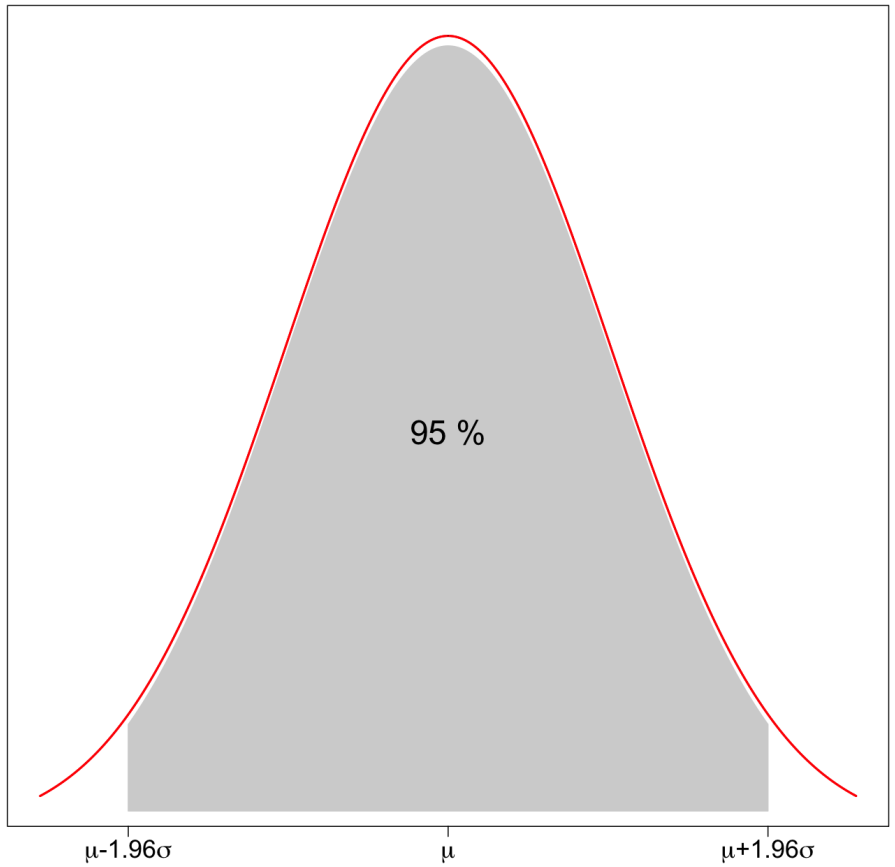
Irske kvinder (*n*=41):

```
 [1]  17.0  18.0  18.5  24.3  24.5  26.2  26.9  28.3  30.5  33.2  34.4  34.6
[13]  35.2  37.6  40.4  40.9  41.1  43.3  43.6  43.7  44.8  45.0  47.1  47.1
[25]  50.2  51.7  53.0  56.9  57.7  57.7  62.7  63.2  66.0  66.7  67.2  67.4
[37]  72.0  74.5  75.7  89.1 110.4
```

Normalområde fra 18 til 89.1 nmol/l.

R / SAS / SPSS (NB andre fraktiler i SPSS (17.1-109.3))

Referenceområde i normalfordeling



Fraktiler:

- 2.5% fraktil: $\mu - 1.96\sigma$
- 97.5% fraktil: $\mu + 1.96\sigma$

Referenceområde baseret på normalfordeling

Normalområdet udregnes som

$$\bar{y} \pm \text{cirka } 2 \times s$$

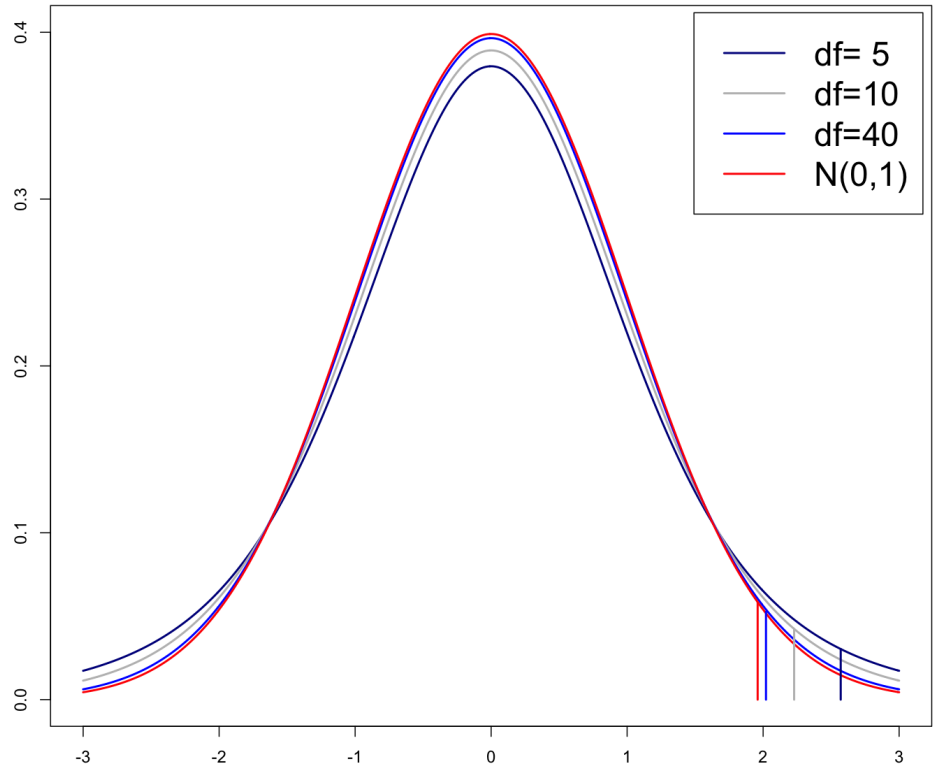
Cirka 2 er $\sqrt{1 + \frac{1}{n}} t_{97.5\%}(n-1)$

- her $\sqrt{1 + \frac{1}{41}} \times 2.02 = 2.07$

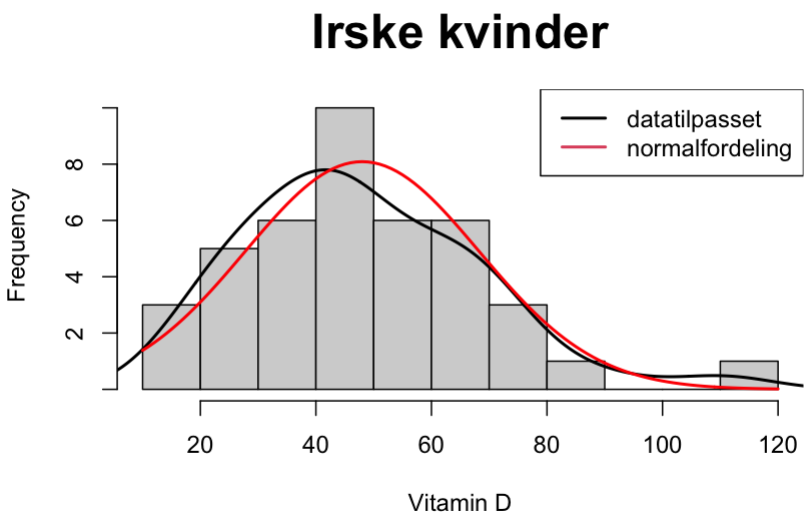
For irske kvinder:

$48.0 \pm 2.07 \times 20.2 = (6.1, 89.9)$

R / SAS / SPSS



Er kvinderne oppe på det ønskede niveau på 25 nmol/l?



Referenceområder

- Normalfordelings: (7.6,88.4)
- Fraktil: (18,89.1)

Antal kvinder med vitamin D under 25:
5 svarende til 12%.

R / SAS / SPSS

Beregning af referenceområde

Store datasæt

- Brug fraktiler

Mellemstore datasæt:

- Brug en rimelig fordelingsantagelse, typisk normalfordelingen (evt. efter transformation)
  Her er normalfordelingsantagelsen vigtig

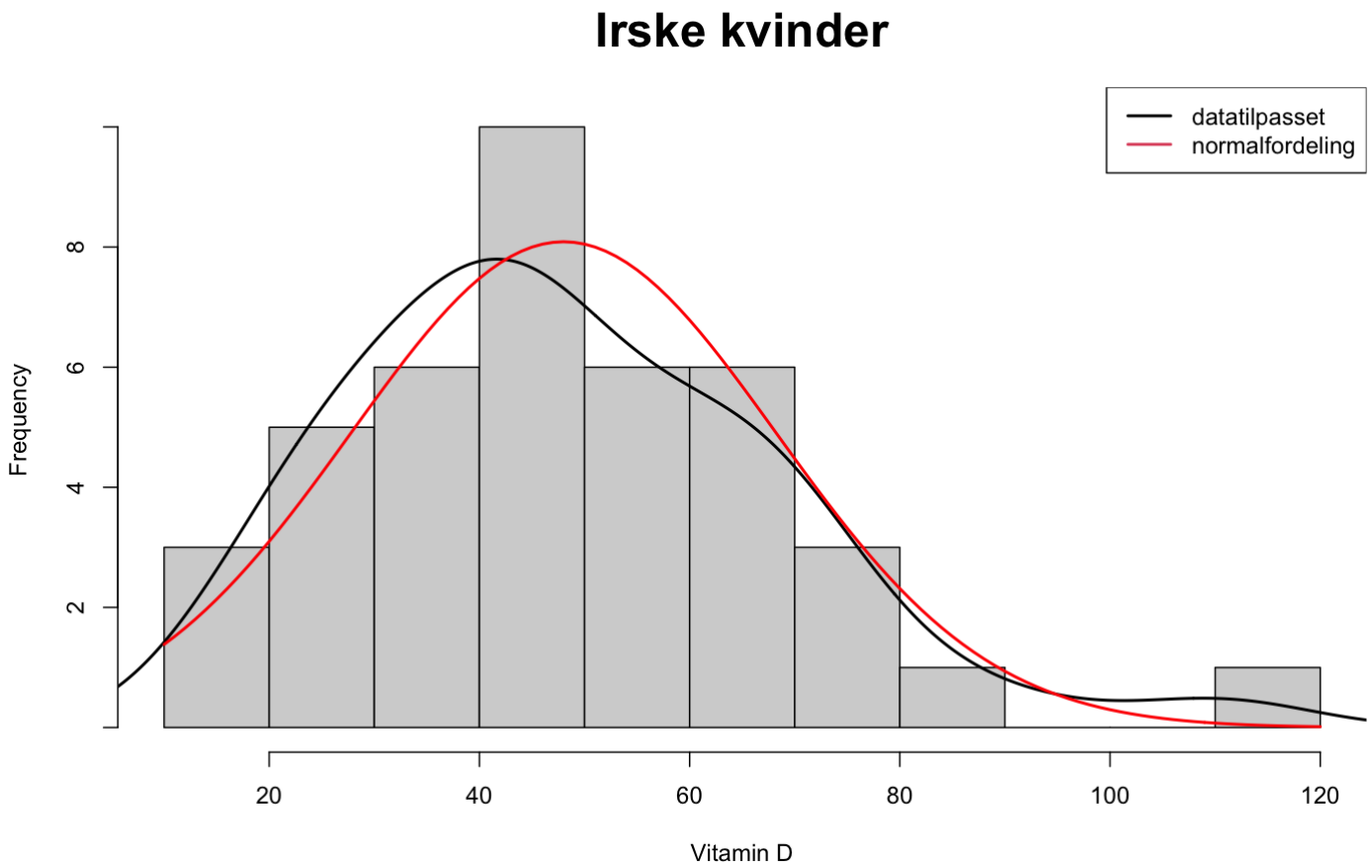
Små datasæt

- Lad være!

Vurdering af normalfordeling

Histogram

Test for normalfordeling - aldrig!



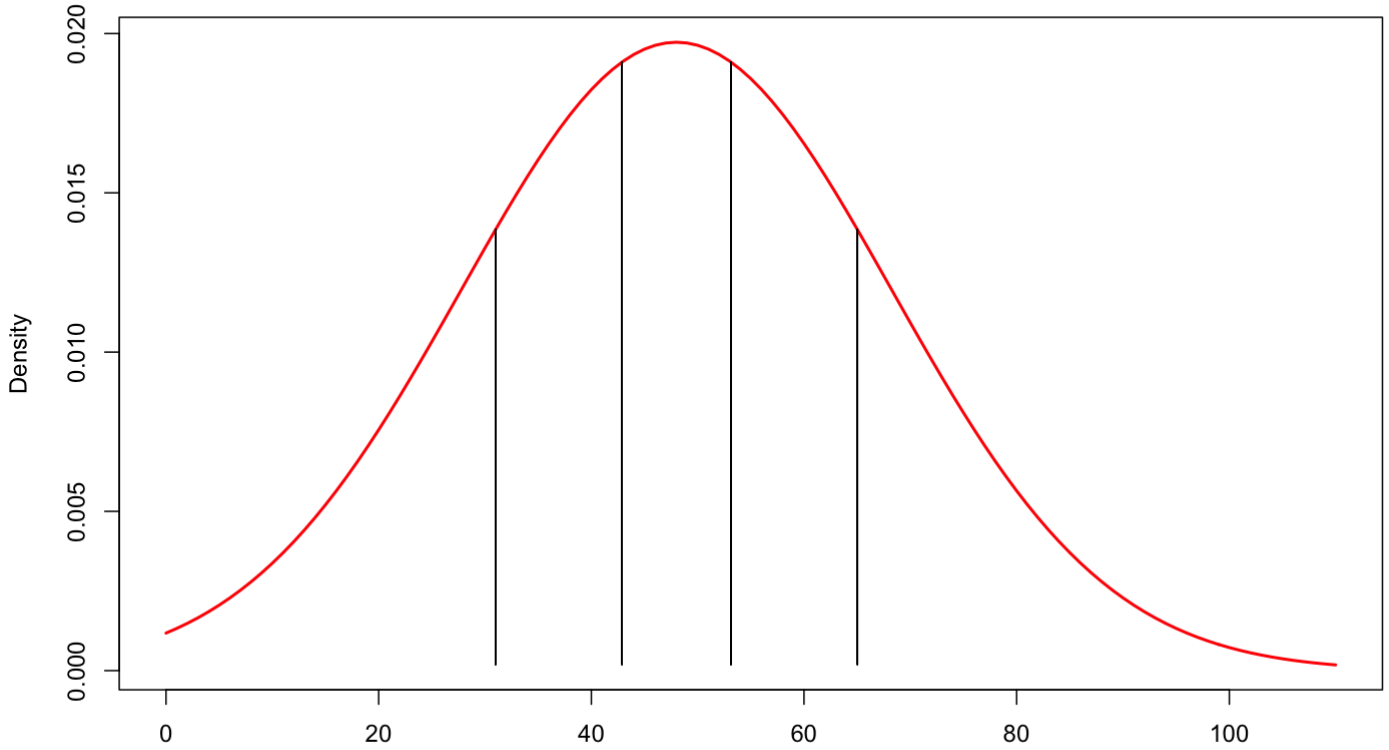
Vurdering af normalfordeling I

Hvordan ser normalfordelingen $\mathcal{N}(\mu = 48.01, \sigma^2 = 20.22^{\,2})$ ud?

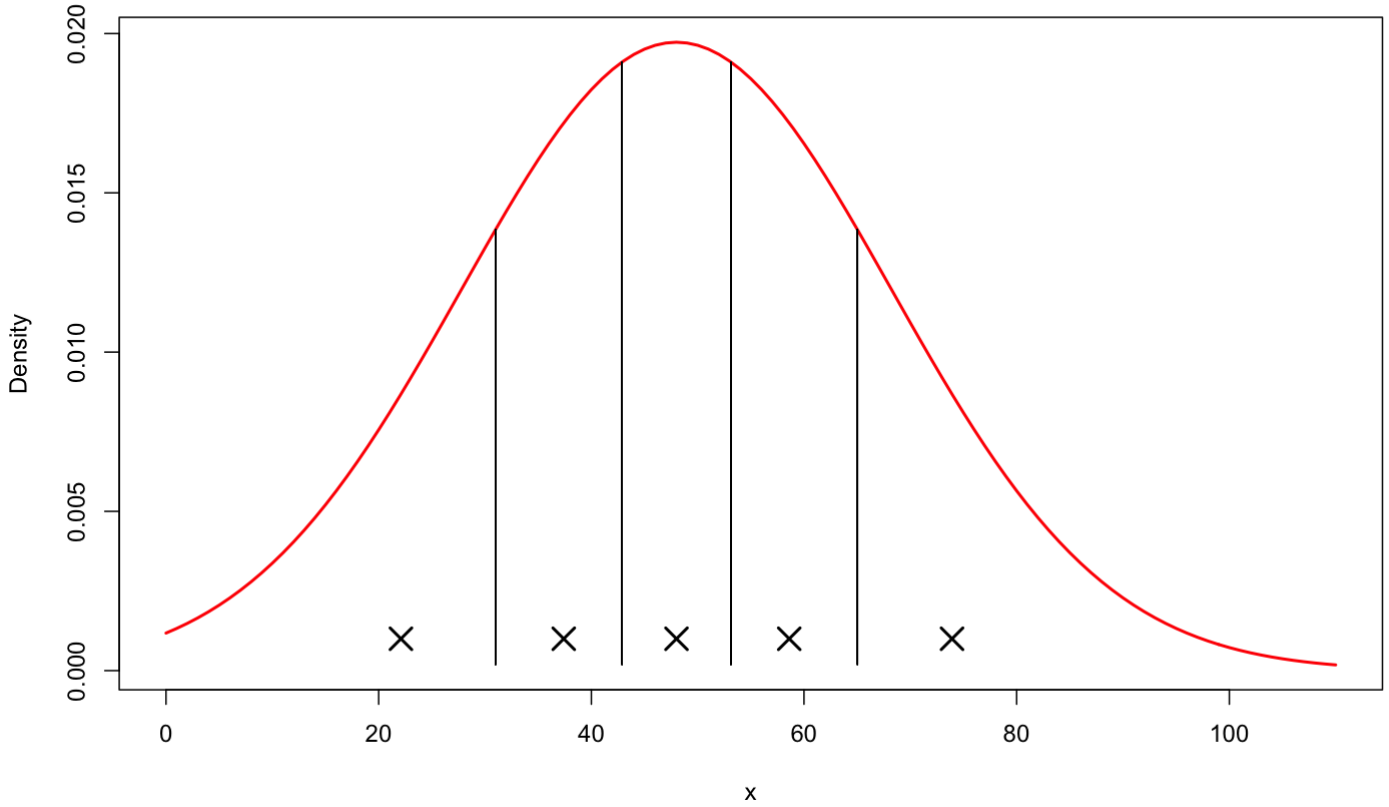
Med 5 (sorterede) observationer forventer vi observation ...

- nr 1 mellem $-\infty$ og 1/5-fraktilen
- nr 2 mellem 1/5 og 2/5-fraktilen
- ...
- nr 5 mellem 4/5-fraktilen og $+\infty$ (=5/5-fraktilen)

Vurdering af normalfordeling II

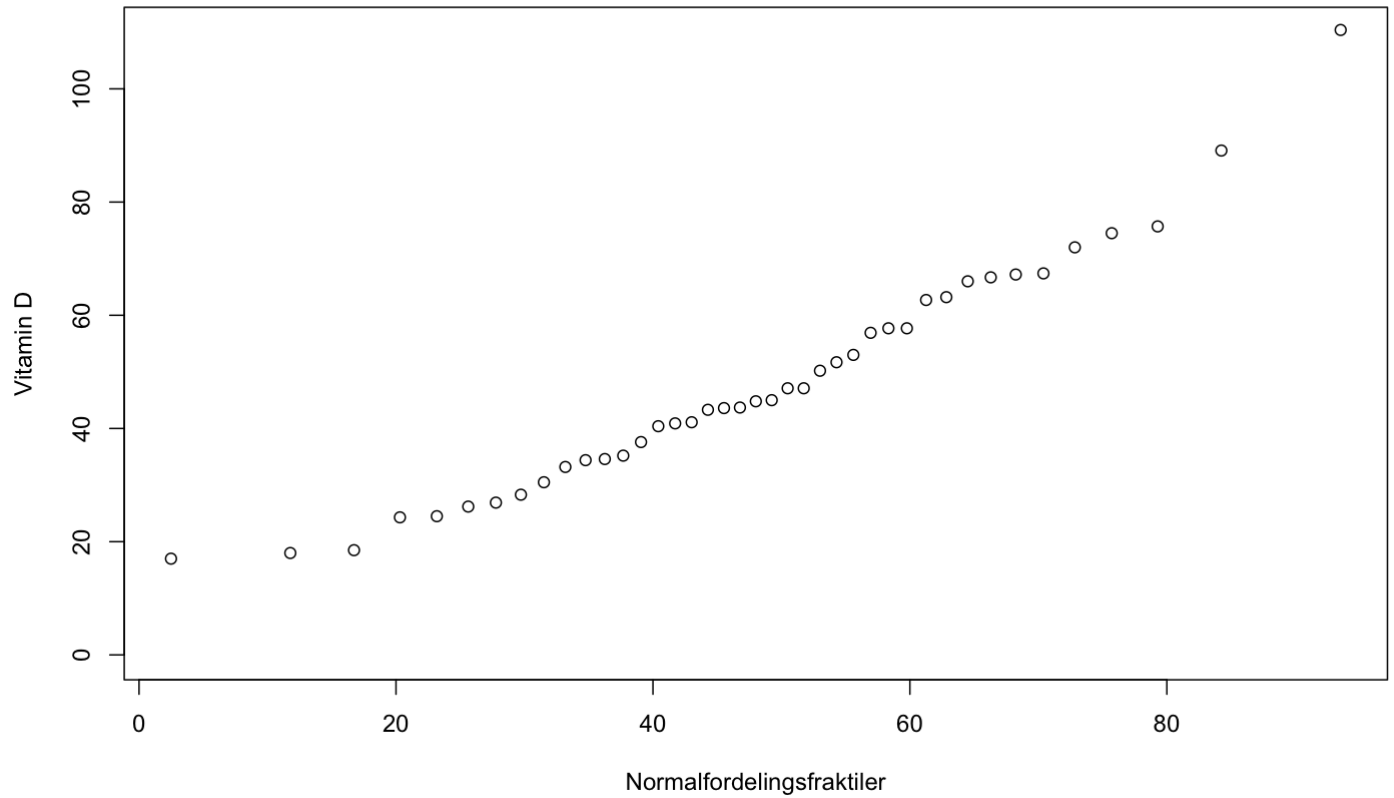


Vurdering af normalfordeling II



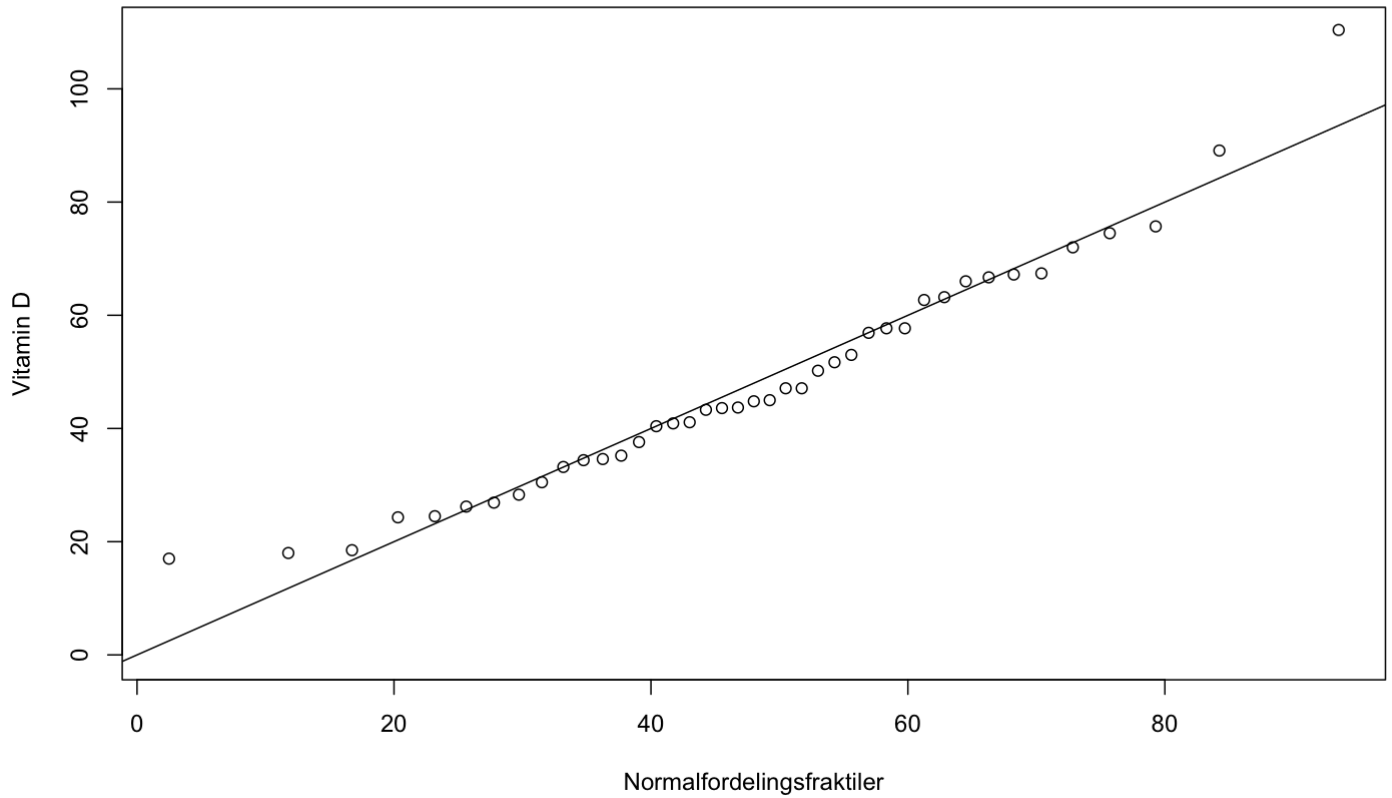
QQ-plot

Quantile-Quantile plot (fraktildiagram):



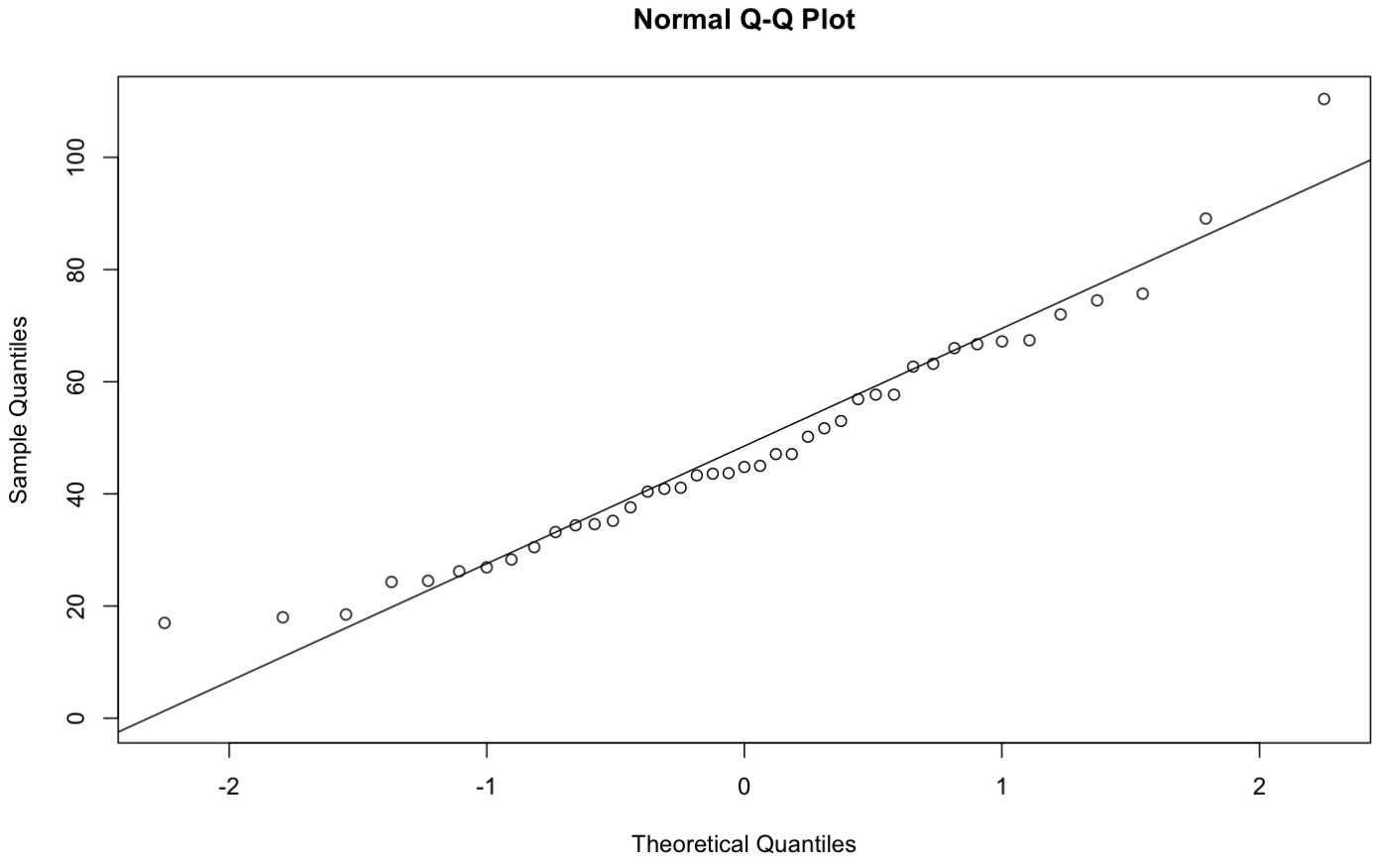
QQ-plot

Ret linje? Se eksempler her



QQ-plot

Quantile-Quantile plot med normerede fraktiler: R / SAS / SPSS



Højreskæve fordelinger

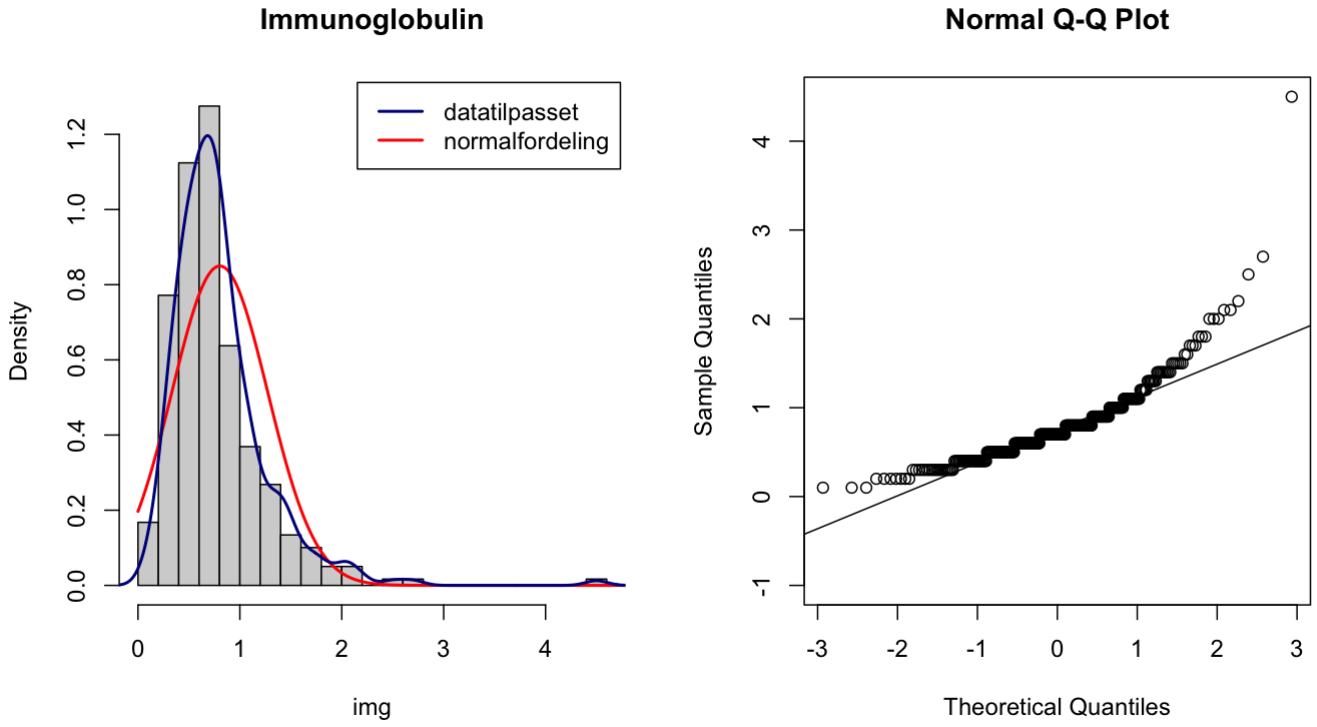
Logaritmen

Nyt eksempel - immunoglobulin

Data: R / SAS / SPSS

```
      mean median         SD  Q1 Q3 min max
 0.8030201    0.7 0.4694982 0.5  1 0.1 4.5
```

Højreskæv fordeling



I begge haler er observationerne for store - **hængekøjeform** i QQ-plot

Højreskæv fordeling

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

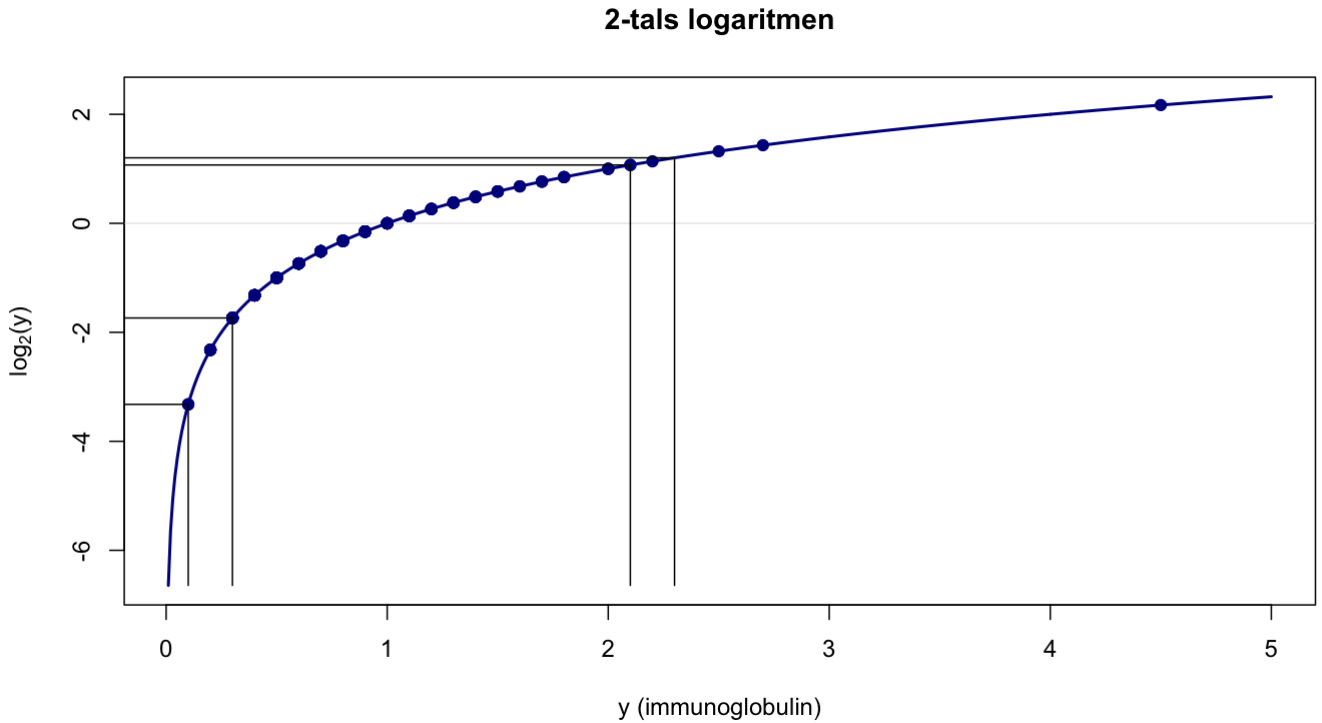
Løsning:

Transformér med en logaritme

- ligegyldig hvilken: naturlig, 10-tals, 2-tals, 1.1-tals, ...

$\log_2(y) = z$: "Hvilket tal z skal jeg opløfte 2 i for at få y?"

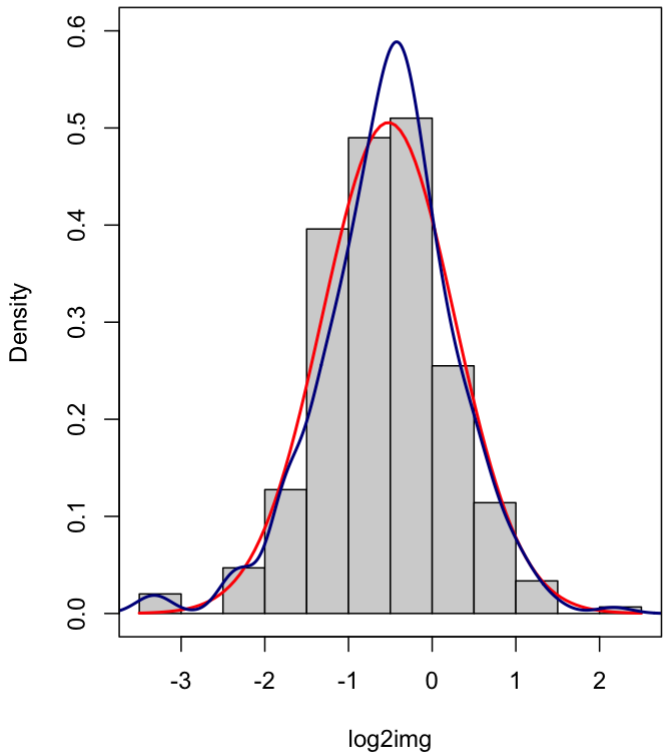
Logaritmen

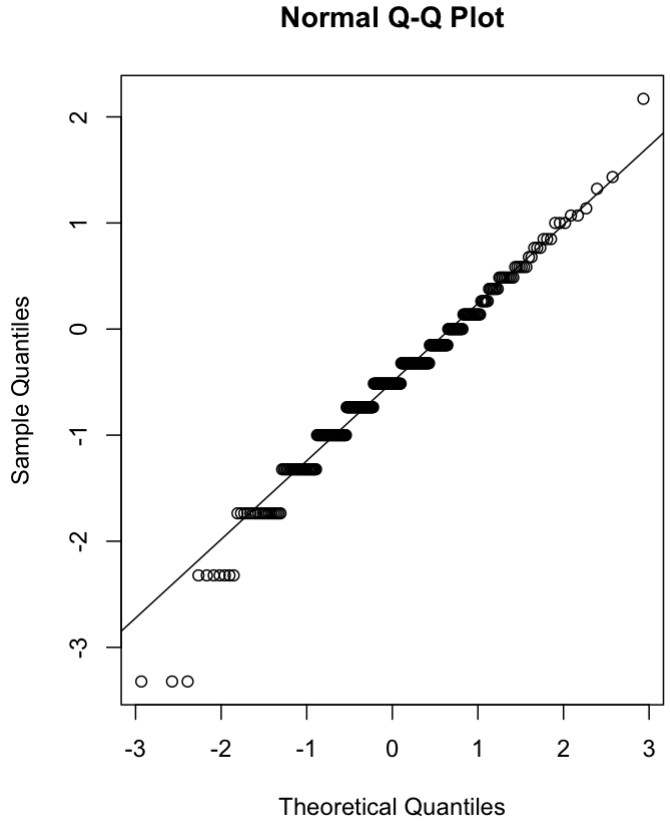


Store observationer 'trykkes sammen', små observationer 'spredes'.

Logaritmetransformeret immunoglobin

Beregn ny variabel log2img = $\log_2(\text{img})$ i datasættet





Anti-log

Når man har regnet "noget" færdigt på log-skala transformeres tilbage med den **samme anti-logaritme**, dvs.

- e^{noget} (eksponentialfunktionen, e=2.72=`exp(1)` Eulers konstant,),
- 10^{noget}
- 2^{noget}

Dvs med $\log_2(y) = z$ er invers funktion (antilog) : $y = 2^z$.

Normalområde immunoglobulin

Data	gennemsnit	median	spredning	referenceområde
utransformeret	0.803		0.469	-0.135 til 1.741
log2-transformeret	-0.524		0.789	-2.102 til 1.054
tilbagetransformeret		0.695	-	0.233 til 2.076
empiriske fraktiler		0.700	-	0.2 til 2.0

Tilbagetransfomationen:

Referenceområde på $\log_2$-skala: (-2.102, 1.054)

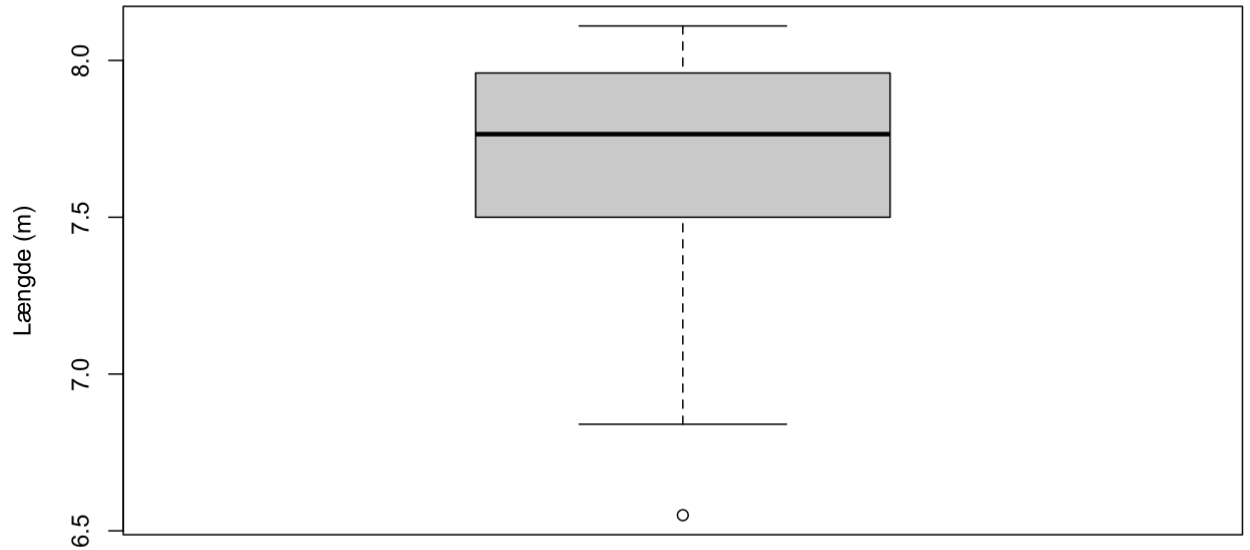
Tilbagetransformeret: $(2^{-2.102}, 2^{1.054})$ =(0.23, 2.08)

Lad være med at tilbagetransformere spredningerne!

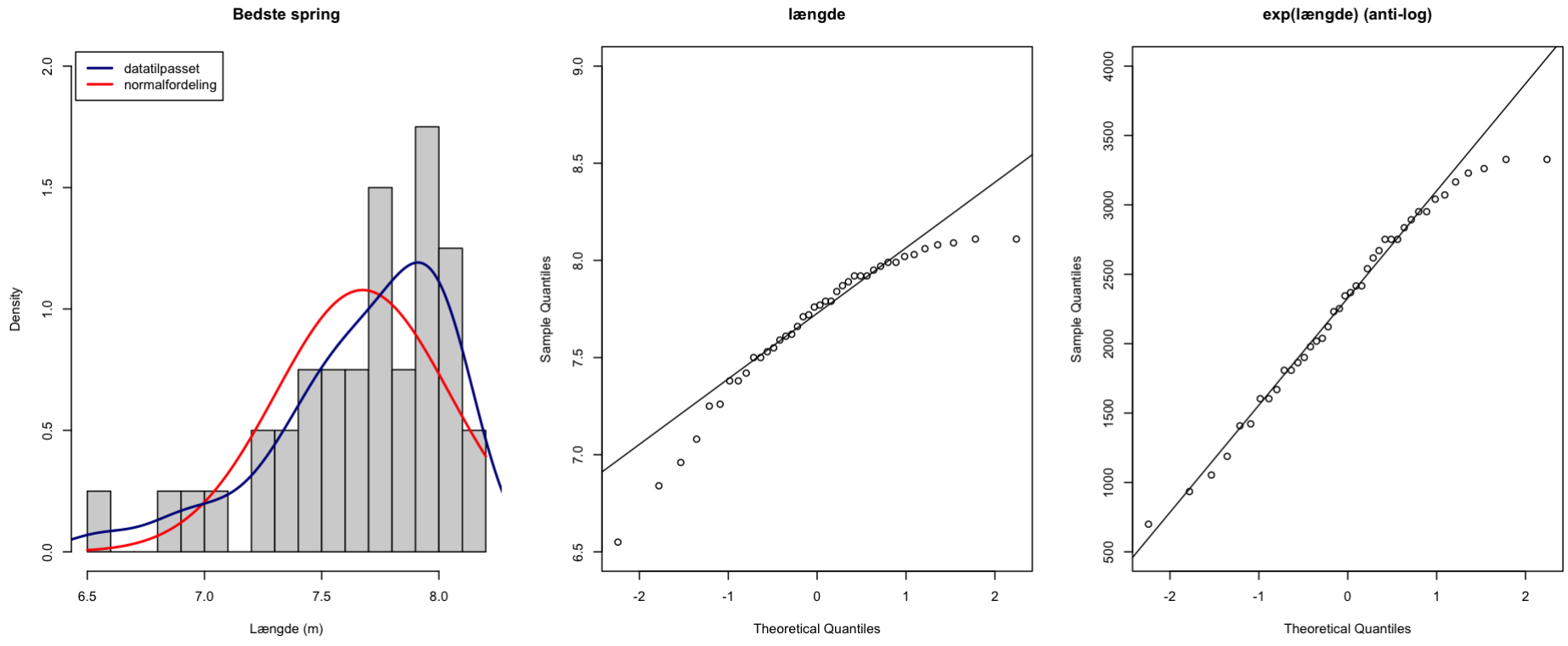
Nyt eksempel - længdespring

40 mandlige atleter, kvalifikation OL 2012, bedste spring af 3. Data: R / SAS / SPSS

mean	median	SD	Q1	Q3	min	max
7.6745	7.765	0.3700031	7.5	7.955	6.55	8.11



Venstreskæv fordeling



I begge haler er observationerne for små. Ingen "mirakel"-transformation ...

Fordeling af gennemsnit

Normalfordelte data

Vi observerer n observationer $Y_1, \ldots, Y_n$ fra $\mathcal{N}(\mu, \sigma^2)$

Gennemsnittet $\bar{Y}$ er normalfordelt med:

$$\text{mean}(\bar{Y}) = \mu.$$

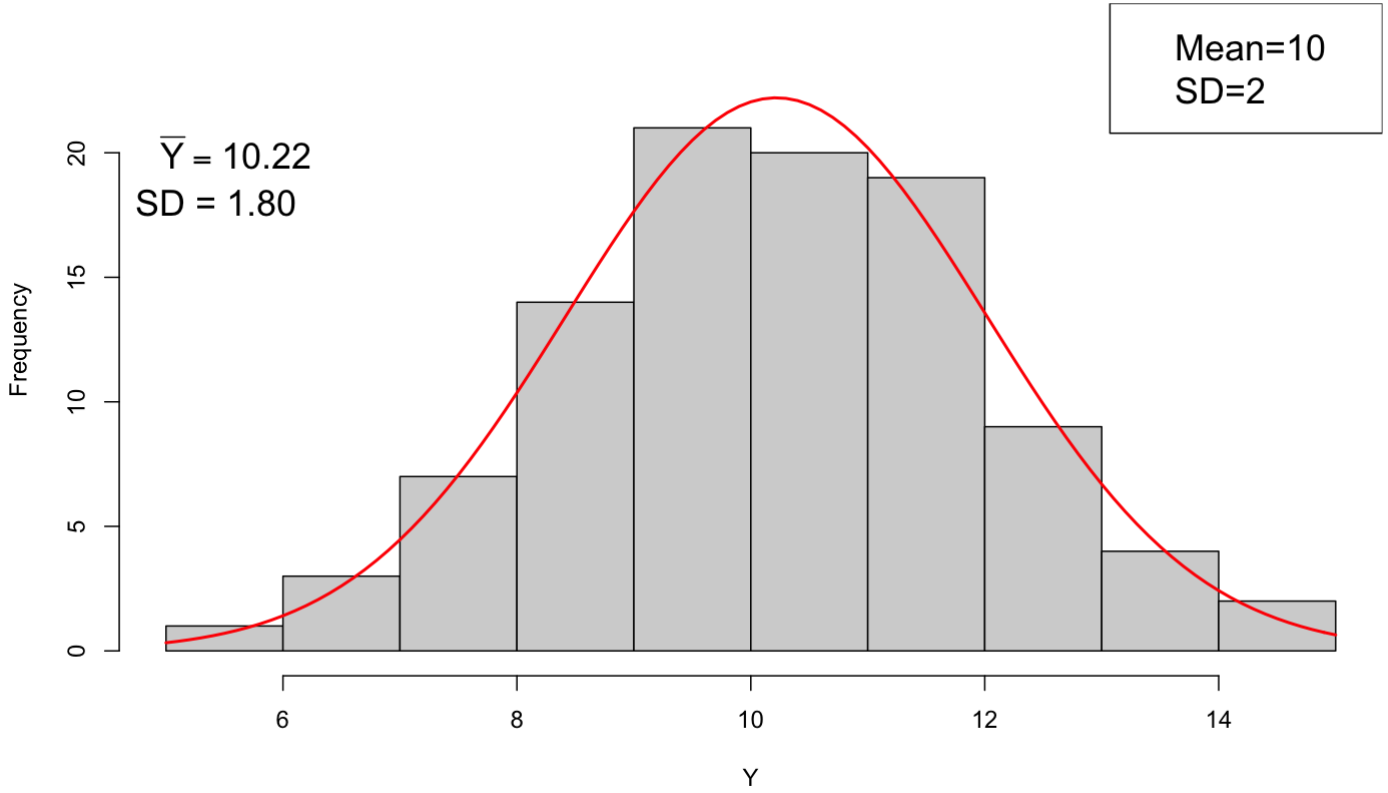
$$\text{SD}(\bar{Y}) = \frac{\sigma}{\sqrt{n}}$$

Denne SD kaldes også **standard error of the mean** (SE eller SEM).

Gennemsnittet har altså en fordeling.

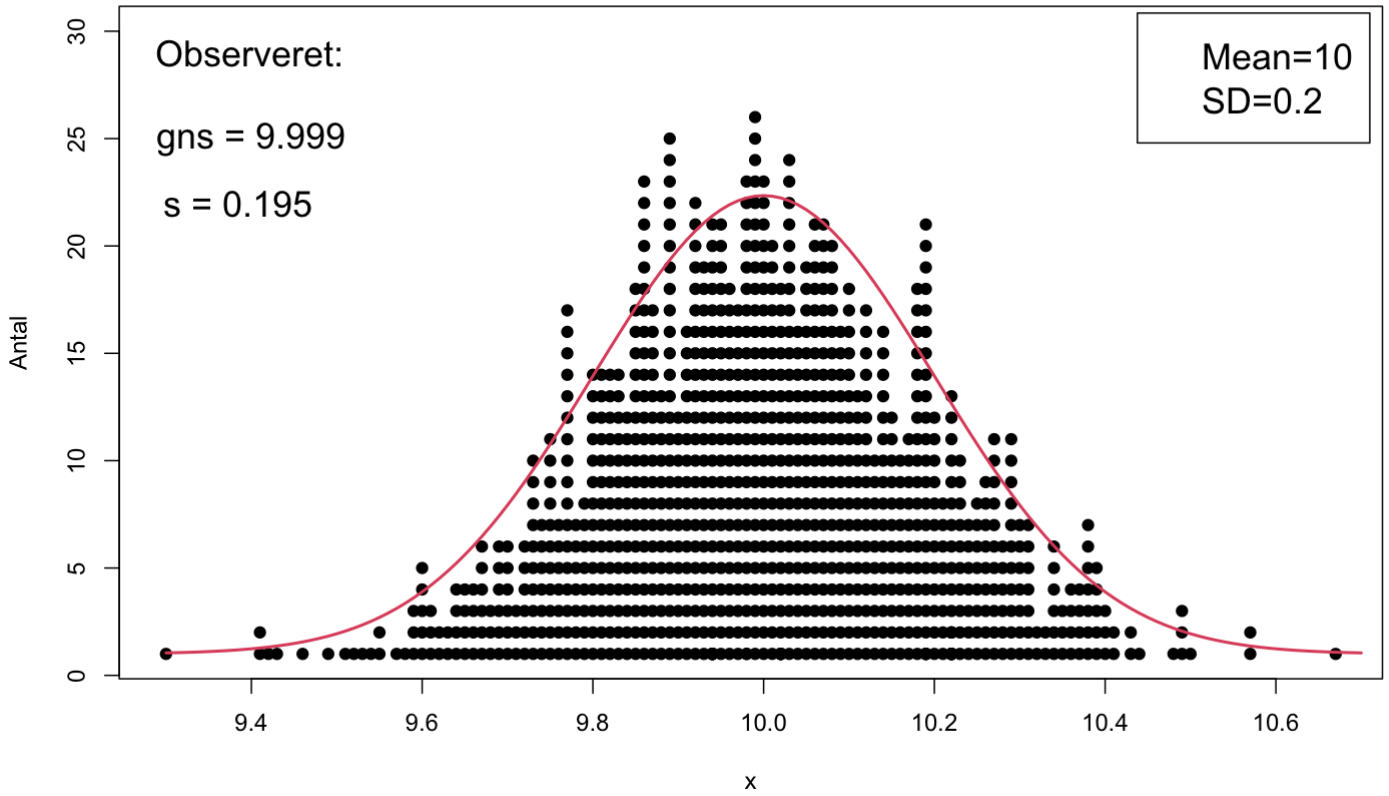
Én normalfordeling

Simulerede data, $n = 100$.



Fordelingen af mange gennemsnit

1000 simulerede datasæt, hver med $n = 100$.



Ikke-normalfordelte data

Vi observerer n observationer $Y_1, \ldots, Y_n$ fra samme fordeling med middelværdi μ og spredning σ

Den centrale grænseværdisætning:

Gennemsnittet $\bar{Y}$ er **approksimativt** normalfordelt med:

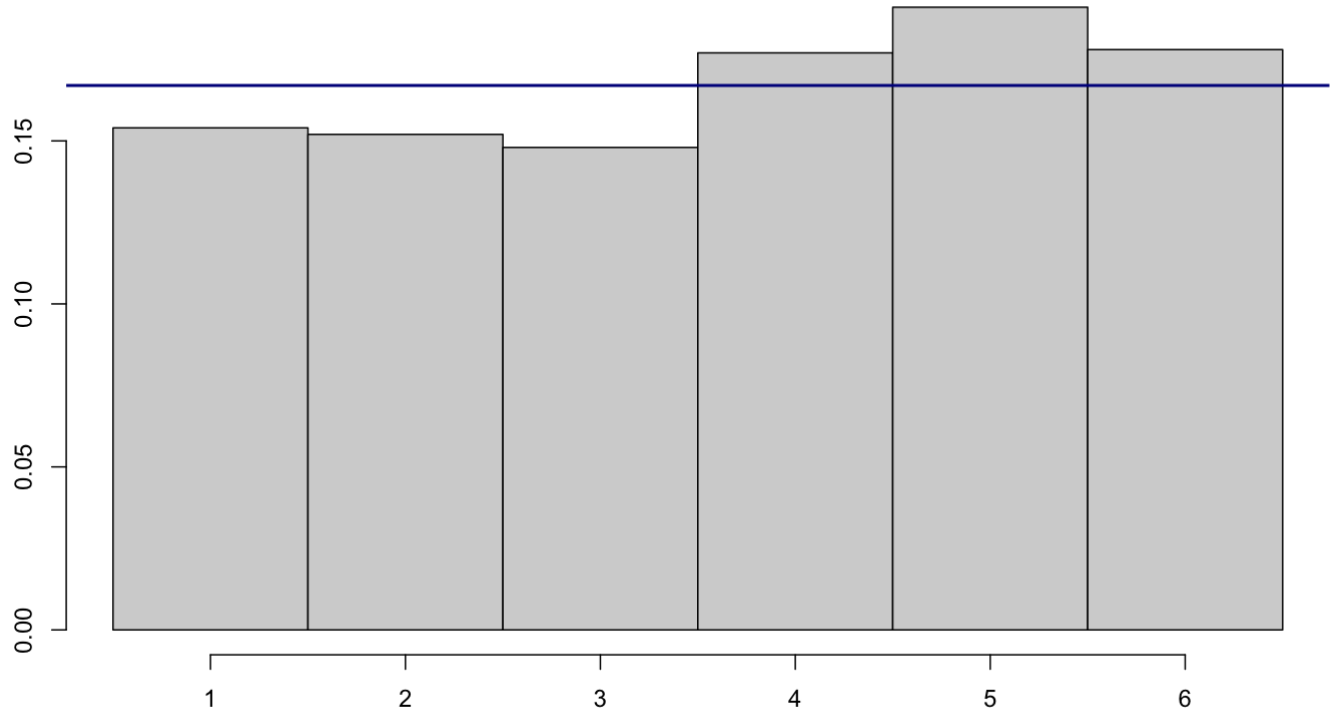
$$\text{mean}(\bar{Y}) = \mu.$$

$$\text{SD}(\bar{Y}) = \frac{\sigma}{\sqrt{n}}$$

Eksempel: Gennemsnitligt antal øjne kast med 5 terninger har $\mu = 3.5$, $\sigma/\sqrt{5} = 0.76$.

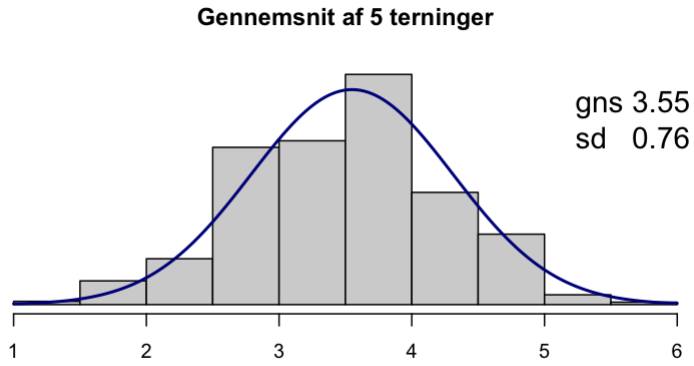
Fordeling af 1 terningekast

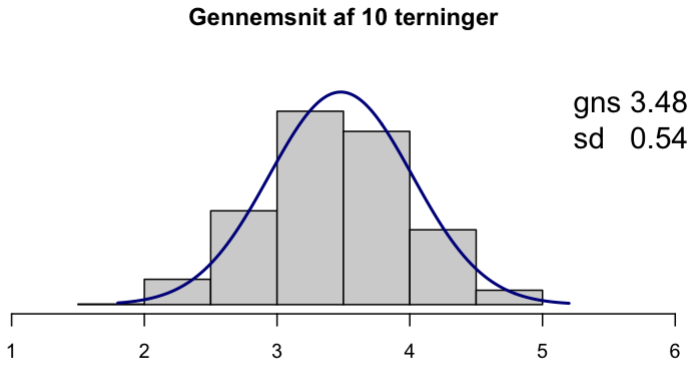
1000 observationer (kast) $Y_1, \ldots, Y_n$:

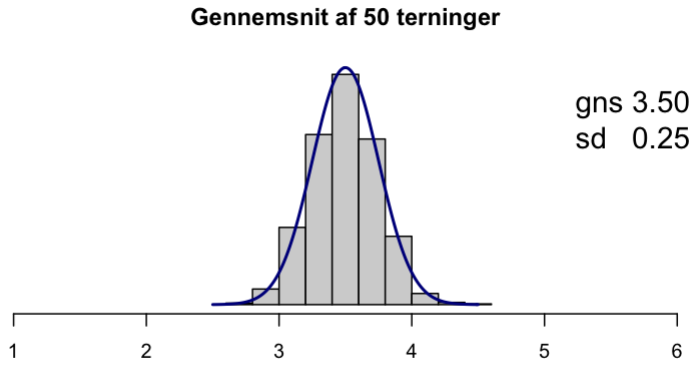


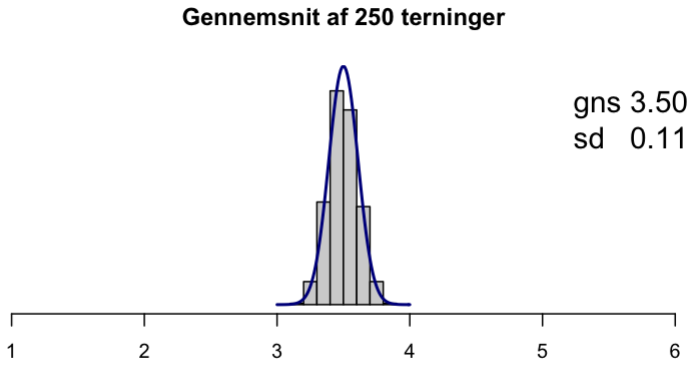
Fordeling af gennemsnit af terningekast

1000 gentagelser:





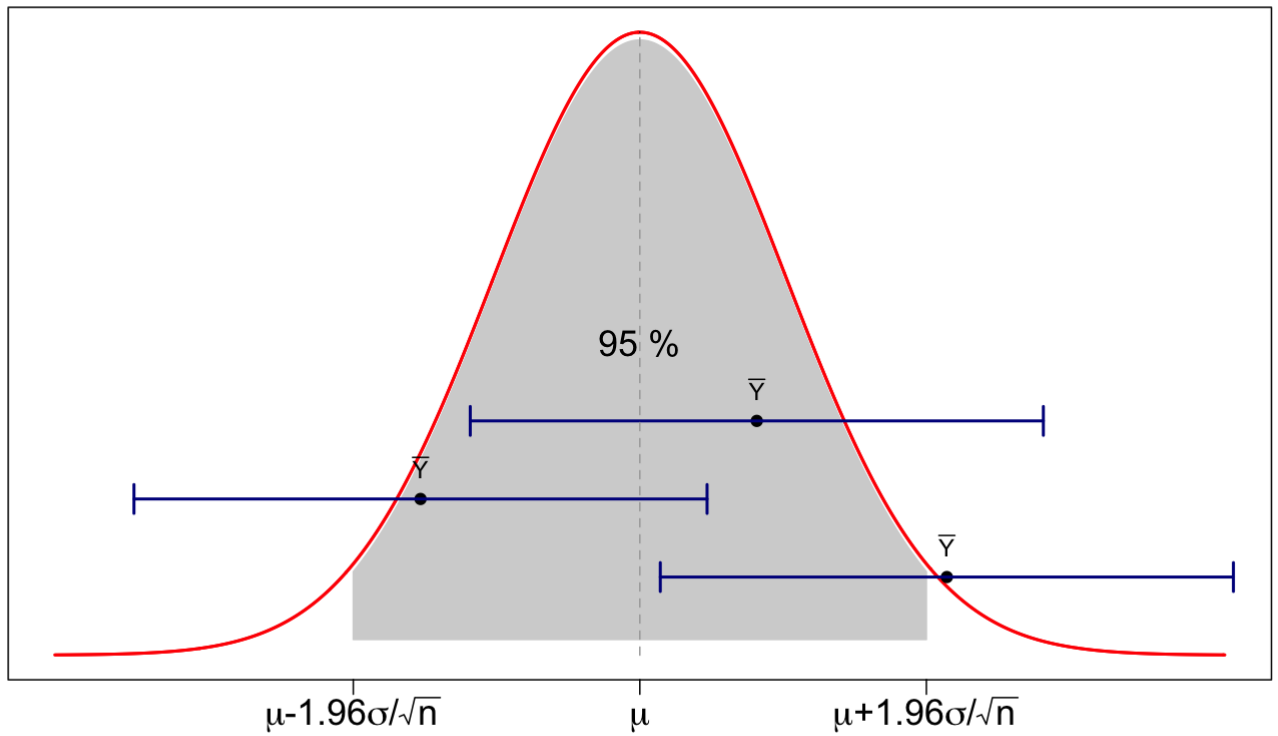




Konfidensintervaller

Fordeling af gennemsnit

Gennemsnittet er (approksimativt) normalfordelt med middelværdi μ og spredning $\sigma/\sqrt{n}$



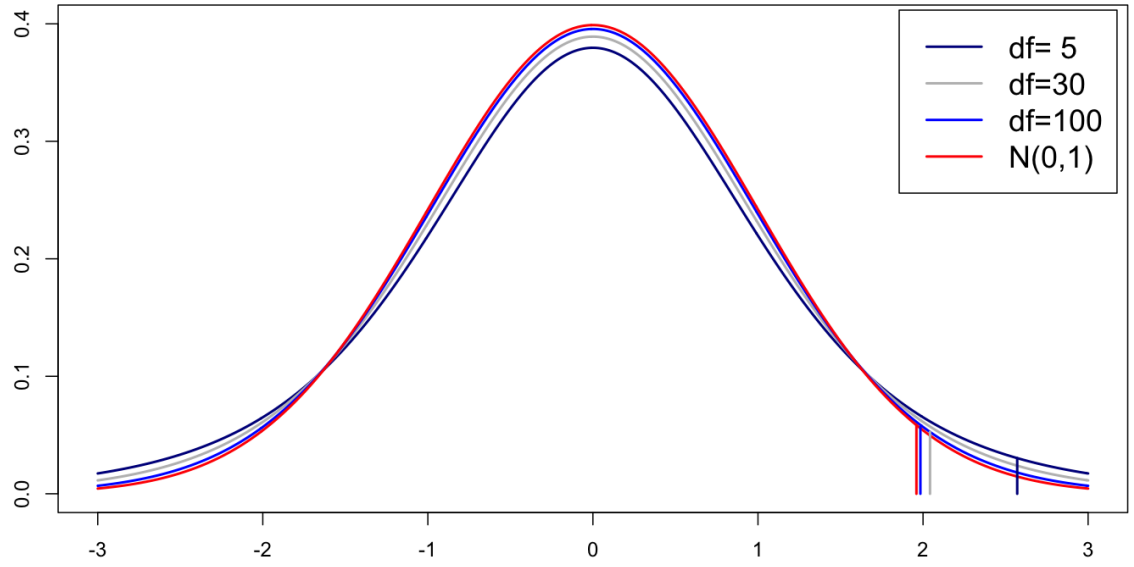
Konfidensinterval for middelværdien

$$\bar{Y} \pm q_{t,0.975}(n-1) \times \frac{\text{SD}}{\sqrt{n}}$$

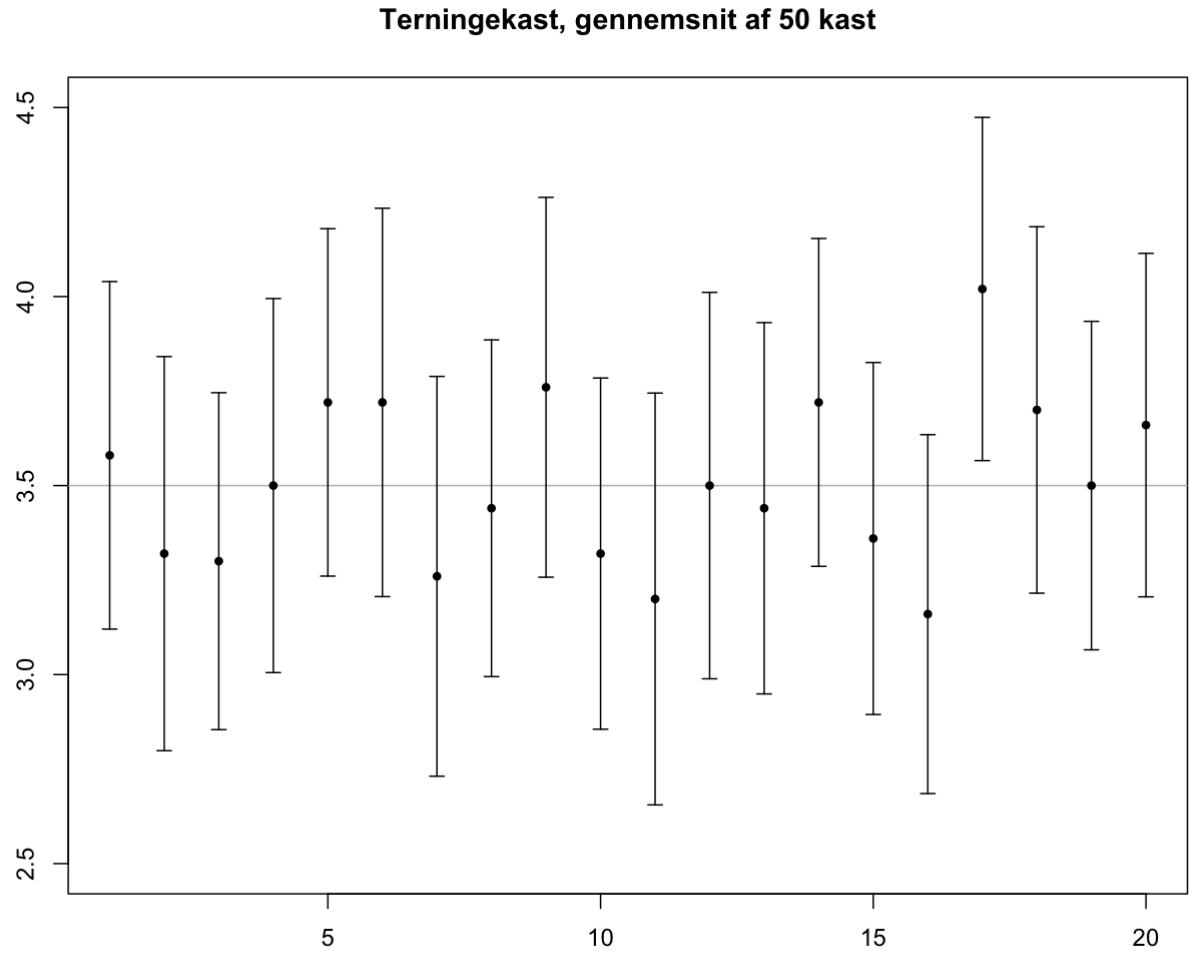
Her $q_{t,0.975}(n-1)$ 97.5%-fraktilen i t-fordelingen med $n-1$ frihedsgrader.

F.eks. med $n = 100$:

$q_{t,0.975}(99) = 1.984$



Fortolkningen af konfidensintervaller



Hvorfor er konfidensintervaller vigtige?

Vi ønsker at bestemme en **parameter**, e.g.

- gennemsnitligt vitamin D niveau
- median niveau af immunoglubolin

Baseret på en stikprøve kommer vi med et kvalificeret gæt (et *estimat*)

- vi er ikke helt sikre på vores gæt og foreslår et interval af plausible værdier
- vi ønsker at intervallet er smalt men ...
- ... også at vi har en stor sandsynlighed (95%) for at gætte rigtigt.

I må ***aldrig*** angive et estimat uden konfidensinterval!

Vigtigheden af normalfordelingen

afhænger af formålet med undersøgelsen

- **Vigtigt**
  - ved beskrivelser
  - i særdeleshed ved konstruktion af referenceområder
- **Knap så vigtig**
  - ved sammenligninger (sammenligning af gennemsnit, lineær regression) hvor antallet af observationer er tilstrækkeligt højt
- **Overhovedet ikke vigtigt**
  - for kovariater / forklarende variable (regressionsanalyse)

Meget mere om dette senere ...

Parrede sammenligninger

Parrede data:

- Observationer fra en situation 'er parret med' observationer fra en anden

Eksempler:

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

Sammenligning af målemetoder

Eksempel: Ultralydsundersøgelse af hjertet.

To metoder til bestemmelse af slagvolumen:

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

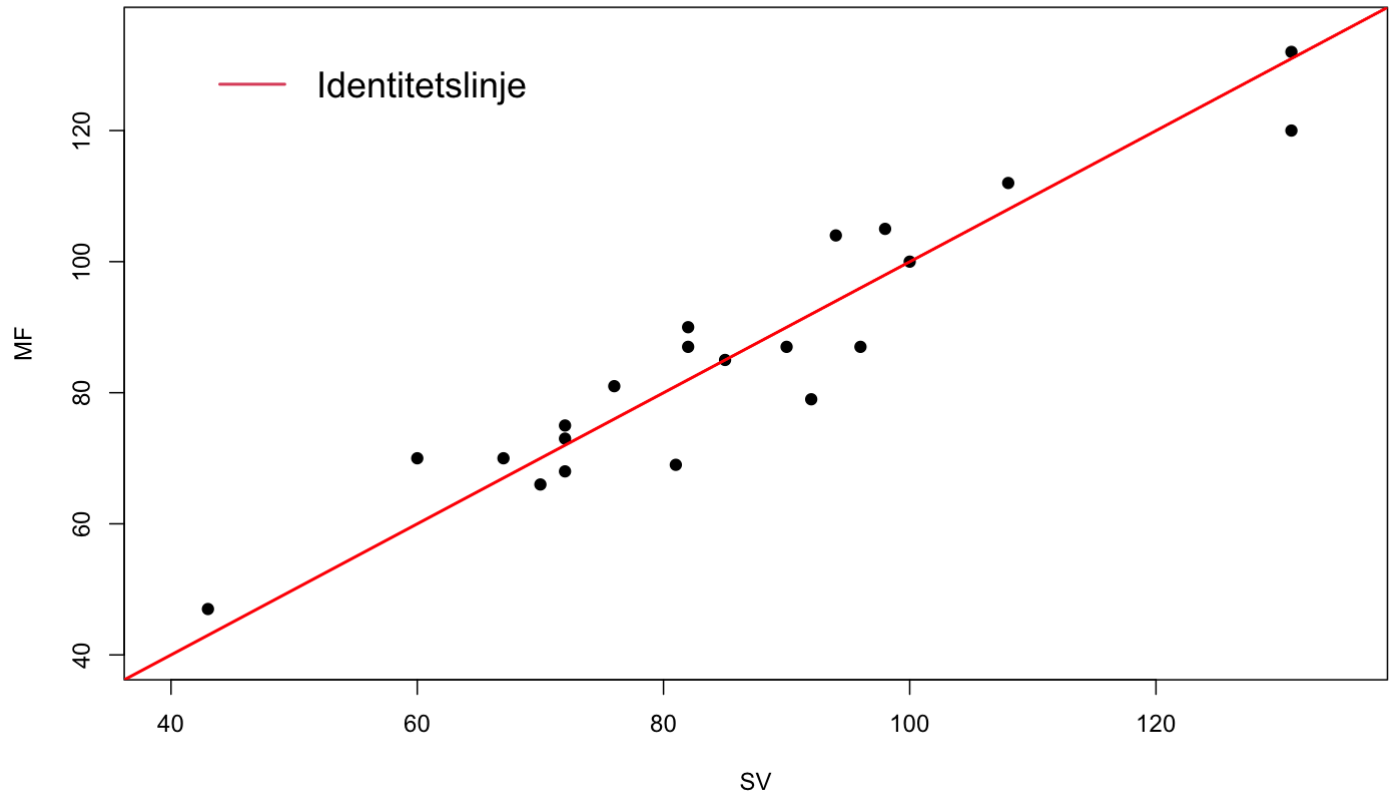
Måler de to målemetoder 'det samme'?

Data: R / SAS / SPSS

```
  mf sv
1 47 43
2 66 70
3 68 72
```

Scatter plot af MF vs. SV

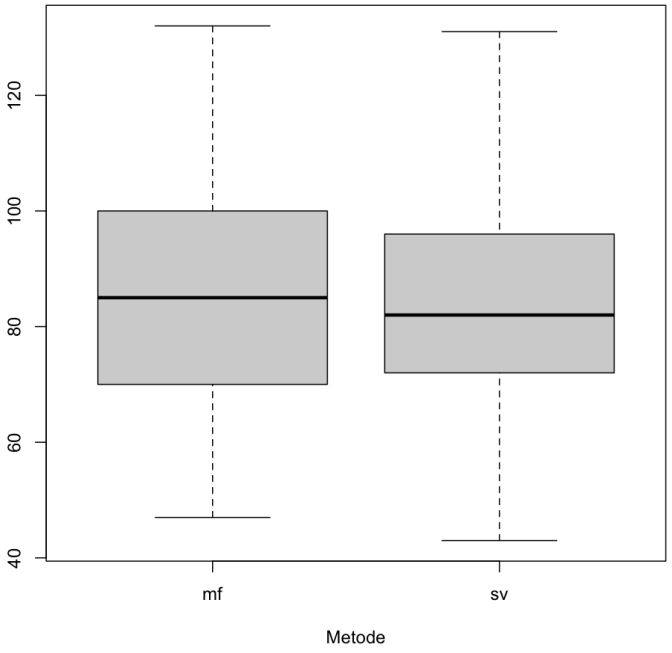
Lineær sammenhæng mellem de to målemetoder? Er de rimeligt ens? R / SAS / SPSS



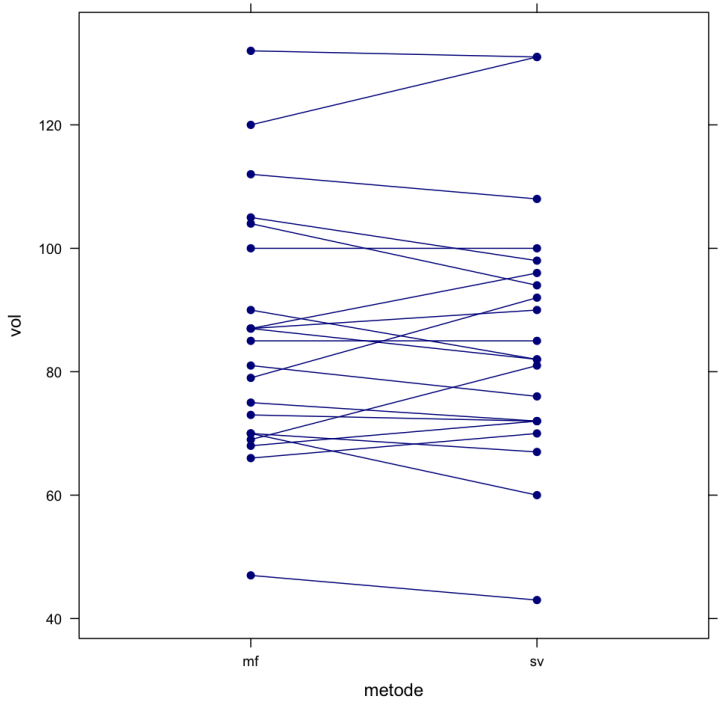
Yderligere illustration

Man skal kunne se parringen! R / SAS / SPSS

Forkert tegning



Rigtig tegning



Analyse af parrede data

- Personen er **sin egen kontrol** - giver stor styrke til at opdage 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 $\rightarrow$ se på differenserne på en logaritmisk skala.
- Undersøg om differenserne har middelværdi 0:
    - parret T-test (= one-sample T-test på differenserne)

Parret T-test / one-sample T-test

Statistisk model for parrede data

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

Differenser $D_i = X_i - Y_i$ er

- uafhængige
- normalfordelte $\mathcal{N}(\delta, \sigma^2)$
  - med middelværdi δ
  - samme spredning σ for alle personer

Bemærk:

- Kun antagelser om *differenserne*

Statistisk analyse

Med udgangspunkt i indsamlede data, hvad kan vi så sige om den sandsynligheds-mekanisme (model, herunder de ukendte parametre), der har frembragt disse data?

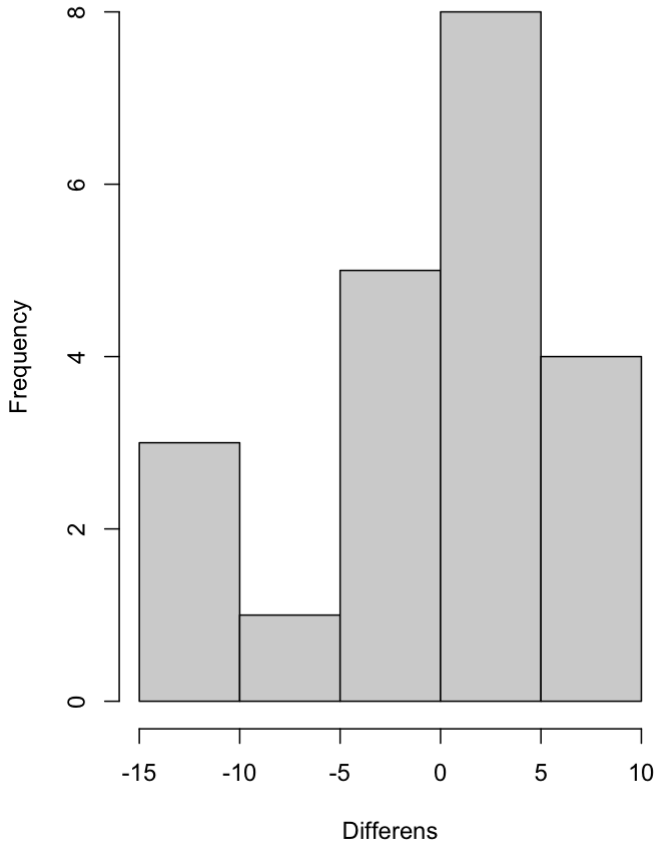
- Estimation:
  - Hvad kan vi så sige om de to ukendte parametre δ og σ ?
- Bias:
  - Ser der ud til at være systematisk forskel på de to metoder, dvs. er $\delta = 0$?
- Prædiktion:
  - Hvor store forskelle kan vi forvente i praksis?

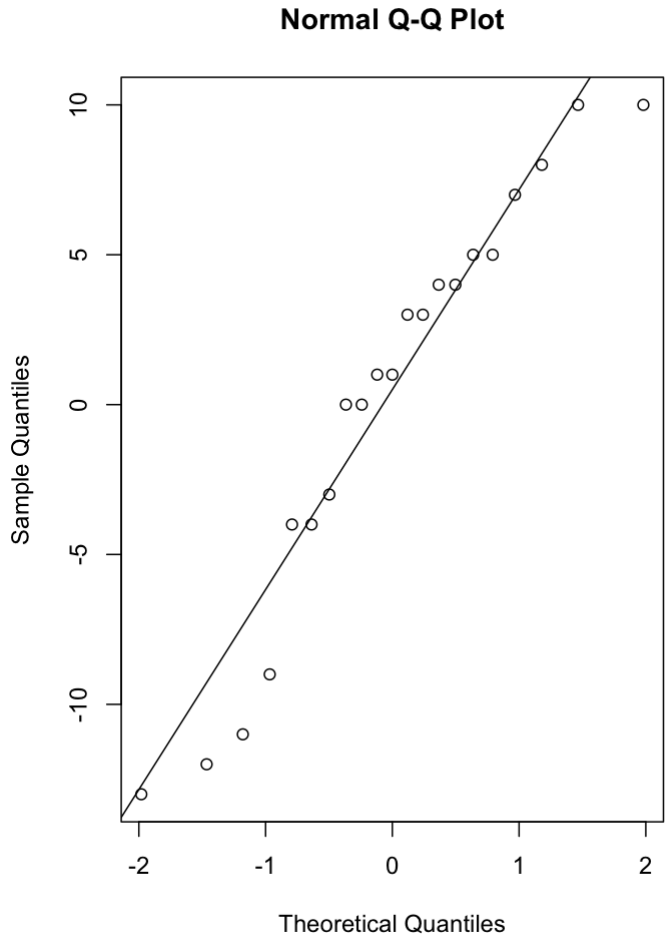
Antagelser for den parrede sammenligning

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

- er **uafhængige**
- har **samme spredning** (spaghettiplot / Bland-Altman plot af differenser mod gennemsnit)
- er **normalfordelte**
  - histogram / QQ-plot (men umuligt at sige med kun 21 observationer)
  - test for normalfordeling (aldrig!)
  - nogle gange vigtig, andre gange ikke

Differenser normalfordelt?





Estimation

Model: $D_i = \text{MF}_i - \text{SV}_i \sim \mathcal{N}(\delta, \sigma^2)$

Parametrene estimeres ved

- $\hat{\delta} = \bar{D} = 0.238$ (gennemsnittet på differenserne)
- $\hat{\sigma} = \text{SD} = 6.96$ (SD af differenserne)

$$\text{CI} = 0.238 \pm 2.09 \times \frac{6.96}{\sqrt{21}} = (-2.93, 3.41)$$

HUSK:

- SD er spredningen på outcome (data / D_i'erne).
- SE (eller SEM) er spredningen på estimatet / gennemsnittet ($\text{SE} = \frac{\text{SD}}{\sqrt{n}}$) $= 1.52$

Fortolkning af konfidensinterval

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. $3 - 3.4$ (til hver side)

Test af 'ingen bias' mellem MF og SV

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

Vi benytter et T-test på differenserne

$$\frac{\text{estimat} - \text{hypoteseværdi}}{\text{standard error for estimat}}$$

som under H_0 er T-fordelt med $n - 1$ frihedsgrader.

Dette er også et eksempel på et one-sample T-test ...

T-test for MF vs. SV

Her finder vi teststørrelsen:

$$t = \frac{\hat{\delta} - 0}{\text{SEM}} = \frac{0.24 - 0}{6.96/\sqrt{21}} = 0.16 \sim t(20)$$

Vi bruger p-værdier til at vurdere teststørrelsen.

P-værdier

Hvis hypotesen er sand og vi gentager forsøget mange gange:

Hvor ofte vil vi få en T-teststørrelse, som er **mindst** lige så stor som den observerede på 0.16?

Vi beregner p-værdien som

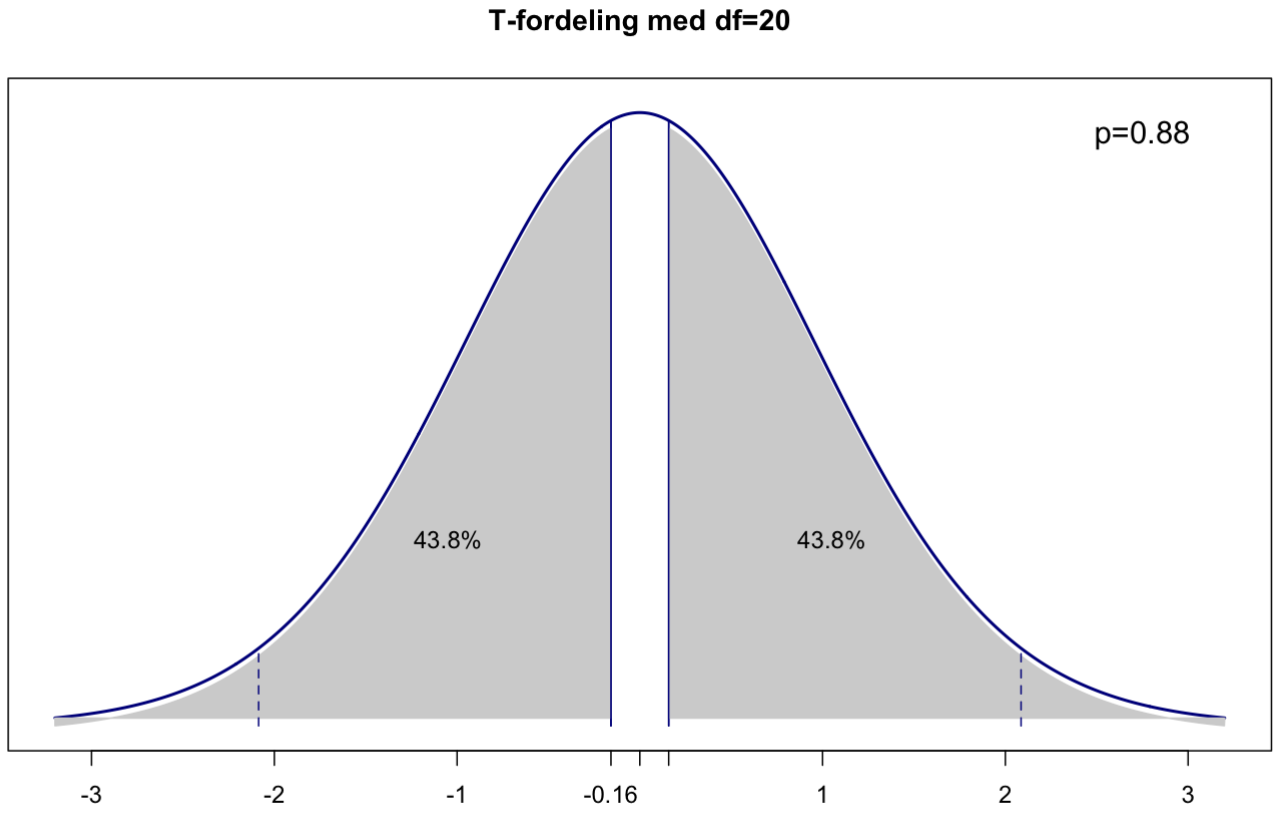
$$P(|\text{T-test størrelse}| > |\text{observeret T-teststørrelse}|)$$

under *antagelsen om at hypotesen er sand*.

Er p-værdien lille (<5%), svarende til at den observerede T-teststørrelse er "usandsynlig", konkluderer vi at det er **usandsynligt** at nul-hypotesen er sand og **afviser** H_0

Fordeling af T-teststørrelsen

Hvis H_0 er *sand*, følger T-teststørrelsen en t-fordeling med $n - 1 = 20$ frihedsgrader (df).



Test vs. konfidensinterval

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

- Hvis konfidensintervallet (sikkerhedsintervallet) indeholder H_0, er testet ikke signifikant
- Hvis konfidensintervallet (sikkerhedsintervallet) ikke indeholder H_0, er testet signifikant

Her er CI (-2.93, 3.41) og $p = 0.88$ dvs vi **kan ikke afvise hypotesen om middelværdi 0 for differenserne**.

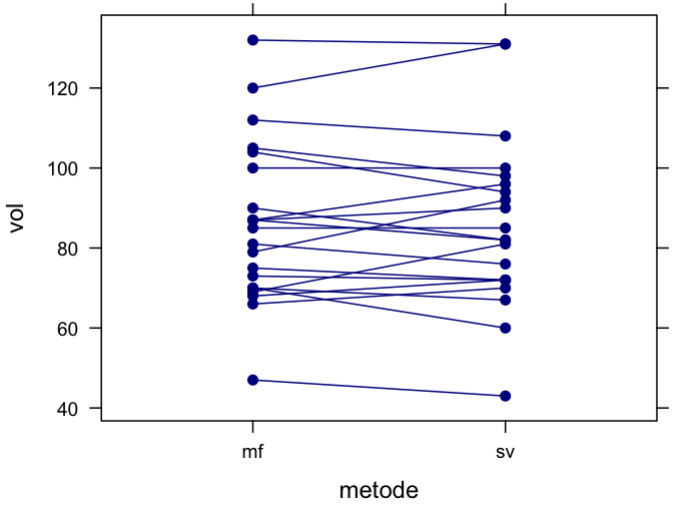
Men var det alt, hvad vi gerne ville vide?
Nej, vi vil gerne vide, hvor store forskellene typisk er....

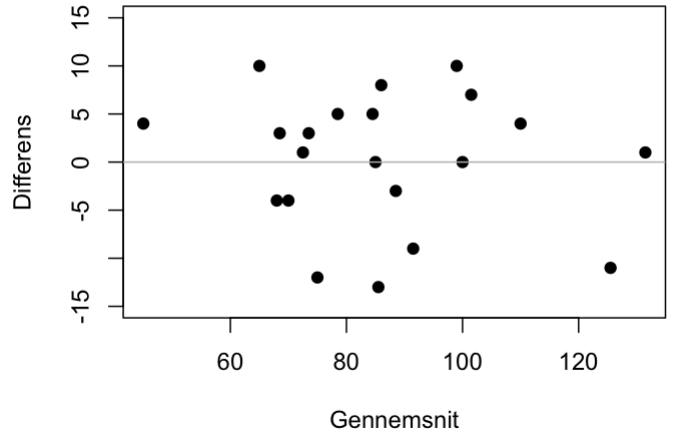
Bland-Altman plot

Limits Of Agreement

Bland-Altman plot

Plot af differenser `dif=mf-sv` mod gennemsnit `gns=(mf+sv)/2`:





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

R / SAS / SPSS

Limits Of Agreement

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

Limits Of Agreement (LOA) er **normalområdet for differenserne**:

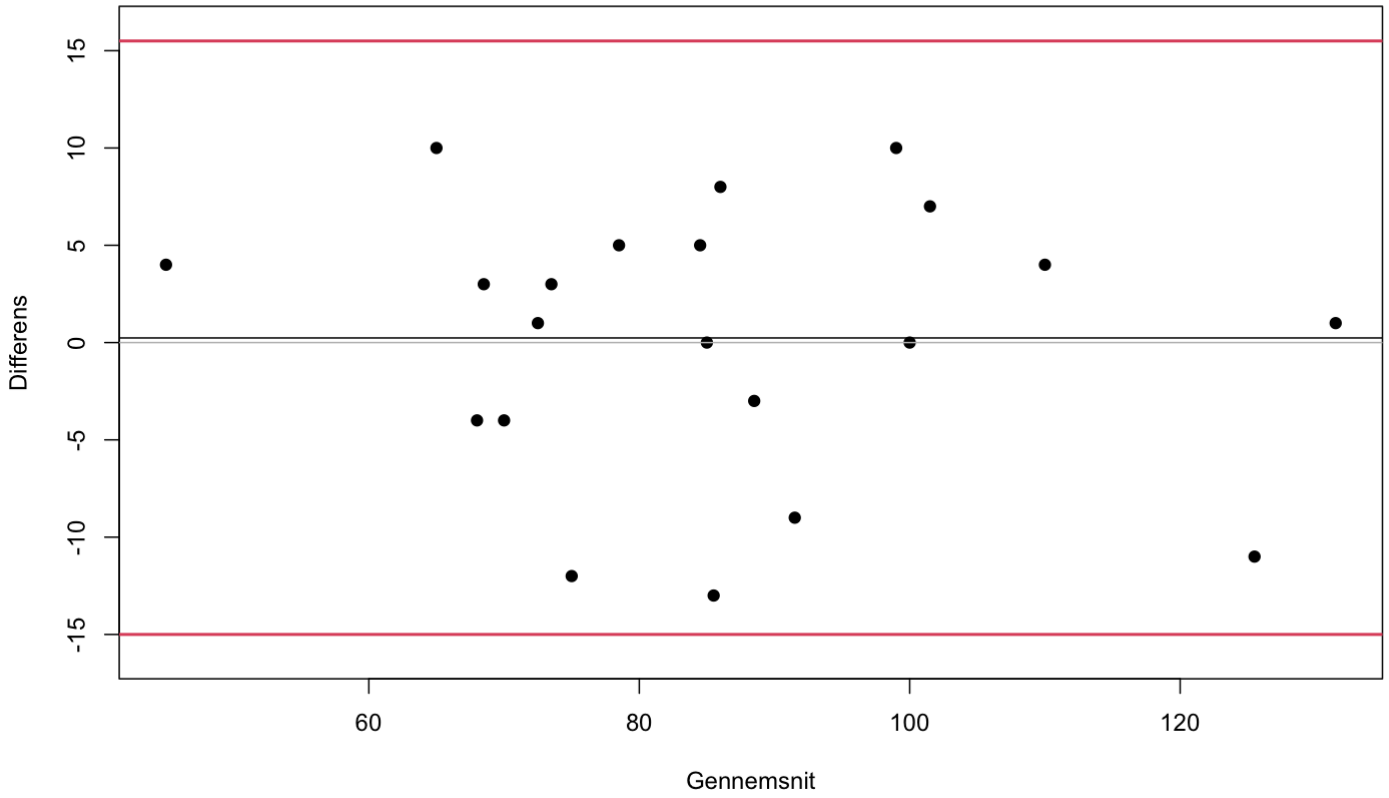
$$\bar{D} \pm \text{'cirka 2'} \times SD = 0.24 \pm 2 \times 6.96 = (-13.68, 14.16)$$

Disse grænser er vigtige for at afgøre om to målemetoder kan erstatte hinanden.

- Det er **ikke** nok, at der ikke er nogen systematisk forskel

NB: Cirka 2 er $\sqrt{1 + \frac{1}{n}} t_{97.5\%}(n-1)$

Limits Of Agreement illustreret



Eksempel: To slags termometre

- Måler de det samme?
  - Det er i hvert fald intentionen....
- Men er de ens?
  - Nej, vi kan se, at de ganske ofte afviger fra hinanden
- Det kunne skyldes måleusikkerhed, men er der også en konsekvent=systematisk forskel?
  - Lav parret T-test på passende skala og kvantificér forskel
- Vi fandt ingen signifikant forskel, er de så ens?
  - Måske ...

Eksempel: To slags termometre II

- Hvad kan vi så overhovedet sige?
    - Vi kan sige, at middelværdiforskellen næppe er større end det, som konfidensintervallet angiver.
- Er det overhovedet interessant? Det udtaler sig jo om populationer... Hvad med Fru Hansen, kan der være systematiske forskelle for hende?
    - Det kan vi kun se, hvis vi måler mange gange på Fru Hansen
- Men kan vi leve med de forskelle, vi ser hos Fru Hansen og alle de andre?
    - **Dette er ikke et statistisk spørgsmål!**

Summary

Du har nu lært at

- beskrive data grafisk og ved deskriptiv statistik
- vurdere om data er normalfordelt
- bestemme normalområder
- gennemsnit er (approksimativt) normalfordelte
- bestemme konfidensintervaller for gennemsnit
- kende forskellen på normalområde og konfidensinterval
- sammenligne to målemetoder ved CI, parret (one-sample) T-test og LOA