Epidemiologi og **biostatistik**.
Uge 1, tirsdag.
Erik Parner, Institut for Biostatistik.

**Generelt om statistik**

**Dataanalysen**
- Deskriptiv statistik
- Statistisk inferens

**Sammenligning af to grupper med kontinuerte data**
- Gennemsnit og spredning
- Parametre
- Estimater
- Sikkerhedsintervaller

**Deskriptiv statistik**



**Eksempel: PEFR**

Sammenligning af to grupper med kontinuerte data

**Udgangspunkt:** Vi ønsker at sammenligne lungefunktion for mænd og kvinder.

**Indsamling af data (stikprøve):** PEFR-målinger for

- 14 tilfældigt udvalgte kvinder
- 16 tilfældigt udvalgte mænd

**Data:**

- Kvinder: 522, 383, 428, 442, 500, 548, 540, 475, 540, 475, 510, 470, 485, 480
- Mænd: 580, 560, 460, 600, 600, 515, 550, 640, 550, 620, 510, 547, 540, 570, 430, 575



Hvorfor er der brug for statistik ?

- Data/observationer er underlagt **tilfældig variation**.
- Behov for at kvantificere hvor meget skyldes **tilfældig** og hvor meget skyldes **systematisk** variation.
- Behov for at resumere **mange** enkelte **observationer** i nogle **få tal**.
- Kvantificere at konklusioner baseret på **meget** data er mere **præcise** end konklusioner baseret på få data.



Formålet med den statistiske analyse er ofte at estimere en **ukendt konstant (parameter)**, som fx.:

- **Middel PEFR**
- **Middel PEFR** for en 30 årig kvinde
- **Forskel i (middel) PEFR** mellem mænd og kvinder
- Den **relative risiko** for SIDS forbundet med maveleje

**Bemærk:** disse parametre omhandler ikke kun vores stikprøve, men hele den population vi betragter.

Det kan være en større opgave, at beslutte sig til **hvilken** størrelse man ønsker at estimere:

Hvordan beskriver man sammenhængen mellem kost og kræft ?



**Hvorfor stikprøver (samples)?**

- hurtigere
- billigere
- umuligt at undersøge alle
- mere præcist (indsamling af data/ homogenitet)
- statistiske metoder kan bruges til at vurdere usikkerhed

**Dataanalysen** kan opdeles i

- deskriptiv statistik
- statistisk inferens



**Dataanalyse: deskriptiv statistik**

**Beskrivelse af data fra stikprøven:**

- 'Data summary':
  - gennemsnit / median / percentiler
  - hyppigheder / relativ risiko / oddsratio
  - varians / spredning
  - korrelationer
- Tegninger/figurer: Vigtig!

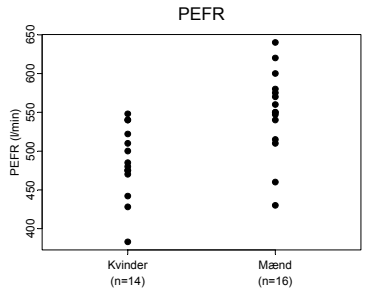
## Dataanalyse: statistisk inferens

**Fra stikprøve til population:**

- **Model** / Antagelser angående variationen i data.
- **Estimation** af relevante parametre i populationen (f.eks. middelværdi eller forskel mellem to grupper) ud fra stikprøven med tilhørende **sikkerhedsintervaller**.
- Opstilling af statistiske **hypoteser**, statistiske **test**
- **Statistiske konklusioner**
- **Faglig konklusioner**

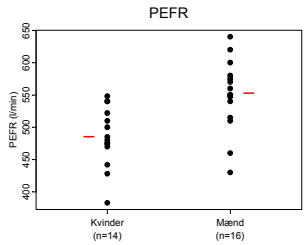


## Eksempel - deskriptiv statistik




Figuren er god, men kan man beskrive disse data med få tal?






Gennemsnittet for hver grupper er markeret med —

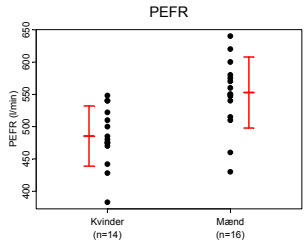
Kvinder: 485.6 l/min
Mænd: 552.9 l/min

$$\text{Gennemsnit} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = (x_1 + x_2 + \cdots + x_n)/n$$

(Summen af tallene divideret med antallet)

**Gennemsnittet beskriver midten / det generelle niveau / den centrale tendens af data.**






Kvinder: sd=46.6 l/min
Mænd: sd=55.0 l/min

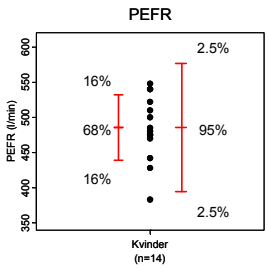
Jo mere data varierer jo større sd.

På figuren er vist gennemsnit +/- sd.

$$\text{sd} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

**Som et mål for variationen bruges ofte spredningen**
(standard afvigelsen / standard deviation / sd)






Hvad siger spredningen?

Hvis data er fordelt rimeligt symmetrisk omkring gennemsnittet (**normalfordelt**), da vil intervallet

gennemsnit ± sd

dække ca. 68% af data, og

gennemsnit ± 1.96*sd

dække ca. 95% af (kommende) data.

Vi vender tilbage disse intervaller (prædiktionsintervaller) næste gang.



**Parametre:** vi har lavet et gæt på parametrene

$\mu$ = middelværdi = gennemsnittet for hele populationen
$\sigma$ = spredning = sd udregnet for hele populationen

Normalfordelingen er beskrevet ved de to parametre: middelværdi og spredning.

Der kommer mere om normalfordelingen næste gang.

**Estimation:** Kvinder: $\hat{\mu}_K$ = gennemsnit = 485.6 l/min
$\hat{\sigma}_K$ = sd = 46.6 l/min
Mænd : $\hat{\mu}_M$ = gennemsnit = 552.9 l/min
$\hat{\sigma}_M$ = sd = 55.0 l/min

^: Dette er et estimat, dvs. et gæt beregnet på basis af data.

Hvor godt passer de observerede gennemsnit med de sande værdier?

Hvis vi havde taget 16 andre mænd og målt deres PEFR ville vi ikke få et gennemsnit på 552.9 l/min…

For at beskrive usikkerheden på estimatet bruger man ofte et (sikkerheds-) interval omkring estimatet.

**Sikkerhedsintervallet er de parameter-værdier der er forenelige (i en eller anden forstand) med data.**



## Sikkerhedsinterval

For et givet estimat (f.eks. gennemsnittet) kan man beregne en tilhørende usikkerhed / spredning (se).

Hvis antallet af data, n, er stor da vil intervallet

$$\text{Estimat} \pm 1.96 \cdot \text{se}(\text{Estimat})$$

være (approximativt) et **95% sikkerheds- / konfidensinterval for estimatet.**

Usikkerheden på gennemsnittet er:

$$\text{se}(\hat{\mu}) = \text{sem} = \frac{\text{sd}}{\sqrt{n}}$$

sem: **S**tandard **e**rror of the **m**ean



**Fortolkningen af et 95% sikkerhedsinterval:**

Hvis vi udtager mange stikprøver og beregner et sikkerheds- eller konfidensinterval for hver stikprøve da vil den sande værdi ligge i 95% af disse intervaller.

Sagt på en anden måde:

Sikkerhedsintervallet indeholder den sande værdi med 95% sandsynlighed.



## Eksempel – beregning af sikkerhedsinterval

**Kvinder:** $n = 14$, $\hat{\mu}_K = 485.6$ l/min, $\hat{\sigma}_K = 46.6$ l/min

$$\text{se}(\hat{\mu}_K) = \frac{46.6}{\sqrt{14}} = 12.4$$

$\text{CI}(\mu_K)$: $485.6 \pm 1.96 \cdot 12.4$

dvs. $(461.2; 510.0)$ l/min

CI: Confidence Interval

**Mænd:** $\text{se}(\hat{\mu}_M) = 13.7$ l/min

$\text{CI}(\mu_M) = (526.0; 579.9)$ l/min



Den estimerede **forskel** mellem mænd og kvinder:

$$\hat{\mu}_M - \hat{\mu}_K = 552.9 - 485.6 = 67.4 \text{ l/min}$$

Usikkerheden på forskellen i gennemsnittene:

$$\text{se}(\hat{\mu}_M - \hat{\mu}_K) = \sqrt{\text{se}(\hat{\mu}_M)^2 + \text{se}(\hat{\mu}_K)^2}$$

$$= \sqrt{12.4^2 + 13.7^2} = 18.5 \text{ l/min}$$

Sikkerhedsintervallet for forskellen bliver

$$\text{Estimat} \pm 1.96 \cdot \text{se}(\text{Estimat})$$

$\text{CI}(\mu_M - \mu_K)$: $67.4 \pm 1.96 \cdot 18.5$

dvs. $(31.0; 103.7)$ l/min

Der er altså statistisk signifikant forskel i PEFR mellem mænd og kvinder!



## Statistisk test

En anden måde at undersøge om der er forskel i PEFR mellem mænd og kvinder er vha **et statistisk test**.

Mere om dette næste gang.

## Eksempel - resultater

**PEFR niveau:**

Kvinder: $\hat{\mu}_K = \text{gennemsnit} = 486$ l/min

$\text{CI}(\mu_K) = (461;510)$ l/min

Mænd : $\hat{\mu}_M = \text{gennemsnit} = 553$ l/min

$\text{CI}(\mu_M) = (526;580)$ l/min

**Variation i PEFR:** Kvinder: $\hat{\sigma}_K = \text{sd} = 47$ l/min

Mænd : $\hat{\sigma}_M = \text{sd} = 55$ l/min

**Forskel i PEFR niveau:**

$\text{Forskel} = \hat{\mu}_M - \hat{\mu}_K = 67$ l/min

$\text{CI}(\mu_M - \mu_K) = (31;104)$ l/min



**Konklusion:**

Mænd har (statistisk signifikant) højere PEFR niveauet end kvinder!

Forskellen i PEFR er mellem 31 og 104 l/min.

Vores bedste bud på forskellen er 67 l/min.

**Bemærk:** konklusionen vedrører hele populationen, og ikke kun den stikprøve vi har undersøgt.



## Sammenligning af to grupper med kontinuerte data

generelt

**Statistisk model:** Antag

- at **variationen** i hver gruppe er **symmetrisk** (data er normalfordelt)
- observationerne indenfor hver gruppe er **uafhængige** (ingen søskene indenfor grupperne)
- de to sæt af observationer er **uafhængige** (ingen søskene, ikke par af målinger i de to grupper)

**Estimation:**

$\hat{\mu}_i = \text{gennemsnit}$ (beskriver niveauet i gruppen)

$\hat{\sigma}_i = \text{sd}$ (beskriver variationen i gruppen)

($i = 1,2$ svarende til gruppenummer)



**Sikkerhedsinterval for middelværdien:**

$$\text{se}(\hat{\mu}_i) = \frac{\hat{\sigma}_i}{\sqrt{n_i}}$$

$$\text{CI}(\mu_i): \ \hat{\mu}_i \pm 1.96 \cdot \text{se}(\hat{\mu}_i)$$

**Sikkerhedsinterval på forskellen:**

$$\text{se}(\hat{\mu}_1 - \hat{\mu}_2) = \sqrt{\text{se}(\hat{\mu}_1)^2 + \text{se}(\hat{\mu}_2)^2}$$

$$\text{CI}(\mu_1 - \mu_2): \ \hat{\mu}_1 - \hat{\mu}_2 \pm 1.96 \cdot \text{se}(\hat{\mu}_1 - \hat{\mu}_2)$$

**Bemærk:** Formlen for se gælder generelt for alle parametre forudsat de to grupper er **uafhængige**.



## Et nyt, større studie

Et større studie for den samme population gav følgende resultat:

| | n | Gennemsnit | CI |
|---|---|---|---|
| Kvinder | 43 | 474 | (459;489) |
| Mænd | 58 | 568 | (552;584) |
| Forskel | | 94 | (72;116) |

Til sammenligning fik vi tidligere:

| | n | Gennemsnit | CI |
|---|---|---|---|
| Kvinder | 14 | 486 | (461;510) |
| Mænd | 16 | 553 | (526;580) |
| Forskel | | 67 | (31;104) |

**Sikkerhedsintervallerne bliver mindre jo større studiet er!**

(Vi bliver klogere jo mere data vi samler ind...)



## Dataanalysen: deskriptiv statistik

**Numeriske metoder til beskrivelse af kontinuerte data:**

Hvor ligger typiske data (det generelle niveau):

| | |
|---|---|
| Gennemsnit (aritmetrisk): | data skal være rimelig symmetrisk fordelt |
| Gennemsnit (geometrisk): (f.eks koncentrationer) | logaritme-transformerede data skal være rimelig symmetrisk fordelt |
| Median (50 percentil): | skæve fordelinger |

**Numeriske metoder til beskrivelse af kontinuerte data:**

Hvor meget afviger de fra hinanden (variation)?

| | |
|---|---|
| Spredning/varians: | data skal være rimelig symmetrisk fordelt |
| Variationskoefficient (f.eks koncentrationer) | logaritme-transformerede data skal være rimelig symmetrisk fordelt |
| Percentiler / kvartiler: Range/ max / min | skæve fordelinger |



**Percentiler:**

5 percentilen er der hvor der er 5% af data der er mindre (og 95% større)

25 percentilen er lig 1. kvartile

50 percentilen er lig 2. kvartil der er lig medianen

95 percentilen er der hvor der er 95% af data der er mindre (og 5% større)

og generelt

X percentilen er der hvor der er X% af data der er mindre (og 100-X% større)



**Kvartiler:**

0. kvartil er lig det **mindste** tal

1. kvartil = 25 percentilen

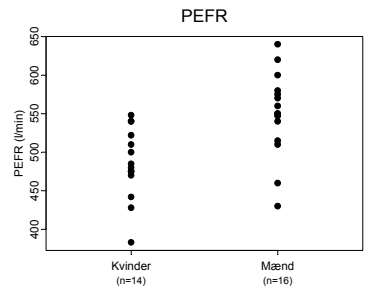
2. kvartil = 50 percentilen = medianen

3. kvartil = 75 percentilen

4. kvartil er lig det **største** tal

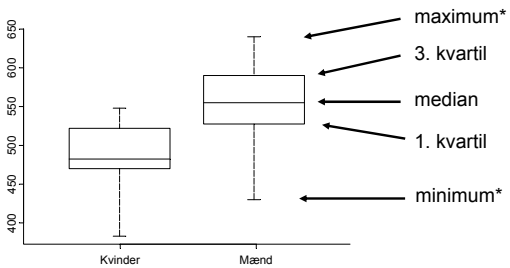


**Præsentation af numeriske data: scatterplot**





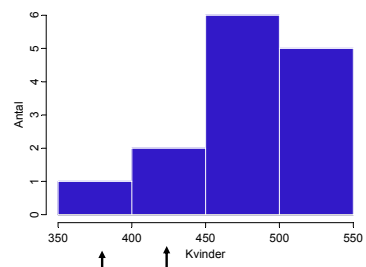

**Præsentation af numeriske data: boxplot**




* det varierer lidt hvordan man definerer den øvre og nedre grænse.



**Præsentation af numeriske data: histogram**




to observationer mellem 400-450

én observation mellem 350-400

Histogrammet beskriver stikprøvens fordeling.

Vi vil senere bruge histogrammet til at undersøge om data er normalfordelt.



**Typer af data**

Vi skal i dette kursus kigge på statistiske analyser af

- kontinuerte [ PEFR ]
- to kategorier (dichotom) [ Syg/rask ]
- flere kategorier [ Hvilket amt man bor ]
- ventetidsdata [ Tid til død eller tilbagefald af sygdom ]

**Den statistiske analyse afhænger af typen af data og hvordan data er indsamlet.**

Vi vil også kigge på statistiske analyser som kan besvare spørgsmål som:

Kan forskellen i mænd og kvinder's PEFR værdi forklares ved at mænd er højre end kvinder?



**Resumé**

- Dataanalysen
- Sammenligning af to grupper med kontinuerte data:
  - Gennemsnit og spredning
  - Estimater
  - Sikkerhedsintervaller
- Deskriptiv statistik

1

Epidemiologi og **biostatistik**

Uge 1, torsdag 6. februar 2003

Morten Frydenberg, Institut for Biostatistik.

- **Bronkitis og hoste**
  - estimation
  - sikkerhedsintervaller
  - antagelser
- **Normalfordelingen**
- **Prædiktion**
- **Statistisk test (udfra estimat og standard error)**
- **Sikkerhedsintervaller og statistiske tests**

2

**Lungefunktions data fra tirsdags**

| Køn | Gennemsnit l/min | se l/min |
|---|---|---|
| Kvinder | 485.6 | 12.5 |
| Mænd | 552.9 | 13.8 |

Udfra dette kunne vi beregne **sikkerhedsintervaller** for:

- Middelværdien for hvert køn
- **Differensen** mellem middel PEFR for mænd og kvinder

95% **sikkerhedsinterval** : CI: Estimat ± 1.96×se(Estimat)

3

**Bronkitis og hoste**

Har bronkitis i den tidlige barndom betydning senere i livet?

**Observeret !**

| | Hoster om natten som 14-årig | | |
|---|---|---|---|
| Bronkitis som 5-årig | Ja | Nej | Total |
| Ja $(+B)$ | 26 | 247 | 273 |
| Nej $(-B)$ | 44 | 1002 | 1046 |

Lad os først se på de, der ikke har haft bronkitis.

$\pi_{-B}$ = Sandsynlighed for at hoste om natten givet man ikke har haft bronkitis **Ukendt !**

**Estimat:** $\hat{\pi}_{-B} = \frac{44}{1046} = 0.04207$

Bedste bud: 4.2% af de, der **ikke** har haft bronkitis, hoster om natten.

4

| | Hoster om natten som 14-årig | | |
|---|---|---|---|
| Bronkitis | Ja | Nej | Total |
| Ja | 26 | 247 | 273 |
| Nej | 44 | 1002 | 1046 |

$\hat{\pi}_{-B} =$ 0.04207

Hvad er usikkerheden, se, på estimatet?

$$\text{se}(\hat{\pi}_{-B}) = \sqrt{\hat{\pi}_{-B}(1-\hat{\pi}_{-B})/n_{-B}}$$
$$= \sqrt{0.04207(1-0.04207)/1046}$$
$$= 0.00621$$
$$\text{CI}(\pi_{-B}) = \hat{\pi}_{-B} \pm 1.96 \cdot \text{se}(\hat{\pi}_{-B})$$
$$= 0.04207 \pm 1.96 \cdot 0.00621$$
$$= (0.02990; 0.05423)$$

5

| | Risiko for hoste om natten | | |
|---|---|---|---|
| Bronkitis | Estimate | se | CI |
| Ja | **0.095**24 | 0.01777 | **0.060; 0.130** |
| Nej | **0.042**07 | 0.00621 | **0.030; 0.054** |

**Konklusion (**På basis af disse data **):**

- Risiko for at et barn, der **ikke** har haft bronkitis, hoster ligger et sted mellem 3.0% og 5.4% - bedste bud er 4.2%.
- Risiko for at et barn, der **har** haft bronkitis hoster, ligger et sted mellem 6.0% og 13.0% - bedste bud er 9.5%.
- Noget tyder på større risiko for at hoste om natten, når man har haft bronkitis.

6

| | Risiko for hoste om natten | | |
|---|---|---|---|
| Bronkitis | Estimate | se | CI |
| Ja | 0.09524 | 0.01777 | 0.060; 0.130 |
| Nej | 0.04207 | 0.00621 | 0.030; 0.054 |

**Risikodifferens:** $RD = \pi_{+B} - \pi_{-B}$

$$\widehat{RD} = \hat{\pi}_{+B} - \hat{\pi}_{-B} = 0.09524 - 0.04207 = 0.05317$$

$$\text{se}\left(\widehat{RD}\right) = \sqrt{\text{se}(\hat{\pi}_{+B})^2 + \text{se}(\hat{\pi}_{-B})^2} = \sqrt{0.01777^2 + 0.0621^2} = 0.01882$$

$$\text{CI}(RD) = 0.05317 \pm 1.96 \cdot 0.01882 = (0.016; 0.090)$$

7

| | Risiko for hoste om natten | | |
|---|---|---|---|
| Bronkitis | Estimate | se | CI |
| Ja | **0.095**24 | 0.01777 | **0.060; 0.130** |
| Nej | **0.042**07 | 0.00621 | **0.030; 0.054** |
| Risiko Differens | **0.053**17 | 0.01882 | **0.016; 0.090** |

**Konklusion:**
Risikoen for hoste om natten er et sted mellem 1.6 og 9.0 *procentpoint* højere, hvis man har haft bronkitis som 5-årig.

**Bemærk**

- se er mindst for 'Nej' gruppen, da der er langt flere børn i denne gruppe.
- Usikkerheden på differensen er større end den største usikkerhed for de to grupper.

8

Hvilke **antagelser** ligger bag beregningerne?

Antagelse 1: **Uafhængighed** mellem grupper

Antagelse 2: Data i hver gruppe er **binomial-fordelt**

**Uafhængighed mellem grupper:**

Denne antagelse er **nødvendig** for at man kan bruge formlen:

$$\mathrm{se}\left(\widehat{RD}\right)=\sqrt{\mathrm{se}\left(\hat{\pi}_{+B}\right)^2+\mathrm{se}\left(\hat{\pi}_{-B}\right)^2}$$

Er den **rimelig** i bronkitis eksemplet ?

**Ja**, data stammer for to forskellige grupper børn.

Et **muligt problem** kunne være hvis der var to **søskende** i hver sin gruppe. Så vil der pga. arv/miljø være en sammenhæng mellem hvorvidt de to børn hoster.

9

**Data i hver af grupperne er binomial-fordelt:**

Denne antagelse er **nødvendig** for, at man kan bruge formlen:

$$\mathrm{se}\left(\hat{\pi}\right)=\sqrt{\hat{\pi}\left(1-\hat{\pi}\right)/n}$$

| Data er binomialfordelt hvis: | Opfyldt? |
|---|---|
| 1 **Uafhængige** 'delforsøg'. | Ingen søskende i samme gruppe. |
| 2 Præcist **to mulige udfald** (hoster/ikke hoster, død/levende). | Klar definition af hoste. |
| 3 **Sandsynligheden** for succes, $\pi$, er **den samme** for alle delforsøg. | Grupperne kan betragtes som homogene. |
| 4 **Antal, *n* ,** delforsøg man betragter **afhænger ikke** af udfaldene. | Der er ikke snydt under data indsamlingen. |

10

**Normalfordelingen**

En vigtig fordeling af to forskellige grunde:

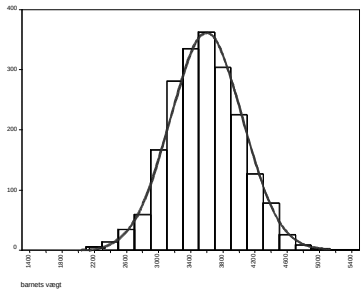
- Mange slags **data** er næsten normalfordelte (muligvis efter en transformation).
- Mange **estimater** er næsten normalfordelte, hvis de er baseret på **mange observationer** (muligvis efter en transformation).

Ingenting er helt normalfordelt, men mange gange er det en rigtig god approksimation !

**Relative** størrelser som **Odds Ratio**, **Relative Risiko** og **Rate Ratio** skal analyseres på **log-skala** (ln).

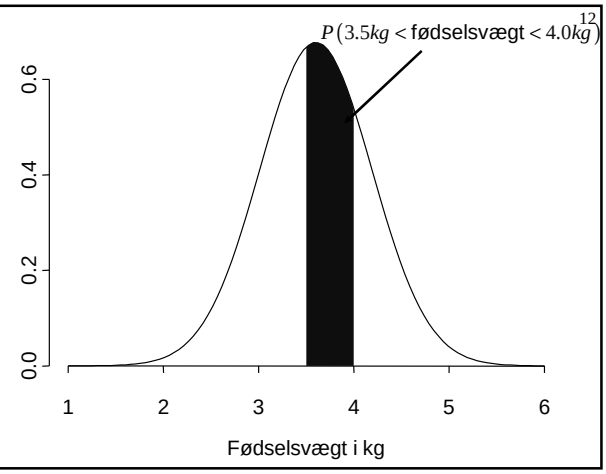
11

Fødselsvægt for 2031 børn




Normalfordeling:
en god approksimation

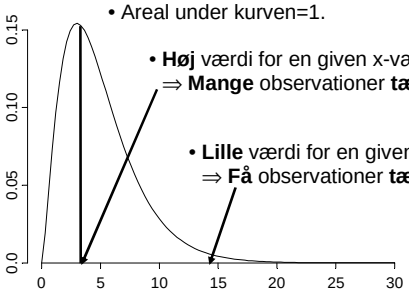
12



13

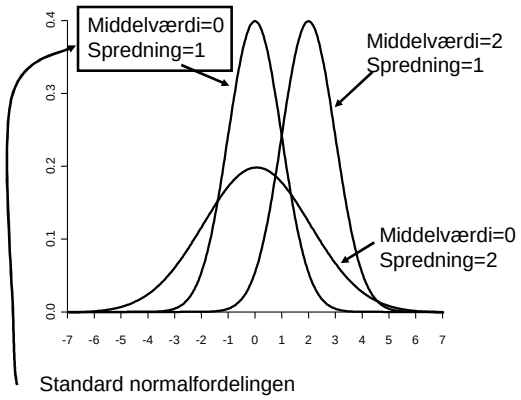
Tæthedsfunktion:

- Sandsynlighed for en observation i et interval = areal under kurven.
- Areal under kurven=1.
- **Høj** værdi for en given x-værdi ⇒ **Mange** observationer **tæt** ved denne værdi.
- **Lille** værdi for en given x-værdi ⇒ **Få** observationer **tæt** ved denne værdi.




14

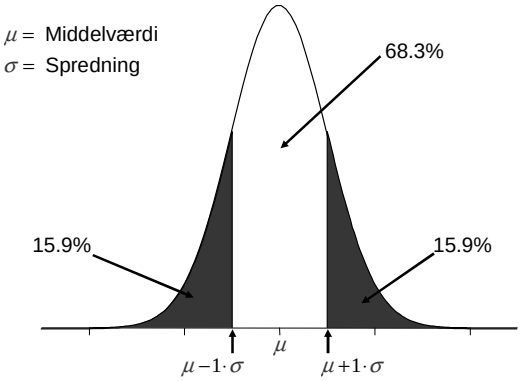
**Forskellige normalfordelinger:**




Standard normalfordelingen

15

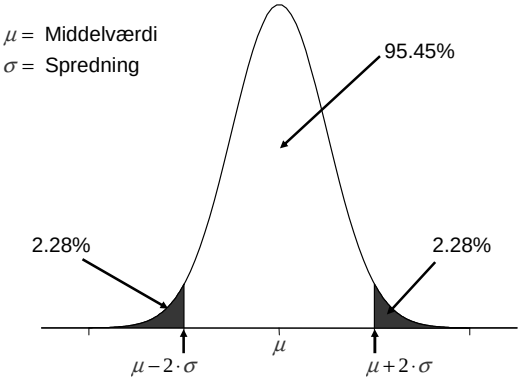
$\mu =$ Middelværdi
$\sigma =$ Spredning




Bland side 109

16

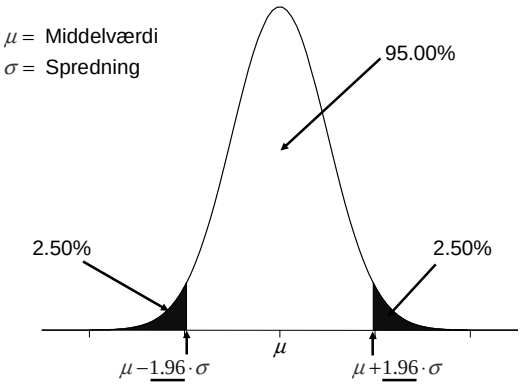
$\mu =$ Middelværdi
$\sigma =$ Spredning




Bland side 109

17

$\mu =$ Middelværdi
$\sigma =$ Spredning




Bland side 110

18

**Tabel over standardnormalfordelingen**

Bland side 109

| $z$ | $P(Z<z)$ | $z$ | $P(Z<z)$ | $z$ | $P(Z<z)$ |
|---|---|---|---|---|---|
| -3.0 | 0.1% | -1.0 | 15.9% | 1.0 | 84.1% |
| -2.9 | 0.2% | -0.9 | 18.4% | 1.1 | 86.4% |
| -2.8 | 0.3% | -0.8 | 21.2% | 1.2 | 88.5% |
| -2.7 | 0.3% | -0.7 | 24.2% | 1.3 | 90.3% |
| -2.6 | 0.5% | -0.6 | 27.4% | 1.4 | 91.9% |
| -2.5 | 0.6% | -0.5 | 30.9% | 1.5 | 93.3% |
| -2.4 | 0.8% | -0.4 | 34.5% | 1.6 | 94.5% |
| -2.3 | 1.1% | -0.3 | 38.2% | 1.7 | 95.5% |
| -2.2 | 1.4% | -0.2 | 42.1% | 1.8 | 96.4% |
| -2.1 | 1.8% | -0.1 | 46.0% | 1.9 | 97.1% |
| -2.0 | 2.3% | 0.0 | 50.0% | 2.0 | 97.7% |
| -1.9 | 2.9% | 0.1 | 54.0% | 2.1 | 98.2% |
| -1.8 | 3.6% | 0.2 | 57.9% | 2.2 | 98.6% |
| -1.7 | 4.5% | 0.3 | 61.8% | 2.3 | 98.9% |
| -1.6 | 5.5% | 0.4 | 65.5% | 2.4 | 99.2% |
| -1.5 | 6.7% | 0.5 | 69.1% | 2.5 | 99.4% |
| -1.4 | 8.1% | 0.6 | 72.6% | 2.6 | 99.5% |
| -1.3 | 9.7% | 0.7 | 75.8% | 2.7 | 99.7% |
| -1.2 | 11.5% | 0.8 | 78.8% | 2.8 | 99.7% |
| -1.1 | 13.6% | 0.9 | 81.6% | 2.9 | 99.8% |
| -1.0 | 15.9% | 1.0 | 84.1% | 3.0 | 99.9% |

19

Sandsynlighed for mere end 1.96 spredninger fra middelværdi:

5%

i en normalfordeling!

1 ud af 20 observationer: Mere end 1.96×sd fra middelværdi

***s**tandard **d**eviation* (spredning)

95% af observationerne fra en normalfordeling :

$$\text{middelværdi} - 1.96 \cdot \text{sd} \le \text{observation} \le \text{middelværdi} + 1.96 \cdot \text{sd}$$

95% **prædiktionsinterval** for en **observation**

20

Dvs. der er 95% chance for:

$$-1.96 \le \frac{\text{observation - middelværdi}}{\text{sd}} \le 1.96$$

Middelværdi **ukendt,** men sd **kendt**

95% **sikkerhedsinterval** for **middelværdien:**

$$\text{observation} - 1.96 \cdot \text{sd} \le \text{middelværdi} \le \text{observation} + 1.96 \cdot \text{sd}$$

Baseret på **én** observation!

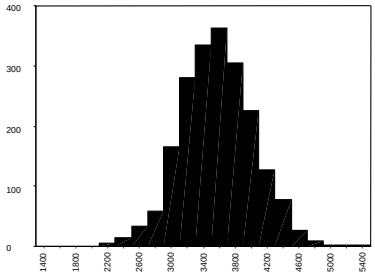
Baseres det på basis af ***n*** observationer fås:

$$\text{gennemsnit} - 1.96 \cdot \text{sem} \le \text{middelværdi} \le \text{gennemsnit} + 1.96 \cdot \text{sem}$$

$$\text{sem} = \frac{\text{sd}}{\sqrt{n}}$$ **S**tandard **e**rror of the **m**ean

21

Tilbage til fødselsvægtene:




Godt beskrevet ved en normalfordeling!

$n = 2031$

$\bar{x} = 3558g$

$\text{sd} = 446g$

Et 95% **prædiktionsinterval** for fødselsvægten:

$$3558g \pm 1.96 \cdot 446g = (2683; 4432)\,g$$

**Konklusion:** 95% af børn fra en tilsvarende population vil have en fødselsvægt mellem 2.7 og 4.4 kg .

22

**Statistisk test**

Risikodifferensen for hoste blandt børn, der har/ikke har haft bronkitis.

Risikodifferensen, *RD*, er **ukendt!**

**Men** vi har et estimat : $\widehat{RD} = 0.05317 \;\; se\left(\widehat{RD}\right) = 0.01882$

**Spørgsmål:** Er disse data forenelige med at *RD*=0.0? Dvs. ingen sammenhæng med bronkitis.

Der gælder at estimatet, $\widehat{RD}$ , er (næsten) normalfordelt

**Med spredning**=se=0.01882

**middelværdi** *RD*

Under **hypotesen** er *RD* =0

23

Normalfordeling med:

**middelværdi** 0 ← **Hypotesen!**

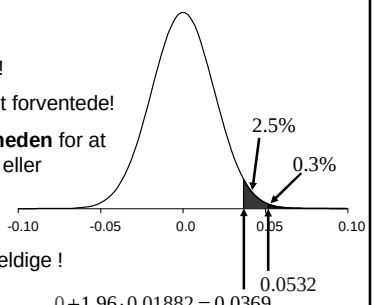
**spredning**=se=0.01882

Vi har observeret 0.0532 !

Det afviger (noget) fra det forventede!

Hvor stor er **sandsynligheden** for at observere en lige så stor eller større afvigelse?

0.3% !!




Vi har godt nok været uheldige !

Det tror jeg ikke vi har !

$0 + 1.96 \cdot 0.01882 = 0.0369$

Så må **hypotesen** være forkert !

Vi **forkaster hypotesen** : "Risikodifferensen er 0"

24

Hvad var nu det ?

Vi sammenlignede vores **estimat** (0.0532) med **hypotesen** 0.

Som 'målestok' brugte vi usikkerheden på estimatet: se=0.1882

Estimat Hypotese

$$\frac{\widehat{RD} - RD_0}{\text{se}\left(\widehat{RD}\right)} = \frac{0.0532 - 0}{0.01882} = 2.83$$

Usikkerheden på estimatet

Dvs. estimatet ligger 2.83 se'er fra det forventede !

Hvor ofte vil dette ske ?

Svar : "Tabelopslag" giver $0.6\% = 2 \times \mathbf{0.3}\%$ ← Fra forrige side

25

**Estimat:** $\widehat{RD} = 0.05317$ **Hypotese:** $RD$=0

**Teststørrelse:** $z = 2.83$ **P-værdi:** 0.06%

**Konklusion:**

**Hvis hypotesen er sand**, så er der kun 0.6% chance for at få et estimat, der ligger så lige så langt eller længere væk fra hypotesen end det vi har observeret.

Det er med andre ord **næsten usandsynligt** at observere det vi har set hvis hypotesen er sand.

Men *vi har jo* observeret det vi har observeret **ergo må hypotesen være falsk.**

Husk CI: (0.016;0.90) 0 **ligger ikke** i intervallet !

Overensstemmelse mellem test og sikkerhedsinterval !

26

**Estimat:** $\widehat{RD} = 0.05317$ **Hypotese:** $RD$=0.05

$z = (0.0532 - 0.05)/0.01882 = 0.167$

**Teststørrelse:** $z = 0.167$ **P-værdi:** 86% = 2×43%

**Konklusion:**

**Hvis hypotesen var sand**, så er der 86% chance for at få estimatet, der ligger så lige så langt eller længere væk fra hypotesen end det vi har observeret.

Data **strider** således **ikke** mod hypotesen.

Hypotesen **kan** akcepteres.

På basis af disse data kan vi **ikke afvise** at risikoen for hoste er 5% højere for børn, der har haft bronkitis !

Husk CI: (0.016;0.090) 0.05 **ligger** i intervallet !

Overensstemmelse mellem test og sikkerhedsinterval !

27

**Generelt**

Lad $\theta$ betegne den ukendte størrelse man ønsker at kende.

Den relevante statistiske analyse bør bestå af beregning af to tal :

$\hat{\theta}$ og $\text{se}(\hat{\theta})$

$\hat{\theta}$ : Et estimat af (gæt på) $\theta$

$\text{se}(\hat{\theta})$: Et estimat af (gæt på) usikkerheden af estimatet

Et **approksimativt** 95% sikkerhedsinterval : $\hat{\theta} \pm 1.96 \cdot \text{se}(\hat{\theta})$

Formlerne for estimatet og *se* afhænger af **den statistiske model** og kan være meget komplicerede.

I langt de fleste tilfælde bruges **computer programmer**.

28

**Generelt**

Hvis man er interesseret i **differensen** mellem to parametre:

$\delta = \theta_1 - \theta_2$

så er estimatet: $\hat{\delta} = \hat{\theta}_1 - \hat{\theta}_2$

Hvis to estimater $\hat{\theta}_1$ og $\hat{\theta}_2$ er **uafhængige** så er:

$$\text{se}(\hat{\delta}) = \sqrt{\text{se}(\hat{\theta}_1)^2 + \text{se}(\hat{\theta}_2)^2}$$

**HUSK!**
**Relative** størrelser som **Odds Ratio**, **Relative Risiko** og **Rate Ratio** skal analyseres på **log-skala** (LN).

29

Hoster om natten

| Bronkitis | Ja | Nej | Total |
|---|---|---|---|
| Ja | 26 | 247 | 273 |
| Nej | 44 | 1002 | 1046 |

**Associationsmål: relativ risiko**

$RR = \frac{\pi_{+B}}{\pi_{-B}}$ $\quad \widehat{RR} = \frac{\hat{\pi}_{+B}}{\hat{\pi}_{-B}} = \frac{0.09524}{0.04207} = 2.26385$

$\ln(\widehat{RR}) = \ln(2.26385) = 0.81707$

$$\text{se}(\ln(\widehat{RR})) = \sqrt{\frac{1}{26} - \frac{1}{273} + \frac{1}{44} - \frac{1}{1046}} = 0.23784$$

95% CI($\ln(RR)$): $0.81707 \pm 1.96 \cdot 0.23784 = (0.35089; 1.28324)$

95% CI($RR$): $(\exp(0.35089); \exp(1.28324)) = (1.42; 3.61)$

Formlerne kan findes på de sidste sider.

30

**Generelt: Et statistisk test**

**Data/estimat:** $\hat{\theta}$ med $\text{se}(\hat{\theta})$

**Hypotese:** $\theta = \theta_0$

**Beregn:** $z = \frac{\hat{\theta} - \theta_0}{\text{se}(\hat{\theta})}$

p-værdi $= 2 \cdot P(Z < -|z|)$ i standard normalfordeling

**Approksimativ**

**Konklusion:** Hvis p-værdien er **lille** er data ikke forenelig med hypotesen og hypotesen må **forkastes**.

Oftes sættes grænsen til 5%

**Bemærk:** Man kan bruge en anden se, når man tester, end den man bruger til beregning af CI (se Bland afsnit 8.6). Dette vil vi **ikke gøre** i dette kursus.

31

**Få data ⇒ dårlige approksimationer**

**Eksempel,** Streptomycin, Bland Table 13.7

15 personer deraf har 13 fået det bedre

Data kan antages at være **binomial**-fordelt.

$\hat{\pi} = \frac{13}{15} = 0.867$, $\text{se}(\hat{\pi}) = \sqrt{0.867 \cdot (1-0.867)/15} = 0.0878$

**Approks**. 95% CI: $0.867 \pm 1.96 \cdot 0.0878 = (0.695, 1.039)$

**Dårlig approksimation !** Ups!

**Eksakt/korrekt** 95% CI (findes vha. af tabel eller computer)
$(0.594, 0.983)$

**Morale**: Hvis der er få eller mange hændelser, så er approksimationerne ikke gode !

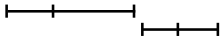
**Men:** For nogle modeller findes der **eksakte** metoder.

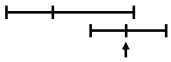
32

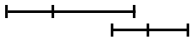
**Sikkerhedsintervaller og test.**

1 95%-sikkerhedsintervallet indeholder hypotesen hvis og kun hvis p-værdien er større end 5%.

**2** Ved sammenligning af **to** parametre baseret på to uafhængige data sæt, tre situationer:

A: Intet overlap:  så p-værdi < 5%

B: Et estimat i det andet CI:  så p-værdi >5%

Hverken A eller B:  så: p-værdi =?

33

Risiko for hoste om natten

| Bronkitis | Estimate | se | CI |
|---|---|---|---|
| Ja | 0.09524 | 0.01777 | 0.060; 0.130 |
| Nej | 0.04207 | 0.00621 | 0.030; 0.054 |
| Risiko Differens | 0.05317 | 0.01882 | 0.016; 0.090 |

Sammenligning af de to grupper:

| | | |
|---|---|---|
| 0 | **ikke med** i CI | p= 0.6% < 5% |
| 0.05 | **med** i CI | p= 86% > 5% |

De to sikkerhedsintervaller **overlapper ikke** p= 0.6% < 5%

34

**Associationsmål i 2x2 tabeller: Risiko differenser**

| | Status | | | *Sandsyn-lighed* |
|---|---|---|---|---|
| Population | 1 | 0 | | |
| 1 | *a* | *b* | $n_1$ | $\pi_1$ |
| 2 | *c* | *d* | $n_2$ | $\pi_2$ |

$\hat{\pi}_1 = \frac{a}{n_1}$ $\hat{\pi}_2 = \frac{c}{n_2}$ $\text{se}(\hat{\pi}_i) = \sqrt{\hat{\pi}_i(1-\hat{\pi}_i)/n_i}$

**R**isiko **D**ifferens: $RD = \pi_1 - \pi_2$

$$\widehat{RD} = \hat{\pi}_1 - \hat{\pi}_2 = \frac{a}{n_1} - \frac{c}{n_2}$$

$$\text{se}\left(\widehat{RD}\right) = \sqrt{\text{se}(\hat{\pi}_1)^2 + \text{se}(\hat{\pi}_2)^2} = \sqrt{\frac{a \cdot b}{n_1^3} + \frac{c \cdot d}{n_2^3}}$$

Bland p 130

35

**Eksempel: Bland side 130**

| Bronkitis som 5 årig | Hoster som 14 årig: Ja | Nej | Total | Obs. Risk |
|---|---|---|---|---|
| Ja. | 26 | 247 | 273 | 0.09524 |
| Nej | 44 | 1002 | 1046 | 0.04207 |

$$\widehat{RD} = 0.09524 - 0.04207 = 0.05317$$
$$\text{se}(\hat{\pi}_1) = \sqrt{0.09524 \cdot (1-0.09524)/273} = 0.01777$$
$$\text{se}(\hat{\pi}_2) = \sqrt{0.04207 \cdot (1-0.04207)/1046} = 0.00621$$
$$\text{se}\left(\widehat{RD}\right) = \sqrt{0.01777^2 + 0.00621^2} = 0.01882$$
$$= \sqrt{\frac{26 \cdot 247}{273^3} + \frac{44 \cdot 1002}{1046^3}} = 0.01882$$

95% CI(*RD*): $0.05317 \pm 1.96 \cdot 0.01882 = (0.01628; 0.09006)$

36

**Associationsmål i 2x2 tabeller: Relativ risiko**

| | Status | | | *Sandsyn-lighed* |
|---|---|---|---|---|
| Population | 1 | 0 | | |
| 1 | *a* | *b* | $n_1$ | $\pi_1$ |
| 2 | *c* | *d* | $n_2$ | $\pi_2$ |

**R**elativ **R**isiko: $RR = \frac{\pi_1}{\pi_2}$

$$\widehat{RR} = \frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{a \cdot n_2}{n_1 \cdot c}$$

$$\text{se}\left(\ln\left(\widehat{RR}\right)\right) = \sqrt{\frac{1}{a} - \frac{1}{n_1} + \frac{1}{c} - \frac{1}{n_2}}$$

Bland p 131

37

**Eksempel: Bland side 131**

| Bronkitis som 5 årig | Hoster som 14 årig: Ja | Nej | Total | Obs. Risk |
|---|---|---|---|---|
| Ja. | 26 | 247 | 273 | 0.09524 |
| Nej | 44 | 1002 | 1046 | 0.04207 |

$$\widehat{RR} = 0.09524/0.04207 = 2.26385$$

$$\ln\left(\widehat{RR}\right) = \ln(2.26385) = 0.81707$$

$$\text{se}\left(\ln\left(\widehat{RR}\right)\right) = \sqrt{\frac{1}{26} - \frac{1}{273} + \frac{1}{44} - \frac{1}{1046}} = 0.23784$$

95% CI($\ln(RR)$): $0.81707 \pm 1.96 \cdot 0.23784 = (0.35089; 1.28324)$

95% CI($RR$): $(\exp(0.35089); \exp(1.28324)) = (1.42; 3.61)$

38

**Associationsmål i 2x2 tabeller: Odds ratio**

| Population | Status 1 | 0 | | *Sandsynlighed* |
|---|---|---|---|---|
| 1 | $a$ | $b$ | $n_1$ | $\pi_1$ |
| 2 | $c$ | $d$ | $n_2$ | $\pi_2$ |

**O**dds **R**atio:

$$OR = \frac{\pi_1}{1-\pi_1} \Big/ \frac{\pi_2}{1-\pi_2} = \frac{\pi_1 \cdot (1-\pi_2)}{(1-\pi_1) \cdot \pi_2}$$

$$\widehat{OR} = \frac{\hat{\pi}_1}{1-\hat{\pi}_1} \Big/ \frac{\hat{\pi}_2}{1-\hat{\pi}_2} = \frac{a \cdot d}{b \cdot c}$$

$$\text{se}\left(\ln\left(\widehat{OR}\right)\right) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Bland p 240

39

**Eksempel: Bland side 240-241**

| Bronkitis som 5 årig | Hoster som 14 årig: Ja | Nej | Total | Odds |
|---|---|---|---|---|
| Ja. | 26 | 247 | 273 | 0.10526 |
| Nej | 44 | 1002 | 1046 | 0.04391 |

$$\widehat{OR} = \frac{26 \cdot 1002}{44 \cdot 247} = 2.39718$$

$$\ln\left(\widehat{OR}\right) = \ln(2.39718) = 0.87429$$

$$\text{se}\left(\ln\left(\widehat{OR}\right)\right) = \sqrt{\frac{1}{26} + \frac{1}{44} + \frac{1}{147} + \frac{1}{1002}} = 0.25736$$

95% CI($\ln(OR)$): $0.87429 \pm 1.96 \cdot 0.25736 = (0.36986; 1.37872)$

95% CI($OR$): $(\exp(0.360986); \exp(1.37872)) = (1.45; 3.97)$

40

**Sikkerhedsinterval for en enkelt rate**

| Events | Risikotid | Rate |
|---|---|---|
| $Y$ | $T$ | $IR$ |

$$\widehat{IR} = \frac{Y}{T} \qquad \text{se}\left(\ln\left(\widehat{IR}\right)\right) = \sqrt{\frac{1}{Y}}$$

41

**Eksempel: Analytisk epidemiologi side 86**

| Emigrations alder | Antal nye tilfælde | Risikotid (år) | Rate (antal per 100 000 år) |
|---|---|---|---|
| <15 år | 4 | 530 999 | 0.75330 |

$$\widehat{IR} = \frac{4}{530999\text{år}} = 0.75330 / {}_{100000\text{år}}$$

$$\ln\left(\widehat{IR}\right) = \ln(0.75330) = -0.28330$$

$$\text{se}\left(\ln\left(\widehat{IR}\right)\right) = \sqrt{\frac{1}{4}} = 0.50$$

95% CI($\ln(IR)$): $-0.28330 \pm 1.96 \cdot 0.50 = (-1.26330; 0.69670)$

95% CI($IR$): $(\exp(-1.26330); \exp(0.69670)) = (0.28; 2.01) / {}_{100000\text{år}}$

42

**Sammenligning af to rater: Rate ratio**

| Population | Events | Risikotid | Rate |
|---|---|---|---|
| 1 | $Y_1$ | $T_1$ | $IR_1$ |
| 2 | $Y_2$ | $T_2$ | $IR_2$ |

**I**ncidence **R**ate **R**atio

$$IRR = \frac{IR_1}{IR_2}$$

$$\widehat{IRR} = \frac{\widehat{IR}_1}{\widehat{IR}_2} = \frac{Y_1 \cdot T_2}{T_1 \cdot Y_2}$$

$$\text{se}\left(\ln\left(\widehat{IRR}\right)\right) = \sqrt{\frac{1}{Y_1} + \frac{1}{Y_2}}$$

43

**Eksempel: Analytisk epidemiologi side 86**

| Emigrations alder | Antal nye tilfælde | Risikotid (år) | Rate (antal per 100 000 år) |
|---|---|---|---|
| <15 år | 4 | 530 999 | 0.75330 |
| 15-29 år | 28 | 790 000 | 3.54430 |

$$\widehat{IRR} = \frac{28 \cdot 530999}{4 \cdot 790000} = \frac{3.54430}{0.75330} = 4.70505$$

$$\ln\left(\widehat{IRR}\right) = \ln(4.70505) = 1.54864$$

$$\text{se}\left(\ln\left(\widehat{IRR}\right)\right) = \sqrt{\frac{1}{4} + \frac{1}{28}} = 0.53452$$

95% CI($\ln(IRR)$): $1.54864 \pm 1.96 \cdot 0.53452 = (0.50097; 2.59630)$

95% CI($IRR$): $(\exp(0.50097); \exp(2.59630)) = (1.65; 13.41)$

44

**Sammenligning af to rater: Rate differens**

| Population | Events | Risikotid | Rate |
|---|---|---|---|
| 1 | $Y_1$ | $T_1$ | $IR_1$ |
| 2 | $Y_2$ | $T_2$ | $IR_2$ |

**I**ncidens **R**ate **D**ifferens

$$IRD = IR_1 - IR_2$$

$$\widehat{IRD} = \widehat{IR}_1 - \widehat{IR}_2 = \frac{Y_1}{T_1} - \frac{Y_2}{T_2}$$

$$\text{se}\left(\widehat{IRD}\right) = \sqrt{\frac{Y_1}{T_1^2} + \frac{Y_2}{T_2^2}}$$

45

**Eksempel: Analytisk epidemiologi side 86**

| Emigrations alder | Antal nye tilfælde | Risikotid (år) | Rate (antal per 100 000 år) |
|---|---|---|---|
| <15 år | 4 | 530 999 | 0.75330 |
| 15-29 år | 28 | 790 000 | 3.54430 |

$$\widehat{IRD} = (3.54430 - 0.75330) / {}_{100000\text{år}} = 2.79101 / {}_{100000\text{år}}$$

$$\text{se}\left(\widehat{IRD}\right) = \sqrt{\frac{4}{530999\text{år}^2} + \frac{28}{790000\text{år}^2}}$$

$$= \sqrt{\frac{4}{5.30999^2} + \frac{28}{7.90000^2}} / {}_{100000\text{år}}$$

$$= 0.76845 / {}_{100000\text{år}}$$

95% CI($IRD$): $2.79101 \pm 1.96 \cdot 0.76845 = (1.28; 4.30) / {}_{100000\text{år}}$

1

Epidemiologi og **biostatistik**.
Uge 2, torsdag d. 13. februar 2003
Morten Frydenberg, Institut for Biostatistik.

**Type 1 og type 2 fejl**

**Statistisk styrke**

**Nogle specielle metoder:**

- **Test i RxC** tabeller
- **Test i 2x2** tabeller
  **Fishers** eksakte test
- **Normalfordelte** data : **t-test**
  **eksakte** sikkerhedsintervaller

2

**Resumé:**

En statistisk analyse resulterer ofte i :

Et estimat $\hat{\theta}$ med en tilhørende $\text{se}(\hat{\theta})$

for den ukendte størrelse, $\theta$,som man er interesseret i.

Et **approksimativt** 95% sikkerhedsinterval :

$$\hat{\theta} \pm 1.96 \cdot \text{se}(\hat{\theta})$$

En specifik **hypotese** om at $\theta = \theta_0$ kan testes ved

$$z = \frac{\hat{\theta} - \theta_0}{\text{se}(\hat{\theta})} \text{ eller } z^2 = \left( \frac{\hat{\theta} - \theta_0}{\text{se}(\hat{\theta})} \right)^2$$

Den vender vi tilbage til !

Store værdier af $|z|$ (eller $z^2$) er kritiske!

p-værdi via standard normalfordeling eller $\chi^2(1)$ -fordeling
Approksimation

3

**Nogle statistiske begreber**

**Type 1 fejl:** At **forkaste** hypotesen, selvom den er **sand**.

**Type 2 fejl:** At **acceptere** hypotesen, selvom den er **falsk**.

**Signifikansniveau:** Den grænse man sætter for den største p-værdi, der leder til, at man forkaster hypotesen.

Som regel sættes **signifikansniveauet til 5%.**

**Hvis hypotesen er sand:**
Sandsynligheden for **type 1 fejl**
=sandsynligheden for **forkaste hypotesen**
=**signifikansniveauet**

M.a.o. **sandsynligheden for type 1 fejl** er kendt og lig signifikansniveauet **(=5%).**

4

ikke forkaste

**Type 2 fejl:** At **acceptere** hypotesen, selvom den er **falsk**.

Hvad er **sandsynligheden for type 2 fejl** ?

**Afhænger af:**

Hvad der så er **sandt** !

**Informationsmængden** !

Sandheden **langt fra** hypotesen ⇒ **lille** ss. for type 2 fejl

Sandheden **tæt på** hypotesen ⇒ **stor** ss. for type 2 fejl

**Meget** information/data ⇒ **lille** ss. for type 2 fejl

**Lidt** information/data ⇒ **stor** ss. for type 2 fejl

**Statistisk styrke = 1 - sandsynlighed for type 2 fejl**

**= sandsynlighed for at forkaste den falske hypotese**

5

**Styrkeovervejelser i forbindelse med planlægning af et studie.**

Planlægning af et **follow-up** studie:

Antagelser:

KIP blandt ikke eksponerede = **1%.**

**Sand relativ risiko** = **2.0 .**

**1500** eksponerede og **1500** ikke eksponerede.

Når data er indsamlet vil man teste hypotese **RR=1.**

Sandsynligheden for at få data, der leder til **accept af dette (Type 2 fejl) = 39%,** dvs. en styrke på 61 %.

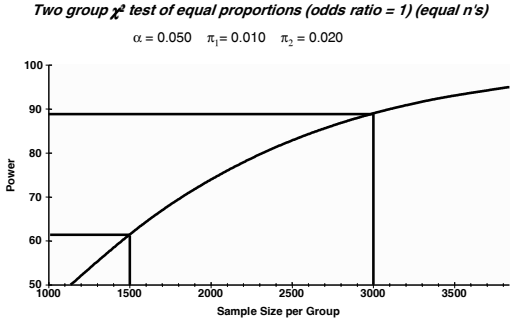
M.a.o. **lille chance** for at få bekræftet, at der en sammenhæng.

**Ikke besværet værd !**

6

**Øges** deltagerantallet til **2*3000** bliver chancen for type 2 fejl reduceret til **11%**, dvs. styrken er 89%.

**Styrken som funktion af gruppestørrelsen :**



7

**Statistisk styrke**
**Nogle kommentarer**

- Afhænger af **designet.**
- Afhænger af **statistisk metode.**
- Relevant i **planlægningsfasen.**
- Når data **er indsamlet** er bredden af sikkerhedsintervaller udtryk for informationsmængden.

8

**Test i RxC tabeller**

Bland table 13.1. **Boligform og for tidlig fødsel** :

| Housing tenure | Preterm | Term | Total |
|---|---|---|---|
| Owner-occupier | 50 | 849 | 899 |
| Council tenant | 29 | 229 | 258 |
| Private tenant | 11 | 164 | 175 |
| Lives with parents | 6 | 66 | 72 |
| Other | 3 | 36 | 39 |
| Total | 99 | 1344 | 1443 |

**Hypotese**: Ingen sammenhæng.

Hvis denne er **sand** bliver det **forventede** antal **preterm** fødsler blandt de, der bor i **egen bolig**:

$$\frac{99}{1443} \cdot 899 = 61.7$$

9

**Test i RxC tabeller**

**Forventet** under hvis hypotesen er sand:

| Housing tenure | Preterm | Term | Total |
|---|---|---|---|
| Owner-occupier | 61.7 | 837.3 | 899 |
| Council tenant | 17.7 | 240.3 | 258 |
| Private tenant | 12.0 | 163.0 | 175 |
| Lives with parents | 4.9 | 67.1 | 72 |
| Other | 2.7 | 36.3 | 39 |
| Total | 99.0 | 1344.0 | 1443 |

Et mål for **forskel** mellem **observeret** og **forventet**:

$$X^2 = \sum_{\text{alle celler}} \frac{(\text{observeret} - \text{forventet})^2}{\text{forventet}}$$

Er **stor** ved **dårlig** overensstemmelse ! $X^2 = 10.5$

10

Vi har fået $X^2$=10.5

Hvor ofte vil man få noget større ?

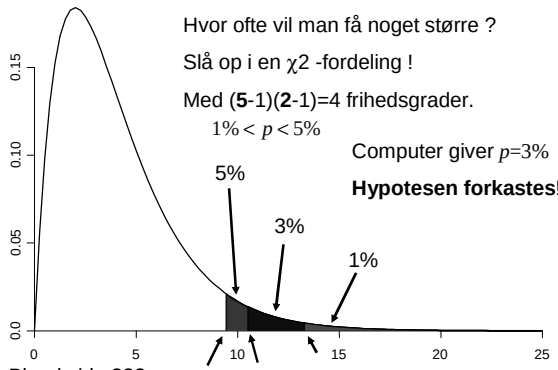
Slå op i en $\chi 2$ -fordeling !

Med (**5**-1)(**2**-1)=4 frihedsgrader.

$1\% < p < 5\%$

Computer giver $p$=3%

**Hypotesen forkastes**!




Bland side 233

11

**Test for ingen association i RxC tabeller**
**Generelt**

Hypotese:
**Ingen sammenhæng** melllem de to inddelingskriterier

$$\text{forventet} = \frac{\text{rækkesum} \times \text{søjlesum}}{\text{total}}$$

$$X^2 = \sum_{\text{alle celler}} \frac{(\text{observeret} - \text{forventet})^2}{\text{forventet}}$$

En **stor værdi** af $X^2$ er **kritisk**.

p-værdi findes i en $\chi^2$ **- fordeling** med **(R-1)(C-1)** frihedsgrader.

12

**Test for ingen association i 2×2 tabeller**

| Svangerskabs-længde | Køn: Dreng | Køn: Pige | Total |
|---|---|---|---|
| 38 | 316 | 260 | 576 |
| 40 | 1040 | 991 | 2031 |
| Total | 1356 | 1251 | 2607 |

**Hypotese**:

***Ingen sammenhæng mellem køn og svangerskabslængde***

Teststørrelsen kan let beregnes i hånden som:

$$X^2 = \frac{(316 \cdot 991 - 260 \cdot 1040)^2 \cdot 2607}{576 \cdot 2031 \cdot 1356 \cdot 1251} = 2.40 < 3.84$$

13

Vi har fået $X^2$=2.4

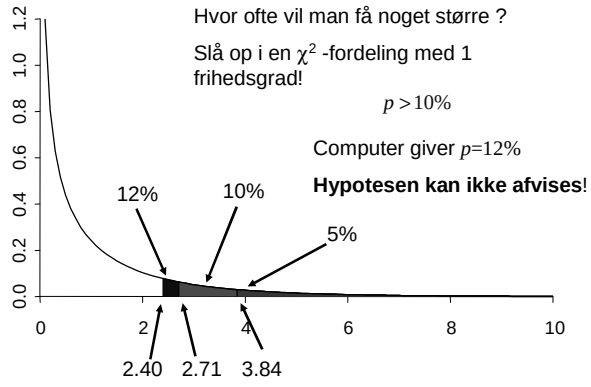
Hvor ofte vil man få noget større ?

Slå op i en $\chi^2$ -fordeling med 1 frihedsgrad!

$p > 10\%$

Computer giver $p$=12%

**Hypotesen kan ikke afvises**!




14

**2×2 tabeller**

| | Status | | |
|---|---|---|---|
| Population | 1 | 0 | |
| 1 | $a$ | $b$ | $n_1$ |
| 2 | $c$ | $d$ | $n_2$ |
| | $s_1$ | $s_0$ | $N$ |

**Hypotese:** "Ingen association"

**Test:**

$$X^2 = \frac{(a \cdot d - b \cdot c)^2 \cdot N}{n_1 \cdot n_2 \cdot s_1 \cdot s_0}$$

Slåes op i en $\chi^2$-fordeling med 1 frihedsgrad.

15

**2×2 tabeller : Fishers eksakte test**

**Amning og tandstilling:**

**Hypotese:**
Ingen sammenhæng

| | Problemer med tandstilling | | |
|---|---|---|---|
| Amning | Nej | Ja | Sum |
| Bryst | 4 | 16 | 20 |
| Flaske | 1 | 21 | 22 |
| Sum | 5 | 37 | 42 |

**For få data** til at approksimationer kan bruges !

**Løsning:** Fishers eksakte test (computer).

Resultat (kun) **en p-værdi** !

Her: p-værdi=29%

**Konklusion:** Data strider ikke mod : "Ingen sammenhæng"

16

**Kommentarer til test for ingen association i 2×2 tabeller**

- Hvis der er **5 eller mindre** i en af cellerne, så bør man bruge **Fisher's eksakte test** .
- Nogle anvender et **kontinuitets** (eller **Yates** ) korrigeret version af $X^2$ -testet:

$$X_C^2 = \frac{\left(|a \cdot d - b \cdot c| - N/2\right)^2 \cdot N}{n_1 \cdot n_2 \cdot s_1 \cdot s_0}$$

Det giver lidt større p-værdier.

Der er mange argumenter for og imod dette valg.

**Brug jeres tid på noget mere fornuftigt !!!**

17

**Eksakt analyse af normalfordelte data**

Lungefunktions data fra i tirsdag i uge 1:

| Køn | $n$ | Gennemsnit l/min | sd l/min | sem l/min |
|---|---|---|---|---|
| Kvinder | 14 | 485.6 | 46.6 | 12.5 |
| Mænd | 16 | 552.9 | 55.0 | 13.8 |

**Approksimativt** $\text{CI}(\mu_K): 485.6 \pm 1.96 \cdot 12.5 = (461; 510)$

Stort set det samme

Under antagelse af normalfordeling :

**Eksakt** 95% CI for $\mu_K$: $485.6 \pm 2.16 \cdot 12.5 = (459; 513)$

Hvor kommer de **2.16** fra ?

Fra **t-fordelingen** !!

18

Tabel over tosidige halesandsynligheder i t-fordelingen

Bland side 158

| df | 10% | 5% | 1% | 0.10% | df | 10% | 5% | 1% | 0.10% |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 6.31 | 12.71 | 63.66 | 636.62 | 16 | 1.75 | 2.12 | 2.92 | 4.01 |
| 2 | 2.92 | 4.30 | 9.93 | 31.60 | 17 | 1.74 | 2.11 | 2.90 | 3.97 |
| 3 | 2.35 | 3.18 | 5.84 | 12.92 | 18 | 1.73 | 2.10 | 2.88 | 3.92 |
| 4 | 2.13 | 2.78 | 4.60 | 8.61 | 19 | 1.73 | 2.09 | 2.86 | 3.88 |
| 5 | 2.02 | 2.57 | 4.03 | 6.87 | 20 | 1.72 | 2.09 | 2.85 | 3.85 |
| 6 | 1.94 | 2.45 | 3.71 | 5.96 | 21 | 1.72 | 2.08 | 2.83 | 3.82 |
| 7 | 1.89 | 2.36 | 3.50 | 5.41 | 22 | 1.72 | 2.07 | 2.82 | 3.79 |
| 8 | 1.86 | 2.31 | 3.36 | 5.04 | 23 | 1.71 | 2.07 | 2.81 | 3.77 |
| 9 | 1.83 | 2.26 | 3.25 | 4.78 | 24 | 1.71 | 2.06 | 2.80 | 3.75 |
| 10 | 1.81 | 2.23 | 3.17 | 4.59 | 25 | 1.71 | 2.06 | 2.79 | 3.73 |
| 11 | 1.80 | 2.20 | 3.11 | 4.44 | 30 | 1.70 | 2.04 | 2.75 | 3.65 |
| 12 | 1.78 | 2.18 | 3.05 | 4.32 | 40 | 1.68 | 2.02 | 2.70 | 3.55 |
| **13** | 1.77 | **2.16** | 3.01 | 4.22 | 60 | 1.67 | 2.00 | 2.66 | 3.46 |
| 14 | 1.76 | 2.14 | 2.98 | 4.14 | 120 | 1.66 | 1.98 | 2.62 | 3.37 |
| 15 | 1.75 | 2.13 | 2.95 | 4.07 | Uendelig | 1.64 | **1.96** | 2.58 | 3.29 |

95%=(100-5)%

$n$–1=14–1=13 *frihedsgrader* (**d**egrees of **f**reedom)

t=2.16

Uendelig mange frihedsgrader = Standard normalfordeling

19

**Eksakt analyse af normalfordelte data**
**Sikkerhedsinterval**

**Model/antagelse:**

Data er $n$ **uafhængige** observationer fra en normalfordeling med **ukendt** middelværdi, $\mu$, og spredning, $\sigma$:

**Estimaterne** for disse er :

$$\hat{\mu} = \bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \qquad \hat{\sigma} = \mathrm{sd} = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\mathrm{sem} = \mathrm{se}(\hat{\mu}) = \mathrm{se}(\bar{x}) = \hat{\sigma} / \sqrt{n} = \mathrm{sd} / \sqrt{n}$$

Et **eksakt** CI for $\mu$: $\bar{x} \pm t_{n-1} \cdot \mathrm{sem}$

$t_{n-1}$ findes i en tabel over **t-fordelingen**

20

**Eksakt analyse af normalfordelte data**
**One sample t-test**

**Hypotese :** $\mu = \mu_0$

**Test :** $z = \dfrac{\bar{x} - \mu_0}{\mathrm{sem}}$

**p-værdi:** Slå op i en **t-fordeling** med $n$-1 frihedsgrader

**(ikke i en standard normalfordeling)**

**PEFR-eksemplet :**

**Hypotese:** Middel PEFR hos kvinder er 500 l/min

$$z = \frac{485.6 - 500}{12.5} = -1.16$$

**Eksakt** p-værdi >10%
(computer p=26.8%)

**Konklusion:** Data strider ikke mod hypotesen.

21

**Eksakt analyse af to sæt (uafhængige) normalfordelte data**

| Køn | $n$ | Gennemsnit | sd | se |
|---|---|---|---|---|
| Kvinder | 14 | 485.6 | 46.6 | 12.5 |
| Mænd | 16 | 552.9 | 55.0 | 13.8 |

Estimat for spredningen blandt mænd

Estimat for spredningen blandt kvinder

Et **fælles** estimat for spredningen :

$$\mathrm{sd}_F = \sqrt{\frac{(n_K - 1) \cdot \mathrm{sd}_K{}^2 + (n_K - 1) \cdot \mathrm{sd}_M{}^2}{n_K + n_M - 2}}$$
$$= \sqrt{\frac{(14-1) \cdot 46.6^2 + (16-1) \cdot 55.0^2}{14 + 16 - 2}}$$
$$= 51.3 \text{ l/min}$$

22

Estimat for fælles spredning: $\mathrm{sd}_F = 51.3$

**Nyt bud** på sem'erne:

$$\mathrm{sem}_K = \mathrm{sd}_F / \sqrt{n_K} = 51.3 / \sqrt{14} = 13.7 \text{ l/min}$$
$$\mathrm{sem}_M = \mathrm{sd}_F / \sqrt{n_M} = 51.3 / \sqrt{16} = 12.8 \text{ l/min}$$

| Køn | $n$ | Gennemsnit | sd | sem | sem(fælles) |
|---|---|---|---|---|---|
| Kvinder | 14 | 485.6 | 46.6 | 12.5 | 13.7 |
| Mænd | 16 | 552.9 | 55.0 | 13.8 | 12.8 |

$$\mathrm{se}_F(\hat{\mu}_M - \hat{\mu}_K) = \sqrt{\mathrm{sem}_M^2 + \mathrm{sem}_K^2} = \sqrt{12.8^2 + 13.7^2} = 18.8 \text{ l/min}$$

95% **eksakt** CI for forskel i middel PEFR, $\mu_M - \mu_K$:

$$(\hat{\mu}_M - \hat{\mu}_K) \pm t \cdot \mathrm{se}(\hat{\mu}_M - \hat{\mu}_K)$$
$$= (552.9 - 485.6) \pm 2.05 \cdot 18.8 = (29; 106) \text{ l/min}$$

Fra t-fordeling med $n_M + n_K - 2 = 28$ frihedsgrader

23

**Analyse af to sæt (uafhængige) normalfordelte data**
**Two sample t-test**

**Hypotese:** $\mu_M - \mu_K = \delta_0$

$$z = \frac{(\hat{\mu}_M - \hat{\mu}_K) - \delta_0}{\mathrm{se}_F(\hat{\mu}_M - \hat{\mu}_K)}$$

**p-værdi:** Slå op i en **t-fordeling** med $n_M + n_K - 2$ frihedsgrader

**(ikke i en standard normalfordeling)**

**PEFR-eksemplet :**

**Hypotese:** Forskel i middel PEFR er 0 l/min.

$$z = \frac{(552.9 - 485.6) - 0}{18.8} = \frac{67.3 - 0}{18.8} = 3.59$$

**Eksakt** p-værdi=0.1%

**Konklusion:** Data strider mod hypotesen.

24

**Kommentarer**

Hvis **antagelsen** om normalfordeling er rimelige :

- Fordelingen kan beskrives ved blot to tal : **Middelværdi** og **spredning** !
- **Eksakte** CI og p-værdier - **ingen approksimationer** !
- Også mulighed for at **sammenligne spredninger** (dækkes ikke på dette kursus)
- Mere komplicerede modeller og analyse metoder :
  - **Variansanalyse** (**ANOVA)**
  - **Lineær regressionsmodeller**
  - **Ikke-lineær regressionsmodeller**
  - **Faktoranalyse**
  - **+meget mere**

25

**Flere kommentarer**

- Metoderne til analyse af en stikprøve fra en normalfordeling bruges ofte hvis man har **parrede** data:

  To målinger per patient, før/efter behandling.

  Beregn efter-før=*obs. Behandlingseffekt.*

  **Hvis** disse kan antages at være normalfordelte,

  **så** analyse som en stikprøve fra en normalfordeling.

  Dette kaldes **Parret t-test.**

- Hvordan checker man antagelsen om normalfordeling ?

  **Plot** data - histogrammer, normal plots (Q-Q plots).

  Hvad siger **erfaringen** om tilsvarende data ?

26

**En sidste kommentar til analyse vha. af t-fordelingen**

- Det er kun hvis man har **små stikprøver** at denne metode giver noget væsentligt andet end den sædvanlige/approksimative metode.
- Metoden er meget udbredt, men vi vil kun **undtagelsesvis** bruge den i dette kursus!

27

**Komponenter i middelværdi og variation**

**Altid** mindst to komponenter i middelværdi og variation:

Disse skyldes egenskaber ved

**”populationen”**

**”målemetoden”**

| | | |
|---|---|---|
| **Middelværdi** | **=** | **Middelværdi i populationen + Systematisk målefejl** |
| **Variation** | **=** | **Variation i populationen + Tilfældig målefejl** |

Epidemiologi og **biostatistik**.
Uge 3, torsdag.
Erik Parner, Institut for Biostatistik.

**Lineær regressionsanalyse**

- Simpel lineær regression
- Multipel lineær regression



## Regressionsanalyse

**Regressionsanalyser** bruges til

- Beskrive sammenhængen mellem to variable.
  Eks: Kvantificere sammenhængen mellem blodtryk og alder.
- Prædiktere værdien af en variabel hvis værdien af én eller flere andre variable er kendt (referencemodel).
  Eks: Forudsige blodtrykket for en 50 årig person.
- Korrektion for potentielle confoundere.
  Eks: Hvad er alderseffekten på blodtrykket korrigeret for BMI?

Den **lineære regressionsanalyse** kan anvendes når responsen er **kontinuert**.



## Eksempel: Systolisk blodtryk

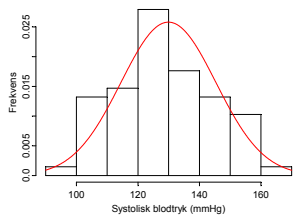
**Udgangspunkt:** Vi ønsker at prædiktere det systoliske blodtryk hos en gruppe af personer.

**Data:** Systolisk blodtryk-målinger and andre baggrunds-variable for 68 personer.

| $i$ | $y_i$ | $x_i$ |
|---|---|---|
| Obs. no. | Syst. blodtryk | Alder |
| 1 | 155 | 45 |
| 2 | 134 | 55 |
| 3 | 135 | 46 |
| . | . | . |
| 68 | 140 | 48 |



## Prædiktionsinterval




$\bar{y} = 129.9,\ sd_{\text{Total}} = 15.5$
$(n = 68)$

Hvis vi antager blodtryk er normalfordelt fås

PI: $129.9 \pm 1.96 \cdot 15.5$
$=(99.6; 160.2)$ mmHg

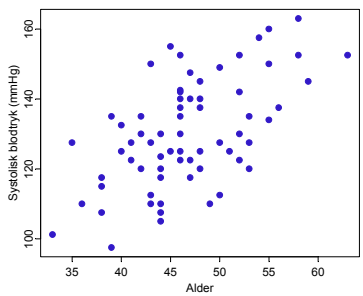
**Fortolkning:** Personernes systoliske blodtryk er mellem 99.6 og 160.2 mmHg.

Bemærk: Vores bedste bud på en persons systoliske blodtryk er altså intervallet (99.6;160.2) mmHg.

Der er dog relativt stor variation i det systoliske blodtryk!



Vil vores bud på personens systoliske blodtryk afhænge af persones alder?



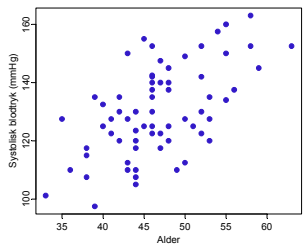

Ja, yngre personer har et lavere blodtryk end ældre personer!

Vi kan lave et mere præcist prædiktionsinterval, hvis vi bruger oplysningen om personens alder.



Én løsning er, at inddele i aldersgrupper og beregne prædiktionsintervaller indenfor hver aldersgrupper.

En anden løsning er en **regressionsanalyse,** hvor personens præcise alder inddrages.




En regressionsmodel er en model for sammenhængen mellem blodtryk og alder.

Der ser ud til at være en **lineær sammenhæng** mellem blodtryk og alder.

## Simpel lineær regression

En simpel lineær afhængighed mellem $y_i$ og $x_i$:

$$y_i = \underbrace{\alpha + \beta \cdot x_i}_{\text{Formlen for en ret linie!}} + \underbrace{E_i}_{\text{Beskriver afvigelsen fra linien.}}$$

Variablen $E_i$ beskriver den tilfældige/uforklarede variation omkring linien, og antages at have middelværdi $0$ og spredning $\sigma_{Res}$ (Res=Residual).

En simpel lineær regressionsmodel har tre parametre:

- $\alpha$ = afskæringen med y-aksen (**intercept**)
- $\beta$ = hældningen (**regressionskoefficient**)
- $\sigma_{Res}$ = et mål for variationen omkring linien.



## Terminologi:

$y$ = responsvariabel = afhængige variabel
= Systolisk blodtryk

$x$ = uafhængig variabel=forklarende variabel
= Alder

## Fortolkning af parametrene:

$\beta$ er forskellen i middel systolisk blodtryk mellem to personer med en aldersforskel på 1 år.

(Fortolkningen er **ikke** den forventede stigning i det systoliske blodtryk når man bliver et år ældre!)

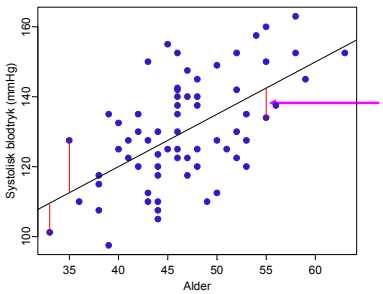
$\alpha$ har i denne situation ingen fornuftig fortolkning.

(Middel blodtrykket for en 0 år gammel person?)

$\sigma_{Res}$ et mål for variationen omkring linien.



## Estimation af $\alpha$, $\beta$ og $\sigma_{Res}$ :




$$r_i = y_i - (\hat{\alpha} - \hat{\beta} \cdot x_i)$$

**Residual**
= afvigelsen af observationen fra linien.

Regressionslinien bestemmes ved **mindste kvadrates metode**, der minimerer (kvadratet på) afstandene fra observationerne til linien.

$\sigma_{Res}$ estimeres ved standard deviationen af residualerne.



Estimation af $\alpha$, $\beta$ og $\sigma$ og se'er m.v. er kompliceret, men kan laves af de fleste statistikprogrampakker.

**Resultat:**

| | Estimat | se | CI | |
|---|---|---|---|---|
| Intercept | 60.3 | 12.0 | (36.3;84.2) | (mmHg) |
| Regression | 1.5 | 0.3 | (1.0;2.0) | (mmHg/år) |
| $sd_{Res}$ | 12.6 | | | (mmHg) |

Regressionsanalysen beskriver sammenhængen mellem (systolisk) Blodtryk og Alder som

$$\text{middel Blodtryk} = 60.3 + 1.5 \cdot \text{Alder}$$

**Eksempel 1:** Middelblodtrykket for 50 årige personer er

$60.3 + 1.5 \cdot 50 = 135.0$ mmHg.

CI kan vi ikke udregne på basis af ovenstående tal!



**Eksempel 2:** Forskellen i middelblodtryk for 40 årige personer og 50 årige personer er

$$\text{Forskel} = (\hat{\alpha} + \hat{\beta} \cdot 50) - (\hat{\alpha} + \hat{\beta} \cdot 40)$$
$$= \hat{\beta} \cdot (50 - 40) = 1.5 \cdot 10 = 14.9 \text{ mmHg}$$

$$\text{se}(10 \cdot \hat{\beta}) = 10 \cdot \text{se}(\hat{\beta})$$
$$= 10 \cdot 0.26 = 2.6 \text{ mmHg}$$

$$\text{CI(Forskel)}: \; 14.9 \pm 1.96 \cdot 2.6 = (9.9; 20.0) \text{ mmHg}$$

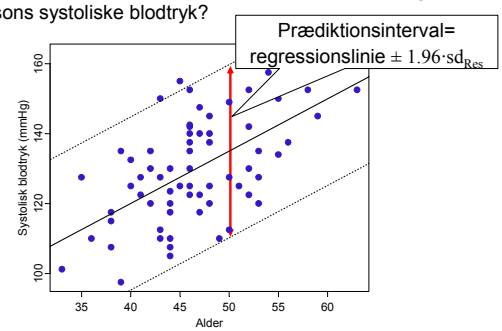
Middelforskellen mellem to personer med en aldersforskel på 10 år er mellem 9.9 og 20.0 mmHg.

Generelt: Forskellen i middelblodtryk mellem 2 personer med en aldersforskel på $\Delta$ år er

$$\text{Forskel} = \Delta \cdot \hat{\beta}, \quad \text{se}(\Delta \cdot \hat{\beta}) = |\Delta| \cdot \text{se}(\hat{\beta})$$



**Eksempel 3:** Hvad er vores bedste bud på en 50 årig persons systoliske blodtryk?




$$\text{PI}(x) = (\hat{\alpha} + \hat{\beta} \cdot x) \pm 1.96 \cdot \text{sd}_{Res}$$

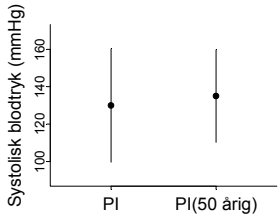
Prædiktionsinterval for de 50 årige personer bliver således

Middelblodtryk: $\hat{\alpha} + \hat{\beta} \cdot 50 = 135.0$ mmHg

$sd_{Res} = 12.6$ mmHg

PI(50 årige): $135.0 \pm 1.96 \cdot 12.6 = (110.2; 159.8)$ mmHg

Det generelle prædiktionsinterval (uden hensyntagen til alder) var PI: (99.6;160.2) mmHg.






**Andel forklaret variation**

Prædiktionsintervallet fra regressionsanalysen er smallere end det generelle prædiktionsinterval
($sd_{Res}$ er mindre end $sd_{Total}$).

Vi har **forklaret** noget af variationen i Blodtryk ved variationen i Alder. Men hvor meget?

Den relative reduktion i variationen er

$$R^2 = \left(15.5^2 - 12.6^2\right) \Big/ 15.5^2$$
$$= 0.34 = 34\%$$

Vi har således forklaret 34% af variationen i blodtryk ved variationen i alderen.

$R^2$ = andel forklaret variation af den totale variation (coefficient of determination).



**Antagelser bag den simple lineære regressionsanalyse**

Den **statistiske model** bygger på følgende antagelser:

- **Uafhængige** par af observationer $(x_1,y_1),\ldots,(x_n,y_n)$.
- **Lineær sammenhæng** mellem $x_i$ og $y_i$ :

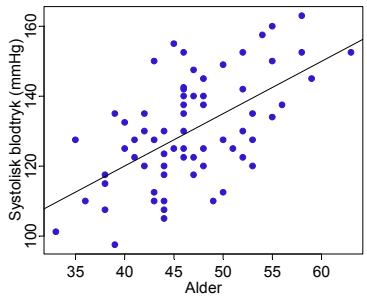
$$y_i = \alpha + \beta \cdot x_i + E_i$$

- Variationen omkring linien, $E_i$, er **normalfordel** med middelværdi 0 og spredning $\sigma_{Res}$.

Variationen omkring linien afhænger **ikke** af den forklarende variabel $x_i$



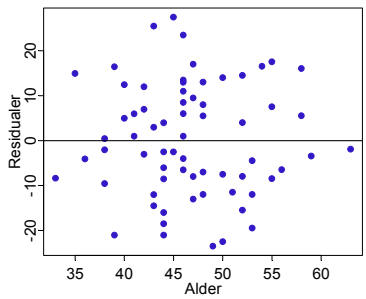
**Modelkontrol: lineær sammenhæng**




Det ser ud til, at den lineære sammenhæng er en rimelig beskrivelse!



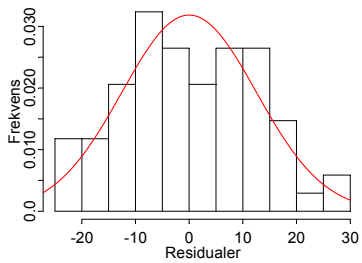
**Modelkontrol: konstant variation**




Residualerne viser symmetri omkring 0 og konstant variation uafhængig af Alder.



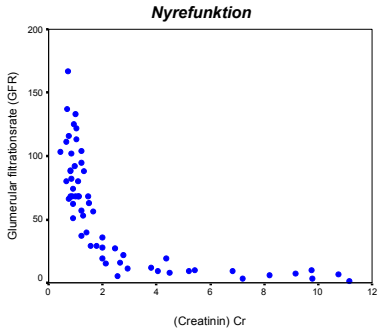
**Modelkontrol: normalfordeling**



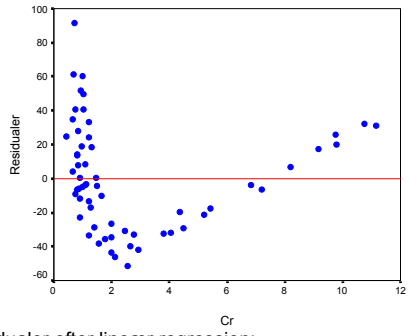

Residualerne kan antages at være normalfordelt!

**Antagelserne bag den lineære regressionsanalyse synes at være opfyldt !**

## Eksempel på en ikke-lineær sammenhæng





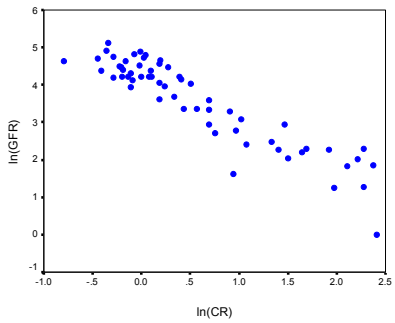




Residualer efter lineær regression:

- mangel på symmetri / systematisk afvigelser fra 0.
- ikke konstant variation.



Ln-transformation af nyrefunktion:




Her er antagelserne bag regressionsanlysen opfyldt.



## Hypoteser omkring $\beta$

Foregår som sædvanlig!

Hvis vi f.eks. ønsker at teste

**Hypotese:** $\beta = 0$ (ingen sammenhæng mellem Blodtryk og Alder)

$$z = \frac{\hat{\beta} - 0}{\text{se}(\hat{\beta})} = \frac{1.2 - 0}{0.2} = 5.1, \qquad p < 0.001$$



## Multipel lineær regression

"Effekten" af alder er beskrevet ved hældningen (fra tidligere)

$\hat{\beta} = 1.5$ mmHg/år (CI: $1.0 - 2.0$)

Hældningen beskriver middelforskellen i systolisk blodtryk mellem to personer med en aldersforskel på 1 år.

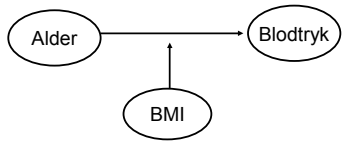
Blodtrykket afhænger også af BMI.

Afhænger alderseffekten af personens BMI?
Mao. er BMI en **effektmodifikator** for alderseffekten?

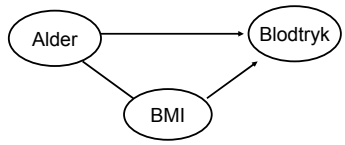
Hvis BMI **ikke** er en effektmodifikator for alderseffekten:
Er BMI en **confounder** for alderseffekten?



## Effektmodifikator?




## Confounder?



**Data:** Samme data fra før, nu suppleret med BMI oplysninger.

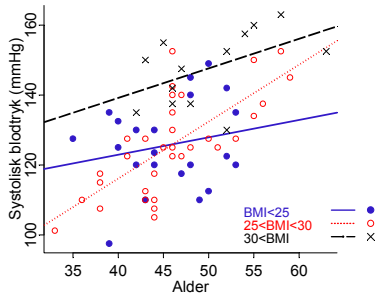
| Obs. no. | Syst. blodtryk | Alder | BMI | BMI gruppe |
|---|---|---|---|---|
| 1 | 155 | 45 | 43.6 | 3 |
| 2 | 134 | 55 | 25.3 | 2 |
| 3 | 135 | 46 | 27.9 | 2 |
| . | . | . | . | . |
| 68 | 140 | 48 | 23.0 | 1 |

BMI er inddelt i 3 grupper:

BMI gruppe = 1 hvis BMI $\leq$ 25
= 2 hvis 25 < BMI $\leq$ 30
= 3 hvis 30 < BMI



**En regressionsanalyse for hver BMI gruppe:**




Er effekten af alderen den samme i de 3 BMI grupper?



**Er BMI en effektmodifikator?**

| | BMI | Hældning | CI |
|---|---|---|---|
| Strata | <25 | 0.5 | (-0.6;1.6) |
| | 25-30 | 1.6 | (1.1;2.2) |
| | <30 | 0.8 | (-0.1;1.8) |

Estimaterne er noget usikre!

**Hypotese:** Samme alderseffekt i de 3 BMI grupper
(BMI er ikke en effektmodifikator)

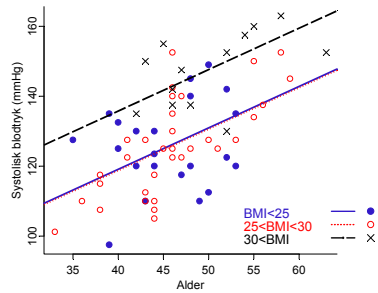
Hypotesen testes vha. en multipel regressionsanalyse, p=0.10.

Vi accepterer dermed hypotesen om den samme alderseffekt i de 3 BMI-grupper.

**Vi kan antage, at BMI er ikke en effektmodifikator.**



En multipel regressionsanalyse med samme alderseffekt (hældning) i de 3 BMI-grupper:




**Modelkontrol:** Som i den simple lineære regressionsanslyse, dog her noget mere kompliceret.



**Resultat:**

| | Estimat | se | CI | p |
|---|---|---|---|---|
| Intercept | 71.7 | 10.8 | (50.1;93.3) | 0.000 |
| Alder | 1.2 | 0.2 | (0.7;1.6) | 0.000 |
| BMI≤25 | 0 | | | |
| 25<BMI≤30 | -0.4 | 3.0 | (-6.4;5.7) | 0.905 |
| BMI>30 | 16.7 | 4.0 | (8.7;24.6) | 0.000 |
| $sd_{Res}$ | 11.0 | | | |

**Hvordan skal vi fortolke dette resultat?**

$$\text{middel Blodtryk} = 71.7 + 1.2 \cdot \text{Alder} - 0.4 \cdot \text{BMI}_{25\text{-}30} + 16.7 \cdot \text{BMI}_{30+}$$



**Eksempel 4: beregning af det forventede blodtryk**

Betragt en person med følgende data:
Alder=50 år, BMI=27 kg/m$^2$

$\text{BMI}_{25\text{-}30} = 1$
$\text{BMI}_{30+} = 0$

Middelblodtrykket udregnes til

Middelblodtryk
$= 71.7 + 1.2 \cdot \text{Alder} - 0.4 \cdot \text{BMI}_{25\text{-}30} + 16.7 \cdot \text{BMI}_{30+}$
$= 71.7 + 1.2 \cdot 50 - 0.4 \cdot 1 + 16.7 \cdot 0$
$= 130.6$ mmHg

Et prædiktionsinterval kan udregnes som tidligere

PI(50 årige, 25<BMI $\leq$ 30): $130.6 \pm 1.96 \cdot 11.0$
$= (109.0;152.2)$ mmHg

## Eksempel 5: effekten af Alder

Betragt to personer:

Person 1: $\text{Alder}_1$=40 år, $\text{BMI}_1$=31 kg/m$^2$
Persen 2: $\text{Alder}_2$=50 år, $\text{BMI}_2$=31 kg/m$^2$

$\text{BMI}_{25\text{-}30} = 1$

$\text{BMI}_{30+} = 0$

Forskellen i middelblodtrykket er

$$\begin{aligned}
&\text{Middelblodtryk}_2 - \text{Middelblodtryk}_1 \\
&= (71.7 + 1.2 \cdot \text{Alder}_2 - 0.4 \cdot \text{BMI}_2^{25\text{-}30} + 16.7 \cdot \text{BMI}_2^{30+}) \\
&\quad - \left(71.7 + 1.2 \cdot \text{Alder}_1 - 0.4 \cdot \text{BMI}_1^{25\text{-}30} + 16.7 \cdot \text{BMI}_1^{30+}\right) \\
&= 1.2 \cdot (\text{Alder}_2 - \text{Alder}_1) \\
&= 1.2 \cdot 10 = 11.8 \text{ mmHg}
\end{aligned}$$

CI(Forskel): $(10 \cdot 0.72; 10 \cdot 1.65) = (7.2; 16.5)$ mmHg



## Eksempel 6: effekten af BMI

Betragt to personer:

Person 1: $\text{Alder}_1$=40 år, $\text{BMI}_1$=21 kg/m$^2$
Persen 2: $\text{Alder}_2$=40 år, $\text{BMI}_2$=27 kg/m$^2$

Forskellen i middelblodtrykket er

$$\begin{aligned}
&\text{Middelblodtryk}_2 - \text{Middelblodtryk}_1 \\
&= \hat{\beta}_{25\text{-}30} \cdot \text{BMI}_2^{25\text{-}30} \\
&= -0.4
\end{aligned}$$

CI(Forskel): (-6.4;5.7)



Betragt to nye personer:

Person 1: $\text{Alder}_1$=40 år, $\text{BMI}_1$=27 kg/m$^2$
Persen 2: $\text{Alder}_2$=40 år, $\text{BMI}_2$=32 kg/m$^2$

Forskellen i middel blodtrykket er

$$\begin{aligned}
&\text{Middel blodtryk}_2 - \text{Middel blodtryk}_1 \\
&= \hat{\beta}_{30+} - \hat{\beta}_{25\text{-}30} \\
&= 16.7 - (-0.4) \\
&= 17.0
\end{aligned}$$

CI(Forskel) kan vi ikke udregne fra på basis af denne analyse.

Sikkerhedsintervallet kan findes ved at lave en ny regressionsanalyse med BMI gruppe nr. 2 som referencegruppe.



## Er BMI en confounder for alderseffekten?

Fra den simple lineære regressionsanalyse fik vi

$\hat{\beta}_{\text{Crude}} = 1.50$ CI($\beta_{\text{Crude}}$): (1.00,1.99) mmHg/år

Fra den multiple lineære regressionsanalyse hvor også BMI-gruppe indgik i modellen fik vi

$\hat{\beta}_{\text{Adjusted}} = 1.18$ CI($\beta_{\text{Adjusted}}$): (0.72,1.65) mmHg/år

Hvis $\beta_{\text{Crude}} \neq \beta_{\text{Adjusted}}$ så er BMI en confounder.

**Det tyder således på, at BMI er en confounder for alderseffekten.**

Epidemiologi og **biostatistik**.
Uge 4, torsdag
Erik Parner, Institut for Biostatistik.

**Logistisk regressionsanalyse**
- Generelt om logistisk regressionsanalyse
- Eksempel 1

**Kliniske målinger**
- Kliniske målinger og variationskilder
- Estimation af størrelsen af de tilfældige variationskilder (eksempel 2)
- Sammenligning af to målemetoder/målinger:
  - kontinuerte målinger (eksempel 3)
  - kategoriske målinger (eksempel 4)

**Korrelation**

1

## Logistisk regressionsanalyse

Responsen ($y$) er en **dichotom variabel**, f.eks.

- operation for diskusprolaps: sucess/ikke-sucess.
- i live efter 6 mdr: ja/nej.
- fødselsvægt < 2500 gram: ja/nej.

Den logistiske regressionsmodel beskriver hvordan *sandsynligheden for hændelsen* ($p$) afhænger af forklarende variable $x_1,...,x_m$ via logaritmen til *odds for hændelsen* ($o$)

$$\text{logit}(p) = \ln(\frac{p}{1-p}) = \ln(o)$$

ved

$$\ln(o) = \alpha + \beta_1 \cdot x_1 + ... + \beta_m \cdot x_m$$

(Bland Kapitel 17.8)

2

## Lineær- versus logistisk regressionsanalyse

**Lineær regressionsanalyse:**

Responsen ($y$) er en kontinuert variabel, f.eks. blodtryk, PEFR eller FEV1.

Responsen afhænger af forklarende variable $x_1,...,x_m$ ved

$$y = \alpha + \beta_1 \cdot x_1 + ... + \beta_m \cdot x_m + \text{"tilfældig variation"}$$

**Logistisk regressionsanalyse:**

Responsen ($y$) er en dichotom variabel og logaritmen til *odds for begivenheden* ($o$) afhænger af de forklarende variable $x_1,...,x_m$ ved

$$\ln(o) = \ln(\frac{p}{1-p}) = \alpha + \beta_1 \cdot x_1 + ... + \beta_m \cdot x_m$$

3

## Eksempel 1 (Bland side 322-323)

Prædiktion af sandsynligheden for kejsersnit.

**Responsvariabel**:
kejsersnit: ja/nej
($p$ = sandsynligheden for kejsersnit)

**Forklarende variable**:
BMI: kontinuert variabel
Induction: ja/nej (ja=1, nej=0)
Prev. vag. del.: ja/nej (ja=1, nej=0)

**Preliminære analyser viser**:
- BMI associeret med kejsersnit
- Induction associeret med kejsersnit
- Prev. vag. del. associeret med kejsersnit

4

**Formål med regressionsanalysen i eksemplet**:
Er der stadig en associationen mellem BMI og kejsersnit når vi korrigerer for Induction og Prev. vag. del. (PVD)?

**Resultat**:

| | Coef. | Std. Err. | z | p-value | 95% CI |
|---|---|---|---|---|---|
| BMI | 0.088 | 0.200 | 4.42 | <0.001 | 0.049 to 0.128 |
| Induction | 0.647 | 0.214 | 3.02 | 0.003 | 0.228 to 1.067 |
| PVD | -1.796 | 0.298 | -6.03 | <0.001 | -2.381 to -1.212 |
| Intercept | -3.700 | 0.534 | -6.93 | <0.001 | -4.747 to -2.653 |

**Hvordan skal vi fortolke dette resultat?**

$$\ln(\hat{o}) = -3.700 + 0.088 \cdot \text{BMI} + 0.647 \cdot \text{Induction} - 1.796 \cdot \text{PVD}$$

5

## Eksempel: sandsynligheden for kejsersnit

$$\ln(\hat{o}) = -3.700 + 0.088 \cdot \text{BMI} + 0.647 \cdot \text{Induction} - 1.796 \cdot \text{PVD}$$

Betragt en kvinde med:

BMI=25 kg/m$^2$, Induction=0, PVD=0

Indsættes dette i regressionsligningen fås:

$$\ln(\hat{o}) = -3.700 + 0.088 \cdot 25 + 0.647 \cdot 0 - 1.796 \cdot 0$$
$$= -1.493$$

$$\hat{o} = \exp(-1.493) = 0.225$$

$$\hat{p} = \frac{\hat{o}}{1+\hat{o}} = 0.184$$ CI kan I ikke udregne!

Hvis Induction=1:

$$\ln(\hat{o}) = -3.700 + 0.088 \cdot 25 + 0.647 \cdot 1 - 1.796 \cdot 0$$
$$\ldots \hat{o} = 0.429 \quad \ldots \quad \hat{p} = 0.300$$

6

**Eksempel: effekten af Induction**

Betragt to kvinder:

Kvinde 1: $BMI_1$=25 kg/m$^2$, $Induction_1$=0, $PVD_1$=0
Kvinde 2: $BMI_2$=25 kg/m$^2$, $Induction_2$=1, $PVD_2$=0

$OR$ kan estimeres ud fra de to odd's fra før:

$$\widehat{OR} = \frac{\hat{o}_2}{\hat{o}_1} = \frac{0.429}{0.225} = 1.9$$ CI?

Kvinde 2 har altså dobbelt så stor risiko (odds) for kejsersnit i forhold til kvinde 1.

Hypotese: Kunne OR være 4?



Odds ratioen kan også udregnes som:

$$\widehat{OR} = \frac{\hat{o}_2}{\hat{o}_1}$$

$$= \frac{\exp(-3.700 + 0.088 \times BMI_2 + 0.647 \times Induction_2 - 1.796 \times PVD_2)}{\exp(-3.700 + 0.088 \times BMI_1 + 0.647 \times Induction_1 - 1.796 \times PVD_1)}$$

$$= \frac{\exp(0.647 \times Induction_2)}{\exp(0.647 \times Induction_1)}$$

$$= \frac{\exp(0.647 \times 1)}{\exp(0.647 \times 0)}$$

$$= \exp(0.647) = 1.9$$ ← Samme OR som før!

CI($\widehat{OR}$): $(\exp(0.228), \exp(1.067)) = (1.3, 2.9)$

Vi får udregnet CI!



- Der gælder altså:

$$\beta = \ln(OR)$$ !!!

- Resultatet bliver det samme uanset hvad BMI og PVD er!

Der er med andre ord i regressionsligningen

$$\ln(\hat{o}) = -3.700 + 0.088 \times BMI + 0.647 \times Induction - 1.796 \times PVD$$

antaget **ingen effektmodifikation** mellem BMI, Induction og PVD!



**Eksempel: effekten af BMI**

Betragt to kvinder:

Kvinde 1: $BMI_1$=25 kg/m$^2$
Kvinde 2: $BMI_2$=27 kg/m$^2$

hvor alt andet er lige

$$\widehat{OR} = \frac{\exp(27 \cdot 0.088)}{\exp(25 \cdot 0.088)} = \exp((27-25) \cdot 0.088) = \exp(2 \cdot 0.088) = 1.2$$

CI($\widehat{OR}$): $(\exp(2 \cdot 0.049), \exp(2 \cdot 0.128)) = (1.1, 1.3)$

Betragt to andre kvinder:

Kvinde 3: $BMI_3$=18 kg/m$^2$
Kvinde 4: $BMI_4$=20 kg/m$^2$

hvor alt andet er lige

$$\widehat{OR} = \exp(2 \cdot 0.088) = 1.2$$

dvs. samme OR!!!



Vi har i regressionsmodellen antaget, at effekten af en BMI forskel på 2 kg/m$^2$ er uafhængig af størrelsen på BMI.

Er det rimeligt?



**Tabel for OR**

| | Odds ratio | p-value | 95% CI |
|---|---|---|---|
| BMI | 1.092 | <0.001 | 1.050 to 1.136 |
| Induction | 1.910 | 0.003 | 1.256 to 2.906 |
| PVD | 0.166 | <0.001 | 0.096 to 0.298 |

OR'en for BMI svarer til en BMI forskel på 1 kg/m$^2$.
Ofte vil det være tabellen for OR, som er angivet i en artikel.

OR'en svarende til en BMI forskel på 2 kg/m$^2$ fås ved

$$\widehat{OR} = 1.092^2 = 1.2$$

CI($\widehat{OR}$): $(1.050^2, 1.136^2) = (1.1, 1.3)$

Mere generelt gælder

$$\widehat{OR} = \widehat{OR}_{BMI} \cdot \widehat{OR}_{Induction} \cdot \widehat{OR}_{PVD}$$ CI kan I ikke udregne!

**Kommentarer til logistisk regressionsanalyse**

- Estimation af $\alpha, \beta_1, \ldots, \beta_m$ og se'er m.v. er kompliceret, men kan laves af de fleste statistikprogrampakker.
- Den logistiske regressionsanalyse bør kun anvendes hvis antallet af observationer er rimeligt stort.

  En tommelfingerregel er:

  - **der bør være MINDST 10 ja'er og 10 nej'er for responsen (men helst 20 af hver), for hver forklarende variabel i modellen** (Bland side 323).
- En variant, **betinget logistisk regressionsanalyse**, anvendes for matchede data. Analysen anvendes ofte i case-control studier hvor cases og kontroller er matchede mht. potentielle risikofaktorer.

13

**Kliniske målinger**

Hvad influerer på en (klinisk) måling ?

**Individ-relateret:**
- person
- helbredstilstand
- tidspunkt (sæson, døgn)
- ???

**Metode-relateret:**
- målemetode
- apparat
- kalibrering af apparat
- observatør
- hospital
- ???

variationskilder

(Bland Kapitel 10.2 + 10.12 + 11.9-11 + 15.1-3)

14

**Klinisk måling:** vi forestiller os, at der er en underliggende/ukendt **sand værdi**, som vi forsøger at **måle**.

(for en given person, med en given helbredstilstand, til et givet tidspunkt etc)

Ved ***gentagen*** måling med ***samme metode*:**

en lidt anden værdi (som regel), fordi:

- metoden har en indbygget usikkerhed (**tilfældig fejl**)

Kan ofte beskrives ved en normalfordeling → "målefejl"

15

Mange målemetoder vil ud fra metodens underliggende fysiske og/eller kemiske principper være gode (lille systematisk og tilfældig fejl)

- lungefunktion: måling af rumfang
- kemiske analyse: kromatografisk metode

Måling på noget "biologisk" introducerer en række nye og måske ukendte variationskilder, f.eks. :

- fastende
- i hvile
- instruktion af patienten før målingen

→ "intra"

- patienterne er forskellige

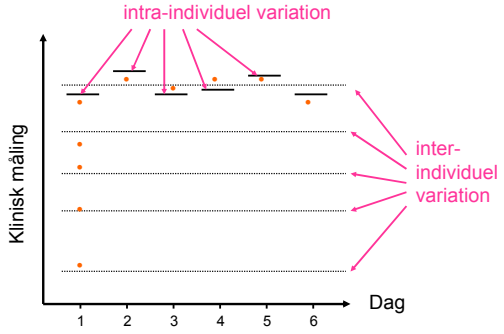
→ "inter"

16

**Variationskomponenter**

- **Inter-individuelle variation**: Hver person har et underlæggende niveau (sande værdi). Den inter-individuelle variation beskriver variationen i personernes niveau. Også kaldet den biologiske variation.
  Eks: personens niveau kunne være gennemsnittet af målingerne over mange dage.
- **Intra-individuelle variation**: Personens sande værdi afhænger af under hvilken omstændighed den bliver målt. Variationen i de sande værdier indenfor personen kaldes for den intra-individuelle variation.
  Eks: den intra-individuelle variation kunne være dag-til-dag variationen i de sande værdier.
- **Målefejl**: Variationen af målingerne hvis vi måler flere gange lige efter hinanden.

17

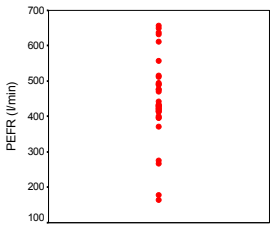



18

## Eksempel 2

Estimation af størrelsen af de tilfældige variationskilder

En stikprøve af PEFR målinger, målt med Wright.




Mean=450
SD=116
PI: 222-678

Hvor meget af variationen i PEFR (målt med Wright) skyldes
- **variation mellem personer** (inter-individuel)
- **variation indenfor person** (intra-individuel+målefejl)**?**

19

## Nyt forsøg:

PEFR (l/min) målt med Wright meter

| Person | 1. måling | 2. måling |
|---|---|---|
| 1 | 494 | 490 |
| 2 | 395 | 397 |
| 3 | 516 | 512 |
| 4 | 434 | 401 |
| : | : | : |
| 14 | 478 | 492 |
| 15 | 178 | 165 |
| 16 | 423 | 372 |
| 17 | 427 | 421 |

målt 2 forskellige dage

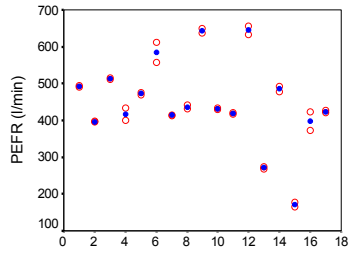
↓

Ingen systematisk forskel mellem de 2 målinger

(Bland Table 15.1, side 270)

20

## Variationskilderne




○ PEFR
● Gennemsnit

Dette design kan **ikke adskille intra-**individuel variation og **målefejl**

**Inter-individuel variation**
= variationen af gennemsnittene
– usikkerhed på gennemsnittene

**Intra-individuel+målefejl**
= variationen omkring gennemsnittene

21

En **Variansanalyse** kan kvantificere de systematiske og tilfældige kilder til variation:

$s_b$ = spredningen mellem personer (between)
$= 112.4$ l/min

$s_w$ = spredningen indenfor person (within)
$= 15.3$ l/min

$s_{\text{Een måling}} = \sqrt{s_b^2 + s_w^2} = 113.4$ l/min

Resultaterne kan bruges til at besvare spørgsmål som:

A. Hvor stor en **andel** udgør den biologiske variation?

$\text{Andel} = 112.4^2 / 113.4^2 = 98\%$

22

B. Prædiktionsinterval for forskel mellem 2 målinger på **samme** person på 2 **forskellige** dage:

$$\pm 1.96 \cdot \sqrt{s_w^2 + s_w^2} = \pm 2.77 \cdot s_w = \pm 42.4 \text{ l/min}$$

$1.96 \cdot \sqrt{2}$

C. Teste hypotesen: $\sigma_b = 0$

Ensidet variansanalyse (kommer til øvelserne!)

23

## Eksempel 3

Sammenligning af to kontinuerte målinger

**Eksempler på metodeforskelle:**

| Systematisk forskel: | Tilfældig variation: |
|---|---|
| • generelt niveau ← | • forskellige måleusikkerhed |
| • kun ved små/store værdier | • større ved store værdier |

Analysen afhænger af den forskel man vil beskrive.

(Hvad man måler på - standardiseret prøve, raske personer eller patienter - afhænger hvad/hvem man ønsker at generalisere til)

24

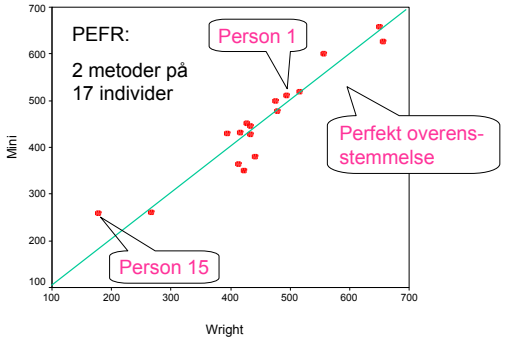
**Data:**

PEFR målt med Wright og Mini meter

| Person | Wright | Mini | AVG | Dif (W-M) |
|---|---|---|---|---|
| 1 | 494 | 512 | 503 | -18 |
| 2 | 395 | 430 | 412.5 | -35 |
| 3 | 516 | 520 | 518 | -4 |
| 4 | 434 | 428 | 431 | 6 |
| : | : | : | : | : |
| 14 | 478 | 477 | 477.5 | 1 |
| 15 | 178 | 259 | 218.5 | -81 |
| 16 | 423 | 350 | 386.5 | 73 |
| 17 | 427 | 451 | 439 | -24 |

(Bland Tabel 15.4, side 273)

25

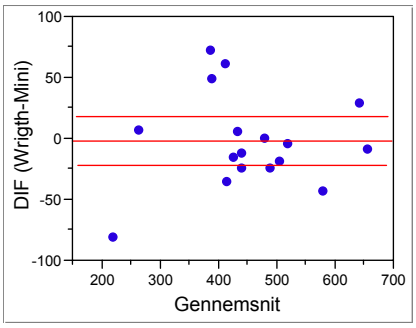



Lineær regression ikke interessant!!

26

**Afhænger forskellen af personens niveau?**

**Bland-Altman plot** (PEFR: Mini vs Wright)



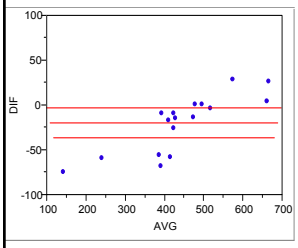

DIF forløber **parallelt** med x-akse

DIF's variation **konstant**

27

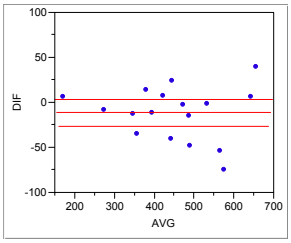
Bland-Altman plot: hvis forskellen afhænger af niveau

Eksempel A




Den **systematiske** forskel afhænger af niveau

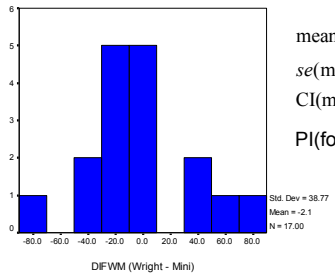
Eksempel B




Den **tilfældige** variation afhænger af niveau

28

**Analyse – Mini versus Wright (fortsat)**




$\text{mean} = -2.1$

$se(\text{mean}) = 38.8/\sqrt{17} = 9.40$

CI(mean)=( -20.5; 16.3)

PI(forskel): $-2.1 \pm 1.96 \cdot 38.8$

$= (-78.1; 73.9)$ l/min

Forudsat, differenser normalfordelt

PI(forskel) : **Limits of agreement**

Prædiktionsinterval for forskel mellem værdier målt med de 2 metoder på samme person

29

**Eksempel 4**

Sammenligning af to kategoriske målinger

***Jern i knoglemarv*** (Nanna M. Jensen, Randers Centralsygehus)

To observatører har (uafhængigt af hinanden) bedømt indholdet af jern i den samme prøve af knoglemarv fra i alt 75 patienter med "jernmangel" (bedømt ud fra blodprøve)

| | Observatør | |
|---|---|---|
| patient | 1 | 2 |
| 1 | Intet | Intet |
| 2 | Intet | Nedsat |
| 3 | Normalt | Normalt |
| 4 | Nedsat | Normalt |
| 5 | Nedsat | Intet |
| : | : | : |

30

| | Observatør 2 | | | | |
|---|---|---|---|---|---|
| Observatør 1 | Intet =1 | Nedsat =2 | Normalt =3 | Øget=4 | Total |
| Intet =1 | 31 | 2 | 1 | 1 | 35 |
| Nedsat =2 | 9 | 4 | 1 | 0 | 14 |
| Normalt =3 | 5 | 5 | 5 | 6 | 21 |
| Øget =4 | 0 | 0 | 1 | 4 | 5 |
| Total | 45 | 11 | 8 | 11 | 75 |

| | Observatør 2 | | | |
|---|---|---|---|---|
| Observatør 1 | Intet | Nedsat | Normalt | Øget |
| Intet | O1=O2 | O1=O2 -1 | … | |
| Nedsat | O1=O2 +1 | O1=O2 | O1=O2 -1 | |
| Normalt | … | O1=O2 +1 | O1=O2 | O1=O2 -1 |
| Øget | | | O1=O2 +1 | O1=O2 |

Antager at forskel mellem Intet og Nedsat svarer til forskel mellem Nedsat og Normalt etc.

31

Er der systematisk uenighed mellem de 2 observatører ?

| | | | | O 1 - O2 | | | |
|---|---|---|---|---|---|---|---|
| | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| Antal | 1 | 1+0 | 2+1+6 | 31+4+5+4 | 9+5+1 | 5+0 | 0 |
| Sum | 1 | 1 | 9 | 44 | 15 | 5 | 0 |

Nulhypotese: Er der symmetri omkring 0 ?

Men: data ikke normalfordelt ???

Signed Wilcoxon test, Forelæsning 5 !

32

**Korrelation**
Pearson korrelationen

Mål for afhængigheden (associationen) mellem 2 variable.

I regression er der **præference** mellem de 2 variable: en responsvariabel og en forklarende variabel. For korrelation er der ikke en præference.

***BMI og Kolesterol (fra Uge 3, regression):***

Næppe interessant

- prædiktere kolesterol vha BMI eller omvendt?

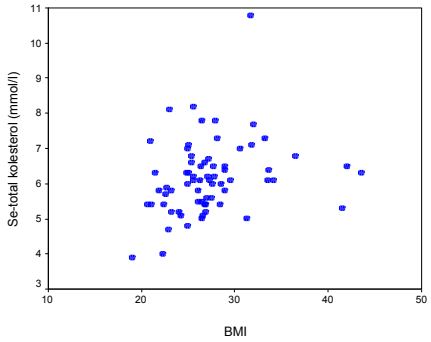
Association mellem BMI og Kolesterol:

- en underliggende fælles "årsag", f.eks. gener, livsstil, m.v.

Vi ønsker at kvantificere associationen med ét tal!

33

Er der en stigende tendens ?




34

Begrebet korrelation har i statistisk forstand en præcis (sandsynlighedsteoretisk) definition.

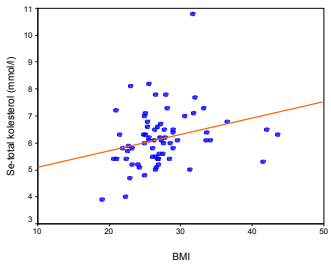
Korrelationskoefficienten ($\rho$) er et tal mellem -1 og 1. Den måler den ***lineære*** afhængighed mellem 2 variable ($x$ og $y$).

Hvis $\rho$ = 1 : så ligger *x* og *y* på en ret linie med positiv hældning.
Hvis $\rho$ = -1: så ligger *x* og *y* på en ret linie med negativ hældning.

I praksis ligger observationerne aldrig på en ret linie !

35

Hvis både $x$ og $y$ er kontinuerede variable, kan korrelationskoefficienten estimeres ved *Pearsons* korrelationskoefficient $r = \hat{\rho}$




$r = 0.26$
(Pearson)

$p$=0.035

$R^2 = 0.068$
$= r^2$
$= 0.26^2$

$R^2$ = andel "forklaret" variation = $r^2$

36

*Hypotesen*: $\rho = 0$ (ingen association)

kan testes vha Pearsons korrelations-koefficient, men det kræver en række forudsætninger opfyldt.

Både x og y normalfordelt og linearitet

*se* og sikkerhedsinterval kan udregnes !

men I slipper ! Kap. 11.10, 18.6

Hvis den ene (eller begge) variabel er en kategorisk variabel med *ordnede* kategorier (f.eks. NYHA I, II, III og IV eller en smerte-score) kan man **ikke** beregne Pearsons korrelationskoefficient. Mere om det i Forelæsning 5.

37

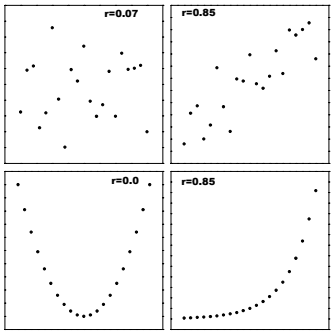
**Misforståelser om korrelationen**

Tolkning af denne koefficient ($r$) giver anledning til **mange** misforståelser. F.eks.:

- der er ingen sammenhæng mellem *x* og *y*, hvis *r* er tæt ved *0*
- korrelationen beskriver overensstemmelsen mellem målemetoder *x* og *y*.
  F.eks. "der er en god overensstemmelse mellem *x* og *y* hvis $r$ er tæt ved 1".

38

**Korrelation og "sammenhæng"**




Husk: korrelationen måler den **lineære** sammenhæng!

39

**Korrelation og sammenligning af målemetoder**

Kan vi sammenligne 2 metoder vha korrelation ???

Nej!

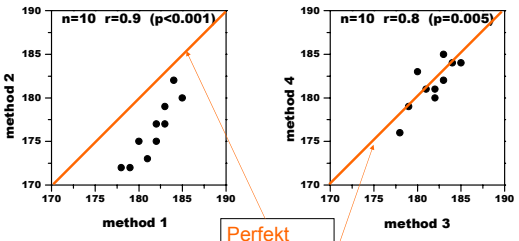
**Korrelation måler ikke størrelsen af:**

1. den systematiske forskel
2. den tilfældige forskel

Det vil vi nu se eksempler på…

40

**Eksempel.** 4 metoder til måling af Højde (cm). Stemmer method 1 og 2 mere overens end method 3 og 4 ? ***NEJ !***



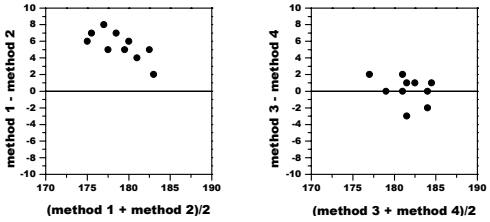

Klar syste-matisk forskel

Perfekt overens-stemmelse

Ingen syste-matisk forskel

41

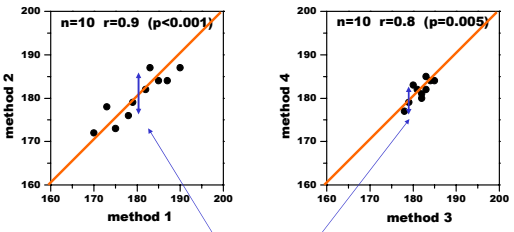
Bland-Altman plot:




Korrelationen måler ikke størrelsen af den **systematiske** forskel.

42

**Eksempel.** 4 metoder til måling af Højde (cm).
Stemmer method 1 og 2 mere overens end method 3 og 4 ? ***NEJ !***



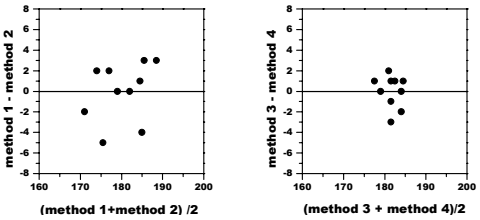

Stor tilfældig forskel

Variation i forskellen mellem de 2 metoder

Lille tilfældig forskel

43

Bland-Altman plot:




Korrelationen måler ikke størrelsen af den **tilfældige** forskel.

44

1

Epidemiologi og **biostatistik**.
Uge 5, torsdag 6. marts 2002
Morten Frydenberg, Institut for Biostatistik.

Det statistiske **modelbegreb**

**Modelselektion**

Ikke **parametrisk statistiske test** :

- Ideen bag
- **To grupper**: Mann-Whitney / Wilcoxon testet
- **Parret data** (symmetrisk fordeling): Wilcoxon signed rank
- **Association**: Test baseret på Spearman's rang korrelation

2

Analyse af **overlevelsesdata** (ventetidsdata)

- **Censurering** (højre + andet)
- **Kaplan-Meyer** kurver
- **Cox' proportional** hazard model

3

**Statistiske modeller**

Bag **alle** beregninger af:

Estimater, se, sikkerhedsintervaller, test og p-værdier

ligger en **statistisk model.**

Modellen er en **approksimation til virkeligheden**.

**Valget af model** er et valg mellem:

**kompliceret** model ← **ofte god approksimation**

**simpel** model ← **ofte dårlig approksimation**

**kompliceret** model → **svær at forstå og analysere**

**simpel** model → **let at forstå og analysere**

En model skal vælges så **kompliceret,** at den **ikke** er helt **forkert** og så **simpel,** at den er til at **analysere** og **forstå**.

4

Modellen er typisk baseret på **antagelser,** så som:

- de enkelte observationer er **uafhængige.**
- målefejlen er **normalfordelt.**
- variationen mellem individer er **normalfordelte.**
- Ln(odds) kan skrives som en **sum** af forskellige bidrag.
- bidraget fra alder **afhænger ikke af** personens køn. (**ingen effektmodifikation**)
- OR stiger **eksponentielt** med forskellen i BMI.

Hvis antagelserne ikke er (næsten) rigtige

bliver **resultaterne værdiløse.**

Derfor bør al statistisk analyse inkludere **modelkontrol**.

5

**Modelselektion**

Ofte er den model man får præsenteret i en artikel **ikke den eneste** forfatterne har fittet til data.

Man får kun præsenteret den **"bedste".**

Modellen er **selekteret** (udvalgt).

**Bevidst** eller **ubevidst**.

**Manuelt** eller **automatisk** (PC: *Find den bedste model* !).

**Modelselektion** har (desværre) **betydning** for resulterne:

**Estimaterne** er typisk for **store** (for langt væk fra nul).

**Sikkerhedsintervallerne** for **smalle.**

**P-værdierne** for **små**.

6

**Ikke-parametrisk statistiske test**

**Hidtil (parametrisk statistik):**

- **Ukendt** størrelse (**parameter**) OR, middelfødselsvægt eller lign.
- **Estimat** og **standard error.**
- **Sikkerhedsinterval.**
- **Hypotese** (fx OR=1).
- **Test** baseret på $z = (estimat\text{-}hypotese)/se$.
- Resultaterne bygger på en **statistisk model.**

7

**Ikke-parametriske test**

- Baseres **ikke** på et parameter estimat.
- Men ofte på de **rangordnede** data.
- Bygger også på **statistiske modeller.**
- Men der er **svagere antagelser** bag dem.
- Ofte **rang-test:**

De **præcise** observationsværdier har **ingen betydning**.

Men det har **rangene:**

| | |
|---|---|
| Hvilken observation er **mindst** | **Rang=1** |
| Hvilken observation er **næstmindst** | **Rang=2** |
| Hvilken observation er **trediemindst** | **Rang=3** |

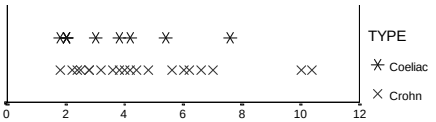
**osv**

8

**Et eksempel på Mann-Whitney-Wilcoxon test**

**Table 12.3** Biceps skinfold thickness (mm)

| Crohn's Disease | | | | Coeliac Disease | |
|---|---|---|---|---|---|
| 1.8 | 2.8 | 4.2 | 6.2 | 1.8 | 3.8 |
| 2.2 | 3.2 | 4.4 | 6.6 | 2.0 | 4.2 |
| 2.4 | 3.6 | 4.8 | 7.0 | 2.0 | 5.4 |
| 2.5 | 3.8 | 5.6 | 10.0 | 2.0 | 7.6 |
| 2.8 | 4.0 | 6.0 | 10.4 | 3.0 | |




**Er der forskel ?**

9

| RANGE | Crohn's Disease | | | Coeliac Disease | |
|---|---|---|---|---|---|
| **1.5** 1.8 | **9.5** 2.8 | **17.5** 4.2 | **24** 6.2 | **1.5** 1.8 | **14.5** 3.8 |
| **6** 2.2 | **12** 3.2 | **19** 4.4 | **25** 6.6 | **4** 2.0 | **17.5** 4.2 |
| **7** 2.4 | **13** 3.6 | **20** 4.8 | **26** 7.0 | **4** 2.0 | **21** 5.4 |
| **8** 2.5 | **14.5** 3.8 | **22** 5.6 | **28** 10.0 | **4** 2.0 | **27** 7.6 |
| **9.5** 2.8 | **16** 4.0 | **23** 6.0 | **29** 10.4 | **11** 3.0 | |
| Gennemsnitsrang: | **16.52** | | | **11.61** | |

Vi ønsker at teste **hypotesen**: *Ingen forskel i skinfold.*

**Idé:** Sammenlign **gennemsnitsrangen** blandt Coeliac patienter med **gennemsnitsrangen** blandt Crohn patienter.

En **stor forskel** vil være **kritisk** for hypotesen.

**P-værdi** = sandsynligheden for at observere en større forskel *under antagelse af hypotesen er sand* !

10

Beregning ved hjælp af **computer** eller **tabel**.

p-værdi=**0.15.**

**Konklusion:** Data strider ikke mod hypotesen.

Hypotesen **kan** accepteres !

**Bemærk**

- Præcist samme test hvis vi regnede på ln-data.
- Eller kvadratroden af data.
- Eller en hvilken som helst monoton transformation.
- Kun **rangordningen** betyder noget.
- Testet hedder **Mann-Whitney U-test** eller **Wilcoxon two-sample test.**

11

**Generelt:** **Mann-Whitney U-test**

ækvivalent med

**Wilcoxon two sample test**

**Data: To uafhængige** sæt observationer.

**Hypotese:** De to **fordelinger** er **ens.**

**Alternativ:** De to fordelinger er **forskudt** i forhold til hinanden.

**Ide:** Hvis **alternativet** er sandt vil **gennemsnitsrangen** være **forskellig** i de to grupper.

Hvis **hypotesen** er sand så vil **gennemsnitsrangene** være næsten **ens**.

**P-værdi** vha. af computer eller tabel.

12

**Et eksempel på signed Wilcoxon test**

Del af **Table 10.2** Capillary density (per $mm^2$) in feet of ulcerated patients. **Differens** mellem værste og bedste fod.

| | | | | | | | | ~~0~~ | ~~0~~ | ~~0~~ | ~~0~~ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Obs | -13 | -7 | -6 | -5 | -5 | -5 | -1 | | | | | 1 | 5 | 5 | 6 | 12 |
| \|Obs\| | 13 | 7 | 6 | 5 | 5 | 5 | 1 | | | | | 1 | 5 | 5 | 6 | 12 |
| Rang | 12 | 10 | 8.5 | 5 | 5 | 5 | 1.5 | | | | | 1.5 | 5 | 5 | 8.5 | 11 |

Sum af range: **47** **31**

**Hypotese:** *Ingen forskel mellem de to fødder,*
mao. **positive** og **negative** differenser har **ens fordelinger**.

**Idé:** Se på **forskellen** i **sum af rangene** af de **positive** og **negative** differenser.

**Bemærk:** Rangene beregnes **uden fortegn**.

**P-værdi** = sandsynligheden for at observere en større forskel *under antagelse af hypotesen er sand* !

13

Beregning ved hjælp af **computer** eller **tabel**.

**p-værdi=0.53.**

**Konklusion:** Data strider ikke mod hypotesen.

Hypotesen **kan** accepteres !

**Bemærk**

- **Ikke samme** test hvis vi transformerede data **inden** vi beregnede differensen.
- Fx et andet resultat hvis vi så på **relative** forskelle.
- Testet hedder **Wilxocon signed-rank test.**

**Et andet eksempel på signed Wilcoxon test**

14

Fra torsdag i uge 4 :

| | Observatør 2 | | | | |
|---|---|---|---|---|---|
| Observatør 1 | Intet | Nedsat | Normalt | Øget | Total |
| Intet | 31 | 2 | 1 | 1 | 35 |
| Nedsat | 9 | 4 | 1 | 0 | 14 |
| Normalt | 5 | 5 | 5 | 6 | 21 |
| Øget | 0 | 0 | 1 | 4 | 5 |
| Total | 45 | 11 | 8 | 11 | 75 |

| O1-O2 | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|
| Antal | 1 | 1 | 9 | 44 | 15 | 5 | 0 |

**Hypotese**: Ingen **systematisk** forskel mellem de to observatører.

**p-værdi**=11%.

**Konklusion:** Hypotesen kan accepteres.

15

**Generelt:** **Wilcoxon signed rank test.**

**Data:** Et sæt **uafhængige** observationer**.**

**Hypotese:** Fordelingen er **symmetrisk** om **0**.

**Alternativ:** Fordelingen er **ikke symmetrisk** om **0**.

**Ide:** Hvis **alternativet** er sandt vil **rangsummene** for de **positive** og **negative** tal være forskellige.

Hvis **hypotesen** er sand så vil **rangsummene** være næsten **ens**.

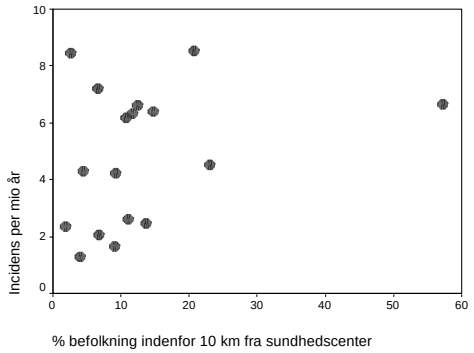
**P-værdi** vha. af computer eller tabel.

Bruges ofte ved **parrede** data - der regnes på **differensen** !

**Et eksempel på test for ingen sammenhæng**

16

**Table 12.7** Incidens af Kaposi's sarcoma i Tanzania




**Er der en sammenhæng/association ?**

17

Forudsætninger for **lineær regression** ikke opfyldt !

(Derfor) beregning af **Pearson** korrelation **uden mening**.

**Hvad så !**

**Kan vi nøjes med** et test **?**

**Til en start: Ja !?**

**Hypotese** (som sædvanlig): ***Ingen*** *sammenhæng***.**

**Idé:** Rangordne *x*'erne samt *y*'erne og beregn korrelation mellem **rangene**.

Korrelation ´langt væk´ fra 0 **kritisk**.

**P-værdi** = sandsynligheden for at observere en korrelation længere væk fra 0 under *antagelse af hypotesen er sand* !

18

Beregning ved hjælp af **computer** eller **tabel**.

Korrelation mellem rangene =0.38.

**p-værdi=0.14**

**Konklusion:** Data strider ikke mod hypotesen.

Hypotesen **kan** accepteres !

**Bemærk**

- Præcist samme test hvis vi regnede på $\ln(x)$ og $y$.
- Eller $\ln(x)$ og $y^2$.
- Eller en hvilkensomhelst monoton transformation.
- Kun **rangordningerne** betyder noget**.**
- Testet "hedder" **Spearman´s rang korrelation**

19

**Generelt:** **Test for ingen association baseret på Spearman´s korrelation**

**Data: Uafhængige par** (x,y) af observationer.

**Hypotese: Ingen association** mellem x og y .

**Alternativ: Monoton** association.

**Ide:** Hvis **alternativet** er sandt vil **rangene** af *x*'erne være **korrelerede** med **rangene** af *y*'erne.

**Spearman's** korrelation beregnes.

Hvis **hypotesen** er sand så vil denne **korrelation** være tæt på **0**.

**P-værdi** vha. af computer eller tabel.

**Spearman's korrelation** er ikke mulig at fortolke !

Men testet er godt nok !

20

**Ikke parametrisk test: Godt eller skidt ??**

**For:**

- **Svage** antagelser.
- Kan også bruges på **ordinal** data som
  meget godt; godt; rimeligt; dårligt; meget dårligt
  CIN 1;CIN 2; CIN 3; Cancer.
- Stort set **lige så stærke** som parametriske **test** (gælder dog ikke hvis man har få data).

**Imod:**

- Der er tale om **test**, ingen **estimater** med **CI**.
- Bruges ofte bevidstløst (svage antagelser=ingen antagelser).
- Kan kun bruges til **simple problemstillinger**.

21

**Overlevelses (ventetids) data**

Data der involverer ventetider:

**Tid til død** af kræft efter kræft diagnose.

**Ventetid til** operation.

Tid **mellem** galdestensoperation og fund af ny galdesten.

er ofte **censureret.**

Personerne dør af **anden årsag** end kræft.

Personerne er i live da **studiet slutter.**

Den opererede får **aldrig galdesten igen.**

Den opererede får ikke galdesten inden **studiet slutter.**

Den opererede **flytter** til et andet amt/land.

**=Højre censurering**: Vi ved hvornår personen sidst var rask/i live

22

Ventetids data kan således være:

- **Højre censureret**: Vi **ved**, at personen **ikke** har oplevet begivenheden **før sidste gang vi ser ham**.

**Men kan også være:**

- **Venstre censureret**: Vi **ved**, at personen har oplevet begivenheden **inden vi ser ham første gang**, men ikke hvornår.
- **Interval censureret:** Vi **ved**, at personen har oplevet begivenheden **i givet tidsinterval**, men ikke hvornår.

Data er ofte **interval censurerede**:

- Vi **ved**, patienten var rask ved **forrige** kontrol, men nu er han syg. Vi **ved ikke**, **hvornår** han blev det.
- **Interval censurerede** data er svære at analysere.

23

Der kan også være andre problemer med data:

- Vi **ved ikke** om personen har oplevet begivenheden **inden vi ser ham første gang.**
- Vi **ved ikke** om personen har oplevet begivenheden **i et givet tidsinterval**.

  Patienter var rask ved forrige kontrol og er det også nu.
  Har han været syg i mellemtiden ?
- Personer indgår kun hvis de **har overlevet** .

24

Det er kun **højre censurede** data, der er let at analysere !

**Metoderne** er:

- **Kaplan-Meyer plot**: Metode til at beregne/tegne **ventetidsfordelingen** under hensyntagen til højre censureringen.
- **Log-rank** test: Tester hypotesen:

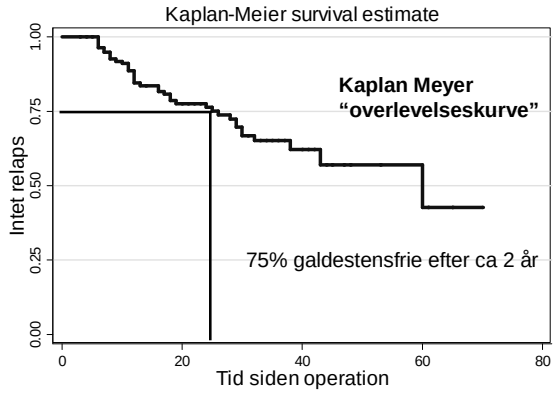
  *Samme ventetidsfordeling i to grupper.*
- **Cox's proportional** hazard model:

  Regressions analyse af ventetids data.

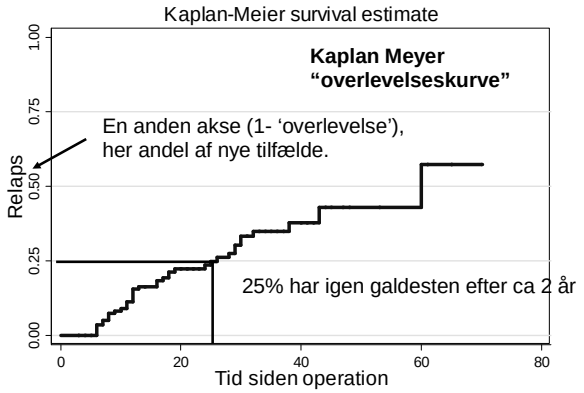
  Modellerer den **relative risiko** på log skala.

  Minder meget om **logistisk regression**.

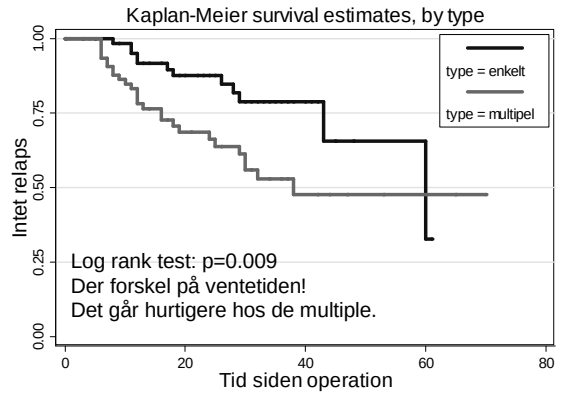
**Table 15.9** 25




26




27




Kurverne meget usikre efter 30-40 mdr.

28

**Cox's proportionale hazard model**
**ultra kort**

$h(t)$: hazard/intensitet til tidspunktet $t$.

$$h(t) = \frac{\text{sands. for at 'dø' inden } t+\Delta t \text{ givet man er i live til tid } t}{\Delta t}$$

**Model:** $h(t) = h_0(t) \cdot \exp\left(\beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_p \cdot x_p\right)$

**Baseline hazard** **Log hazard ratio (relativ risiko)**

29

**Tid mellem galdestensoperation og næste galdesten**

**Prediktorer:**

Flere galdesten fjernet

Diameter af største galdesten

Den tid det tog at opløse galdesten(ene) i mdr.

| Variable | B | S.E. | Exp(B) |
|---|---|---|---|
| Flere sten | .8384 | .4007 | 2.3127 |
| Diameter | -.0226 | .0361 | .9776 |
| Opløsningstid | .0445 | .0168 | 1.0455 |

Risikoen er **2.3** gange større, når **flere sten** er fjernet.

Risikoen stiger med **5%** per måned det tog at **opløse** stenene.