

Statistisk analyse af beskrivende undersøgelser

Morten Frydenberg
Institut for Biostatistik®

Introduktion til nogle statistiske begreber

- **Estimat**
- **Sikkerhedsinterval**

Analyse af proportioner (hyppigheder)

- Prævalens- og kumuleret incidens proportion

Analyse af rater

Lexis diagrammer

Beskrivelse af ventetider

- **Kaplan-Meier kurver**

Morten Frydenberg MPH, Introduktionsmodul - Uge 3 - Dag 1

1

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.

Morten Frydenberg

MPH, Introduktionsmodul - Uge 3 - Dag 1

2

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

- **Prævalensen** af patienter med liggesår en tilfældig dag på en given afdeling
- Den **kumulerede incidens** af SIDS (vuggedød) indenfor de første 2 mdr.
- Den **kumulerede incidens** af SIDS indenfor de første 6 mdr.
- Den **relative risiko** for SIDS (1 år) forbundet med maveleje i forhold til rygleje.

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

Hvordan beskriver man sammenhængen mellem kost og kræft ?

Morten Frydenberg

MPH, Introduktionsmodul - Uge 3 - Dag 1

3

Resultatet af den statistiske analyse bør bestå af:

- Et **estimat**, dvs. et **databaseret gæt** på den ukendte konstant.
- Et **sikkerhedsinterval**, dvs. et **databaseret** interval over hvilke værdier af den ukendte konstant, der er **forlignelig** med det data man har observeret.
- Eventuelt et statistisk **test**, der undersøger om en given **hypotese**, dvs. værdi af den ukendte konstant, er **forlignelig** med det data man har observeret.

Morten Frydenberg

MPH, Introduktionsmodul - Uge 3 - Dag 1

4

Eksempel 1: Rygning blandt gravide

Formål: at udtale sig om prævalensen af rygning blandt gravide ved første kontakt med jordmoderen.

Man er således interesseret i **sandsynligheden** for at en gravid ved det første besøg hos jordmoderen siger at hun ryger.

Denne sandsynlighed er ukendt!!!!

Men kan bruge **data** til at komme den nærmere:

Man spurgte derfor **106 gravide** på Skejby sygehus om hvorvidt de røg.

27 svarede ja og resten (**79**) nej.

På basis af disse tal vil vi **estimere** sandsynligheden til:

$$\frac{27}{106} = 0.2547 = 25.47\%$$

Morten Frydenberg

MPH, Introduktionsmodul - Uge 3 - Dag 1

5

Rygning blandt gravide

106 gravide af disse er **27** rygere

bedste gæt (**estimat**): **25%** af gravide ryger

Da dette gæt kun er baseret på oplysninger for 106 gravide er det **noget usikkert**.

Sikkerheden på estimatet kan repræsenteres vha. af et **sikkerhedsinterval**:

$$CI: (0.1751; 0.3486) = (18; 35)\%$$

Konklusion på basis af **disse data**:

Den **ukendte** sandsynlighed for at ryge ligger et sted mellem **18** og **35 %** - det bedste bud er **25%**.

Confidence Interval

Morten Frydenberg

MPH, Introduktionsmodul - Uge 3 - Dag 1

6

Ryging blandt gravide

I et andet studie, samme sted, havde man spurgt 425 gravide og 121 havde svaret ja.

Regner man på disse tal får man:

Estimat 28% og CI: (24;33)%

Konklusion på basis af studie 2:

Den ukendte sandsynlighed for at ryge ligger et sted mellem 24 og 33 % - det bedste bud er 28%.

Bemærk:

1. På grund af den tilfældige variation er der en forskel på estimerterne .
2. Fordi det andet studie er større end det første, så er sikkerhedsintervallet mindre. **Eksempel slut**

Morten Frydenberg

MPH, Introduktionsmodul - Uge 3 - Dag 1

7

Analyse af proportioner

Prævalens proportioner og kumuleret incidens proportioner er fra et statistisk synspunkt den samme type tal - det er sandsynligheder.

Formelt er de givet ved:

Kumuleret incidens proportion=

P(Opleve begivenheden i det givne tidsrum)

Prævalens proportion=

P(Besidde den givne egenskab på det givne tidspunkt)

Hvor P(.....) betyder sandsynligheden af

Derfor er de statistiske metoder identiske!

Morten Frydenberg

MPH, Introduktionsmodul - Uge 3 - Dag 1

8

Analyse af proportioner

Lad π betegne den ukendte sandsynlighed vi er interesseret i at udtale os om

Det kan så være en prævalens proportion eller en KIP, det er for et hvad angår den statistiske analyse ligegyldigt, - men essentielt når resultaterne skal fortolkes!

Data:

Vi betragter en gruppe objekter (personer)

n : antal personer i gruppen

y : antal personer i gruppen, der oplever begivenheden/besidder egenskaben

Estimatet bliver så $\hat{\pi} = \frac{y}{n}$

$\hat{\pi} = \frac{27}{106} = 0.2547$

Morten Frydenberg

MPH, Introduktionsmodul - Uge 3 - Dag 1

9

Eksempel 1

(n=106)

(y=27)

 $\hat{\pi} = \frac{27}{106} = 0.2547$

Analyse af proportioner

Sikkerhedsinterval for en proportion

To metoder:

- Eksakt - kun vha. computer
- Approksimativ - kan laves i hånden

Den approksimative metode for et 95% sikkerhedsinterval:

Beregn:

$$se(\hat{\pi}) = \sqrt{\hat{\pi} \cdot (1 - \hat{\pi}) / n}$$

CI for π :

$$\hat{\pi} \pm 1.96 \cdot se(\hat{\pi})$$

Eksempel 1

$$se(\hat{\pi}) = \sqrt{0.2547 \cdot (1 - 0.2547) / 106} = 0.0423$$

$$CI(\pi) = 0.2547 \pm 1.96 \cdot 0.0423 = 0.2547 \pm 0.0829 = (0.1718; 0.3377)$$

Morten Frydenberg

MPH, Introduktionsmodul - Uge 3 - Dag 1

10

Analyse af proportioner

Eksakt 95% sikkerhedsinterval i STATA (eksempel 1):

Kommando: cii 106 27

Output:

```
-- Binomial Exact --
Variable | Ob   Mean   Std. Err.   [95% Conf. Interval]
-----+-----
          | 106   .2547   .04232     .1751     .3486
```

Det eksakte sikkerhedsinterval : (17.5; 34.9)%

afviger lidt fra det approksimative: (17.2; 33.8)%

Men ikke meget!

De approksimative sikkerhedsintervaller afviger ikke væsentligt hvis man har en rimelig mængde data.

Når man regner 'i hånden' benyttes de approksimative.

Morten Frydenberg

MPH, Introduktionsmodul - Uge 3 - Dag 1

11

Approksimative sikkerhedsintervaller generelt

Hvordan fandt vi det (approks.) 95% sikkerhedsinterval?

Ved hjælp af et estimat og en standard error!

$$\hat{\pi} = 0.2547 \text{ og } se(\hat{\pi}) = 0.0423$$

Sikkerhedsinterval blev beregnet som:

$$\hat{\pi} \pm 1.96 \cdot se(\hat{\pi})$$

Da sikkerhedsintervallet afhænger af data vil det være forskelligt (variere) fra studie til studie.

95% af sikkerhedsintervallerne vil indeholde den sande ukendte værdi af π !

Morten Frydenberg

MPH, Introduktionsmodul - Uge 3 - Dag 1

12

Eksempel 2: Kræft blandt sygeplejersker

En måde at beskrive forekomsten af mave-tarm kræft er ved **incidensraten**, dvs. antal nye tilfælde per risiko tidsenhed, fx antal nye tilfælde per 1000 personår.

I studiet af mave-tarm kræft blandt 78,586 sygeplejersker i USA fandt man 602 nye (incidente) tilfælde i 10års perioden fra 1988 til 1998.

Population var fulgt i 758,903 personår, dvs. ikke alle kvinder var fulgt i 10 år.

Den estimerede rate var $\frac{602}{758,903 \text{ år}} = 0.00079$ per år
 $= 79$ per 100,000år

Med et sikkerhedsinterval på: CI: (73;86) per 100,000 år

Morten Frydenberg

MPH, Introduktionsmodul - Uge 3 - Dag 1

13

Kræft blandt sygeplejersker

Hvis vi lader **IR** betegne den **ukendt incidens rate**, så gav den statistisk analyse:

$$\widehat{IR} = 79 (73;86) \text{ per } 100,000 \text{ år}$$

Konklusion på basis af **disse data**:

Den **ukendte** rate af mave/tarmkræft ligger et sted mellem 73 og 86 per 100,000 år - det bedste bud er 79.

Eksempel slut

Morten Frydenberg

MPH, Introduktionsmodul - Uge 3 - Dag 1

14

Analyse af rater

Lad λ betegne den **ukendte** rate vi er interesseret i at udtale os om.

Det kan så være en incidens-, mortalitets- eller anden form for rate, det er for et hvad angår den **statistiske analyse** ligegyldigt,

- men essentielt når resultaterne skal **fortolkes!**

Data:

Vi betragter personer over tid

Eksempel 2

T : Observeret tid i risiko

$$(T = 758,903 \text{ år})$$

y : antal begivenheden (events)

$$(y = 602)$$

Estimatet bliver så $\hat{\lambda} = \frac{y}{T}$

$$\hat{\lambda} = 602/758,903 \text{ år} = 79.3 \text{ per } 100,000 \text{ år}$$

Morten Frydenberg

MPH, Introduktionsmodul - Uge 3 - Dag 1

15

Analyse af rater**Sikkerhedsinterval for en rate:**

Tre metoder:

- Eksakt (computer)
- Approksimativ original skala
- Approksimativ logaritmisk skala

Eksakt metode i STATA (eksempel 2)

Kommando: cii 758903 602,po

Output:

Variable	Exposure	Mean	Std. Err.	-- Poisson --	Exact --
				[95% Conf. Interval]	
	758903	.0007933	.0000323	.0007311	.0008592

Morten Frydenberg

MPH, Introduktionsmodul - Uge 3 - Dag 1

16

Analyse af rater

Approksimative sikkerhedsinterval - **Original** skala

Eksempel 2

$$\hat{\lambda} = \frac{y}{T}$$

$$se(\hat{\lambda}) = \frac{\sqrt{y}}{T}$$

95% sikkerhedsinterval:

$$CI(\lambda) = \hat{\lambda} \pm 1.96 \cdot se(\hat{\lambda})$$

$$\widehat{IR} = \frac{602}{758,903 \text{ år}} = 79.3 \text{ per } 100,000 \text{ år}$$

$$se(\widehat{IR}) = \frac{\sqrt{602}}{758,903 \text{ år}} = 3.23 \text{ per } 100,000 \text{ år}$$

$$CI(IR) = 79.3 \pm 1.96 \cdot 3.2 \text{ per } 100,000 \text{ år} = (73.0; 85.7) \text{ per } 100,000 \text{ år}$$

Morten Frydenberg

MPH, Introduktionsmodul - Uge 3 - Dag 1

17

Analyse af rater

Approksimative sikkerhedsinterval - **LN** skala

Eksempel 2

$$\ln(\hat{\lambda}) = \ln(y/T)$$

$$\ln(\widehat{IR}) = \ln(79.3) = 4.374$$

$$se(\ln(\hat{\lambda})) = 1/\sqrt{y}$$

$$se(\ln(\widehat{IR})) = 1/\sqrt{602} = 0.0407$$

95% sikkerhedsinterval:

$$CI(\ln(\lambda)) = \ln(\widehat{IR}) \pm 1.96 \cdot se(\ln(\widehat{IR})) = 4.374 \pm 1.96 \cdot 0.0407 = (4.294; 4.453)$$

Tilbage til original vha. eksponentialfunktionen

$$CI(IR) = (\exp(4.294); \exp(4.453)) = (73.2; 85.9) \text{ per } 100,000 \text{ år}$$

Morten Frydenberg

MPH, Introduktionsmodul - Uge 3 - Dag 1

18

Analyse af rater

Sikkerhedsintervaller for incidens raten per 100,000 år:

- Eksakt (73.1; 85.9)
- Approk. original skala (73.0; 85.7)
- Approk. logaritmisk skala (73.2; 85.9)

Stort set ens!

Generelt vil de tre metoder give det samme hvis antal events (begivenheder) er større end 20.

Anbefaling:

Få events brug den eksakte metode

Ellers brug den på ln-skala.

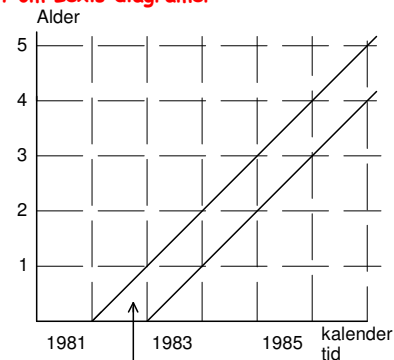
Morten Frydenberg

MPH, Introduktionsmodul - Uge 3 - Dag 1

19

Lidt om Lexis diagrammer

Et Lexis diagram er et vigtigt redskab ved hjælp ved planlægning, analyse og beskrivelse af et followup studie.



"Livslinjerne" for 1982 kohorten forløber mellem disse linier

Morten Frydenberg

MPH, Introduktionsmodul - Uge 3 - Dag 1

20

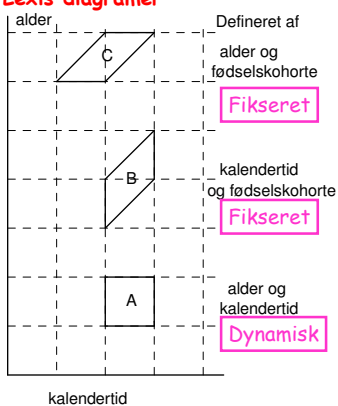
Lidt om Lexis diagrammer

Tre forskellige måder at definere hvordan man vil samme information.

A: "Personer som var mellem 30 og 50 år i perioden 2000-2002"

B: "Personer som var født i perioden 1950-60 i perioden 2000-2002"

C: "Personer som var født i perioden 1950-60 i alderen 60-70 år"



Fikseret

Fikseret

Dynamisk

Morten Frydenberg

MPH, Introduktionsmodul - Uge 3 - Dag 1

21

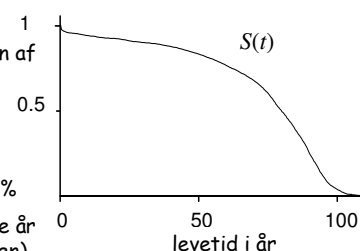
Lidt om overlevelsesfunktioner

Andel i live, som funktion af levetid:

Overlevelsesfunktion

Karakteristika:

- Starter ved 1, dvs. 100%
- Falder noget det første år (spædbørnsdødeligheden)
- Meget lille fald fra 2 til 45 år
- Derefter "ned ad bakke"
- Meget få i live efter 100 år



Morten Frydenberg

MPH, Introduktionsmodul - Uge 3 - Dag 1

22

Vi ønsker at udtale os om:

$$S(t) = P(\text{at være i live til tid } t)$$

fx

$$S(30 \text{ år}) = P(\text{at være i live 30 år efter}).$$

En måde til at finde denne er at udnytte følgende:

$$P(\text{at være i live til 30 år})$$

$$= P(\text{at være i live til 29 år}).$$

$$P(\text{ikke at dø før 30 år givet man er i live 29 år})$$

$$= P(\text{at være i live til 29 år}).$$

$$[1 - P(\text{at dø før 30 år givet man er i live 29 år})]$$

$$S(t) = S(u) \cdot [1 - P(\text{at dø inden } t \text{ givet ilive til } u)] \text{ , hvor } u < t$$

Morten Frydenberg

MPH, Introduktionsmodul - Uge 3 - Dag 1

23

Eksempel: Opfølgning efter operation.

21 pt. følges efter operation.

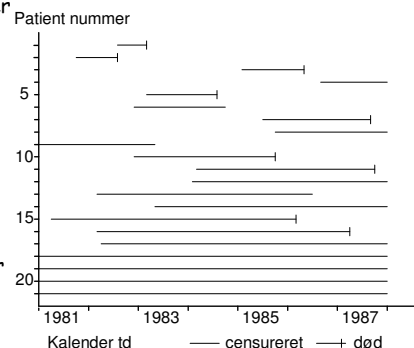
Tidsskala kalendertid.

Endpoint : Død

Data: Ventetid.

Nogle bliver censureret!

Dvs. opfølgning er ikke mulig.



Ønskes: Et estimat af overlevelsesfunktionen.

Morten Frydenberg

MPH, Introduktionsmodul - Uge 3 - Dag 1

24

