Intro til statistik

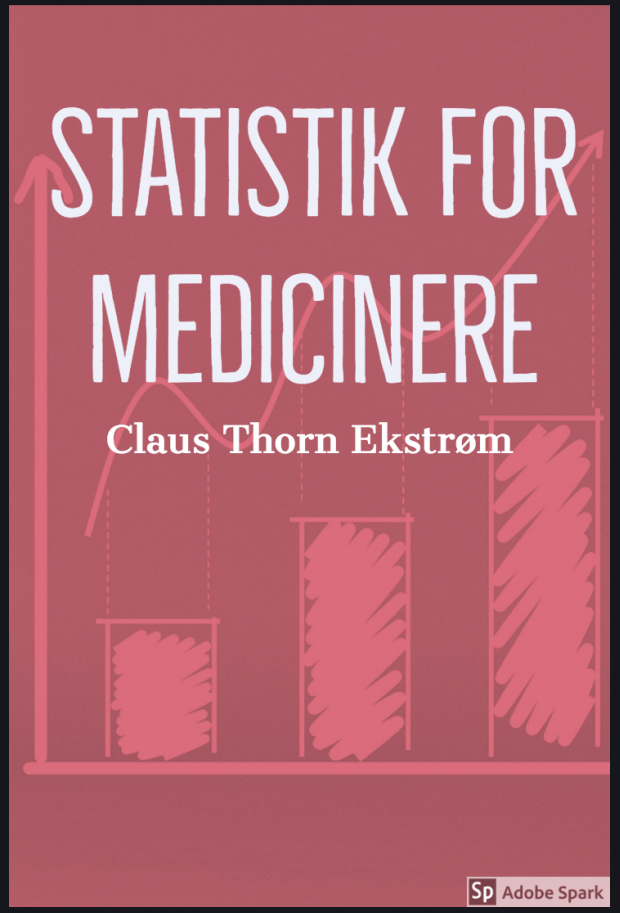
I. Deskriptiv statistik og R

Claus Thorn Ekstrøm
Biostatistik, KU
ekstrom@sund.ku.dk

Onsdag 29. april 2020
Slides @ biostatistics.dk/puff/

Velkommen!

- 5 undervisningsgange. 3 timer ~ ca 1 times snak og 2 timers øvelser
- Statistik med R - DataCamp
- Hjemmeside: `biostatistics.dk/puff/`
- Arbejdsbelastning: temmelig stor



Overordnet indhold

- **Begrebsforståelse**: kende til statistisk terminologi og standardanalyser
- **Praktisk erfaring**: kunne lave simple analyser
- **Artikellæsning**: kunne læse og forstå de statistiske overvejelser, analyser og resultater i en artikel

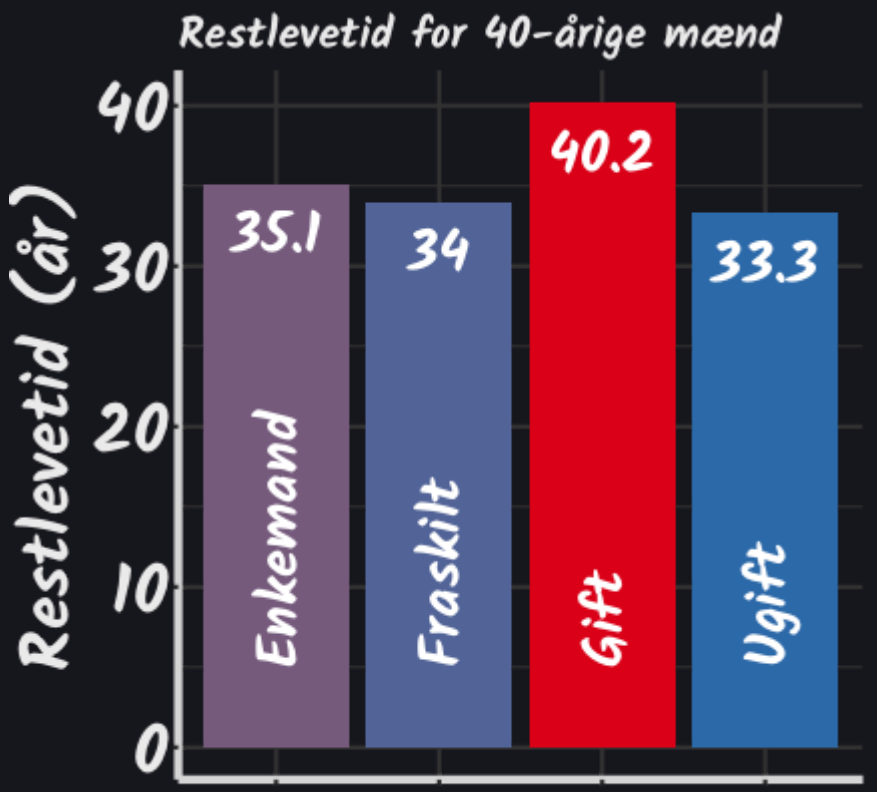
De 5 undervisningsgange

1. Deskriptiv statistik
2. Lineære modeller 1
3. Lineære modeller 2
4. Kategoriske data 1
5. Kategoriske data 2


166
2497
cph
AU 82 692

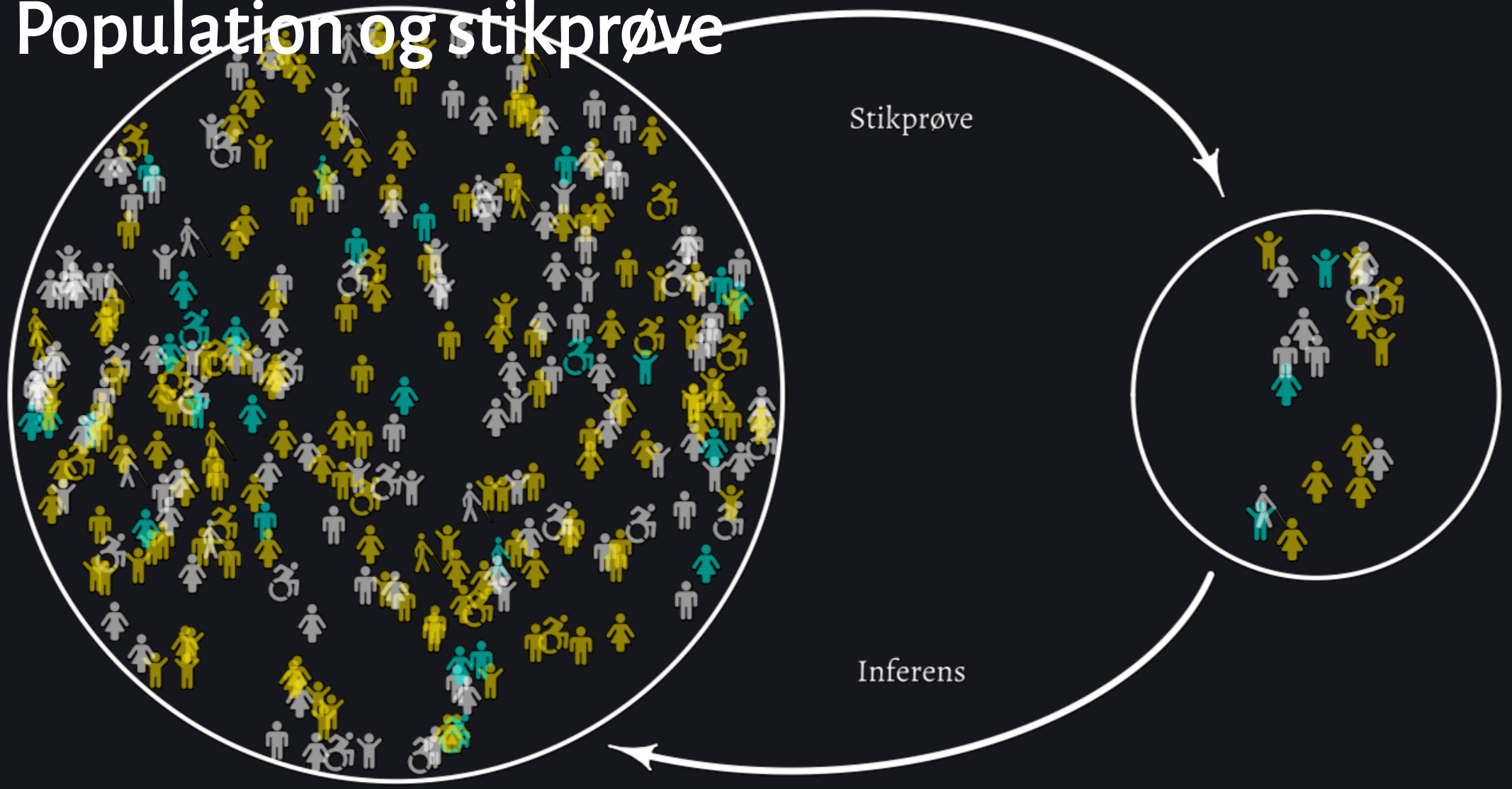
Hvad bruger vi statistik til?

- **Mønstre**.
  Hvad ser vi?
- **Prædiktion**.
  Hvad forventer vi ved ny observation?
- **Kausalitet**.
  Hvorfor?



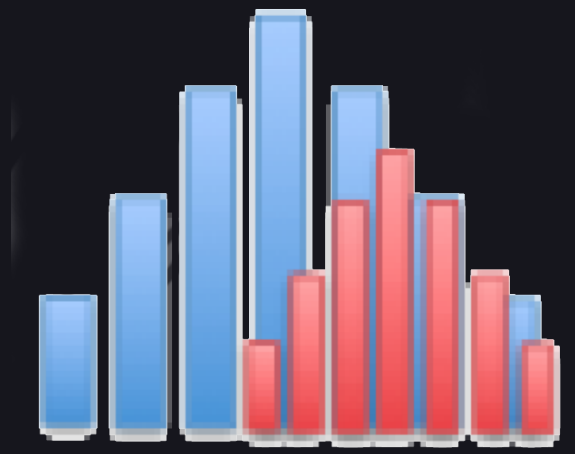
Den videnskabelige proces

Population og stikprøve



Data-genererende proces

Sandsynlighedsregning
Givet model, forudsig data
Deduktiv



Statistik
Givet data og modelklasse - find optimale parametre
Induktiv

Skal barnet ligge på ryggen eller maven?

Sundhedsstyrelsens anbefalinger:

- 1961: "Skiftevis på siderne og på ryggen".
- 1970: "Skiftevis på siderne, men ikke på ryggen".
- 1980: "Skiftevis på siderne og på maven - aldrig på ryggen".
- 1990: "Skiftevis på siderne og på maven".
- 1992: "Små børn, som ikke kan vende sig selv, må aldrig ligge på maven".
- 2001: "Læg altid spædbarnet til at sove på ryggen".

Videnskabelige artikler

Forskningsresultater publiceres i artikler. Struktur:

- **Abstract**: En kort sammenfatning af det hele
- **Introduction**: Problemstillingen, og hvorfor den er interessant
- **Methods**: Den videnskabelige process: indsamling af data, statistisk analysis, osv.
- **Results**: De fundne resultater herunder de statistiske analyser
- **Discussion**: Hvordan er vi nu blevet klogere, og hvilken betydning har det?

Den store tabel 1

Formålet med tabellen er at **beskrive studiepopulationen** (den type personer man undersøger). Generaliserbarhed. Repræsentativitet.

Dette gøres ved at præsentere **deskriptiv statistik** af de forskellige variable, man har målt i stikprøven.

- Hvad var den gennemsnitlige alder?
- Hvor stor variation i alder var der?
- Hvor mange procent kvinder var der?
- Hvor mange tilfælde havde de i gennemsnit?

Datatyper

Overordnet set to typer:

- **kvantitative** (numeriske/kontinuerte) variable
- **kvalitative** (kategoriske) variable

Eksempler: alder, højde, hårfarve, dagligt indtag af D-vitamin, køn, socialklasse, rygehistorik.

Deskriptiv statistik for kvantitative variable

Kvantitative variable beskrives ved *typisk værdi* og *variation*

- Til typiske værdier bruges *gennemsnit* eller *median*.
- Til variationen bruges enten *spredning* (standardafvigelse / standard deviation) eller *IQR* (inter-quartile range).

Gennemsnit og spredning hører sammen og bruges oftest til **symmetriske** fordelinger (kan checkes ved boxplot). Ellers bruges median og IQR.

Spredningen beskriver variationen omrking gennemsnittet og fortolkes løst som *den gennemsnitlige afstand fra gennemsnittet*!

Deskriptiv statistik for kategoriske variable

Kategoriske variable beskrives via *frekvens* og *andel* (også kaldet *relativ frekvens*)

Eksempel: I et studie på 515 personer var der 206 mænd. Frekvensen af mænd var derfor 206, og den relative frekvens (andelen af mænd) var

$$\frac{206}{512} = 0.4 = 40$$

Table 1. Baseline Characteristics of the Study Participants.*

Characteristic	Insulin Glargine (N=6264)	Standard Care (N=6273)
Demographic and clinical characteristics		
Age — yr	63.6±7.8	63.5±7.9
Female sex — no. (%)	2082 (33.2)	2304 (36.7)
Prior cardiovascular event — no. (%)†	3712 (59.3)	3666 (58.4)
Prior myocardial infarction — no. (%)	2221 (35.5)	2208 (35.2)
Prior stroke — no. (%)	805 (12.9)	851 (13.6)
Hypertension — no. (%)	4974 (79.4)	4989 (79.5)
rrent smoker — no. (%)	781 (12	
New di		395 (6.3)
Impaired glucose tolerance or impaired fasting glucose — no. (%)	735 (11.7)	717 (11.4)
Duration of diabetes — yr	5.5±6.1	5.3±5.9
Fasting plasma glucose — mg/dl		
Median	125	124
Interquartile range	109–148	108–148
Glycated hemoglobin — %		

Reproducerbar forskning

- R
- R studio
- Rmarkdown