INDFØRING I PRINCIPPER OG METODER TIL KVANTITATIV EFFEKTMÅLING

Læringsseminar 12

Tine Hjernø Lesner
Rasmus Højbjerg Jacobsen

13. september 2018

Hvem er vi?

Tine Hjernø Lesner

Ph.d., specialkonsulent i Center for Data, Analyse og Metode, Socialstyrelsen

Rasmus Højbjerg Jacobsen

Ph.d., projektchef i VIVE – Det Nationale Forsknings- og Analysecenter for Velfærd

Program for læringsseminaret

Del 1

- Hvad er en effekt?
- Evidensstigen
- Lodtrækningsforsøg (RCT)

PAUSE – ca. 10 min.

Del 2

- Difference-in-differences (DiD)
- Regression discontinuity design (RDD)

PAUSE – ca. 10 min.

Del 3

- Matching
- Case

Hvorfor vil vi gerne effektmåle?

Vi ønsker at vide, om en given indsats rent faktisk har haft den virkning, som var hensigten med indsatsen.

Og vi vil også gerne sikre os, at indsatsen ikke har en skadelig virkning.

EKSEMPEL:

Vi vil gerne vide, om en iværksat indsats over for unge kriminelle har reduceret deres kriminalitet.

Hvad er en effekt?

Når man taler om *effekt*, taler man også automatisk om *kausalitet.*

At en indsats har en (positiv) effekt, betyder:

1) at de personer, der har fået indsatsen har oplevet et positivt skift, som de ikke ville have oplevet, hvis de ikke havde fået indsatsen
2) at det er indsatsens indhold, der har forårsaget det positive skift

Hvad er en effekt?

Der er ofte mange tal i medier mm., som ikke er en effekt, fx

ERVISNING. Eleverne Tobias, Lasse og Leo fra 4.d på Brønshøj Skole i København er trætte i sidste skoletime
or, de ligger ned. De har matematik og er i gang med at måle areal af en presenning.

8 ud af 10 elever synes, at deres skoledage er for lange

82 procent af eleverne i folkeskolen synes, at deres dage er al
lange. Elever og forældre savner variation i undervisningen. M
er klar med flere penge og ny lov.

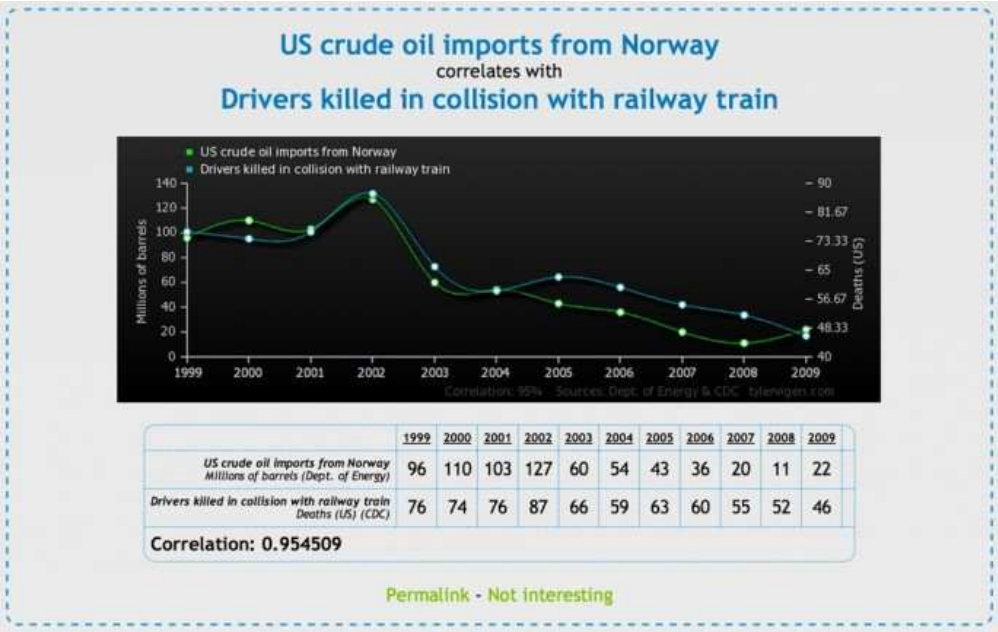
EKSEMPEL, fortsat:

2 pct. af alle 10-17-årige har været mistænkt eller sigtet for noget kriminelt i 2015 (Justitsministeriet, 2016).

Hvad er en effekt?

Det er altså kun en effekt, når vi kan tage højde for/vurdere, hvad der ville være sket, hvis der ikke havde været en indsats.

Det er relativt nemt at finde ud af, om der er en statistisk sammenhæng (korrelation) imellem indsats og udfaldsmål.



	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
US crude oil imports from Norway Millions of barrels (Dept. of Energy)	96	110	103	127	60	54	43	36	20	11	22
Drivers killed in collision with railway train Deaths (US) (CDC)	76	74	76	87	66	59	63	60	55	52	46

Correlation: 0.954509

Permalink - Not interesting

Hvad er en effekt?

Dvs. man skal have en sammenligningsgruppe (et kontrafaktisk udfald) for at kunne fastslå en effekt.

Sammenligningsgruppen kan dannes ud fra mange metoder, fx

- Før-efter-måling (man er sin egen sammenligning)
- Matching
- Lodtrækning
- Etc.

Vi skal se på nogen af dem i dag.

Udfaldsmål

I mange tilfælde er udfaldsmålet for en effektmåling relativt let at bestemme:

- Beskæftigelsesindsats
- Medicinsk behandling
- Stofmisbrugsbehandling

Men nogle udfald lader sig ikke så let måle (fx trivsel), og i andre tilfælde kan der være intermediære udfald, som er vigtige at få med (måske især hvis der kan være tvivl om, hvorvidt endemålet vil blive nået).

EKSEMPEL:

Vi vil gerne vide, om en iværksat indsats over for unge, der måske er ved at påbegynde en kriminel løbebane har reduceret deres kriminalitet.

Lad os forestille os, at der i 2010 blev iværksat en indsats over for målgruppen.

Vi vil måle på, om deres kriminalitet er reduceret i 2015.

Som udfaldsmål vælges fx: hvorvidt de unge er blevet dømt for kriminalitet.

Evidensstigen

Systematisk litteraturoversigt

Randomiseret forsøg (RCT)

Design, der renser for målbare faktorer

Før- og eftermålinger

Eftermålinger

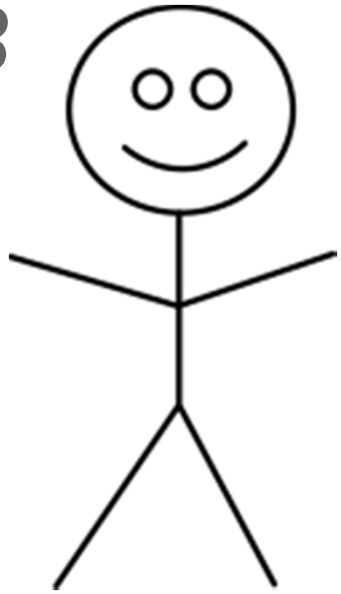
Det randomiserede kontrollerede forsøg (RCT)

Opsamling: Evalueringsproblem 1
Den kontrafaktiske situation

Indsats

Preben

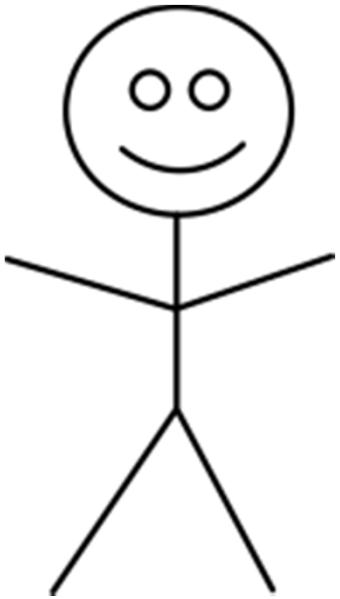
§



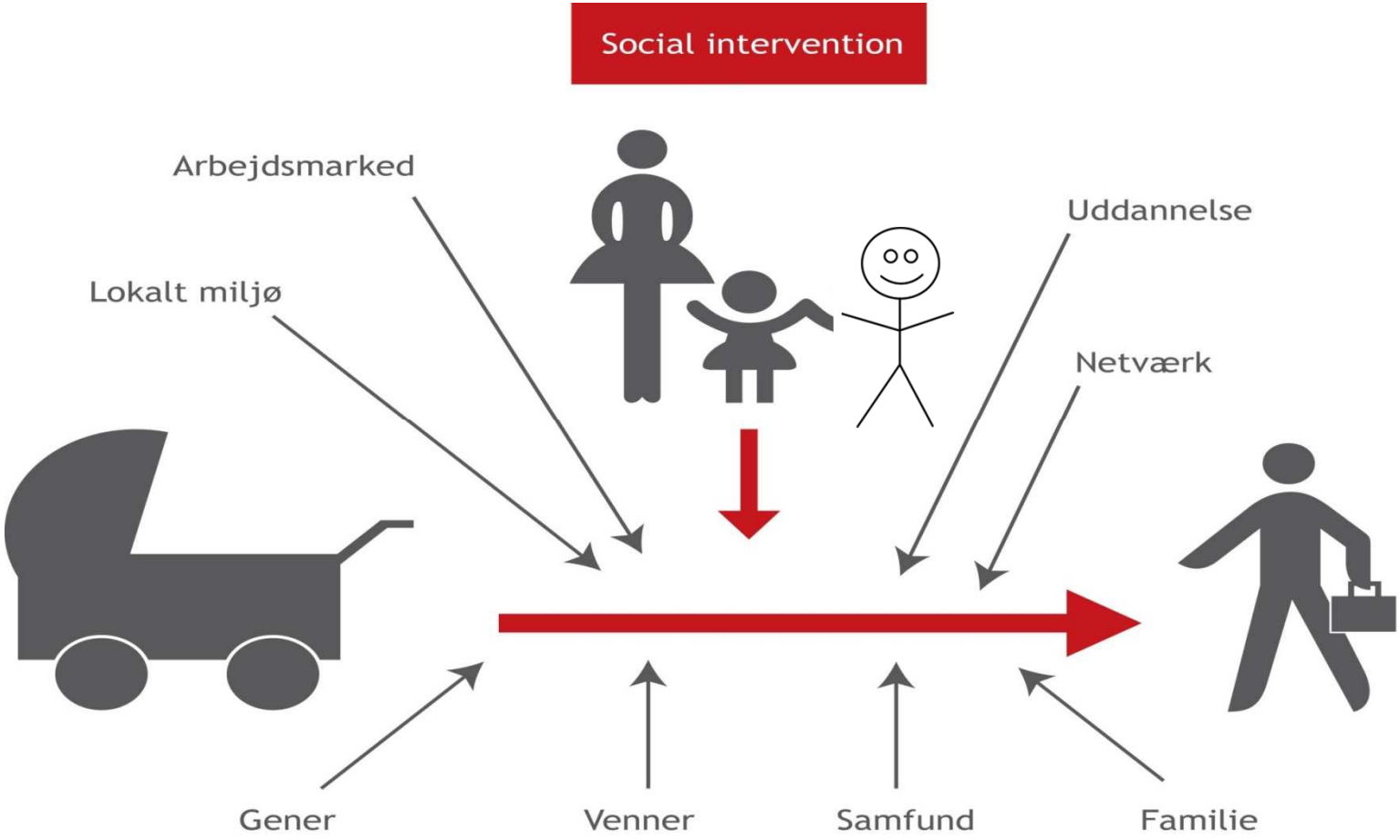
VS.

Kontrol

Kontrafaktisk
Preben



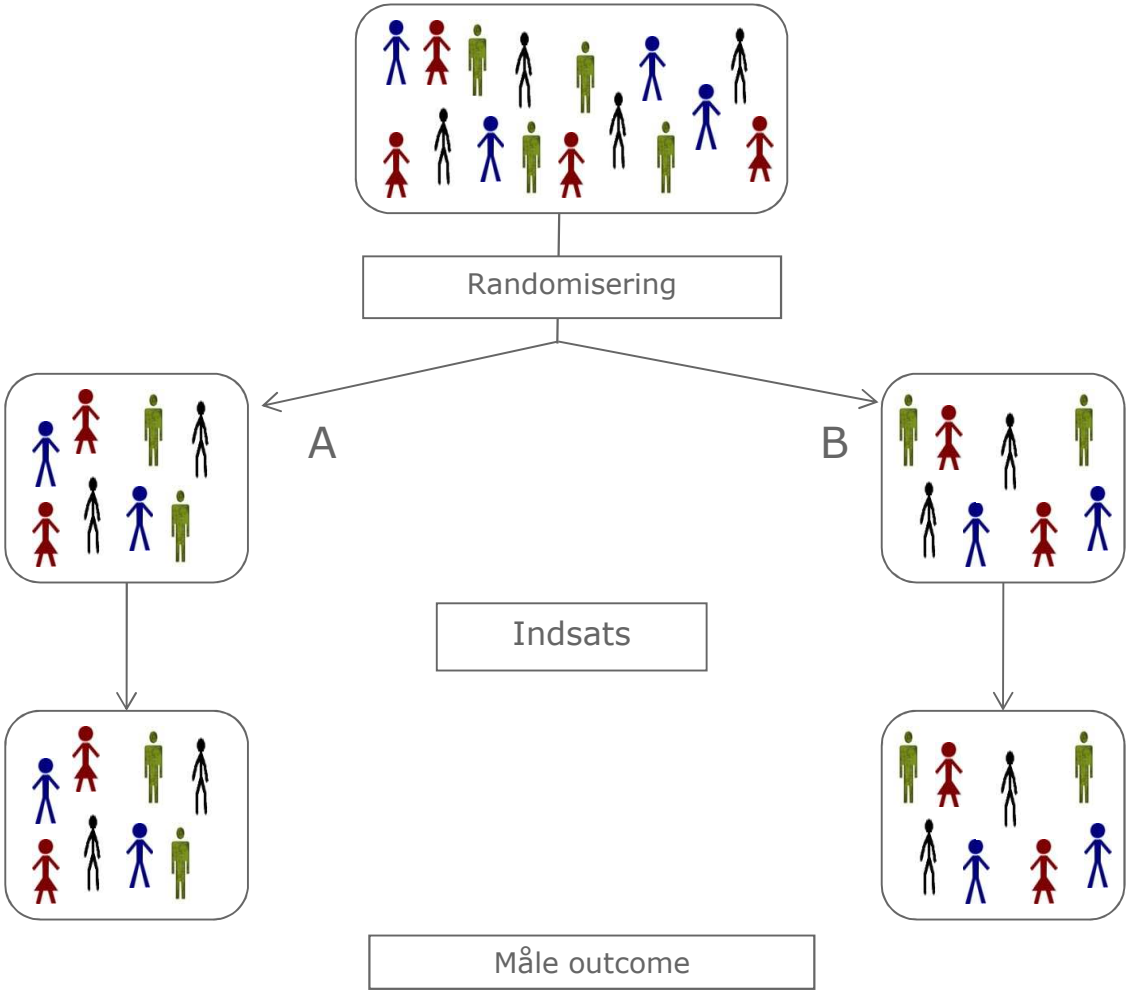
Opsamling: Evalueringsproblem 2
Preben påvirkes af andet end indsatsen



DET RANDOMISEREDE KONTROLLEREDE FORSØG

- Kaldes også lodtrækningsforsøg eller *et eksperiment*

- Den tilfældige tildeling i indsats- og kontrolgruppe sikrer, at grupperne er ens på både målbare og ikke-målbare faktorer
- Løser evalueringsproblem 1 og 2
  - Der findes en kontrafaktisk situation
  - Confounders er ens for begge grupper
- Kan fastslå et kausalitetsforhold


Randomisering
A
B
Indsats
Måle outcome

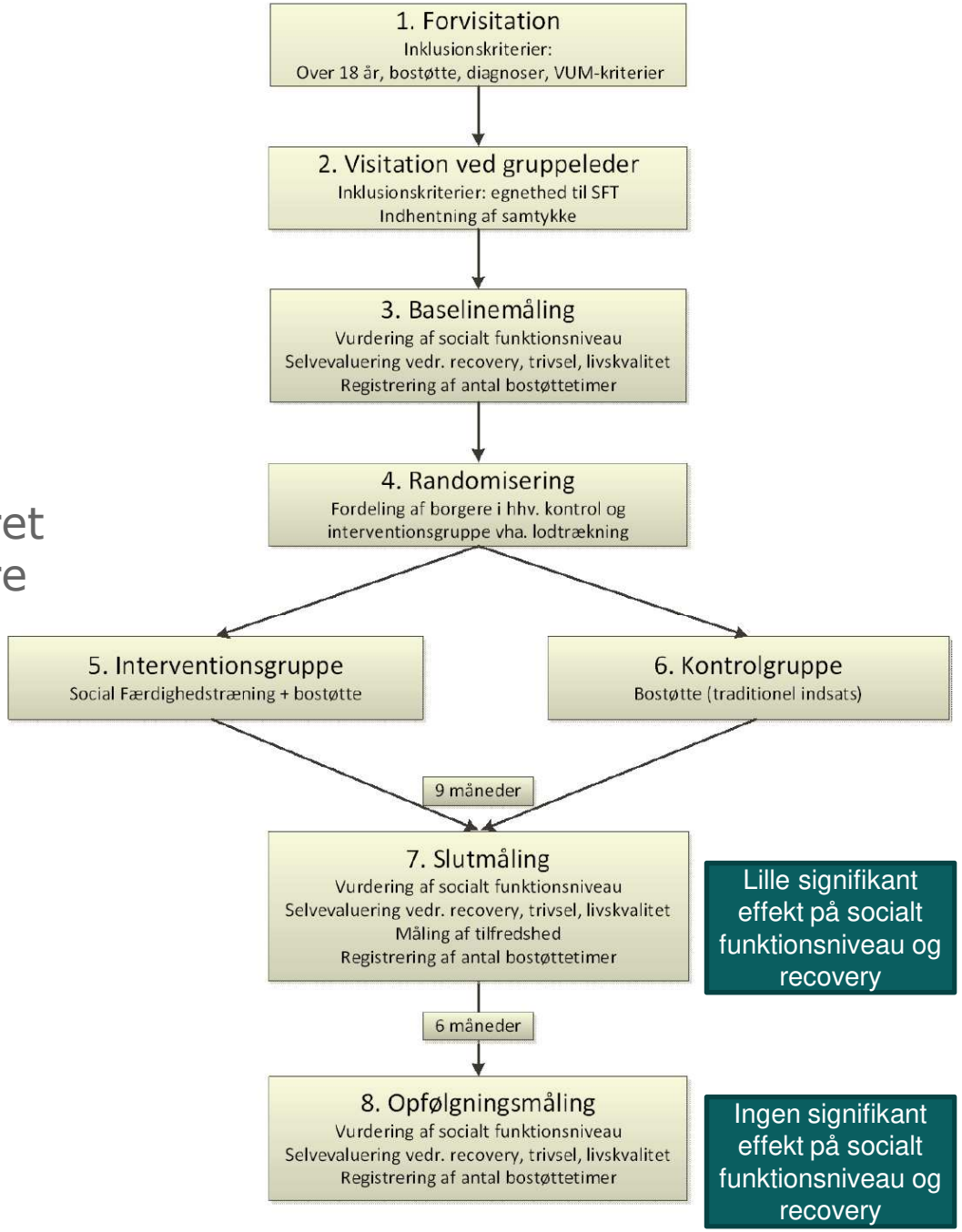
Eksempel på RCT:

Social Færdighedstræning

En velafprøvet og manualiseret metode til borgere med svære psykiske lidelser.

Evalueret af DEFACTUM for Socialstyrelsen.

304 borgere fra 9 kommuner blev randomiseret til indsats- eller kontrolgruppe.



Eksempel på RCT

Kærlighed i Kaos

KiK er et forældretræningsprogram, der henvender sig både til familier, hvor barnet har en ADHD-diagnose, men også familier, hvor barnet har ADHD-lignende vanskeligheder.

Et projekt med en klart defineret indsats og målgruppe.

Ventelistedesign: Familierne randomiseres til at modtage indsatsen nu eller senere.

- Fordel: Alle får indsatsen
- Ulempe: Ingen mulighed for langtidsopfølgning

Refleksion

Har I erfaringer med lodtrækningsforsøg?

Er der områder hos jer, hvor I kunne forestille jer at bruge det? Hvorfor/hvorfor ikke?

Hvor oplever/forudser I særligt problemer?

ANTAL DELTAGERE

- Vigtigt: Indsats- og kontrolgruppe skal være ”store nok”.
- Hvor store er ”store nok”?

- Beregning af stikprøvestørrelse er vigtigt fordi:
  - Stikprøven ikke må være for lille: Det er nødvendigt at have mange nok i indsats- og kontrolgruppen for at kunne identificere kausal effekt af en indsats.
  - Stikprøven ikke bør være for stor: Jo større stikprøve, desto dyrere er studiet.

- Hvor mange personer vi skal have med i forsøget hænger sammen med den effektstørrelse, vi forventer at finde:

  - Jo større effektstørrelse ⇒ jo større styrke/power.

  - Jo større effektstørrelse ⇒ jo færre personer behøver vi for at kunne vise en forskel

ANTAL DELTAGERE

- Tommelfingerregler (ikke absolut!):

- Stor effekt: 25 personer i hver gruppe (Cohen's d=0.80)
- Moderat effekt: 64 personer i hver gruppe (Cohen's d=0.50)
- Lille effekt: 400 personer i hver gruppe (Cohen's d=0.20)

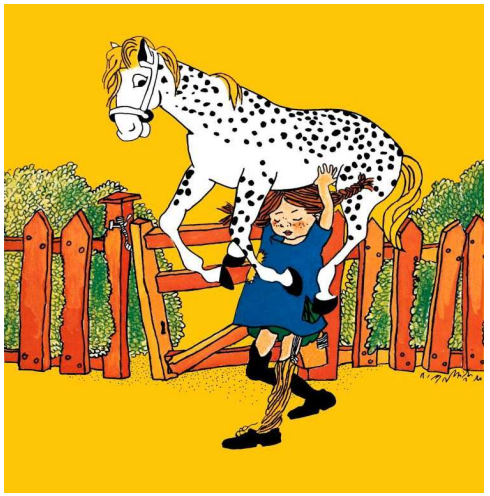
- OBS på "klyngeeffekter" og frafald

STYRKEBEREGNINGER

Der findes forskellige værktøjer til styrkeberegninger på nettet.

Prøv eventuelt:

Optimal Design eller http://www.uccs.edu/~lbecker/



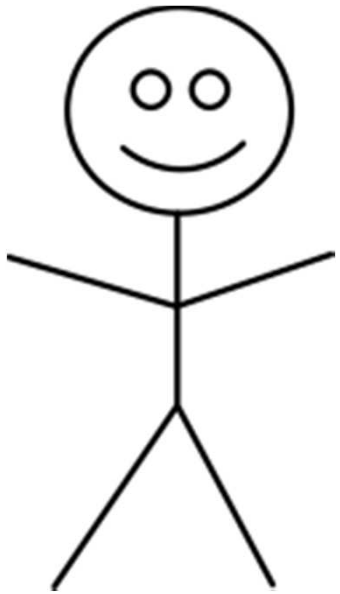
Før-efter-måling

- Fra toppen til bunden på evidensstigen

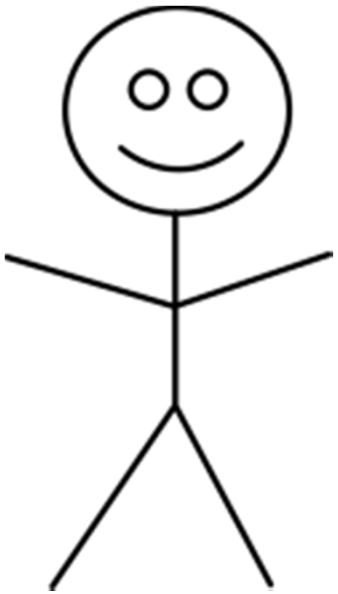
Før-efter-måling

I en før-efter-måling er Preben sin egen kontrol. Dog er der tale om Preben før indsatsen.

Kontrol Preben (år 0)

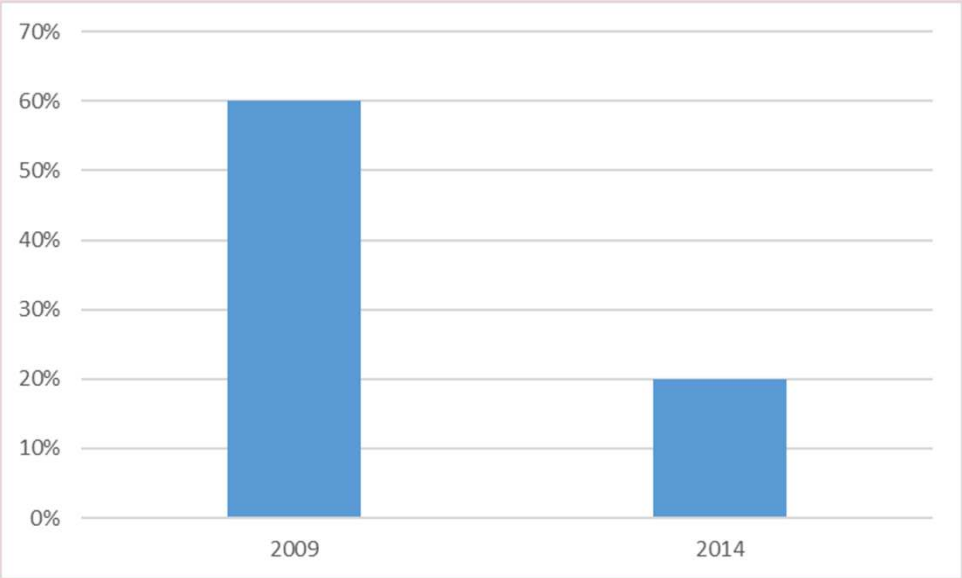


Indsats Preben (år 1)



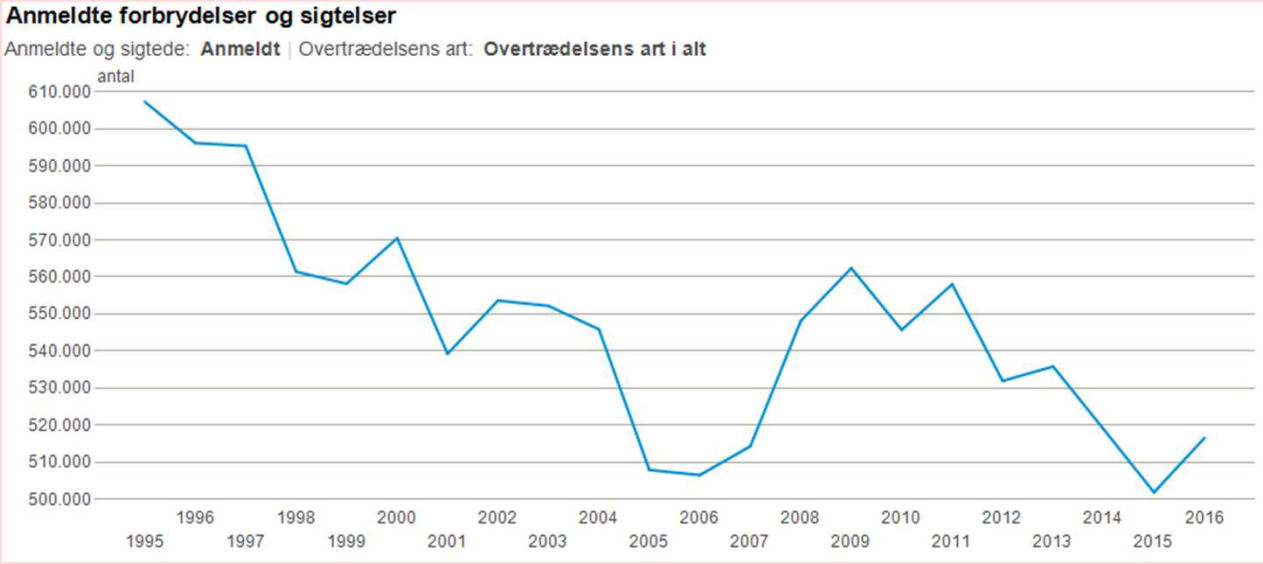
EKSEMPEL, fortsat:

Ved en før-efter-måling vil man opgøre, hvor mange af de unge, der har begået kriminalitet, før indsatsen igangsættes, og efter den afsluttes.



EKSEMPEL:

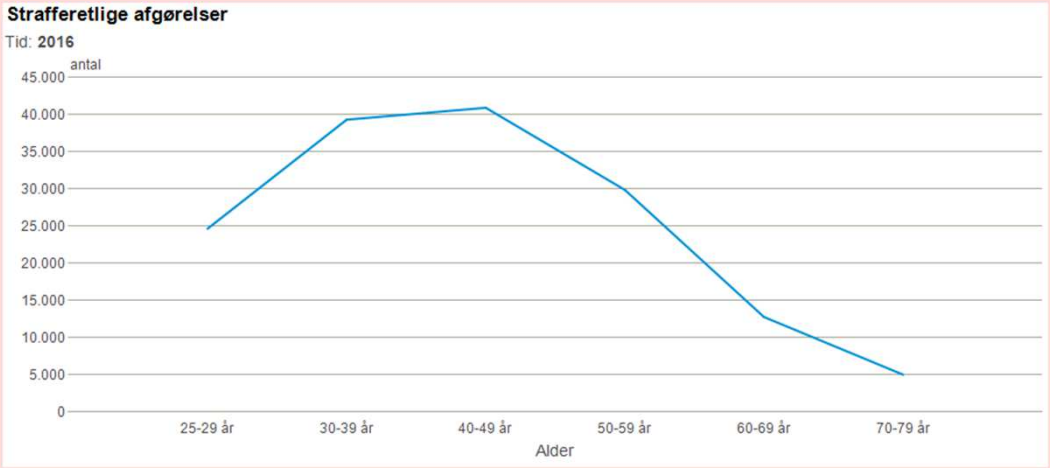
Der har været et stort fald i anmeldte forbrydelser fra 2010 til 2015:



Et fald i kriminaliteten for deltagerkunne altså lige så godt været forårsaget af det generelle fald i kriminaliteten.

EKSEMPEL:

Ud over problemet med et fald i kriminaliteten over tid, er der også en klar sammenhæng imellem kriminalitet og alder:



Måske er kriminaliteten blandt deltagerne ændret, blot fordi de er blevet ældre?

EKSEMPEL:

I dette eksempel kunne man altså relativt let stå med en kvantitativ måling, der viser, at kriminaliteten er faldet for deltagerne.

Så indsatsen må altså have været en succes? Nej, ikke nødvendigvis.

Dette eksempel er et godt eksempel på, hvorfor en før-efter-måling ikke nødvendigvis giver et retvisende billede.

Før-efter-måling kan ikke tage højde for faktorer, der ændrer sig over tid som fx alder og generelle tidstrends.

Derfor rangerer før-efter-måling ikke så højt på evidensstigen.

Før-efter-måling

Fordel:

- Nemmere at gennemføre, da man ikke behøver en kontrolgruppe

Ulempe:

- Kan ikke afgøre, om en udvikling skyldes indsatsen, eller om det er en anden tidsvarierende faktor

MEN: Vælg kun før-efter-måling, hvis det ikke er muligt at komme længere op på evidensstigen, og det er ret sikkert, at der ikke er andre tidsvarierende faktorer, der kunne påvirke outcome.

Refleksion

Arbejder I med før-efter-målinger?

Var det det bedste design, I kunne have brugt?

PAUSE

Difference-in-differences (DiD)

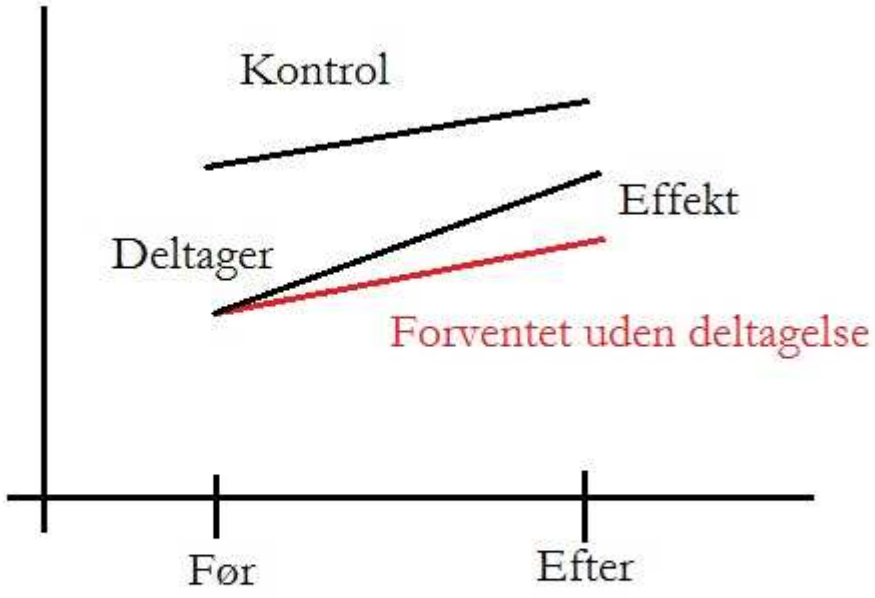
Difference-in-differences

Er i virkeligheden den avancerede før-efter-måling.

Ideen er at sammenligne med en kontrolgruppe og se på udviklingen over tid for begge grupper.

DiD - princip bag metode

Princippet ved DiD illustreres godt ved denne figur:



Centrale forudsætninger

Parallel trend

Dvs. udviklingen over tid ville have været den samme for grupperne, hvis der ikke havde været nogen indsats.

Indsats uafhængig af baseline

Niveauet for outcome-variablen i før-målingen må ikke være bestemmende for tildeling af indsats

DiD – fordele og ulemper

Fordele:

- Kan etablere kausalitet
- Intuitiv fortolkning
- Kontrol- og deltagergruppe behøver ikke nødvendigvis at have samme udgangspunkt
- Tager højde for, at outcomevariable kan være påvirket af andre tidsvarierende faktorer end indsatsen

DiD – fordele og ulemper

Ulemper:

- Kræver data fra før indsatsen igangsættes
- Kræver både deltager- og kontrolgruppe
- Kræver at parallel trend antagelsen er opfyldt

Eksempel DiD

Projekt om udeskole – evalueres af VIVE

Deltagergruppe: elever på de årgange, der deltager i udeskoleprojektet

Kontrolgruppe: elever fra andre kommuner, men på samme årgange og med samme baggrundskarakteristika

Målevariable: Testscore i de nationale test i dansk og matematisk samt den nationale trivselsmåling.

Regression Discontinuity Design

Regression discontinuity (RD)

Grundlæggende er RD en kvasieksperimentel metode, dvs.

- vi vil gerne opnå nogle af de samme statistiske egenskaber som en lodtrækning
- der bliver ikke foretaget lodtrækning

Idéen er, at der findes en underlæggende allokeringsvariabel, som påvirker individets deltagelse i en indsats, men hvor der ved en given tærskelværdi er forskel på, om man får indsatsen eller ej. Denne variabel kaldes også ”forcing”-variablen.

Regression discontinuity

Det er vigtigt, at allokeringsvariablen er uafhængig af udfaldsvariablen.

Allokeringsvariablen kan fx være alder, testscore, geografiske grænser mv.

Typisk er det sådan, at hvis *D* angiver deltagelse, så

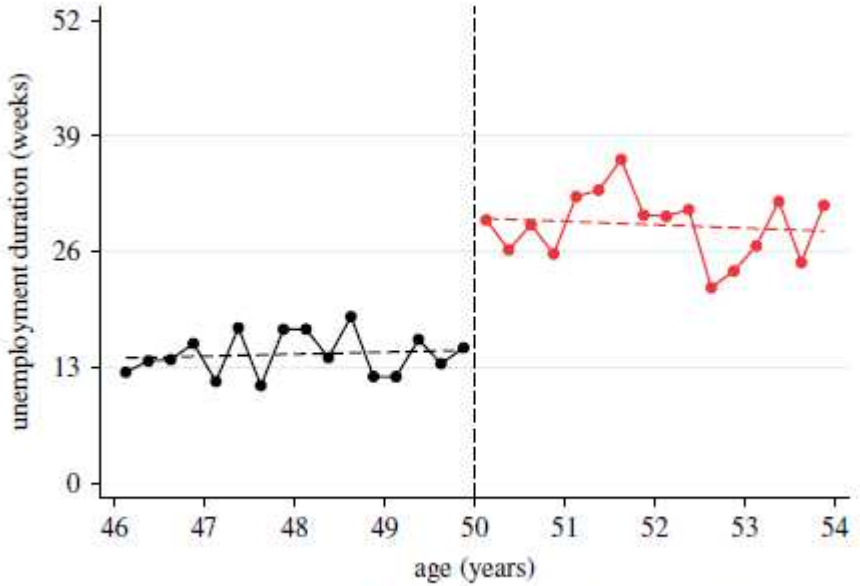
$$D = \begin{cases} 0 \; hvis \; x \leq c \\ 1 \; hvis \; x > c \end{cases}$$

hvor *c* er tærskelværdien for allokeringsvariablen.

Regression discontinuity

Eksempel (Lalive, 2008):

Østrig indførte et program, der tildelte personer over 50 år ret til længere arbejdsløshedsunderstøttelse:



Regression discontinuity

Generelt findes sådanne diskontinuiteter inden for

- Skattesystemet
- Offentlige overførsler
- Skolesystemet

Det er relativt let at se, om der er en ikke-kontinuitet ved at lave en figur, hvor allokeringsvariablen er på X-aksen og udfaldsvariablen er på Y-aksen.

Regression discontinuity

Det er vigtigt, at

- personerne ikke har fuld kontrol over, om de befinder sig på den ene eller den anden side af grænsen. Oftest vil fx skattesystemet kunne give incitamenter til at vælge, hvis man kan.

Hvis denne antagelse er opfyldt, er der imidlertid store fordele ved RD:

- Allokeringen til indsats/kontrol er helt tilfældig lige omkring tærskelpunktet. Det svarer til lodtrækning.
- Man kan teste, om antagelsen om tilfældighed er opfyldt. Hvis den er det, må der ikke være systematisk variation imellem indsats- og kontrolgruppe.
- Metoden kan (som et startpunkt) let lade sig fremstille i en grafisk præsentation. MEN: det siger ikke noget om effekt.

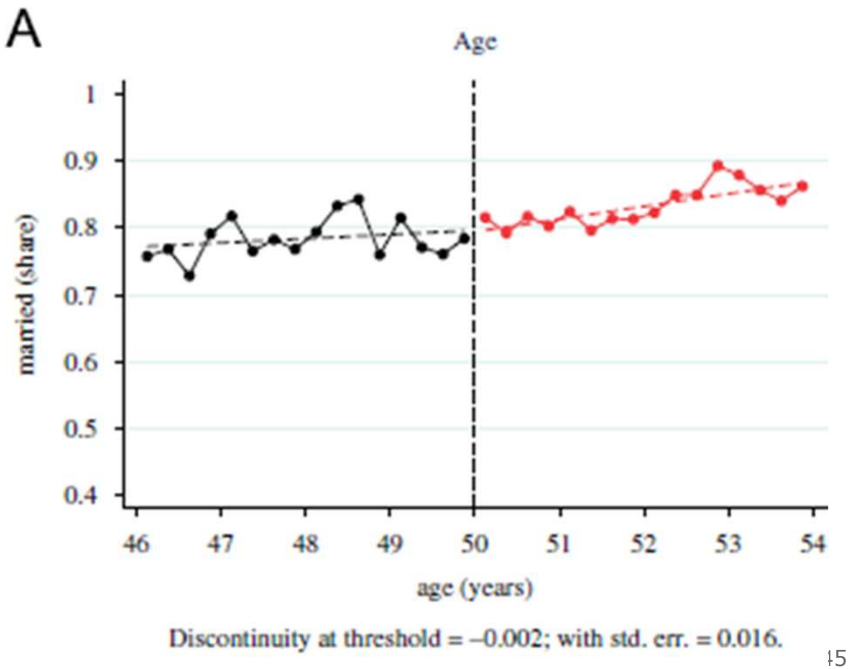
Regression discontinuity

En måde at undersøge, om individerne på hver side af tærskelværdien er tilfældigt fordel, er ved at lave graf med allokeringsvariablen på X-aksen og forklarende baggrundsvariable (fx køn, alder, uddannelse osv.) på Y-aksen.

Disse må ikke vise et spring ved tærskelværdien.

Figur fra Lalive (2008), andel gifte som funktion af alder.

Man skal huske at teste statistisk!



Fortolkning af RD-estimater

Der er to mulige fortolkninger af RD-estimatoren:

1. Hvis man finder en effekt ved tærskelværdien, så gælder den effekt, der observeres kun for individer med værdier meget tæt på tærskelværdien, fx for elever med en testscore tæt på værdien.

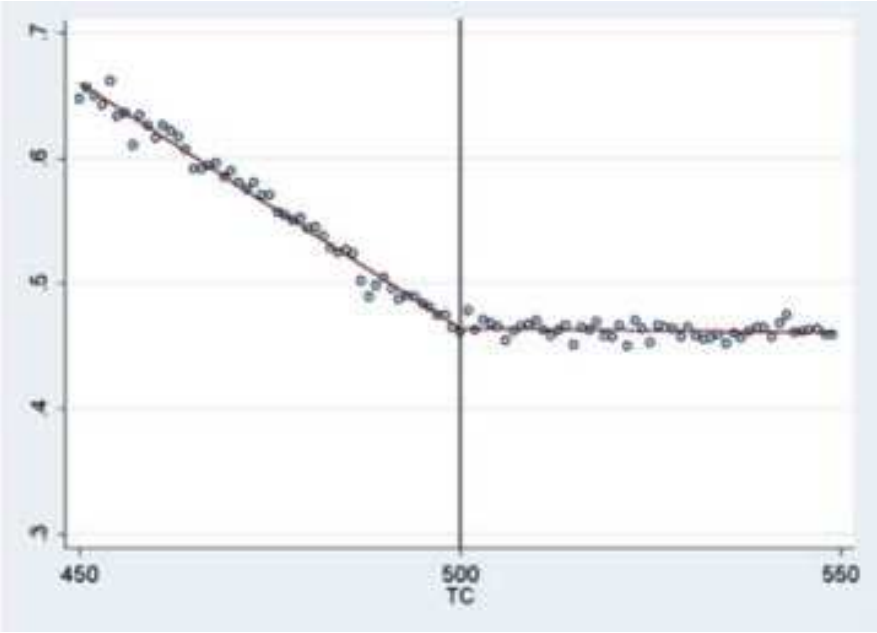
2. Estimatet kan fortolkes som en mere bred effekt af deltagelse ved at se estimatet som en vægtet værdi, med højere vægt, jo tættere man er på tærskelværdien.

Regression kink design

Regression kink design er et specialtilfælde af regression discontinuity, hvor der i stedet for en ikke-kontinuitet er et "knæk" for den variabel, der undersøges. Dvs. der er en ikke-kontinuitet i de første afledede mht. forcing variablen:

Fx viser Simonsen et al. (2015)

Totale omkostninger ved receptpligtig medicin.



Afsluttende pointer om RD og RK

1. Lav altid en grafisk præsentation og inspektion. Dette giver en god idé om, hvorvidt der er en ikke-kontinuitet, og om hvilken form denne har. Er den fx lineær eller ikke-lineær?
2. Er der visuelt et tegn på, at der optræder en effekt ved tærskelværdien? Dvs. er der forskel på, hvordan individer på hver side af grænsen optræder? Hvis ikke, så er der næppe nogen effekt.
3. Husk altid at teste for, om grupperne på hver side af tærsklen er ens. Man kan evt. starte med en ”grafisk test”.

Eksempler på RD i Danmark

Skolestart: Efter reglerne skal man starte i skole i det år, man fylder 6 år. Derfor vil der være nogle, der er født lige omkring d. 1. januar, som enten starter i skole eller ikke gør det (Simonsen et al., 2015).

Efterløn: Under indfasning af efterlønsreformen var der aldersgrænser for, om man var berettiget til efterløn eller ej.

Jobpræmie: Ledige, der havde været ledige mere end 47 uger kunne få en præmie på op til 4% af jobbets månedsløn (Kolodziejczyk og Arendt, 2017).

Geografi: Visse udkantskommuner har forhøjet befordringsfradrag for pendlere.

Eksempler på RD

Jobpræmie: Et statsfinansieret skattefrit beløb på 4 % af den arbejdsmarkedsbidragspligtige indkomst, eller maksimalt 600 kr. pr. måned.

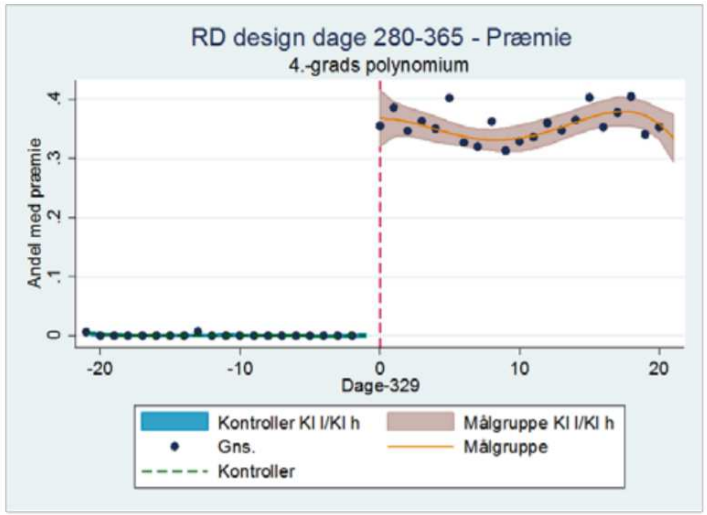
Målgruppe: Enlige forsørgere eller langvarige kontanthjælpsmodtagere med mindst 47 ugers modtagelse af offentlige ydelser set over et år.

RD er relevant, fordi der er en klar allokeringsvariabel, nemlig at man har været på offentlig forsørgelse mindst 47 uger.

Eksempel på RD

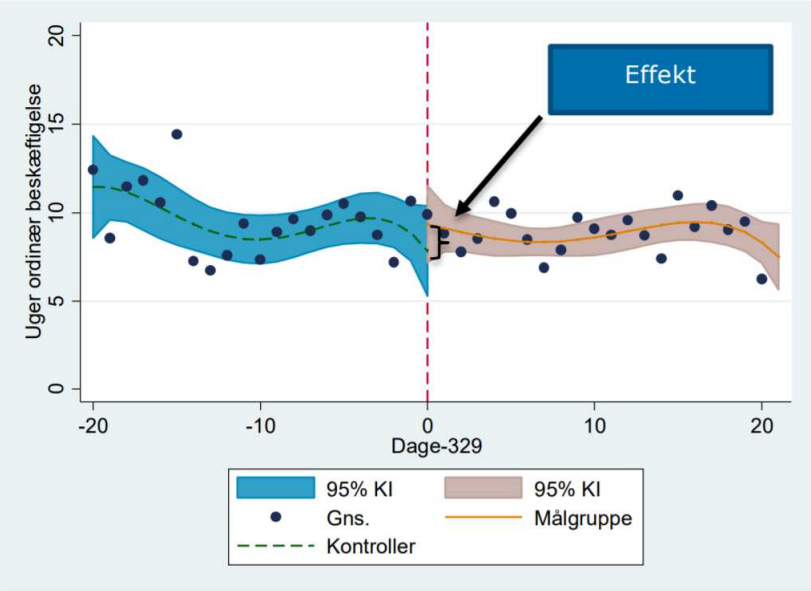
Grafisk illustration af allokeringsvariablen:

Bilagsfigur 5.1 Udbetaling af jobpræmie i kontrol- og målgruppe



Eksempel på RD

Figur 5.1 Beskæftigelseseffekt af jobpræmie for langvarige kontanthjælpsmodtagere



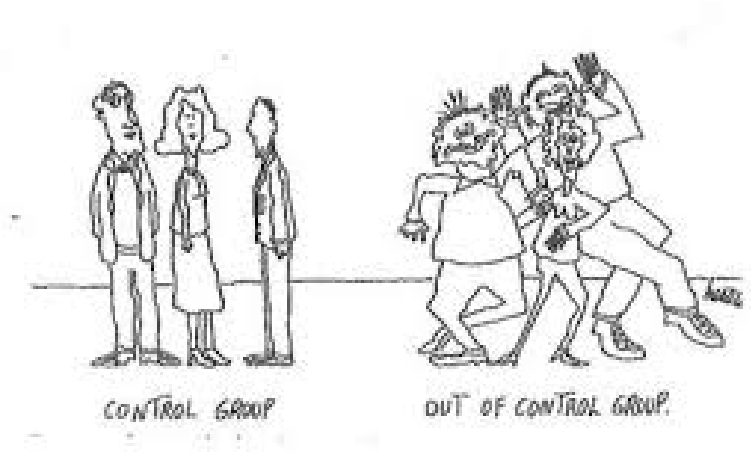
Note: Figuren viser uger i ordinær beskæftigelse i forsøgsperioden i forhold til dage fra kvalifikationskravet til ordningen. Hvert punkt er et gennemsnit, og der er estimeret et 4. grads polynomium gennem disse gennemsnit for både mål- og kontrolgruppen på hver side af tærsklen for kvalifikation til ordningen i population med 308-350 på indkomstydelse i kvalifikationsperioden. De skraverede felter er 95 %-konfidensintervaller.

Ikke nogen statistisk signifikant effekt (det gælder også de øvrige undersøgte områder: løntilskud, aktivering og lønindkomst)

PAUSE

Matching

MATCHING - princippet

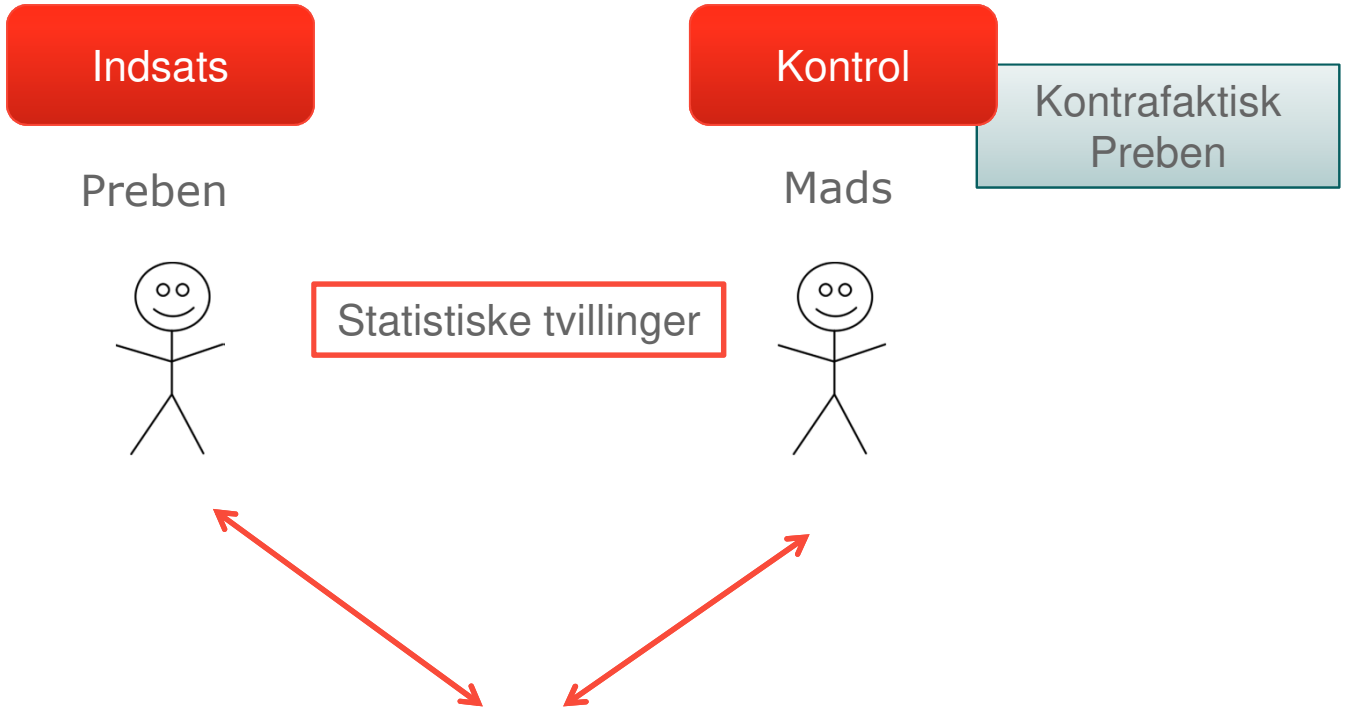


Kontrafaktisk Preben

Matching er en statistisk metode til at **konstruere** en kontrolgruppe,
som **ligner** en bestemt indsatsgruppe så meget,
at det er muligt at **sammenligne** outcomes for de to.

Preben

Matching I
Eksakt matching - den statistiske tvilling



Ens på observerbare karakteristika



FORDELE OG ULEMPER VED EKSAKT MATCHING

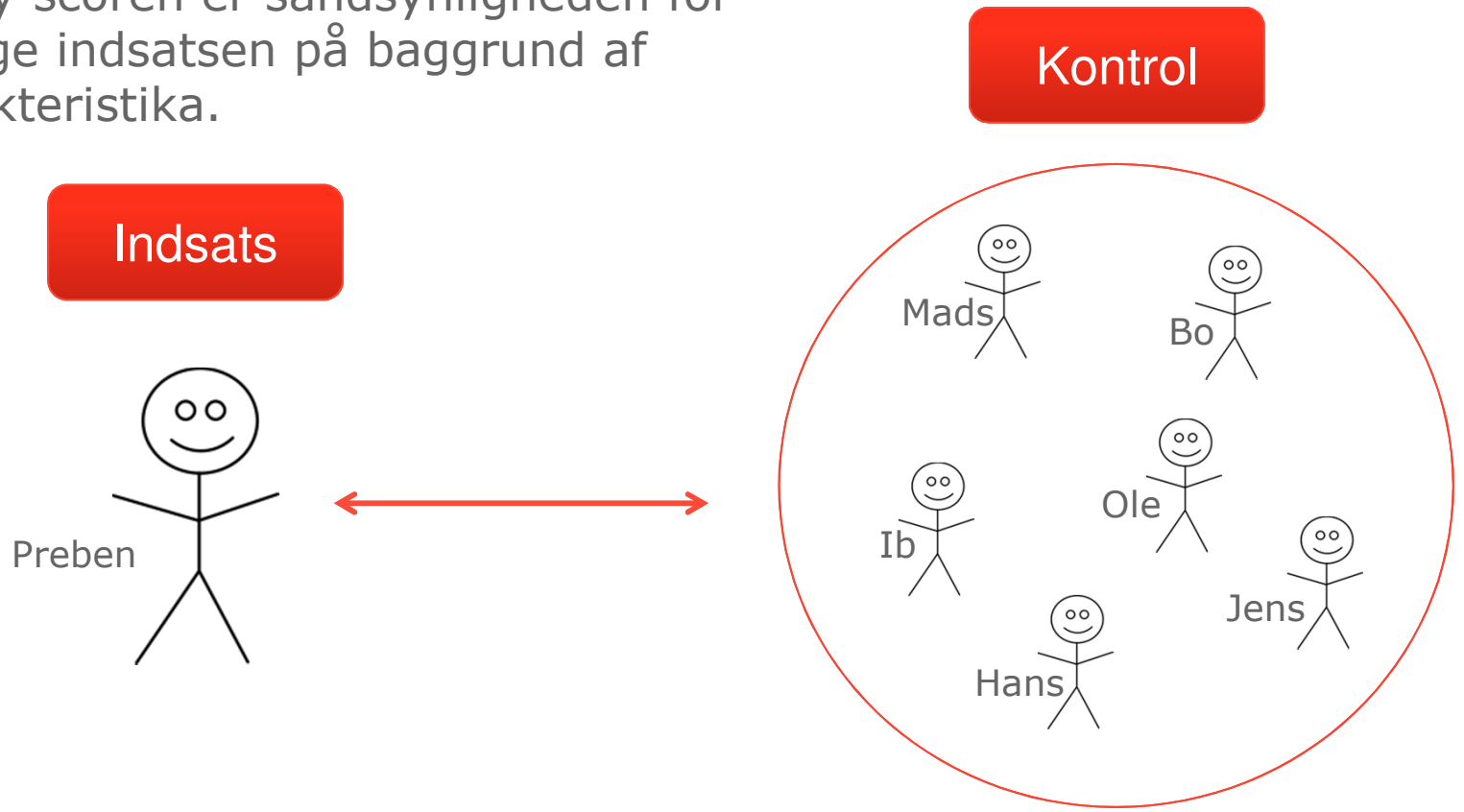
- Metoden er gennemskuelig

- Det bliver hurtigt svært at finde et eksakt match!!

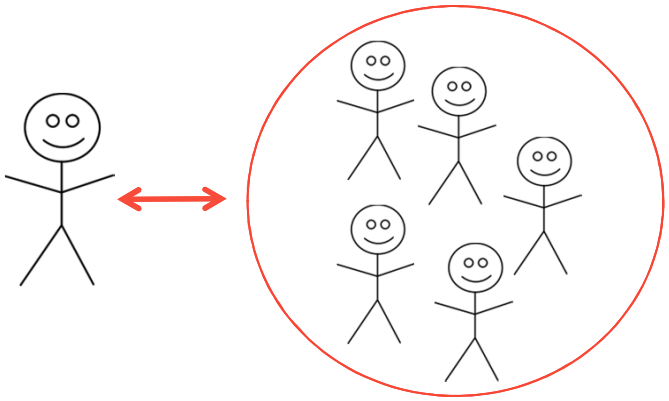
- Kræver at der faktisk er en gruppe der ligner, og derfor bedst til problemstillinger der findes i ”almindelige” (store) grupper

Matching II
Propensity score matching

Propensity scoren er sandsynligheden for at modtage indsatsen på baggrund af dine karakteristika.



PROPENSITY SCORE MATCHING



Hvilke variable skal der matches på (beregnes propensity score på)?

Dem, der er vigtige for selektion
- Fx alder, køn, bopæl, etnicitet etc.

Husk: Man kan ikke matche på sit outcome-mål (men gerne på målet før indsatsen).

FORDELE OG ULEMPER VED PS MATCHING

- Man kan finde en Kontrolgruppe, selvom der ikke er mange, der ligner 1:1 på individuelle faktorer

- Variable vægtes efter betydning

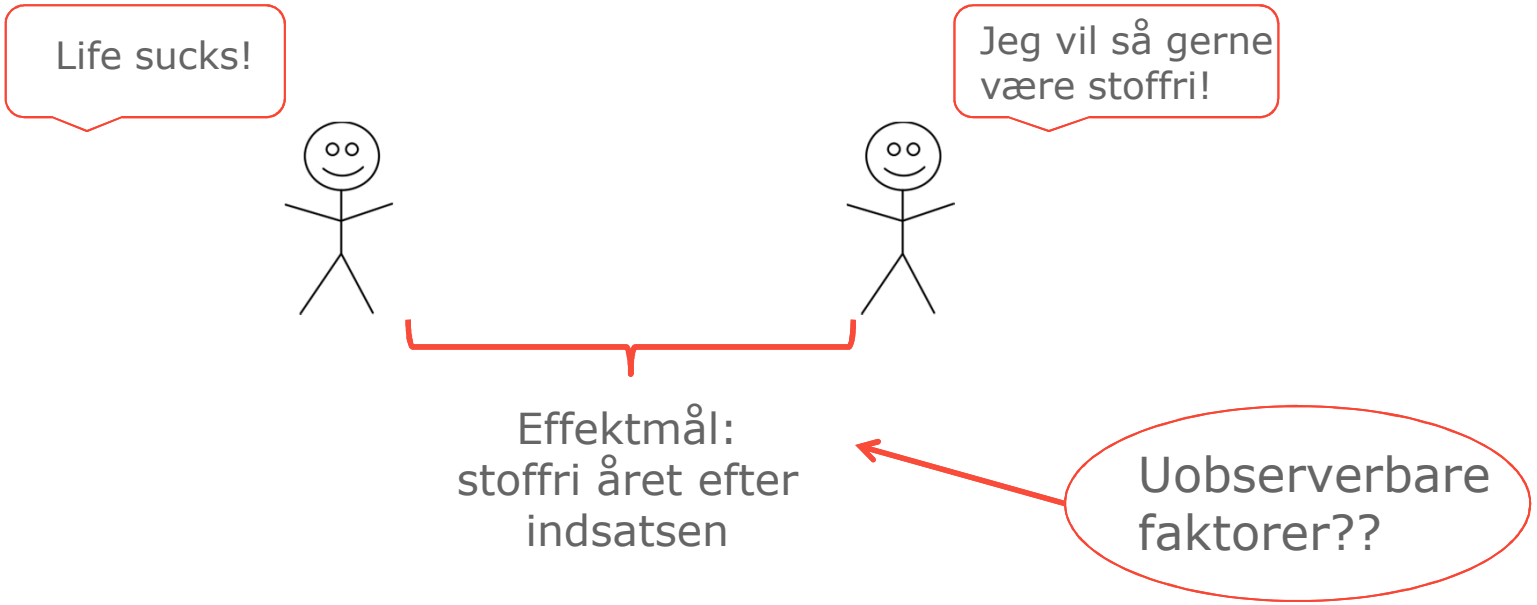
- Propensity scoren i sig selv er uigennemskuelig og svær at forklare intuitivt

- Metoden er datasulten

- De parametre, der indgår i PS, kan ikke indgå i selve analysen

VIGTIGE ANTAGELSER I

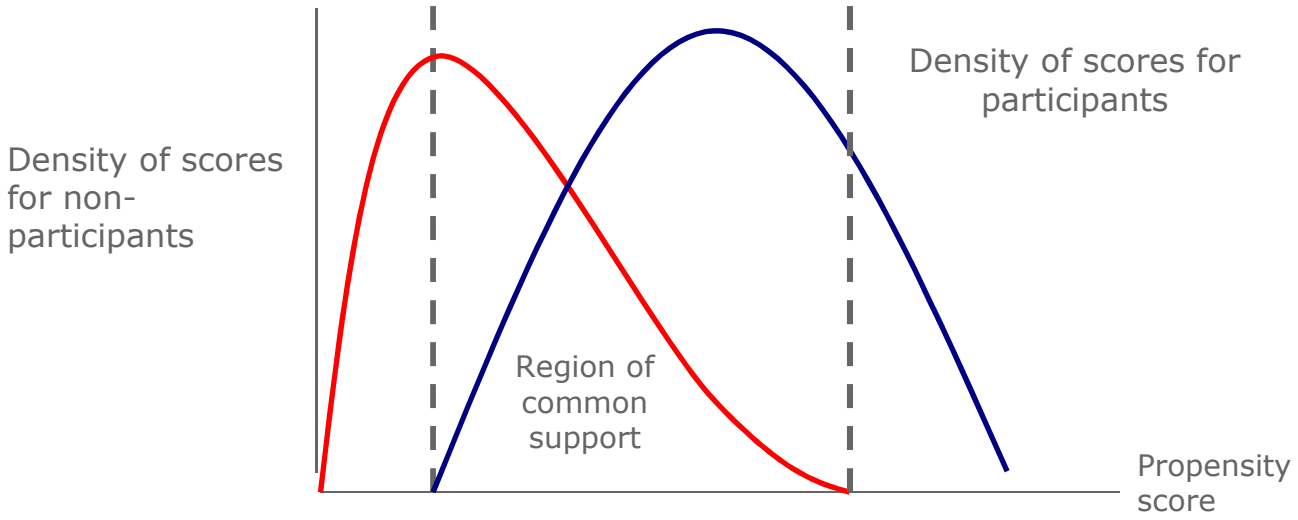
Conditional Independence Assumption (CIA)



Hvis uobserverbare faktorer har betydning for outcome og er forskellige for indsats og kontrol, holder CIA ikke => bias.
Typiske ”unobservables”: evne, motivation, drive…

VIGTIGE ANTAGELSER II

- Common support
  - For enhver værdi af matchingvariablene skal der være en positiv sandsynlighed for både at være treated og kontrol.
    - Sikrer at vi kun sammenligner sammenlignelige observationer
    - Testes ved visuel analyse af fordelingen af propensity scoren for treatment og kontrolgruppe



VIGTIGE ANTAGELSER III

- Balance på kovariater
- Har den estimerede propensity score haft succes med at eliminere forskelle i baggrundsvariable for indsats- og kontrolgruppen? Er de to grupper faktisk sammenlignelige?
- Tjekkes ved simpel statistisk test for hver variabel og samlet set – der må ikke være statistisk signifikant forskel på de to grupper efter matching.

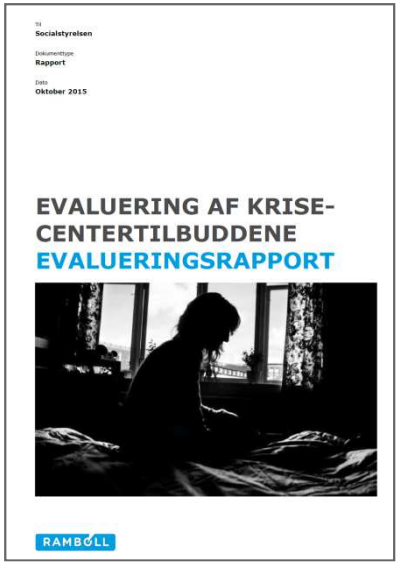
	Indsats	Kontrolgruppe før matching	Kontrolgruppe efter matching	% bias efter matching
Baggrundskarakteristika				
Alder (år)	14,687	14,884	14,750	-4,4
Pige	0,397	0,419	0,391	1,3
1. eller 2. generationsindvandrer (0/1)	0,157	0,116	0,161	-1,3
Bor med enlig forældre (0/1)	0,634	0,721	0,621	2,8
Antal børn i familien	1,911	1,169	1,935	-1,9
Mor lønmodtager (0/1)	0,522	0,419	0,533	-2,3
Mor modtager af overførselsindkomst (0/1)	0,253	0,288	0,257	-0,9

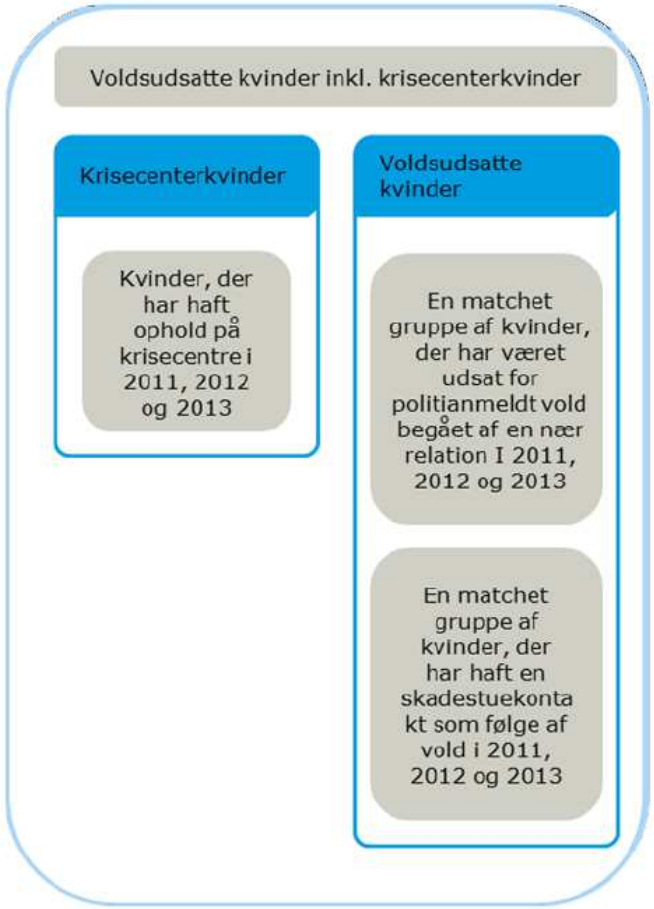
HVORNÅR KAN MAN ELLERS KOMME I PROBLEMER?

- Altid! Men det hjælper hvis:
- Indsatsen er veldefineret og velafgrænset
- Indsatsen er velimplementeret (ikke i pilotfase)
- Der ikke er mange indsatser samtidig
- Målgruppen ikke er for lille
- Man har masser af data på individniveau på matchingvariable og outcomes
  - ➢ Registrene er meget velegnede til matching

Eksempel I: Evaluering af kvindekrisecentre Socialstyrelsen og Rambøll (2015)

Effektmål:
- Vold
- Uddannelse
- Beskæftigelse
- Sundhed





Variable i propensity score beregning

- Alder
- Etnicitet
- Bruttoindkomst
- Lønindkomst i året før opholdet
- Gennemsnitlig ledighedsgrad i tre år før opholdet
- Uddannelsesniveau
- Antal partnere de seneste 10 år før opholdet
- Antal flytninger de seneste 10 år før opholdet
- Om kvinden bor sammen med en partner primo året for krisecenteropholdet
- Antal børn
- Om kvinden er dømt for ikke-færdselsrelateret kriminalitet de seneste 10 år før opholdet

Summeopgave - krisecentre

- Designet i denne evaluering er ikke perfekt

- Har I kommentarer til?
  - Populationen: Udvælges kontrolgruppen fra en hensigtsmæssig population? Er der nogen problemer?
  - Kun 75% af alle kvinder på krisecenter oplyser deres CPR-nummer – og indgår derfor i indsatsgruppen. Er det et problem for analysen?
  - Holder Conditional Independence Assumption eller er der mon uobserverbare faktorer, som ikke tages højde for? Evalueringen er baseret på registerdata.

Gruppearbejde

- Diskuter casen