Socialstyrelsen

SFI DET NATIONALE FORSKNINGSCENTER FOR VELFÆRD

INTRODUKTION TIL KVANTITATIV EVALUERING

Helle Hansen, SFI

Tine Lesner, Socialstyrelsen

PROGRAM

10.00-10.45	Velkomst Hvad er randomiserede kontrollerede forsøg? - Når det går godt – og når det går knap så godt.. Før- og eftermålinger
10.45-11.00	Pause
11.00-11.45	Naturlige eksperimenter Regression Discontinuity Design
11.45-12.00	Pause
12.00-13.00	Matching Hvordan fortolker og formidler vi resultaterne?

PRÆSENTATION AF OS OG JER

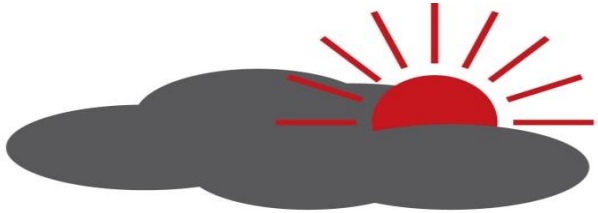
HVAD ER DIT NAVN?

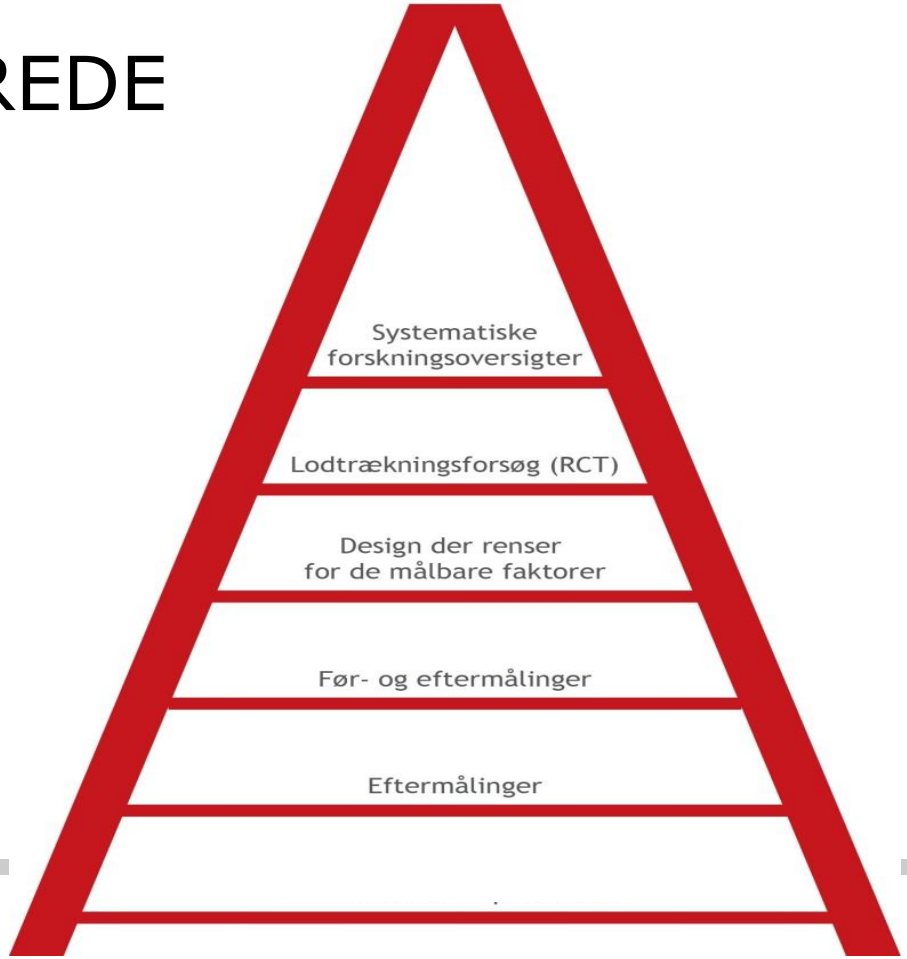
HVOR ARBEJDER DU?

HVORFOR HAR DU MELDT DIG TIL LÆRINGSSEMINARET?

HVORFOR SNAKKER VI SÅ MEGET OM RANDOMISEREDE FORSØG?

DET RANDOMISEREDE FORSØG

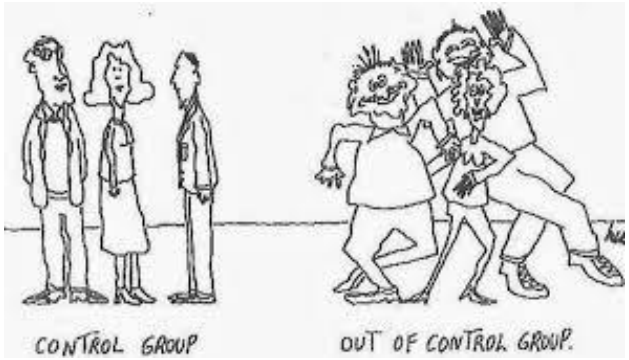




DET RANDOMISEREDE KONTROLLEREDE FORSØG

Kaldes også lodtrækningsforsøg eller *et eksperiment*

- Det bedste design til effektmåling - Pga. den tilfældige tildeling af indsats.
- Sikrer at grupperne er ens på både målbare og ikke-målbare faktorer
- Kan håndtere kompleksitet
- Kræver færrest deltagere



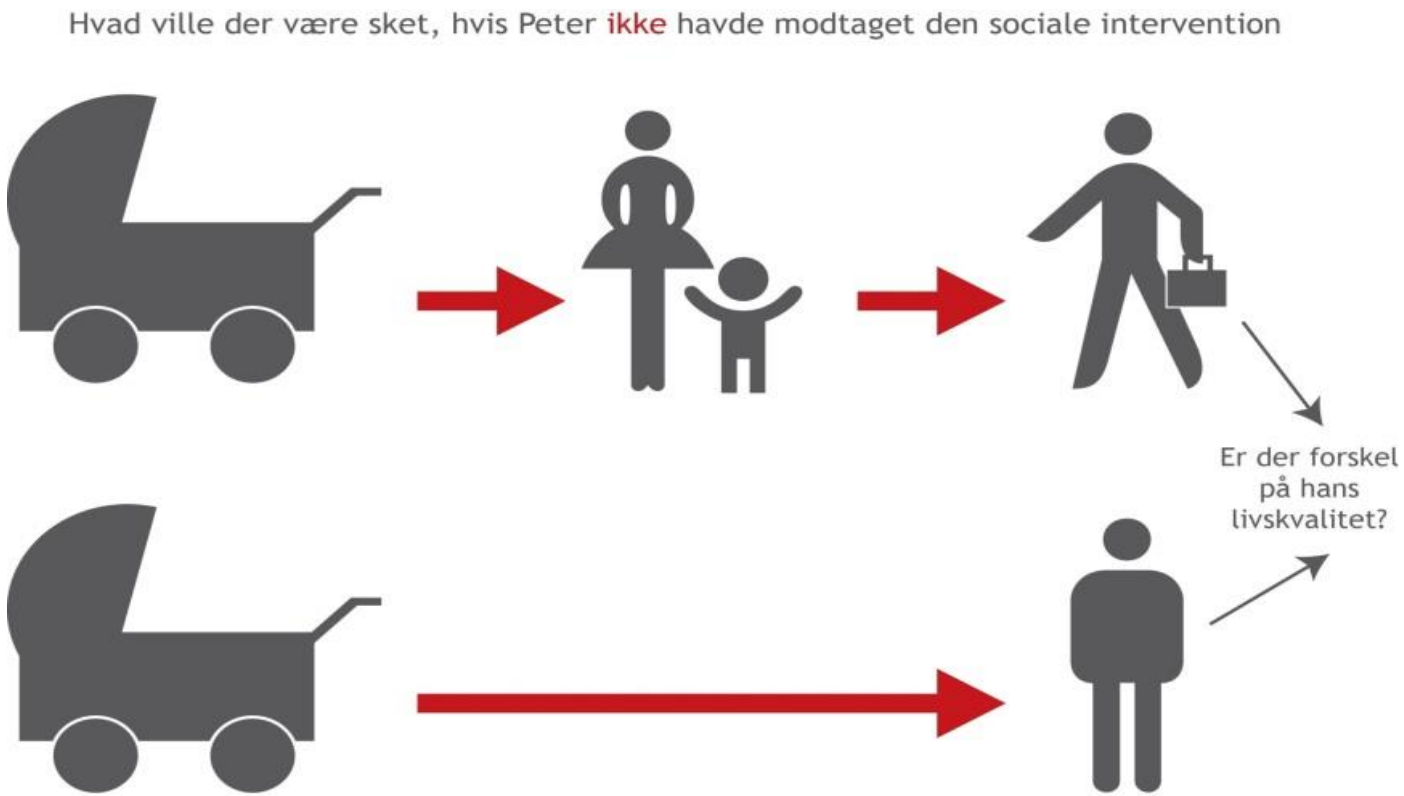
HVORFOR RCT?

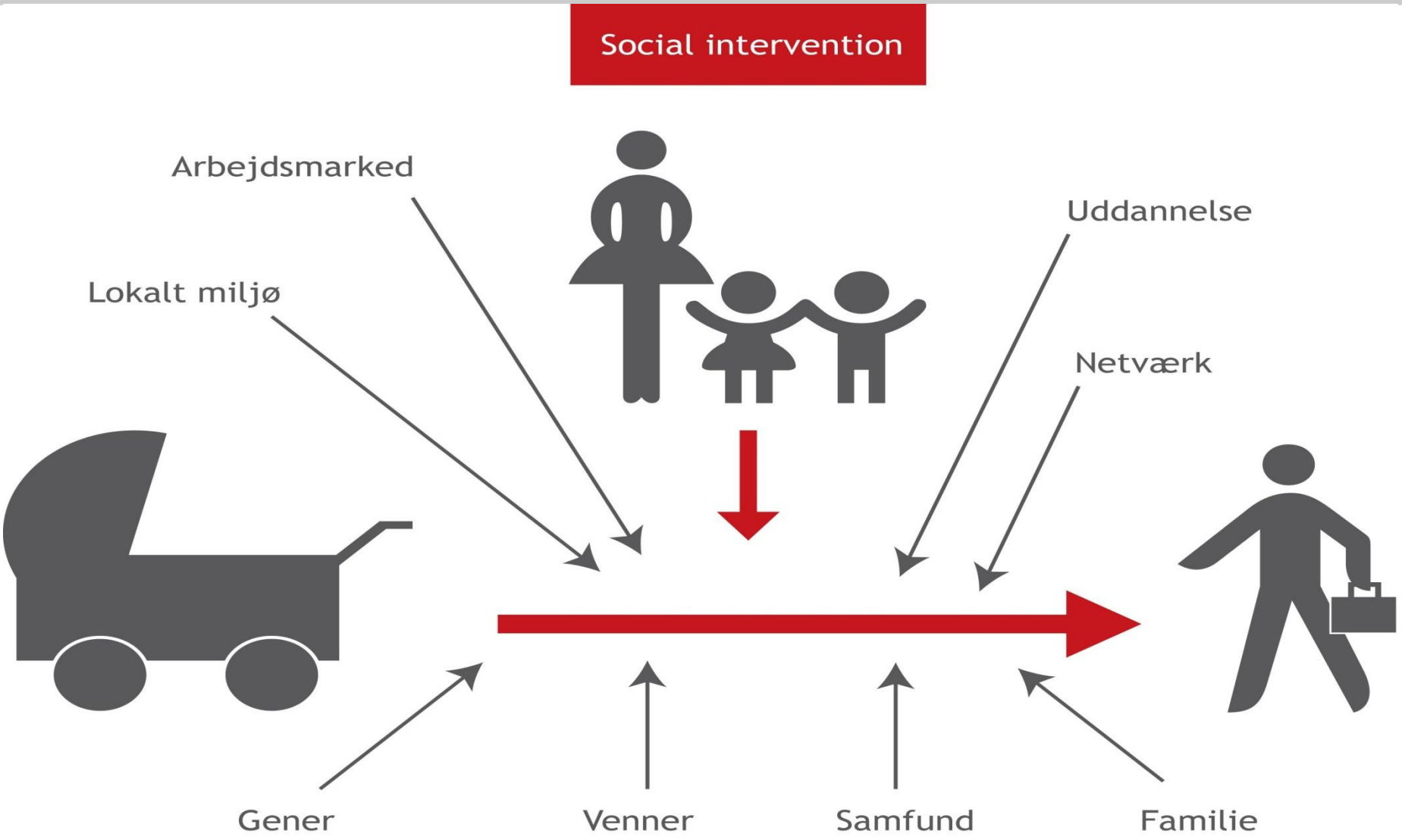
Vi kan ikke blot se en *sammenhæng* eller *korrelation* mellem indsats og effektmål.

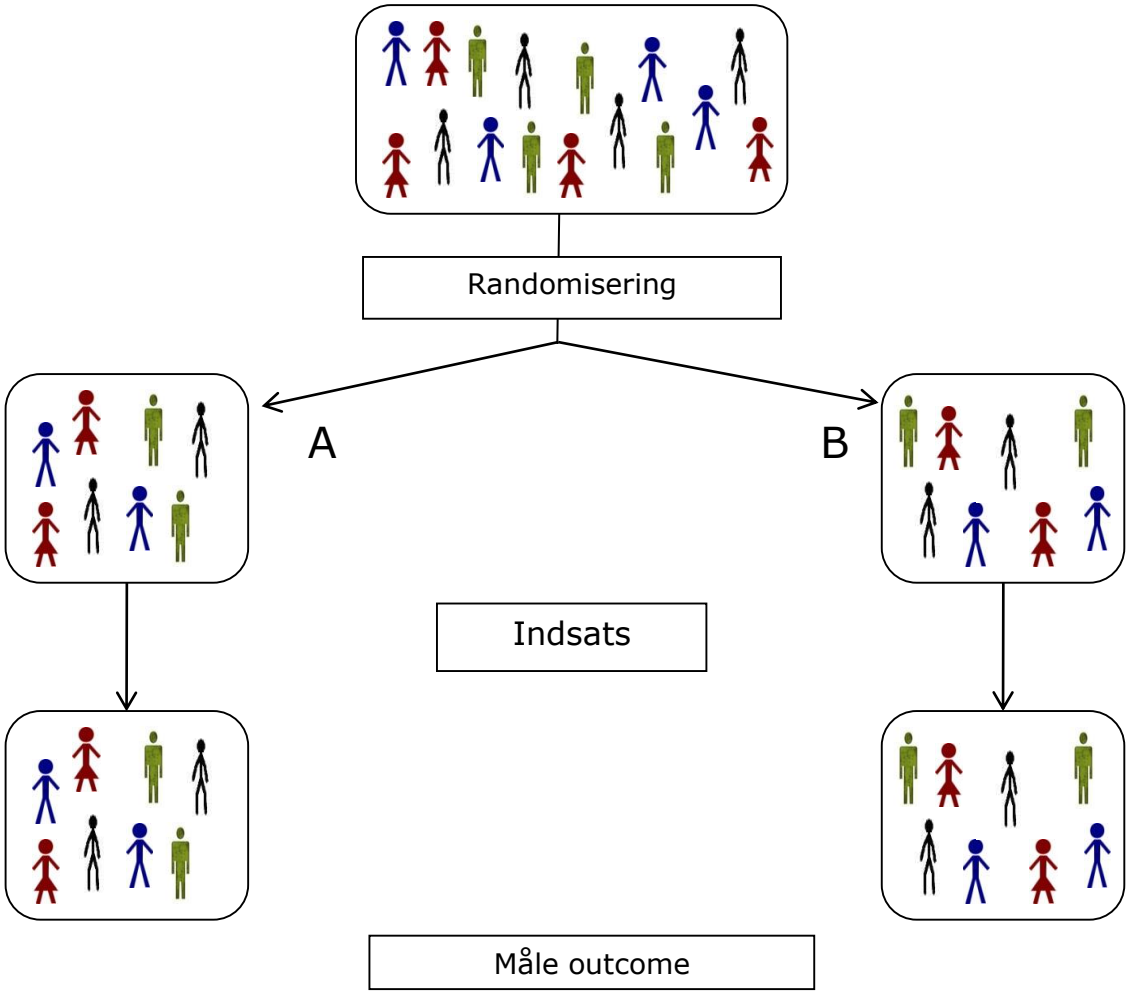
Fordi vi ved, at den eneste forskel på indsats- og kontrolgruppe er indsatsen, kan vi fastslå et *kausalitetsforhold*

Hvis indsatsgruppe klarer sig bedre/dårligere end kontrolgruppen, så kan vi konkludere, at det er på grund af indsatsen

At finde den kausale effekt af indsats på udfaldsmål




Social intervention
Arbejdsmarked
Uddannelse
Lokalt miljø
Netværk
Gener
Venner
Samfund
Familie


Randomisering
A
B
Indsats
Måle outcome

ANTAL DELTAGERE

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 at kunne vise en forskel ⇒ jo færre personer behøver vi for
- Jo mindre effektstørrelse ⇒ jo flere personer skal vi bruge

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)

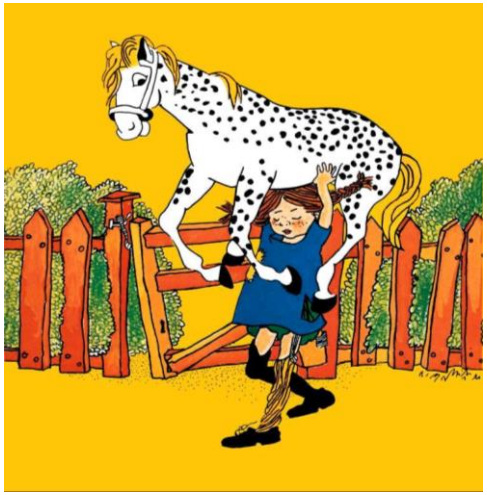
Generelt siger vi gerne 100-150 i alt

STYRKEBEREGNINGER

Der findes forskellige værktøjer til styrkeberegninger.

Prøv eventuelt:

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



EKSEMPEL: KÆRLIGHED I KAOS (KIK)

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

EKSEMPEL: STØTTE TIL UDSATTE BØRNEFAMILIER

'Praktisk Pædagogisk Støtte' og 'Familiebehandling' er de to familiebevarende foranstaltninger i Serviceloven, som familier oftest visiteres til i Danmark.

Designet var simpel randomisering (1:1).

Vi havde 8 kommuner med, men kun 43 familier!

Konklusionen blev: 'Effektmålingen viser ingen signifikante forskelle på effekten af 'Praktisk Pædagogisk Støtte' og 'Familiebehandling'. Dette kan skyldes, at der ikke er familier nok i studiet til at måle en signifikant forskel'.

HVORFOR GIK DET GALT?

- I nogle kommuner var indsatser mere eller mindre de samme, derfor kunne vi ikke finde nogen forskelle.
- I nogle kommuner var målgruppen til de to indsatser vidt forskellige og derfor kunne vi ikke rekruttere.
- Sagsbehandlerne var imod randomiseringen.

ETISKE PROBLEMSTILLINGER

Er det okay at trække lod om indsatser?

Er det muligt at lave randomiserede forsøg der, hvor du arbejder?

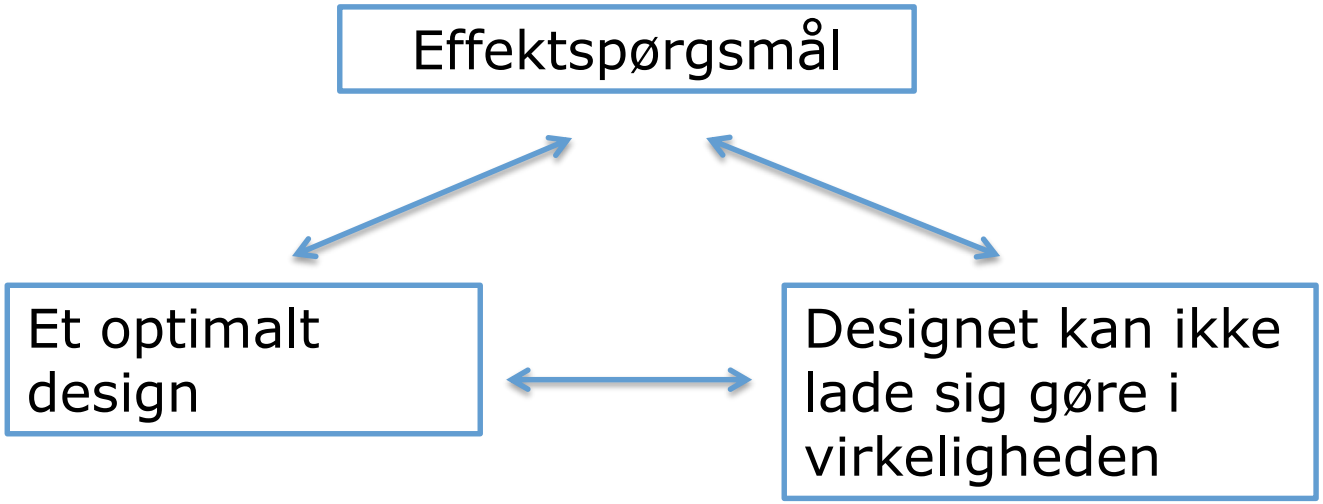
Hvilke typer af begrænsninger vil der være for at anvende eksperimenter inden for dit område?

ETISKE PROBLEMSTILLINGER

Argumenter FOR randomisering på det sociale område:

- Store regionale forskelle i behandling/tilbud
- Ikke enighed om, hvad der bør tilbydes
- Meget sjældent at kontrolgruppen ikke bliver tilbudt nogen behandling/tilbud overhovedet (oftest standard behandling)
- Ofte kan det, man tilbyder indsatsgruppen, betragtes som noget ekstra oven i standardbehandlingen.

EVALUERINGENS BERMUDATREKANT



Et risikabelt sted at være, hvis man er en evaluering

Peter Dahler-Larsen (2013)

UD AF BERMUDATREKANTEN!

Når vi ikke kan lave ”rigtige” eksperimenter, må vi gøre noget andet:

- Undvære kausalitet og nøjes med korrelation
  - Før/efter måling uden kontrolgrupper
- Bruge et naturligt eksperiment – hvis vi kan finde et
  - Regression discontinuity design m.fl.
- Bruge observationelle data og avanceret statistik (reparere den omstændighed at vi ikke har et eksperiment – eller et elendigt eksperiment)
  - Matching

EKSEMPEL:
STØTTE TIL UDSATTE BØRNEFAMILIER

Når der ikke er nok deltagere eller randomiseringen ikke virker:

- Så har vi en før- og eftermåling af to indsatser.

Derfor lød resten af konklusionen:

'Før- og eftermålinger viser, at alle familier gennemgik en positiv udvikling, hvor mødrene fx fik færre depressionssymptomer, og børnene udviste mindre problemadfærd. Disse ændringer kan dog ikke med sikkerhed tilskrives foranstaltningerne'.

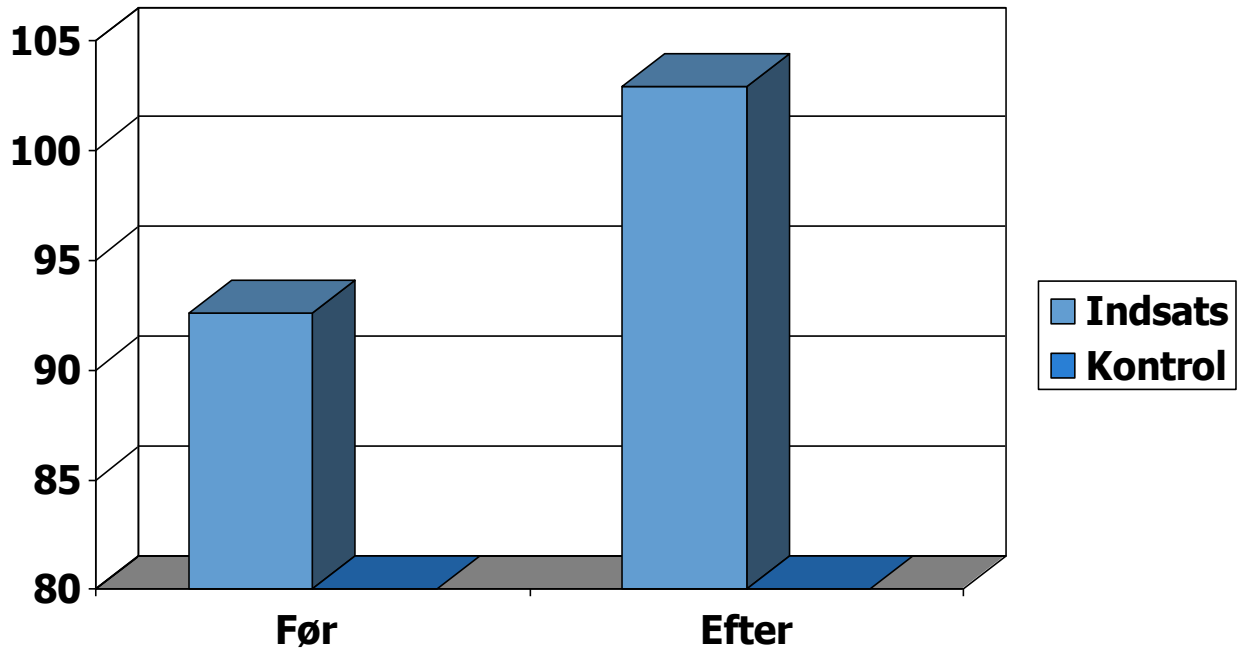
FØR- OG EFTERMÅLINGER

Viser hvordan borgeren UDVIKLER/ÆNDRER sig over tid.

Kan bruges når der ikke kan laves en effektmåling
- ikke muligt pga. fx økonomi,
- for lille målgruppe
- ingen mulighed for kontrolgruppe

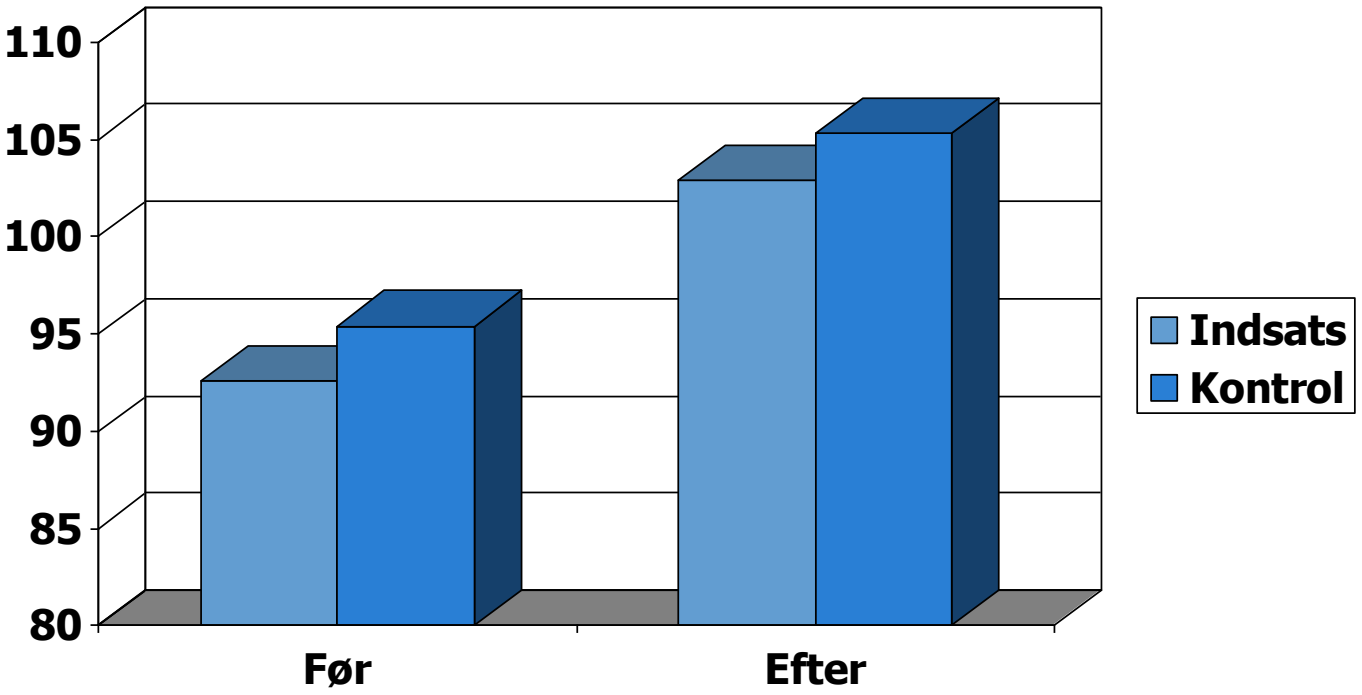
Men resultaterne kan ikke med sikkerhed tilskrives indsatsen – man ved ikke noget om den kausale sammenhæng.

FØR-EFTERMÅLING AF LÆSE-INTERVENTION



Statistisk signifikant forskel $p < 0.001$

FØR-EFTERMÅLING AF LÆSE-INTERVENTION



Forskel mellem grupper er IKKE signifikant

FØR- OG EFTERMÅLINGER

Et godt sted at starte!

Mange evalueringer bygger på dette design.

Særligt når der ikke er en naturlig målgruppe eller hvis man starter en evalueringskultur op i en organisation.

PAUSE!

NATURLIGE EKSPERIMENTER

Logikken er den samme som i det menneskeskabte eksperiment.

- Vi vil have noget ”naturlig” eller kvasi-eksperimentel variation i den indsats, som vi er interesseret i.
- Vi lader naturen (eller tilfældighed) kaste terningerne og observerer resultatet …

Det unaturlige eksperiment:

1. intervention foretaget af undersøgeren
2. forskerrolle: designe interventioner
3. klare kontrol- og forsøgsgrupper
4. random assignment
5. Mange forskellige designs: Labeksperimenter, surveyeksperimenter, felteksperimenter.

Det naturlige eksperiment:

1. intervention fra “naturens” hånd
2. forskerrolle: afsøge interventioner
3. ofte uklare kontrol- og forsøgsgrupper
4. “as if” random assignment
5. Mange forskellige designs: RDD, IV, matching, Diff-in-diff, Interupted time series mv.

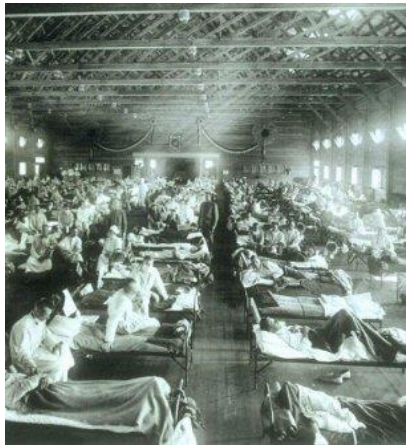
Asmus Leth-Olsen (2013)

EKSEMPEL

- Har sundhed som barn effekt på, hvordan man klarer sig som voksen?

- Problem: Sunde mødre får sunde børn. Men sunde mødre har også andre ressourcer, der påvirker deres børns outcomes.

- Hvordan kan vi isolere effekten af sundhed?
- Vi skal bruge et skud tilfældig tildeling af sundhed!



- Eksempel: Den spanske syge i 1918
  - Ramte tilfældigt gravide kvinder
  - Børn hvis mødre var smittet under graviditet tjente 5-9% mindre gennem livet end børn hvis mødre ikke var smittet

ANDRE EKSEMPLER

Kan I komme på eksempler på evalueringsspørgsmål, som kunne besvares med et naturligt eksperiment?

(man må gerne være kreativ☺)

NÅR PLAN B ER BEDRE END PLAN A - ET EKSEMPEL

Evalueringsspørgsmål: Klarer elever i store klasser sig dårligere end elever i små klasser?

Kan vi bare sammenligne små og store klasser?

Evalueringsproblemet: Klassestørrelse samvarierer med uobserverbare karakteristika ved forældre/børn, der påvirker børns læring → vi måler ikke kausal effekt af klassestørrelse!

KAN VI LAVE RCT?

JA! Og det har man gjort i Project STAR

- Involverede 11.600 elever, 1.300 lærere og 76 skoler i Tennessee, USA
- Elever fordelt ved lodtrækning i enten (1) små (13-16 elever), (2) almindelige (22-26 elever) og (3) almindelige klasser med en ekstra lærer
- Lærere også fordelt til de tre klassetyper ved lodtrækning
- Resultat: Elever i små klasser klarede sig bedre mht. karakterer og ssh. for videregående uddannelse.

MEN

- Elever, lærere og børn vidste, at de var med i et eksperiment (Rosenthaleffekt). De vidste også, om de havde været så heldige at komme i små klasser
- Succeskriteriet var kendt for alle. Hvad med dem, der blev sure over, at de havnede i en stor klasse?
- Ekstern validitet: Ville eksperimentet give samme resultat, hvis det blev lavet et andet sted?
- Der var en kausal effekt, men eksperimentet kostede 70 mio. kr. at udføre og den "kvantitative" effekt var ikke særlig stor. Kan det overhovedet betale sig at reducere klassestørrelsen relativt til andre tiltag, der også forbedrer elevers læring?

Kan vi finde et naturligt eksperiment, der påvirker klassestørrelsen men ikke har noget med de individuelle elever at gøre?

ET BERØMT EKSEMPEL: MAIMONIDE'S RULE

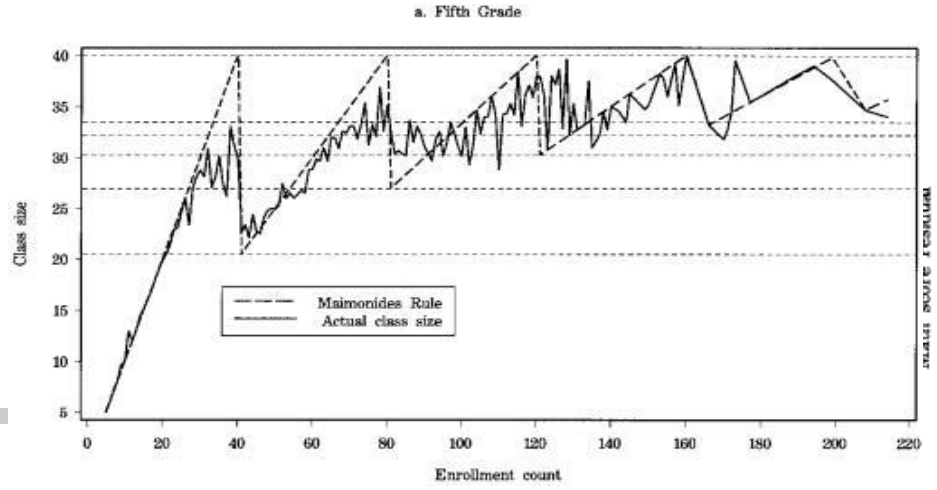
(Angrist & Lavy, 1999)

Maimonide var rabbiner i Israel i det 12. århundrede. Han tolkede den jødiske bibels regler for klassestørrelse således:

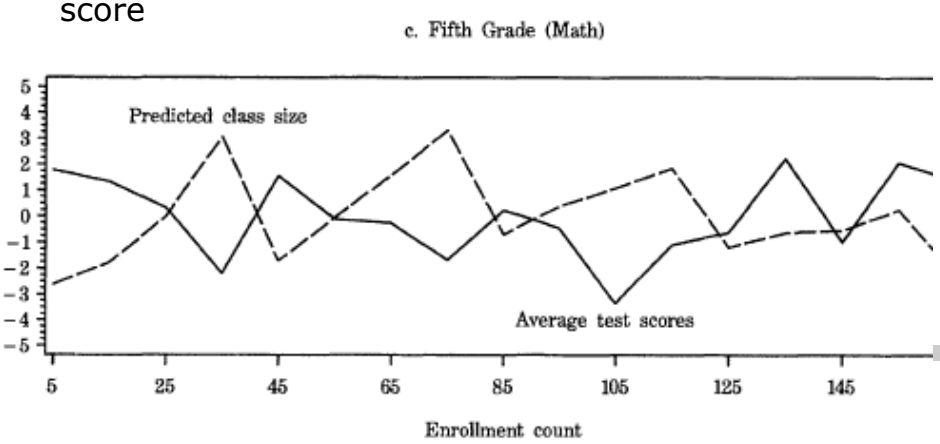
- En lærer må undervise 25 elever. Hvis der er mere end 25 men mindre end 40 elever skal han have en hjælpelærer. Hvis der er flere end 40 elever, skal klassen deles.

Reglen har været i brug i det israelske skolesystem siden 1969.

➢ Eksogen variation i klassestørrelse!



Negativ sammenhæng mellem klassestørrelse og test score



Ideen er brugt i flere andre studier – også i Danmark.

- De fleste finder negativ effekt af klassestørrelse på elev-outcomes. Men ikke alle!

Det er ok! "Situationsbundet" fortolkning af den kausale effekt, som eksperimentet identificerer viser, at sandheden ikke nødvendigvis er endegyldigt fundet med ét studie (LATE).

"Hvem virker eksperimentet på? Af dem, der får en pille, er der nogen, der (1) altid spiser pillen, (2) kun spiser pillen hvis den smager af lakrids, (3) altid - men grinende - skyller pillen ud i toilettet og (4) er sure over, at de ikke fik pillen og køber en der minder om den nede på den lokale bodega".

(Mads Jæger, SFI)

FORDELE VED DET NATURLIGE EKSPERIMENT

- Ingen etiske problemer – ”naturen”/tilfældigheder bestemmer, hvem der kommer i indsats og kontrol
- Ingen politiske problemer – ingen kan forhindre dig i at undersøge dem
- Færre logistiske udfordringer
- Billigere
- Historiske data kan bruges – ikke kun fremtidige

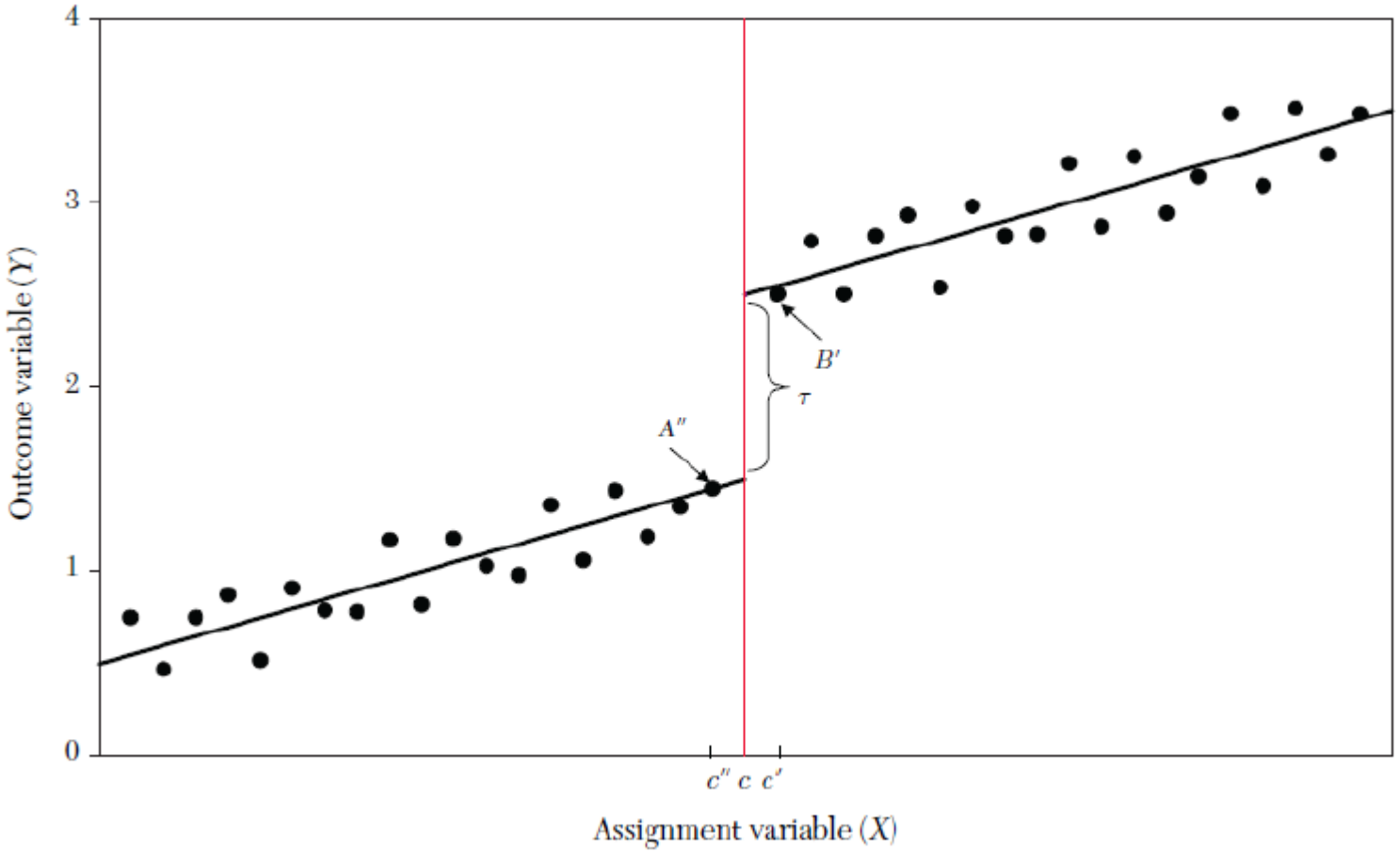
Regression discontinuity design (RD)

En måde at måle effekt ved naturlige eksperimenter

Eksempel: Et amerikansk program, der giver økonomisk støtte til skoler for at løfte uddannelsesniveauet i udsatte områder.

Hvordan kan man måle effekten?

- Skoler i hvert skoledistrikt kan få økonomisk støtte, hvis antallet af elever fra fattige familier er over distriktsgennemsnittet.
  - Vi skal bruge denne selektionsmekanisme!
  - Er ”as if random” – hverken skoler eller elever har direkte kontrol over, om de kommer i indsats- eller kontrolgruppen
- Vi kan sammenligne skoler lige over og under distriktsgennemsnittet


4
3
2
1
0
Outcome variable (Y)
A″
B′
τ
c″ c c′
Assignment variable (X)

Begreber

En *observerbar* tærskelvariabel opdeler *deterministisk* observationer i indsats- og kontrolgrupper

- Fx fattigdomsraten

Tærskelværdien (cut-off) er den værdi af tærskelvariablen, der bestemmer hvornår treatment "slås til"

- Fx den gennemsnitlige fattigdomsrate i skoledistriktet
  - Indsats: Skoler med over 25% fattige
  - Kontrol: Skoler med under 25% fattige

Antagelser

Omkring tærskelværdien er det "as if random" om observationer er havnet i indsats eller kontrol

Dvs. der må ikke være mulighed for selvselektion eller aktiv selektion fra andre

Hvis det holder, kan vi sammenligne observationer lige over og under tærskelværdien, da de må være ens på alle punkter, undtagen om de er i indsats eller kontrol.

Et typisk problem i RD: Tærskelværdien (assignment rule) overholdes ikke strengt – og så bliver det svært at sammenligne.

- "Fuzzy RD"

Eksempler på tærskelvariable og tærskelværdier

1. Alle med et snit på 11,2 kunne læse medicin på KU i 2016 – alle andre kan ikke
2. Alle over 18 år kan købe alkohol – alle andre kan ikke
3. Folk født før 1. juli 1954 pensioneres et halvt år før alle andre
4. Personer over den kriminelle lavalder kan straffes

Har I eksempler på *skarpe* og *målbare* regler eller kriterier indenfor jeres område?

Hvilke evalueringsspørgsmål kan de ovennævnte tærskelvariable fx bruges til at svare på?

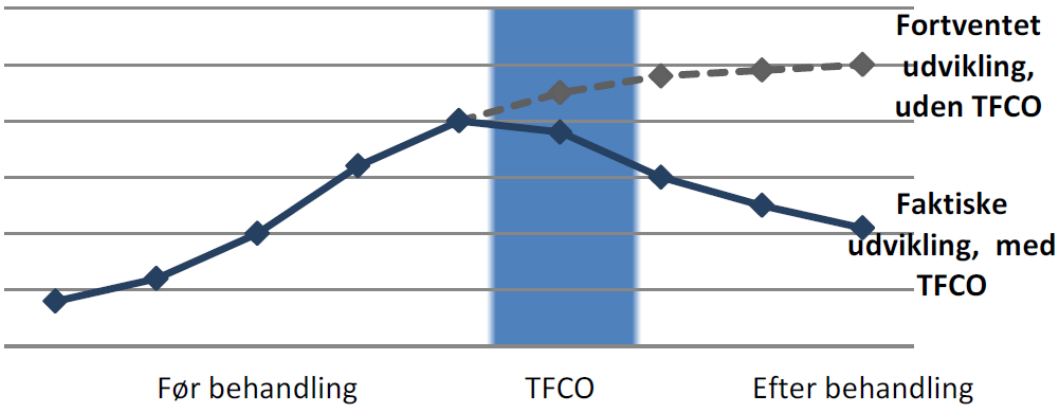
Kinked regression discontinuity design

En lidt anderledes variant:

Ingen ”treatment” og ”control” – hvert individ er sin egen kontrol.

Princip: Man estimerer et forventet udviklingsforløb for hvert individ baseret på pre-treatment variable. Det forventede udviklingsforløb sammenlignes med det faktiske for at se, om indsatsen ”knækker kurven”

Figur 1: Eksempel på et udviklingsforløb (for eksempel kriminel løbebane, med og uden TFCO)



SFI og Socialstyrelsen, 2016

PAUSE!

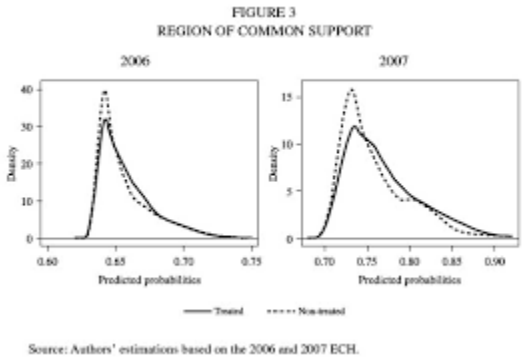
MATCHING

Når vi ikke kan lave et eksperiment af hverken den ene eller anden slags!

…må vi bruge ikke-eksperimentielle data

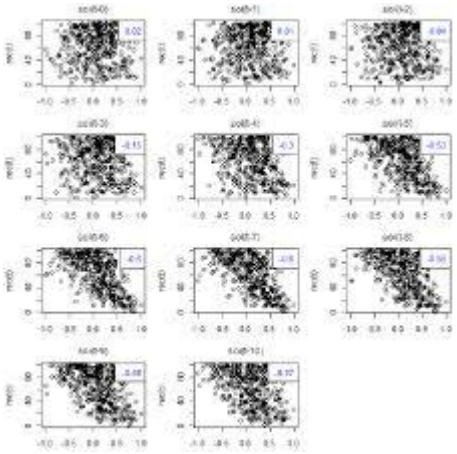
…til at finde en måde at konstruere en kontrolgruppe

Med risiko for at det hele kommer til at se ud som noget à la dette:



Source: Authors' estimations based on the 2006 and 2007 ECH.

$$
\begin{aligned}
L_N(\theta) &= \frac{1}{N}\sum_{i=1}^{N}\ln f\left(y_i, x_i;\ (\beta,\ \gamma)\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\ln g\left(y|x;\ (\beta,\ \gamma)\right) + \frac{1}{N}\sum_{i=1}^{N}\ln h\left(x|(\mu,\ \delta^2)\right) \\
&= \left\{\frac{-1}{2N}\sum_{i=1}^{N} X_i\gamma - \frac{1}{2N}\sum_{i=1}^{N}\frac{\left[y_i - \exp\left(X_i\beta\right)\right]^2}{\exp\left(X_i\gamma\right)}\right\} \\
&\quad + \left\{-\frac{N}{2}\ln\delta^2 - \frac{1}{2\delta^2}\sum_{i=1}^{N}\left[x_i - \mu\right]^2\right\} - \ln(2\pi) \\
&\equiv L_N^c(\beta,\gamma) + L_N^m(\mu,\delta^2).
\end{aligned}
$$



MATCHING - PRINCIPPET

Den bærende idé i matching er – med nogle statistiske kneb – at skabe en kontrolgruppe, som ligner en bestemt indsatsgruppe så meget, at det er muligt at sammenligne outcomes for de to.

Matching er ofte velegnet ved små målgrupper.

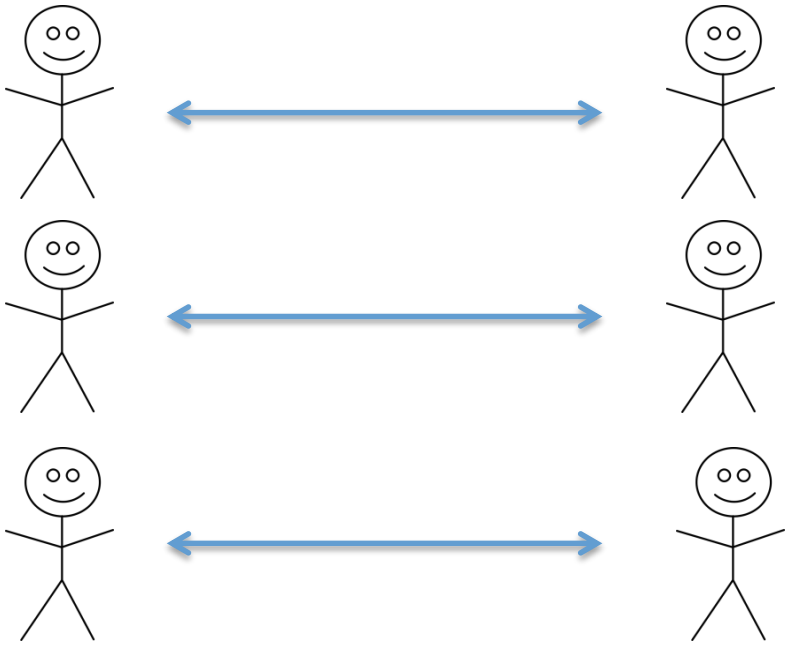
To varianter

1. Eksakt matching
2. Propensity score matching

EKSAKT MATCHING – STATISTISK TVILLING

Indsats

Kontrol



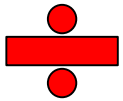
Eksempler på matching-variable:

- Køn
- Alder
- Etnicitet
- Bopæl
- Uddannelse
- Sundhed

Formål:
At lave en statistisk tvilling, hvor den eneste forskel er, at en har fået indsatsen og den anden ikke har.

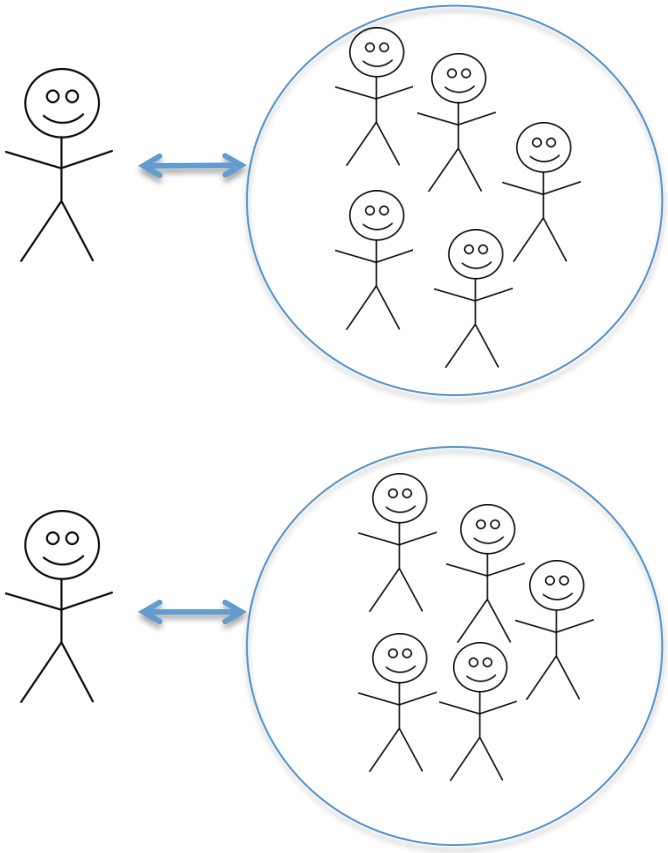
FORDELE OG ULEMPER VED EKSAKT MATCHING

- Metoden er gennemskuelig
- God ved store casegrupper



- 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

PROPENSITY SCORE MATCHING



Hvad er en propensity score?

En sammensat variabel (bestående af flere underliggende variable), der angiver *sandsynligheden* for deltagelse i indsatsen

De underliggende variable:
- vigtige for selektion
- Fx alder, køn, bopæl, etnicitet etc.

Husk: Man kan ikke matche på sit outcome-mål (KRAP).

PROPENSITY SCORE

Udregnes for alle individer i populationen – både indsatsgruppen og den population, som kontrolgruppen skal findes i (med en probit/logit regression)

Sandsynligheden for at tilhøre indsatsgruppen

De fleste individer i indsatsgruppen vil have en høj propensity score

For restpopulationen er det typisk omvendt

Kontrollerne udvælges, så deres PS er så tæt på indsatspersonerne som muligt.

PREBENS PROPENSITY SCORE

Her er Preben 

Preben tilhører en udsat familie i Assens kommune. Assens vil gerne hjælpe Preben og giver ham en smart amerikansk indsats på 3 bogstaver. Assens kommune vil gerne vide, om indsatsen virker.

Kan vi finde den kontrafaktiske Preben?

Ved en regressionsanalyse, hvor alle observerbare karakteristika tages i betragtning findes Prebens propensity score til at være PS=0,9615.

Hvis Assens var en stor kommune fandtes måske en anden mand, Mads, med nøjagtig samme PS, men som ikke fik indsatsen (et eksakt match). Men Assens er en lille kommune med få udsatte.

PREBENS PROPENSITY SCORE II

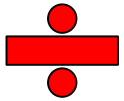
- Hvis man bruger ”nearest neighbour” som match, vil Preben blive matchet med Søren, som har PS=0,9601
- Eller i en 1-2 (1 case, 2 kontroller) matching med Søren og Asger (PS=0,9636)

- Hvordan har Søren og Asger fået deres propensity score?
  - De har måske samme arbejdsmarkedstilknytning, civilstand og uddannelse som Preben, men er nogle år ældre eller yngre
  - Alle risikofaktorer vægtes i propensity scoren, derfor er Søren og Asger ens med Preben på de parametre, der virkelig betyder noget (fx arbejdsmarkedstilknytning) og derfor vægtes højere.

FORDELE OG ULEMPER VED PS MATCHING

- Man kan finde en kontrolgruppe selvom der ikke er mange der ligner på individuelle faktorer

- Variable vægtes efter betydning

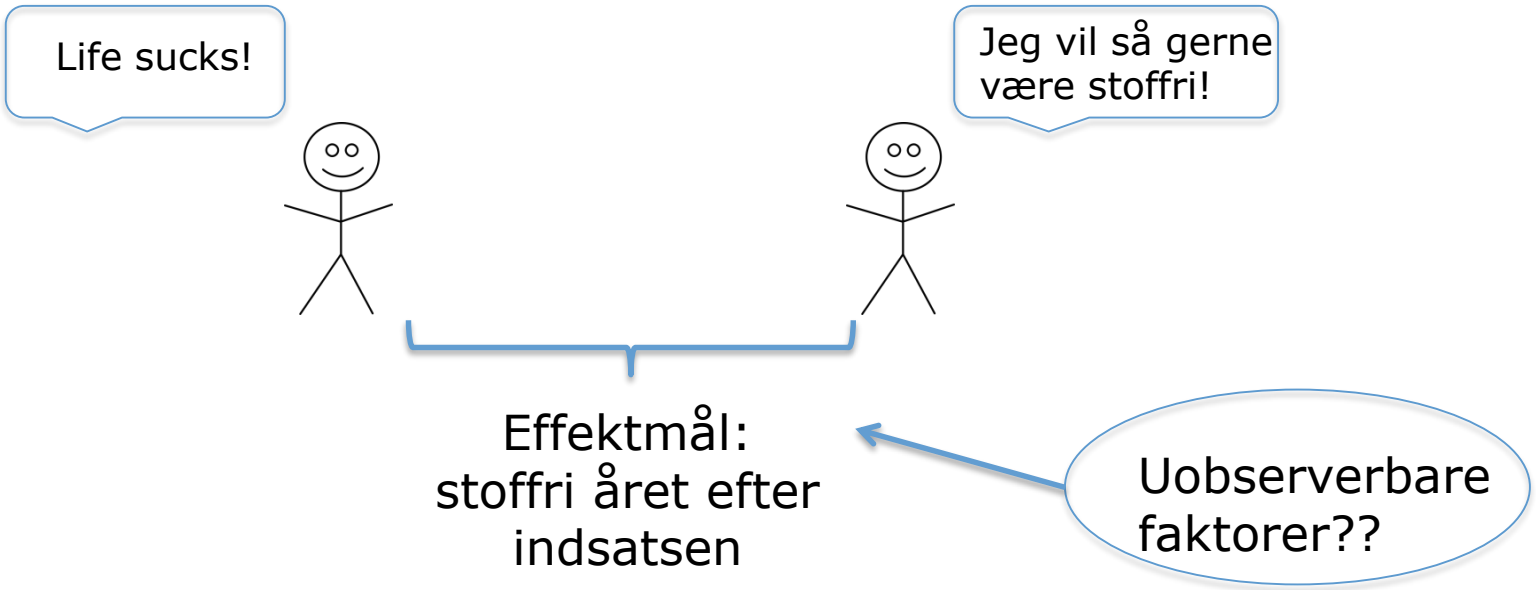


- Propensity scoren i sig selv er uigennemskuelig og svær at forklare intuitivt

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

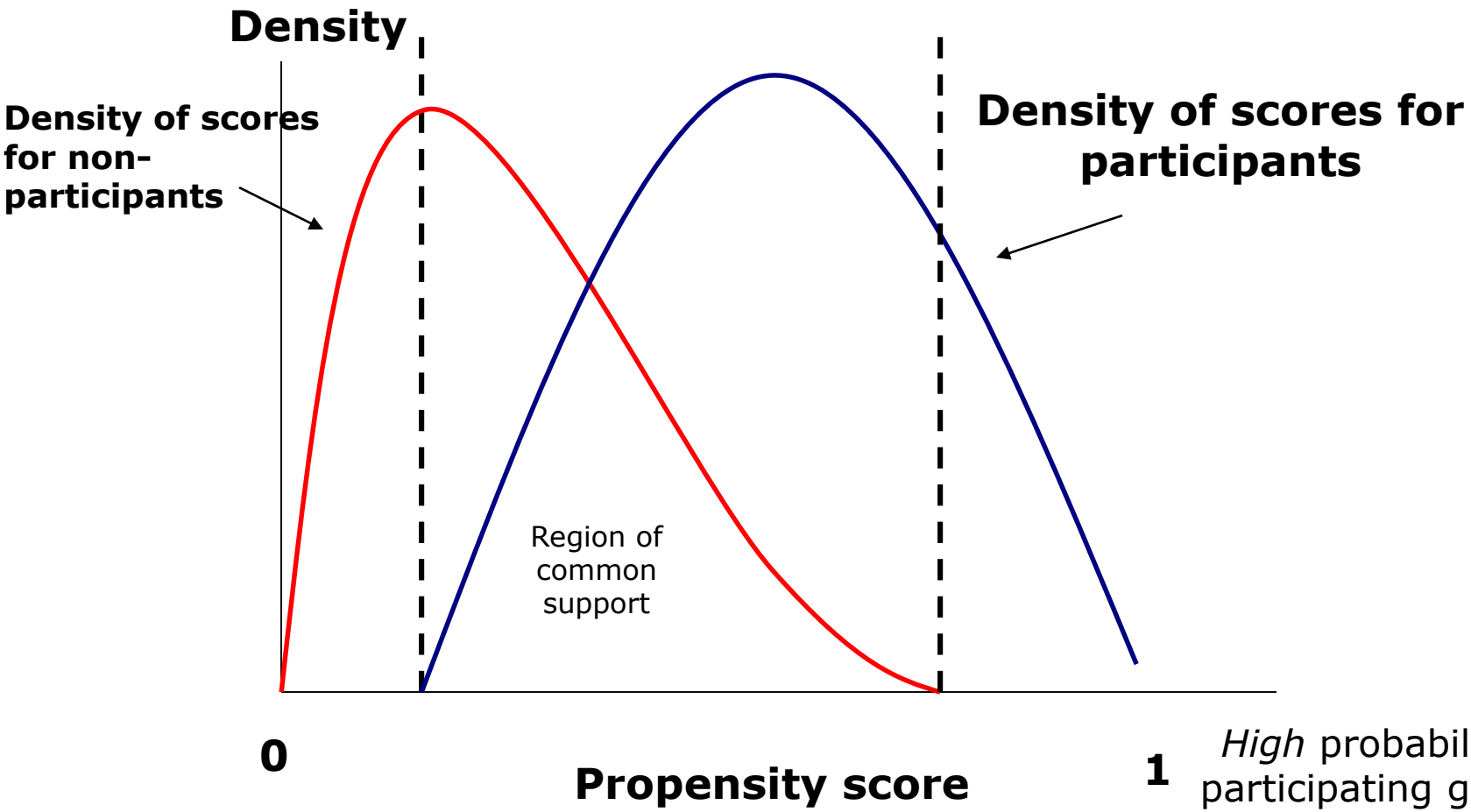
VIGTIGE ANTAGELSER

Conditional Independence Assumption (CIA)



Hvis uobserverbare faktorer har betydning for outcome og er forskellige for indsats og kontrol, holder CIA ikke => bias.
Det er vanskeligt at teste - og gøre noget ved!
Typiske ”unobservables”: evne, motivation, drive…

Common Support



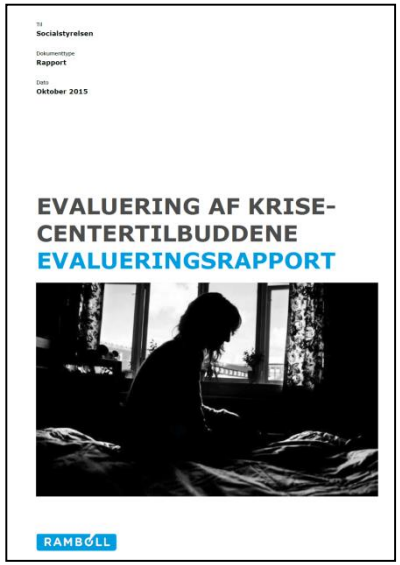
HVORNÅR KAN MAN ELLERS KOMME I PROBLEMER?

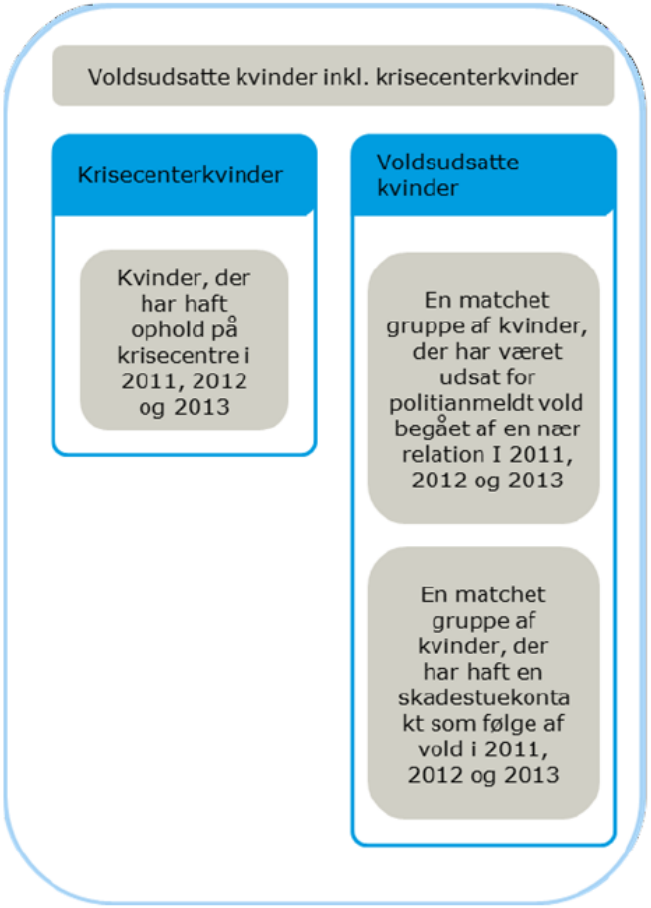
Altid! Men det hjælper hvis:

- 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 kriscenter 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

Hvordan ser resultaterne af et matching studie ud?

Hvordan fortolker jeg resultaterne?

Hvad skal man være særligt opmærksom på?

Andre metodiske kneb

Bruge avancerede statistiske metoder, når designet er dårligt☺

Bedre tjent med at løse problemer i designfasen (selv om det kan være svært)

Se uddelt ark for oversigt.

SPØRGSMÅL OG KOMMENTARER?!