Mudelid põhjuslikele mõjudele

Mis on põhjuslik seos (mõju)?

Kontrafaktid (*Counterfactuals*)

- Jaan suitsetas ja suri noorena.
- Kui Jaan poleks suitsetanud, poleks ta noorena surnud.

Järelikult eksisteerib suitsetamise põhjuslik mõju Jaanile (suitsetamine on Jaani varase surma põhjuseks).

Jaani varase surma põhjusteks on ka mittegeeniusest perearst (sest geeniusest perearst oleks suutnud kopsuvähi varases staadiumis diagnoosida); halvad geenid (sest heade geenide korral poleks suitsetamise tagajärjel temal kopsuvähki tekkinud); …

Keskmine põhjuslik mõju

Neyman–Rubin causal model

Nimi	$Y^{suits=0}$ (ei suitseta)	$Y^{suits=1}$ (suitsetab)
Jaan	0	1
Kalle	1	1
Malle	0	0
Ants	1	0
AnnaBellaTrullala	0	1
Benita	1	1
Giuseppe	0	1
Taavi	0	1
Kristel	0	0
Sven	0	1

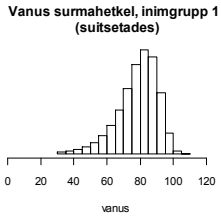
$$E(Y^{suits=0}) = 0,3$$

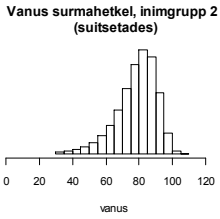
$$E(Y^{suits=1}) = 0,7$$

Keskmine põhjuslik mõju puuduks, kui $E(Y^{suits=1} - Y^{suits=0}) = 0$

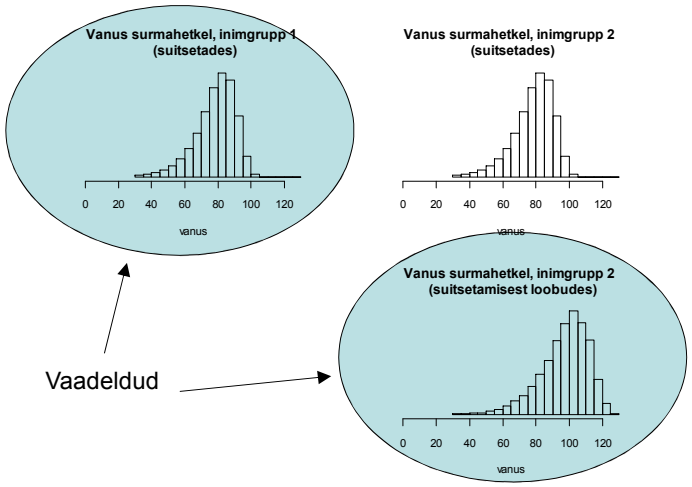
Keskmine põhjuslik mõju: $E(Y^{suits=1} - Y^{suits=0}) = 0,4$

Üksikisiku puhul saavutamatu, inimeste grupi puhul saavutatav





Üksikisiku puhul saavutamatu, inimeste grupi puhul saavutatav

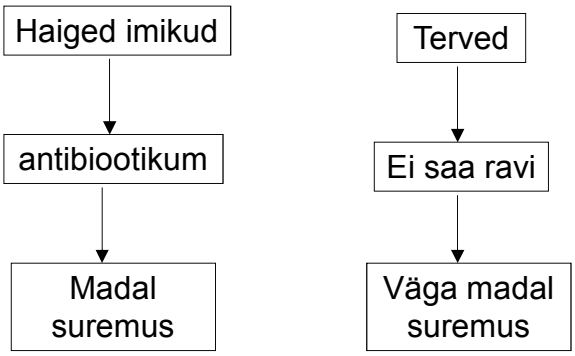


Näide – randomiseerimine muudab grupid võrreldavaks

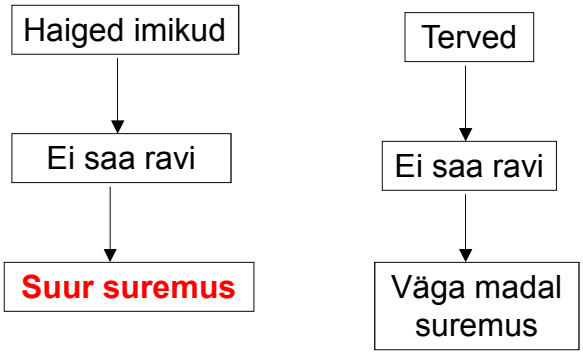
Grupi suurus (*n*)	naiste %		keskmine vanus	
	grupp A	grupp B	grupp A	grupp B
10	20%	40%	66,2	64,7
50	46%	34%	64,3	64,5
100	42%	47%	65,4	65,5
500	40,2%	40,8%	64,9	65,1
3000	40,2%	39,6%	65	64,9

40

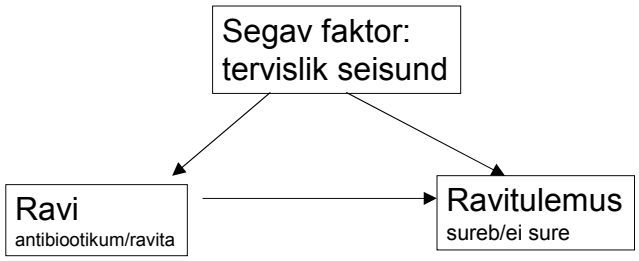
Mitterandomiseeritud uuring
(Vaatlusandmete pealt tehtud järeldused…)



Mitterandomiseeritud uuring
(Vaatlusandmete pealt tehtud järeldused…)

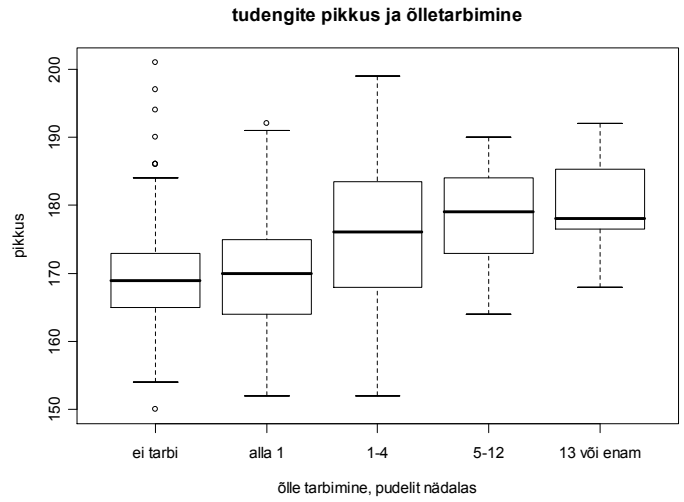


Segav faktor



Näiteid sagedastest segavatest faktoritest:
sugu; rikkus; vanus; rahvus (populatsioon); …

Mis on segavaks faktoriks siin?



44

Kui eesmärgiks on kirjeldada põhjuslikke mõjusid…

Näide genereeritud andmetega (tõde on teada…)

Andmeid genereerinud mudel:

$$y = 12 - 1 \cdot x + s + \varepsilon$$

Soovime teada, kuidas tunnus *x* mõjutab tunnuse *y* väärtust…

Kui eesmärgiks on kirjeldada põhjuslikke mõjusid…

$$y = 12 - 1 \cdot x + s + \varepsilon$$

```
> summary(lm(y~x))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.00465    0.02056  583.90   <2e-16 ***
x            0.34213    0.03413   10.02   <2e-16 ***

Residual standard error: 1.126 on 2998 degrees of freedom
Multiple R-squared:  0.03243,   Adjusted R-squared:  0.0321
F-statistic: 100.5 on 1 and 2998 DF,  p-value: < 2.2e-16
```

Kui eesmärgiks on kirjeldada põhjuslikke mõjusid…

$$y = 12 - 1 \cdot x + s + \varepsilon$$

```
> summary(lm(y~x+s))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.01175    0.01803  666.22   <2e-16 ***
x           -1.02831    0.05457  -18.84   <2e-16 ***
s            0.99739    0.03321   30.03   <2e-16 ***

Residual standard error: 0.9874 on 2997 degrees of freedom
Multiple R-squared:  0.2563,    Adjusted R-squared:  0.2558
F-statistic: 516.3 on 2 and 2997 DF,  p-value: < 2.2e-16
```

Kui eesmärgiks on kirjeldada põhjuslikke mõjusid…

$$y = 12 - 1 \cdot x + s + \varepsilon$$

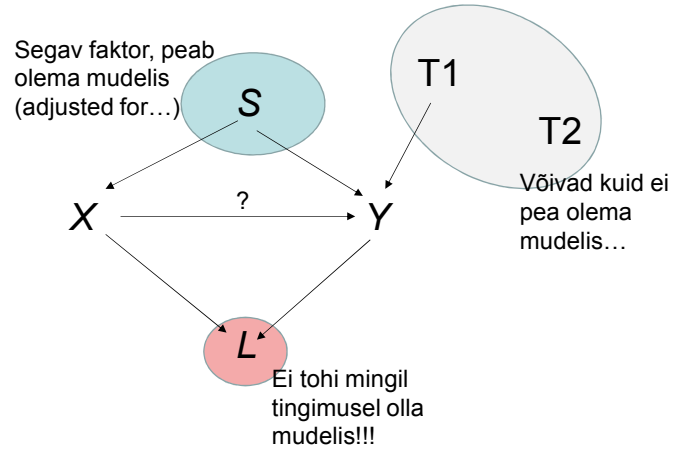
```
> summary(lm(y~x+s+ muu1))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.01176    0.01803 666.158   <2e-16 ***
x           -1.02677    0.05463 -18.796   <2e-16 ***
s            0.99643    0.03325  29.971   <2e-16 ***
muu1        -0.01143    0.01754  -0.652    0.515

Residual standard error: 0.9875 on 2996 degrees of freedom
Multiple R-squared:  0.2564,    Adjusted R-squared:  0.2556
F-statistic: 344.3 on 3 and 2996 DF,  p-value: < 2.2e-16
```

Kui eesmärgiks on kirjeldada põhjuslikke mõjusid…

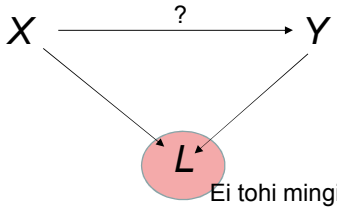
$$y = 12 - 1 \cdot x + s + \varepsilon$$

```
> summary(lm(y~x+s+ muu2))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.424641   0.087885   27.59   <2e-16 ***
x           -4.193544   0.037815 -110.90   <2e-16 ***
s            0.197090   0.016546   11.91   <2e-16 ***
muu2         0.399408   0.003646  109.55   <2e-16 ***

Residual standard error: 0.4414 on 2996 degrees of freedom
Multiple R-squared:  0.8514,    Adjusted R-squared:  0.8513
F-statistic:  5723 on 3 and 2996 DF,  p-value: < 2.2e-16
```

Kui eesmärgiks on kirjeldada põhjuslikke mõjusid…



Kui eesmärgiks on kirjeldada põhjuslikke mõjusid…



Ei tohi mingil tingimusel olla mudelis!!!

X – naise nahavärv (must või valge)
Y – mehe nahavärv (must või valge)

L – lapse nahavärv (valge, mulatt, must)

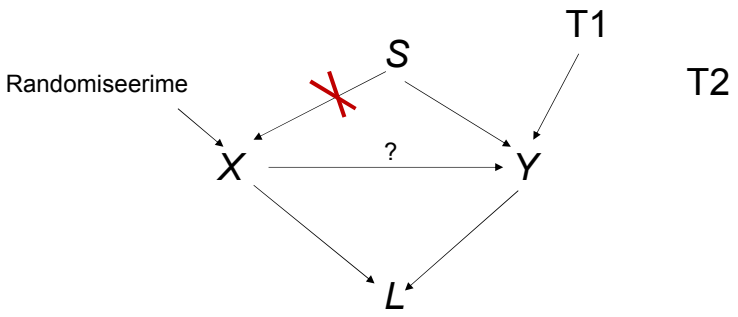
„Poliitkorrektne“ maailm – abikaasa valikut nahavärv ei mõjuta…

Kui teame lapse nahavärvi (mulatt), kas siis ema nahavärvi lisaks teadmine aitab prognoosida isa nahaväri?

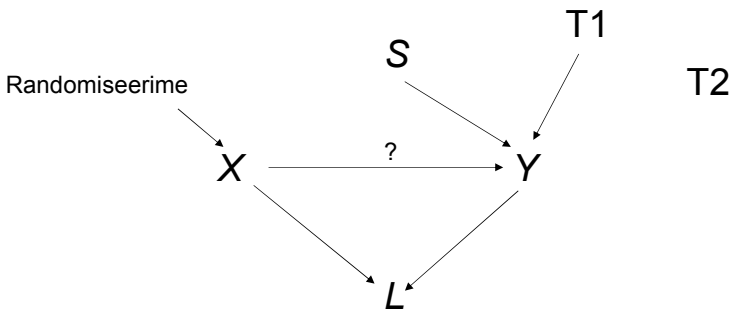
Kui eesmärgiks on kirjeldada põhjuslikke mõjusid…

Kui kasutatakse randomiseeritud katse abil kogutud andmeid…

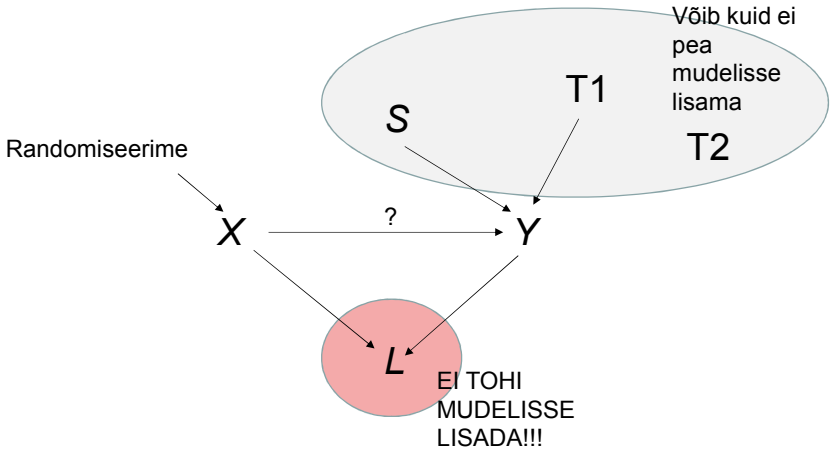
Kui eesmärgiks on kirjeldada põhjuslikke mõjusid…



Kui eesmärgiks on kirjeldada põhjuslikke mõjusid…



Kui eesmärgiks on kirjeldada põhjuslikke mõjusid…



Segavate faktorite/tunnuste näiteid

Inimestel:

Sugu ja geenid (X ja Y tunnus vaevalt suudavad mõjutada inimese sugu või tema geene, küll aga võivad tema sugu ja geenid mõjutada erinevate tunnuste väärtuseid…)

Võimalikud: Optimism. Optimist: majanduslik seis on mul hea (kuigi töötu ja kohustused kaelas); tervis samuti hea (kuigi põed vähki, aga näed, täna on kuidagi hea päev, ei valutagi kuskilt eriti…); pessimist: majanduslik seis halb (minu omanduses olevate firmade väärtus arvatavasti langeb mõne peatselt saabuva kriisi tõttu); tervis halb (lõin hommikul varba ära – väga valus oli…)

….

```
> mudel1=lm(pikkus~factor(olu))
> drop1(mudel1, test="F")
Model:
pikkus ~ factor(olu)
            Df Sum of Sq   RSS    AIC F value    Pr(>F)
<none>                   41098 2736.8
factor(olu)  4    5413.1 46512 2810.4  21.568 < 2.2e-16 ***
```

```
> mudel2=lm(pikkus~factor(olu)+factor(sugu))
> drop1(mudel2, test="F")
Model:
pikkus ~ factor(olu) + factor(sugu)
             Df Sum of Sq   RSS    AIC  F value Pr(>F)
<none>                    23836 2379.2
factor(olu)   4     129.1 23965 2374.8   0.8855 0.4722
factor(sugu)  1   17262.9 41098 2736.8 473.6606 <2e-16 ***
```