Tõenäosusteooria ja statistika II

Loeng 15

Statistiline seos ja põhjuslik mõju

Märt Möls

Mis on statistiline seos?

Matemaatiline definitsioon (pideva tunnuse jaoks):

Tunnused Y ja X on sõltumatud, kui

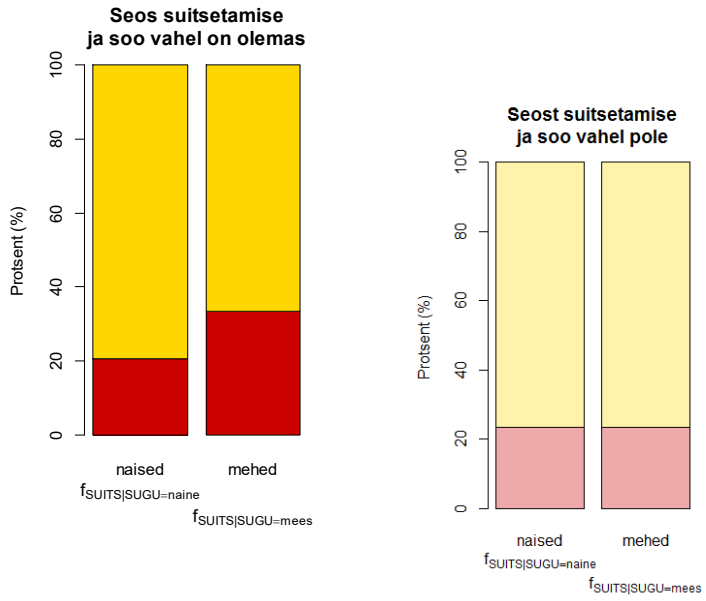
$$f_{Y|X=c} = f_Y \quad \forall c$$ (mida Y võib omandada)

tunnuse Y tinglik jaotus ei sõltu valitud X-tunnuse väärtusest

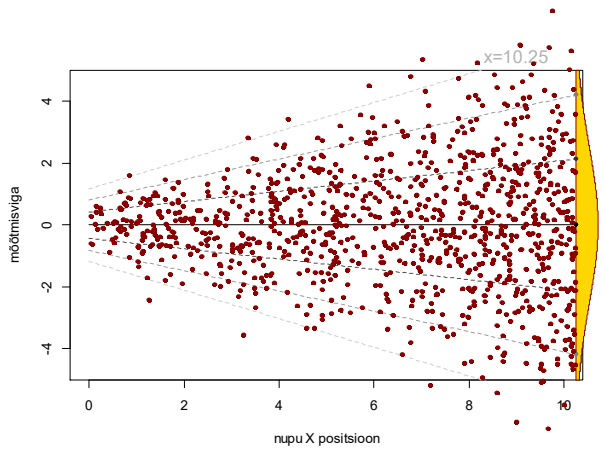
Tunnused Y ja X on sõltuvad, kui

$$\exists c, \quad f_{Y|X=c} \neq f_Y$$

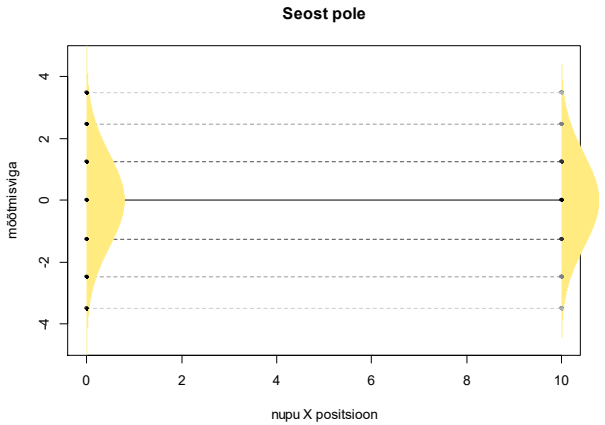
tunnuse Y tinglik jaotus sõltub valitud X-tunnuse väärtusest



Näide stat. seosest kahe pideva tunnuse vahel



Statistilist seost pole



Statistiline seos kahe mittearvulise tunnuse vahel. Hii-ruut test.

Näide

kas esineb seos tudengi tervisehinnangu ja tema soo vahel?

Tabel (arstiteaduskonna 2. kursus):

	hinnang tervisele			
sugu	v.hea	hea	keskmine/halb	kokku
naine	83 (13%)	404 (62%)	161(25%)	648 (100%)
mees	35 (18%)	105 (55%)	50(26%)	190 (100%)

Mida tähendab seose olemasolu kahe tunnuse vahel? Siin: seos on olemas, kui erinevast soost inimeste tervisehinnangute jaotus on erinev.

Küsime: milline oleks oodatud tervisehinnangute jaotus, kui hinnang tervisele ei sõltuks soost? (Nullhüpoteesiks on siin, et tervisehinnangu jaotus tabeli igas veerus on sama.)

Vaatame, milline on tervisehinnangute jaotus valimis kokku:

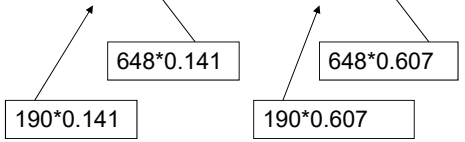
tervis (%)		
v.hea	hea	keskmine/halb
118(14,1%)	509 (60,7%)	211 (25,2%)

Nullhüpoteesi täidetuse korral peaks see jaotus olema sama nii meestel kui naistel. Seega 14% naistest ja sama suur osa, ehk siis samuti 14% meestest, peaks arvama, et nende tervis on väga hea, 61% nii meestest kui naistest, et nende tervis on hea, jne.

Leiame, kui palju see teeks arvuliselt.

Vaadeldud ja eeldatav (sulgudes) tervisehinnangute jaotus meestel ja naistel, kui hinnang ei sõltuks tudengi soost:

	tervis		
sugu	v.hea	hea	keskmine/halb
naine	83 (91)	404 (394)	161 (163)
mees	35 (27)	105 (115)	50 (48)



Meie näites:

$$\chi 2 = (83 - 91)^2/91 + (404 - 394)^2/394 + \ldots + (50 - 48)^2/48 = 4{,}6$$

Leitud statistik on χ2- jaotusega, vabadusastmete arvuga

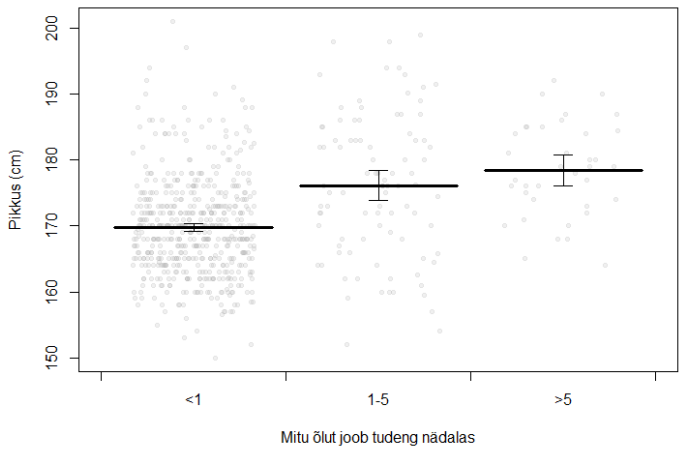
$$df = (r - 1) \times (v - 1) = r\,v - r - v + 1,$$

kus *r* on ridade ja *v* veergude arv uuritavas tabelis.

Vabadusastmete arvuks on siin 2 ja seega ei saa antud juhul seose olemasolu tõestada (χ2-statistiku kriitiline väärtus df = 2 korral on 5,99; olulisustõenäosuseks tuleb p = 0,10)

Eeldustest

Hii-ruut test on ligikaudne test – hii-ruut statistik on vaid ligikaudu hii-ruut jaotusega. Mida suuremad on H_0 eeldusel leitud ootuspärased sagedused N_1, N_2, … N_k, seda paremini kirjeldab hii-ruut jaotus teststatistiku tegelikku jaotust. Kui sagedused N_1, N_2, … N_k on väiksemad kui **5**, siis ei pruugi hii-ruut jaotust kasutav hii-ruut test anda usaldusväärseid tulemusi.



Statistilisest seosest

Vaatluseid genereeriv protsess:

```
x <- rnorm(n, 0, 1)
y <- 2 +2*x +rnorm(n)
```

$X \sim N(0, 1)$

$Y|X \sim N(2+2X, 1)$

$$F_{X,Y} = F_X \cdot F_{Y|X} \quad (\neq F_X \cdot F_Y) = F_{X|Y} \cdot F_Y$$

$Y \sim N(2, 5)$

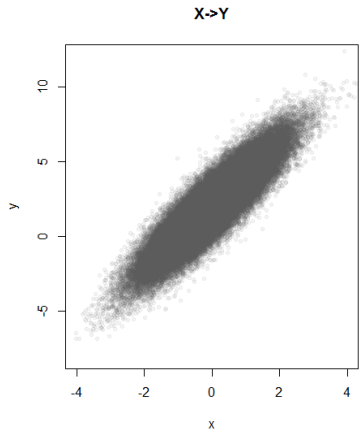
$X|Y \sim N(-4/5+2/5Y, 1/5)$

$$F_{X,Y} = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{[y-(2+2x)]^2}{2}\right)$$
$$= \frac{1}{\sqrt{2\pi \cdot 5}} \exp\left(-\frac{(y-2)^2}{2 \cdot 5}\right) \cdot \ldots$$
$$= \frac{1}{\sqrt{2\pi \cdot 5}} \exp\left(-\frac{(y-2)^2}{2 \cdot 5}\right) \cdot \frac{1}{\sqrt{2\pi \cdot 1/5}} \ldots$$
$$= \frac{1}{\sqrt{2\pi \cdot 5}} \exp\left(-\frac{(y-2)^2}{2 \cdot 5}\right) \cdot \frac{1}{\sqrt{2\pi \cdot 1/5}} \exp\left(-\frac{[x-(2/5y-4/5)]^2}{2 \cdot 1/5}\right)$$

Statistilisest seosest

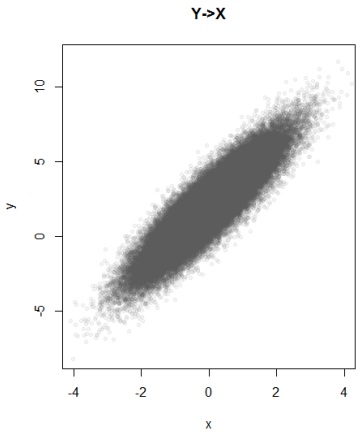
Vaatluseid genereeriv protsess:

```
x <- rnorm(n, 0, 1)
y <- 2 +2*x +rnorm(n)
```



Vaatluseid genereeriv protsess:

```
y <- rnorm(n, 2, 5)
x <- 2/5*y-4/5 +rnorm(n)
```

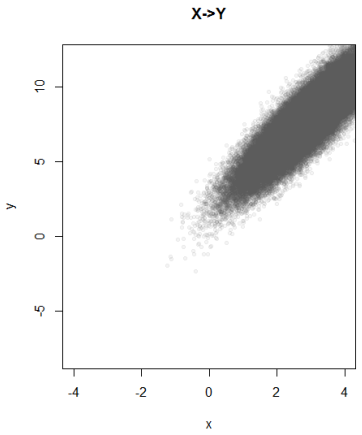


Statistilisest seosest

SEKKUMINE?
Suurendame X-tunnuse väärtust

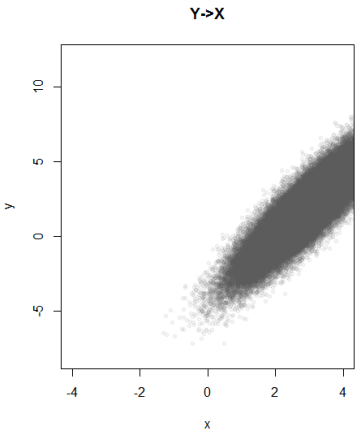
Vaatluseid genereeriv protsess:

```
x <- rnorm(n, 0, 1)+3
y <- 2 +2*x +rnorm(n)
```



Vaatluseid genereeriv protsess:

```
y <- rnorm(n, 2, 5)
x <- 2/5*y-4/5 +rnorm(n) +3
```



Järeldusi

- Tunnuste ühisjaotuse teadmine pole tegijatele (maailma muutjatele) piisav;
- Põhjuslik seos on ebasümmeetriline;
- Vajame lisainformatsiooni (ja täiendavaid tähistusi lisainformatsiooni kirjapanekuks)

Vanim ja kõige pikemate traditsioonidega lahendus – teediagrammid. Näita nooltega milline tunnus mõjutab millist:

$X \longrightarrow Y$ $\qquad$ $Y \longrightarrow X$

Mis on põhjuslik mõju?

Jaan suitsetas ja suri noorelt.

Kas suitsetamine põhjustas Jaani enneaegse surma?

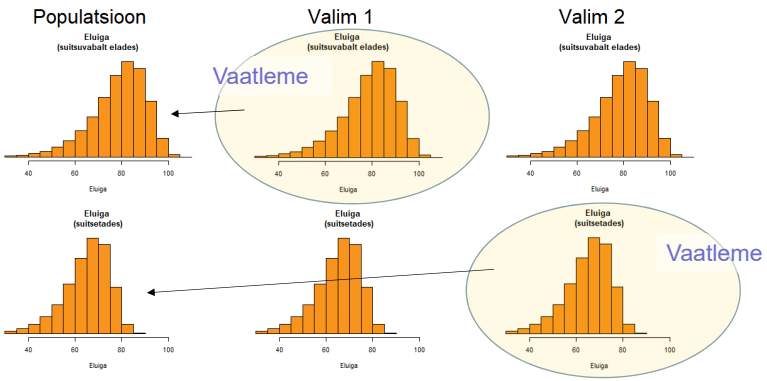
Sellele küsimusele vastamiseks vajame kontrafakte (*counterfactuals*):

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

Järeldus: suitsetamine põhjustas Jaani surma.

Jaani surma pole kunagi täie kindlusega võimalik suitsetamise süüks ajada – sest meie võimuses pole lasta tal kaks korda elada (ükskord suitsetades ja teine kord ilma suitsetamata)

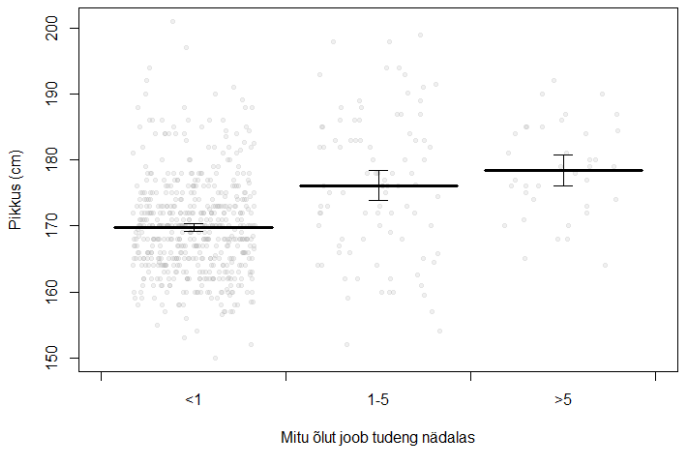
Kuidas leida põhjuslikku mõju (teoreetiliselt)?

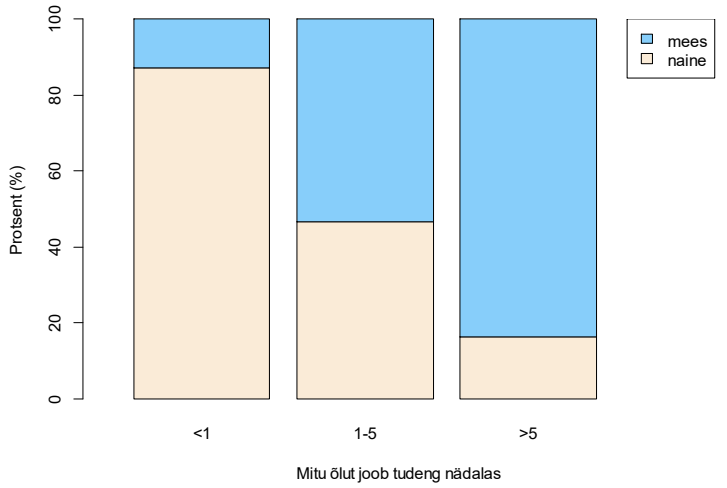


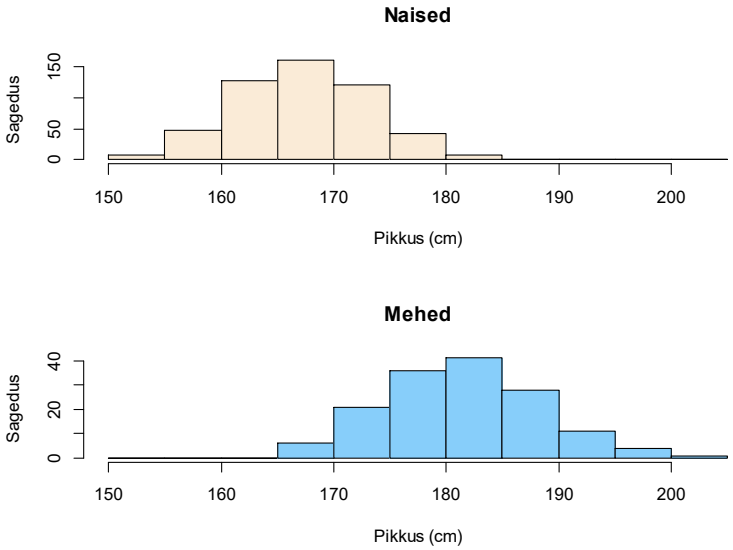
Kuidas leida põhjuslikku mõju (teoreetiliselt)?

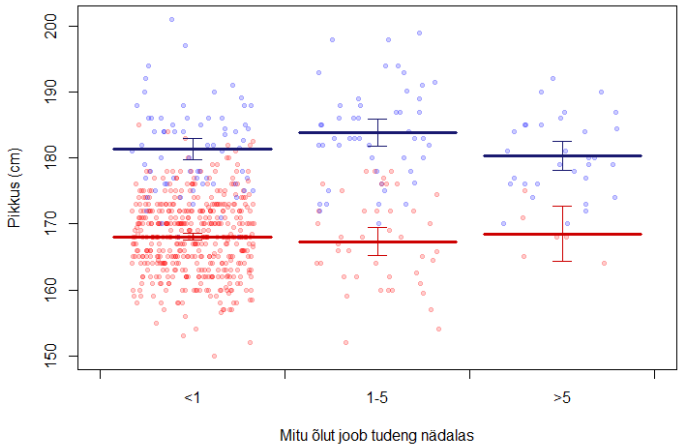
- Kahe valimi võtmise asemel võetakse enamasti üks valim ja sinna valimisse sattunud inimesed jagatakse juhuslikult kahte (või vajadusel enamasse) gruppi – randomiseeritakse.
- Taolist uuringut tuntakse randomiseeritud uuringu (või randomiseeritud katse) nime all.
- Näiteks selleks, et uus ravim saaks kasutusloa Euroopas, peab olema eelnevalt randomiseeritud uuringu abil tõestatud, et tal on soovitud põhjuslik mõju inimeste tervisele või elueale.
- Randomiseeritud uuringud on ideaal, mille poole pürgida ja millega erinevaid statistikameetodeid võrreldakse (kas jõuaksime sama tulemuseni?). Praktikas ei pruugi aga randomiseeritud uuringute korraldamine alati lihtne olla (randomiseeritud uuring intressimäärade mõju kindlakstegemiseks riigi majandusele?)

Kas on alternatiive randomiseeritud uuringule?



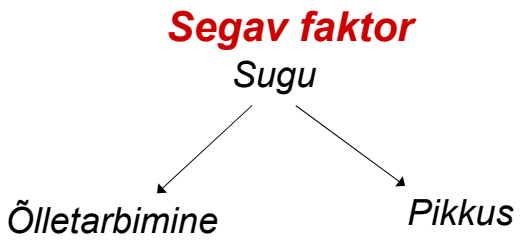






$X = a_1 S + e_X$
$Y = a_2 S + e_Y$

cov(*X*, *Y*) = ….



Põhjuslik mõju vs statistiline seos

$F_{Y|X} \neq F_Y$

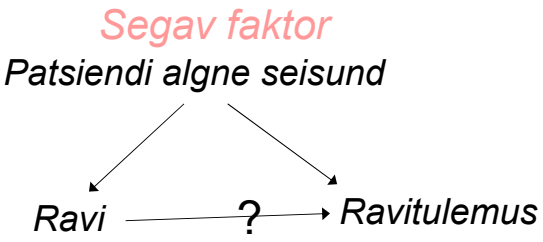
Eksisteerib statistiline seos *X* ja *Y* vahel

$F_{Y|X,S_1,\ldots,S_k} \neq F_{Y|S_1,\ldots,S_k}$

$Y \nrightarrow X$

Eksisteerib *X-i* põhjuslik mõju *Y*-le

Kui ravi ei mõjuta tulemust

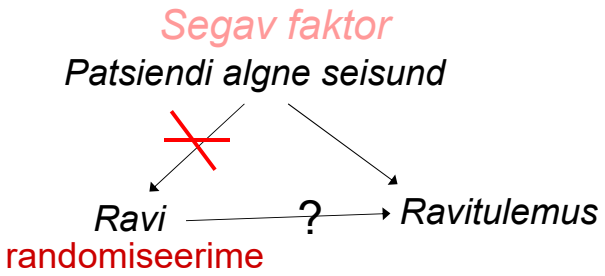


cov(*Ravi, Ravitulemus*) = cov($a_1\ S + e_1$; $a_2\ S + e_2$) = $a_1\ a_2$ D(S)

Kui ravi ei mõjuta tulemust,

Ravitulemus <- $f(S_1, S_2, ...)$

siis on tunnused *ravi* ja *ravitulemus*
teineteisest sõltumatud...

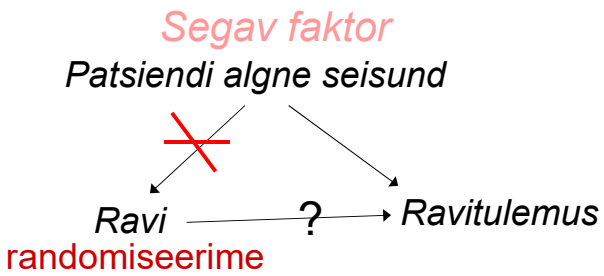


cov(*Ravi, Ravitulemus*) = cov($a_1\ S + e_1$; $a_2\ S + e_2$) = $a_1\ a_2$ D(S) = 0

Kui ravi mõjutab tulemust

Ravitulemus <- f(*Ravi*, S_1, S_2, ...)

siis eksisteerib statistiline sõltuvus tunnuste
ravi ja *ravitulemus* vahel...



cov(*Ravi, Ravitulemus*) = cov(*Ravi*; $a_2\ S + b\ Ravi + e_2$) = b D(*Ravi*)
= e_1