

Statistiline seos (statiline sõltuvus)

statistical dependence; statistical association

Mis on statistiline seos?

Matemaatiline definitsioon:

Tunnused X ja Y on
sõltumatud, kui

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

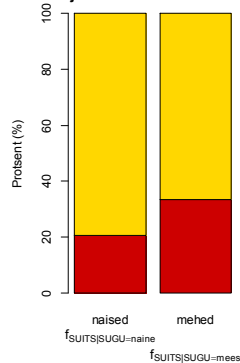
tunnuse X tinglik
jaotus ei sõltu valitud
 Y -tunnuse väärtusest

Tunnused X ja Y on
sõltuvad, kui

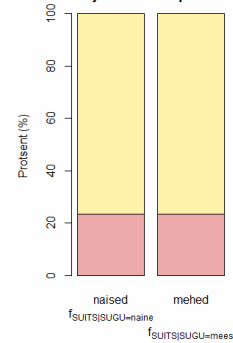
$$\exists c, \quad f_{X|Y=c} \neq f_X$$

tunnuse X tinglik
jaotus sõltub valitud
 Y -tunnuse väärtusest

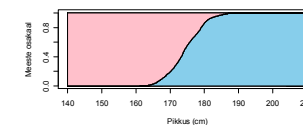
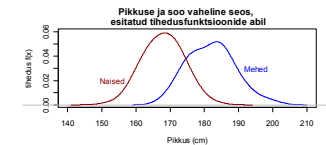
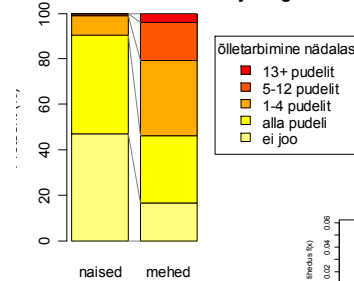
Seos suitsetamise
ja soo vahel on olemas

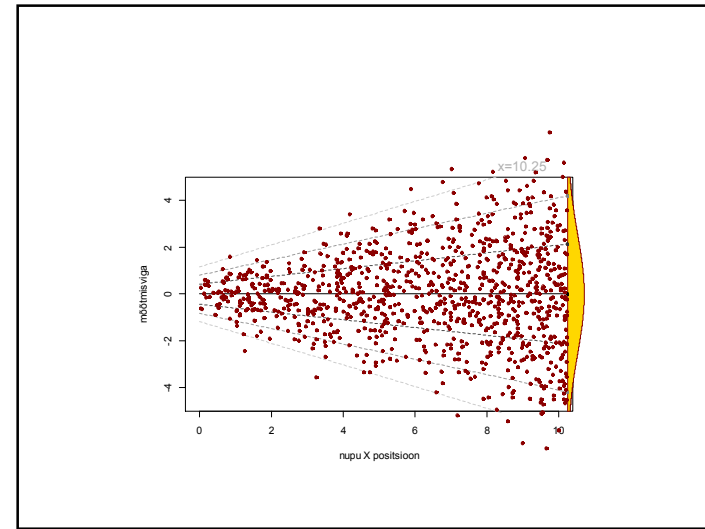
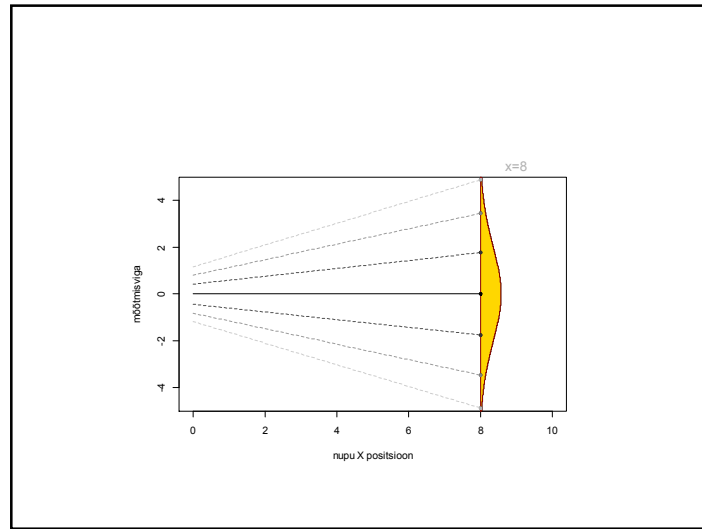


Seost suitsetamise
ja soo vahel pole

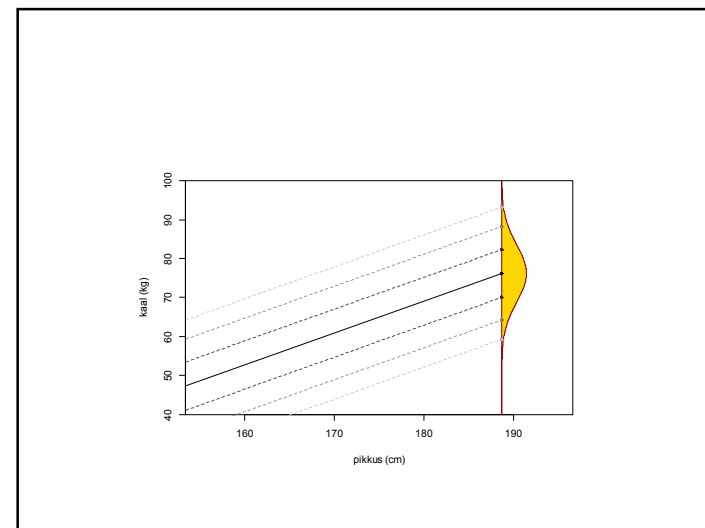
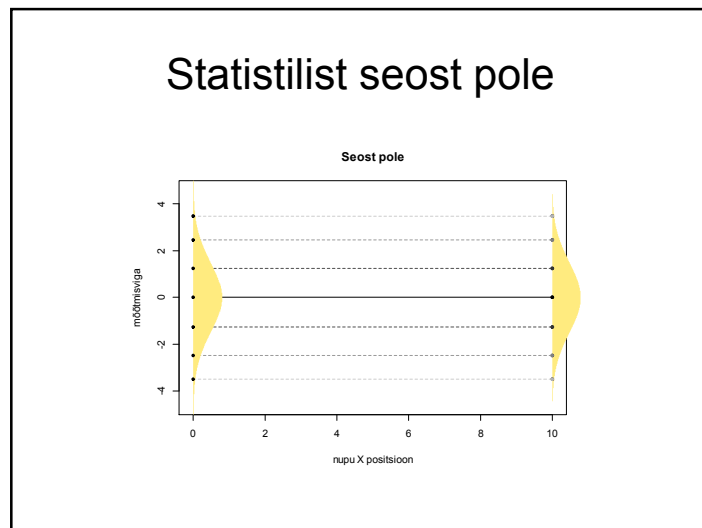


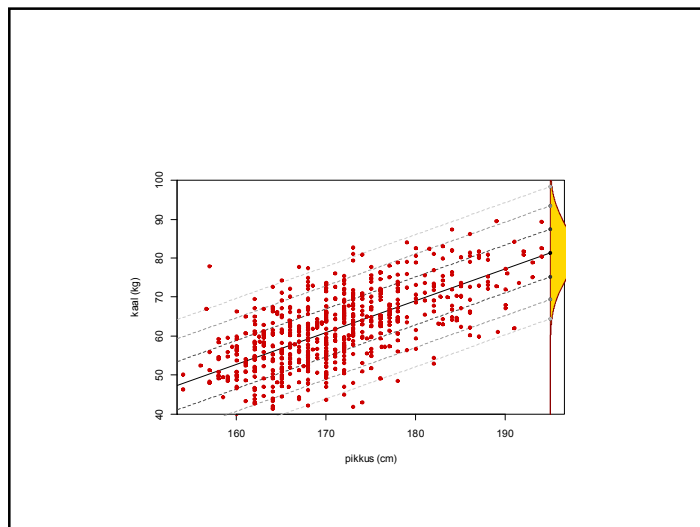
Õlle tarbimine ja sugu



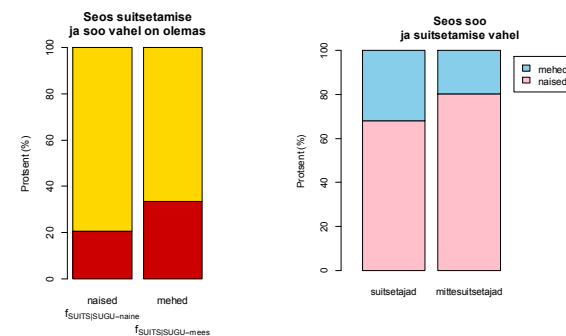


Statistilist seost pole

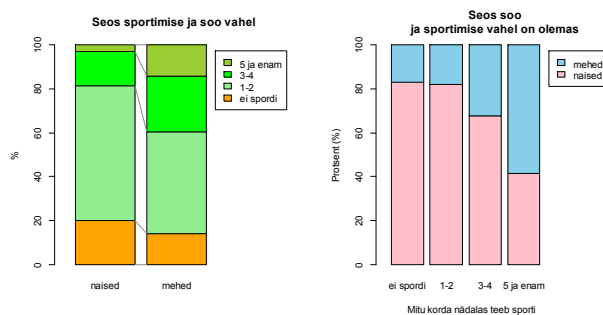




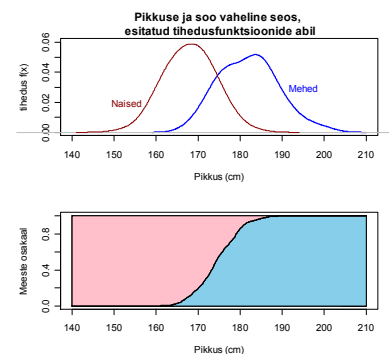
Statistiline seos on sümmeetriline – kui on seos tunnuste X ja Y vahel, siis eksisteerib ka seos tunnuste Y ja X vahel!



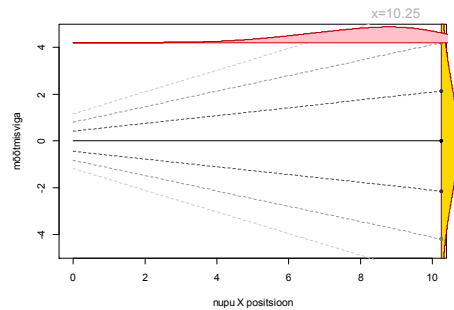
Statistiline seos on sümmeetriline – kui on seos tunnuste X ja Y vahel, siis eksisteerib ka seos tunnuste Y ja X vahel!



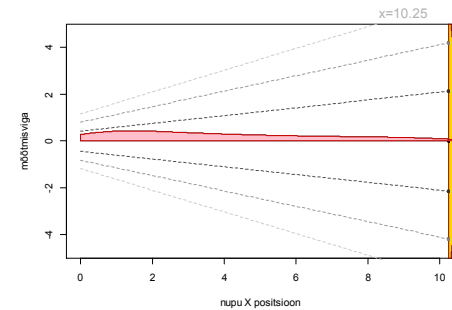
Statistiline seos on sümmeetriline – kui on seos tunnuste X ja Y vahel, siis eksisteerib ka seos tunnuste Y ja X vahel!



Statistiline seos on sümmeetriline – kui on seos
tunnuste X ja Y vahel, siis eksisteerib ka seos
tunnuste Y ja X vahel!



Statistiline seos on sümmeetriline – kui on seos
tunnuste X ja Y vahel, siis eksisteerib ka seos
tunnuste Y ja X vahel!



Statistiline seos tunnuste X ja Y
vahel eksisteerib, kui:

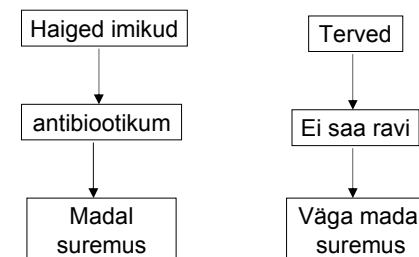
$X \rightarrow Y$ (X mõjutab põhjuslikult Y -t)

$X \leftarrow S \rightarrow Y$ eksisteerib segav faktor
S-haigus, X-antibiootikum,
Y-imiku elulemus

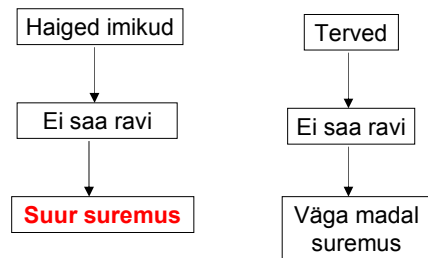
$X \rightarrow S \rightarrow Y$ X - SNP/geen,
S – inimese pikkus
Y – kopsumaht

$Y \rightarrow X$ (Y mõjutab põhjuslikult X -i)

Antibiootikume saanute seas
on suremus suurem...



Statistiline seos ei pruugi olla põhjuslik seos!!!



Põhjuslik seos (Causal relationship)

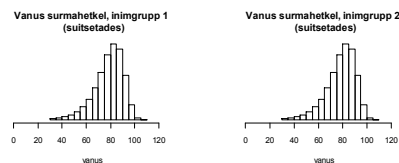
Counterfactuals (kontrafaktid)

Jaani suitsetas ja suri noorena.

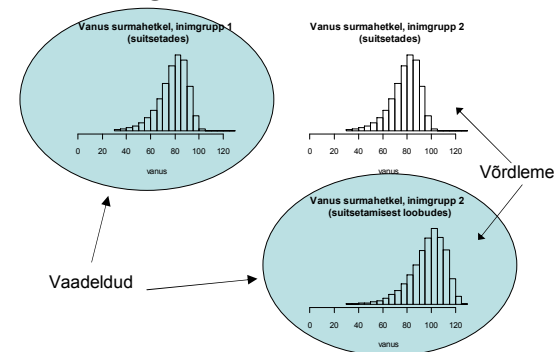
Kui Jaan poleks suitsetanud, poleks ta noorena surnud.

Järelikult põhjustas suitsetamine Jaani surma.

Üksikisiku puhul saavutamatu, inimeste grupi puhul saavutatav



Üksikisiku puhul saavutamatu, inimeste grupi puhul saavutatav

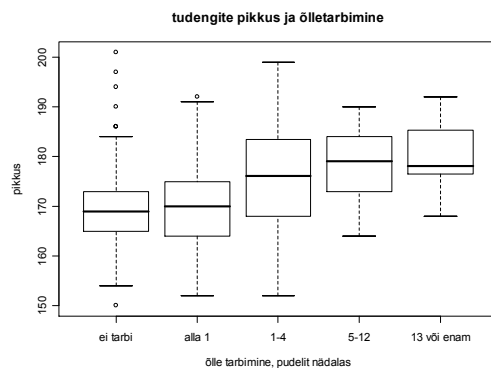


Suurte arvude seadus: randomiseerimine
muudab grupid võrreldavaks

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

Näiteid (võimalikest) segavatest faktorites

- Vanus – vanemad inimesed eelistavad ühte ravi; noored teist – aga vanemad inimesed tervevad aeglasemalt;
- Raviarst – innukam raviarst eelistab ehk moodsamat ravi kui väheminnukam arst. Ühtlasi hoolitseb innukam arst ka muul viisil patsientide eest paremini, mistõttu tema patsiendid võivad paraneda kiiremini;
- Haiguse raskusaste arsti poole pöördumisel – arst võib eelistada raskematele haigetele määrata ühte ravi ja kergematele teist ravi;
- Geenid; arstiabi kättesaadavus; patsientide rahalised võimalused; ...

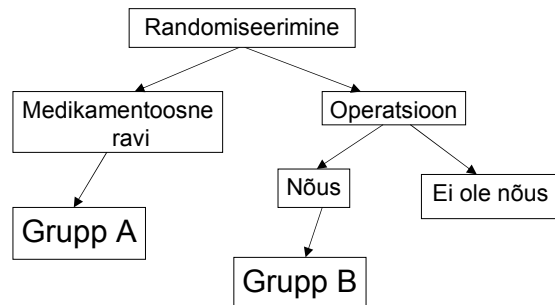


Vead randomiseerimisel

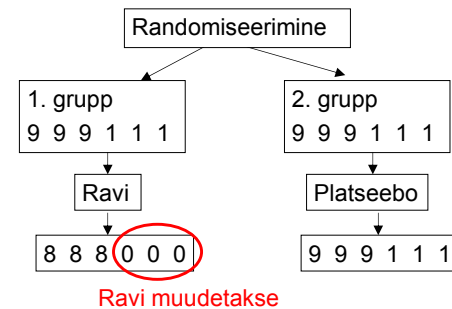
Randomiseerimine tagab kahe võrreldava
grupi tekke.

Kui peale randomiseerimist lisatakse ühte
gruppi täiendavaid objekte või
eemaldatakse vaatluseid, võivad grupid
muutuda mittevõrreldavaks.

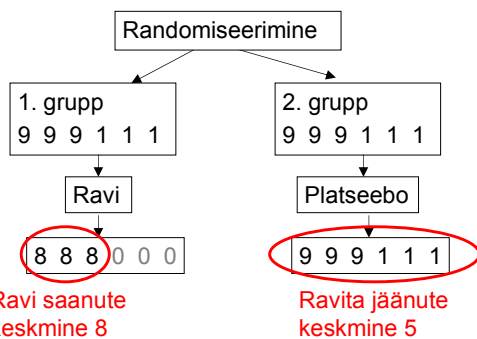
Näide veast – liiga hiline nõusoleku küsimine uuringus osalemiseks



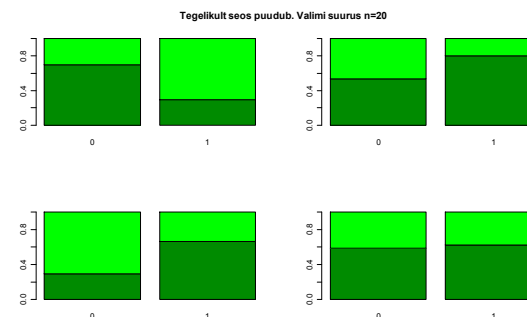
Näide veast – randomiseerime, aga võrdleme ainult ravi tegelikult saanud patsiente



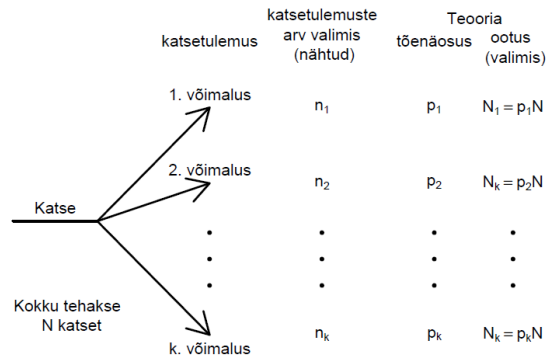
Näide veast – randomiseerime, aga võrdleme ainult ravi tegelikult saanud patsiente



Kas seos on tegelik või näiline (valimi juhuslikkus petab meid)?



Hii-ruut test I



$$\chi^2 = \sum_{i=1}^k \frac{(n_i - N_i)^2}{N_i}$$

$$\chi^2 \sim \chi^2_{df=k-1} \quad (l\text{-hinnatud parameetrite arv})$$

Näide

- Testiti 100000 SNP'i seost fenotüübiga
- 6000 SNP'i puhul oli p-väärtus väiksem kui 0,05
- 5500 SNP'i puhul oli p-väärtus 0,05...0,1
- 88500 SNP'i puhul oli p-väärtus suurem kui 0,1.

Kas geenide ja vaadeldud fenotüübi vahel on seos?

Kontrollitava teooria ennustus

- Kui ühelgi SNP'il poleks seost uuritava fenotüübiga, peaks 5% SNP'ide puhul p-väärtus tulema väiksem kui 0,05: $p_1=0,05$
- ja 10% SNP'ide puhul peaks p-väärtus tulema väiksem kui 0,1, seega p-väärtust 0.05...0.1 peaksime samuti nägema 5% SNP'ide korral. $p_2=0,05$
- 0,1-st suuremat p-väärtust peaksime nägema 90%-l juhtudest, $p_3=0,9$

$N=100000$

$$N_1 = 5000 \quad n_1 = 6000$$

$$N_2 = 5000 \quad n_2 = 5500$$

$$N_3=90000 \quad n_3=88500$$

$$\begin{aligned}\chi^2 &= \frac{(6000-5000)^2}{5000} + \frac{(5500-5000)^2}{5000} + \frac{(88500-90000)^2}{90000} \\ &= 200 + 50 + 25 \\ &= 275\end{aligned}$$

Hii-ruut statistiku kriitiline väärtus ($df=3-1=2$ korral): 5,99, järelikut on H_0 kummutatud (p -väärtus tuleb arvutustäpsuse piires 0).

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

Näide

kas esineb seos tudengi tervisehinnangu ja tema soo vahel?

Tabel (arstiteaduskonna 2. kursus aastatel 2001-2005):

	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.

Küsimine: 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(60,74%)	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.

Leiname, 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)	

		tervis (%)		
	v.hea	hea	keskmine/halb	
	14.1%	60.74%	25.2%	

Meie näites:

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

Leitud statistik on χ^2 - jaotusega, vabadusastmete arvuga

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

kus r on ridade ja v veergude arv uuritavas tabelis.

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