Statistiline seos

ja Hii-ruut test

Deterministlik seos

Ühtede muutujate/katsetingimuste/tunnuste väärtuste muutmisel muutub ka meid huvitava tunnuse või näitaja väärtus. Meid huvitava näitaja väärtuse saab üheselt leida arvutusvalemi abil, juhuslikust pole.

Näide:

Minnes rahavahetaja juurde sooviga vahetada oma eurod Läti lattideks vaatame kurssi (näiteks 0,729) ja saame leida, mitu latti saame, kui vahetame *x* eurot:

Raha (lattides) = 0,729 *x*

Saadud kroonide arv pole (antud päeval) juhuslik – see sõltub deterministlikult vahetatavate eurode arvust. Tegemist on deterministliku seosega.

Statistiline seos (*association*)

Ühtede muutujate/katsetingimuste/tunnuste väärtuste muutmisel muutub ka meid huvitava tunnuse või näitaja jaotus.

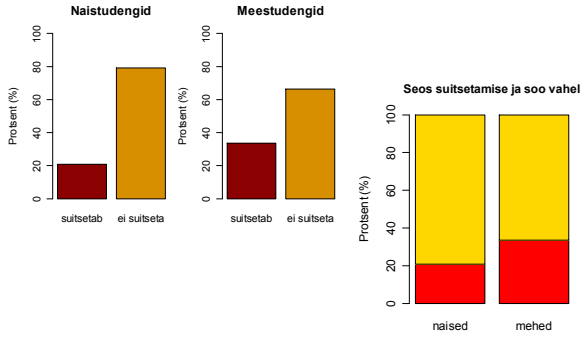
Näiteks: katsetingimuste teadmine ei pruugi meil veel täpselt öelda, milline tuleb katsetulemus; aga teatud katsetingimuste juures on mõned katsetulemused tõenäolisemad kui teiste katsetingimuste korral.

Konkreetne näide:

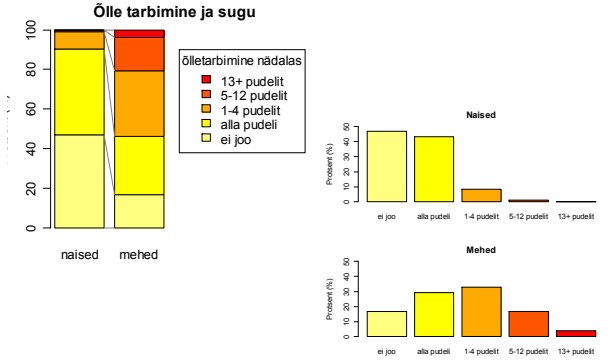
Koheselt peale munemist inkubaatorisse paigutatud linnumunast koorub linnupoeg tõenäosusega 0,9; kui aga oodata 8 päeva peale munemist ja alles siis muna inkubaatorisse paigutada, koorub temast linnulaps vaid tõenäosusega 0,75 – seega on tunnuste „ooteperioodi pikkus“ ja „koorumisedukus“ (koorub/ei kooru) vahel statistiline seos.

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

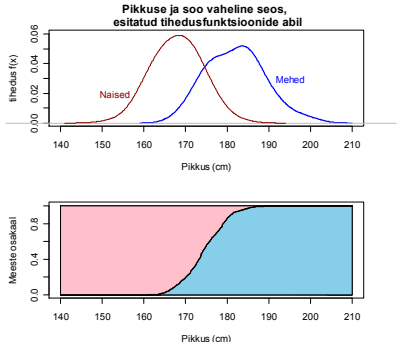
Näide 1 – seos kahe binaarse tunnuse vahel

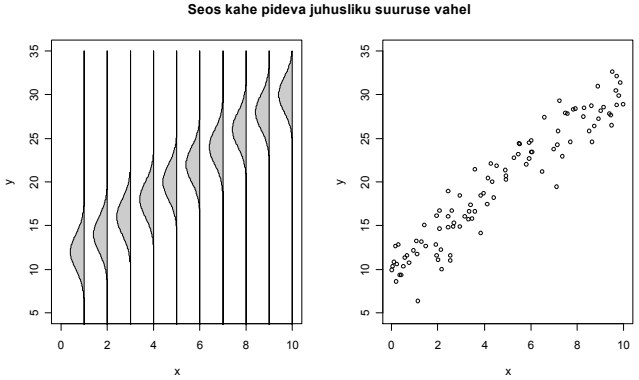


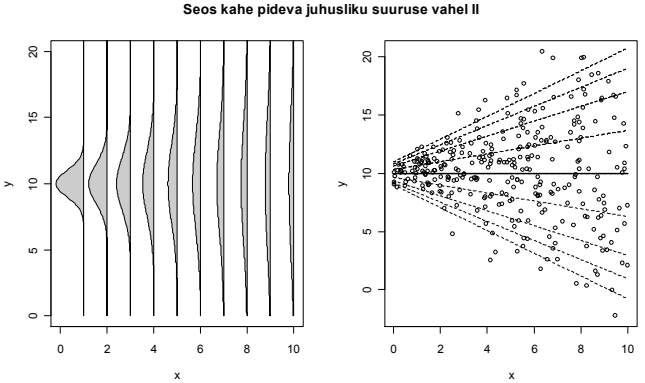
Näide 2: seos õlle tarbimise ja soo vahel



Seos pideva ja binaarse tunnuse vahel (seos soo ja pikkuse vahel)







prognoosimine vs toimuva mõjutamine. Põhjuslik seos.

Tulekahjut kustutama sõitnud tuletõrjeautode arv ja põlengu poolt tekitatud kahju suurus on tugevas seoses – mida enam autosid, seda suurem kahju

Kas nimetatud seose teadmisest on kasu, kui olete:

- Päästeameti direktor (saata välja vähem tuletõrjeautosid?)
- Ajakirjanik (kumba kahest tulekahjust kajastamiseks valida?)

„Kured läinud, kurjad ilmad“

Kas kurgede jõuga kinnihoidmine Eestis aitaks talve vältida?

Prognoosimine vs toimuva mõjutamine II

Toimuva mõjutamiseks peame teadma, mis midagi põhjustab. Muutes algpõhjuseid saame esile kutsuda teistsugust homset.

Kõrvaltvaatajana toimuva prognoosimiseks pole enamasti vaja teada, kas mingi seos tunnuste vahel on põhjuslik või mitte. Erandiks võib osutuda arukate olendite (inimeste) tegevuse prognoosimine – kui inimesed taipavad, millist mittepõhjuslikku seost me prognoosimisel kasutame, võivad nad meid kerge vaevaga eksiteele viia.

“*Uut ravi saanud patsiendid tervenesid kiiremini/elasid kauem kui vanal viisil ravitud patsiendid.”*

Kas sellistest andmetest järeldub, et uus ravi on parem?

Näide (hormoonasendusravi)

Haigestumine südame isheemiatõppe (Coronary Heart Disease) on 40-50% väiksem nendel postmenopausi perioodis naistel, kes saavad hormoonasendusravi.

Stampfer M, Colditz G. Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiologic evidence. *Prev Med*. 1991;20:47-63.

Grady D, Rueben SB, Pettiti DB, et al. Hormone therapy to prevent disease and prolong life in postmenopausal women. *Ann Intern Med*. 1992;117: 1016-1037.

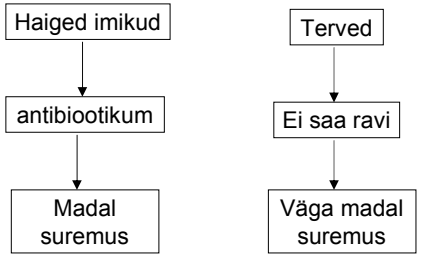
Rijpkema AH, van der Sanden AA, Ruijs AH. Effects of postmenopausal estrogen-progesterone therapy on serum lipids and lipoproteins: a review. *Maturitas*. 1990;12:259-285.

Adams MR, Kaplan JR, Manuck SB, et al. Inhibition of coronary artery atherosclerosis by 17-beta estradiol in ovariectomized monkeys: lack of an effect of added progesterone. *Arteriosclerosis*. 1990;10: 1051-1057.

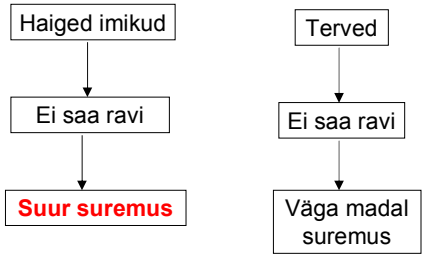
Miks tervenevad uut ravi saanud patsiendid kiiremini?

- Sest uus ravi on parem?
- Sest uut ravi anti tervematele patsientidele?
- Sest uut ravi kirjutasid enamasti välja ülihoolitsevad arstid, kes ka muudel viisidel aitasid patsiente rohkem (parem põetus vms)?

Antibiootikume saanute seas on suremus suurem...

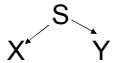


Ravi saanute/mittesaanute grupid pole juba algselt võrreldavad



Statistiline seos tunnuste X ja Y vahel eksisteerib, kui:

X → Y (X mõjutab põhjuslikult Y-t)



eksisteerib segav faktor
S-haigus, X-antibiootikum,
Y-imiku elulemus

X → S → Y
X - SNP/geen,
S – inimese pikkus
Y – kopsude võimekus

Y → X (Y mõjutab põhjuslikult X'i)

Põhjuslik seos (*Causal relationship*)

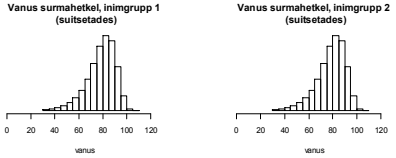
Counterfactuals (kontrafaktid)

Jaan suitsetas ja suri noorena.

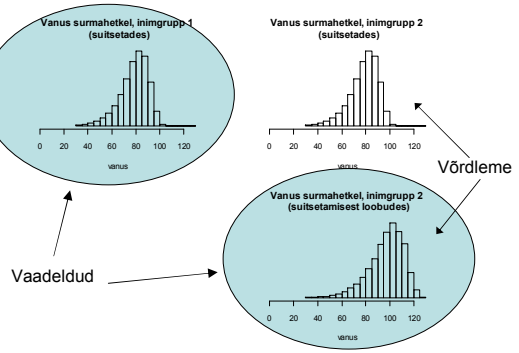
Kui Jaan poleks suitsetanud, poleks ta noorena surnud.

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

Hormoonasendusravi näide

Mis juhtub hormoonasendusravi kasulikkusega, kui alustame võrreldavate gruppidega (randomiseeritud kliiniline katse)?

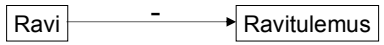
Südame isheemiatõppe (Coronary Heart Disease) haigestumine on suurem nendel patsientidel, kes saavad hormoonasendusravi.

Risks and Benefits of Estrogen Plus Progestin in Healthy Postmenopausal Women: Principal Results From the Women's Health Initiative Randomized Controlled Trial *JAMA*. 2002;288(3):321-333 (doi:10.1001/jama.288.3.321)

Confounder – segav tunnus

Confundo (ladina. k) – sassi, segamini ajama; moonutama, tundmatuseni muutma

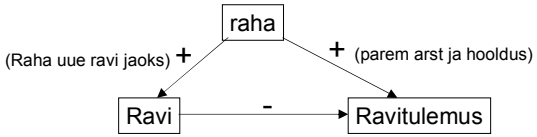
Segamine (confounding) – kui meid huvitava tunnuse (protsessi) mõju pole võimalik eristada teiste tunnuste (protsesside) mõjust, öeldakse, et tegemist on segamisega (segavate tunnuste poolt).



Confounder – segav tunnus

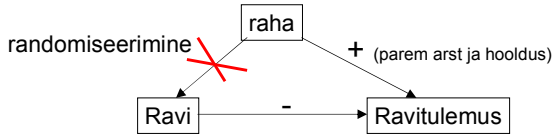
Confundo (ladina. k) – sassi, segamini ajama; moonutama, tundmatuseni muutma

Segamine (confounding) – kui meid huvitava tunnuse (protsessi) mõju pole võimalik eristada teiste tunnuste (protsesside) mõjust, öeldakse, et tegemist on segamisega (segavate tunnuste poolt).



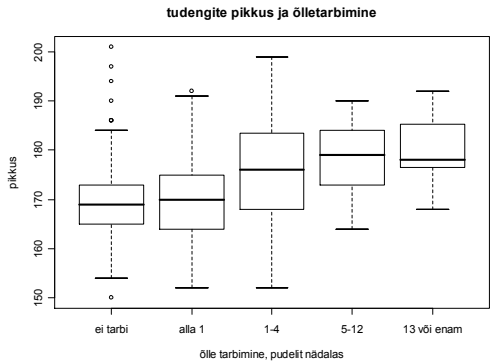
Randomiseerimine lõpetab segava faktori mõju (muudab grupid võrreldavaks)

Selleks, et mingi tunnus oleks segavaks tunnuseks (segavaks faktoriks), peab ta mõjutama nii ravi kui ravitulemust. Tänu randomiseerimisele kaob patsiendi rahakoti suuruse mõju raviviisi valikule – ja tunnus raha pole enam segavaks faktoriks!



Näiteid (võimalikest) segavatest faktoritest

- Vanus – vanemad inimesed eelistavad ühte ravi; noored teist – aga vanemad inimesed tervenevad 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; ...

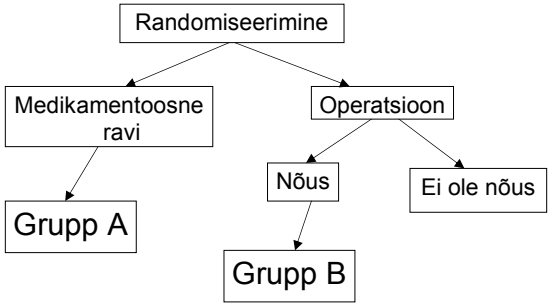


Vead randomiseerimisel

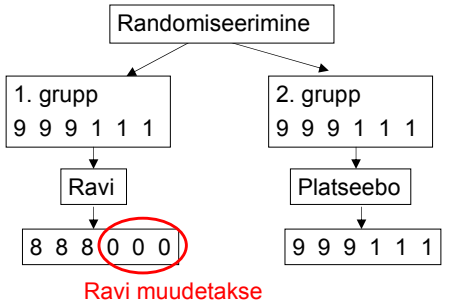
Randomiseerimine tagab kahe võrreldava grupi tekke.

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

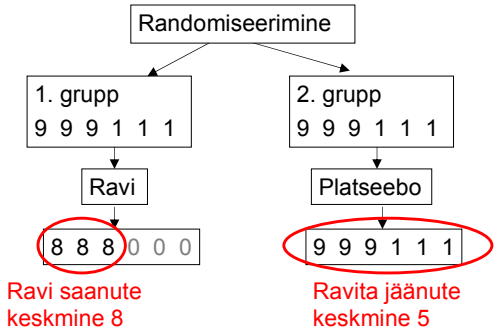
Näide veast – liiga hiline nõusoleku küsimine uuringus osalemiseks



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



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

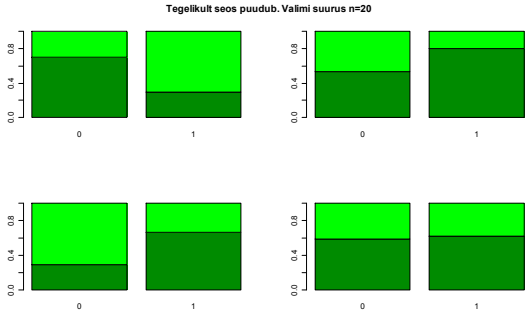


Ravikavatsuse põhimõte – Intention to treat analysis

Võrrelda tuleb ravirühma randomiseerituid kontrollgruppi randomiseeritutega. Ka siis, kui patsient tegelikult ravi ei saanud, tuleb ta lugeda ravirühma kuuluvaks (võrdle neid, keda kavatsesid ravida, nendega, keda kavatsesid jätta ravita).

Märkus: ravikavatsuse põhimõtte jälgimine ekvivaentsuskatsete (kas odav koopiaravim on samahea kui kallis originaal) juures ei pruugi olla soovitatav.

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

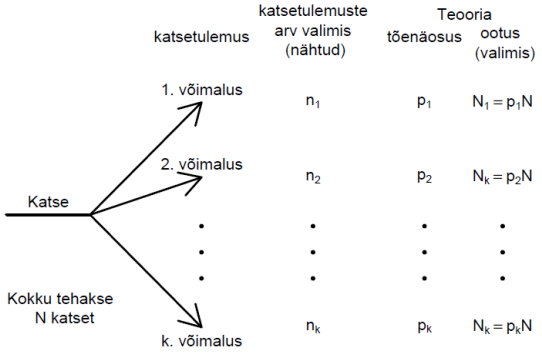


Seose olemasolu testimine

	normaaljaotus	**pidev (pole normaaljaotus)**	**binaarne**	**nominaalne**
pidev	**regressioon-analüüs**	**regressioon-analüüs;** üldistatud lineaarased mudelid,...	logistiline regressioon	diskriminant-analüüs
binaarne	**t-test**	Kolmogorov-Smirnovi test; Wilcoxoni test; **t-test (suur valim)**	Fisheri täpne test, **hii-ruut test**	
nominaalne	**dispersioon-analüüs (ANOVA)**	Krusal-Wallis	logistiline regressioon; **hii-ruut test**	**hii-ruut test**

Suuremas kirjas need testid ja analüüsid, mida me antud kursuse raames oleme käsitlenud / saame käsitlema

Hii-ruut test I



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

$\chi^2 \overset{H_0}{\sim} \chi^2_{df = k-l}$ (*l*-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% juhtudest, p_3=0,9

N=100000

N_1= 5000 n_1= 6000

N_2= 5000 n_2= 5500

N_3=90000 n_3=88500

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

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

Statistiline seos kahe mittearvulise 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.

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:

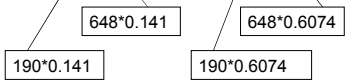
	tervis (%)		
v.hea	hea	keskmine/halb	
118(14,1%)	509 (60,74%)	211 (25,2%)	

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.

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 $\chi 2$- 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 tostada ($\chi 2$-statistiku kriitiline väärtus df = 2 korral on 5,99; olulisustõenäosuseks tuleb p = 0,10)