

Sissejuhatus.
Mis on kvalitatiivsed andmed?
Mida nendega teha ja mida karta?

Tunnuste liigitamine

Mittearvulised ehk
kvalitatiivsed tunnused

- Nominaalsed tunnused**
silmaäär (sinisilm; pruunid silmad; rohelised; hallid; ...)
liik (*Taraxacum officinale*; *Anthemis arvensis*; *Trifolium repens*; ...)
rahvus (eestlane; venelane; soomlane; ...)
parteiline eelistus (Reformierakond; EKRE; Keskerakond; ...)
- Järjestustunnused**
hinnangud (väga hea / hea/ keskpärane / jube);
Likerti skaala (ei nõustu üldse; pigem ei nõustu; ükskõik; pigem nõustun;
nõustun täielikult)
haridus (*algharidus*; *keskharidus*; *kõrgharidus*; *doktorikraad*)

Arvulised ehk
kvantitatiivsed tunnused

- Diskreetsed tunnused**
sõprade arv; terade arv viljapeas; ...
- Pidevad tunnused**
Pikkus; Kaal; kontsentratsioon; temperatuur; ...

Tunnuste liigitamine

Kas tegemist on nominaalse
tunnusega, järjestustunnusega,
diskreetse või pideva
tunnusega?

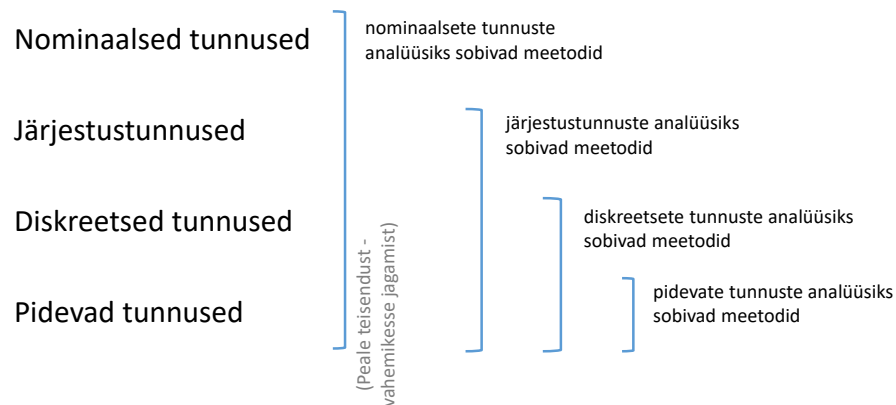
	tunnus 1	tunnus 2	tunnus 3	tunnus 4
1	1	1	1	1
2	2	1	2	2
2	3	1	1	
3	1	2	3	
1	1	3	...	
2	2	1		

Tunnuste liigitamine

Kas tegemist on nominaalse
tunnusega, järjestustunnusega,
diskreetse või pideva
tunnusega?

	kasvukoht	soomlastest sõpru	vastaja rahvus	kodutee pikkus (km)
1	1	1	1	1
2	2	1	2	2
2	3	1	1	
3	1	2	3	
1	1	3	...	
2	2	1		

Statistikameetodid



Tunnuste liigitamine – teine vaatenurk

Uuritav tunnus (Y)
sõltuv tunnus
funktsioonitunnus
response variable
dependent variable

Selle tunnuse jaotust tahame kirjeldada; teda prognoosida või uurida mis võiks seda tunnust mõjutada

seletav tunnus või tunnused (X)
sõltumatu tunnus
argumenttunnus
explanatory variable
independent variable

Aitab seletada, miks uuritava tunnuse jaotus muutub – võimaldab kas sekkuda (muutes X-i väärtust teatud viisil, saame mõjutada Y-tunnuse väärtuseid meile soovitavas suunas) või valmistuda selleks mis tuleb (ilmad lähevad külmaks – inimesed ostavad rohkem kasukaid ja saapaid ja vähem kingi)

Kuidas on andmed kogutud?

Uuringute liigitamine

Ristlõikeline uuring (*Cross-sectional study*)

Küsitlеме juhuslikku valimit: kas suitsetad? kas sul on kopsuhaigus?

Prospektiivne ehk ettesuunatud uuring (*Prospective study*)

Randomiseeritud uuring/katse (*Randomized trial/study; Clinical trial*)

Uuritavad jagatakse juhuslikult gruppidesse. Iga grupp saab erinevat ravi või mõjutust (X). Teatud aeg hiljem – näiteks aasta möödudes – mõõdetakse uuritava tunnuse väärtus (Y).

Kohortuuring (*Cohort study*)

Fikseeritakse mingi grupp katsealuseid ja jälgitakse mis nendega tulevikus juhtub.

Retrospektiivne ehk tagasivaatav analüüs (*Retrospective study*)

Juht-Kontrolluuring (*Case-Control Study*)

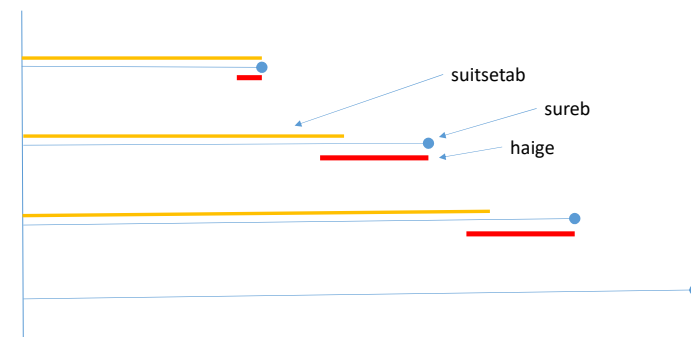
Valitakse välja mingi arv juhte (näiteks haigeks jäänud inimesi) ja mingi arv kontrolle (näiteks terveid inimesi) ja otsitakse minevikust millised olid nende X-tunnuste väärtused (näiteks kas nad lapsena tegid sporti või mitte)

Kohortuuring (*Cohort study*)

Fikseeritakse mingi grupp katsealuseid ja uuritakse, mida nad minevikus tegid.

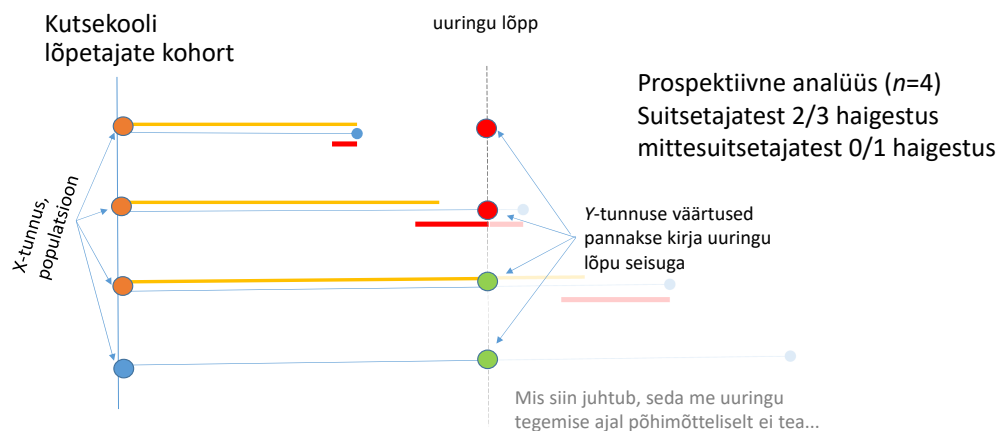
Samad inimesed, sama küsimus – erinevad uuringud

Kutsekooli lõpetajad



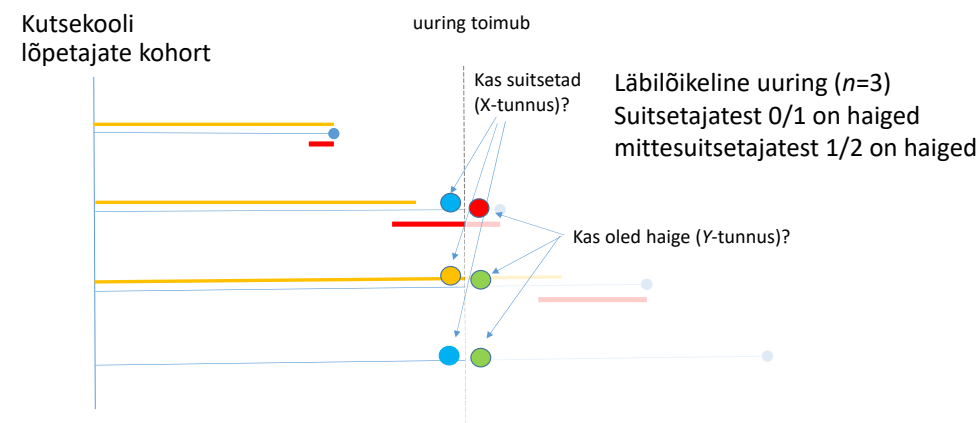
Samad inimesed, sama küsimus – erinevad uuringud

I Ettesuunatud uuring (Prospektiivne uuring)



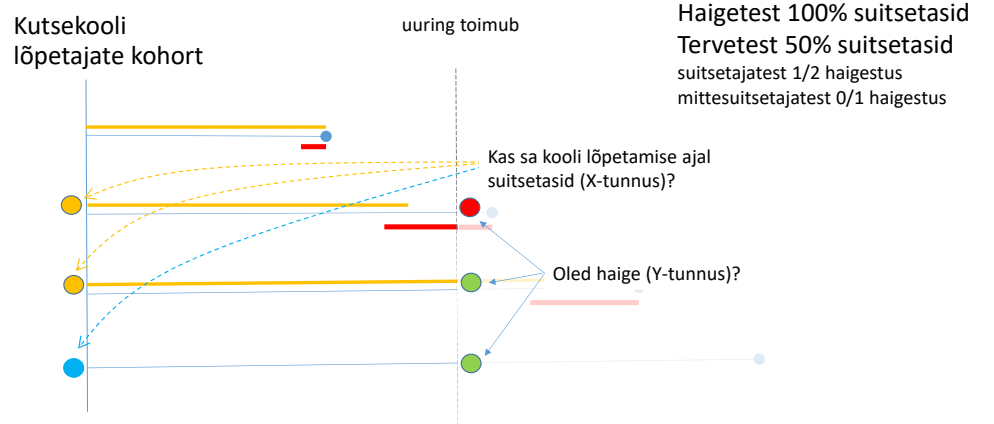
Samad inimesed, sama küsimus – erinevad uuringud

II Läbilõikeline uuring (Cross-sectional study)



Samad inimesed, sama küsimus – erinevad uuringud

III Tagasivaatav uuring (Retrospektiivne)



Kogutud andmed – kindel ajahetk, kõikne uuring

Prospektiivne uuring

	haigus	terve	
suitsetaja	2	1	Suitsetajad jäid märksa sagedamini haigeks (66,7% suitsetajatest haigestus)
mittesuitsetaja	0	1	

Läbilõikeline uuring

	haigus	terve	
suitsetaja	0	1	Mittesuitsetajatel on tõenäolisemalt haigus
mittesuitsetaja	1	1	

Retrospektiivne uuring

	haigus	terve	
suitsetaja	1	1	Praegused haiged suitsetasid kooli lõpus suurema tõenäosusega kui hetkel terved koolilõpetajad
mittesuitsetaja	0	1	

Ettesuunatud, tagasivaatav või läbilõikeline uuring?

Võetakse 1000 esimest Eestis covid-vaktsiini saanud inimest ja vaadatakse, kes neist 6 kuu jooksul peale vaktsineerimist jäid covid-19-sse ja kui paljud neist surid.

Veneaegsete sõjakomissariatide dokumentide põhjal leitakse inimesed, kes saadeti Tšernobõli radioaktiivset saastust küürima. Võrdluseks võeti sama arv tollel ajal samas vanuses mehi, keda Tšernobõli ei saadetud. Kõigi uuritavate kohta märgitakse ülesse kas nad on elus või surnud, kas neil on praegu vähk või on nad hoopis hetkel igati terved.

Tartu haiglasse sattus 15 toidumürgitusega patsienti. Igaühele neist otsiti kolm paarilist – sama vanusega ja samast soost tervet inimest. Kõigi käest küsiti, kus nad eelmisel nädalal söömas on käinud või milliseid toiduaineid on tarbinud.

Soolevähi sõeluuringus uuriti, kas inimesel on soolevähk või mitte. Lisaks pandi kirja ka mitmete teiste tunnuste – näiteks inimese vanus, sugu jne – väärtused.

Jaotus ja tema hinnang

Tunnuse „kiirabi kutsumise põhjus“ jaotus

infarkt	õnnetus	kõhuvalu	jutusoov
π_1	π_2	π_3	π_4

Kus $\pi_1 = P(\text{Põhjus} = \text{„infarkt“})$ jne.

Kvalitatiivsete andmete puhul ei saa leida uuritava tunnuse keskmist või standardhälvet, küll aga saame uurida tema jaotust – või edaspidi – ka seda, kuidas see jaotus võib muutuda.

Hindame juhusliku suuruse jaotust

Kui palju kiirabi väljakutseid mingi põhjus nädala aja jooksul tekitas

infarkt	õnnetus	kõhuvalu	jutusoov
n_1	n_2	n_3	n_4

Leiame hinnangud tõenäosustele – kasutame suhtelisi sagedusi

infarkt	õnnetus	kõhuvalu	jutusoov
n_1/n	n_2/n	n_3/n	n_4/n

Hinnang on juhuslik

infarkt	õnnetus	kõhuvalu	jutusoov
n_1/n	n_2/n	n_3/n	n_4/n

teadlane Kalle:

infarkt	õnnetus	kõhuvalu	jutusoov
0,12	0,24	0,4	0,24

teadlane Malle:

infarkt	õnnetus	kõhuvalu	jutusoov
0,09	0,21	0,45	0,25

Miks on hinnang juhuslik?

infarkt	õnnetus	kõhuvalu	jutusoo
N_1	N_2	N_3	N_4

Sagedused N_i on juhuslikud suurused – järgmises valimis võivad nad omada teistsuguseid väärtuseid.

Selleks, et saaksime oma hinnangute täpsust kirjeldada peaksime esmalt mõistma seda, kuidas võivad juhuslike suuruste N_i väärtused kõikuda – milline on nende jaotus (üle erinevate valimite).

Jaotusmudelid sageduste jaoks

• Binoomjaotus

$$N_1 \sim B(n; \pi_1)$$

Kui operatsiooni õnnestumise tegelik tõenäosus oleks 0,65, siis nähtud andmestiku (sagedustabeli) nägemise tõenäosus on

$$\begin{aligned} P(N_1 = 7) \cdot P(N_2 = 3 | N_1 = 7) &= \\ &= C_{10}^7 0,65^7 (1 - 0,65)^3 \cdot 1 \\ &= 0,25 \dots \end{aligned}$$

Sellised on meie andmed

Võimalikud alternatiivsed andmestikud mida oleksime võinud näha.

Operatsiooni tulemus:	
õnnestus	patsient suri
7	3
6	4
8	2

Vaatame fikseeritud suurusega valimeid (teeme alati täpselt n vaatlust). Valimi suurus ehk vaatluste arv n pole juhuslik, on eelnevalt paika pandud. Siin näites: Arst teeb igal nädal 10 teatud tüüpi südameoperatsiooni – ja nende 10 operatsiooni tulemused ongi meie andmed.

Jaotusmudelid sageduste jaoks

• Multinomiaalne jaotus

$$N_1, \dots, N_k \sim \text{Mult}(n; \pi_1, \dots, \pi_k)$$

Operatsiooni tulemus:

õnnestus	nii-ja-naa	patsient suri
3	4	3
4	6	0
4	5	1

Paneme kirja informatsiooni kõigi operatsioonide kohta, mis nädala aja jooksul toimusid – summa fikseeritud.

• Multinomiaalne jaotus

$$N_1, \dots, N_k \sim \text{Mult}(n; \pi_1, \dots, \pi_k)$$

$$\begin{aligned} P(N_1 = x_1, \dots, N_k = x_k) &= \\ &= \frac{n!}{x_1! \cdot \dots \cdot x_k!} \cdot \pi_1^{x_1} \cdot \dots \cdot \pi_k^{x_k} \end{aligned}$$

Kui

$$N_1, \dots, N_k \sim \text{Mult}(n; \pi_1, \dots, \pi_k)$$

siis

$$N_1 | (N_1 + \dots + N_k) = n \sim B(n, \pi_1)$$

Operatsiooni tulemus:

õnnestus	nii-ja-naa	patsient suri
3	4	3

$$P(N_1 = 3, N_2 = 4, N_3 = 3)$$

$$\begin{aligned} &1 \ 1 \ 1 \ 2 \ 2 \ 2 \ 2 \ 3 \ 3 \ 3 \quad \pi_1^3 \pi_2^4 \pi_3^3 \\ &1 \ 1 \ 2 \ 1 \ 2 \ 2 \ 2 \ 3 \ 3 \ 3 \quad \pi_1^3 \pi_2^4 \pi_3^3 \\ &\dots \quad \dots \quad \dots \quad \dots \\ &3 \ 3 \ 3 \ 2 \ 2 \ 2 \ 2 \ 1 \ 1 \ 1 \quad \pi_1^3 \pi_2^4 \pi_3^3 \end{aligned}$$

$$\frac{10!}{3! \ 4! \ 3!} \cdot \pi_1^3 \pi_2^4 \pi_3^3$$

Jaotusmudelid sageduste jaoks

• Poissoni jaotus

$$N_i \sim \text{Poi}(\lambda_i)$$

$$P(N_i = k) = \frac{\lambda_i^k e^{-\lambda_i}}{k!}$$

Kui keskmiselt sünniks aastas 564 poissi ja 537 tüdrukut, siis vaadeldud andmete nägemise tõenäosus on

$$\begin{aligned} P(N_1 = 493) \cdot P(N_2 = 493) &= \\ &= \frac{564^{493} e^{-564}}{493!} \cdot \frac{537^{493} e^{-537}}{493!} \\ &= 4,7 \cdot 10^{-7}. \end{aligned}$$

Sellised on meie andmed

Võimalikud alternatiivsed andmestikud mida oleksime ka võinud kohata tänu juhusele

Tartus aasta (2021) jooksul sündinud lapsed:

poisse	tüdrukuid
493	493

poisse	tüdrukuid
485	509

poisse	tüdrukuid
565	609

Miks pole poiste arv binoomjaotusega?

Aasta jooksul sündivate laste arv pole eelnevalt fikseeritud – ühel konkreetsel aastal võib Tartus sündida kas natuke rohkem või natuke vähem lapsi...

Jaotusmudelid sageduste jaoks

• Poissoni jaotus

$$N_i \sim \text{Poi}(\lambda_i)$$

$$P(N_i = k) = \frac{\lambda_i^k e^{-\lambda_i}}{k!}$$

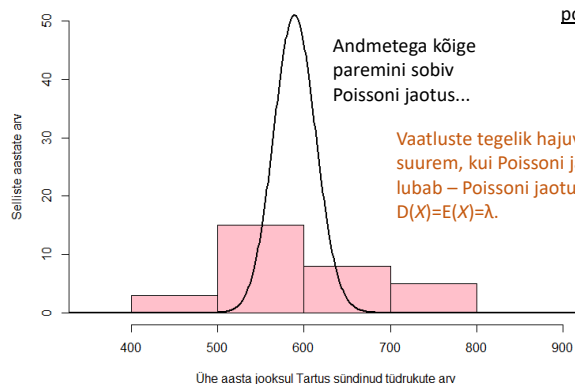
Poissoni jaotuse omadused:

Kui $N_1 \sim \text{Poi}(\lambda_1)$ ja $N_2 \sim \text{Poi}(\lambda_2)$ on sõltumatud:

- $N_1 + N_2 \sim \text{Poi}(\lambda_1 + \lambda_2)$
- $N_1 | N_1 + N_2 \sim B(N_1 + N_2; \lambda_1 / (\lambda_1 + \lambda_2))$
- $E(N_1) = D(N_1) = \lambda_1$
- Kui $\lambda_1 \rightarrow \infty$, siis ligikaudu $N_1 \sim N(\lambda_1; \sigma^2 = \lambda_1)$

Jaotusmudelid sageduste jaoks

Probleemid Poissoni jaotusega....



Tartus aasta (2021) jooksul sündinud lapsed:

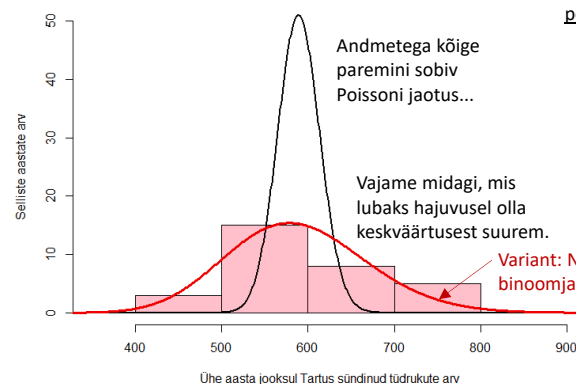
poisse	tüdrukuid
493	493

Varasematel aastatel Tartus sündinud tüdrukute arvud

509
609
531
550
533
537
582
564
614
.....

Jaotusmudelid sageduste jaoks

Probleemid Poissoni jaotusega....



Tartus aasta (2021) jooksul sündinud lapsed:

poisse	tüdrukuid
493	493

Varasematel aastatel Tartus sündinud tüdrukute arvud

509
609
531
550
533
537
582
564
614
.....

Jaotusmudelid sageduste jaoks

Negatiivne binoomjaotus

$$X \sim \text{NB}(r, p)$$

ebaõnnestumiste arv mida peame kogema, enne kui saavutame r „õnnestumist“. Ühe katse õnnestumise tõenäosus on p

Kui $X \sim \text{NB}(r=1,75; p=0.7)$, siis kuidas mõista parameetri $r=1,75$ väärtust?

$$X \sim \text{NB}(\mu; \theta)$$

$\lambda \sim \Gamma$ ← Esmalt tõmbame juhusliku λ gamma jaotusest – kas võiks olla hea „tõdrukaasta“ või mitte, $E(\lambda) = \mu$.

$X|\lambda \sim \text{Pois}(\lambda)$ ← Kasutades saadud λ väärtust genereerime Poissoni jaotusest sel aastal sündivate tõdruka arvu

$$X \sim \text{NB}(\dots)$$

Lisaks eksisteerib veel viis sagedamini kasutatavat parametrisatsiooni

`dnbinom(2, size=1.75, prob=0.7)`

`dnbinom(2, mu=0.75, size=1.75)`

kus $\mu = r \cdot (1-p)/p$

$$D(X) = \mu + \mu^2/\theta$$

$$E(\lambda) \quad D(\lambda)$$

Kui $\theta \rightarrow \infty$ siis muutub negatiivne binoomjaotus Poissoni jaotuseks

Töökorraldus

kodulehekül: <https://www-1.ms.ut.ee/mart/kval2026/>

Viimane loeng: **6. aprill**

Tarvis esitada ja kaitsta **projekt** (esitamise tähtaeg **19. aprill**; kaitsmised **20. aprillil** kell 10:15; 12:15; 14:15; 16:15 ruumis 2030 ja **21. aprillil** kell 10.15; 12:15 ruumis 1026)

Eksam

27. aprill kell 14:15-16:00

<- võimalusel eelista

1. juuni kell 14:15-16:00

Taustatõestused I

Loengus lahmitud väidete tõestused

Kui $N_1 \sim \text{Poi}(\lambda_1)$ ja

$N_2 \sim \text{Poi}(\lambda_2)$ on

sõltumatud, siis

$$N_1 + N_2 \sim \text{Poi}(\lambda_1 + \lambda_2)$$

Kui ühes kuus sündinud poiste arv on Poissoni jaotusega siis on ka aasta jooksul sündinud poisilaste arv Poissoni jaotusega juhuslik suurus...

$$\begin{aligned} P(N_1 + N_2 = k) &= \sum_{i=0}^k P(N_1 + N_2 = k | N_1 = i) P(N_1 = i) \\ &= \sum_{i=0}^k P(N_2 = k - i | N_1 = i) \frac{\lambda_1^i e^{-\lambda_1}}{i!} \\ &= \sum_{i=0}^k \frac{\lambda_2^{k-i} e^{-\lambda_2}}{(k-i)!} \cdot \frac{\lambda_1^i e^{-\lambda_1}}{i!} \\ &= \frac{e^{-(\lambda_1 + \lambda_2)}}{k!} \sum_{i=0}^k \frac{k!}{(k-i)! i!} \lambda_1^i \lambda_2^{k-i} \\ &= \frac{e^{-(\lambda_1 + \lambda_2)}}{k!} \sum_{i=0}^k C_k^i \lambda_1^i \lambda_2^{k-i} \quad (a+b)^k = \dots \\ &= \frac{e^{-(\lambda_1 + \lambda_2)}}{k!} (\lambda_1 + \lambda_2)^k \quad \leftarrow \text{Poissoni jaotus parameetriga } \lambda_1 + \lambda_2 \end{aligned}$$

Taustatõestused II

Loengus lahmitud väidete tõestused

Kui $N_1 \sim \text{Poi}(\lambda_1)$ ja

$N_2 \sim \text{Poi}(\lambda_2)$ on

sõltumatud, siis

$$N_1 | (N_1 + N_2) = n \sim B(n; \lambda_1 / (\lambda_1 + \lambda_2))$$

Kui keskendume vaid neile aastatele, mil sündis täpselt n last, siis on poiste arv nendel aastatel binoomjaotusega juhuslik suurus.

$$\begin{aligned} P(N_1 = k | N_1 + N_2 = n) &= \frac{P(N_1 = k) P(N_1 + N_2 = n | N_1 = k)}{P(N_1 + N_2 = n)} \\ &= \frac{P(N_1 = k) P(N_2 = n - k)}{(\lambda_1 + \lambda_2)^n e^{-(\lambda_1 + \lambda_2)} / n!} \quad \leftarrow \text{Poissoni jaotusege j.s. summa on Poissoni jaotusega} \\ &= \frac{\lambda_1^k e^{-\lambda_1} / k! \cdot \lambda_2^{n-k} e^{-\lambda_2} / (n-k)!}{(\lambda_1 + \lambda_2)^n e^{-(\lambda_1 + \lambda_2)} / n!} \\ &= \frac{\lambda_1^k / k! \cdot \lambda_2^{n-k} / (n-k)!}{(\lambda_1 + \lambda_2)^n / n!} \\ &= \frac{n!}{k! (n-k)!} \cdot \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^k \cdot \left(\frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^{n-k} \\ &= \frac{n!}{k! (n-k)!} \cdot \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^k \cdot \left(1 - \frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^{n-k} \\ &\quad \leftarrow \text{Binoomjaotus } B(n; \lambda_1 / (\lambda_1 + \lambda_2)) \end{aligned}$$