Sissejuhatus matemaatilise statistika erialasse

Tõenäosusteooria kui statistika alus

Krista Fischer

Tartu, oktoober 2019

Sissejuhatuseks: Monty Halli probleem

- Mäng (nt teleshow):

Mängijal on võimalus võita auhind, milleks on kas kits või (uhke ja kallis) auto. Tal palutakse valida kolme ukse vahel, millest kahe taga on kits ja ühe taga auto. Mängija teeb valiku.

Kas soovid oma algset valikut muuta?

Kas soovid oma algset valikut muuta?

Sinu lõplik valik?

Kas algse valiku muutmine pärast ühe ukse avamist:

- Suurendab auto võitmise tõenäosust?
- Vähendab auto võitmise tõenäosust?
- Ei muuda seda tõenäosust?

https://www.mathwarehouse.com/monty-hall-simulation-online/

Õige vastus/lahendus?

- Alguses valivad 1/3 inimestest autoga ukse ja 2/3 kitsega ukse.

- Kui inimene oma valikut ei muudaks, siis on auto võitmise tõenäosus 1/3

- Kui aga kõik muudaksid oma valikut pärast ühe kitsega ukse avamist, siis valiksid kõik algselt kitsega ukse valinud inimesed (ehk 2/3) endale auto (ja algselt auto valinud inimesed, ehk 1/3, saaksid kitse)! Seega, auto saamise tõenäosus oleks 2/3!

Loo moraal: tõenäosuste puhul võib algne intuitsioon vahel petta – tõenäosusteooria reeglite teadmine aitab õigete lahendusteni jõuda.

Tõenäosus

Tõenäosus iseloomustab sündmuse toimumise võimalikkust skaalal 0-st 1-ni.

Tõenäosus 0: võimatu sündmus, tõenäosus 1: kindel sündmus..

Tihti saab tõenäosust tõlgendada kui mingi sündmuse või nähtuse suhtelist sagedust (osakaalu või protsenti).

Matemaatiliselt saab tõenäosust vaadelda kui mõõtu, mis igale sündmusele teatud sündmuste ruumist seab vastavusse arvu 0-st 1-ni. Tõenäosuse korrektseks defineerimiseks kasutatakse nn Kolmogorovi aksioome

Tõenäosus – tähistused ja põhimõtted

Et rääkida tõenäosustest, on vaja selgelt määratleda nn sündmuste ruum – millised sündmused saavad toimuda.

Sündmuse A tõenäosust tähistatakse P(A)

Kui A on kindel sündmus („hommikul tõuseb päike“), siis P(A)=1

Võimatu sündmuse korral P(A)=0

Kui sündmuse A tõenäosus $P(A) = p$, siis vastandsündmuse tõenäosus

$$P(\bar{A}) = 1 - p$$

Kuidas me teame tõenäosuseid?

- Kindlalt määratletud katsed/olukorrad – nt täringu- või kaardimängud, mündivise, loterii, jne.
  - Nt. kui A: „täringuviske tulemus on kuus silma“, siis P(A)=1/6

  (6 võrdset võimalust, neist ühel juhul toimub huvipakkuv sündmus)
  - Sarnaseid näiteid leiab ka geneetikas!
- Andmete põhjal suhtelise sagedusena hinnatud tõenäosused
  - A: juhuslikult valitud isik on vasakukäeline
  - P(A) = (kõigi vasakukäeliste arv rahvastikurühmas)/(rahvastikurühma suurus)
  - Tihti pole meil andmeid kogu rahvastikurühma kohta ja hinnang saadakse valimi põhjal
- Subjektiivselt hinnatud tõenäosused (eksperthinnangud)

- **Klassikaline tõenäosus**. Kui juhuslikul katsel on üldse N võrdvõimalikku tulemust ja nende hulgas sündmuse A toimumiseks soodsaid tulemusi on N_A, siis sündmuse A tõenäosus avaldub suhtena

$$P(A) = \frac{N_A}{N}$$

- Arvude N ja N_A leidmine võib olla küllalt keeruline (vahel vaja kombinatoorika valemeid)

Operatsioonid tõenäosustega

Kahe sündmuse, A ja B, korral kehtivad reeglid:

- Liitmisreegel. Kui kaks sündmust on üksteist välistavad (ei saa samaaegselt toimuda), siis tõenäosus, et vähemalt üks neist kahest toimub, on leitav kui nende sündmuste tõenäosuste summa:

$$P(A \cup B) = P(A) + P(B), \quad \text{juhul kui } P(A \cap B) = 0$$

P(A või B)

Tõenäosus, et täringuviske tulemuseks on kas 1 või 2, on

$$\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

Operatsioonid tõenäosustega

- Korrutamisreegel. Kui kaks sündmust on sõltumatud (ühe sündmuse toimumine või mittetoimumine ei mõjuta teist sündmust), siis tõenäosus, et nad samaaegselt toimuvad on leitav kui vastavate sündmuste tõenäosuste korrutis:

$$P(A \cap B) = P(A) \cdot P(B), \text{ kui } A \perp B$$

P(A ja B)

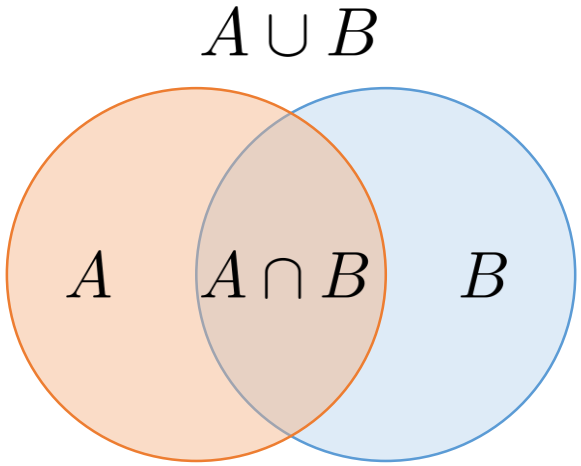
Tõenäosus, et kahe täringu viskamisel tuleb kaks „kuut“, on $\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} \approx 0,028$

Tõenäosus, et kaks juhuslikult valitud isikut on mõlemad sündinud oktoobris, on $\frac{1}{12} \cdot \frac{1}{12} = \frac{1}{144} \approx 0,0069$

Operatsioonid tõenäosustega

- Üldiselt aga:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



Tõenäosus, et kahe täringu viskamisel saadakse vähemalt üks „kuus“?

A: esimese täringu viskamisel tuleb „6“
B: teise täringu viskamisel tuleb „6“

$$P(A \cap B) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$$

Mõlemad „kuued“

$$P(A \cup B) = \frac{1}{6} + \frac{1}{6} - \frac{1}{36} = \frac{11}{36}$$

Vähemalt üks „kuus“

Näide geneetikast (kuigi sama probleem esineb ka mujal)

Geneetilist laktoositalumatust esineb 28%-l inimestest. Kui suur on tõenäosus, et 5-inimeselises grupis vähemalt ühel on laktoositalumatus?

- Idee: arvutame tõenäosuse, et mitte ühelgi pole laktoositalumatust

Tõenäosus, et inimesel ei ole laktoositalumatust, on 1-0.28=0.72

Korrutamisreegel (eeldame et inimesed on sõltumatud):

Tõenäosus, et 5-st inimesel mitte ühelgi pole, on

$0.72^5 \approx 0.19$

Seega tõenäosus, et vähemalt ühel on laktoositalumatus, on ligikaudu 1-0.19=0.81, ehk 81%

Kuidas lahendada sünnipäevaprobleemi?

Milline on tõenäosus, et kahel inimesel 30-st on samal kuupäeval sünnipäev?

- Teeme eelduse, et sünnid jagunevad ühtlaselt kõigi kuupäevade vahel ja et aastas on 365 päeva
- Arvutame vastandsündmuse tõenäosuse – tõenäosuse, et kõik on sündinud erinevatel kuupäevadel

Esimese isiku sünnikuupäev võib olla suvaline

Teisel võib olla üks 364 ülejäänud päevast – tõenäosus et tal on esimesest erinev sünnikuupäev, on 364/365

Kolmandal üks 363 ülejäänud päevast – tõenäosus 363/365

Jne

Korrutamisreegel

$p = (364/365)*(363/365)*\ldots*(336/365)$ - 29 tegurit

Arvutame R-i abil!

Tõenäosus, et kõigil on erinevad sünnikuupäevad:

```
prod(336:364)/(365^29)   # prod(x) - x elementide korrutis
[1] 0.2936838
```

Vastandsündmus – vähemalt 2-l samal päeval sünnipäev:

```
 1- prod(336:364)/(365^29)
[1] 0.7063162
```

Ehk siis saame selleks tõenäosuseks ligikaudu 71%

Äkki oli mõttekäigus viga?

Kontrollime simulatsiooni abil

```
k = 0                    # loendur k=0
for (i in 1:10000)         # 10000 kordust
{
x = sample(1:365,size=30, replace=T)  # valime 365 kuupäevast 30
k = k +  (sum(duplicated(x))>0)   # kui esineb duplikaate, siis k=k+1
}

k/10000

[1] 0.7121  # juhuslik varieeruvus on ootuspärane, ligikaudu siiski õige
```

sample(x) võtab juhuvalimi vektorist x,
duplicated(x) tekitab TRUE/FALSE väärtustega vektori, vastavalt sellele kas mõningaid x elemente esineb mitmekordselt – summeerimisel on TRUE väärtus 1 ja FALSE 0.

Tinglik tõenäosus

(Conditional probability)

P(B|A) : sündmuse B tõenäosus, juhul kui on teada et sündmus A on toimunud

(mõnikord vaadeldav kui tõenäosus mingis alamhulgas)

Näide: tõenäosus, et inimesel esineb värvipimedus on 0.05 juhul kui on tegemist mehega, ja 0.0025, juhul kui tegu on naisega.

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

Tinglik tõenäosus – veel näiteid

Väga tihti seotud tunnustevahelise seosega – milline on tõenäosus, et inimene kaalub üle 100kg, kui on teada et ta on 185cm pikk meesterahvas (kindlasti erinev sellest, mis oleks see tõenäosus 165cm pika naisterahva korral)?

Diagnostiliste testide omadused

- Oletame et on üks test T, mille eesmärgiks on teatud haiguse või seisundi D olemasolu testimine. Tulemused saab esitada järgmises tabelis:

Test / Haigus	Positiivne: T+	Negatiivne: T-
Esineb, D+	Õige positiivne testitulemus	Valenegatiivne testitulemus
Ei esine, D-	Valepositiivne testitulemus	Õige negatiivne testitulemus

Kui hea on test?

- Ideaalne test ei anna valepositiivseid ega valenegatiivseid tulemusi
- Ideaalseid teste esineb üliharva
- Testi omaduste iseloomustamiseks kasutatakse tundlikkuse ja spetsiifilisuse mõisteid

Näide (diagnostiliste testide omadused)

Et iseloomustada konkreetse testi täpsust, kasutatakse järgmisi omadusi:

- Testi tundlikkus (ingl. *sensitivity*) on tinglik tõenäosus, et test annab uuritava seisundi olemasolul positiivse tulemuse:

  $P(T+|S+)$

- Testi spetsiifilisus (ingl. *specificity*) on tinglik tõenäosus, et test annab uuritava seisundi puudumisel negatiivse tulemuse:

  $P(T-|S-)$

Seega saame ka testi eksimise tõenäosused:

Valenegatiivse testitulemuse tõenäosus = 1 – tundlikkus

$P(T-|S+) = 1 - P(T+|S+)$ (miks?)

Valepositiivse testitulemuse tõenäosus = 1 – spetsiifilisus

$P(T+|S-) = 1 - P(T-|S-)$

Testi tundlikkust ja spetsiifilisust saab hinnata katseliselt, kasutades uuritavaid kellel seisundi olemasolu või puudumine on teada

Olgu meil M uuritava seisundiga katsealust ja N uuritava seisundita katsealust (seisundi olemasolu või puudumine määratud nn kindla meetodiga).
Testides neid katsetatava testiga, saame tulemused esitada järgmise tabelina.

Test / Seisund	Positiivne, T+	Negatiivne, T-	Kokku	Testi omadus
Jah, S+	m	M - m	M	Tundlikkus = m/M
Ei, S-	n	N - n	N	Spetsiifilisus = (N-n)/N

Näide

Testi omaduste hindamine juhtkontrolluuringu (case-control study) tingimustes

Biomarkeri test / Haigus	Positiivne	Negatiivne	Kokku
Jah	190	10	200
Ei	20	180	200
kokku	210	190	400

Uuringusse on valitud 200 isikut, kellel esineb uuritav haigus ja 200 nn kontrollisikut, kellel seda haigust kindlasti ei ole

Tundlikkus: 190/200=0,95 ehk 95%

Spetsiifilisus: 180/200=0,90 ehk 90%

Testi prognoosiväärtus

... Võib olla praktikas olulisem kui tundlikkus ja spetsiifilisus

Positiivne prognoosiväärtus *(positive predictive value)*, PPV:

P(S+|T+) – uuritava seisundi tõenäosus positiivse testitulemuse korral (sarnaselt on P(S-|T-) negatiivseks prognoosiväärtuseks ehk NPV)

Kuidas leida P(S+|T+), kui on teada P(T+|S+) ja P(T-|S-)?

(miks seda ei saa teha nt eelnevas näites toodud juhtkontrolluuringu põhjal?)

Näide

A: uuringus 200 haiget ja 200 kontrolli („tervet“)

Test / Haigus	Positiivne	Nega-tiivne	Kokku
Jah	190	10	200
Ei	20	180	200
kokku	210	190	400

Tundlikkus: 190/200=0,95 ehk 95%
Spetsiifilisus: 180/200=0,90 ehk 90%
PPV = 190/210 = 0.905 ehk 90.5%
NPV = 180/190= 0.947 ehk 94.7%

B: uuringus 200 haiget ja 2000 kontrolli

Test / Haigus	Positiivne	Nega-tiivne	Kokku
Jah	190	10	200
Ei	200	1800	2000
kokku	390	1810	2200

Tundlikkus: 190/200=0,95 ehk 95%
Spetsiifilisus: 1800/2000=0,90 ehk 90%
PPV = 190/390 = 0.487 ehk 48.7%
NPV = 1800/1810= 0.994 ehk 99.4%

Muutes kontrollide arvu, jäävad tundlikkuse ja spetsiifilisuse hinnangud samaks (reaalsuses erineksid ehk veidi juhusliku varieeruvuse tõttu), kuid positiivse ja negatiivs prognoosiväärtuse hinnangud muutuvad olulisel määral

Veel üks näide....

Oletame, et 10000 sportlast testiti dopingu suhtes testiga, mille tundlikkus on 10% ja spetsiifilisus 99.9%. Kui umbes 2% sportlastest kasutaks dopingut, siis võiks saada sellised tulemused:

Test / Doping	Positiivne	Negatiivne	Kokku
Jah	20	180	200
Ei	10	9790	9800
kokku	**30**	**9970**	**10000**

PPV = 0.667 NPV= 0.982

Seega 1/3 positiivse testitulemuse saanutest tegelikult dopingut ei tarbinud, kusjuures 1.8% negatiivse tulemuse saanutest kasutas dopingut

Testi prognoosiväärtus

... Võib olla praktikas olulisem kui tundlikkus ja spetsiifilisus

Positiivne prognoosiväärtus *(positive predictive value)*, PPV:

P(S+|T+) – uuritava seisundi tõenäosus positiivse testitulemuse korral (sarnaselt on P(S-|T-) negatiivseks prognoosiväärtuseks ehk NPV)

Kuidas leida P(S+|T+), kui on teada P(T+|S+) ja P(T-|S-)?

Ehk siis kas ja kuidas saaks tinglikku tõenäosust „tagurpidi pöörata“?

See on võimalik, kui lisaks on teada veel seisundi levimus, ehk P(S+)

Muutes eelnevas näites kontrollide arvu, muutus ka seisundi levimus selles uuringus – tegelikkuses me sooviks teada PPV ja NPV väärtuseid mitte uuringus, vaid rahvastikus, kus seda testi kasutada soovitakse.

Bayesi valem (lihtsal kujul)

Thomas Bayes
(1701-1761)

Tõenäosuste korrutamise lausest:

$$P(A \cap B) = P(B|A)P(A) = P(A|B)P(B)$$

saame, et:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

See ongi Bayesi valem (tegelikult selle lihtsustatud kuju).

Edasi saame veel arvestada, et sündmus A saab toimuda kahel teineteist välistaval juhul: kas samaaegselt sündmusega B või nii, et B ei toimu, ehk siis samaaegselt B vastandsündmusega $\bar{B}$.

Ehk siis $P(A) = P(A \cap B) + P(A \cap \bar{B}) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B})$

Testi prognoosiväärtus

Positiivne prognoosiväärtus *(positive predictive value)*, PPV:

P(S+|T+) – uuritava seisundi tõenäosus positiivse testitulemuse korral (sarnaselt on P(S-|T-) negatiivseks prognoosiväärtuseks ehk NPV)

Positiivse prognoosiväärtuse saab leida Bayesi valemi abil:

$$\mathrm{P}(\mathrm{S}^+|\mathrm{T}^+) = \frac{\mathrm{P}(\mathrm{T}^+|\mathrm{S}^+)\mathrm{P}(\mathrm{S}^+)}{\mathrm{P}(\mathrm{T}^+)} = \frac{\mathrm{P}(\mathrm{T}^+|\mathrm{S}^+)\mathrm{P}(\mathrm{S}^+)}{\mathrm{P}(\mathrm{T}^+|\mathrm{S}^+)\mathrm{P}(\mathrm{S}^+) + \mathrm{P}(\mathrm{T}^+|\mathrm{S}^-)\mathrm{P}(\mathrm{S}^-)}$$

$$= \frac{\mathrm{P}(\mathrm{T}^+|\mathrm{S}^+)\mathrm{P}(\mathrm{S}^+)}{\mathrm{P}(\mathrm{T}^+|\mathrm{S}^+)\mathrm{P}(\mathrm{S}^+) + [1 - \mathrm{P}(\mathrm{T}^-|\mathrm{S}^-)] \cdot [1 - \mathrm{P}(\mathrm{S}^+)]}$$

$$= \frac{\text{tundlikkus} \cdot \text{levimus}}{\text{tundlikkus} \cdot \text{levimus} + (\text{1-spetsiifilisus}) \cdot (\text{1-levimus})}$$

Kokkuvõte

- Tõenäosusteooria on statistika ja ka moodsa andmeteaduse aluseks
- Praktikas uuritakse tihti tunnustevahelisi seoseid ja seejuures on oluline tunda tingliku tõenäosuse mõistet, samuti saada aru sündmuste (tunnuste) sõltumatusest
- Bayesi valemile tugineb terve statistika koolkond – Bayesi statistika. See on siiski kasulik valem ka nende jaoks, kes eelistavad jääda nn klassikalise frekventistliku statistika pärusmaale.

Veel üks ülesanne

Juku läheb pimedasse tuppa sahtlist sokke otsima. Tal on sahtlis 6 sokki, mis on kas mustad või valged. Tõenäosus, et juhuslikult valides saab ta kaks valget sokki, on 2/3.

Mis on tõenäosus, et ta saab kaks musta sokki?