Teoreetilise Bioloogia Kevadkool, 2021

Mis asi on tõenäosus ja miks me sellest hoolime

Ülo Maiväli
ymaivali@ut.ee
Tartu Ülikooli Tehnoloogia Instituut

Matemaatika trügib bioloogiasse põhiliselt kahel viisil: deterministlike mudelite näol, näiteks Lotka-Volterra kiskja-saaklooma võrrandid, ja stohhastiliste mudelitena, sageli regressiooni/ANOVA mudelite kujul, milliste väljundiks on alati tõenäosused. Nood stohhastilised mudelid on bioloogias sedavõrd laialt levinud, et ilma nendeta ei ilmu tänapäeval enam praktiliselt ühtki teadustööd. See on tekitanud olukorra, kus ka bioloogil ei jää muud üle, kui õppida rakendama statistilisi mudeleid oma mitte-statistilise teaduse vankri ette. Ja see on raskem kui võiks nende meetodite populaarsust arvestades arvata (Päll et al. 2021).

Töötaval teadlasel on kaks põhilist vastutust. Esiteks, isegi olukorras kus ta kasutab professionaali abi oma mudelite formuleerimisel ja jooksutamisel, vastutab just tema selle eest, et on valinud mõistlikud mudelid, mille eeldustest ja piirangutest ta aru saab. Sellega me käesolevas artiklis ei tegele. Ja teiseks on tema kohustus tõlgendada mudeldamise tulemusi oma teaduse kontekstis. Stohhastiline mudeldamine on keeruline asi ja oleks kummaline (või küüniline) arvata, et teadlane teeks seda vabatahtlikult mingil muul eesmärgil kui oma teaduse hüvanguks. See, kuidas täpselt mudeldamine kinnitab või vähendab teie teaduslike järelduste usaldusväärsust, peab avaldatud publikatsioonidest selgelt ja ühemõtteliselt välja tulema. Vastasel korral valitseb oht, et formaalne mudeldamine täidab teie teaduses maagilist rolli või, mis veel hullem, on pelgalt veel üks viis teie teaduse lehmakoogistamiseks.

Paraku ei ole ühte kõigi poolt aktsepteeritud viisi, kuidas mudeldamise tulemusi teaduses kasutada (aga vaata Prommik ja Maiväli 2021). See sõltub nii sellest, milline on mudeli väljund (p väärtus, usaldusintervall või järeljaotus), kui sellest, kuivõrd on mudeli struktuur inspireeritud teadusliku teooria struktuurist, kui ka sellest, kui relevantne on mudeldatav

nähtus ise teie teadusliku probleemi seisukohast. Nagu ka mudeldamise eesmärgist: olgu selleks parameetriväärtustele (nagu ravimi mõju suurus) ebakindluse määra omistamine, põhjuslike hüpoteeside kinnitamine/falsifitseerimine, uute hüpoteeside loomine või hoopis *in silico* andmepunktide ennustamine hüpoteetilistes olukordades, mida me ei saa või ei taha looduses mõõta. Siiski on kõigile neile võimalustele ühine, et tõenäosuse mõistest aru saamata (või vähemalt selle osas arvamust omamata) pole suurt lootust neist ühtki mõistlikul viisil kasutada.

Järgnevalt vaatame otsa tõenäosusmudelite väljundile, milleks on tõenäosused: mis nad on, mida nad tähendavad ja milline on nende vormirikkus bioloogilises andmeanalüüsis. Siinkohal peab lugejat hoiatama, et kuigi matemaatilised tehted tõenäosustega on väga hästi mõistetud, on tõenäosuste päris-maailmaga haakuvad tõlgendused olnud matemaatikute, teadlaste ja filosoofide diskussiooniobjektiks juba alates Leibnizist kuni Keynesi, Ramsey, de Finetti, von Misese, Popperi, Fisheri, Savage, Jeffreysi ja Jaynesini välja (Hacking, 1975, 1990, Maiväli, 2015). Seega püüame ka meie oma väiteid illustreerida näidetega ja jätta lugejale mõtlemis- ja vaidlemisruumi.
Tõenäosusel on kaks poolt, mis on üsna erinevad, aga mõlemad teadlasest kasutajale võrdselt vajalik. Tõenäosus on aksiomaatiline matemaatiline teooria, aga samal ajal ka teaduslik/statistiline/filosoofiline mõiste. Esimene on vajalik tõenäosuste arvutamiseks ja teine on vajalik nende kasutamiseks teadusliku argumendi osana. Samas on tutvus mõlemaga vajalik, kui tahta mõista teaduslike ja statistiliste meetodite olemust ja omavahelisi keerukaid suhteid.

Tõenäosusteooria kui matemaatika haru

Tõenäosusteooria on matemaatika haru, mis põhineb kolmel aksioomil ja ühel definitsioonil, mis kehtestati Andrei Kolmogororovi poolt 1933 aastal. Nagu iga aksiomaatikat, võib ka tõenäosusteooriat vaadelda mõttemänguna, millel võib, aga ei pruugi olla, rakendusi maailmas. Siiski leidub asjaolusid, mis räägivad tõenäosusteooria erilise tähtsuse kasuks nii matemaatikale kui teaduslikule meetodile.
Kõigepealt, Kolmogorov tuletas oma aksioomid hulgateooriast (sellele pani aluse Georg Cantor 1874. aastal), mida on omakorda vaadeldud kui kogu matemaatika alusteooriat, aga

ka kui lõpmatuse matemaatilise uurimise alust. Sellest sõltumatult tuletas Richard Threlkeld Cox 1946 aastal samad aksioomid järgmistest postulaatidest, mis võiksid vastata ratsionaalse mõtlemise miinimumnõuetele:

1. Propositsiooni usutavus (*plausibility*) on pidev suurus, mis sõltub informatsioonist, mida meil selle propositsiooni kohta on. (Iga lisanduv infokild, ükskõik kui väike, suudab seda usutavust muuta, ükskõik kui vähe.)
2. Propositsioonide usutavused sõltuvad tavamõistuslikul viisil usutavuse hinnangutest mudelis. (Kui me saame propositsiooni kehtimise kohta lisainfot, siis selle propositsiooni usaldusväärsus kasvab, mitte ei kahane).
3. Kui propositsiooni usutavust on võimalik arvutada mitmel erineval viisil, siis peavad need kõik viima välja samale arvulisele väärtusele (konsistentsuse eeldus). Sellele lisandub totaalse informatsiooni printsiip, mis ütleb, et propositsiooni usaldusväärsuse hindamisel peame arvesse võtma kogu relevantse informatsiooni.

Nendest postulaatidest (küll mõnevõrra formaalsemal kujul) tuletas Cox sellesama tõenäosusteooria. See on tähtis, sest Coxi postulaadid seovad tõenäosused klassikalise lausearvutusliku loogikaga, mis tähendab, et tõenäosusteooria ei ole mitte ainult hulgateooria pikendus, vaid ka loogika formaalne edasiarendus juhtudele, kus informatsioon on piiratud või ebakindel (Jaynes 2003). Seega saame tõenäosusteooriat näha tõenäosusloogikana, mida omakorda võib vaadata kui ratsionaalse mõtlemise mudelit ebakindluse tingimuses, mis hõlmab endas olulist osa inimlikust ratsionaalsusest – või äärmuslikult, isegi kogu normatiivset ratsionaalset mõtlemist oma täies ilus. Igal juhul tähendab see, et tõenäosusteooria on intiimses seoses teadusliku mõtlemisega, ehkki mitte tingimata tavamõtlemisega (Tversky ja Kahneman. 1974).

Tõenäosusteooria seosele ratsionaalsusega viitab ka Frank Ramsey avastus (1926), et tõenäosusteooria aksioomidega vastuolulised kihlveod viivad kindla kaotuseni (Vineberg, 2016). Kehtib ka vastand-teoreem, mille kohaselt senikaua kuni teie kihlveod on tõenäosusteooriaga kooskõlas on teil alati vähemalt teoreetiline võimalus kasumisse jääda. Lisaks normatiivsele rollile teaduslikus mõtlemises (või vähemalt selle analüüsil) on tõenäosusteoorial ka üha kasvav roll statistiliste mudelite käitamisel, mille väljundeid – tõenäosusi – omakorda kasutatakse teaduslikus argumenteerimises. Me räägime siin bayesiaanlikust statistikast, mis ei ole oma olemuselt midagi muud, kui tõenäosusteooria

praktiline rakendamine statistilise analüüsi teenistusse (Päll ja Maiväli 2019, Prommik ja Maiväli 2021).

Kolmogorovi aksioomid:

1. P(A) >= 0, ehk A tõenäosus ei saa olla negatiivne
2. P(Ω) = 1, ehk üksteist välistavate ja ammendavate sündmuste tõenäosuste summa on 1, ehk hüpoteesiruumi Ω tõenäosus on 1, ehk loogiliselt tõsikindla sündmuse tõenäosus on 1.
3. P(A või B) = P(A) + P(B), ehk kahe või enama üksteist välistava sündmuse korral võrdub neist ühe esinemise tõenäosus nende sündmuste tõenäosuste summaga.

Tingliku tõenäosuse definitsioon on P(A I B) = P(A & B)/P(B), kus esimene liige tähendab „A tõenäosus juhul kui B on tõene“ (see ei tähenda, et B peab päriselt kehtima – me küsime, milline oleks A tõenäosus siis, kui peaks juhtuma, et B kehtib). Footnote: See definitsioon tuleb otse hulgateooriast. Meil on n elemendist koosnev hulk S, ehk n(S), mis sisaldab kahte alamhulka A ja B, millel on ühisosa A & B. Siis on P(A I B) sama, mis ühisosa elementide arv suhtena B elementide arvule: P(A I B) = n(A & B)/n(B). Kui me jagame murrujoone pealmise ja alumise läbi n(S)-ga, siis konverteerime absoluutväärtused suheteks ehk tõenäosusteks ja saamegi oma definitsiooni. Definitsioonist on lihtne tuletada tõenäosusteooria usutavasti tähtsaim valem, Bayesi teoreem: P(A I B) = P(A)P(B I A)/P(B). Bayesi teoreemi tähtsus seisneb eelkõige selles, et see annab loogiliselt parima viisi seostamaks erinevat informatsiooni, mis võimaldab arvutada andmepõhise ratsionaalse usu määra ühe või teise hüpoteesi kehtimisse. See on aga just see, mille eest teadlased palka saavad.
Kuidas Bayesi valem töötab? Kõigepealt, A ja B ei ole muud kui tühjad kestad, millele võime anda mistahes sisu – senikaua kui nõnda tähistatud väärtused käituvad kooskõlaliselt Kolmogorovi aksioomidega on tegu tõenäosustega. Meie jaoks võivad A ja B tähendada näiteks sündmusi, andmeid, hüpoteese, parameetriväärtusi. Seega võime Bayesi teoreemi ümber kirjutada nii:

$$P(H_1 \mid a) = P(H_1)P(a \mid H_1)/P(a)$$

kus *a* tähistab andmeid ja H_1 tähistab meid huvitavat hüpoteesi. Vaatame igat selle valemi osa ükshaaval:

- $P(H_1 \mid a)$ on järeltõenäosus (*posterior*). See on meid huvitava hüpoteesi tõenäosus, mis arvestab nii andmetega (tõepära kaudu), kui selle hüpoteesi eeltõenäosusega, kui kogu hüpoteesiruumiga.
- $P(a \mid H_1)$ on tõepära (*likelihood*) ehk andmete tõenäosus H_1 kehtimise korral. Siinkohal on tõenäosuse mõiste kasutamine veidi eksitav, sest hüpoteesiruumi elementide tõepärad ei pea summeeruma ühele, mis tähendab, et tõepära ei ole normaliseeritud, ehk see ei ole päris tõenäosus (ehkki individuaalsed tõepärad jäävad alati 0 ja 1 vahele).
- $P(H_1)$ on meid huvitava hüpoteesi eeltõenäosus (*prior*), ehk selle hüpoteesi andmetest sõltumatu kehtimise tõenäosus. Siin ei ole ajalist konnotatsiooni: andmetes leiduv info tohib kajastuda ainult tõepäras ja kõik muu relevantne info ainult eeltõenäosuses.
- P(a) on normaliseerimiskonstant, mis tagab, et järeltõenäosused summeeruvad ühele ja selle arvutamine sõltub sellest, kuidas on Ω jagatud individuaalseteks hüpoteesideks. Kui Ω on jagatud kaheks, $\{H_1, H_2\}$, kus $P(H_1) + P(H_2) = 1$, siis $P(a) = P(a \,\&\, H_1) + P(a \,\&\, H_2) = P(H_1)P(a \mid H_1) + P(H_2)P(a \mid H_2)$. Kui jagame Ω nii, et sellel on 3 liiget, siis tuleb ka murrujoone alla 3 liiget, jne.

Bayesi teoreem on triviaalne tuletus tõenäosusteooria aksioomidest, milles pole midagi maagilist. See ei ole automaatne meetod, mis tagaks inimkonna teadmiste kasvu, vaid lihtsalt parim võimalik viis andmemudeli ja taustateadmiste mudeli ühendamiseks ja normaliseerimiseks tinglikuks tõenäosuseks (hüpoteesi tõenäosus meie andmete ja taustateadmiste korral). Edasi sõltub kõik mudelite, andmete ja taustateadmiste kvaliteedist. Päriselu küsimus on, kas see teoreetiliselt täiuslik meetod ei ole äkki praktikas halvem kui mõni teine lihtsalt hea meetod?

Näide: Bayesi teoreem läheb kohtusse.

Siin ei arvuta me mitte hüpoteesi tõenäosust vaid kahe hüpoteesi tõenäosuste suhet – ehk kui palju on tõenäolisem, et kohtualune on süüdi kui, et ta on süütu. Meil on kaks teineteist

välistavat ja ammendavat hüpoteesi: H_1 = süüdi ja H_2 = süütu, ning 2 ühikut tõendusmaterjali: DNA ja tunnistaja ütlus. Lisaks on süüdistataval nõrk motiiv ja nõrk alibi. Tõenäosuste suhte arvutamine on lihtne:

$$\frac{\mathrm{P(H1|tõendid)}}{\mathrm{P(H2|tõendid)}} = \frac{\mathrm{P(DNA|H1)}}{\mathrm{P(DNA\ I\ H2)}} \times \frac{\mathrm{P(ütlus|H1)}}{\mathrm{P(ütlus|H2)}} \times \frac{\mathrm{P(motiiv|H1)}}{\mathrm{P(motiiv|H2)}} \times \frac{\mathrm{P(alibi|H1)}}{\mathrm{P(alibi|H2)}}$$

Kus alibi ja motiivi tõepärasuhete korrutis annab eeltõenäosuste suhte. Iga juurde tulev tõend läheb sisse uue liikmena, ja olles arvesse võtnud kogu relevantse tõendusmaterjali saame teada, mitu korda on süüaluse süü tõenäolisem kui süütus. Sõltuvalt sellest, kas see number ületab 50, 100, või kasvõi 10000, võime võtta vastu otsuse inimene süüdi mõista. Huvitaval kombel on anglo-ameerika õigussüsteemis lubatud kohtule esitada tõepärasuhteid, näiteks DNA- tõendite puhul, aga mitte tõepärasuhteid omavahel läbi korrutada. Võib arvata, et peale sellist tehet kaoks vajadus vandemeeste järele ning kohtuotsus muutuks algoritmiliseks. Kohtuniku roll oleks siis otsustada, millist tõendusmaterjali milliste tõepäranumbritega arvutusse sisse panna ja millist mitte. Tõenäosusteooria ütleb paraku, et kui me välistame mõne olulise tõendi näiteks sellepärast, et see omandati süüaluse õigusi rikkudes, siis on ratsionaalne kohtuotsus loogiliselt võimatu.

Mõned järelmid Bayesi teoreemist:

1. Ilma eeljaotuseta pole võimalik arvutada järeljaotust (aga tõepärata saab hakkama, mis annab mugava viisi oma eeljaotuste kontrollimiseks).
2. Kui andmepunkti lisamisega muutub ühe hüpoteesi järeltõenäosus, mõjutab see ka kõigi teiste hüpoteesiruumi hüpoteeside järeltõenäosusi.
3. Kuna järeltõenäosused sõltuvad sellest, kuidas me hüpoteesiruumi osadeks jagame, siis halvasti jagatud hüpoteesiruumis sooritatud arvutused ei vii meid lähemale universaalsele tõele. Seega, kuni me teaduses ei küsi õigeid küsimusi, ei ole meil kindlust selle kohta, et me päriselt suudame andmetest õppida. Pessimistlik meta-induktsioon väidab, et kuna kõik senini teaduse ajaloos püstitatud hüpoteesid on varem või hiljem ümber lükatud (ehkki mõned neist püsisid aastatuhandeid), pole

meil põhjust uskuda, et ka praegusest hüpoteesisaagist midagi püsima jääks (Psillos, 2018). Seega pole lähiajal oodata bayesiaanliku paradiisi saabumist.

4. Ühest ja nullist erinevate eeltõenäosuste ja tõepärade korral ei jõua järeltõenäosused kunagi 0 ega 1-ni. Ebakindlad andmed ei vii kunagi loogiliselt kindlate uskumusteni.
5. Ehkki andmed nihutavad alati hüpoteeside tõenäosusi, mida lähemal on eeltõenäosus 0 või 1-le, seda rohkem on vaja andmeid, et järeltõenäosus erineks oluliselt eeltõenäosusest.
6. Kui eeltõenäosused ja tõepärad piirata kahe väärtusega, 0 ja 1, siis taandub tõenäosusteooria lausearvutuslikule loogikale. Seega on klassikaline loogika tõenäosusteooria kitsas erijuht.
7. Kui hüpoteesi eeltõenäosus on null või üks ja/või tõepära on null, siis ei suuda seda nihutada ükskõik kui suur uute andmehulk (nulliga korrutamise lõbu; kui P(H1) = 1, siis P(mitte-H1) peab olema null, et rahuldada tõenäosusteooria nõudmisi).

<u>Mida tähendab tõenäosus teadlasele?</u>

Alustuseks mõned küsimused, et intuitsioonipumpa käima tõmmata.

1. Monte Carlo kasiinos ruletil mängides, millise tõenäosusega jääb kuulike mustale väljale?
2. Koduses lasteruletis, millise tõenäosusega jääb kuulike mustale väljale?
3. Kui ilmajaam ütleb, et homme on vihma tõenäosus 60%, siis mida see tähendab?
4. Kui Pets vaatab taevasse ja ütleb, et homme on vihma tõenäosus 60%, siis mida see tähendab?
5. Kui suure tõenäosusega sajab Tartus vihma täpselt kuu aja pärast?

Vaatame lähemalt. Esimesel juhul saame tõenäosuse otse tõenäosusteooriast, jagades ruletiratta mustade väljade arvu läbi kõigi väljade arvuga (vastus on 0.486). See on võimalik, kuna tegu on õnnemänguga, mille reeglid on eelnevalt paika pandud, ja mis vähemalt sama tähtis, Monte Carlo kasiinod teevad märkimisväärseid kulutusi, et oma ruletirattaid tasakaalus hoida, seeläbi vältides, et kliendid neid pankrotti ajavad.

Teisel juhul on tegu odava ruletiga ja tõenäosuse leidmiseks peame mängureeglitest välja arvutatud sagedust adjusteerima empiiriliste andmete põhjal – ehk mängima ja võtma arvesse ka tegelikud tulemused. Mida rohkem kordi me ratast keerutame, seda vähemtähtsaks muutuvad mängureeglid ja seda suurema kaalu omandab empiiriline suhe. Mida arvata kolmanda juhu kohta? Siin saadakse vihma tõenäosus jooksutades ilmamudelit mitu korda stohhastiliselt varieeritud andmetel. Kui näiteks 6-l juhul 10-st näitab mudel vihma, siis vihma tõenäosus on 60%. Selline tõenäosus ei tähenda, et vihma sajaks 60% ajast või 60% pindalast, ega midagi muud päris maalima kohta. See iseloomustab pelgalt meie mudelipõhiste teadmiste määra homse vihma kohta. Hea meteoroloog adjusteerib mudelipõhist tõenäosust lähtuvalt oma kogemustest ja teadmistest antud paikkonna ilma kohta ja on näidatud, et hoolimata ilmamudelite pidevast paranemisest on see jätkuvalt vajalik kvaliteetse ilmaennustuse tagamiseks (Silver 2012).

Kuidas on siis neljanda juhuga? Mida tähendab 60% tõenäosus, mis on saadud ilma formaalse mudeldamiseta? Kui Pets on tõenäosusteooriat õppinud, siis tähendab 60% vihma tõenäosus tema jaoks õnnemängu. Kui keegi müüks piletit, mille ette näitamisel homse vihma toimumisel makstakse välja 1 EUR, siis on ta nõus selle eest maksma mitte rohkem kui 60 senti (saades võidu korral minimaalselt 40 senti kasumit). Jällegi, selline tõenäosus ei käi ilma kohta, vaid kajastab Petsi vaimuseisundit, mis omakorda põhineb arusaamisel ilmast ja selle ennustamisest selle järgi, kui madalalt lendavad pääsukesed ja kui punaselt loojub päike. Selline mängurlik 60% on siiski sama tähendusega, mis kolmanda juhu mudelipõhine 60%, sest mõlemal juhul põhineb tõenäosus ilmaandmetega žongleerimisel, lihtsalt erinevatel viisidel. (Sellega ei taha me öelda, et Petsi ennustus on sama usaldusväärne kui ilmajaama oma.)

Aga kuidas on lood kahe esimese juhuga, mille tõenäosused on pikaajalised suhtelised sagedused, olgu see siis mudelist (1. juht) või empiiriline (2. juht). Neid tõenäosusi võib vaadelda samamoodi, meie vaimuseisundite kirjeldustena, mis võtavad arvesse nii taustateadmised mängureeglite kohta, meie uskumused kasiino ja lasteruleti tootja erinevatest majanduslikest huvidest, kui empiirilised andmed ruletikeerutuste kohta, mida oleme vaadelnud. Kuid neid saab vaadelda ka hoopis teistmoodi – kui objektiivseid, päriselt maailmas esinevaid suhtelisi sagedusi. See kitsam vaade on sageduslik tõenäosuse tõlgendus (vt viimane peatükk).

Tuleme veel korraks viienda juhu juurde (vihma tõenäosus kuu aja pärast). Paljud inimesed nimetaksid seda lahendamatuks probleemiks, sest ükski ilmamudel sellist ennustust teha ei suuda. Selle vaate järgi on tegu näitega fundamentaalsest ebakindlusest, mida ei suuda leevendada ei andmed ega teooria ja millele tõenäosuste arvutamine oleks lihtsalt kohatu (Kay and King 2020). Tegelikult pole asi sugugi nii hull, sest me teame nii mõndagi ajaloolise ilma kohta Tartus sellel kuupäeval. Siin on meie parim ennustus lihtsalt ajalooline vihma sagedus (adjusteerituna kliima muutumisele, kui soovite).
Siit nähtub, et kuigi meie näited algasid väga täpsest tõenäosusehinnangust ja lõppesid väga ebatäpsega, võib siiski kõiki tõenäosusi tõlgendada põhimõtteliselt samal viisil, kui episteemilisi tõenäosusi, mis kirjeldavad meie usu määra mingi sündmuse toimumisse või teooria tõesusesse. Siinkohal tuleb tunnistada teatud ebakõla teadusliku ja statistilise praktika vahel. Kui me paneme teaduslikud probleemid, millele me lahendust otsime, sarnasesse suureneva komplekssuse & ebakindluse skaalasse, siis kipuvad need pahatihti langema suurema ebakindluse poolele. Samas, lineaarsed statistilised mudelid, mida me nende lahendamisel kasutame, lähevad kuhugi meie 1. juhu kanti, kus me neid mudeldame justnagu õnnemänge, mille mängulauad on hoolikalt rihti aetud.

<u>10 viisi tõenäosuste esitamine biomeditsiinilises kirjanduses</u>

1. **Risk** on sündmuse toimumise tõenäosus.
2. **Suhteline risk** (relative risk, risk ratio, RR) on kahe tõenäosuse suhe. Näiteks surmarisk suitsetajatel jagatuna surmariskiga mittesuitsetajatel.
3. **Riskitiheduste suhe** (Hazard ratio, HR) võrdleb samuti sündmuse tõenäosust kontrollgrupis selle tõenäosusega katsegrupis, aga selle järgi saab hinnata, kas katsegrupi patsientidel kulub sündmuse saabumiseni rohkem või vähem aega kui kontrollgrupi patsientidel. HR on sündmuse RR ajaühikus t. HR = 3 tähendab et igas ajavahemikus nähakse 3 korda rohkem sündmusi kui katsegrupis ja HR = 0.33 tähendab, et katsegrupis nähakse 3 korda vähem sündmusi kui kontrollgrupis. RR = 3 tähendab, et surma kumulatiivne risk on kolm korda kõrgem; HR = 3 tähendab, et risk on kolm korda kõrgem igal ajahetkel.
4. **Suhteline riski langus** (relative risk reduction) näitab kui palju risk langeb katsegrupis võrrelduna kontrollgrupiga: (KONTR - KATSE)/KONTR, kus KONTR on kontrollgrupi

tõenäosus ja KATSE on katsegrupi tõenäosus. Kui suri 70% kontrollgrupist ja 35% katsegrupist, siis *(0.7 – 0.35)/0.7 = 0.5*, ehk surmade sagedus katsegrupis on poole väiksem kui kontrollgrupis.

5. **Absoluutne riski langus** (risk difference, absolute risk reduction): KONTR – KATSE, ehk 70 – 35 = 35, mis tähendab, et absoluutne riski langus on 35 protsendipunkti.
6. **Šansid** (odds) on sündmuse tõenäosus jagatuna mitte-sündmuse tõenäosusega, ehk p/(1 - p). Kui šansid on 1/3, siis tõenäosus on ¼ ja kihlveo sõlmimisel nende šanssidega saab võidu korral ühe euro sissepanekul 3 eurot kasumit.
7. **Šansisuhe** (odds ratio, OR) on näiteks vähišansid suitsetajatel jagatud vähišansid mitte-suitsetajatel. OR > RR, ja see võib ka suurusjärkude võrra suurem olla. OR puudub intuitiivne tõlgendus, mis RR-il on selgelt olemas. Nende kasutamine võiks piirduda case-control uuringutega, kus ei ole võimalik ilma lisateadmisi kasutamata RR-i arvutada. Kuna OR-i on lihtne arvutada logistilise regressiooni beta-koefitsiendist, seda eksponentides, siis kasutatakse seda kirjanduses liiga palju.
8. **Tõenäosusintervall** (credible intervall, CI): 90% CI tähendab, et mudeli arvates on parameetri tõene väärtus tõenäosusega 0.9 selles vahemikus. See on bayesiaanlik statistik, mis arvestab eeljaotustega.
9. **Usaldusintervall** (confidence intervall, CI): 90% CI tähendab, et kui me kordaksime oma katset paljudel sõltumatutel juhuvalimitel, siis 90% arvutatud CI-dest hõlmab endas tõest parameetri väärtust. See on sageduslik statistik, millel sellisena puudub teaduslikus mõttes intuitiivne tõlgendus (see, et 90% katsetest hõlmab CI mingit väärtust ei tähenda, et meie katses on see väärtus 90% tõenäosusega sees). Samas, kui eeltõenäosus = 0.5, siis tuleb usaldusintrvall numbriliselt sama, mis normaalsele tõepärale arvutatud vastav tõenäosusintervall. Eeltõenäosusele hea põhjuseta kindla väärtuse omistamine on väga tugev eeldus.
10. **Tõepärasuhe** (likelihood ratio, LR) on kahe tõepära suhe. Tõepärad ei ole päris tõenäosused, sest need ei summeru ühele. Tõepärasuhe $P(a \mid H_1)/P(a \mid H_2)$ mõõdab tõendusmaterjali hulka, mida andmed annavad H_1 kasuks suhtena sellesse, mida need annavad H_2 kasuks. LR = 1 tähendab, et andmed on ebarelevantsed. Ehkki LR ütleb, kui tugevalt kõnelevad andmed H_1 kasuks võrreldes H_2-ga, võib H_3 ikkagi olla palju tõenäosem kui H_1 ja H_2. Seega ei ütle LR midagi H_1 ja H_2 kehtimise tõenäosuste kohta. Selleks on vaja LR läbi korrutada nende kahe hüpoteesi eeltõenäosuste

suhtega $P(H_1)/P(H_2)$, mille tulemusel me saame **Bayesi faktori** (Bayes Factor, BF). BF annab meile, mitu korda on H_1 tõenäolisem kui H_2.

Tõenäosuse sageduslik tõlgendus viib null-hüpoteesi testimisele

Erijuhtudel, kus meie relevantsed teadmised piirduvad sagedusinfoga, võib tõenäosusele anda ka päris maailma kohta käiva tähenduse, mis muudab tõenäosuse identseks sündmuste pikaajalise suhtelise sagedusega. Aga kui see on ainus tõlgendus, mida te olete nõus tõenäosusele andma, siis peate tõenäosusi arvutades paratamatult ignoreerima kõike, mis ei ole sagedusinfo. Sageduslik tõlgendus eeldab ka, et igale tõenäosusele vastab mingi üha uuesti ja uuest toimuv statsionaarne protsess: st, et teatud muutumatud põhjuslikud elementaarprotsessid tekitavad „katseid", milledest osad viivad „sündmusteni" ja teised mitte. Paraku ei ole teaduslikud hüpoteesid ja teooriad, nagu ka hinnangud parameetriväärtustele, sellised kasiino-protsessid. Seega ei saa sageduslikku tõenäosuse tõlgendust kasutades isegi küsida küsimust „millise tõenäosusega jääb keskmine efekti suurus vahemikku ...", ehk üldistatult: „milline on teaduslikult huvitava hüpoteesi (H_1) tõenäosus, kui meil on andmed ...", ehk $P(H_1 \mid a)$.

Et sellest raskusest mõõda hiilida, õppis R.A. Fisher 1920-ndatel aastatel küsima veidi teistsugust küsimust: milline on meie andmetega sarnaste andmete sagedus/tõenäosus juhul kui peaks kehtima null-hüpotees (H_0), mille kohaselt katsetöötluse mõju puudub (Fisher uuris sel ajal mineraalväetiste mõju viljasaagikusele). Kui me ei saa omistada sagedusi hüpoteesidele/parameetri väärtustele, võime selle asemel kujutleda, et meie andmed moodustavad ühe paljudest võimalikest juhuvalimitest statistilisest populatsioonist. Seega on tõenäosus valimitel, mitte hüpoteesidel, mida nende abil kontrollitakse. Kuna me uurime ikkagi ühte reaalset valimit, peame ülejäänud valimid simuleerima, mida on kõige lihtsam teha null-hüpoteesi abil, mis postuleerib väetisele null-mõju. Seega on meil (i) oma ainsa valimi põhjal arvutatud statistik (näiteks N põllulapi viljakuste keskväärtus) ja (ii) (lõpmata) suur hulk kujuteldud H_0-st võetud juhuvalimitest arvutatud statistikuid. Need H0 statistikud moodustavad tõenäosusjaotuse, ja kui (i) punktväärtus jääb kaugele selle jaotuse kõrgemast osast, siis saame väikese p väärtuse. Seega arvutame $P(a \mid H_0)$ ehk p väärtuse, mitte $P(H_1 \mid a)$.

See meetod eeldab, et me oleme korjanud juhuvalimi statistilisest populatsioonist ning et see juhuvalim peegeldab adekvaatselt populatsiooni (seega ei tohi valim olla liiga väike). Tõsi küll, vähemalt molekulaarbioloogias, kus p väärtusi kasutatakse rohkem kui tualettpaberit, on ebaselge, mida „juhuvalim“ võiks üldse tähendada, ning valimi suurused kipuvad sageli olema üliväikesed (N = 2-4).

Peenet nalja saab ka p väärtuste teaduslikult huvitava tõlgendamisega. Kui meie poolt arvutatud $P(a \mid H_0)$ on 0.01, ehk $p = 0.01$, ja me võtame lisaeeldusena arvesse, et taustateadmiste pinnalt hinnatuna on ravimi null-mõju tõenäosus 0.5 e 50%, siis saame Bayesi teoreemist, et ka $P(H_0 \mid a) = 0.01$, ehk nullhüpoteesi tõenäosus meie andmete korral on (ligikaudu) 0.01. Seega on vastandhüpoteesi tõenäosus $1 – 0.01 = 0.99$, mille võib lahti kirjutada nii: 99% tõenäosusega on teie ravimimõju kas naeruväärselt väike, või siis usutava ja meditsiiniliselt kasuliku suurusega, või hoopis absurdselt suur, või on hoopistükkis midagi väga viltu teie valimi varieeruvusega. Ilmselt ei ole selle hüpoteesi tõenäosuse arvutamine päris see, mida teadlane tahtis saavutada p väärtust arvutades.

Eelnev ei tähenda, et p väärtuste arvutamisest kunagi kasu ei võiks tõusta – vähemalt Roland Fisher kasutas neid hästi ja sai lõikas sellest teaduslikku profiiti. Küll aga tuleb veelkord meelde tuletada, et statistiliste meetodite kasutamine mitte-luksuma-ajaval moel on puhtalt ja eranditeta teadlase vastutusel. Statistik saab siin ainult nõu anda.

Kokkuvõtteks:

Tõenäosusteooria annab meile ratsionaalse viisi kujundamaks oma uskumusi ebatäielike ja ebatäiuslike andmete põhjal. See on loogiliselt täiuslik viis erinevate infokildude ühendamiseks, mille edukus väljaspool matemaatilist nirvaanat sõltub nii informatsiooni kvaliteedist kui sellest, kui hästi meie mudelid peegeldavad maailmas andmeid genereerivaid protsesse. Statistiliste mudelite väljundiks on tõenäosused, mille kasutamine teadusliku argumendi osana eeldab võimet neist mõelda mitte-matemaatiliselt, st anda tõenäosusele maailma uurimise seisukohalt mõtekaid tõlgendusi. Teadlasel, kes ei suuda, pole ka põhjust tõenäosusi arvutada.

Hacking, Ian. 1975, The Emergence of Probability, Cambridge University Press.

Hacking, Ian. 1990, The Taming of Chance, Cambridge University Press.

Kay, John ja King, Mervin. 2020. Radical Uncertainty. The Bridge Street Press, London.

Maiväli, Ülo. 2015. Interpreting Biological Science. Elsevier.

Prommik, Pärt, Maiväli, Ülo. 2021. Bayesi regressioonist praktiku pilgu läbi. Matemaatika Seltsi Aastaraamat. Valmimisel.

Psillos, Stathis, 2018. "Realism and Theory Change in Science", The Stanford Encyclopedia of Philosophy (Summer 2018 Edition), Edward N. Zalta (ed.) https://plato.stanford.edu/archives/sum2018/entries/realism-theory-change/

Päll, Taavi; Luidalepp, Hannes; Tenson, Tanel; Maiväli, Ülo. 2021. A field-wide assessment of differential high throughput sequencing reveals widespread bias. bioRxiv. DOI: 10.1101/2021.01.04.424681

Päll, Taavi ja Maiväli, Ülo. 2019. Bayesi statistika kasutades R keelt. Loengumaterjalid. https://rstats-tartu.github.io/bayesiraamat/

Ramsey, Frank.P (1926) "Truth and Probability", in Ramsey, 1931, The Foundations of Mathematics and other Logical Essays, Ch. VII, p.156-198, edited by R.B. Braithwaite, London: Kegan, Paul, Trench, Trubner & Co., New York: Harcourt, Brace and Company. http://fitelson.org/probability/ramsey.pdf

Silver, Nate. 2012. The Signal and the Noise. The Penguin Press.

Tversky, Amos., & Kahneman, Daniel. 1974. Judgment under Uncertainty: Heuristics and Biases. Science, 185(4157), 1124–1131.

Vineberg, Susan, "Dutch Book Arguments", The Stanford Encyclopedia of Philosophy (Spring 2016 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2016/entries/dutch-book/