

STATISTILINE OLULISUS VALIMIVÕTT JA VALIMIMAHT

STATISTILINE OLULISUS. STATISTILISE ANALÜÜSI MEETODID EPIDEMIOLOOGILISTES UURINGUTES

1. Mida annab meile statistiline analüüs?

Statistilisel analüüsil on kaks põhilist väljundit:

1) **Mõju e. efekti e. tulemi suurus** - näitab kui suur on ühe või teise teguri toime populatsiooni v. isendi parameetritele.

Näit.: antibiootikumiravi efekt kinnisperioodil karja keskmisele somaatiliste rakkude arvule

2) **Statistiline olulisus** - näitab, milline on tõenäosus, et resultaat või efekt, mille te avastasite, on saadud puht juhuse läbi.

Statistilist olulisust väljendab "**p**" väärtus, mis mahub 0 ja 1 piiridesse (kuna on **tõenäosus**). Väike p väärtus (0.01) näitab, et juhuse tõenäosus on väike (antud juhul vaid 1%). Järelikult võite p väärtuse 0,01 puhul olla 99% kindel, et tulemus ei ole juhuslik.

Tavaliselt loetakse $p=0,05$ tulemuste statistilise olulisuse piiriks

Kummatigi ei tähenda $p>0,05$ ilmingimata, et täheldatud tulemus on ka **sisuliselt tähtis**. Siin on koht, kus valemeid ei saa pimesi usaldada. Hinnang saadud tulemuste tähtsusele uurimuse seisukohast võttes tuleb anda inimesel. Siit tulenevalt ei maksa automaatselt kõrvale heita ka tulemusi mille p väärtus on pisut suurem kui 0,05.

2. Statistiline olulisus, valimi suurus ja statistiliste vigade tüübid

Tulemuse statistiline olulisus sõltub kahest tegurist:

(1) mõju suurusest

(2) valimi suurusest

Isegi teguri suur mõju, kui seda on uuritud väikesel arvul isenditel ei pruugi olla statistiliselt oluline. Ja vastupidi, väike efekt võib osutuda statistiliselt oluliseks, kui seda on täheldatud piisavalt suurel arvul loomadel.

Teguri mõju hinnatakse katse ja kontrollrühma parameetrite võrdlemise teel. Võrdluse tulemusel me kas avastame erinevuse või mitte. Samuti tegelikkuses erinevus võib eksisteerida või mitte. Sellest lähtuvalt võime esitada järgmise tabeli uurimistulemuste ja tegeliku olukorra vahekorra:

		TEGELIK OLUKORD	
		ERINEVUST POLE	ERINEVUS ON
UURIMISTULEMUSED	ERINEVUST POLE	TÕENE TULEM	II tüüpi viga
	ERINEVUS ON	I tüüpi viga	TÕENE TULEM

I tüüpi viga - me saame uurimistulemuseks erinevuse seal kus seda tegelikult ei eksisteeri. Olulisust tähistava p-väärtuse hoidmine madalal tasemel enamasti kaitseb meid esimest-tüüpi vigade eest.

II tüüpi viga - erinevus eksisteerib, kuid meie uurimistulemused seda ei väljenda – on eelkõige seotud valimi väiksusega.

3. Statistilise analüüsi meetodi sõltuvus andmete tüübist

Meie poolt kogutavad andmed saavad olla peamiselt kaht tüüpi:

(1) Kvantitatiivsed (pidevad, diskreetsed) - st. tunnuse väärtus võib olla mistahes arv nullist lõpmatuseni. Sellisteks andmeteks on kaalu, pikkust, läbimõõtu kirjeldavad andmed, aga ka loendusandmed (loomade arv, somaatiliste rakkude arv/ml)

(2) Kategoorialised – andmed on klassifitseeritud teatud kategooriatesse

Kategoorialised tunnused jaotatakse omakorda:

a) Järjestusandmed (ordinaalsed): tunnused, mida on võimalik järjestada kasvavas järjekorras (haiguse kulgu: krooniline, alaäge, äge)

b) Nominaalne: ei ole võimalik järjestada (asulate nimed, tõug). Nominaalsete tunnuste eri liik on: Kaheväärtuselised, e. binaarsed, e. dihhotoomsed tunnused; (sugu; haige-terve jne.)

Märkus: Tuletame meelde, et kvantitatiivseid andmeid on võimalik muuta kategoorialisteks.

Statistilises analüüsis on vajalik veel üks tunnuste klassifitseerimine. Nimelt klassifitseeritakse tunnused (muutujad) **sõltumatuteks** ja **sõltuvateks**. **Sõltuv muutuja** on see mida mõõdetakse. **Sõltumatu muutuja** see, mille mõju mõõdetakse.

Märkus: Üks ja sama muutuja võib ühes uuringus olla sõltuv ja teises sõltumatu. Muutuja sõltuvus ei tulene otseselt muutuja omadustest.

Epidemioloogilistes uuringutes sagedamini kasutatavad statistilist olulisust hindavad testid sõltuvalt andmete tüübist:

(Väljund) SÕLTUV TUNNUS	SÕLTUMATU MUUTUJA			
	Kategoorialine			Kvantitatiivne
	Võrreldakse 2 sõltumatut rühma (näit. katse ja kontrollrühm)	Võrreldakse 2 sõltuvat rühma (näit. enne ja pärast ravi)	Võrreldakse >2 rühma	
Pidev (normaaljaotusele vastav)	Paaritu t-test	Paaris t-test	Dispersioonanalüüs (ANOVA)	Korrelatsioon ja regressioonanalüüs
Nominaalne binaarne	χ^2 -test	McNemar'i χ^2 -test	χ^2 -test	Logistiline regressioon
Nominaalne > 2 kategooriat	χ^2 -test	Cochran'i Q-test	χ^2 -test	diskriminantanalüüs

4. Mõju suurust iseloomustavad parameetrid

Mõju suurust iseloomustavad populatsiooni kirjeldavad statistikud.

Kvantitatiivsete andmete puhul on selleks **aritmeetiline ja geomeetriline keskmine**.

Geomeetrilist keskmist kasutatakse andmete puhul, mis ei ole normaalselt jaotunud. Geomeetrilise keskmise kalkuleerimine käib andmete logaritmitamise ja saadud tulemuste keskmise anti-logaritmitamise teel.

Miks vajame andmete geomeetrilist transformeerimist ja geomeetrilist keskmist? Kuna enamik statistilise analüüsi meetodeid töötab usaldusväärselt vaid normaalselt jaotuvate andmete puhul.

Märkus: Nagu selgub eelnevast, on oluline määrata enne andmete analüüsimist nende jaotumine (histogramm).

Mõju suuruse iseloomustamiseks on oluline lisaks keskmistele vajalik välja tuua ka andmete hajuvuse näitajad, nagu haare, dispersioon ja standardhälve ning keskmise usaldatavust iseloomustavad näitajad nagu keskmise standard viga ja usaldusvahemik

Kategoorialiste andmete puhul on mõju suurust kirjeldavateks näitajateks sündmuse esinemissageduse näitajad ning ka sageduste suhtelised näitajad, nagu levimuste suhe, suhteline risk ja šansside suhe.

5. Testid statistilise olulisuse hindamiseks

T-test

T-testi kasutatakse kahe loomarühma **kvantitatiivse tunnuse keskmiste võrdlemiseks**.

T-testi kasutamise eelduseks on, et andmed on enamvähem normaalse jaotusega ja dispersioon kummaski rühmas on enamvähem võrdne.

On olemas kahte tüüpi t-testi

1) T-test sõltuvate rühmade võrdlemiseks (paired t-test)

Valimid on sõltuvad, kui andmed, mida me võrdleme pärinevad ühest ja samast loomarühmast.

Näiteks:

a) kahte erinevat ravikuuri kohaldatakse samadel loomadel mõningase ajavahega ning võrreldakse ravi efekti või muutusi füsioloogilistes näitajates esimese ja teise ravikuuri järgselt.

b) võrreldakse näitajaid enne ja pärast ravimi manustamist

c) võrreldakse teatud näitajate muutumist elu jooksul (näiteks piimatoodang eri laktatsioonides)

Kõigil neil juhtudel huvitavad meid loomaga toimuvad muutused ja t-test selgitab, kas muutus erineb oluliselt 0-st.

2) T-test sõltumatute valimite jaoks

Valimid on sõltumatud, kui andmed, mida me võrdleme pärinevad kahelt erinevalt loomarühmalt.

a) võrreldakse raviefekti kahes loomarühmas, kellele on kohaldatud erinevaid kuure:

b) võrreldakse ravitud ja ravimata loomarühma.

Seega huvitavad meid rühmadevahelised erinevused ja t-test näitab kas erinevus on oluliselt suurem 0-st.

Dispersioonanalüüs (ANOVA)

Kui meil on vaja võrrelda enam kui kahte rühma, kasutame me t-testi asemel dispersioonanalüüsi. Võimalik on dispersioonanalüüs nii sõltuvatele kui sõltumatutelerühmadele, kuid et esimene on väga harvaesinev juht, siis keskendume viimasele.

Nagu ka t-testis nii eeldatakse ka ANOVA korral, et andmed on normaaljaotusega ja kõigis rühmades sarnase dispersiooniga.

ANOVA puhul iseloomustab rühmadevahelist erinevust F-statistiku väärtus, mis oma statistiliselt sisult on sarnane t-staistikule. F-väärtuse statistilist olulisust iseloomustab p väärtus. madal p väärtus ütleb meile, et rühmadevaheline variatsioon on statistiliselt oluline, kuid selle alusel ei saa öelda, milliste rühmade vahel erinevus oli oluline. Selle selgitamiseks tuleb rühmi võrrelda paarikaupa t-testi abil.

Katagoorialiste andmete analüüs

Kvalitatiivsete andmete analüüsimeetodeid rakendatakse juhul kui nii sõltumatu kui sõltuv tunnus on kas nominaalsed või ordinaalsed. Me vaatleme siinjuures kvalitatiivsete andmete analüüsi juhul kui nii sõltumatu kui sõltuv muutuja on binaarsed (kaheväärtuselised).

Sellise analüüsi aluseks on χ^2 -test ja 2x2 tabel.

2x2 tabeli analüüs võrdleb kahe sündmuse esinemissagedust kahes loomarühmas, kes on jaotatud kaheks ühe sõltumatu tunnuse alusel ja kus sündmus ise kujutab endast sõltuvat tunnust. Seega võrreldakse sõltumatu tunnuse mõju sündmuse esinemissagedusele.

Näit.: ketoosi esinemine kahel veisetõul

	ketoos 0	ketoos 1	Kokku
Tõug 1	72	64	136
Tõug 2	241	105	346
Kokku	313	169	482

Tabelist selgub, et ketoosi esinemissagedus tõul 1 on 47% ja tõul 2 - 30 %.

Lugedes tõu 1 riskiteguriks ketoosi haigestumise suhtes võime teguri mõju suurust väljendada suhtelise riski (levimuste suhte) ja šansside suhte kaudu. Suhteline risk on antud näite puhul $SR = 47/30 = 1,56$

Selgitamaks, kas saadud erinevus on ka statistiliselt oluline analüüsitakse andmeid χ^2 testi abil.

Antud näite puhul on χ^2 väärtuseks 11,98. Vastavast tabelist on võimalik leida p väärtus sellisele χ^2 väärtusele 1 vabadusastme juures (kaheväärtuselised andmed - järelikult $2 - 1 = 1$ vabadusaste). Tabel näitab, et $p=0,0005$, mis tõestab, et saadud erinevus on väga väikse tõenäosusega juhuslikku laadi.

Oluline on märkida siinjuures, et 2x2 tabeli kasutamine eeldab piisavalt suurt valimit, kuna igas lahtris peab eeldatav väärtus olema ≥ 5 . Vastasel korral ei ole analüüsi tulemused piisavalt usaldusväärsed.

Olgu veel lisatud, et 2x2 tabelit on võimalik kasutada ka sõltuvate rühmade analüüsil ja rohkem kui 3 kategooriaga tunnuste korral, kuid neid juhte me siin ei käsitle.

Korrelatsioonanalüüs

Korrelatsioon näitab kuivõrd on seotud kaks kvantitatiivset tunnust omavahel. S.t. kuivõrd ühe tunnuse muutus mõjutab teise tunnuse väärtust.

Korrelatsiooni väljendab korrelatsiooni koefitsient r , mille väärtus võib olla $-1 \dots +1$.

-1 ütleb, et kahe tunnuse vahel valitseb täiuslik negatiivne korrelatsioon.

+1 ütleb, et kahe tunnuse vahel valitseb täiuslik positiivne korrelatsioon.

0 näitab, et mingisugust seost ei eksisteeri.

r^2 näitab milline osa teise tunnuse variatsioonist on seletatav või põhjustatud esimese tunnuse poolt.

Korrelatsioonanalüüsil põhineb ka **regressioonanalüüs**, mille sügavustesse me siinjuures ei sisene.

Kokkuvõte

Nagu eelnevast selgub, sõltub statistilise analüüsi meetod lisaks andmete tüübile ka rühmade arvust uuringus, samuti sellest, kas rühmad on teineteisest sõltumatud või sõltuvad. Lisaks sellele tuleb arvestada nominaalsete andmete puhul, kui mitme kategooriaga on tegemist ja kui palju indiviide on meil igas kategoorias. Lõpetuseks sõltub meetodi valik kvantitatiivsete andmete puhul ka sellest, kas meie andmete jaotus vastab normaaljaotusele. Eelkirjeldatud testid (t-test, ANOVA, korrelatsioon- ja regressioonanalüüs, nimetatakse ka parameetrilised testid) on rakendatavad vaid normaaljaotusega andmete korral. Vastasel juhul on vaja andmeid teisendada viimaks nende jaotuse vastavaks normaaljaotusele (näit geomeetrilise keskmise arvutamise abil) või kasutada mitteparameetrilisi teste, mida kasutatakse kategoorialiste andmete analüüsil (näiteks on siin juures χ^2 – test) või rakendada Bays'i teoreemil põhinevat analüüsi.

Parameetrilisi teste saab kasutada ainult kvantitatiivsete andmete analüüsil. Parameetrilise testi kasutamisel on lisaks normaaljaotusele vastamise nõudele veel kaks eeldust. Need on järgmised: (1) rühmadesse võetud isendid on valitud populatsioonist juhuslikult (2) dispersioonid rühmade andmetes on võrdsed. (3) rühmades on piisav arv üksusi (rohkem kui 6).

Mitteparameetrilisi teste kasutatakse eelkõige nominaalsete ja ordinaalsete andmete analüüsil, kuid saab kasutada ka kvantitatiivsete andmete analüüsil. Mitteparameetrilise testi rakendamisel on eeldusi vaid üks – rühmadesse võetud isendid on valitud populatsioonist juhuslikult. Need testid ei sõltu andmete jaotusest.

Oma statistiliselt jõult on mitteparameetrilised testid nõrgemad kui parameetrilised testid, see tähendab piltlikult seda, et erinevused rühmade vahel peavad olema oluliselt suuremad selleks, et mitteparameetiline test tunnistaks erinevused statistiliselt oluliseks võrreldes parameetrilise testiga.

VALIMIVÕTT JA VALIMIMAHT

SISSEJUHATUS

Loomaarstid tahavad sageli teada saada haiguse või muu olulise tootmisparameetri levimust või haiguse olemasolu loomade populatsioonis.

Selleks võib kasutada **loenduslikku meetodit** (*census*); loenduse puhul on iga üksus vaadeldavast populatsioonist haaratud valimisse. Loendus kätkeb andmeid kõikide loomade kohta populatsioonis, näiteks kõikide loomade kohta teatud karjas. Sellisel juhul on populatsioon **eksplitsiitne**.

Kui populatsioon on väike, siis on see kõige täpsem ja efektiivsem viis informatsiooni kogumiseks.

Siiski on enamikul juhtudest uuritav populatsioon suur ning loendused muutuvad raskesti läbiviidavateks ja äärmiselt kulukateks. Samuti võivad andmed olla kättesaadavad ainult vaadeldava populatsiooni teatud osast ja/või viimane ei pruugi olla kergesti piiritletav füüsiline üksus. Näiteks saab piirkonna karjade ühe osa andmete alusel hinnata maakonna kõikide karjade keskmist toodangut. Tapamajas on võimalik teatud hulga lihakehade testimisel ravimijääkide sisalduse suhtes kogutud andmete alusel teha järeldusi kõikide ülejäänud lihakehade seisundi kohta.

Tuleb täheldada, et kuigi kõikide lihakehade või kõikide karjade populatsioon on tõeline, ei ole seda nii kerge määratleda kui kõiki lehmi teatud karjas. Sellisel juhul on populatsioon **implitsiitne**.

Sageli eelistatakse **valimiuuringut** loendusele kulude, ressursside ja aja kokkuhoiu mõttes.

Valim on esemete või isendite kogum, mis on valitud suuremast kogumist või **populatsioonist**, mille kohta me informatsiooni saada soovime. Et valimisse kuulub ainult osa populatsioonist, on mõõdetud üksuste arv suhteliselt väike. Peale selle, et valimi moodustamine on oluliselt säästvam, võib valimiuuring anda täpsemaid tulemusi kui loenduslik uuring. Seda seetõttu, et väiksem ülesannete maht võimaldab paremat kontrolli vaatluste ja mõõtmiste üle.

Siiski võivad valimiuuringu kõik eelised väheneda, kui uuritavate üksuste populatsioonist valimise viis on hooletult valitud või kui andmete analüüs ei vasta kasutatud valimi moodustamise meetodile.

Keskseks näitajaks on see, kuidas andmed on populatsioonist saadud. Kasutatud valimi moodustamise protseduur määrab põhimõtteliselt korratavuse ja valimilt populatsioonile tehtava ekstrapolatsiooni laadi.

Populatsiooni kirjeldused koosnevad sageli kahest osast. Esimene on hinnang populatsiooni mõnele omadusele, mida mõnikord nimetatakse **punkthinnanguks** (*point estimate*). See vastab sellistele küsimustele nagu: milline oli teatud tekitaja vastu suunatud antikehade levimus uurimise ajal?

Olles teadlikud variatsioonist, ei saa me olla kindlad valimi punkthinnangu korrektsuses. Järelikult tahame me lisaks väljendada seda, kui suur on punkthinnangu tõenäoline viga. Üks viis seda teha on täiendada punkthinnangut **intervallhinnanguga**. Selle alusel saame näiteks öelda, et valimi põhjal oleme me kindlad haiguse X vastu suunatud antikehade levimuses vahemikus a kuni b protsenti.

Kindel olemise all peame me silmas seda, et näiteks 95%lise tõenäosusega sisaldab meie valimi intervall – a kuni b protsenti – tegelikku, kuid tundmatut populatsiooni levimust.

TERMINOLOOGIA

1. **Mittetõenäosuslik valik** (*nonprobability sampling*) on meetodid, mis ei kasuta formaalseid juhuslikkuse tehnikaid valimisse kaasatavate valimiüksuste määratlemiseks.

Näideteks on mugavusvalim, eesmärgist lähtuv valim ja hinnanguline valim.

Mugavusvalimisse (*convenience sampling*) on üksused valitud “mugavalt”, näiteks hõlpsa kättesaadavuse, leitavuse, käsitletavuse, koostöövalmiduse tõttu.

Eesmärgist lähtuva valiku (*purposive sampling*) puhul põhineb üksuste võtmine valimisse nende eksponeeritusel või tervisestaatusel.

Hinnangulise valiku (*judgement sampling*) korral valitakse üksused, mida uurija hindab kui “tüüpilisi”.

Mittetõenäosuslikud valimid põhjustavad sageli nihkeid, kuna tüüpilised või mugavad valimiliikmed võivad sageli osutada ebatüüpilisteks või mitteesinduslikeks.

2. **Tõenäosuslik/juhuslik valik** (*probability/random sampling*) on järgmiste omadustega:

- populatsiooni igal liikmel on teatud tõenäosus saada valimisse kaasatud,

- valik on tehtud mõne juhusliku valiku meetodiga, mis on kooskõlas nende tõenäosustega,
- valimist hinnangute kalkuleerimisel võetakse arvesse valiku tõenäosused.

Lihtne juhuvalim, süstemaatiline juhuvalim, kiht-juhuvalim, klastervalim ja mitmeastmeline valim on tõenäosusliku valiku meetodid.

3. **Valimiüksused** (*sampling unit*) on esemed, asukohad, inimesed, loomad jne, kes/mis valitakse individuaalselt valiku meetodi abil. Valikuprotsessi erinevatel tasemetel võivad olla erinevat tüüpi üksused, mida nimetatakse primaarseteks, sekundaarseteks, tertsiaarseteks valimiüksusteks. Näide: maakonnad on primaarsed, vallad sekundaarsed ja farmid tertsiaarsed üksused.
4. **Valikuloend** (*freim*) (*sampling frame*). Tavaliselt nimekiri, mis sisaldab populatsiooni kõiki valimiüksusi. See võib olla tegelik nimekiri, näiteks piirkonnas registreeritud kõikidest farmidest või karja kõikidest loomadest jne. Alati ei ole vaja omada täielikke nimekirju. Tegelikult jaotatakse tavaliselt suuremad valiku loendid väiksemateks, st piirkond jaotatakse väiksemateks administratiivseteks üksusteks, kust seejärel valitakse farmid.

TÕENÄOSUSLIKU/JUHUSLIKU VALIKU MEETODID

(Otte, 1991)

Lihtne juhuslik valik (*simple random sampling*)

Lihtsa juhusliku valiku esimeseks sammuks on valiku loendi moodustamine. Teine samm on valiku loendist valimiüksuste valimine mõne formaalse juhuslikkuse meetodiga, näiteks juhuslike arvude tabeli kasutamine või numbrite võtmine mütsi seest ilma asendamiseta, kuni valimi maht on saavutatud. Lihtsa juhusliku valiku puuduseks on, et seda ei saa teha ilma piisavalt täpse ja üksikasjaliku valiku loendita, mis kindlustab valimiüksuste valimise võrdse tõenäosusega. Põllumajandusloomade uuringutes, kus sageli tuleb uurida üksikuid loomi suuremates või väiksemates geograafilistes piirkondades, on üsna ebatavaline, et omatakse piirkonna kõikide loomade nimekirja. Välja arvatud väikestes populatsioonides (näiteks veisekari või väga lihtsa struktuuriga populatsioonid) on lihtne juhuslik valik sageli liiga kulukas või ebaefektiivne ning vajalikuks osutuvad palju komplekssemad valikuprotseduurid.

Süstemaatiline juhuslik valik (*systematic random sampling*)

Siia valitakse valimi üksused populatsioonist korrapäraste intervallide järel, näiteks iga viies loom, kui on vajalik 20% valim. Kui kasutatakse süstemaatilisi meetodeid, valitakse alguspunkt esimesele intervallile formaalse juhuslikkuse alusel. Süstemaatilise valiku peamiseks eeliseks on, et seda on kergem teostada kui lihtsat juhuslikku valikut ning see võimaldab valimit võtta populatsioonist, mille täpne suurus ei ole teada ja mille liikmed ei ole individuaalselt kindlaks tehtud. Selle meetodi peamiseks puuduseks on, et valimiüksused võivad valimi loendis või populatsioonis olla häiritud regulaarsete intervallide järel, mis kattuvad valiku intervallidega, ja sellisel juhul võib valimist tulenev hinnang olla väga eksitav. Seda peaks esinema siiski harva ja analüüsis tuleb andmeid käsitleda, nagu oleksid need kogutud lihtsa juhusliku valikuna.

Juhuslik kihtvalik (*stratified random sampling*)

Juhusliku kihtvaliku puhul jaotatakse valiku loend enne üksuste valimist teatud alampopulatsioonideks ehk kihtideks. Kihtide määratlemise aluseks on tunnused, mis tõenäoliselt mõjutavad mõõdetavat omadust (näiteks vanus, tõug, sugu). Seejärel võetakse igast kihist lihtne juhuslik või süstemaatiline juhuvalim. Juhuslik kihtvalik on paindlikum kui lihtne juhuslik valik, kuna valimi võtmisel võib erinevates kihtides kasutada erinevat valikuprotsenti. Seda tehnikat tuntakse kui kihitamist varieeruvate valimifraktsioonidega ja tehnika kasulikkus seisneb selles, et ta lubab koondada meie käsutuses olevad vahendid populatsiooni nendele osadele, mis meile iseäranis huvi pakuvad. Lisaks saab parandada valimi hinnangu korratavust, kuna juhusliku kihtvaliku puhul mõjutab keskmise variatsiooni (standardviga) ainult kihisisene variatsioon. Kihitamise efektiivsus on seda suurem, mida suuremad on erinevused eri kihtide keskmiste väärtuste vahel ja mida sarnasemad on väärtused kihi sees. Kui kogupopulatsiooni hinnangu korratavuse parandamine on kihitamise peamiseks motivatsiooniks, on see harva väärt rohkem kui viie või kuue kihi moodustamist ja sageli saab palju kasu, kasutades vaid kahte või kolme kihti.

Kihtvaliku ilmseks puuduseks on see, et kõikide valimiüksuste staatus kihte moodustavate tegurite suhtes peab olema teada enne valimi võtmist.

Selleks, et võimaldada rohkem paindlikkust valikustrateegiates ning vähendada isendite loendi moodustamise kulusid ja praktilisi raskusi, on sageli alguses kergem võtta valim karjade või loomade

muude loomulike kogumite hulgast sihtpopulatsioonis. Seejärel võib nõutava arvu isendeid valida nendest kogumitest. Selleks otstarbeks kasutatavatest valikumeetoditest on 2 kõige tavalisemat klastervalik ja mitmeastmeline valik.

Klastervalik (*cluster sampling*)

Klastervaliku puhul on esialgne valimiüksus suurem kui uuringuüksus, viimane on tavaliselt üksikisend. Näiteks küla, farmi või karja kõik loomad võivad olla võetud valimisse, seejuures küla, farm või kari moodustavad valimiüksuse ja loomad valimiüksuses on klasteriks. Isendite klasterid tulevad sageli loomulikult esile (näiteks pesakonnad, sulud, karjad). Valiku eesmärkidel võib kunstlike klasteritena kasutada ka administratiivseid üksusi nagu maakonnad.

Klastreid võib valida süstemaatilise või lihtsa juhusliku valiku või juhusliku kihtvaliku meetodil ja kõiki isendid nendes klasterites uuritakse.

Et teha kergemaks mis tahes hinnangu standardvea arvutamist, on soovitatav valida klasterid koos asendusega, see tähendab, et tuleb teha valik täielikust klasterite loetelust, ilma et ühtegi eelnevalt valitud klasterit kõrvaldaks. Sellise meetodi puhul võib sama klaster sattuda valimisse mõnikord rohkem kui üks kord, mis tähendab, et klasterit tuleb uurida rohkem kui üks kord.

Et valimite võtmise peamiseks kuluartiklik on sageli reisikulud, siis on valimi moodustamisel küla, farmi või karja kõikidest loomadest ilmsed eelised (klasterisse peavad olema võetud kõik loomad). Reisikulude alanemise tõttu on arenguriikides klastervalik sageli peamiseks meetodiks epidemioloogilistes uuringutes. Siiski, kuna haiguse levimuse varieerumine on klasterite vahel tõenäoliselt suurem kui klasterite sees, siis annab loomade uurimine klasteri sees vähem informatsiooni kui erinevatest klasteritest pärinevate loomade uurimine. Iseäranis on see nii nakkushaiguste puhul. Seega annab klasterite valim tavaliselt vähem täpse hinnangu, kui oleks võimalik saada sama suurusega lihtsast juhuvalimist. Sellest tulenevaks üheks järelduseks on see, et minimaalne proovide arv haiguse levimuse hindamiseks sihtpopulatsioonis tervikuna etteantud korratavusega on mitmeid kordi suurem kui lihtsa juhusliku valiku korral. Üldiselt, mida suurem on varieerumine klasteri sees ja mida väiksem on varieerumine klasterite vahel, seda efektiivsemaks klastervalik muutub.

Mitmeastmeline valik (*multi-stage sampling*)

See meetod on sarnane klastervalikuga, kuid nagu nimetus mõista annab, sisaldab mitmeastmeline valik populatsioonist valimi võtmist erinevatel tasanditel, kusjuures valimiüksus on igal tasemel erinev. Näiteks karjade ja loomade kaheastmelise valiku korral valitakse karjad (esimene aste) formaalse juhuslikkuse tehnika abil ja seejärel valitakse loomad (teine aste) juhuslikkuse meetoditega ainult nendest karjadest, mis olid valitud esimeses astmes. Seega alustatakse nagu klastervalikulgi primaarsete üksuste valimi võtmisega, näiteks loendi nimekirjas olevate küladega. Seejärel valitakse igas külas sekundaarsete üksuste valim, näiteks loomad.

Mitmeastmelist valikut kasutatakse sageli selle praktiliste eeliste ja paindlikkuse tõttu. Primaarsete ja sekundaarsete üksuste arv võib erineda, et võtta arvesse erinevad kulud valimi võtmisel primaarsel ja sekundaarsel tasandil, nagu ka hinnatava karakteristiku varieeruvuse primaarsete üksuste vahel ja sekundaarsete üksuste vahel primaarse üksuse sees. Mil iganes võimalik, tuleks iga taseme valimiüksus valida tõenäosusega, mis on proportsionaalne isendite valikuüksuste arvuga, mida nad sisaldavad. Nagu klastervaliku puhulgi on mitmeastmelise valiku peamiseks puuduseks see, et tavaliselt on valimisse vaja rohkem loomi, saamaks sama korratavust, kui see oleks ootuspäraselt loomade valikul lihtsa juhusliku valiku teel.

VALIMITE MAHUD

Määrates valimi mahtu tahetakse vältida liiga väikese valimi võtmist, mille korral hinnang oleks liiga ebatäpne ja ei omaks väärtust. Samavõrd tahetakse vältida valimit, mis on liiga suur; see raiskaks väärtuslikke ressursse.

Seetõttu on valimimahtude täpne määramine tähtis, kuid võib olla üsna keeruline ja vajada mitmetel juhtudel statistikute abi.

Valimimahtude määramise kaks peamist kasutusala on monitooring ja seire.

Haiguse **monitoorimisel** eeldatakse haiguse olemasolu ja eesmärk on määrata selle levimust ja/või esinemust. Seega on eesmärk hinnata haiguse esinemise määra.

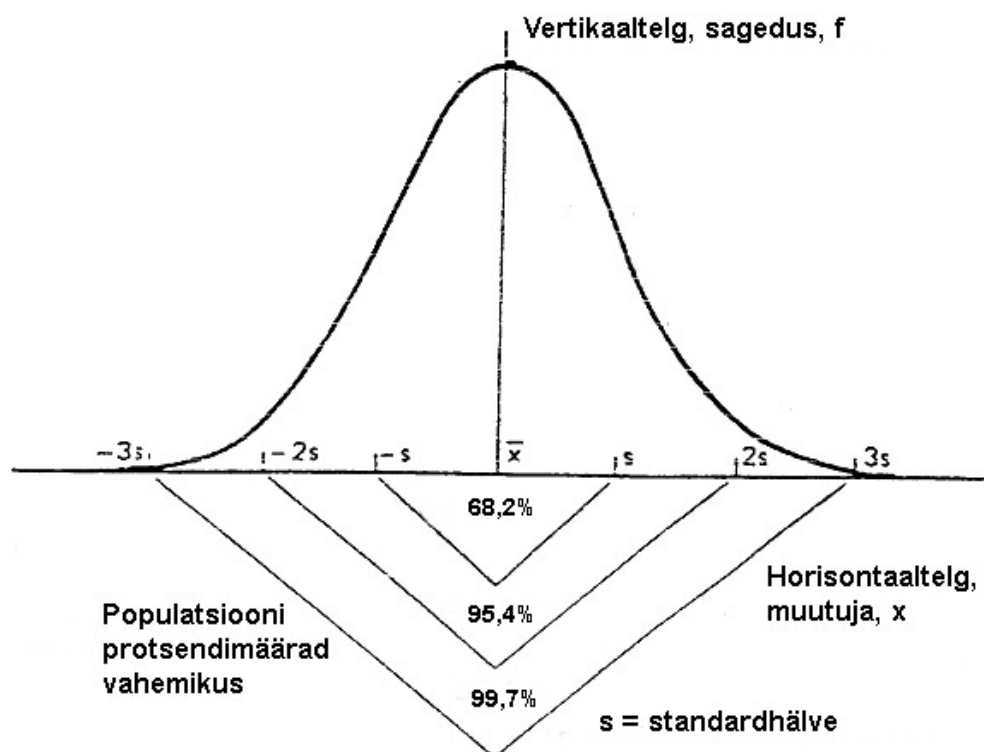
Haiguse **seire eesmärk** on haiguse tunnuste aktiivne otsing. Kui tunnused on leitud, järgneb tegevus. Seire eesmärk võib olla seega, näiteks, demonstreerida haigusvabaduse astet (täielikult

vaba või vaba kindlaks määratud levimuse tasemel või alla selle) või kinnitada, et haigust ei ole uuesti sisse toodud.

Valimimahtude arvutamise meetodid on erinevad sõltuvalt sellest, kas eesmärk on monitooring või seire.

1. Valimimaht hindamaks haiguse levimust suurtes populatsioonides (monitooringu eesmärk)

Kui populatsioonist võetaks korduvalt juhuvalimeid, kasutades ühte ja sama valimi võtmise meetodit, saaksime me iga valimi kohta erineva hinnangu ja võiksime need väärtused kanda graafikule. Tulemuseks on sagedusjaotus, mis näitab, millise sagedusega erinevad hinnangud esinevad. See sagedusjaotus on umbkaudu sama kujuga kui normaaljaotus (joonis 30) ning populatsiooni tegeliku levimuse (mida me tahame hinnata ühest valimist) usalduspiiride hindamiseks saab kasutada normaaljaotuse statistikat.



Joonis 30. Normaaljaotus. Populatsiooni protsendimäärad, mis langevad keskmise teatud vahemikku

Varieeruvuse mõõduks on meil standardvea (SE) valem:

$$SE = \sqrt{p(100 - p) / n} \text{ või } SE = \sqrt{p \times q / n},$$

kus p on levimus protsentides,

n on valimimaht,

$q = (100 - p)$.

Seega, analoogselt standardhällbega, paikneb 68% valimi hinnangutest vahemikus tegelik levimus miinus 1 standardviga kuni tegelik levimus pluss 1 standardviga. Samamoodi, 95 protsenti hinnangutest erineksid tõesest väärtusest vähem kui kahekordne standardviga ja 99% hinnangutest paikneks eeldatavasti tegeliku levimuse (p) 3 standardvea piirides.

Järelikult võime näiteks öelda, et me sooviksime valimimahtu, mille puhul saame olla 95% kindlad, et meie hinnang paikneb tegeliku levimuse (p) 1%lise vea piires.

Pidage meeles: 95% = 2 SE.

See tähendab, me tahame, et 2 SE ei oleks suurem kui 1% või 1 SE ei oleks suurem kui 0,5%.

Teisendades ülemise valemi ümber n -i (valimimaht) arvutamiseks, saame:

$$n = p \times (100 - p) / SE^2 \text{ ehk } n = pq / SE^2.$$

Näide.

Oletame, et meil on kahtlus, et umbes 20 protsendil lehmpopulatsioonist esinevad antikehad brutsella vastu ja me tahame olla 95% kindlad, et meie valimi hinnang saab paiknema tegeliku levimuse 4% sees.

2 SE mitte suuremad kui 4%,
1 SE mitte suurem kui 2%.

Kasutades valemit

$$n = 20 \times (100 - 20)/2^2 = 400$$

on meie valimimaht 400.

Kui meie valimimaht moodustab suure osa kogupopulatsioonist (N), ütleme, et n on suurem kui 10% kogupopulatsioonist, siis kasutame valemit

$$n = \frac{p(100 - p)}{[SE^2 + p(100 - p)/N]}.$$

2. Vajalikud valimimahud, et avastada haiguse olemasolu populatsioonis (seire eesmärk)

Mõnikord on oluline määrata, kas haigus üldse esineb populatsioonis.

Siin ei ole enam probleemiks võtta valim mahus, mis on piisav, andmaks populatsioonis esineva haiguse tegeliku levimuse kehtivat hinnangut, vaid see, et valimimaht oleks piisav (piisavalt väike), leidmaks vähemalt üks haige loom populatsioonis, kui haigus selles esineb.

Tabelis 10 on toodud vajalikud valimimahud, et olla 95/99% kindel, et vähemalt üks loom valimis on haige juhul, kui haiguse levimus on 1%, 5%, 10% või 50%.

Eelduseks on, et olemasolevate haigete loomade minimaalne arv on üks.

Märkus. Kui soovitud usaldusnivoo peab olema saavutatud, tuleb kasutada formaalset juhusliku valiku meetodit, kusjuures valimiüksuseks on isend.

Nende valimimahude (= avastamise tõenäosuse jaoks) arvutamise põhimõte on väga lihtne (A. James, 1998).

Oletame, et nakatunud on 5 looma 50liikmelisest populatsioonist (levimus on 10%). Kui testimiseks valitakse juhuslikult üks loom, on tõenäosus, et ta ei ole nakatunud $45/50 = 0,90$ (90%). See tähendab, et avastamise luhtumise tõenäosus on 0,90 ja avastamise tõenäosus on $1 - 0,90 = 0,10$ (10%).

Kui populatsioonist on testimiseks juhuslikult valitud kaks looma, on tõenäosus, et esimene neist ei ole nakatunud, $45/50$ (= 0,90 (90%)). Kui see loom ei ole nakatunud, on tõenäosus, et teine loom ei ole nakatunud, $44/49 = 0,898$.

Summaarne tõenäosus, et nakatunud ei ole kumbki loom, on $0,90 \times 0,898 = 0,8082$. Avastamise tõenäosus on järelikult $1 - 0,8082 = 0,1918$ (19,18%).

Tõenäosus, et kolmas loom on negatiivne, oleks $43/48 = 0,8958$, summaarne tõenäosus, et kõik 3 on negatiivsed, on $0,90 \times 0,898 \times 0,8958 = 0,7214$. Avastamise tõenäosus on aga $1 - 0,7214 = 0,276$ (27,6%).

Seda protsessi võib korrata, et saavutada nõutud avastamise tõenäosus (tabel 10).

Tabel 10. (Jamesi, 1998, järgi)

Valimimaht	Mitteavastamise tõenäosus (p)	Avastamise tõenäosus (1 - p)
1	$45/50 = 0,90$	0,10
2	$45/50 \times 44/49 = 0,8082$	0,1918
3	$45/50 \times 44/49 \times 43/48 = 0,724$	0,276
4	$45/50 \times 44/49 \times 43/48 \times 42/47 = 0,647$	0,353
5	$45/50 \times 44/49 \times 43/48 \times 42/47 \times 41/46 = 0,5766$	0,4234
6	$45/50 \times 44/49 \times 43/48 \times 42/47 \times 41/46 \times 40/45 = 0,5126$	0,4874
7	$45/50 \times 44/49 \times \dots \times 39/44 = 0,4543$	0,5457
8	$45/50 \times 44/49 \times \dots \times 38/43 = 0,4015$	0,5985
9	$45/50 \times 44/49 \times \dots \times 37/42 = 0,3537$	0,6463
10	$45/50 \times 44/49 \times \dots \times 36/41 = 0,3106$	0,6894
11	$45/50 \times 44/49 \times \dots \times 35/40 = 0,2717$	0,7283
12	$45/50 \times 44/49 \times \dots \times 34/39 = 0,2369$	0,7631
13	$45/50 \times 44/49 \times \dots \times 33/38 = 0,2057$	0,7943
14	$45/50 \times 44/49 \times \dots \times 32/37 = 0,1779$	0,8221
15	$45/50 \times 44/49 \times \dots \times 31/36 = 0,1532$	0,8468
16	$45/50 \times 44/49 \times \dots \times 30/35 = 0,1313$	0,8687
17	$45/50 \times 44/49 \times \dots \times 29/34 = 0,112$	0,888
18	$45/50 \times 44/49 \times \dots \times 28/33 = 0,095$	0,905
19	$45/50 \times 44/49 \times \dots \times 27/32 = 0,0802$	0,9198
20	$45/50 \times 44/49 \times \dots \times 26/31 = 0,0673$	0,9327
21	$45/50 \times 44/49 \times \dots \times 25/30 = 0,056$	0,944
22	$45/50 \times 44/49 \times \dots \times 24/29 = 0,0464$	0,9536

23	$45/50 \times 44/49 \times \dots \times 23/28 = 0,0381$	0,9619
----	---	--------

Seega, saavutamaks 95%st avastamise tõenäosust, peab valimimaht olema 22.

Arvutamine on vaevarikas ja tavaliselt kasutatakse valimimahtude saamiseks tabeleid või selleks ettenähtud arvutitarkvara. (Tabelid: A. James, VEERU, Reading, UK.)

Tabelis 11 olevad väärtused baseeruvad järgmisel arvutusel (Cannon, Roe, 1982), määramaks testimisele kuuluvat valimimahtu (n), et saavutada antud usaldatavus (d) haiguse esinemise kindlakstegemiseks minimaalse levimuse (d/N) juures populatsioonis (N):

$$n = (1 - (1 - \alpha)^{1/d}) \times (N - d/2) + 1,$$

kus N on populatsiooni suurus,

d – positiivsete arv populatsioonis,

n – valimisse valitute arv,

α – soovitud usaldatavus (see on vähemalt ühe positiivse leidmise tõenäosus valimis).

Näide. Meil on 250 seaga kari, mille puhul me tahame olla kindlad, et 99% on vabad haigusest X. Nakatunud karjas tabandab haigus tavaliselt 10% populatsioonist.

Seega on meil: $N = 250$,

$d = 25$,

$\alpha = 0,99$,

$$n = (1 - (1 - 0,99)^{1/25}) \times (250 - 25/2) + 1 = 0,168 \times 238,5,$$

$$\underline{n = 40}.$$

Kui ühelgi valimisse valitud 40 seast ei ole haigust X, saame me olla 99% kindlad, et haigust ei eksisteeri kogu karjas antud levimuse juures.

Allikad.

Cannon, R. M., Roe, R. T. 1982. Livestock Disease Surveys: A field manual for veterinarians. Australian Government Publishing Service, Canberra.

James, A. 1998. Risk Analysis Course. Veterinary Epidemiology and Economics Research Unit (VEERU), University of Reading, UK.

Otte, J. 1991. The design of epidemiological investigations. Bogota, 1911, German Agency for Technical Cooperation (GTZ), not published.

Tabel 11. Haigestunud loomade mitteavastamise tõenäosus α

Levi- mus %	Valimimaht n														
	5	10	15	20	30	40	50	60	75	100	150	200	250	500	1000
0,1	0,9950	0,9900	0,9851	0,9802	0,9704	0,9608	0,9512	0,9417	0,9277	0,9048	0,8606	0,8186	0,7787	0,6064	0,3677
0,2	0,9900	0,9802	0,9704	0,9608	0,9417	0,9230	0,9047	0,8868	0,8606	0,8186	0,7406	0,6701	0,6062	0,3675	0,1351
0,3	0,9851	0,9704	0,9559	0,9417	0,9138	0,8868	0,8605	0,8350	0,7982	0,7405	0,6372	0,5483	0,4718	0,2226	0,0496
0,4	0,9802	0,9607	0,9417	0,9230	0,8867	0,8519	0,8184	0,7862	0,7404	0,6698	0,5482	0,4486	0,3671	0,1348	0,0182
0,5	0,9752	0,9511	0,9276	0,9046	0,8604	0,8183	0,7783	0,7403	0,6866	0,6058	0,4715	0,3670	0,2856	0,0816	0,0067
0,6	0,9704	0,9416	0,9137	0,8866	0,8348	0,7861	0,7401	0,6969	0,6368	0,5478	0,4055	0,3001	0,2221	0,0493	0,0024
0,7	0,9655	0,9322	0,9000	0,8689	0,8100	0,7550	0,7038	0,6561	0,5905	0,4954	0,3486	0,2454	0,1727	0,0298	0,0009
0,8	0,9606	0,9228	0,8865	0,8516	0,7859	0,7252	0,6692	0,6176	0,5475	0,4479	0,2997	0,2006	0,1343	0,0180	0,0003
0,9	0,9558	0,9136	0,8732	0,8346	0,7624	0,6965	0,6363	0,5813	0,5076	0,4049	0,2577	0,1640	0,1043	0,0109	0,0001
1	0,9510	0,9044	0,8601	0,8179	0,7397	0,6690	0,6050	0,5472	0,4706	0,3660	0,2215	0,1340	0,0811	0,0066	
2	0,9039	0,8171	0,7386	0,6676	0,5455	0,4457	0,3642	0,2976	0,2198	0,1326	0,0483	0,0176	0,0064		
3	0,8587	0,7374	0,6333	0,5438	0,4010	0,2957	0,2181	0,1608	0,1018	0,0476	0,0104	0,0023	0,0005		
4	0,8154	0,6648	0,5421	0,4420	0,2939	0,1954	0,1299	0,0864	0,0468	0,0169	0,0022	0,0003			
5	0,7738	0,5987	0,4633	0,3585	0,2146	0,1285	0,0769	0,0461	0,0213	0,0059	0,0005				
6	0,7339	0,5386	0,3953	0,2901	0,1563	0,0842	0,0453	0,0244	0,0097	0,0021	0,0001				
7	0,6957	0,4840	0,3367	0,2342	0,1134	0,0549	0,0266	0,0129	0,0043	0,0007					
8	0,6951	0,4344	0,2863	0,1887	0,0820	0,0356	0,0155	0,0067	0,0019	0,0002					
9	0,6240	0,3894	0,2430	0,1516	0,0591	0,0230	0,0090	0,0035	0,0008	0,0001					
10	0,5905	0,3487	0,2059	0,1216	0,0424	0,0148	0,0052	0,0018	0,0004						
11	0,5584	0,3118	0,1741	0,0972	0,0303	0,0095	0,0029	0,0009	0,0002						
12	0,5277	0,2785	0,1470	0,0776	0,0216	0,0060	0,0017	0,0005	0,0001						
13	0,4984	0,2484	0,1238	0,0617	0,0153	0,0038	0,0009	0,0002							
14	0,4704	0,2213	0,1041	0,0490	0,0108	0,0024	0,0005	0,0001							
15	0,4437	0,1969	0,0874	0,0388	0,0076	0,0015	0,0003	0,0001							
16	0,4182	0,1749	0,0731	0,0306	0,0054	0,0009	0,0002								
17	0,3939	0,1552	0,0611	0,0241	0,0037	0,0006	0,0001								
18	0,3707	0,1374	0,0510	0,0189	0,0026	0,0004									
19	0,3487	0,1216	0,0424	0,0148	0,0018	0,0002									
20	0,3277	0,1074	0,0352	0,0115	0,0012	0,0001									
21	0,3077	0,0947	0,0291	0,0090	0,0008	0,0001									
22	0,2887	0,0834	0,0241	0,0069	0,0006										

23	0,2707	0,0733	0,0196	0,0054	0,0004
24	0,2536	0,0643	0,0163	0,0041	0,0003
25	0,2373	0,0563	0,0134	0,0032	0,0002
26	0,2219	0,0492	0,0109	0,0024	0,0001
27	0,2073	0,0430	0,0089	0,0018	0,0001
28	0,1935	0,0374	0,0072	0,0014	0,0001
29	0,1804	0,0326	0,0059	0,0011	
30	0,1681	0,0282	0,0047	0,0008	
40	0,0778	0,0060	0,0005		
50	0,0313	0,0010			
60	0,0102	0,0001			

Tõenäosus $\alpha = (1 - \text{levimus}) n$, et teatud kindla levimuse juures ei sisalda valim n ühtegi haiget looma, või kui n suurusega valim ei sisalda ühtegi haiget looma, võib tõenäosusega α välistada, et tegelik levimus (populatsioonis) on $\geq$ eeldatud levimusest, või võib tõenäosusega $1 - \alpha$ järeldada, et tegelik levimus on $\leq$ kui eeldatud levimus.

ÜLESANNE

1. Väike broilerifarm veab broilereid 100pealistes gruppides. Teie soovite uurida lindude soolestikku, et avastada koktsidioosi, ja usute, et kui see üleüldse esineb, siis nakatub vähemalt 8% lindudest. Te tahate olla 95% kindel, et koktsidioosi levimus on väiksem kui 8%.

Kui mitut soolestiku proovi tuleb teil uurida?

2. Te võtate proove 100 veisega karjast, et uurida, kas seal esineb anaplasmoosi. Te testite 10 lehma ja nad kõik on negatiivsed.

Milline on eeldatav väikseim juhtude arv karjas?

Milline on eeldatav suurim juhtude arv karjas, kui te olete testinud 50 looma ja nad kõik on negatiivsed?