Matemaatiline statistika ja modelleerimine

Andmed, tunnused, populatsioon ja valim

EMÜ doktorikool
DK.0007

Tanel Kaart

Milleks meile statistika?

- Kokkuvõtete tegemiseks ja ülevaadete saamiseks.
- Huvipakkuvate seoste ja erinevuste tuvastamiseks.
- Tõestamiseks.
- Prognoosimiseks.
- …

Millest inimesed intuitiivselt aru saavad?

- Sellest, kui palju või kui mitu protsenti midagi on

A'la: 1. novembri seisuga 2012 oli piirivalves 1289 töötajat, kriminaalpolitseis 1109 töötajat ja korrakaitsepolitseis 1829 töötajat, või et 29% inimkannatanutega liiklusõnnetuse põhjustajatest aastail 2006-2011 olid joobes.

- Aritmeetilisest keskmisest

A'la: inimkannatanutega liiklusõnnetuse põhjustajatest aastail 2006-2011 olid 72% mehed, keskmise vanusega 35 eluaastat, ning 70% mootorsõiduki-juhid (keskmine juhistaaž 10,6 a).

- Minimaalsest ja maksimaalsest väärtusest

A'la: liiklusõnnetustes hukkunute keskmine vanus aastal 2011 oli 35 aastat (noorim 10 ja vanim 78 aastat).

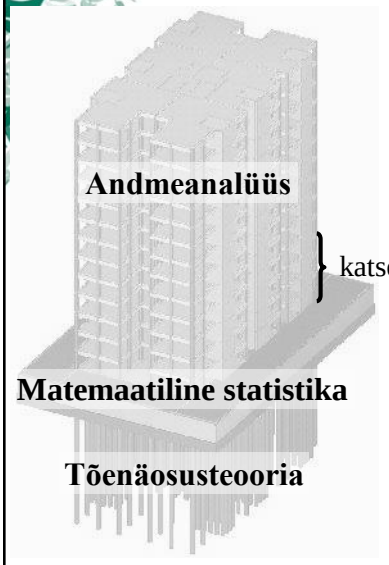
- Korrelatsioonist (= seosest)

A'la: inimeste kuusissetuleku ja elukoha kauguse vahel Tallinnast on tugev negatiivne seos.

Ja millest nad intuitiivselt aru ei saa?

- Negatiivsest korrelatsioonist, korrelatsioonikordaja arvväärtusest
- Mediaanist (kvartiilidest, …)
- Standardhälbest, rääkimata standardveast
- Populatsioonist ja valimist
- Usaldusintervallist
- Statistilisest olulisusest ja p-väärtusest
- …
- Ja sellest, et statistika valetabki alati …

Andmeanalüüsi olemus



Andmeanalüüs teeb teaduslikke järeldusi reaalsete (vaatlustest, katsetest, mõõtmistest pärinevate) andmete põhjal, valides rakendatavad statistikameetodid nii, et need võimalikult hästi andmetega sobiksid.

} katseplaan; andmed ja nende esitus;
kirjeldav statistika;
valim *versus* populatsioon

Matemaatilise statistika tegeleb teoreetiliste andmete $\mathbf{X} = (X_1,\ldots,X_n)$ ja nende funktsioonide $T(\mathbf{X})$ (e statistikute) tõenäosuslike omaduste uurimisega ning statistiliste otsustuste tegemisega.

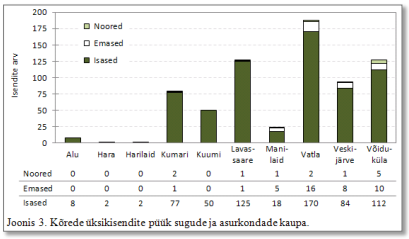
Andmeanalüüsi tüübid

✓ Kirjeldav statistika [*descriptive statistics*] – andmete kokkuvõtlik/ülevaatlik esitamine:

- arvkarakteristikud,
- sagedustabelid,
- joonised.

✓ Analüüsiv statistika [*inferential statistics*] – andmete põhjal üldiste järelduste ja otsustuste tegemine:

- parameetrite hindamine,
- hüpoteeside kontroll,
- mudelite konstrueerimine.



Joonis 3. Kõrede üksikisendite püük sugude ja asurkondade kaupa.

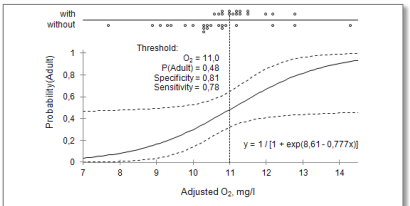


Figure 1. Results of logistic regression analyses predicting the probability of existence of *bufo calamita* adults based on the ponds adjusted water oxygen concentration. Distribution of ponds with and without *bufo calamita* eggs or adults, logistic regression curve (solid line) with 95% confidence interval (dashed lines) are shown. Also the optimal oxygen concentration to distinguish the ponds with and without *bufo calamita* (vertical dotted line), corresponding estimated probability, specificity and sensitivity are presented.

Uuringute tüübid 1

✓ **Valikuuring** – mõõdetakse-kirjeldatakse-analüüsitakse vaid teatud viisil valitud uurimisobjekte, järeldusi tehakse kõigi, ka uuringusse mittekaasatud uurimisobjektide kohta.

Näiteks
- katsepõhised uuringud (põldkatsed, söötmiskatsed jne),
- juhuslikult valitud objektide analüüs,
- küsitlused.

✓ **Kõikne uuring** – mõõdetakse-kirjeldatakse-analüüsitakse kõiki huvipakkuvaid uurimisobjekte.

Näiteks
- rahvaloendus,
- kõigi ülikooli sisseastunute analüüs SAIS-i andmete alusel.

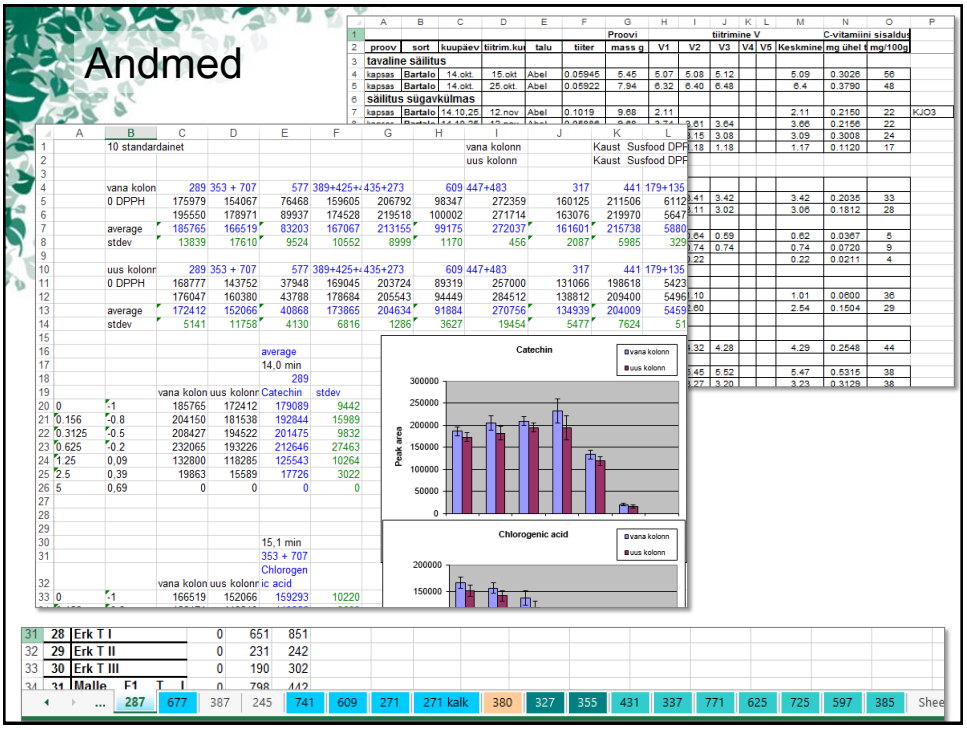
Uuringute tüübid 2

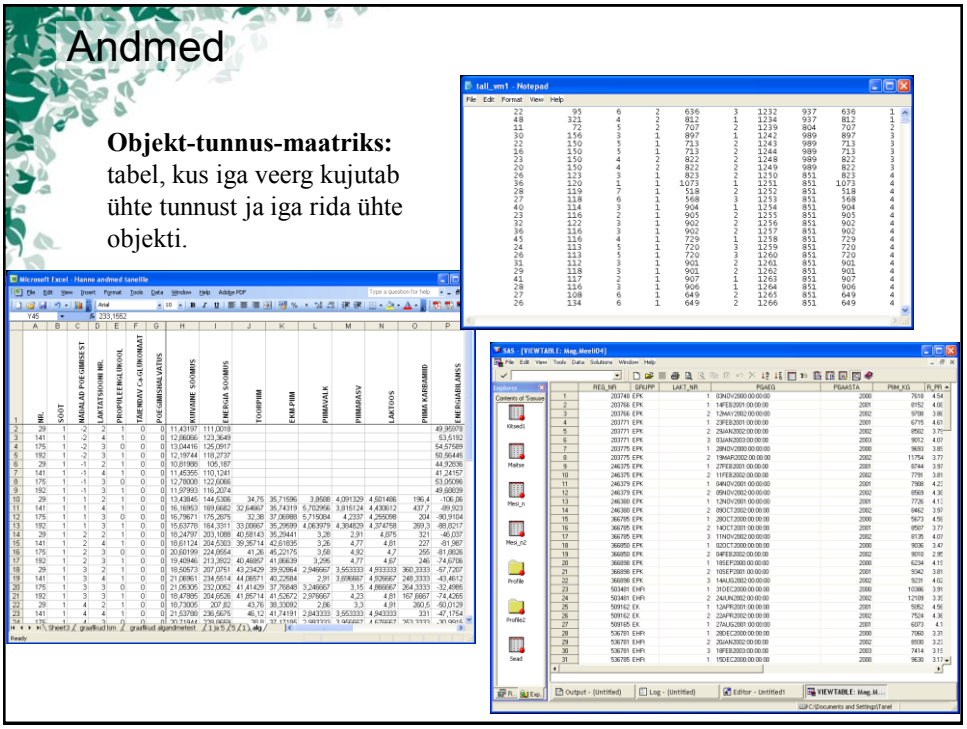
✓ Eksperimendipõhine uuring [*experiment-based study*]

- katse-eelselt määratud ja täpselt kontrollitud tingimused,
- väike ja erinevatele (võrreldavatele) tingimustele vastav sama arv vaatlusi/mõõtmisi,
- analüüsimiseks standardsed statistikameetodid.

✓ Mudelipõhine uuring [*model-based study*]

- mitte ette määratud tingimustes sooritatud vaatlused/mõõtmised,
- sageli suur, segane ja ebaühtlane andmebaas,
- analüüsimeetodid sõltuvad uurija ettekujutusest uuritavaist suurustest ja neid mõjutavaist tegureist, samuti kogutud andmete hulgast ja struktuurist.


Andmed
Catechin
Chlorogenic acid
vana kolonn
uus kolonn
Peak area


Andmed
Objekt-tunnus-maatriks:
tabel, kus iga veerg kujutab ühte tunnust ja iga rida ühte objekti.

Uurimisobjekt

Objekt on uurimisalune ühik, üksikindiviid (näiteks lehm, talu, põllulapp, firma, inimene, punkt metsas või järvel).

Ka samade andmete puhul võib uurimisobjekti valikuks olla mitu erinevat võimalust.

Näiteks:

- kaks pesakonda kutsikaid, ühes 2, teises 6 kutsikat;

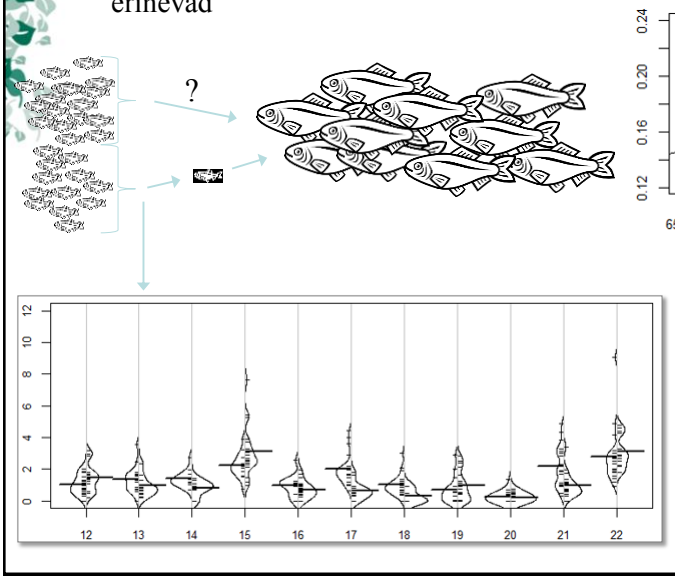
Objekt – kutsikas	
Psk nr	Psk suurus
1	6
1	6
1	6
1	6
1	6
1	6
2	2
2	2
Keskmine psk suurus: 5	

Objekt – pesakond	
Psk nr	Psk suurus
1	6
2	2
Keskmine psk suurus: 4	

- metsa uurides võivad uurimispunktideks olla kas puud või punktid metsas: “80% vaadeldud puudest olid pajud” *vs* “45%-l vaadeldud metsaaladest leidus pajusid”;
- kahjureid uurides võivad uurimisobjektideks olla taimed või põllud: 2,5% taimedest olid kahjustatud *vs* 89,7%-l põldudest leidus kahjureid.

Uurimisobjekt

Teinekord võivad uurimisobjektid olla uuringu erinevatel etappidel erinevad



Objekt ja tunnus

Tunnus on objekti iseloomustav näitaja, mida põhimõtteliselt on võimalik mõõta või vaadelda.

Näiteks päevane piimaand, tõug ja vanus lehmi uurides,
talusid uurides talu aastane sissetulek, töötegijate arv, põllumaa pindala ja kaugus lähimast linnast,
metsapuid uurides võivad mõõdetavateks tunnusteks olla liik, ümbermõõt, kõrgus, vanus jne.

Et statistika näol on tegu matemaatilise distsipliiniga, ei saa siin kuidagi läbi ilma valemiteta.

Traditsiooniliselt esitatakse tunnuste nimed valemeis suurte tähtedega, näiteks *VANUS*, *TÕUG*, *SAAGIKUS*. Sageli kasutatakse ka lühendeid – näiteks tunnuse "lehma aastane tingühikutes mõõdetud väljalüps" võime tähistada sümboliga *X*.

Konkreetsete mõõdetud väärtuste tähistamiseks kasutatakse väikeseid tähti ja soovides täpsustada objekti, kellel/millel see väärtus on mõõdetud, esitakse objekti number alaindeksis:
x_3 on tunnuse *X* väärtus 3. objektil (näiteks 3. lehmal).

Tunnuste tüübid

Mittearvulised e. kvalitatiivsed tunnused [*categorical*]

Järjestustunnuse [*ordinal*] väärtuste vahel on võimalik objektiivne järjestus (hinnangud etteantud skaalal jm).

Näiteks haridus (alg- / kesk- / kõrgharidus / doktorikraad), poegimisraskus, hinnang mulla niiskusele (väga kuiv / kuiv / paras / niiske / liigniiske), hinnang pulli välimusele (niru / normaalne / kaunis), ...

Probleemiks võimalikud subjektiivsed hinnangud (milline pull on kaunis?)!

Nominaalsed tunnused [*nominal*] on mittearvulised tunnused, mille väärtuste vahel ei ole sisulist järjestust.

Näiteks tõug, värvus, farm, kasvukoht, ...

Binaarsed (dihhotoomsed) tunnused on kahe väärtusega nominaalsed tunnused.

Näiteks sugu.

Tunnuste tüübid

Uuriti Lõuna-Eestis asuvaid talusid, kogutud andmed on esitatud järgnevas tabelis.

A suurus (ha)	*B* müügitulu	*C* peamine tegevusala	*D* põllumaa kvaliteet	*E* talupere suurus
25	340000	1	1	5
15	220000	2	2	4
44	700000	1	3	4
12	500000	3	3	2
20	1200000	2	1	2

Tunnused *A* ja *B* on pidevad, *C* nominaalne (kasutatud kodeering: 1-karjakasvatus; 2-viljakasvatus; 3-turism), *D* on järjestustunnus (kasutatud kodeering: 1-väga hea; 2-keskmine; 3-kehv) ning tunnus *E* on diskreetne tunnus.

Tunnuste kodeerimine

Kodeerimine – sõnaliste vastusevariantide arvudega asendamine.

Näiteks tunnuse "arvamus valitsusest" väärtuste sisestamisel võime vastsevariandi "valitsus on hea" asemel sisestada numbri "1", vastusevariandi "valitsus on keskpärane" asemel numbri "2" ja vastusevariandi "valitsus on saast" asemel numbri "3"

✓ Järjestustunnuste kodeerimisel tuleb jälgida, et koodid säilitaksid väärtuste sisulise järjestuse.

✓ Binaarse tunnuse kodeerimisel on eelistatav lihtsaim võimalus, näiteks 0 ja 1 (või ka 1 ja 2, kui see on sisuliselt mõistetavam).

✓ Nominaaltunnuseid ei ole enamasti vaja arvuliseks kodeerida, ja kui kodeerida, siis koodid sisulist tähendust ei oma (loogiline oleks näiteks järjestada väärtused tähestikulises järjekorras).

Puuduvad väärtused

- Plaanides andmeid analüüsida standardse statistikatarkvara (SAS, R, Statistica, …) või mõne tabelarvutussüsteemi (MS Excel, Open Office, …) abil, on mõistlik jätta puuduvale väärtusele vastav lahter tühjaks.
- Arvu 0 puuduva väärtuse tähisena võiks vältida.

?

Piiriületaja vanus	Sugu	Kuupäev	Sigarettide arv
31	M	12.11.2012	2
25	N	30.11.2012	0
0	M	13.10.2012	1
56	M	0	0
29	0	25.11.2012	2

- Aruandes/kokkuvõttes (aga mitte analüüsitavas andmetabelis!) võib selguse mõttes kirjutada puuduva väärtuse kohale punkti („.“) või kriipsu („-“).

Tabel 2. IT-vahendite toel toime pandud kuriteod 2005–2011

Arvutikuritegu	2005	2006	2007	2008	2009	2010	2011	Kokku
KarS § 157[2]	–	–	–	–	3	55	65	123
KarS § 157[1]	–	–	1	–	2	1	1	5
KarS § 206[1]	–	–	–	1	–	–	4	5
KarS § 217[1]	–	–	1	–	1	2	1	5

Allikas: ALIS. 16.01.2012, eriväljavõte

Statistiline andmestik

Märkusi ja soovitusi – ankeedivastused

✓ Andmetabeli igasse lahtrisse sisestatakse üks arv või sõna.

Talu	Tegevusala
1	karjakasvatus
2	karjakasvatus, viljakasvatus
3	viljakasvatus
4	turism
5	viljakasvatus, turism

Talu	Tegevusala		
	karjakasvatus	viljakasvatus	turism
1	1	0	0
2	1	1	0
3	0	1	0
4	0	0	1
5	0	1	1

Statistiline andmestik

Märkusi ja soovitusi

Nimi	Regnr	Lakt.	1.seemen-duse aeg	Seemen-duste arv	Aeg1	Glükoos1 (mg/dl)	Aeg2	Glükoos2 (mg/dl)	Aeg3	Glükoos3 (mg/dl)	Aeg4	Glükoos4 (mg/dl)
ALBI	5030	7	64	6	-12	35,1	10	28,2	37	23,7	64	30,4
SEEVIK	5383	5	74	5	-12	34	10	19,7	37	24,7	.	.
RIIBU	5537	4	89	1	-14	46,3	13	23,3	26	24,9	76	29,7

✓ Kui vähegi võimalik, tuleks mõõtmisi sooritada kõigil objektidel ühesuguste, regulaarsete ajavahemike järel.

(siis pole ka mõõtmise aega näitavaid veerge vaja)

Nimi	Regnr	Laktat-sioon	Periood	Esimese see-menduse aeg	Seemen-duste arv	Mõõtmis-aeg	Glükoos (mg/dl)
ALBI	5030	7	1	64	6	-12	35,1
ALBI	5030	7	2	64	6	10	28,2
ALBI	5030	7	3	64	6	37	23,7
ALBI	5030	7	4	64	6	64	30,4
SEEVIK	5383	5	1	74	5	-12	34
SEEVIK	5383	5	2	74	5	10	19,7
SEEVIK	5383	5	3	74	5	37	24,7
RIIBU	5537	4	1	89	1	-14	46,3
RIIBU	5537	4	2	89	1	13	23,3
RIIBU	5537	4	3	89	1	26	24,9
RIIBU	5537	4	4	89	1	76	29,7

Ajas korduvad mõõtmised

Populatsioon *versus* valim

Üldkogum (**populatsioon**) on reaalselt olemasolev või ka abstraheeritud objektihulk, mille/kelle kohta soovitakse uurimistöö tulemusena sisulisi järeldusi teha

Populatsiooni defineerides piiritletakse ära uuritav objekt ajas, ruumis, katsetingimuste kaudu, …

Näiteks:

- Eesti talud 2012. aastal,
- põllu nr 12154 saagikus (nii minevikus, praegu kui ka tulevikus),
- kõik Eestimaa põllud, millel kasvatatakse talirukist,
- Eesti põhjaranniku jõgedes kudevad lõhed,
- Eesti maatõugu veised,
- Eesti sigade pekipaksuse muutus ajavahemikul 1995-2005,
- kõik antud mündiga teha võidavad kulli/kirja viskamised,
- …

Populatsioon *versus* valim

Valim [*sample*] on teatava eeskirja järgi moodustatud hulk üldkogumisse kuuluvatest objektidest e uurimiseks valitud üldkogumi osa.

Populatsioon	Valim
Eesti vetes kudevad lõhed	Kontrollpüükidel püütud 60 lõhet
Eestis jõudluskontrolli all olevad 1. laktatsiooni EHF-lehmad	Jõudluskontrollikeskuse andmebaasist välja valitud 12000 1. laktatsiooni EHF-lehma
Kolm erineva säilitusainega jogurtipartiid	10 proovi igast jogurtipartiist

Statistika olemus:
- võtame teatud reeglite järgi osa üldkogumist (valimi),
- analüüsime seda ja
- teeme järeldusi kogu üldkogumi kohta!

Populatsioon *versus* valim

Valimi põhjal tehtud järeldused on õiged,
- kui valim on **esindav** e representatiivne, st et uuritava tunnuse väärtuste proportsionaalne jaotus valimis on enam-vähem samasugune kui populatsioonis ning
- kui otsuste tegemisel on järgitud matemaatilise statistika reegleid.

Andmete kogumisel tuleb
- vältida teadlikult üksnes soovitud tendentsi peegeldavate vaatluste registreerimist;
- püüda tagada kõigi uuritavate objektide "võrdne kohtlemine" uuringus mittehuvipakkuvate näitajate osas; kui viimane pole võimalik, tuleks need nn segavad faktorid samuti registreerida võimaldamaks hilisemal analüüsil nende (potentsiaalse) mõju arvesse võtmist.

Kui valim ja populatsioon kattuvad, on tegu **kõikse uuringuga**.
NB! Oma loomult välistab kõikne uuring igasugused teaduslikel alustel tehtavad prognoosid!