Kirjeldav ja tõestav (kinnitav) andmeanalüüs

Kirjeldav analüüs (*descriptive, exploratory data analysis*)	**Tõestav analüüs** (*inferencial, deductive, confirmatory analysis*)
❖ Alustatakse andmetest (andmete kogumisest ilma kindla plaanita ja sidumata end kindla teooriaga). ❖ Järeldused on kirjeldavad ja ei anna nende kehtivuse tõenäosust. ❖ Tulemuste kõnekus sõltub (graafilisest) esitusviisist. ❖ Võib olla kiirem, odavam ja paindlikum. ❖ Tulemuslikkus sõltub: 1) üldistamisoskusest (eelnevatest teadmistest), 2) kasutatud andmete esinduslikkusest, 3) objekti stabiilsusest.	❖ Lähtutakse oletusest (hüpoteesist), mida üritatakse tõestada nullhüpoteesi ümberlükkamise abil. ❖ Tugineb rangele teooriale. ❖ Tulemuseks on olulisustõenäosus, mis võimaldab hinnata, kui kindel võib järeldustes olla. ❖ Nõuab katseplaani järgimist. ❖ Tulemus kehtib vaid etteantud eelduste kehtimisel. ❖ Tulemuslikkus sõltub: 1) hüpoteeside püstitamise oskusest (eelnevatest teadmistest), 2) katse (vaatluse) planeerimise oskusest, 3) andmete esinduslikkusest.

Kus kohast hüpotees tuleb?

Tõestava statistilise analüüsi etapid

1. Statistiline hüpotees.
2. **Valimi** moodustamine, katse või vaatluste korraldamine.
3. Saadakse mõõtmistulemused valimis.
4. Arvutatakse arvulise mõõdiku (test-statistiku) väärtus.
5. Võrreldakse test-statistiku väärtust oodatava väärtusega nullhüpoteesi kehtimise (seose või erinevuse puudumise, varieeruvuse juhuslikkuse) korral.
6. Arvutatakse olulisustõenäosus (*p*) ehk **tõenäosus, et test-statistiku väärtuse selline kõrvalekalle oodatavast tekkis juhuslikult.**
7. Statistiline järeldus **üldkogumi** kohta ja eksimise tõenäosus *p*.

Hypothesis (kreeka k.) — teatrietenduse sisukokkuvõte (mis eelneb etendusele.
Tänapäevases **oletus** tähenduses kasutusel alates Newtonist.

Tõestav ja kirjeldav analüüs

Valimeid kasutatakse ka kirjeldava analüüsi puhul.

Tõestava analüüsi erinevus seisneb olulisustõenäosuse *p* arvutamises.

Olulisustõenäosus näitab **eksimuse tõenäosust** sisuka järelduse tegemisel.

Selle järgi saab hinnata tulemuste tõsiseltvõetavust.

Milleks tõestada?

Iga juhuslik väljavõte üldkogumist on reeglina erinev ja annab erineva tulemuse.

Enamasti soovitakse laiemat järelduse kehtimist kui vaid valimi kohta. Statistilise uurimuse eesmärk on üldised teadmised, mitte vaid mõne üksikmõõtmise tulemus.

Kui suudetakse ära mõõta kõik objektid (üldkogum), mille kohta soovitakse järeldusi teha,

ja mõõtmisvigu kindlasti ei ole,

siis pole tulemuste õigsust vaja tõestada

(test-statistikut ja olulisustõenäosust pole tarvis arvutada).

Paraku enamasti soovitakse järeldada laiemalt, kui on otsesed tulemused.

Vastuväited hüpoteeside tõestamisele

- Juhuslikkuse **nullhüpotees on mõttetu**, kuna on ette teada, et kõik on kõigega seotud, miski pole juhuslik ja alati on andmekogumite vahel mingeid erinevusi. (Käsitlus upub mõjude paljususse.)

- Teaduse olulisem ülesanne on seoseid ja erinevusi **mõõta ja selgitada** seoste ja erinevuste **põhjuseid**, mitte nende olemasolu tõestada. (Kas tasub tegeleda põhjustega, kui nähtuse olemasolu pole tõestatud?)

- Statistiline olulisus ei pruugi näidata **põhjuslikku** olulisust. Statistiliste hüpoteeside tõestamisel on mõtet vaid siis, kui on ära näidatud nende seos praktiliste järeldustega. (Kõigepealt peaks tõestama, et nähtus üldse on olemas, siis tasub uurida selle põhjuseid.)

- Üksiku hüpoteesi tõestamine **üksikuuringus** ei tõesta niikuinii midagi lõplikult. Kindel võib olla alles siis, kui samu tulemusi on saanud mitmed vabad ja omavahel sõltumatud uurijad. (Ei eita hüpoteeside vajalikkust.)

Valim ehk väljavõte

Valim

Valikuline uuring ja olulisustõenäosuse arvutamine on vahend, mitte eesmärk.

Valim — väljavõtt üldkogumist. Valimi uurimine on lihtsam, kiirem ja odavam kui üldkogumi uurimine. Mõnikord on üldkogumi uurimine lihtsalt ebareaalne,
näiteks:

1) supelrandade kõigi liivaterade mõõtmine nende keskmise suuruse võrdlemiseks või

2) kõigi pikslite analüüs detailselt fotolt (objekte liiga palju),

3) kõigi elanike küsitlemine (liiga kallis ja aeganõudev).

Valim, mõisted

Katsete jada ehk **katseseeria** — **mingis suhtes** samadel tingimustel korratud katsed või vaatlused. Võib olla ka valimite seeria.

Sõltuvad valimid — samu tunnuseid on mõõdetud korduvalt samadel objektidel. Näiteks testid enne ja pärast objekti mõjutamist või sama päeva ilmavaatlusandmed ühes vaatlusjaamas ja teises vaatlusjaamas (kui võrreldakse vaid sama päeva andmeid).

Sõltumatud valimid — vaatlused valimites ei ole omavahel seotud. Vaatlusalused **ei moodusta uurija poolt planeeritud paare**.

Kõikne valik — uuritakse tervet üldkogumit.

Valim, mõisted

Tagasipanekuga valik — sama objekt võib korduvalt valimisse sattuda.

Tagasipanekuta valik — sama objekti korduvalt valida ei saa.

Esinduslikkus

Valimi põhjal püütakse teha järeldusi üldkogumi kohta.
Et järeldused oleksid põhjendatud, peab valim üldkogumit hästi **esindama** ehk olema **esinduslik** ehk **representatiivne**.

Millest sõltub esinduslikkus?

1) *Valimi mahust,*
2) *valiku meetodist,*
3) *mõõtmistäpsusest,*
4) *tunnuse varieeruvusest,*
5) *oluliste aspektide esindatusest*.

Täielikult ja **igas suhtes esinduslik** on vaid kõikne valim.
Väljavõtete esinduslikkus on suhteline.

Esinduslik valim kajastab üldkogumit piisava täpsusega.
Milline on piisav täpsus, sõltub väärjäreldustest tulenevatest riskidest.

Üldkogum ja valim

Esinduslik ≠ juhuslik valim.

Esinduslik valim **peaks olema**

- piisavalt **suur** (muidu ei esinda varieeruvust ja juhusliku eksimuse risk on suur);
- **kajastama** üldkogumi seesmist **struktuuri** (klasside vahekorda). Korrapärane ja osaliselt korrapärane (kihiline) juhuslik valim võivad olla esinduslikumad kui lihtne juhuslik valik.
- Kõigil üldkogumi objektidel peaks olema (võrdne) **võimalus** valimisse sattuda.

Valikumeetodi otsustamisel peaksid eesmärgiks olema **tõesed järeldused** üldkogumi kohta.

Kas tõesed järeldused peaksid olema ainueesmärgiks?

Töömahu tuleb ka arvestada ja mõistlikes piires hoida.

Esinduslikkus

Ingliskeelses kirjanduses kasutatakse üldist terminit esinduslikkus harva, sagedamini arutletakse valimi võimaliku hälbimise (*biased sample*) ja kirjutatakse tulemuste hälbimisest (*biased results*).

Valim on hälbinud, kui kõigil üldkogumi objektidelt ei ole võrdne võimalus valimisse sattuda. Näiteks kui välivaatlustel kõnnitakse mööda teeradu, siis radadest kaugemal olevad objektid võivad jääda leidmata (valim on hälbinud). Tulemuseks võib olla objektide sageduse hälbega hinnang.
Paraku lisanduvad hälvetele ka mitmesugused juhuslikud (ettearvamatud) vead.

Vead ≠ hälbed.

Lihtsaim ja kindlaim viis vältimaks valimi hälbimist on kasutada juhuvalimeid.

Paraku ei ole juhuslik valim sageli mitmes muus mõttes tõhus.

Otsuste puu valimi esinduslikkuse tagamisel

Igat ülesannet peaks täitma nii hästi kui võimalik.
Valim peab üldkogumit võimalikult hästi esindama.

Uurija ise teab kõige paremini, millised objektid on kõige paremad üldkogumi esindajad, sest ta on asjatundja.
Mina usun, et parimad esindajad on just need. Aga kui teised ei usu?

Valimi koostamine peaks ideaaljuhul olema **uurija suvast sõltumatu**, sest subjektiivsus on individuaalne, on muutlik ja tuleneb laiemalt levinud arvamusest (usust).

Otsene subjektiivne otsus asendatakse **reeglitega**.
Saadakse regulaarne valim.
Reeglid formaliseerivad valiku ja tagavad protsessi korratavuse.
– Aga miks just ühed reeglid ja mitte teised?

Otsene subjektiivne otsus asendatakse **juhusega**, sest juhus on pime.
Saadakse **juhuvalim.**
– Suurte arvude seadus toimib vaid suurte arvude puhul.
Üksik juhuvaatlus ja väike juhuvalim võib olla vägagi erandlik.
• Üksik jalgpallikohtumise tulemus on vähem esinduslik kui korvpalli tulemus.
• Üksik orienteerumisvõistluse tulemus on vähem esinduslik kui pikamaajooksu tulemus.

Juhuvaliku eelised ja puudused

Juhuslikkus ei ole eesmärk, eesmärk on esinduslikkus.

+ +

Juhuvalik väldib süstemaatilisi vigu ja nihkega hinnanguid.
Võimaldab hinnata valimist arvutatud näitajate usaldusväärsust.

– –

Juhuvalik ei pruugi üldkogumi kõiki osi esindada ja võib olla töömahukas.

Mittejuhuslikud valikumeetodid

Planeeritud meetodid (korrapärased ja ekspertvalimid) — uurija otsustab valimisse hõlmamise reeglid, võtab selles osas vastutuse.

• **Kvootide meetod** — valimi struktuur määratakse enne valimist (meeste ja naiste vahekord esinduskogus).

• **Ekspertvalik** — uurija või keegi teine ekspert otsustab, millised on need kõige tüüpilisemad näidised, mida uurida.

• **Klastervalik —** üldkogum jagatakse gruppideks. Valitakse juhuslikud grupid. Analüüsitakse kõik valitud grupi liikmed.

• **Kihiline valik —** üldkogum klassifitseeritakse enne valimi moodustamist mingite uurija meelest oluliste tunnuste järgi kihtidesse. Kihtide piires kasutatakse enamasti juhuslikku valikut. Kihtide moodustamiseks **on vaja eelteadmisi** uuritava üldkogumi kohta.

• **Süstemaatiline** ehk **regulaarne valik —** objekte võetakse valimisse kindla reegli järgi (kindla vahega). Objektid peavad olema mingil moel järjestatud, ruumiobjektide puhul näiteks koordinaatide järgi.

Planeeritud valiku eelised ja puudused

+ +

Planeeritud valikuga saab tagada üldkogumi kõigi osade (piisava) esindatuse. Võib olla lihtsam teostada.

– –

Planeeritud valimitest saadud hinnangute puhul on suurem võimalus, et need on nihkega (hälbega) kui samasuurest juhuvalimist saadud hinnangud. Valimi mahu suurendamine ei pruugi nihet kõrvaldada.

Valim geograafilistest andmetest

Näited uurimisteemast: muld ja mõisakeskmete paiknemine

Temaatiline esinduslikkus

Need teemad, mille mõju uuritakse — Põhiliste muldade esindatus

Need teemad, mille mõju on arvatavasti oluline — Mõisate asutamisaegade esindatus

Ruumiline esinduslikkus

1. Ruumi kaetus — Põhja- ja Lõuna-Eesti esindatus
2. Tiheduse varieeruvuse esindatus — Ida-Virust ja Järvamaalt erinev arv vaatlusi, sest mõisate tihedus on erinev
3. Kaugusvahemike esindatus — Olgu nii suuri kui väikseid vahemaid
4. Lähivaatluste omavaheline sõltuvus — Lähestiku paiknevad vaatlused rikuvad valimi esinduslikkust (sama muld, seotud majandusüksus)

Kas põtrade keskmise tiheduse ja üldarvu arvutamiseks Ida-Viru maakonnas tuleks vaatlusalad paigutada vaid metsa või ka lagerabasse ja põldudele?

Statistilised hüpoteesid

Statistiline hüpotees — üldkogumi kohta tehtud **oletus**.
Koosneb kahest või enamast alternatiivsest oletusest:

1) oletus kehtib — sisukas hüpotees,
2) oletus ei kehti — nullhüpotees.

Nullhüpoteesi ei tõestata, nullhüpoteesi juurde jäädakse.
Nullhüpotees on teoreetiline abivahend, mis tegelikkuses niikuinii ei kehti.
Juhuslikkuse korral oodatavat on võimalik arvutada, tegelikku jaotust me ei tea.

Esimest liiki viga — mittekehtiva sisuka hüpoteesi vastuvõtmine.

Vale järeldus on ränk viga. Nagu süütu inimese karistamine.

Teist liiki viga — mittekehtiva nullhüpoteesi juurde jäämine.

Järelduse tegemata jätmine ei ole nii riskantne. Ehk õnnestub oletust teinekord tõestada. Nagu süüdlase karistamata jätmine. Ehk inimene parandab ennast või õnnestub teinekord süüd tõestada.

Statistiline olulisus ja usalduspiirid

Olulisustõenäosus (*p*) — **eksimuse tõenäosus sisukat hüpoteesi eelistades**.

Olulisuse nivoo (α) ehk **riskitase** ehk riskiprotsent ehk riski **kriitiline** tase ehk uurija seatud **tõke** esimest liiki vea tõenäosusele ehk tase, mille juures saab veel sisuka hüpoteesi vastu võtta

Usaldusnivoo — tõese otsuse tõenäosus (= 1 – α). Eksimuse vastandsündmus on tõene otsus.

Usaldusintervall ehk usalduspiirkond ehk usaldusvahemik sisaldab etteantud tõenäosusega parameetri tegelikku väärtust.
Miks me seda tegelikku väärtust kindlalt ei tea? (valikuline uuring)

Usalduspiirid — usaldusintervalli otsapunktid.

Loodusteadustes kasutatakse esimest liiki eksimuse tõenäosuse vaikimisi piirväärtust (olulisuse nivood) 0,05. Seega lubatakse **esimest liiki vea tõenäosust 0,05 ja teist liiki vea tõenäosust 0,95.**

Miks kasutatakse nii suurt riskide erinevust?

Sest esimest liiki eksimuse tagajärjed on enamasti rängemad.

Kahepoolne ja ühepoolne hüpotees

Kahepoolse hüpoteesi korral on mõlemat pidi erinevus võimalik. Ei ole ette teada, kas suurem on A või B.

Ühepoolse puhul on teistpidi erinevus ette välistatud või ei paku huvi.

Enamikus ülesannetes tahetakse teada, kas valimis mõõdetud erinevuse järgi saab teha järelduse, et need valimid on võetud erinevatest üldkogumitest. Ka samast üldkogumist korduvalt väljavõtteid tehes võib saada iga kord erineva tulemuse. Seega kui küsitakse erinevuse kindlust, tuleks lähtuda kahepoolsest hüpoteesist.

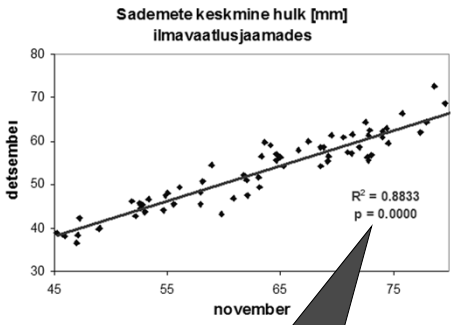
Valimi kirjeldamise meetodid
ja
valimite erinevuste mõõtmise meetodid

On loetletud teooria raamatus ja 2013. a loengu märkmetes.

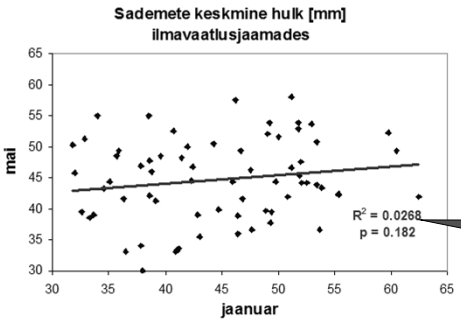
Nende kasutamise harjutamiseks on näidisülesanded.

Seose tugevus ja olulisus I

Korrelatsiooniväli ja lineaarse seose graafik vaatlusjaamades mõõdetud paljuaastase keskmise sademete hulga vahel.



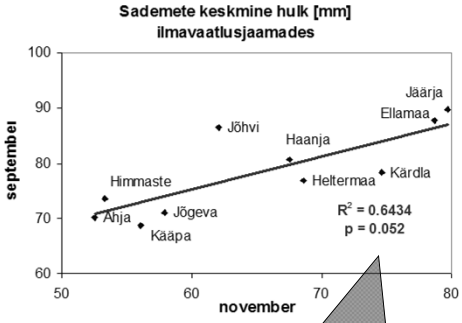
Tugev ja statistiliselt oluline seos



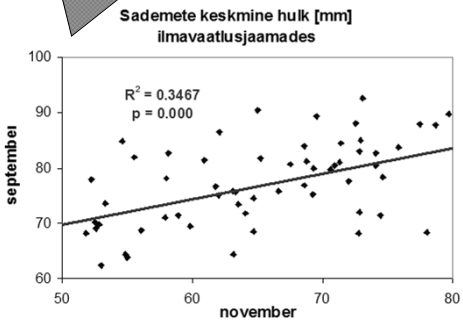
Nõrk ja statistiliselt ebaoluline seos

Seose tugevus ja olulisus II

Korrelatsiooniväli ja lineaarse seose graafik.



Nõrk, aga siiski statistiliselt oluline seos



Üsna tugev, aga statistiliselt ebaoluline seos. Mille poolest see erineb kõrval kujutatud seosest?

Väikestest valimitest saadakse näiliselt tugevad aga vähe usaldatavad seosed

Võiks meelde jääda

- Tõestav analüüs lähtub oletusest ehk hüpoteesist.
- Olulisustõenäosus on eksimise tõenäosus sisuka oletuse õigsuse uskumisel (ning kasutatud meetodi eelduste ja valimi esinduslikkuse korral).
- Tõestav analüüs ei ole ainuvõimalik lähenemisviis probleemidele.
- Valimi esinduslikkus on vahend tõeste järeldusteni jõudmiseks.
- Katse planeerija peab otsustama, mis mõttes esinduslikkus on oluline. Juhuslik valim ei ole igas mõttes.
- Geograafiliste andmete puhul peaks jälgima nii valimi temaatilist kui ka ruumilist esinduslikkust.
- Vale järelduse tegemise (esimest liiki eksimuse) tagajärjed on enamasti rängemad kui järelduse tegemata jätmise tagajärjed.
- Loodusteadustes kasutatakse esimest liiki eksimuse tõenäosuse vaikimisi piirväärtust 0,05.
- Statistilise seose tugevus ja olulisus on kaks erinevat asja.