Kursus: Mitmemõõtmeline statistika

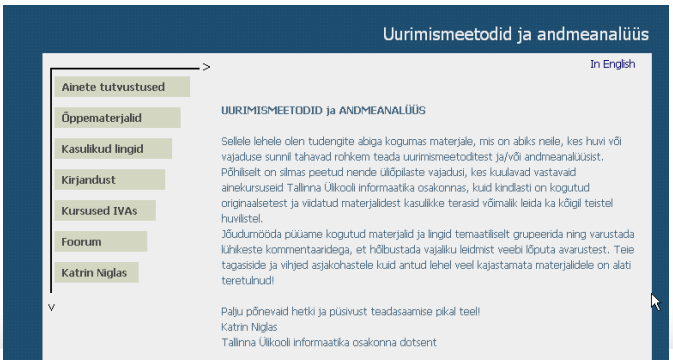
Seminar I: Sissejuhatus ja kordav ülevaade andmeanalüüsi põhitõdedest.

Õppejõud: Katrin Niglas
PhD, dotsent
informaatika instituut

TALLINNA ÜLIKOOL

Kursuse materjalid ...

- Kursusest osavõtjatele: IVA's (Mitmemõõtmeline statistika K2009)
- Osavõtjatele ja kõigile teistele huvilistele: www.tlu.ee/~katrin/



TALLINNA ÜLIKOOL

Mis ja kuidas: vt Aineprogramm!

- Kursuse maht 3 AP e 120 töötundi jaguneb:
  40 tundi auditoorset tööd (loeng-seminar-praktikum)
  40 tundi iseseisvat õppimist (õppematerjalid, õpikud, ...)
  40 tundi arvestusliku koduse töö tegemiseks
- Kursus eeldab andmeanalüüsi põhitõdede valdamist
- Kohal käimist ei kontrollita
  st kursuse võib läbida ka aineprogrammi järgi iseseisvalt õppides
- Arvestuse saamiseks:
  I. iseseisev praktiline töö (eeldab kõigi programmis olevate meetodite praktilist rakendamist andmete analüüsimiseks)
  II. test avatud küsimustega, kus kontrollitakse materjali sisulist mõistmist ja analüüsi tulemuste tõlgendusoskust

TALLINNA ÜLIKOOL

Millest sõltub analüüsimeetodi valik?

I. Küsimuse tüübist

e mis tüüpi on küsimus, millele tahame analüüsiga vastust saada – nt Kas kolm gruppi **erinevad**? Kas kaks nähtust on **seotud**? Kas katsealuseid saaks andmete põhjal **grupeerida**?

II. Andmete tüübist

Kas nimi-, järjestus, arv- või binaarsed tunnused

III. Sihtrühmast

Kui suurt teadlikkust statistiliste meetodite osas võib eeldada? Milline esitlusviis on selle rühma puhul köitev ja sobilik?

TALLINNA ÜLIKOOL

Tunnusetüübid /ka skaalatüübid või andmete tüübid/

Võtmeküsimused:
väärtuste järjestatavus ja skaalavahemike võrdsus!

- Nimitunnused (nt rahvus)

! Nimitunnusel ei ole väärtused üheselt järjestatavad, järjestustunnusel on!

- Järjestustunnused (nt haridustase)

! Järjestustunnusel ei ole väärtuste vahemikud võrdsed, arvtunnusel on!

- Arvtunnused /ka intervalltunnused/ (nt laste arv)
  - Arvtunnused väheste erinevate väärtustega
  - Arvtunnused paljude erinevate väärtustega

! Binaarsel tunnusel on ainult kaks väärtust ja seega järjestamise ja vahede võrdsuse probleemi ei teki!

- Binaarsed tunnused (nt sugu)

TALLINNA ÜLIKOOL

Mõisted

- objekt – tunnus – väärtus ; skaala
- parameetrilised ja mitteparameetrilised meetodid
- sõltuv tunnus
- sõltumatu tunnus
- kirjeldav ja üldistav statistika
- mitmemõõtmeline statistika

TALLINNA ÜLIKOOL

Kuidas oma andmeid kokku võtta?

Struktureeritud andmete esmaseks kokkuvõtuks ning ülevaatlikuks analüüsiks saab kasutada **kirjeldava statistika** meetodeid, mis võib jagada kolme suurde rühma:

- Sagedustabelid (sh risttabelid)
- Arvnäitajad
- Arvjoonised e diagrammid

TALLINNA ÜLIKOOL

Kuidas teada saada, mida võib valimi põhjal väita üldkogumi kohta?

Selleks, et teada saada, milliseid üldistusi (ja kas üldse) saab valimi põhjal üldkogumi kohta teha, saab peale kogutud andmete esmast analüüsi kasutada **üldistava statistika** meetodeid, mis võib jagada kahte suurde rühma :

- Vahemikhinnangud e usaldusintervallid
- Statistilised olulisustestid

PS! Samu meetodeid saab kasutada ka eksperimentaalsetes uuringutes, et kontrollida, kas saadud erinevused jms on tekkinud gruppide erineva mõjutamise tulemusena või võivad olla juhuslikud.

TALLINNA ÜLIKOOL

Statistilise olulisustesti põhisammud:

E I: Analüüsisin olemasolevaid andmeid kirjeldava statistika meetodite abil ning leidsin midagi „huvitavat“ (nt. erinevuse või seose jne)

E II : Tekkis küsimus: „Kas võib üldistada?“

Ü → V → Ü (v eksperimentaalne disain)

I. **Õige olulisustesti valik** (lähtuvalt probleemist ja andmetüübist)

II. **Valitud olulisustesti eelduste kontroll:** **ei** / **jah**

III. **Hüpoteesid:** sisukas hüpotees H_1:

nullhüpotees H_0:

Olulisuse nivoo α

(„Kui väike peab olema H_0 kehtimise tõenäosus, et me võiks ilma suurema riskita ta mittekehtivaks tunnistada?“)

IV. **Arvutused**

eesmärgiks hinnata H_0 kehtimise tõenäosust p

(„Kui suur on tõenäosus, et olukorras, kus H_0 kehtib, tekkis valmis olnud erinevus v seos juhuse tõttu?“)

(*p = Sig = olulisuse tõenäosus*)

V. **Otsus tulemuse kohta**:

$p > \alpha$ H_0 jääb kehtima - statistiliselt mitte oluline (ei üldista)

$p \leq \alpha$ H_1 tõestatud - statistiliselt oluline (võib üldistada)

VI. **Järelduse sõnastamine**

Eelnevast tuttavad meetodid: meetodi valik lähtuvalt küsimusest ja andmete tüübist

Milline analüüsimeetod valida?	**Parameetrilised meetodid** (eeldus: arvtunnused)	**Mitteparameetrilised meetodid** (järjestus- või nimitunnused aga ka arvutunnused)
1 grupp (keskmine tase/ osakaal)	K.st: $\bar{x}$, s, jne Ü.st: vahemikhinnangud (μ, σ)	K.st: sagedustabel, % Ü.st: vahemikhinnangud
2 gruppi ERINEVUSED	K.st: $\bar{x}_1$ $\bar{x}_2$ Ü.st: t-test	K.st: risttabel Ü.st: χ^2-test
3 või enam gruppi ERINEVUSED	K.st: $\bar{x}_1$ $\bar{x}_2$ $\bar{x}_3$ … Ü.st: ANOVA	K.st: $\bar{x}_1$ $\bar{x}_2$ $\bar{x}_3$ … Ü.st: Kruskall-Wallise test
2 või enam tunnust SEOSED	K.st: Pearsoni r (korrelatsioonikordaja) Ü.st: H_0: $r_{ÜK}=0$	K.st: Spearmani ρ Ü.st: H_0: $\rho_{ÜK}=0$ K.st: risttabel Ü.st: χ^2-test

NB! Üks sõltuv ja/või üks sõltumatu tunnus!

TALLINNA ÜLIKOOL

Statistilise olulisuse sisuline tõlgendamine elulisse konteksti

		VALIM	
		suur	väike
ERINEVUS SEOS	suur	Stat.olulisus: +	Stat.olulisus: − (?)
		Elul.olulisus: +	Elul.olulisus: + (!?)
	väike	Stat.olulisus: + (?)	Stat.olulisus: −
		Elul.olulisus: − (!?)	Elul.olulisus: − (!?)

TALLINNA ÜLIKOOL

Efekti e mõju suurus eksperimendis/valimis (*effect size*)

- Probleem: suurte valimite korral on väga väiksed erinevused statistiliselt olulised ja vastupidi, väikeste valimite korral võib suhteliselt suur efekt osutuda statistiliselt mitteoluliseks!
  Kas statistiliselt oluline tulemus ikka on sisuliselt/eluliselt tähenduslik?
- Lahendus: koos olulisustesti tulemustega esitatakse ka sobiv seosekordaja, mida antud kontekstis nimetataksegi **efekti suuruse näitajaks.**
  PS! Kui uuritav nähtus on mõõdetud ühikutes, mis on lihtsalt tajutavad ja eluliselt tähenduslikud (nt. treeningtundide arv nädalas), siis piisab efekti suuruse kirjeldamiseks ka gruppide vahelise erinevuse ära toomisest!
- Seega, eristatakse:
  standardiseeritud efekti suuruse näitajad
  (nt seosekordajad, standardiseeritud regressioonikordajad)
  standardiseerimata efekti suuruse näitajad
  (nt keskmiste erinevus, standardiseerimata regressioonikordajad)

TALLINNA ÜLIKOOL

Efekti e mõju suurus eksperimendis/valimis - enamkasutatud näitajad

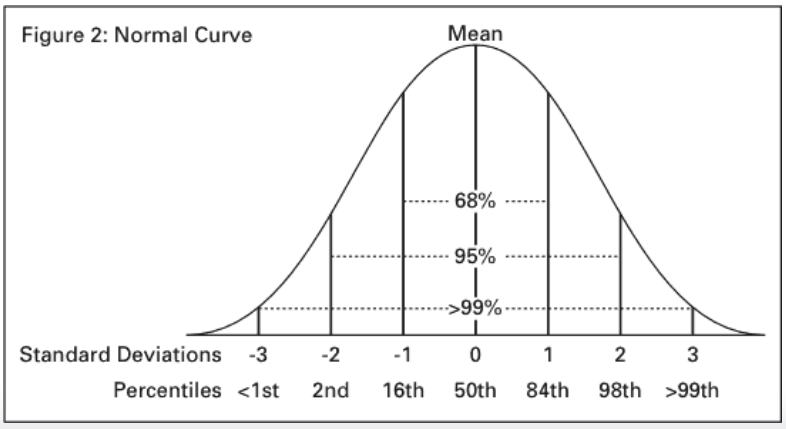
- Arv- ja binaartunnused: Pearsoni r, r^2=d
  sotsiaalteadustes:
  ≈0.2 "väike" efekt, ≈0.4 "keskmine" efekt and ≈0.6 "suur" efekt
- Kahe grupi keskmised (t-test): Coheni d

  sotsiaalteadustes:
  ≈0.3 "väike" efekt, ≈0.5 "medium" efekt and ≈0.9 "suur" efekt
- Rohkem kui kaks gruppi (ANOVA): Coheni f^2, Eta^2, partial Eta^2
- Sagedused (Hii^2-test): Crameri Phi ja V
- Kaks binaarset tunnust: riskisuhe (*Odds ratio*)

NB! Vaata veebist – leiad ka kalkulaatoreid!

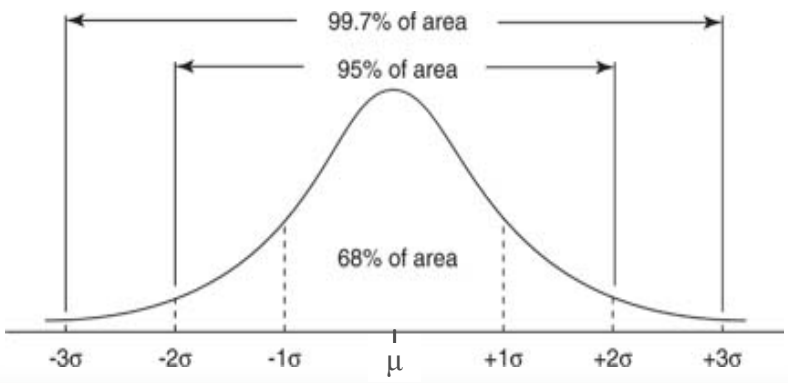
TALLINNA ÜLIKOOL

Proportsioonid normaaljaotuskõvera all



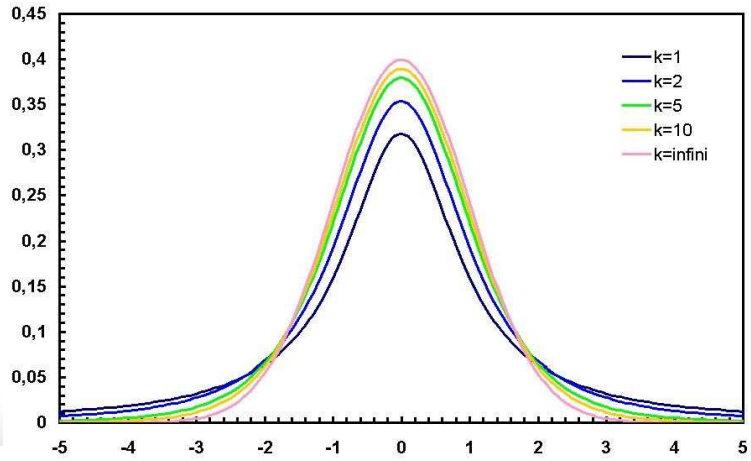
TALLINNA ÜLIKOOL

Proportsioonid normaaljaotuskõvera all



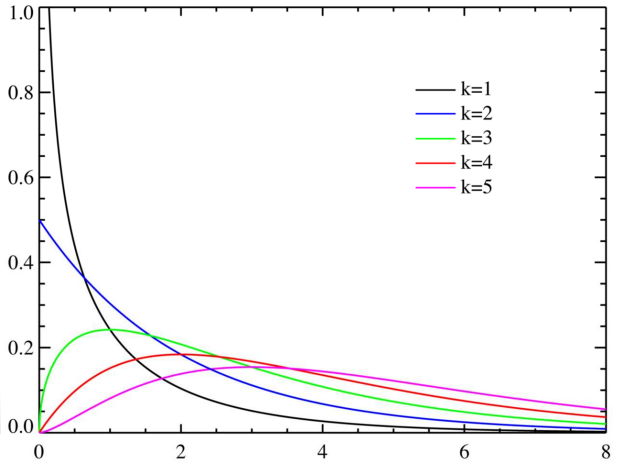
TALLINNA ÜLIKOOL

t-jaotuste pere

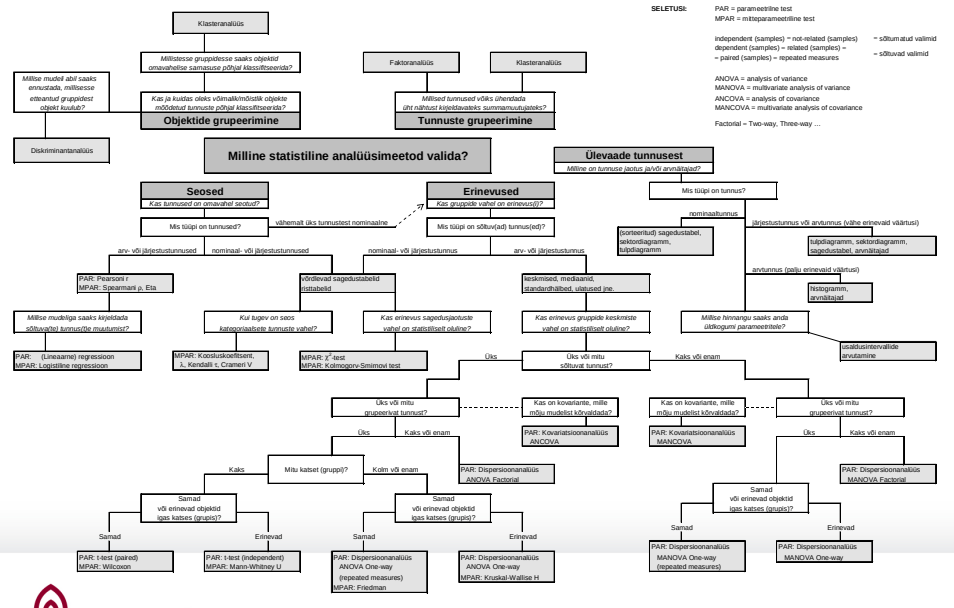


TALLINNA ÜLIKOOL

hii-ruut jaotuste pere



TALLINNA ÜLIKOOL



TALLINNA ÜLIKOOL