Kirjeldav ja tõestav (kinnitav) andmeanalüüs

Kirjeldav analüüs (*descriptive, exploratory data analysis*)

- Alustatakse andmetest (andmete kogumisest ilma kindla plaanita ja sidumata end kindla teooriaga).
- Järeldused on kirjeldavad ja ei anna nende kehtivuse tõenäosust.
- Tulemuste esitusviis on olulisem.
- Võib olla kiirem, odavam ja paindlikum.
- Tulemuslikkus sõltub:
  1) üldistamisoskusest (eelnevatest teadmistest),
  2) kasutatud andmete esinduslikkusest.

{Millest sõltub esinduslikkus?}

Selles loengus käsitletakse kirjeldava analüüsi meetodeid.

Tõestav analüüs (*inferencial, deductive, confirmatory analysis*)

- 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.

1

Kirjeldava uuringu põhilised viisid ja vahendid

Ühe andmekogumi kirjeldamine
keskmiste ja **variatsiooninäitajate** arvutamine ja esitamine,
jaotuste iseloomustamine, kirjeldamine,
vaatluste **järjestamine** mingi tunnuse järgi (ordinatsioon),
tüüpiliste näidiste leidmine,
vaatluste ja tunnuste **klassifitseerimine**.

Andmekogumite võrdlemine
keskmiste ja varieeruvuse näitajate võrdlemine,
jaotuste võrdlemine ja sarnasuste arvutamine.

Seoste kirjeldamine (tunnuste võrdlemine)
korrelatsioonikordajad, seosed aja ja ruumiga (trend ja tsüklilisus),
vastavus (*correspondence*) nominaalsete muutujate korral.

Seose üldistamine ja formaliseerimine (modelleerimine)
regressioon jt statistilised **mudelid**,
klassifitseerimine (klassifikatsioonisüsteemi loomine),
näidiste (tüüpide) komplekti (näidistebaasi) moodustamine.

Statistilisest modelleerimisest on omaette loeng

Tulemuste visualiseerimine ja visuaalne analüüs
graafikute ja kaartide võrdlemine.

Iga analüüsitüübi jaoks on omad vahendid

2

Ühe andmekogumi kirjeldamine

Keskmised
1) aritmeetiline keskmine.
2) kaalutud keskmine (kui vaatlused ei ole võrdsed, esindavad nähtuse erinevat mahtu või on erineva usaldusväärsusega). Näiteks riigi metsasuse arvutamine maakondade metsasuste (osakaalu) andmetest.
3) geomeetriline keskmine (kui olulised on kordsed erinevused).
4) mood (asümmeetrilise jaotuse või nominaalse tunnuse korral).
5) mediaan (tundmatu ebakorrapärase jaotustüübi korral).

Varieeruvuse ehk väärtuste hajuvuse näitajad
1) dispersioon ehk keskmine **ruuthä**
2) standardhälve ehk **ruutkeskmine**
3) varieeruvuse ulatus (teoreetiline j
maksimum, haare, kvantiil

Väärtus muutuja arvteljel, millest ühele ja teisele pole jääb kindel osa vaatlustest.
Mediaan on 50% kvantiil.
Alumisest **kvartiilist** väiksemaid väärtusi on 25%
ja ülemisest suuremaid väärtusi on 25%

Mis on kvantiili mõõtühik?

Sagedused
1) absoluutsagedus.
2) suhteline sagedus ehk osakaal (mis on tervik, mis on ühik?).
3) sagedusjaotus (mitme klassi puhul).

3

Kvartiilid ehk veerandi kvantiilid

Arvuderea
1; 2; 3; 4; 5; 6; 7; 8; 9; 10; 11; 12
alumine kvartiil on **3,5;** ülemine kvartiil **9,5**; keskmine on nende vahel = **6,5**

Kas keskmine võib olla suurem kui ülemine kvartiil?
Kas keskmine võib olla väiksem kui alumine kvartiil?

Arvude 1; 2; 3; 4; 5; 6; 7; 10; 100; 200; 500; 10000
alumine kvartiil on **3,5** ja ülemine kvartiil **150**
keskmine = **903,1667** on kõrgem kui ülemine kvartiil

Arvude 1; 100; 101; 102; 103; 104; 105; 106; 107; 108; 109; 110
alumine kvartiil on **101,5** ja ülemine kvartiil **107,5**
keskmine = **96,3 on neist väiksem.**

4

Standarhälve ja standardviga

Dispersioon ehk keskmine ruuthälve kirjeldab vaatluste varieeruvust tunnuse ühikute ruutudes.

Üldkogumis

$$\sigma^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{N}$$

Ühe vaatluse puhul dispersioon = 0

Valimi järgi hinnates

$$s^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n-1}$$

Ühe vaatluse järgi ei saa dispersiooni hinnata. Nulli jagamine nulliga

Standardhälve kirjeldab vaatluste varieeruvust tunnuse ühikutes.

$$\sigma = \sqrt{\sigma^2}$$

$$s = \sqrt{s^2}$$

Standarviga kirjeldab hinnangute oodatavat varieeruvust, s.o. hinnangu täpsust.

Kõiksel analüüsil on kõik objektid mõõdetud ja viga ei ole.

Tunnuse enda varieeruvus ja mõõtmisvead on erinevad asjad.

Keskmise standardviga valimi järgi hinnates

$$SE = \frac{s}{\sqrt{n}}.$$

5

Milleks ja kuidas klassifitseerida?

Milleks klassifitseerida?

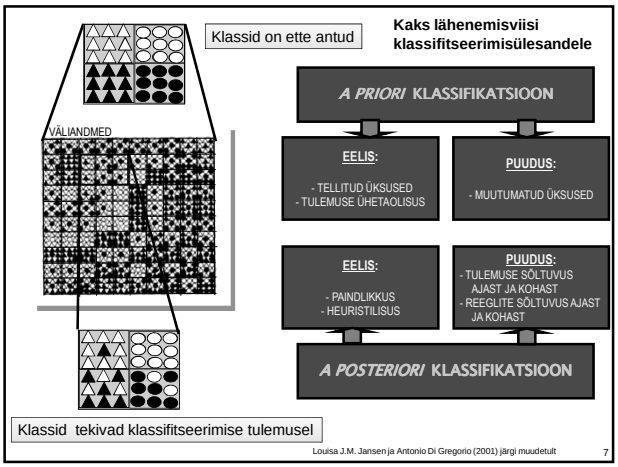
Tunnetuslik vajadus. Nähtuste lõputu mitmekesisus oleks muidu haaramatu.

Andmetöötluslik vajadus. Vaatluseid on palju, kõiki ei jõua kõigega võrrelda, rühmitamise järel piisab rühmakuuluvuse määramisest ja rühmade võrdlemisest.

Kuidas määrata tundmatu objekti klassikuuluvus?
Numbriliste väärtuste prognoosimisel samad võimalused
Klassid on ette antud

1. **Küsida eksperdilt** (asjatundjalt või ekspertsüsteemilt)
2. **Võrrelda objekti ja klassikirjeldustega**
3. **Kasutada otsustamise reegleid** (määramistabelit)
4. **Võrrelda olemasolevate näidistega** (muuseumisäilikutega, andmebaasikirjetega, juhtumitega)
5. Arvutada **mudelite** või ekspertsüsteemide abil.

6



Louisa J.M. Jansen ja Antonio Di Gregorio (2001) järgi muudetult

7

Statistilise seose kirjeldamine

Meetodid:

- **jaotuste võrdlemine** (nominaalsed andmed),
- **korrelatsioon ja astakkorrelatsioon** (järjestatavad),
  - **korrelatsioonimaatriks ja korrelogramm,**
- **regressioonimudelid, s.h. globaalselt ja lokaalselt,**
  - **trend** (regressioonioon aja- või ruumikoordinaatidega),
  - **lokaalselt sobitatud regressioonid** (ka libisev keskmine),
- **näidiste järgi järeldamine.**

Põhilised tulemuste esitusvormid:

- **tabel** (näiteks korrelatsioonimaatriks),
- **joonis** (näiteks korrelogramm),
- **valem** (regressioonivõrrand),
- **tekst,**
- **esitlus multimeedia vahenditega.**

8

Kovariatsioon ja korrelatsioon

Kovariatsioon — hälvete korrutiste keskmine.

Kaks muutujat x ja y

$$\mathrm{cov}(x,y) = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Mille hälbed millest? Kummagi muutuja väärtuste hälbed sama muutuja keskmisest.

Mis on kovariatsiooni ühik? Muutujate ühikute korrutis.

Kas kovariatsiooni väärtus sõltub muutujate mõõtühikutest? Jah, sõltub.

Millal on kovariatsioon positiivne, millal negatiivne, millal null? Null, kui kui muutujate vahel ei ole lineaarset seost; positiivne, kui muutujatel on kalduvus muutuda samas suunas; negatiivne, kui ühe muutuja suurenedes teine enamasti väheneb.

Pearsoni lineaarne korrelatsioonikordaja — standardhälvetega normeeritud kovariatsioon.

$$R = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Milline on korrelatsioonikordaja muutumisvahemik?

Miks kovariatsiooni normeeritakse? Võrreldavus.

Mida korrelatsioonikordaja näitab? Vastastikust seost.

Milliste tunnuste puhul saab korrelatsioonikordajat arvutada, milliste puhul on mõtet arvutada? Arvutada saab numbritest, mõtet on arvutada, kui ruuthälvetel on mõtet.

Kas korrelatsioonikordaja abil saab tõendada seose olemasolu (statistilist olulisust)? Otse ei saa, vaatluste arvu on ka tarvis teada.

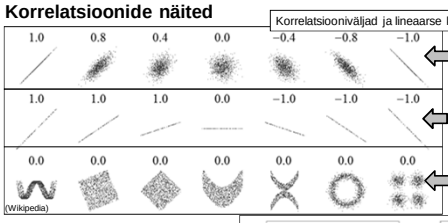
Mis on korrelatsioonikordaja ühik? Ühikuta.

Determinatsioonikordaja — korrelatsioonikordaja ruut R^2. Väljendab mudeliga (korrelatsioonikordaja puhul lineaarse mudeliga) seletatud dispersiooni suhet muutuja kogudispersiooni.

9

Korrelatsioonide näited

Korrelatsioonivälja d ja lineaarse korrelatsiooni tugevus



Korrelatsiooni tugevus esindab punktide hajuvust.

Graafiku tõusunurk ei näita seose tugevust (va seose puudumisel).

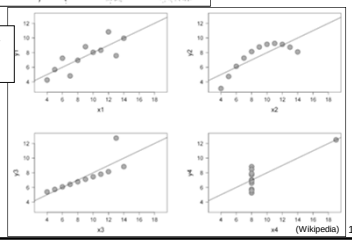
Lineaarne korrelatsioon näitab vaid lineaarset seost.

Korrelatsioonikordaja on üksik üldistav näitaja, graafik on palju informatiivsem.

Mis on korrelatsioonikordaja eelis graafilise kujutamise ees?

Numbriline väärtus on kompaktsem, ja enam üldistatud, mis tagab üldisema võrreldavuse.

Kõigil nelja seose R = 0,816



(Wikipedia)

10

Astakkorrelatsioonid

Astak — järjekorranumber.

Astakute kasutamisel ei ole muutujate jaotus oluline.
Muutujad peavad küll olema **järjestatavad**, aga ei pea olema pidevad ega normaaljaotusega. See on astakkorelatsioonide eelis.
Otsitav **seos** peab olema **monotoonne** (aina kasvav või kahanev).

Kui need tingimused ei ole täidetud, siis saab statistilisi seoseid uurida jaotuste võrdlemise testidega, näiteks χ^2 testiga.

Spearmani astakkorrelatsioonikordaja *ρ (roo)* — **lineaarne korrelatsioonikordaja vaatluste järjekorranumbritest.**

Omadused:
$-1 \le \rho \le 1$;
kui tunnuste vahel on kasvav seos, on $\rho > 0$;
kui tunnuste vahel on kahanev seos, on $\rho < 0$;
kui tunnuste vahel on funktsionaalne seos, siis $|\rho| = 1$;
kui tunnused on sõltumatud, siis $\rho = 0$.

11

Dice-Sørenseni sarnasuskordaja

Ühendi suhteline osa ehk **Dice-Sørenseni sarnasuskordaja**

$$QS = \frac{2C}{A+B}$$

Kokkulangev osa

Ühe vaatluse kogumaht ja teise vaatluse kogumaht

QS = 2·(0,2+0,07)/(0,3+0,2+0,12+0,08+0,5+0,07+0,2) = 0,37

Puude katvused ühes eraldises: Mä 30%, Ka 20%, Ku 12%, Le 8% ⟺ **Katvused teises eraldises: Ka 50%, Ku 7%, muu 20%**

QS = ?

Katvused kolmandas eraldises: Ka 40%, Ku 30% Le 10%

QS = ?

Sarnasuskordaja arvestab korraga mitut tunnust.

Kõigi vaatluspaaride puhul saab leida vaatlustevahelise kauguse. Saadakse
1) sarnasuste maatriks ja
2) kauguste maatriks.

Saab uurida, kas sarnasus sõltub kaugusest ja kui kaugele sarnasus ulatub.

12

Sagedusjaotuste võrdlemine (2 x 2 sagedusjaotus)

Näidisülesanne

Vaadeldi 1090 rasvatihase reaktsiooni raudkulli topisele pesapaiga läheduses. 450st vaadeldud isalinnust reageeris agressiivselt 320, emaslindudest näitas agressiivsust välja 420.

Ülesandes antud sagedused

Vaenulikkus	♀	♂	Kokku
+	420	320	?
–	?	?	?
Kokku	?	450	1090

Seose puudumise korral oodatavad sagedused

Vaenulikkus	♀	♂	Kokku
+	420	320	740
–	220	130	350
Kokku	640	450	1090

Vaenulikkuse osa
♀ **— 420 / 640 = 65,625%**
♂ **— 320/450 = 71,11%**
Paistab, et isased on vaenlase suhtes vaenulikumad.

Sagedustabel kirjeldab, aga ei tõesta seose olemasolu!

Ka juhuslikke arve tabelisse seades saab mingid erinevused.

13

Vigade maatriks

Kontrollandmed

Õpetusandmed	A	B	C	Kokku
A	120	23	12	155
B	11	505	23	539
C	35	45	400	480
Kokku	166	573	435	1174

Ridades ja veergudes on sama tunnus. Mõõtmisviis on erinev.

Diagonaalil on kokkulangevate vaatluste arv.

Vigade maatriksist saab arvutada klassifikatsioonide vastavuse suhtarve, näiteks:

Klassi A eristamise täpsus kasutaja jaoks (*user's accuracy*), **kontrollandmete suhtes** = 120/166 ≈ 72%

Klassi A eristamise täpsus klassifitseerija jaoks (*producer's accuracy*), **õpetusandmete suhtes** = 120/155 ≈ 77%

14

Vastavuse indeks kapa (inglise k *kappa* ehk *KHAT*)

Vigade maatriks

	A	B	C	Σ
A	120	23	12	155
B	11	505	23	539
C	35	45	400	480
Σ	166	573	435	1174

Valemid

$$P_c = \frac{\sum_{i=1}^{k} n(i,i)}{n}$$

$$P_0 = \frac{\sum_{i=1}^{k} (n(+,i) \cdot n(i,+))}{n^2}$$

$$K = \frac{(P_c - P_0)}{(1 - P_0)}$$

Diagonaalil olevate pikslite oodatav **osakaal** juhuslikkuse korral

0 tähistab juhuslikkust

muutub vahemikus 0 … 0.5;

sõltub klasside arvust, paljude ühtlase suurusega klasside puhul väiksem

P_0 = (155 • 166 + 573 • 539 + 435 • 480) / 1174^2 = **0.394**

Tegelikult diagonaalil olevate pikslite osakaal

C tähistab korrektsust

P_C = (120 + 505 + 400) / 1174 = **0.873**

muutub vahemikus 0 … 1

Kui klasse on palju, on diagonaali ruute suhteliselt vähem. Sellepärast õigete osakaal (P_0) ise ei sobi klassifikatsioonide kooskõla hindamiseks.

Kapa koefitsient

K = (**0.873** – **0.394**) / (1 – **0.394**) = **0.79** = 79%

***K* muutub vahemikus −1 … 1**

0 tähistab seose puudumist

Millal on K väärtus −1?

Võiks meelde jääda

- Andmeanalüüs jaguneb kirjeldavaks ja tõestavaks.
- Kirjeldav analüüs
  - otsib tüüpilist (keskmist),
  - kirjeldab tunnuste varieeruvust,
  - iseloomustab seoseid tunnuste vahel ja sarnasust objektide vahel,
  - üldistab ja modelleerib.
- Kvantiil on koht muutuja arvteljel, millest ühel ja teisel pool on kindel osa vaatlustulemusi.
- Korrelatsioon on vastastikune seos tunnuste vahel.
- Korrelatsioonikordaja kirjeldab lineaarset seost ja kasutab ruuthälbeid.
- Mitteparameetrilised korrelatsioonikordajad ei sea tunnuste jaotustüübi eeldusi.
- Sarnasus on vastastikune seos objektide (vaatluste) vahel.
- Kapa kordaja mõõdab vastavust klassifitseerimistulemustes.
- Kapa muutumisvahemik on -1...+1. Null näitab juhuslikkuse korral oodatavat vastavust.

16