Sotsiaalse analüüsi meetodid:
kvantitatiivne lähenemine

Andmete kvaliteedi hindamine: erindid

Indrek Soidla

Erindid

- Erind – teatava kriteeriumi kohaselt skaala teistest väärtusest märkimisväärselt erinev väärtus
  - Milline on märkimisväärne erinevus?
  - Pole üheselt defineeritav
  - On erinevaid kriteeriume (nt erinevus 2, 2,5, 3 vms standardhälvet keskmisest)
  - Millise kriteeriumi ja lävendi kasuks otsustada –hinnanguline, sõltub kontekstist

Erindid: miks oluline?

- Miks on oluline erindid tuvastada ja nende osas midagi ette võtta?
  - Erindite võimalikud põhjused
    - vead andmetes
      - juhuslik kõrvalekalle või süstemaatiline nihe
    - heterogeensus andmetes (erinevad distinktiivsed alamrühmad)
  - Võivad oluliselt mõjutada/kallutada tunnuse põhjal arvutatavaid näitajaid
    - Nt tunnuse keskmine, standardhälve, dispersioon
  - Võivad seoste uurimisel viia I või II tüüpi veani
  - Võivad viia uue sisulise teadmise jälile

Erindid: liigid ja võimalused tuvastamiseks

- Ühemõõtmeline erind – märkimisväärselt erinev väärtus **ühe tunnuse poolest**
  - Nt indiviid, kelle kuusissetulek on 10 000 eurot
  - Tuvastatavad ühemõõtmelise analüüsiga, nt
    - tunnuse enda jaotus
    - tunnuse põhjal arvutatavad jaotusparameetrid ja näitajad
- Mitmemõõtmeline erind – märkimisväärselt erinev väärtus **kahe või rohkema tunnuse kombinatsioonis**
  - Näide:
    - 15-aastane indiviid ei ole erind
    - 2000 eurose kuusissetulekuga indiviid ei ole erind
    - Küll aga on erind 15-aastane indiviid, kes saab kuus 2000 eurot
  - Tuvastatavad mitmemõõtmelise analüüsiga, nt
    - mitmene jaotus
    - regressioonimudeli jääkide analüüs

Erindid: liigid ja võimalused tuvastamiseks

- Näide: ESS 2014

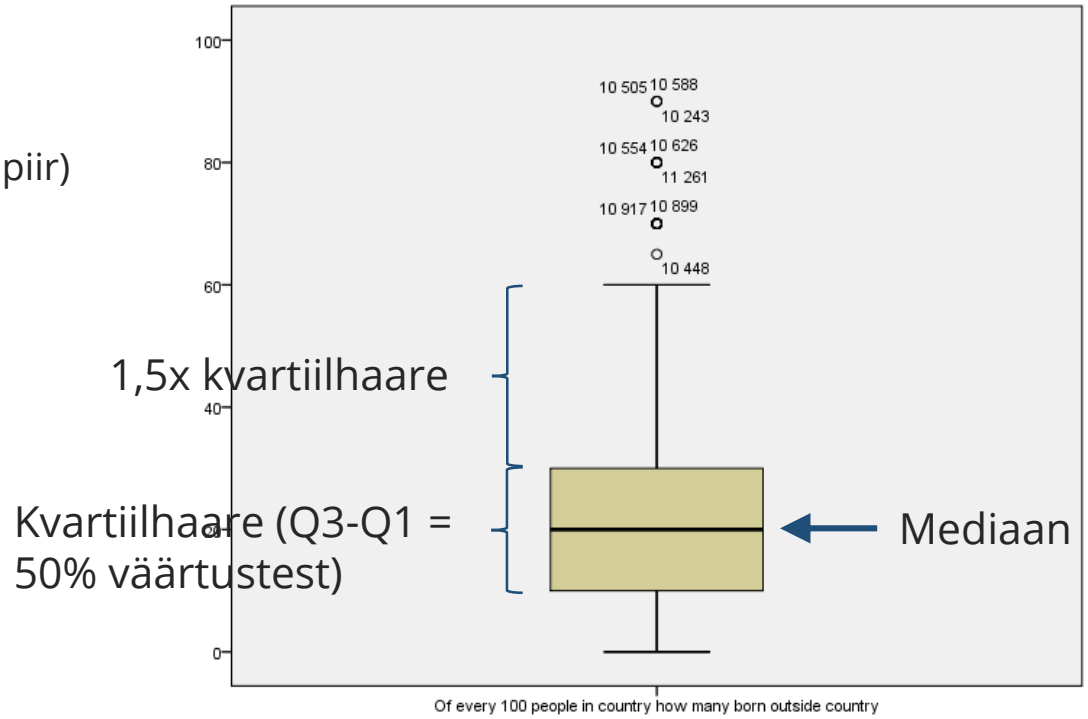
  Kui mitu igast 100st Eestis elavast inimesest on Teie arvates sündinud väljaspool Eestit?

  INTERVJUEERIJA: Kui vastaja ütleb „ei oska öelda“; siis öelge: „Palun andke hinnanguline number“.

- Tunnus noimbro: Of every 100 people in country how many born outside country

Erindid: liigid ja võimalused tuvastamiseks

- Ühemõõtmelised erindid
  - Vihjed erinditeolemasolule: suur asümmeetriakordaja, mediaani ja keskmise suur erinevus, miinimumi, maksimumi ja keskmise võrdlus
  - Tunnuse jaotus tabelis
  - Visuaalne jaotus: histogramm
  - Visuaalne jaotus: karpdiagramm
  - Põhineb variatsioonirea kvartiilidel:
    - Q1 = alumine kvartiil (alumise/esimeste 25% väärtuste piir)
    - Q2 = mediaan
    - Q3 = ülemine kvartiil
    - Q3-Q1 = kvartiilhaare (IQR)
    - Erindid: x_i < Q1 - 1,5*IQR | x_i > Q3 + 1,5*IQR
    - Äärmuslikud erindid: x_i < Q1 - 3*IQR | x_i > Q3 + 3*IQR



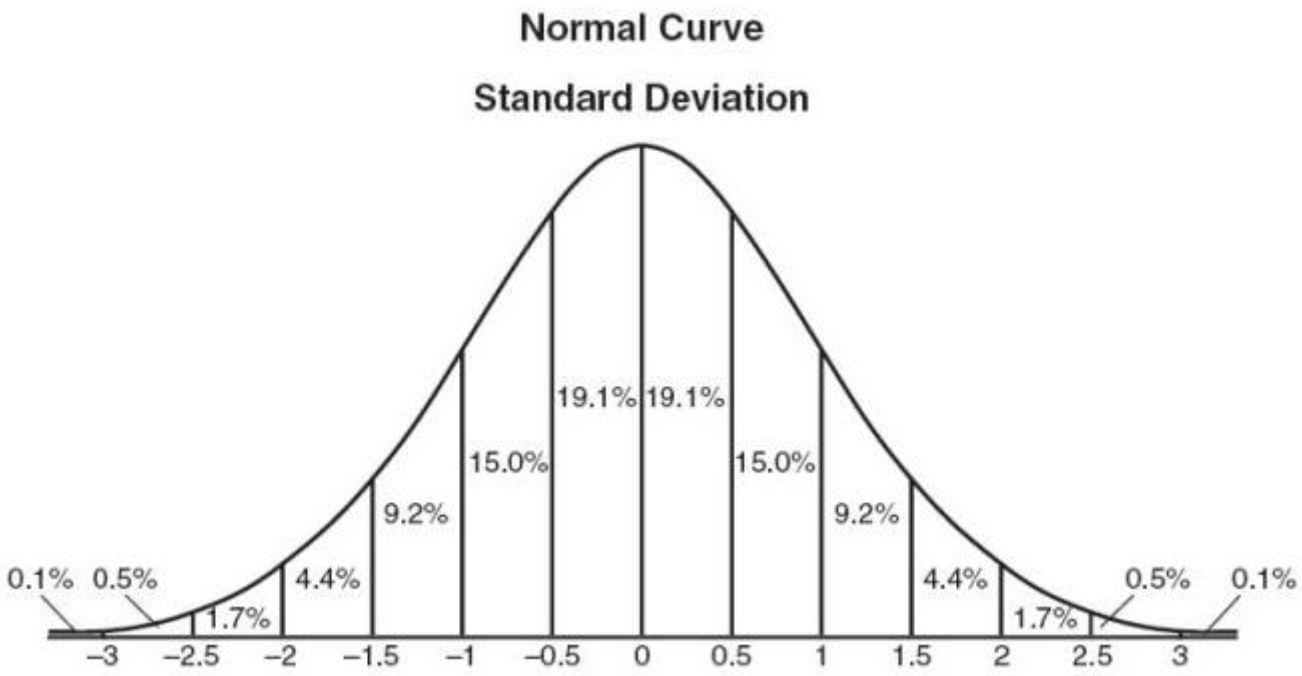
Erindid: liigid ja võimalused tuvastamiseks

- Ühemõõtmelised erindid
  - Z-skoor
    - Tunnus standardiseeritakse: tunnuse väärtuseid nihutatakse nii, et
      - keskmine m_x saab väärtuse 0,
      - tunnuse väärtusi väljendatakse standardhälbe (s_x) ühikutes
      - tunnuses *x* saame indiviid *i* puhul *z*-skoori väärtuseks seega

$$z_i = \frac{x_i - m_x}{s_x}$$

    - Erindi lävendiks on erinevus keskmisest mõõdetuna standardhälvetes, nt
      - väärtus paikneb vähemalt/rohkem kui 2,5 või 3 või 3,5 standardhälbe kaugusel keskmisest

Z-skoor: võimalikud erindi lävendid

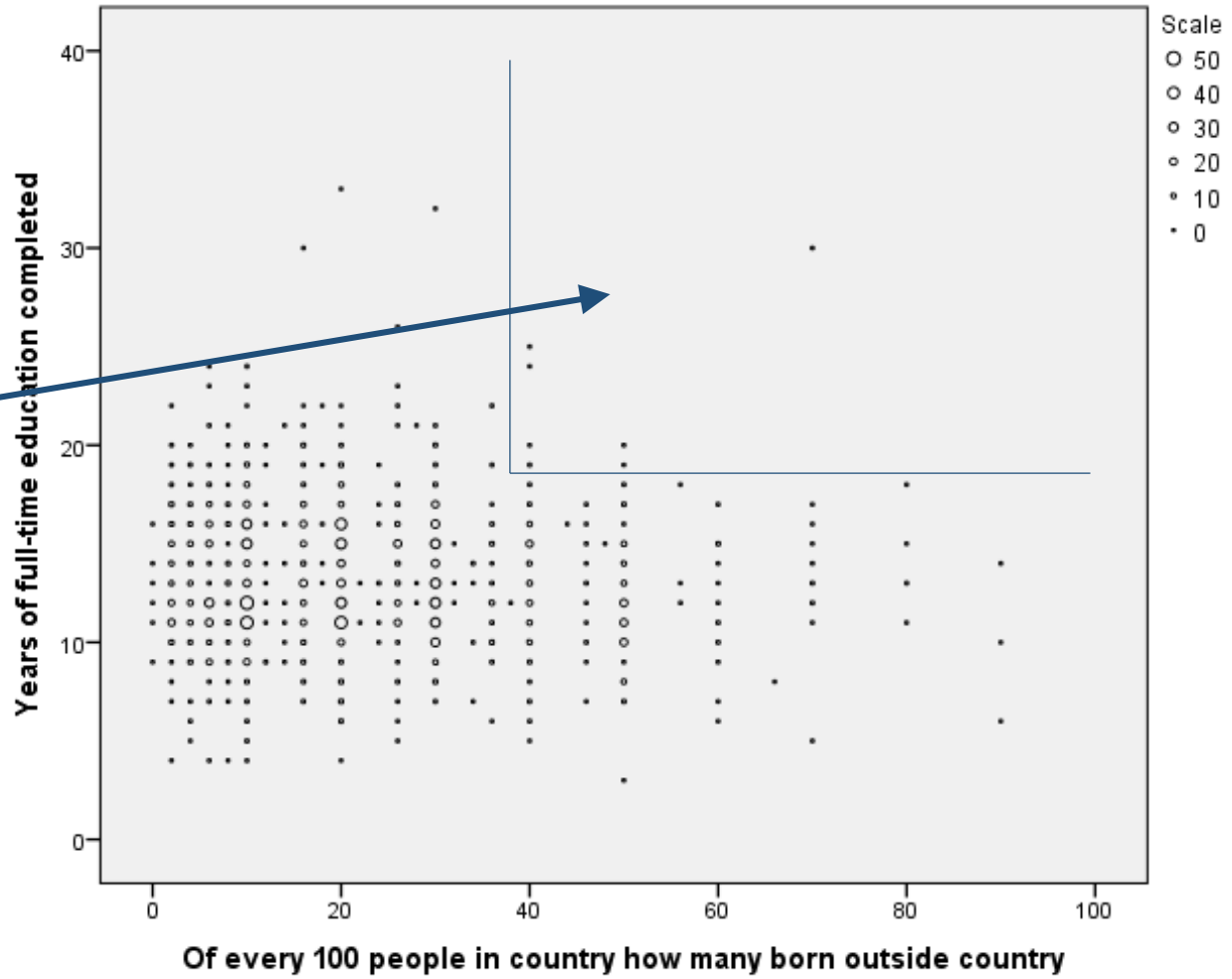


Erindid: liigid ja võimalused tuvastamiseks

- Z-skoor
  - Standardiseeritud väärtused sõltuvad tunnuse jaotusest
  - Sh erinditest =>
  - Erindite olemasolu kahandab z-skooride „erandlikkust“ =>
  - Teatud määral sõltub see erindite tuvastamise meetod erindite olemasolust!

Erindid: liigid ja võimalused tuvastamiseks

- Mitmemõõtmelised erindid
  - (Tunnuste jaotus risttabelis)
  - Visuaalne mitmemõõtmeline jaotus (nt hajuvusdiagramm)
- Põhjused?
  - Iseäralikud indiviidid?
  - Kehva andmekvaliteediga indiviidid?



Erindid: liigid ja võimalused tuvastamiseks

- Mitmemõõtmelised erindid
  - Suured regressioonijäägid regressioonimudelis
  - Erindite algpõhjuste leidmiseks teha indikaatortunnus erindite ja normaalväärtuste eristamiseks => võrrelda muude tunnuste jaotuseid või keskmisi

Erindid: põhjused ja käsitlusviisid

1) Andmesisestusviga
   - teatud ulatuses võimalik järelkontrollida või ennetamiseks seadistada kriteeriumid
   - kui on selge, et tegu sisestusveaga…
     - ja võimalik tuvastada täpne väärtus => sisestada täpne väärtus
     - pole võimalik tuvastada täpset väärtust => vastus võimalik kustutada
2) Andmelünga kood jäetud lüngana defineerimata
   - kontrollimisel reeglina lihtsasti tuvastatav
   - defineerida andmelüngana
3) Ülekaetuse viga
   - respondendi vastused kustutatakse
4) Respondent on sihtpopulatsiooni esindaja, kellel ongi tunnuses ebatavaliselt erandlik väärtus –
   - tunnuse teisendamine mingi matemaatilise funktsiooni abil
     - ainult paari erindi pärast pole mõttekas, kui tunnuse jaotus muidu enam-vähem normaaljaotuse lähedane
   - tunnuse teisendamine arvulisest kategoriaalseks (tunnuse n-ö kirjeldusvõime kahaneb)
   - võimalik respondent alles jätta, aga muuta erindi väärtust *(Winsorising)*
     - väärtusel väiksem mõju analüüsitulemustele
     - subjektiivne
   - ei muuda erindit, vaid kasutame analüüsimeetodeid, mis pole erindite suhtes tundlikud
     - (nt mitteparameetrilisi meetodeid, nt korrelatsioonseose hindamisel Pearsoni korrelatsioonikordaja asemel Spearmani korrelatsioonikordajat)