Datavask

Lena Ringstad Olsen
statistiker
Nasjonalt Servicemiljø for medisinske kvalitetsregistre

Hvorfor datavask ?

- Grunnleggende for all bruk av data er **god datakvalitet**
- Undersøke data og dokumentere datakvalitet før man gjør analyser/forskning
- Avdekke mulige feil, mangler, svakheter, motsigelser, umuligheter
- Må være sikker på at man kan stole på dataene før noe som helst publiseres!
- Søppel inn = Søppel ut

Datakvalitets«dimensjoner»

Relevans

I hvilken grad registeret oppfyller nåværende og fremtidige behov hos brukere av data.

Korrekthet

I hvilken grad registeret reflekterer virkeligheten det skal måle.

Kompletthet

I hvilken grad alle data som burde vært registrert er registrert.

Reliabilitet

I hvilken grad innholdet i registeret er reproduserbart.

Aktualitet

Tid fra hendelsen inntraff til informasjonen er tilgjengelig for brukere av data.

Sammenlignbarhet

I hvilken grad data er sammenlignbare på tvers av tid, geografi og ulike datakilder.

Kontroll på data - dokumentasjon

- ***«Det er ønskelig at registrene skal holde ei fullstendig oversikt over hver enkelt variabel med begrunnelse for hva de har tenkt å bruke variabelen til og hvordan den blir brukt, samt kvaliteten av den»***
- *Etterspør grundig dokumentasjon av variabler*
  - *Hva er et forløp/hendelse?*
  - *Nøyaktig definisjon av hver variabel*
  - *Hva ønsker vi å vise (budskap)*
  - *Kodede variable (heltall, naturlig ordnet)*
  - *Kan navnene føre til misforståelser/feil? RevaskulariseringFoerInfarkt, Kjønn, Blod2, Ms3*
- *Beskrivelse av alle variable (kodebok, metadata…)*
- *Valideringsregler i innregistreringsløsninga*

Definisjonen til folkehelseinstituttet?

Fra behandlingsstart til behandlingsslutt?

Liggetid

Fra operasjon til utskriving?

Fra innskriving til utskriving?

Illustrasjonsfoto: www.colourbox.com

Endringer i data

- Endringer i variabelnavn? (også navneendring for id-variable/koblingsnøkler)
- Har variabler endret innhold, f.eks. fått en ny kategori?
- Er standardverdier endret?
  - Eks. endring i representasjon av tomme felter fra NULL til -1.
- Variabeloversikt (kodebok) skal inneholde:
  - Forklaring av hver enkelt variabel
  - Variabeltype/format
  - Koding av kategoriske variabler 1=brun, 2=blå
  - Min/maksgrense
  - Dato for måling
  - Enhet
- Etterspør dokumentasjon av alle endringer!

Kodebok

Fritekstvariabler

- Bør unngås
- Ikke uvanlig å motta data i fritekst
- Utelukker mulighet for valideringsregler ved innregistrering
- Vanlige problemer:
  - Mellomrom før og etter ord
  - Blanding av små og store bokstaver
  - Ved scanning: 1 tolket som l eller I, O som 0, og motsatt
  - Komma/punktum
  - Encoding: Lesing og visning av æ, ø, å

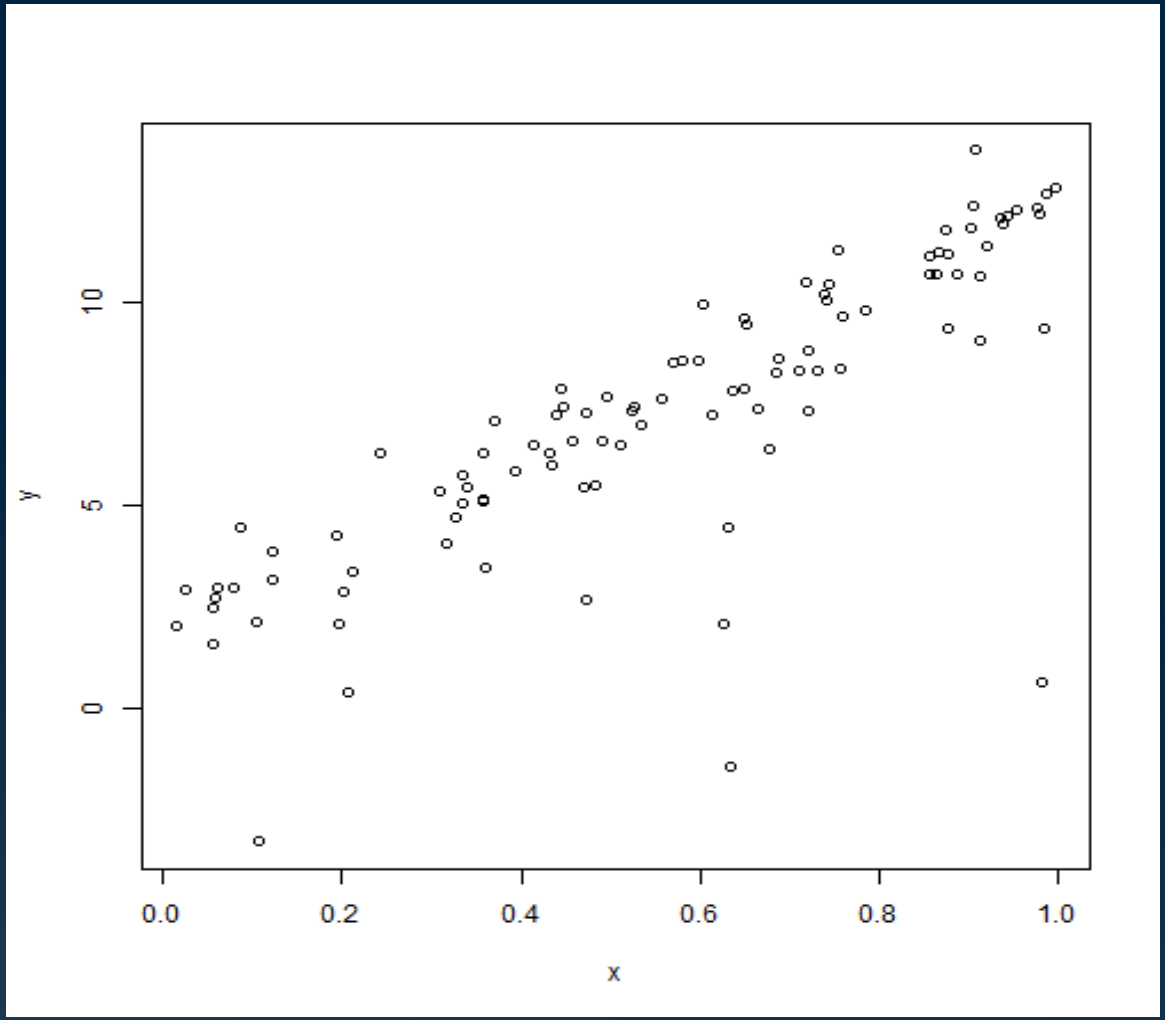
Datoformater

- Kan komme som tekst
- dd.mm.yyyy, yyyy-dd-mm, mm-dd-yyyy
- NB: Excel tolker relativt ofte et desimaltall som dato!
- Studie i 2021 fant at 30% (av 10000) publiserte genetiske studier innehold datafeil pga. autoendringer i dataformat i Excel [ref: Abeysooriya M, Soria M, Kasu MS et al. «Gene name errors: Lessons not learned. https://doi.org/10.1371/journal.pcbi.1008984
- Kan åpne fila i Notepad(++) for se hvordan den egentlig ser ut.

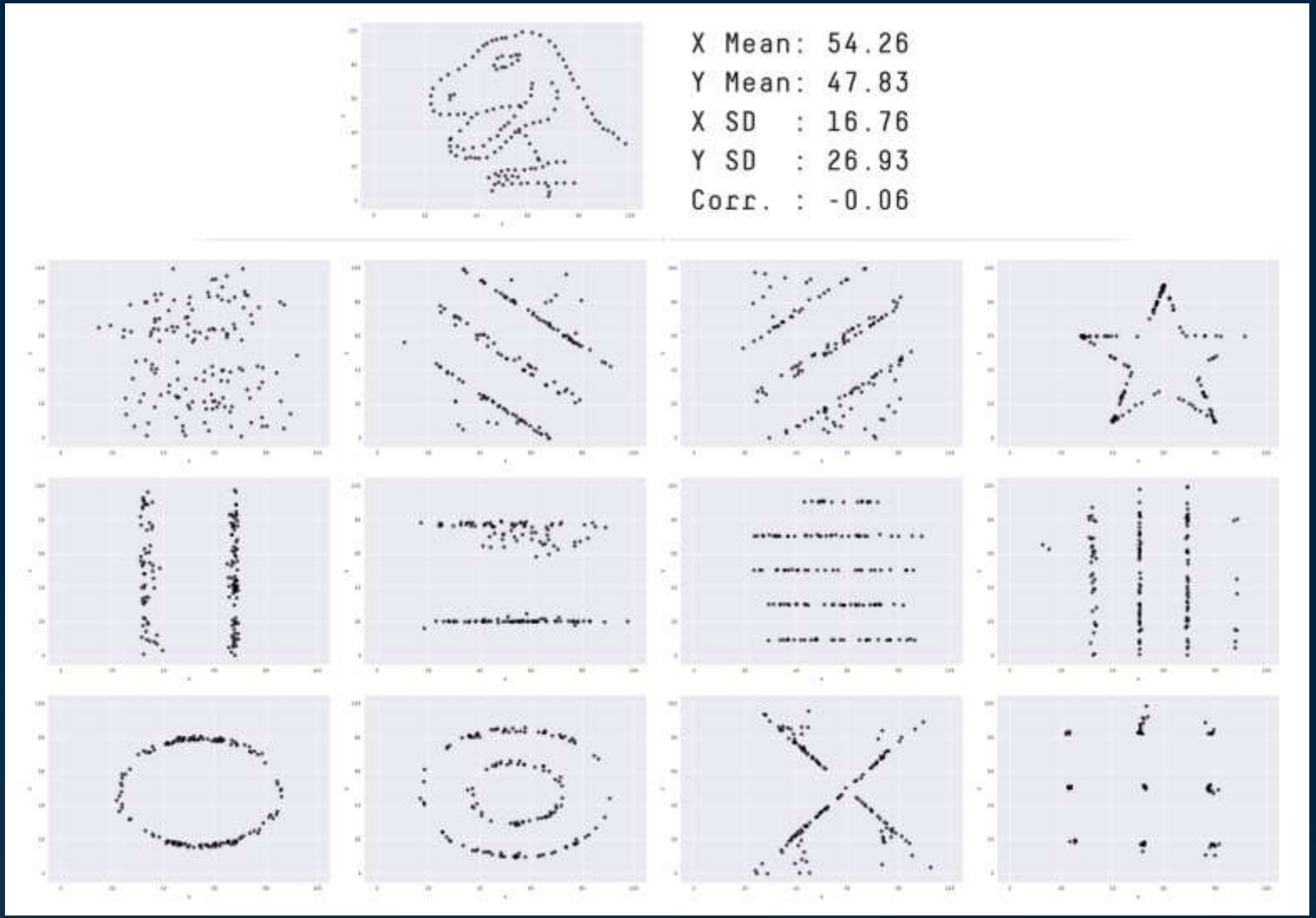
Vurdere data

- Tallbehandling: Verktøy
  - Personlige favorittverktøy…
  - Rene tekstlesere (Notepad)
  - Excel: sortere, gjøre utvalg, sammenlikne variable,
- Visualiser (plott) data

Variabelinnhold



SKDE



https://youtu.be/DbJyPELmhJc?t=4

Mulige problemer med datasett

- «Ukontrollerbar» fritekst (f.eks. fra scannerløsning):
- Datoer: Ulike formater
- Duplikater
- Ugyldige verdier
- Tomme felter
- Misvisende standardverdier
- Uhandterlige variabelnavn
- Tall som tekst
- Upraktisk struktur
- Data levert i flere tabeller, behov for kobling
- Encoding

Datavask, demo

Huskeliste

- Kontrollere innhold i variable.
  - Uhåndterlige variabelnavn?
  - Dobbeltregistreringer?
  - Koding/tekststrenger – stemmer kodeverdiene man finner med det som er oppgitt i kodeboka
  - Inneholder variabelen andre enn de dokumenterte verdiene?
  - Hvordan representeres manglende verdier?
  - Max/min-verdier, er de ok?
  - Format (fritekst, datoer)
  - Encoding
  - Utfyllingsgrad/kompletthet
  - Test logikk (eks innleggelsestidspunkt FØR utskrivingstidspunkt)
  - Avledede variabler –få tak i beregningsmetodene/formlene
- Trenger vi å omstrukturere data?
- Etterspør dokumentasjon!!

Til slutt

- Lek med data og få et nært og «personlig» forhold til dataene du skal analysere
- Ta vare på kode/beregninger så du kan dokumentere kvalitet/vask
- Plott variablene for å få overblikk