



Statistisk Dataanalyse 1: Praktisk info og genopfriskning af R, datatyper og deskriptiv statistik

Anders Tolver
Institut for Matematiske Fag



Dagens program

Velkommen

- Hvad er statistik?
- Praktiske oplysninger
- Datatyper
- Genopfriskning af R: primært i dagens R-program
- Deskriptiv statistik



Eksempler på statistik



Folkesundhed og børneopdragelse

"Min oldefar var en rigtig levemand, der røg cigarer og elskede god, gammel dansk mad. Og han blev 98 år. Min tante dyrkede regelmæssig motion og spiste sundt og varieret. Men hun døde som 59 årig."

(Hørt til familiefesten i Vestjylland)

"Vi har aldrig begrænset børnenes adgang til slik og sodavand, men vores børn vælger altid det sundeste alternativ."

(Irriterende Blogger eller mor på Aula)

- Hvad ønsker vi egentlig at vide?
- Kan vi bruge data/oplysningerne?
- Hvad er det statistiske problem?



Eksempel 1: To-kryds-to tabeller

Situation 1: Vaccine mod miltbrand hos får. Næppe brug for en statistiker i dette tilfælde...

	Vaccineret	Ej vaccineret
Død	0	24
I live	24	0

Situation 2: Forekomst af leversvulster hos mus i forskellige miljøer. Konklusionen er knapt så oplagt.

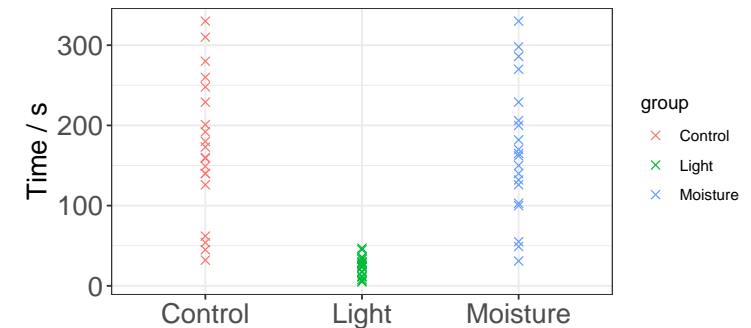
	<i>E.coli</i>	Rent miljø
Leversvulster	8	19
Ingen svulster	5	30



Eksempel 2: Ensidede variansanalyse

60 bænkebidere er blevet placeret i et af tre miljøer, og deres fysiske formåen er blevet testet ved at lade dem løbe en distance.

Er der en effekt af lys hhv. fugtighed? Hvor stor er effekten?



Eksempel 3: Alkohol og studiefrafald

Statistik — Du falder fra, hvis du drikker for meget. Men også, hvis du drikker for lidt. Friske tal viser, at studerende, der ikke drikker alkohol i studiestarten, har lige så stor risiko for frafald i løbet af første studieår, som studerende der drikker meget tæt.

Figure: Publiceret online i Universitetsavisen d. 29/8-2019

Hvorfor skal vi ikke prioritere at servere en stor gratis fadøl for alle studerende på KU hver fredag kl. 16?



Hvad er statistik?

Statistik handler om, hvordan man drager **korrekte konklusioner på baggrund af data**.

Hvorfor er statistik et vigtigt fag? Forskellige holdninger ...

If your experiment needs a statistician, you need a better experiment. (Ernest Rutherford, fysiker)

Ressourcer (tid/økonomi/udstyr) vil dog ofte begrænse mængden og kvaliteten af data.

Hvor længe kan vi vente med at tage en behandling i brug, som ser ud til at virke? (Corona vaccine, grøn omstilling, ...)



Holdninger til statistisk

Vi indtager en mere idealistisk holdning på dette kursus.

Brug af statistik garanterer ikke, at vi drager rigtige beslutninger fra data.

While it is easy to lie with statistics, it is even easier to lie without them.

(maybe Frederick Mosteller, statistician)

Vi bør være bevidste og åbne omkring metoder og antagelser, som vi bruger til at analysere og fortolke data.



Alternative holdninger til statistik/sandheden

Andre har en mere afslappet holdning til brug af data ...

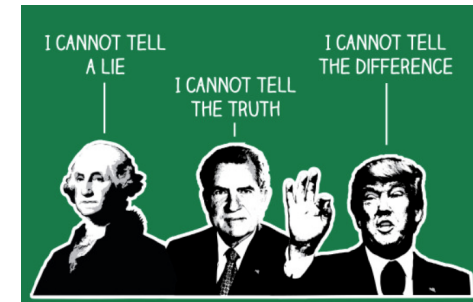


Figure: fra kommerciel t-shirt



Hvad er statistik?

Formålet med statistik er (typisk) at **undersøge sammenhænge mellem flere typer målinger udfra indsamlet data.**

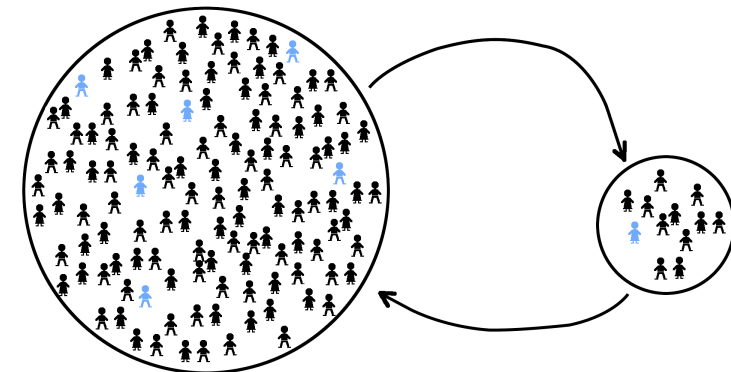
- Er der en sammenhæng? Hvilken?
- Er der en effekt af behandling? I hvilken retning? Hvor stor?

Udfordringer:

- Data behæftet med **usikkerhed**: biologisk variation, målestøj
- Ser kun en **begrænset mængde data**, men ønsker at udtale os om generelle sammenhænge



Populationer og stikprøver



Vi ønsker at udtale os om en population ud fra en stikprøve



Hvad er formålet med dette kursus?

Kursusindhold: Grundlæggende statistiske metoder og beregninger

Kurset giver jer redskaber til at

- forstå og vurdere udsagn givet ved brug af statistik
- lave valide konklusioner udfra egne eksperimenter
- vurdere hvornår det er nødvendigt at søge hjælp hos en statistiker
- arbejde effektivt og struktureret med data



Datatyper



Datatyper

Første skelnen: **Kategoriske data vs kvantitative data**

Kategoriske data:

- **Nominale** — {mand, kvinde}, {gul, grøn, blå}.
- **Ordinale** — {ingen, lav, mellem, høj}, indkomstklasser, graduering af smerter/symptomer.

Kvantitative data

- **Diskrete** — unger pr. kuld, antal familiemedlemmer.
- **Kontinuerte** — længde, højde, alder, vægtændring, indkomst.

StatData1: Vi skal mest bruge *nominale kategoriske* og *kontinuerte kvantitative* data. Ofte siger vi bare kategoriske og kontinuerte.



Hvorfor er datatypen vigtig?

Fordi datatypen er afgørende for hvordan der er relevant at behandle data:

- Hvilke stikprøvestørrelser (summary measures)?
- Hvilke tegninger?
- Hvilke statistiske analyser?



Meningsmålinger

Undersøgelse baseret på telefoninterview med 1.000 repræsentativt udvalgte personer 18 år+ i perioden 22.-28. august 2022. Her ses udpluk af data ...

```
##      stemme
## 1      A-Soc
## 2      C-Kons
## 3      V-Venst
## 4      Æ-Dan
## 5      Æ-Dan
## 6      V-Venst
## 7      V-Venst
## 8      V-Venst
## 9      Øvrige
## 10     D-NyeB
```

(Kilde: [Voxmeter for Ritzaus Bureau](#))



Meningsmålinger

Tænk lidt over følgende

- Hvilke typer variable optræder i datasættet?
- Findes der en bedre måde til at opsummere datasættet?
- Hvad ønsker vi egentlig at vide?
- Hvad er det statistiske problem?
- Hvordan kan vi visualisere data?



Datasæt med katte

Data vedr. 144 katte.

Tre variable: Køn, kropsvægt i kg, vægt af hjerte i gram.

```
##      Sex Bwt Hwt
## 1      F 2.0 7.0
## 2      F 2.0 7.4
## 3      F 2.0 9.5
## 4      F 2.1 7.2
## 5      F 2.1 7.3
## 6      F 2.1 7.6
## 7      F 2.1 8.1
## 8      F 2.1 8.2
## 9      F 2.1 8.3
## 10     F 2.1 8.5
```

Hvilke datatyper er de tre variable i datasættet?



Praktiske oplysninger



Praktisk info

Kurset har en ekstern [hjemmeside](#), hvor du vil kunne finde alle praktiske oplysninger om kurset.

Et stor del af de **Praktiske oplysninger** vil også ligge på kursets Absalonside, hvorfra der også er link til kursushjemmesiden.

En del af undervisningsmaterialet vil kun være tilgængeligt via links på den eksterne hjemmeside.

Skriv til mig, hvis du finder oplagte fejl og mangler på hjemmesiden.

Planen for næste uges øvelser udsendes typisk sent torsdag, og forelæsningslides lægges ofte først ud lige før forelæsningen.



Undervisningsmateriale og ugestruktur

Undervisningsmateriale:

- *Introduction to Statistical Data Analysis for the Life Sciences* af Ekstrøm og Sørensen, 2. udgave
- Slides, opgaver, data, R-programmer mm.
- Quiz'er (ikke nødvendigvis hver uge)

Aktiviteter:

- Forelæsninger (2 x 2 timer)
- Øvelsestimer (2 + 3 timer)
- Video med gennemgang af quiz + opsummering efter behov (Ca. 45 minutter - ikke alle uger)
- **Hjemmearbejde** (mindst 10 timer per uge!)
- 2–3 afleveringsopgaver. Frivilligt, men et godt tilbud!



Undervisningen

Forelæsningerne:

- Jeg gennemgår ikke bogen fra A til Z. Mindre matematik, ofte andre dataeksempler
- Jeg lægger fuldstændige R programmer ud til jer, men kører ikke alt ved forelæsningerne
- Slides kommer som regel på hjemmesiden aftenen før

Øvelsestimerne:

- Det meste af tiden regner I selv de opgaver der er stillet på ugeplanen, med hjælp fra instruktorerne
- Gennemgang af enkelte ting fra foregående timer
- Flere opgaver end I kan nå i timerne. I skal regne hjemme!
- Arbejd sammen i grupper, spørg om hjælp



Hjemmearbejde

Du forventes at bruge i alt mindst **20 timer om ugen** på kurset!

Hvordan timerne bruges bedst er individuelt, men her er et forslag:

- Forelæsninger/video: 5 timer
- Øvelser: 5 timer
- Læse i bogen, læse slides, køre mine R-programmer: 6 timer
- Regne opgaver hjemme: 4 timer

Der kommer facit/besvarelser til det meste efter timerne, men brug dem med omhu. Du skal selv have fingrene ned i skidtet for at lære det!



Eksamen

Du bør evaluere dit eget udbytte af kurset på om du

- forstod hvorfor faget kan være relevant for dit fagområde
- brugte tid på at lære at tænke over statistiske problemstillinger
- lærte at lave simple statistiske analyser med R

Jeres udbytte af kurset evalueres desuden ved en eksamen

- 4 timer skriftlig prøve med alle hjælpemidler på nær internet
- **I skal selv køre R** på data som udleveres i forb. med eksamen
- Der kommer **quizspørgsmål** som dem der bliver stillet til quizzer i løbet af kurset



Om R

- Vi skal bruge R intensivt på kurset
- Installér de **nyeste versioner af R og RStudio**
- Nogle af HS-opgaverne er genopfriskning af R
- På kursushjemmesiden findes en oversigt over relevant R materiale for kurset

Alle R programmer lægges ud i R markdown-format, da det er kedeligt og ufuldstændigt at vise R koder på forelæsningslides.

Anbefaling

- Download R Markdown-filen og følg med under forelæsningsen. Skriv evt. korte noter.
- Kør selv R koden i R Markdown-filerne efter forelæsningsen. Suppler med egne kommentarer.



Øvelsesundervisning

Finder sted

- mandage fra ca. kl. 15:00-16:45 i kursusugerne 1-8
- onsdag fra ca. kl. 13:00-15:45 i kursusugerne 1-7

I er automatisk blevet inddelt på 6 øvelseshold men

- i praksis afholdes øvelserne i 4 lokaler med totalt 6 hjælpelærere
- I må gerne fordele jer jævnt i lokalerne i ønskede arbejdsgrupper

Forslag: Hjælpelærerne laver 4 grupper i Absalon hørende til hvert lokale. Her kan man tilmelde sig (uforpligtende) på det eller de hold, som man regner med at være mest på.



Genopfriskning af R



R

- Konsollen, prompten, kommandoer ved prompten
- Skriv kommandoer i R-program (eller Markdown, mere om det på onsdag)
- Vektorer/variable i R
- Datasæt, observationer, variable
- Variable i datasæt vha. `$`
- Eksempel: Datasættet **cats** i *MASS*-pakken

Se også HS-opgaverne og R-programmet Rprog220905.



Vektorer/variable

Man kan selv definere en vektor/variabel med funktionen `c`:

```
> x <- c(1,2,6)
> x
[1] 1 2 6
> y <- c(4,6,1)
> x+y
[1] 5 8 7
> mean(x)
[1] 3
```



Datasættet **cats**

Datasættet **cats** ligger i pakken *MASS*. Pakke og datasæt skal *loades* før de kan bruges:

```
library(MASS)
data(cats)
```

Data vedr. 144 katte. Tre variable: Køn, kropsvægt i kg, vægt af hjerte i gram.

```
> head(cats, n=3)
  Sex Bwt Hwt
1  F 2.0 7.0
2  F 2.0 7.4
3  F 2.0 9.5
```

Datatyper af de tre variable?



`$`-syntaksen

Vi skal fortælle R at den skal finde variablene i datasættet **cats**.

Dette kan gøres med `$`-syntaks: `datasætnavn$variabelnavn`

```
> Bwt # Virker ikke, da R ikke ved hvor variabelen er
Error: object 'Bwt' not found

> cats$Bwt
[1] 2.0 2.0 2.0 2.1 2.1 2.1 2.1 2.1 2.1 2.1 ...
> mean(cats$Bwt)
[1] 2.723611
```

Advarsel: Det skaber ofte forvirring, hvis man tilfældigvis i R allerede har en variabel ved navn `Bwt`, der intet har at gøre med indholdet i datasættet `cats`.



Deskriptiv statistik



Deskriptiv statistik

Grafer og simple stikprøvestørrelser.

Hvorfor?

- For at give **overblik** over data
- For at give en umiddelbar **kommunikation** af data
- Evt. **finde fejl** i data, fx forkert placering af decimal

Hvordan?

- **Visualisering:** søjlediagrammer, histogrammer, boxplots, scatter plots
- **Simple stikprøvestørrelser:** gennemsnit, spredning, range (min og max), fraktiler
- Altsammen i **R**



Kategoriske data

- **Frekvens** = hyppighed, dvs. antal forekomster
- Hvis n er antallet af observationer er

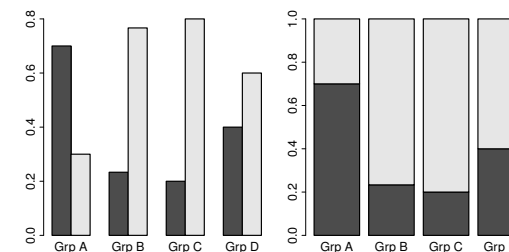
$$\text{Relativ frekvens} = \frac{\text{frekvens}}{n}$$

	Group A	Group B	Group C	Group D	Total
TD present	21	7	6	12	46
TD absent	9	23	24	18	74
Pct present	70	23	20	40	38



Kategoriske data

	Group A	Group B	Group C	Group D	Total
TD present	21	7	6	12	46
TD absent	9	23	24	18	74
Pct present	70	23	20	40	38



R-kode: Se side 18 i bogen.



Kattene igen

Data vedr. 144 katte.

Tre variable: Køn, kropsvægt i kg, vægt af hjerte i gram.

Relevante spørgsmål?

- Sammenhæng mellem vægt af krop og hjerte?
- Fordeling af kropsvægt? Fordeling af hjertevægt?
- Kønsforskelle?

I dagens R program **Rprog220905** beskrives hvordan man kan visualisere kvantitative data ved brug af

- scatterplot
- histogrammer
- boxplot



Stikprøvestørrelser (summary statistics)

Grafer er godt, men vi vil også gerne give nogle **tal** der indeholder information om hvordan fordelingerne ser ud.

- Mål for "**centrum**": Gennemsnit, median
- Mål for **variabilitet**: spredning, range, inter-quartile range (IQR)



Median, kvartiler, IQR

Sortér data efter størrelse (min til max).

Range: Intervallet fra mindste til største observation.

Median: Midterste observation i det sorterede datasæt. Hvis lige antal observationer: Gennemsnit af de to midterste observationer.

Kvartiler deler sættet op i fire grupper. 25% obs. er $\leq Q_1$ (første kvartil), og 75% obs. er $\leq Q_3$ (tredje kvartil).

Altså: De 50% "midterste" data ligger i intervallet fra Q_1 til Q_3 .

Inter quartile range, $IQR = Q_3 - Q_1$



Gennemsnit og stikprøvespredning

Gennemsnit er defineret ved:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{y_1 + \dots + y_n}{n}$$

Stikprøvespredning er defineret ved:

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

For symmetriske data, typisk: Cirka 95% af data ligger i intervallet

gennemsnit $\pm 2 \cdot$ spredning

Gennemsnit og spredning har samme enhed som observationerne.

Stikprøvevariansen: s^2 .



Stikprøvestørrelser for hjertevægt

```
library(MASS)
data(cats)
summary(cats$Hwt)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      6.30   8.95   10.10   10.63   12.12   20.50

mean(cats$Hwt)

## [1] 10.63056

sd(cats$Hwt)

## [1] 2.434636

var(cats$Hwt)

## [1] 5.927451
```



Median eller gennemsnit?

- Median og gennemsnit ens for symmetriske fordelinger, forskellige for skæve fordelinger.
- Ikke-symmetriske fordelinger: Median giver bedre mening end gennemsnit
- Gennemsnit er følsom overfor ekstreme observationer. Median er robust overfor ekstreme observationer.
- Gennemsnittet er "pænere" fra et matematisk synspunkt



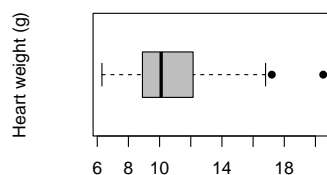
Boxplot

Et **boxplot** illustrerer en fordeling grafisk vha. median og kvartiler.

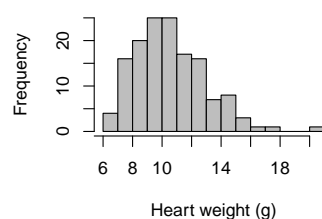
Fed streg er median, kassen går fra Q_1 til Q_3 . Detaljerne er lidt komplicerede...

Boxplots er gode til sammenligning af fordelinger og et groft men fornuftigt alternativ til histogrammer

Boxplot for heart weight



Cats: Heart weight



R til øvelserne i dag

I skal selv indtaste data til vektorer. Kommandoer som nedenstående kan være nyttige:

```
### Indtast relevante værdier
x <- c(2.1, 3.5, 5.3, 1, 9.8)

### Diverse summary statistics
mean(x)
sd(x)
var(x)
median(x)
summary(x)

### Et par figurer
boxplot(x)
hist(x)
```



Opsummering — til eget brug

- Giv eksempler på kategoriske og kvantitative variable. Er de nominale, ordinale, diskrete eller kontinuerte?
- Hvad er medianen, Q_1 og Q_3 ?
- Hvordan beregnes gennemsnit og stikprøvespredning?
- Hvad er et boxplot?
- Hvad sker der med median hhv. gennemsnit hvis der kommer en ny obs. der er ekstremt lille i forhold til de oprindelige?
- Hvordan arbejder man i R?
- Hvordan bruger man en variabel i et datasæt?

