Populationer og stikprøver

Claus Ekstrøm
E-mail: ekstrom@life.ku.dk

Program

- Praktiske oplysninger
- Populationer og stikprøver
- Data
  - Datatyper
  - Visualisering
  - Centrum og spredning af en fordeling

Praktiske oplysninger

- Kursushjemmeside: Absalon
- Program
- Øvelserne
  - Øvelsestimer
  - Afleveringsopgaver
  - Prisopgave
- Cases
- Materiale
  - Lærebogen
- R

Eksempel — vaccine mod miltbrand hos får

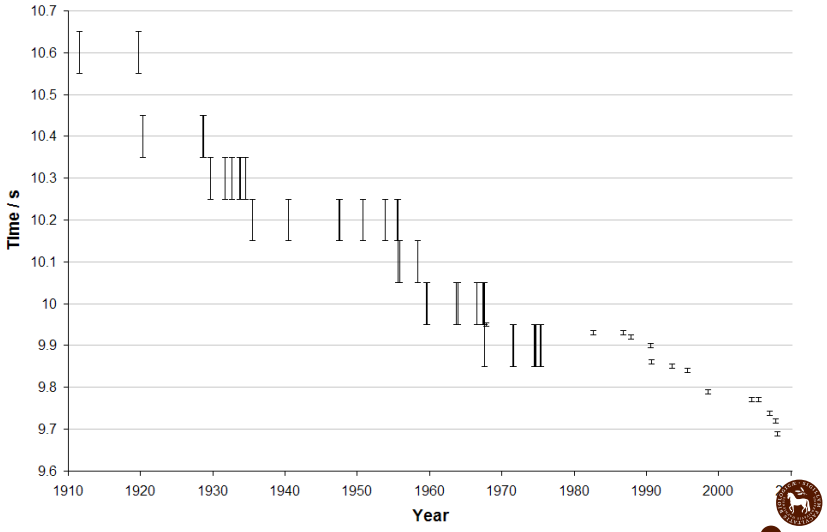
	Vaccineret	Ej vaccineret
Død	0	24
I live	24	0

Eksempel — forekomst af leversvulster hos mus

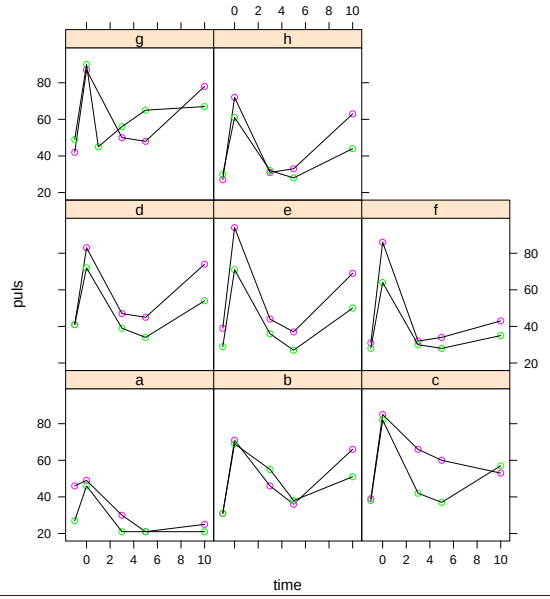
	E.coli	Rent miljø
Leversvulster	8	19
Ingen svulster	5	30

- Er der en effekt af miljø på forekomsten af leversvulster?
- Kan tilfældig variation være skyld i resultatet?
- Hvor stor er effekten?

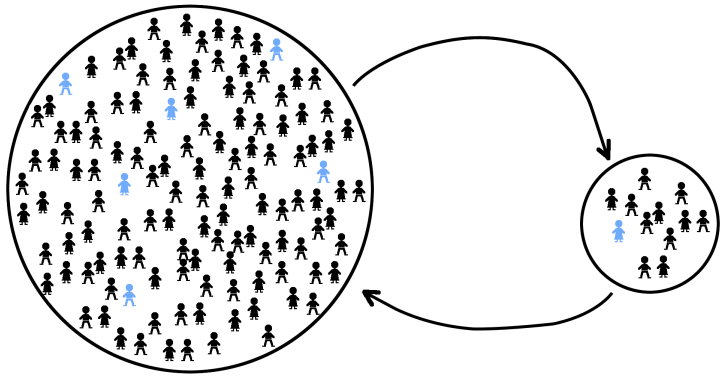
Eksempel — 100 sprint for mænd



Eksempel — LIFE

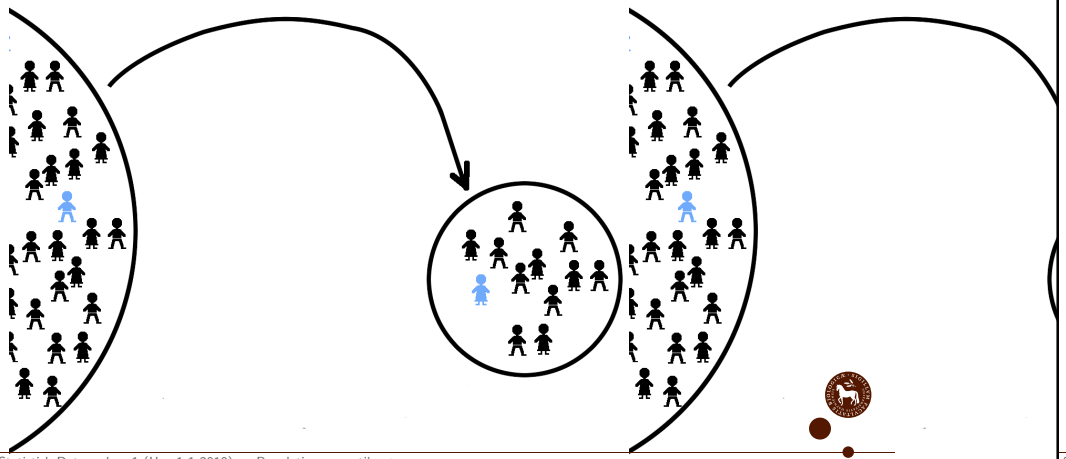


Populationer og stikprøver



Stikprøver

Der er variation i populationen — ikke alle er ens. Der er også variation i stikprøveudtagningen.



Datatyper

- Kategoriske data
  - Nominale — {Mand, kvinde}, {Gul, grøn, blå}.
  - Ordinale — {Ingen, lidt, mellem, meget}, socialklasser.
- Kvantitative data
  - Diskrete — unger pr. kuld, antal familiemedlemmer.
  - Kontinuerte — længde, højde, vægt, alder, ...

Datatyper — eksempel

Halthed hos kvæg 72 timer efter indtagelse af sukker.

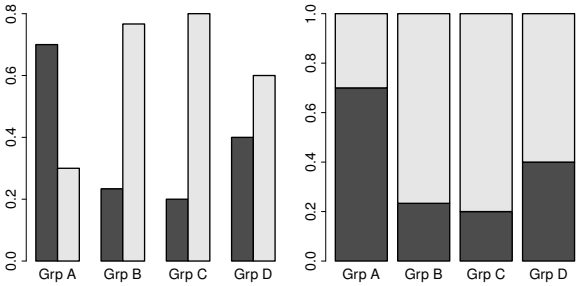
Sted	Vægt (kg)	Halthedsscore	Antal hævede led
I	276	Mildly lame	2
I	395	Mildly lame	1
I	356	Normal	0
I	437	Lame	2
II	376	Lame	0
II	350	Moderately lame	0
II	331	Lame	1
II	331	Normal	0

Visualisering — kategoriske data

Frekvensen = hyppigheden eller antal forekomster.

Hvis n er antallet af observationer er den relative frekvens = $\frac{\text{frekvensen}}{n}$.

	Group A	Group B	Group C	Group D	Total
TD present	21	7	6	12	46
TD absent	9	23	24	18	74



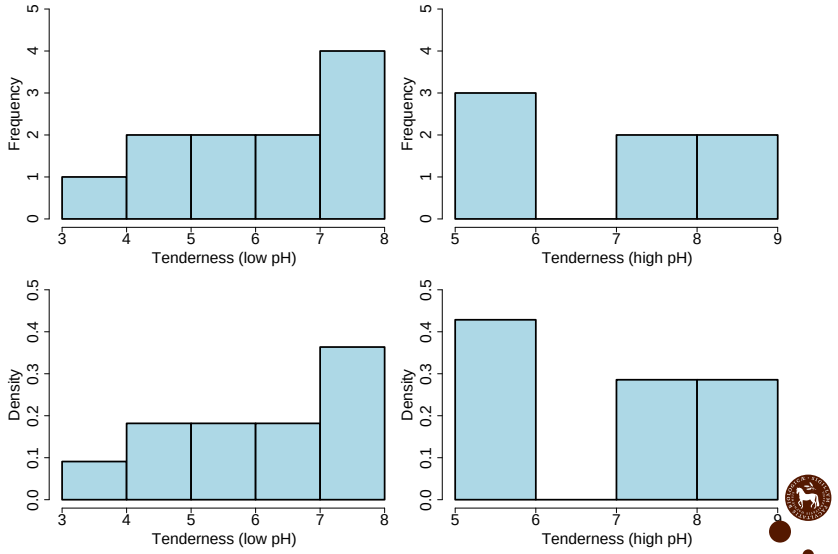
Visualisering — kvantitative data

pH group	Tunnel cooling	Rapid cooling
high	8.44	8.44
high	7.11	6.00
high	6.00	5.78
high	7.56	7.67
low	7.22	5.56
high	5.11	4.56
low	3.11	3.33
high	8.67	8.00
low	7.44	7.00
low	4.33	4.89
low	6.78	6.56
low	5.56	5.67
low	7.33	6.33
low	4.22	5.67
high	5.78	7.67
low	5.78	5.56
low	6.44	5.67
low	8.00	5.33

Mørhed af svinekød ved forskellige frysemetoder.

- Hvordan sammenlignes resultater fra de to metoder lettest?

Visualisering — kvantitative data



Median, range og kvartiler

Medianen er den "midterste observation", hvis man rangordner sine data. Er der et lige antal observationer er medianen midt imellem de to midterste observationer:

$$\text{Median} = \begin{cases} y_{(\frac{n+1}{2})} & \text{hvis } n \text{ er ulige} \\ \frac{1}{2}[y_{(n/2)} + y_{(n/2+1)}] & \text{hvis } n \text{ er lige} \end{cases}$$

Range er defineret som den største minus den mindste observation:

$$\text{Range} = y_{(n)} - y_{(1)}$$

Kvartiler deler datasættet op i fire grupper således at de mindste 25%, 50%, 75% og 100% af observationerne er i hhv. 1., 2., 3., og 4. kvartil (benævnt $Q_1, \ldots, Q_4$).

Inter-quartile range svarer til range over de midterste 50% af observationerne:

$$Q_3 - Q_1$$

Boxplots

En outlier er en observation, der ikke "passer så godt overens" med de øvrige observationer. Formelt defineres outliers som observationer, der er udenfor intervallet

$$[Q_1 - 1.5 \cdot \text{IQR}; Q_3 + 1.5 \cdot \text{IQR}].$$

Et boxplot bruges til at illustrere en fordeling grafisk ved at plotte de 5 mål: minimum, Q_1, median, Q_3 og maximum.

I et modificeret boxplot er minimum og maximum erstattet med hhv. den mindste og største observerede værdi, som er indeholdt i intervallet

$$[Q_1 - 1.5 \cdot \text{IQR}; Q_3 + 1.5 \cdot \text{IQR}].$$

Observationer udenfor intervallet markeres som punkter.

Middelværdi og spredning

Stikprøvemiddelværdien er defineret ved:

$$\bar{y} = \frac{\sum_i y_i}{n} = \frac{y_\bullet}{n}$$

Stikprøvespredningen er defineret ved:

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

Stikprøvevariansen er givet som stikprøvespredningen kvadreret.

Bemærk, at både middelværdien og spredningen har samme enhed som de målinger, der beregnes ud fra.

Infobox 1.1

Lineær transformation af middelværdi og spredning

Let $\bar{y}$ and s be the sample mean and sample standard deviation from observations $y_1, \ldots, y_n$ and let $y_i' = c \cdot y_i + b$ be a linear transformation of the y's with constants b and c. Then $\bar{y}' = c \cdot \bar{y} + b$ and $s' = |c| \cdot s$.

- Simple lineære transformation har *ingen* betydning.
  - Vi kan tillade os at gange og addere uden at det resulterer i "spøjse" ændringer af middelværdien eller spredningen.

Median eller middelværdi?

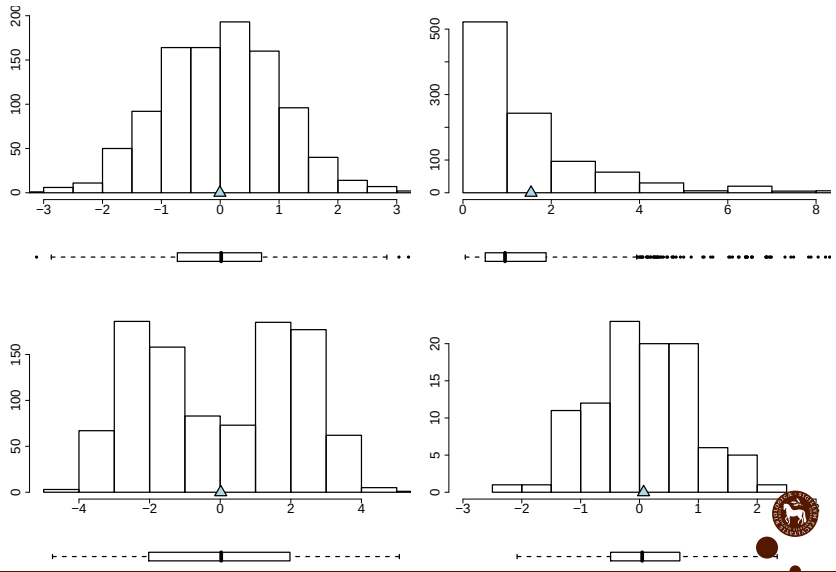
Median / IQR

- Medianen deler datasættet op i to lige store dele
- Ordinale og kvantitative data
- Ikke følsom for outliers
- IQR finder de "midterste 50%" af data.

Middelværdi / spredning

- Middelværdien deler datasættet op så afstanden fra centrum kommer i betragning.
- Kvantitative data (primært symmetriske data).
- Følsom for outliers.
- Spredningen er den "gennemsnitlige afstand til gennemsnittet".

Fire eksempler



Dagens hovedpunkter

- Populationer / stikprøver og inferens.
- Datatyper.
- Visualisering.
- Hvad er median, middelværdi, spredning, og IQR? Hvad fortæller de om data og hvordan fortolkes de?