Den basale statistiske tankegang

Læsning:

Jens Ledet Jensen kap. 1

Intuition om korrelation

- Autokorrelationen $E[X(t) \cdot X(t+\tau)]$
  - siger noget om, hvor meget signalet $X(t)$ ligner sig selv til tiden $X(t+\tau)$.
  - må afhænge af, hvor hurtigt signalet $X(t)$ ændrer sig over tid.
  - må være stor, hvis τ er lille.

- Krydskorrelationen $E[X(t) \cdot Y(t+\tau)]$
  - kan bruges til at lede efter steder (tidspunkter τ), hvor signalet $X(t)$ ligner signalet $\mathrm{Y}(t)$.

Autokorrelation

- Generelt

$$R_X(t_1, t_2) = E[X(t_1) \cdot X(t_2)] = E[X_1 \cdot X_2]$$
$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 \cdot x_2 \cdot f(x_1, x_2)\, dx_1 dx_2$$

- For en stationær proces

$$R_X(t_1, t_2) = R_X(t_1 + T, t_2 + T) = E[X(t_1 + T) \cdot X(t_2 + T)]$$

eller bare $R_X(\tau) = E[X(t) \cdot X(t + \tau)]$

Tidslig autokorrelation

- Generelt

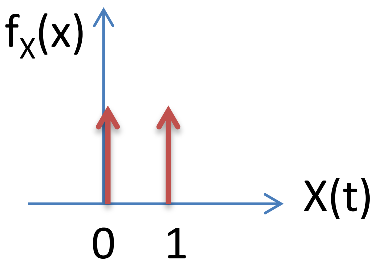
$$\mathcal{R}_X(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t) \cdot x(t+\tau) d\tau$$

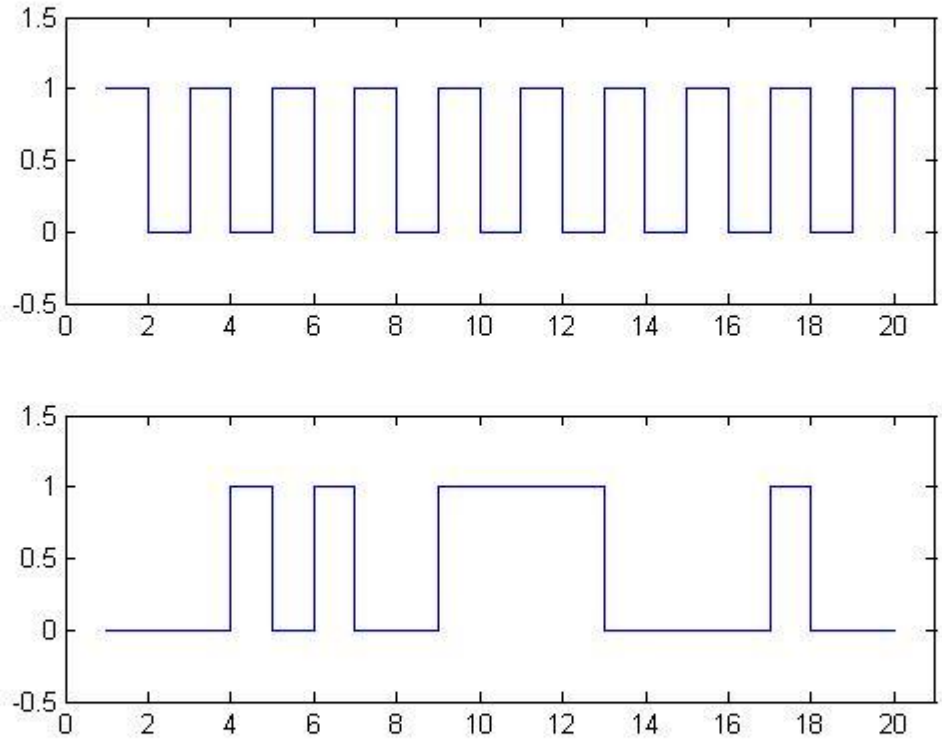
- Hvis processen er ergodisk

$$R_X(\tau) = \mathcal{R}_X(\tau)$$

Deterministisk vs. non-deterministisk

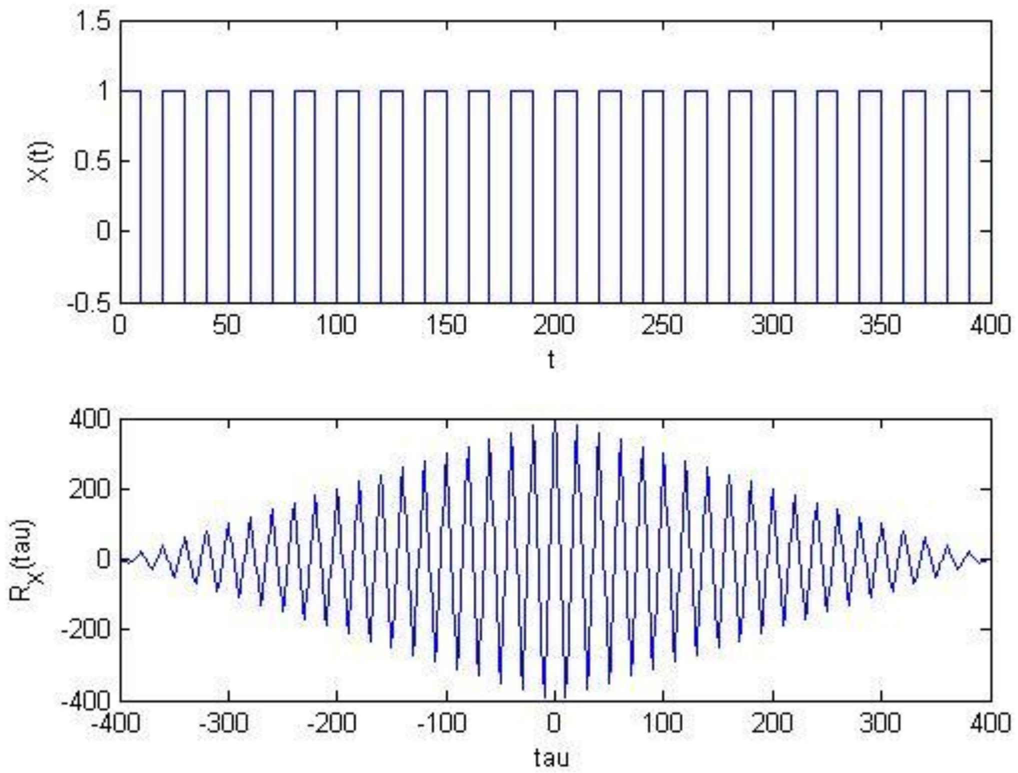
Tæthedsfunktion





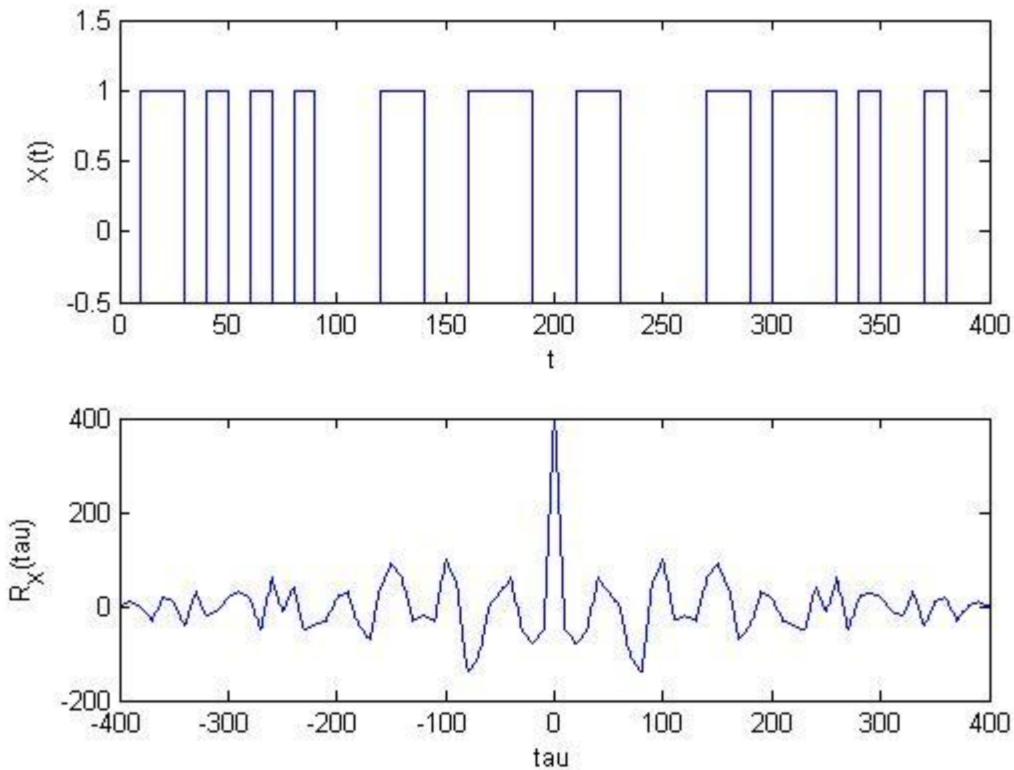
- Bemærk, at de to signaler har samme tæthedsfunktion.

Deterministisk vs. non-deterministisk



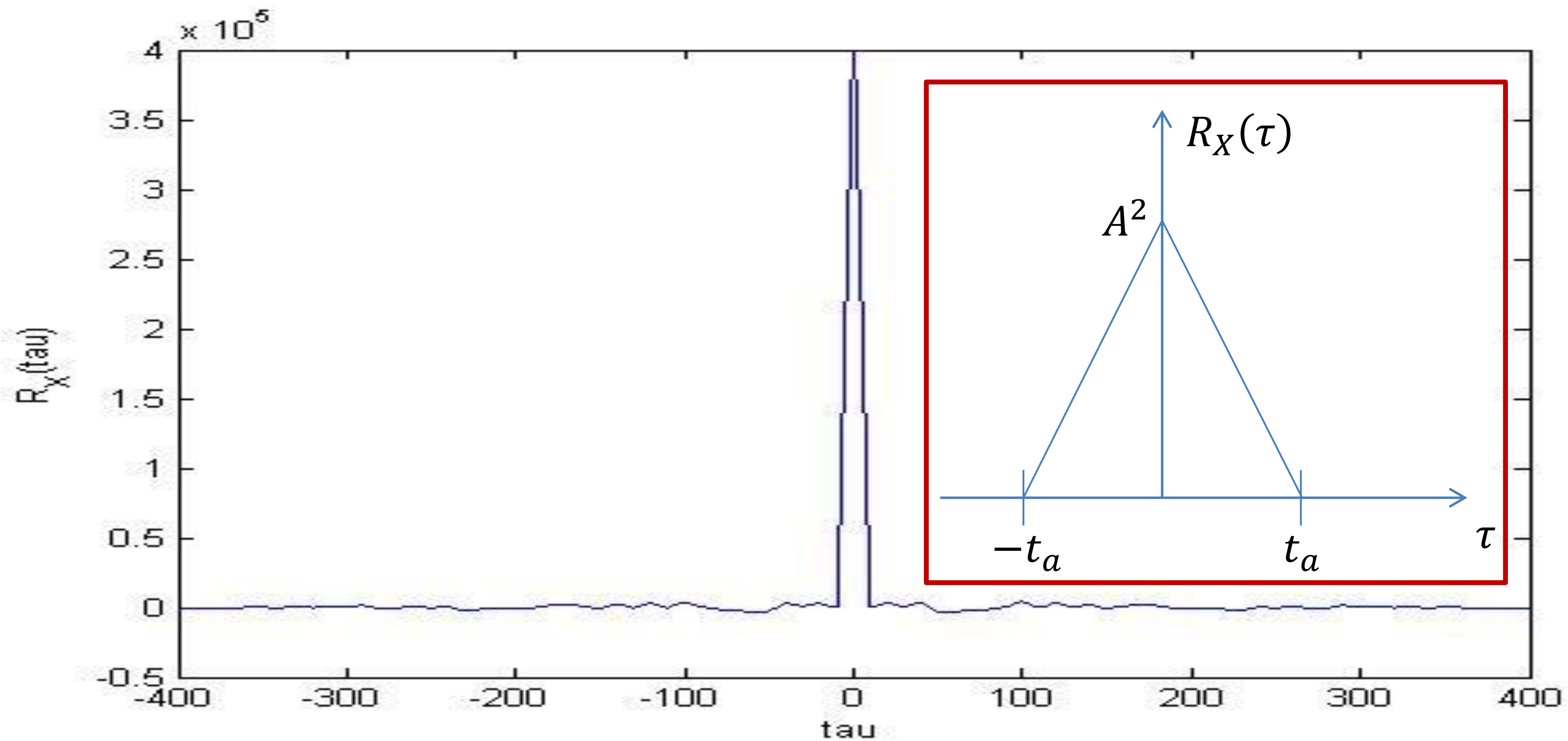

```
Rx = conv(x,fliplr(x));
```

Deterministisk vs. non-deterministisk




```
Rx = conv(x,fliplr(x));
```

Deterministisk vs. non-deterministisk



Autokorrelation midlet over 1000 simulationer

Krydskorrelation

- Generelt

$$R_{XY}(t_1, t_2) = E[X(t_1) \cdot Y(t_2)] = E[X_1 \cdot Y_2]$$
$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 \cdot y_2 \cdot f(x_1, y_2)\, dx_1 dy_2$$

- Hvis X og Y er indbyrdes stationære processer

$$R_{XY}(t_1, t_2) = R_{XY}(t_1 + T, t_2 + T) = E[X(t_1 + T) \cdot Y(t_2 + T)]$$

eller bare $R_{XY}(\tau) = E[X(t) \cdot Y(t + \tau)]$

Tidslig krydskorrelation

- Generelt

$$\mathcal{R}_{XY}(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t) \cdot y(t+\tau) d\tau$$

- Hvis processen er ergodisk

$$R_{XY}(\tau) = \mathcal{R}_{XY}(\tau)$$
$$R_{YX}(\tau) = \mathcal{R}_{YX}(\tau)$$

Opgave 6-8.3

```
% Problem 6-8.3
t1 = 0.0:0.001:0.099;
s1 = sin(100*pi*t1);
s = zeros(1,1000);
s(700:799) = s1;
n1 = randn(1,1000);
x = s + n1;

N = length(t1);
y = fliplr(s1);
z = conv(y,x)/(N+1);
z = z((1:1000)+length(y)/2);
tt = 0:0.001:0.999;
subplot(3,1,1); plot(tt,s); ylabel('Signal without noise');
subplot(3,1,2); plot(tt,x); ylabel('X');
subplot(3,1,3); plot(tt,z(1:1000)); xlabel('Time');
ylabel('R');
```

Opgave 6-8.4

```
% Problem 6_8_4
dt  = 1e-4;
f   = 1/dt;
T   = 1/400;
t   = -4*T:dt:4*T;
N   = length(t);

% a)
x1  = zeros(size(t));
x1(find(abs(t)<T/2))=1;
y1  = sin(2000*pi*t).*x1;
R1=conv(x1,fliplr(y1))/(N+1);
t1=(-(N-1):N-1)*1/f;
subplot(2,2,1), plot(t,x1,t,y1),xlabel('t'),ylabel('x(t) og y(t)')
subplot(2,2,2), plot(t1,R1); xlabel('tau'),ylabel('Ra');

% b)
x2  = sin(2000*pi*t).*x1;
y2  = cos(2000*pi*t).*x1;
R2=conv(x2,fliplr(y2))/(N+1);
t2=(-(N-1):N-1)*1/f;
subplot(2,2,3), plot(t,x2,t,y2),xlabel('t'),ylabel('x(t) og y(t)')
subplot(2,2,4), plot(t2,R2); xlabel('tau'),ylabel('Rb');
```

Vigtige regneregler

Egenskaber ved middelværdier og varianser

$$E[aX+b] = aE[X]+b \quad \text{middelværdien af } aX+b \tag{12}$$

$$\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X) \quad \text{variansen af } aX+b \tag{13}$$

$$E[X+Y] = E[X]+E[Y] \quad \text{linearitet af middelværdi} \tag{14}$$

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X)+\mathrm{Var}(Y)+2\cdot\mathrm{Cov}(X,Y) \tag{15}$$

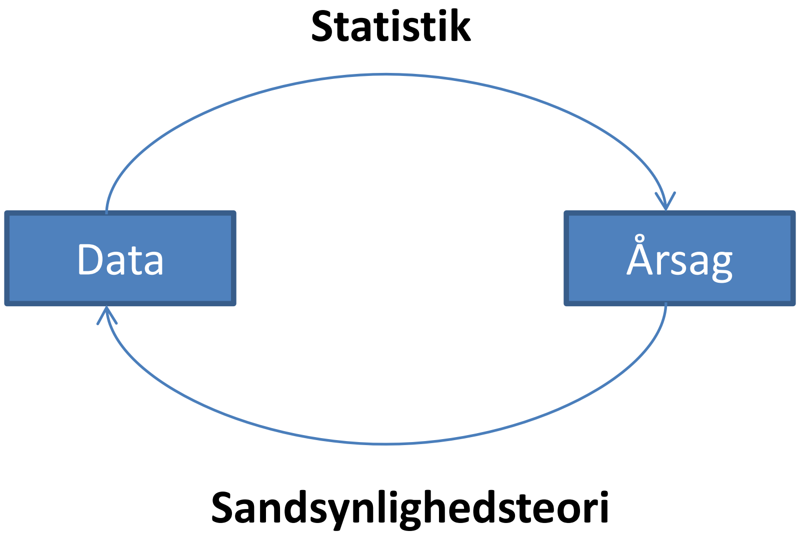
$$E[X\cdot Y] = E[X]\cdot E[Y] \quad \text{når } X \text{ og } Y \text{ er uafhængige} \tag{16}$$

$$\mathrm{Cov}(\mathrm{X},\mathrm{Y}) = E[(X-E[X])(Y-E[Y])] = E[X\cdot Y]-E[X]E[Y] \tag{17}$$

$$\mathrm{Corr}(\mathrm{X},\mathrm{Y}) = \frac{\mathrm{Cov}(\mathrm{X},\mathrm{Y})}{\sqrt{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}} \tag{18}$$

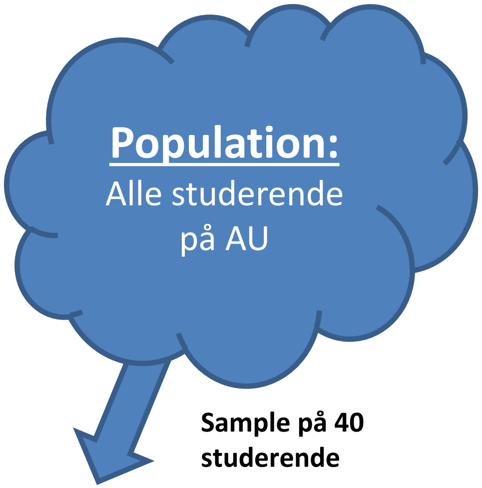
Den basale statistiske tankegang

Givet data, hvad er årsagen?



Givet årsagen, hvordan ser data ud?

Hypotesetest

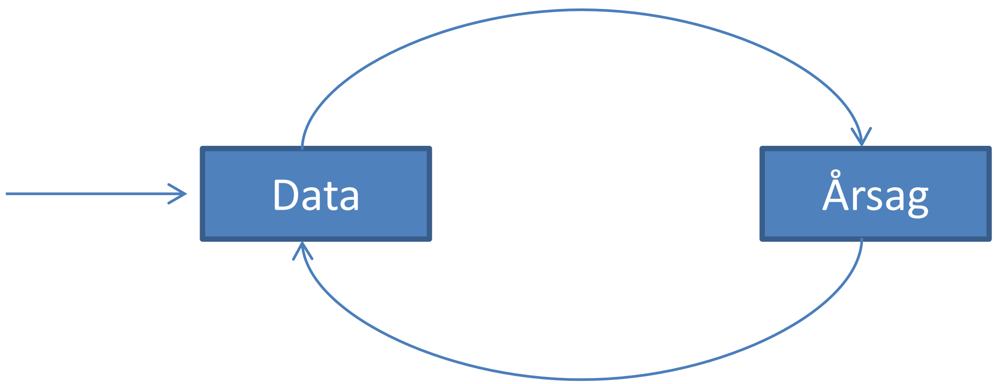


Sample på 40 studerende

Person	Alder
1	22
2	24
3	26
.	.
.	.
.	.
40	25

Hypotese:
Middelværdien af alderen i populationen er 50 år

Statistik



Sandsynlighedsteori

Hvis middelværdien af alderen er 50 år, hvordan bør data så se ud?

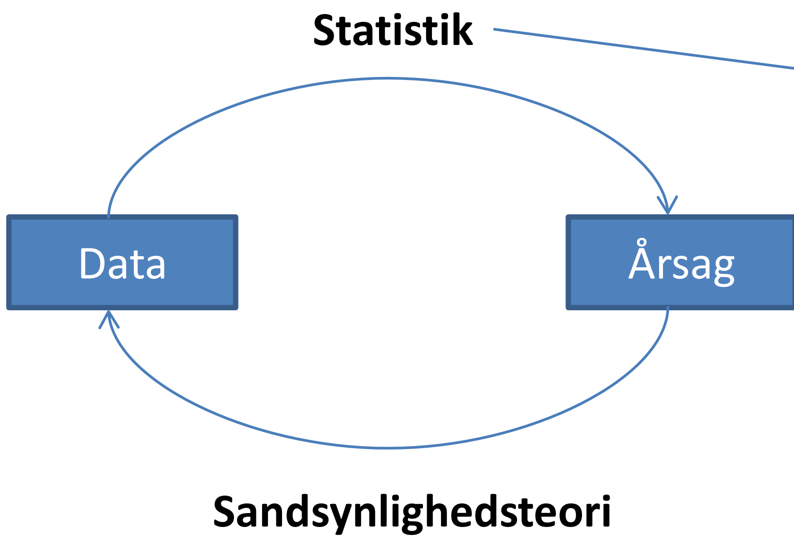
Hypotesetest

Population:
Alle studerende
på AU

Sample på 40 studerende

Person	Alder
1	22
2	24
3	26
.	.
.	.
.	.
40	25

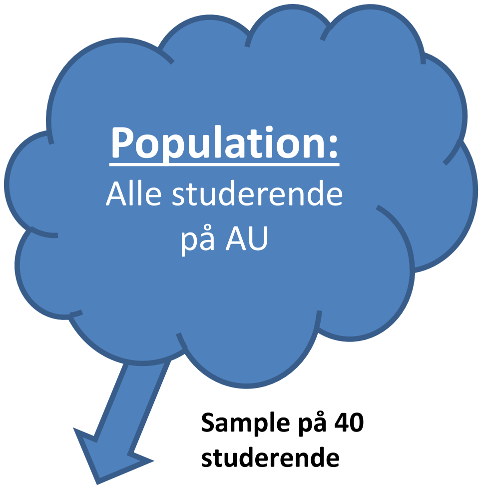
Notation:
H: $\mu = 50$



Teststørrelse (z)

Hvis middelværdien af alderen er 50 år, hvordan bør data så se ud?

Hypotesetest



Sample på 40 studerende

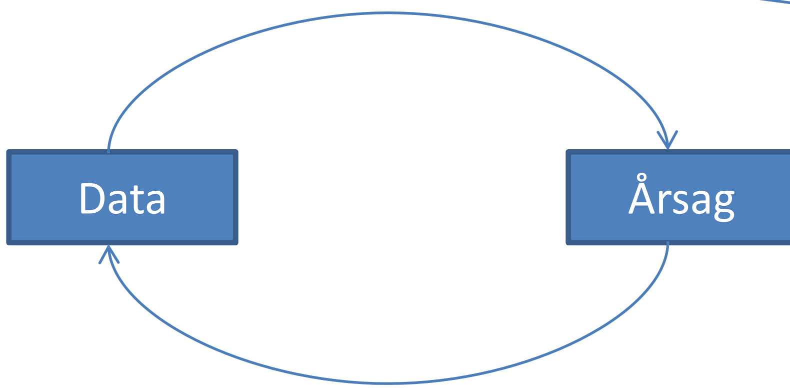
Person	**Alder**
1	22
2	24
3	26
.	.
.	.
.	.
40	25

Notation:
H: $\mu = 50$

Statistik

Gennemsnittet

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$



Sandsynlighedsteori

Hvis middelværdien af alderen er 50 år, hvordan bør data så se ud?

Hypotesetest

Population:
Alle studerende på AU

Sample på 40 studerende

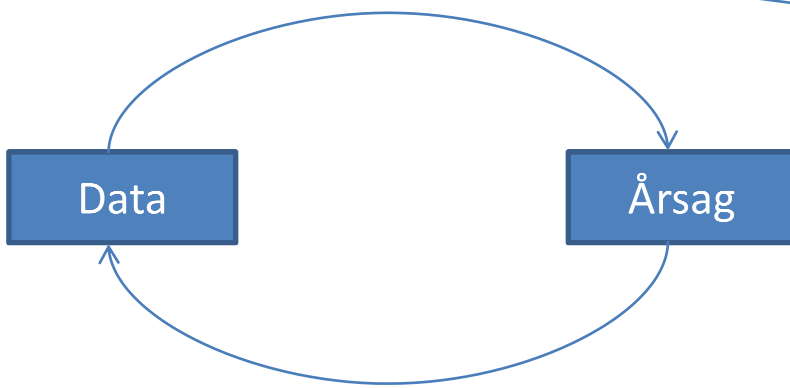
Person	**Alder**
1	22
2	24
3	26
.	.
.	.
.	.
40	25

Notation:
H: $\mu = 50$

Statistik

Gennemsnittet

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$



Sandsynlighedsteori

Fx fordelings-funktion $F_Z(z)$ for teststørrelsen under antagelse af, at hypotesen er korrekt.

Hvis middelværdien af alderen er 50 år, hvordan bør data så se ud?

Hypotesetest

Population:
Alle studerende på AU

Sample på 40 studerende

Person	Alder
1	22
2	24
3	2
.	.
.	.
.	.
40	2

Notation:
H: $\mu = 50$

Statistik

Gennemsnittet

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

ıgs-
z(z) for
sen
agelse af,
sen er

The Central Limit Theorem

If a random sample of size n is drawn from a population with mean μ and variance σ^2, then the sample mean $\overline{X}$ has approximately a normal distribution with mean μ and variance σ^2/n. That is, the distribution function of

$$\frac{(\overline{X} - \mu)}{\sigma/\sqrt{n}}$$

is approximately a standard normal. The approximation improves as the sample size increases.

Hypotesetest

Population:
Alle studerende på AU

Sample på 40 studerende

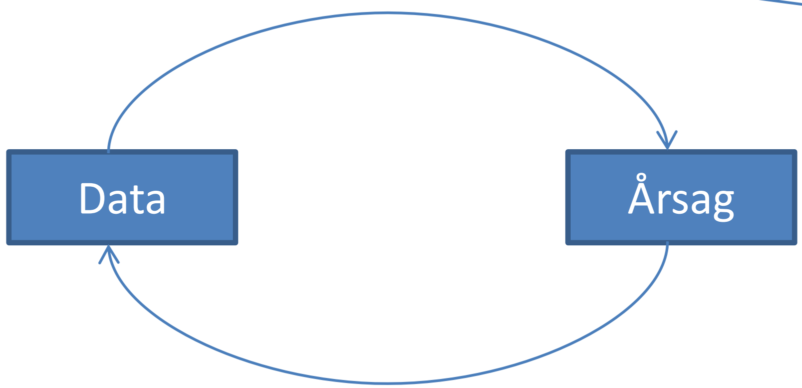
Person	**Alder**
1	22
2	24
3	26
.	.
.	.
.	.
40	25

Notation:
H: $\mu = 50$

Statistik

Gennemsnittet

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$



Sandsynlighedsteori

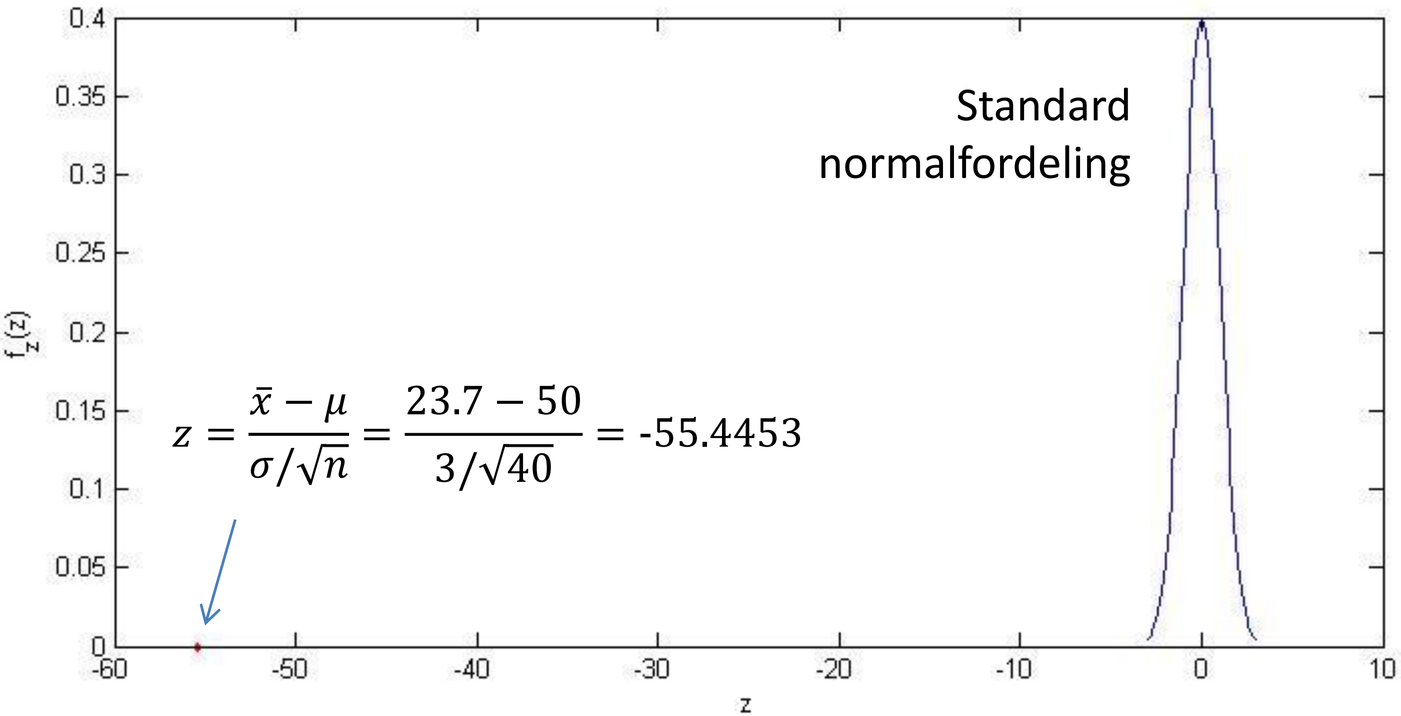
$$z \sim N(0,1)$$

Standard normalfordeling - følger af den centrale grænseværdisætning

Hvis middelværdien af alderen er 50 år, hvordan bør data så se ud?

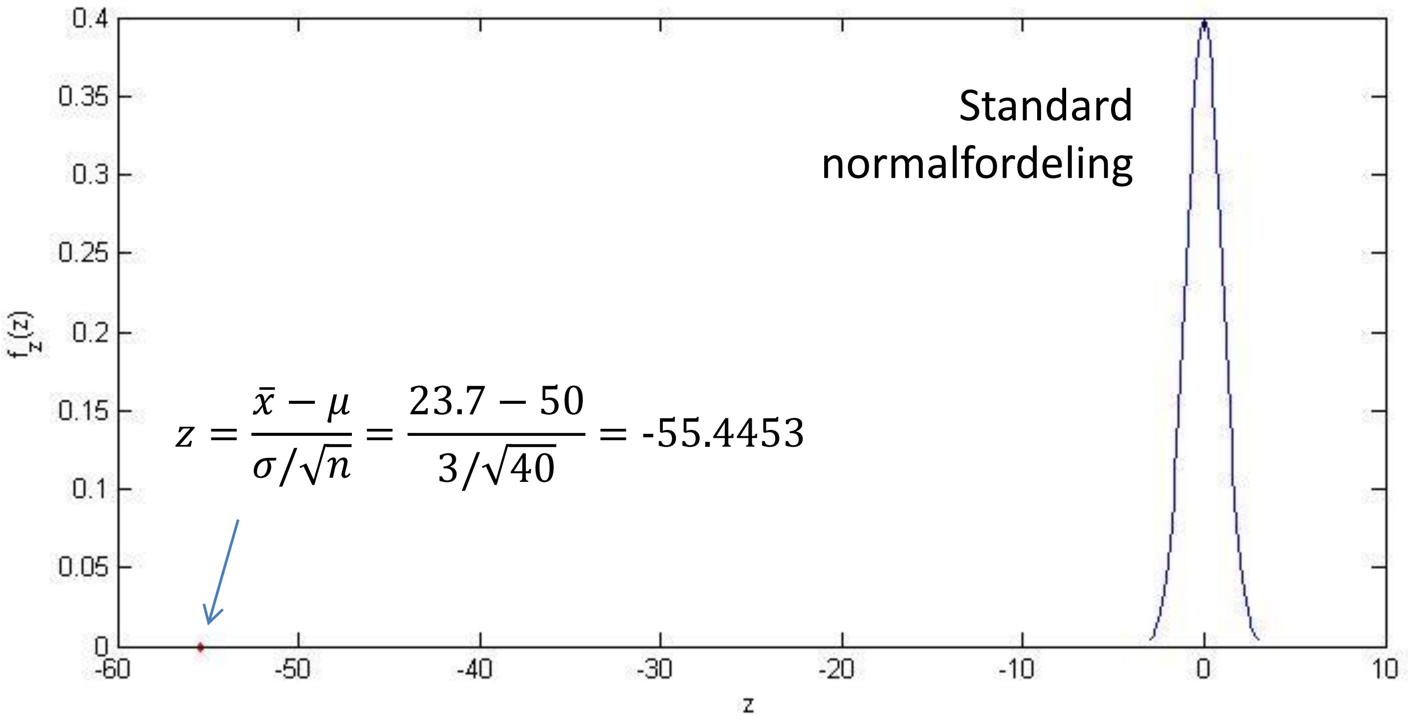
Hypotesetest

```
% Sand middelværdi og std. afvigelse
mu       = 24;
sigma    = 3;
% Samplede data
n        = 40;
x        = randn(1,n)*sigma + mu;
% Gennemsnit
xhat     = mean(x);
% Teststørrelse
mu_hyp = 50;
z = (xhat-mu_hyp)/(sigma/sqrt(n))
```



Hypotesetest

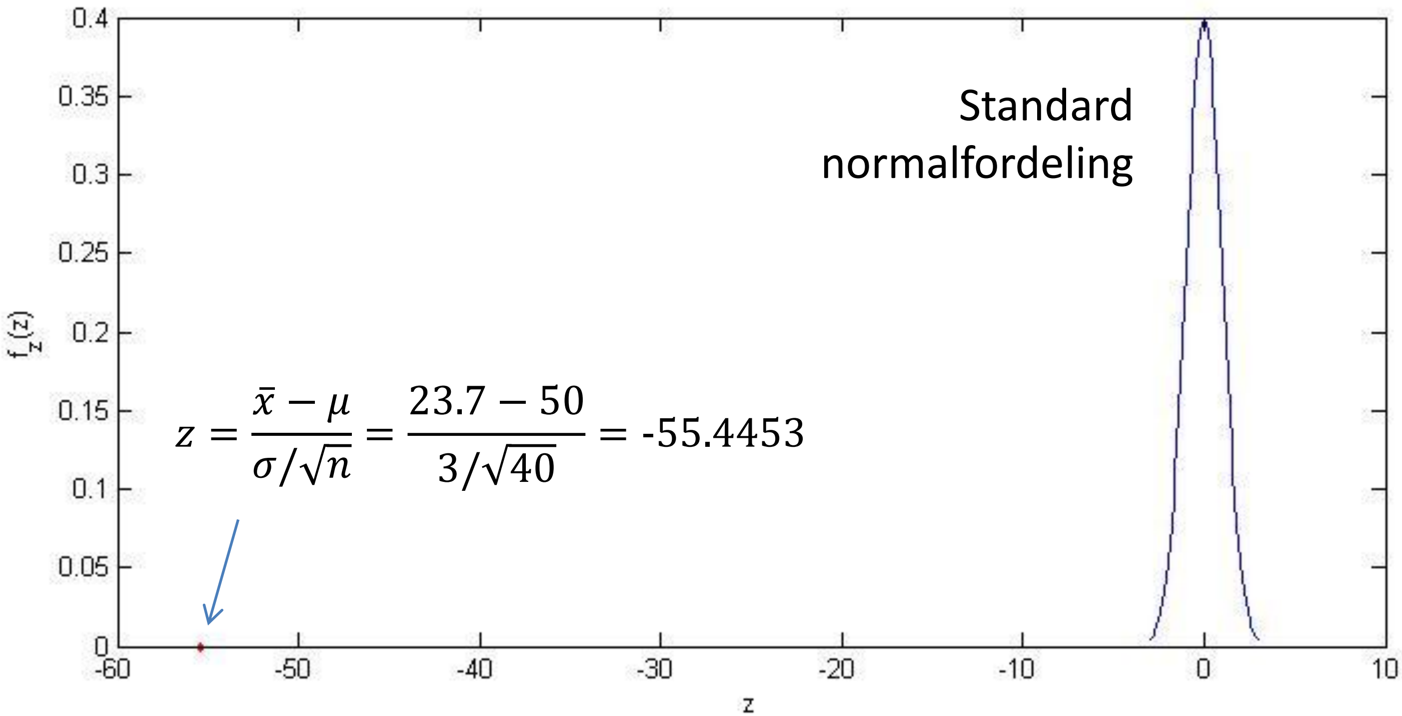
Virker det plausibelt, at vores teststørrelse *z* er et sample fra en standard normalfordeling?



$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{23.7 - 50}{3/\sqrt{40}} = \text{-}55.4453$$

Hypotesetest

<u>Det samme som at spørge:</u>
Hvad er sandsynligheden for at observere en teststørrelse, som er mere ekstrem end den, vi har observeret (altså z = -55.4453)



$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{23.7 - 50}{3/\sqrt{40}} = -55.4453$$

Hypotesetest

<u>Det samme som at spørge:</u>
Hvad er sandsynligheden for at observere en teststørrelse, som er mere ekstrem end den, vi har observeret (altså z = -55.4453)

<u>P-værdi:</u>

$$
\begin{aligned}
&\Pr(Z \leq -|z| \ \cup Z > |z|) \\
&= \Pr(Z \leq -55.4453) + \Pr(Z > 55.4453) \\
&= \Phi(-55.4453) + \left(1 - \Phi(55.4453)\right) \\
&= \left(1 - \Phi(55.4453)\right) + \left(1 - \Phi(55.4453)\right) \\
&= 2\left(1 - \Phi(55.4453)\right) \\
&= 2(1 - 1) \\
&= 0
\end{aligned}
$$

```
>> normcdf(55.4453)
ans =
     1
```

Hypotesetest

<u>Tolkning af p-værdi ≈ 0:</u>
Hvis eksperimentet gentages 100 gange, så vil vi – under hypotesen – aldrig observere en teststørrelse mere ekstrem end *z*.

<u>P-værdi:</u>

$$
\begin{aligned}
&\Pr(Z \leq -|z| \cup Z > |z|) \\
&= \Pr(Z \leq -55.4453) + \Pr(Z > 55.4453) \\
&= \Phi(-55.4453) + \big(1 - \Phi(55.4453)\big) \\
&= \big(1 - \Phi(55.4453)\big) + \big(1 - \Phi(55.4453)\big) \\
&= 2\big(1 - \Phi(55.4453)\big) \\
&= 2(1 - 1) \\
&= 0
\end{aligned}
$$

```
>> normcdf(55.4453)
ans =
     1
```

Hypotesetest

<u>Tolkning af p-værdi ≈ 0:</u>
Hvis eksperimentet gentages 100 gange, så vil vi – under hypotesen – aldrig observere en teststørrelse mere ekstrem end *z*.

<u>Derfor afviser vi hypotesen,</u>

H: $\boldsymbol{\mu = 50}$

Mendels ærteeksperiment

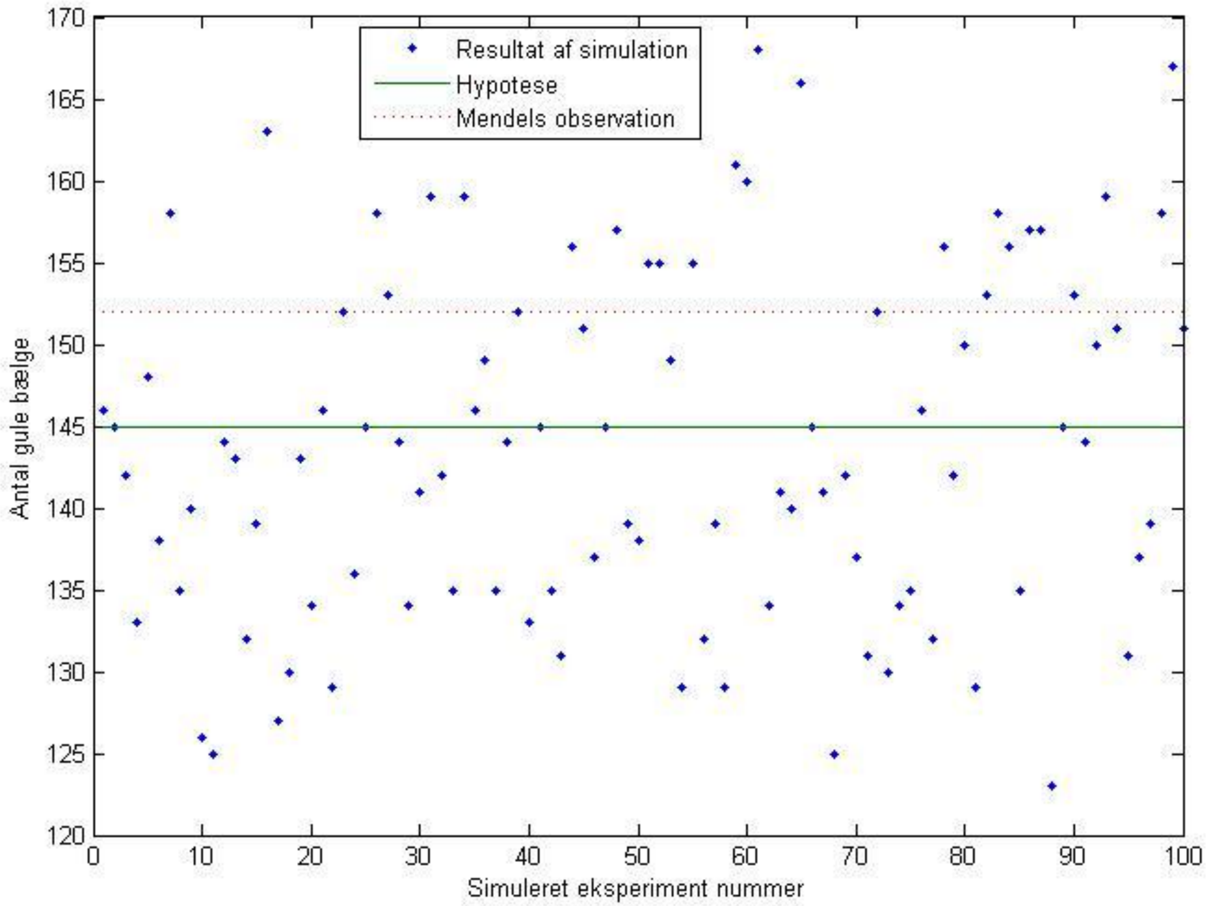
- Farven på den umodne ærtebælg styrres af to alleller:
  - A dominant (grøn)
  - a recessiv (gul)
- Genotyper
  - AA, Aa, aA → grøn bælg
  - aa → gul bælg
- Mendels hypotese:
  - Krydsning af Aa'er med sig selv skal give lige mange af de fire genotyper.
- Mere formel hypotese
  - H: Pr(gul bælg) = ¼
  - H: Pr(grøn bælg) = ¾

Mendels ærteeksperiment

- Mendel foretager et forsøg med 580 planter
- Resultat
  - 152 gule bælge
  - 428 grønne bælge
- Stemmer overens med Mendels hypotese?
- Observation
  - Hvis eksperimentet gentages fx 100 gange, så forventer vi at observere noget forskelligt hver gang.
- Den statistiske metode
  - Hvad ville man typisk observere under forudsætning af, at hypotesen er sand?
  - Svarer data til, hvad man typisk ville observere?

Mendels ærteeksperiment

- Resultat af 100 simulationer i Matlab (under hypotesen)



Målt værdi

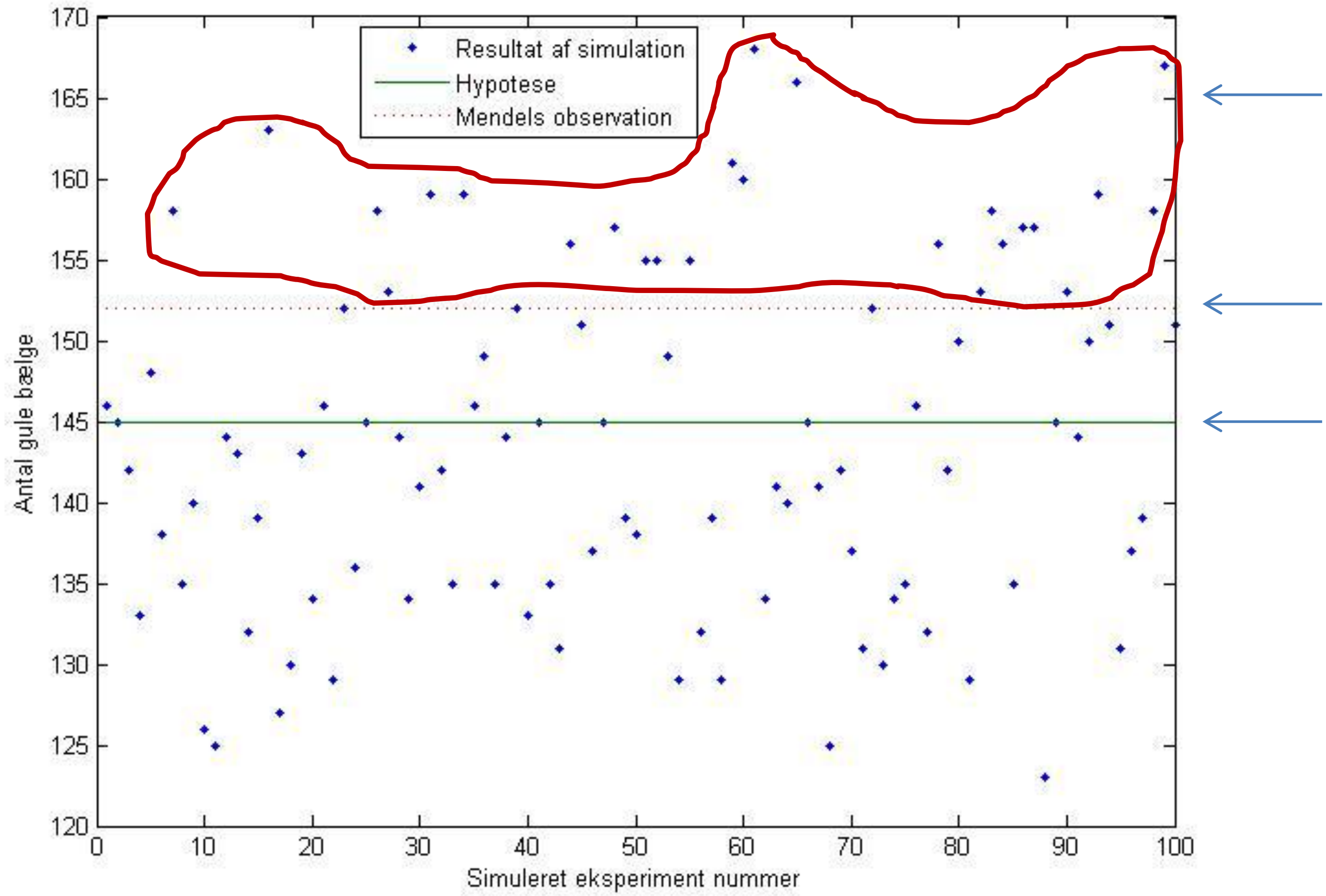
Teoretisk værdi

Mendels ærteeksperiment

```
%% Mendels eksperiment
n = 580;    % Antal planter
p = 1/4;    % Sandsynlighed for gul bælg (under hypotesen)
x = 152;    % Mendels observation
num_sim = 100;
antal_gule_planter = zeros(1,num_sim);
for i = 1:num_sim
    counter = 0;
    for plante = 1:580
        if rand(1)<=p
            counter = counter + 1;
        end
    end
    antal_gule_planter(i) = counter;
end
plot(1:num_sim,antal_gule_planter,'.',...
     1:num_sim,ones(1,num_sim)*n*p,...
     1:num_sim,ones(1,num_sim)*x,':')
xlabel('Simuleret eksperiment nummer')
ylabel('Antal gule bælge')
legend('Resultat af simulation','Hypotese','Mendels observation')
```

Mendels ærteeksperiment

- Resultat af 100 simulationer i Matlab (under hypotesen)



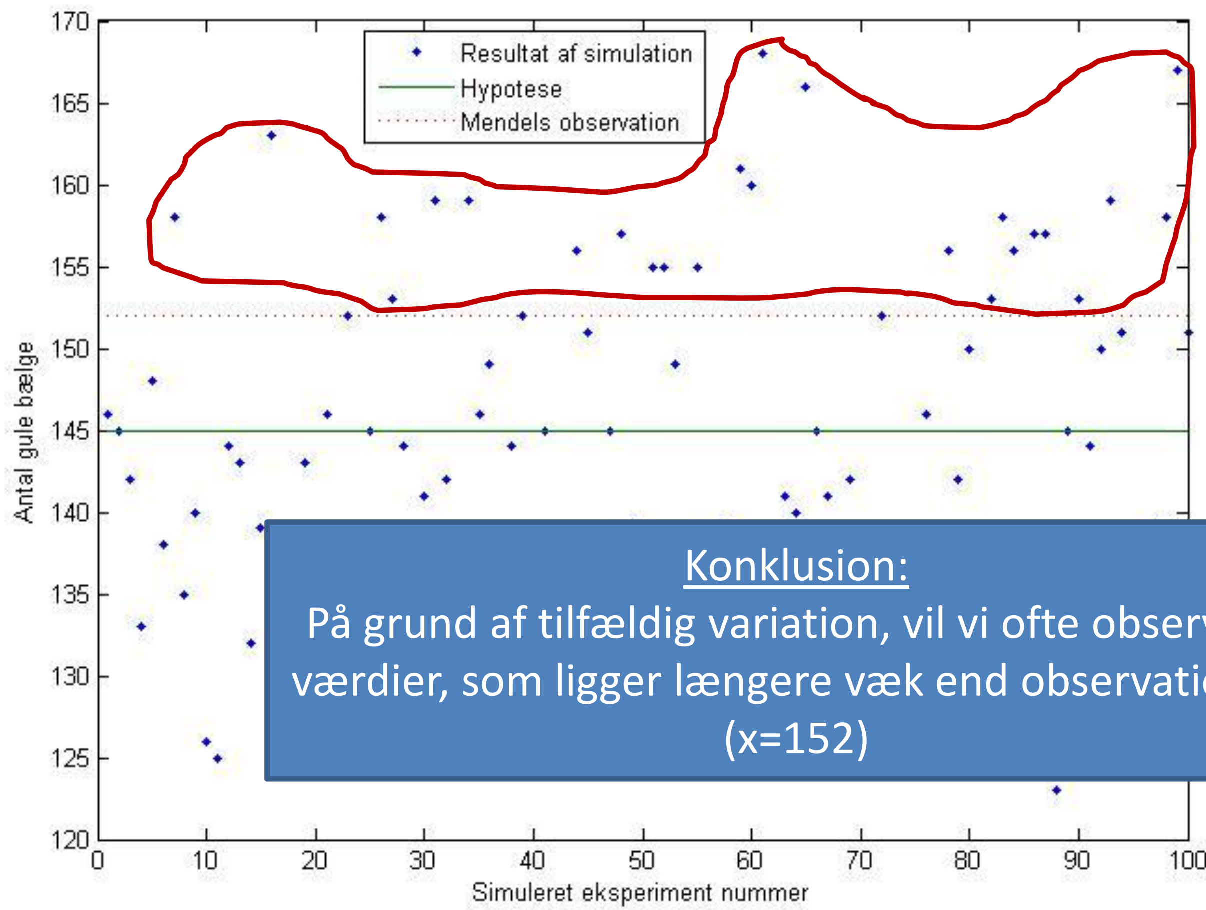
25 ud af 100 simulerede værdier er mere ekstreme end den målte værdi.

Målt værdi

Teoretisk værdi

Mendels ærteeksperiment

- Resultat af 100 simulationer i Matlab (under hypotesen)



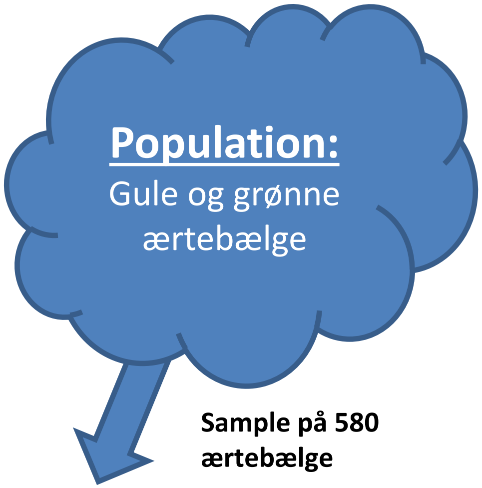
25 ud af 100 simulerede værdier er mere ekstreme end den målte værdi.

Målt værdi

Teoretisk værdi

<u>Konklusion:</u>
På grund af tilfældig variation, vil vi ofte observere værdier, som ligger længere væk end observationen (x=152)

Hypotesetest



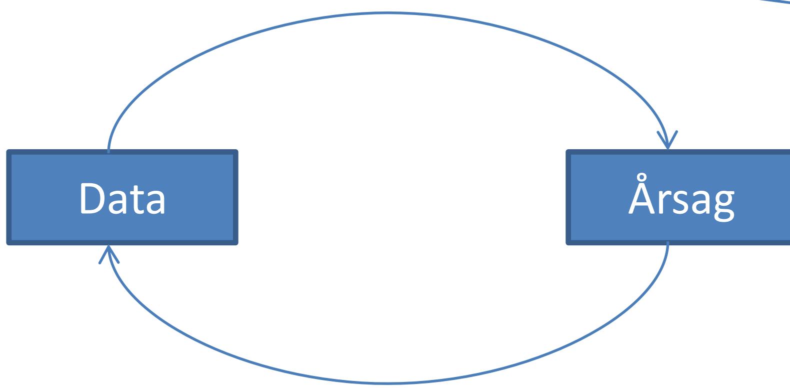
Sample på 580 ærtebælge

Bælg	**Farve**
1	grøn
2	gul
3	grøn
.	.
.	.
.	.
580	grøn

H: $\mathbf{Pr(gul)} = \mathbf{1/4}$

Statistik

Hvilken teststørrelse?



Hvilken fordeling?

Sandsynlighedsteori

Hvis $\mathbf{Pr(gul)} = \mathbf{1/4}$,
hvordan bør data så se ud?

Bernoullifordelingen

- To mulige udfald
  - B = {0,1}

- Sandsynligheder
  - Pr(B=1) = *p* (succes)
  - Pr(B=0) = 1-*p* (failure)

- Notation

$$B \sim bernoulli(p)$$

Binomialfordelingen

- Lad $B_1, B_2, \dots, B_n$ være uafhængige stokastiske variable, hvor

$$B_i \sim bernoulli(p)$$

- Så er antallet af successer

$$X = \sum_{i=1}^{n} B_i$$

binomialfordelt med antalsværdi *n* og sandsynlighedsparameter *p*.

- Notation

$$X \sim binomial(n, p)$$

Binomialfordelingen (fortsat)

- Notation

$$X \sim binomial(n, p)$$

- For en konkret måling (X=k) beregnes sandsynligheden ud fra binomialformlen:

$$\left.\begin{aligned} \Pr(X = k) &= \Pr_n(k) \\ &= \Pr(k \text{ succeser ud af } n \text{ forsøg}) \\ &= \frac{n!}{k!(n-k)!} p^k q^{n-k} \\ &= \binom{n}{k} p^k q^{n-k} \end{aligned}\right\}$$

Formel 1-29
i Cooper/McGillem

Mendels ærteeksperiment

- Data antages binomialfordelte

$$X \sim binomial(n = 580, p = 1/4)$$

Hypotesetest

Population:
Gule og grønne ærtebælge

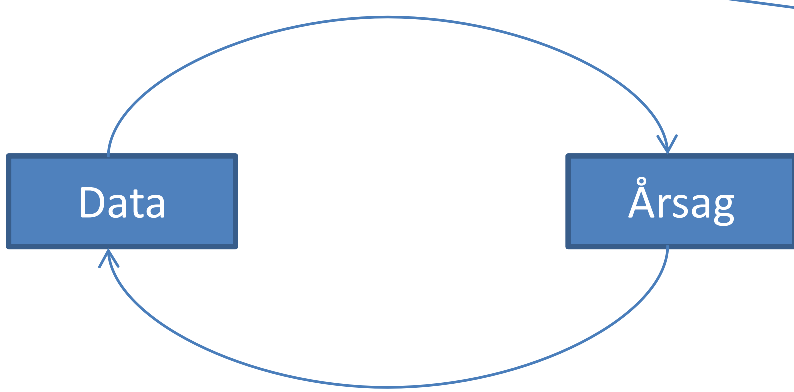
Sample på 580 ærtebælge

Bælg	Alder
1	grøn
2	gul
3	grøn
.	.
.	.
.	.
580	grøn

H: $p = 1/4$

Statistik

x = antal successer = 152



$x \sim binomial(580, 1/4)$

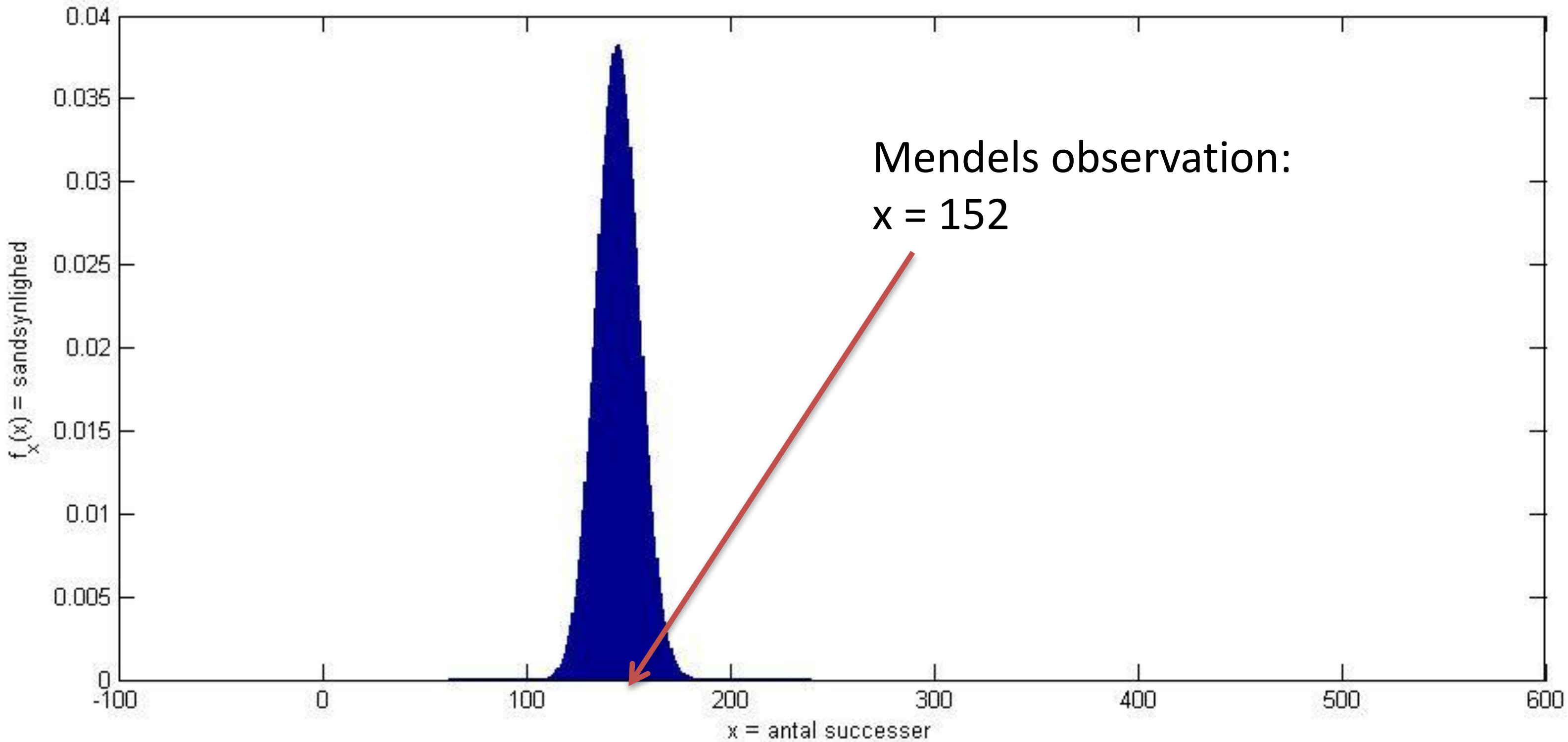
Sandsynlighedsteori

Hvis $p = 1/4$ og $n = 580$, hvordan bør data så se ud?

Hypotesetest

- Data antages binomialfordelte

$$X \sim binomial(n = 580, p = 1/4)$$



P-værdi

$$
\begin{aligned}
&\Pr(X \le np - |np - x| \cup X > np + |np - x|)\\
&= \Pr(X \le 145 - |145 - 152|) + \Pr(X > 145 + |145 - 152|)\\
&= \Pr(X \le 145 - 7) + \Pr(X > 145 + 7)\\
&= \Pr(X \le 138) + \Pr(X > 152)\\
&= F_{binomial}(138) + \left(1 - F_{binomial}(152)\right)\\
&= 0.50
\end{aligned}
$$

```
n = 580;
p = 1/4;
x = 152;     % Observation
lower = n*p-abs(n*p-x)
upper = n*p+abs(n*p-x)
pval = binocdf(lower,n,p) + (1 - binocdf(upper,n,p))

lower =   138
upper =   152
pval =    0.5030
```

Hypotesetest

Tolkning af p-værdi ≈ 0:
Hvis eksperimentet gentages 100 gange, så vil vi – under hypotesen – aldrig observere en teststørrelse mere ekstrem end *z*.

Tolkning af p-værdi ≈ 0.5:
Hvis eksperimentet gentages 100 gange, så vil vi – under hypotesen – halvdelen af gangene observere en teststørrelse mere ekstrem end *z*.

Derfor afviser vi ikke hypotesen,

H: $\boldsymbol{p = 1/4}$

Uddybning af binomialfordelingen

- Hvis

$$X_1 \sim binomial(n_1, p)$$

og

$$X_2 \sim binomial(n_2, p)$$

- Så er

$$X_1 + X_2 \sim binomial(n_1 + n_2, p)$$

- Praktisk anvendelse:
  - Kombinere målinger fra to (eller flere) eksperimenter.

Uddybning af binomialfordelingen

- Lad os først lige kigge på Bernoulli fordelingen

- Middelværdi

$$E[X] = \sum_{z\in\{0,1\}} z \cdot \Pr(X = z)$$
$$= 0 \cdot \Pr(X = 0) + 1 \cdot \Pr(X = 1)$$
$$= 0 \cdot (1 - \mathrm{p}) + 1 \cdot p$$
$$= p$$

- Varians

$$\mathrm{Var}(X) = \sum_{z\in\{0,1\}} (z - p)^2 \cdot \Pr(X = z)$$
$$= (0 - p)^2 \cdot \Pr(X = 0) + (1 - p)^2 \cdot \Pr(X = 1)$$
$$= p^2 \cdot (1 - \mathrm{p}) + (1 - p)^2 \cdot p$$
$$= p(1 - p)$$

Uddybning af binomialfordelingen

- Husk, at

$$X = \sum_{i=1}^{n} B_i$$

hvor $B_i \sim bernoulli(p)$ og uafhængige.

- Middelværdi

$$E[X] = E\left[\sum_{i=1}^{n} B_i\right] = \sum_{i=1}^{n} E[B_i] = n \cdot p$$

- Varians

$$Var(X) = Var\left(\sum_{i=1}^{n} B_i\right) = \sum_{i=1}^{n} Var(B_i) = n \cdot p(1-p)$$

$$\boxed{\begin{aligned} E[X+Y] &= E[X]+E[Y] && \text{linearitet af middelværdi} && (14) \\ \mathrm{Var}(X+Y) &= \mathrm{Var}(X)+\mathrm{Var}(Y)+2\cdot\mathrm{Cov}(X,Y) && && (15) \end{aligned}}$$

Uddybning af binomialfordelingen

- Standardisering
- Hvis

$$X \sim binomial(n, p)$$

og $n \cdot p > 5$ og $n \cdot (1 - p) > 5$.

- Så er

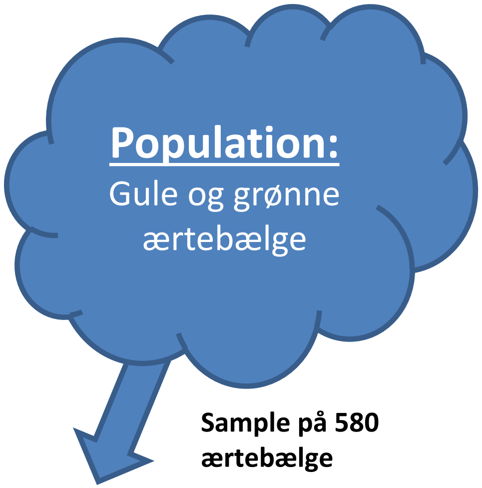
$$\Pr(X \leq k) = F_{binomial}(k) \approx \Phi\left(\frac{k - np}{\sqrt{np(1-p)}}\right)$$

Approximativ p-værdi

$$
\begin{aligned}
&\Pr(X \leq np - |np - x| \cup X > np + |np - x|) \\
&= \Pr(X \leq 145 - |145 - 152|) + \Pr(X > 145 + |145 - 152|) \\
&= \Pr(X \leq 145 - 7) + \Pr(X > 145 + 7) \\
&= \Pr(X \leq 138) + \Pr(X > 152) \\
&\cancel{= F_{binomial}(138) + \left(1 - F_{binomial}(152)\right)} \\
&= \Phi\left(\frac{138 - 1/4 \cdot 580}{\sqrt{580 \cdot 1/4 \cdot (1 - 1/4)}}\right) + \left(1 - \Phi\left(\frac{152 - 1/4 \cdot 580}{\sqrt{580 \cdot 1/4 \cdot (1 - 1/4)}}\right)\right) \\
&= 0.50
\end{aligned}
$$

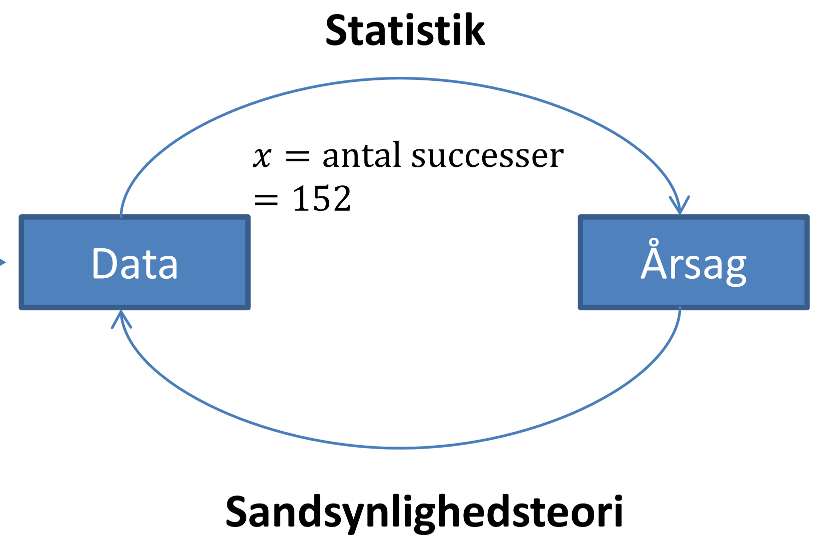
```
%% Approksimativ P-værdi
n = 580;
p = 1/4;
x = 152;    % Observation
lower = n*p-abs(n*p-x)
upper = n*p+abs(n*p-x)
z_lower = (lower-n*p)/sqrt(n*p*(1-p))
z_upper = (upper-n*p)/sqrt(n*p*(1-p))
pval = normcdf(z_lower) + (1 - normcdf(z_upper))
```

Næste gang: Estimation



Sample på 580 ærtebælge

Bælg	**Alder**
1	grøn
2	gul
3	grøn
.	.
.	.
.	.
580	grøn



Givet data, hvad kan vi sige om parameteren p
- Estimat: $\hat{p} = ?$
- Usikkerhed: Konfidensinterval