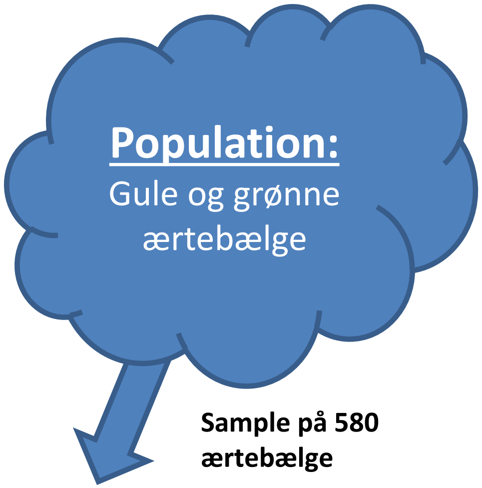
P-værdier og konfidensintervaller

Læsning:

Jens Ledet Jensen kap. 2+3

Hypotesetest

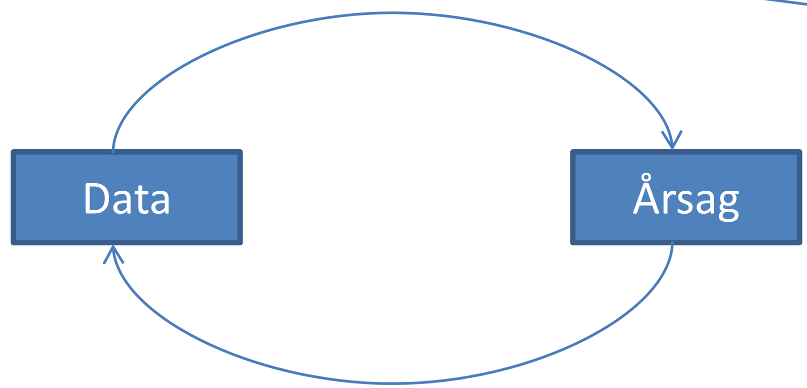


Sample på 580 ærtebælge

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

H: $p = 1/4$

Statistik



Sandsynlighedsteori

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

Teststørrelse:

x = antal successer = 152

Statistisk model:

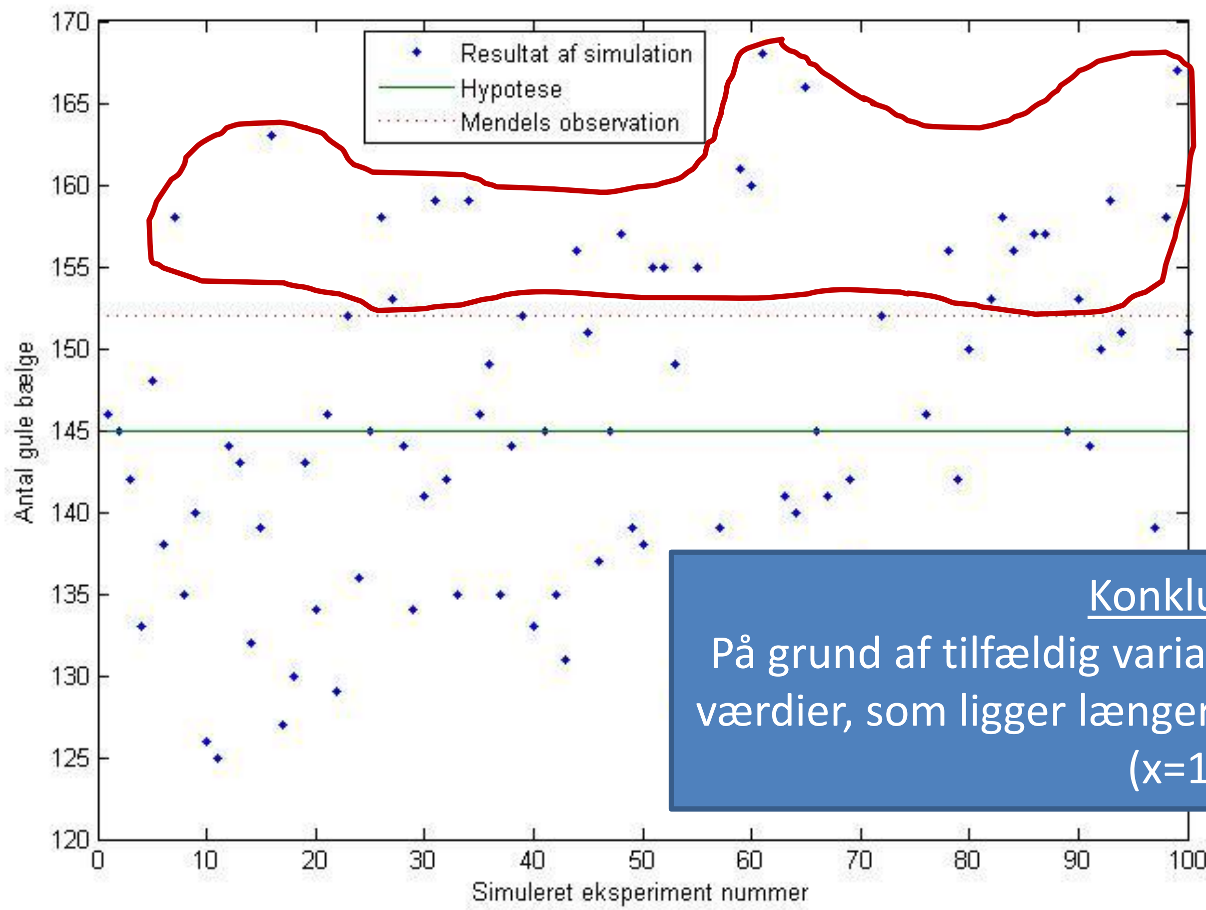
$x \sim binomial(580, 1/4)$

Hypotesetest

- Vi vurderer holdbarheden af en hypotese ved at sammenligne de observerede data med, hvad man typisk vil se, hvis hypotesen er sand.

Data simuleret på baggrund af hypotesen (H: p = ¼)

- Resultat af 100 simulationer i Matlab



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

Målt værdi

Middelværdi

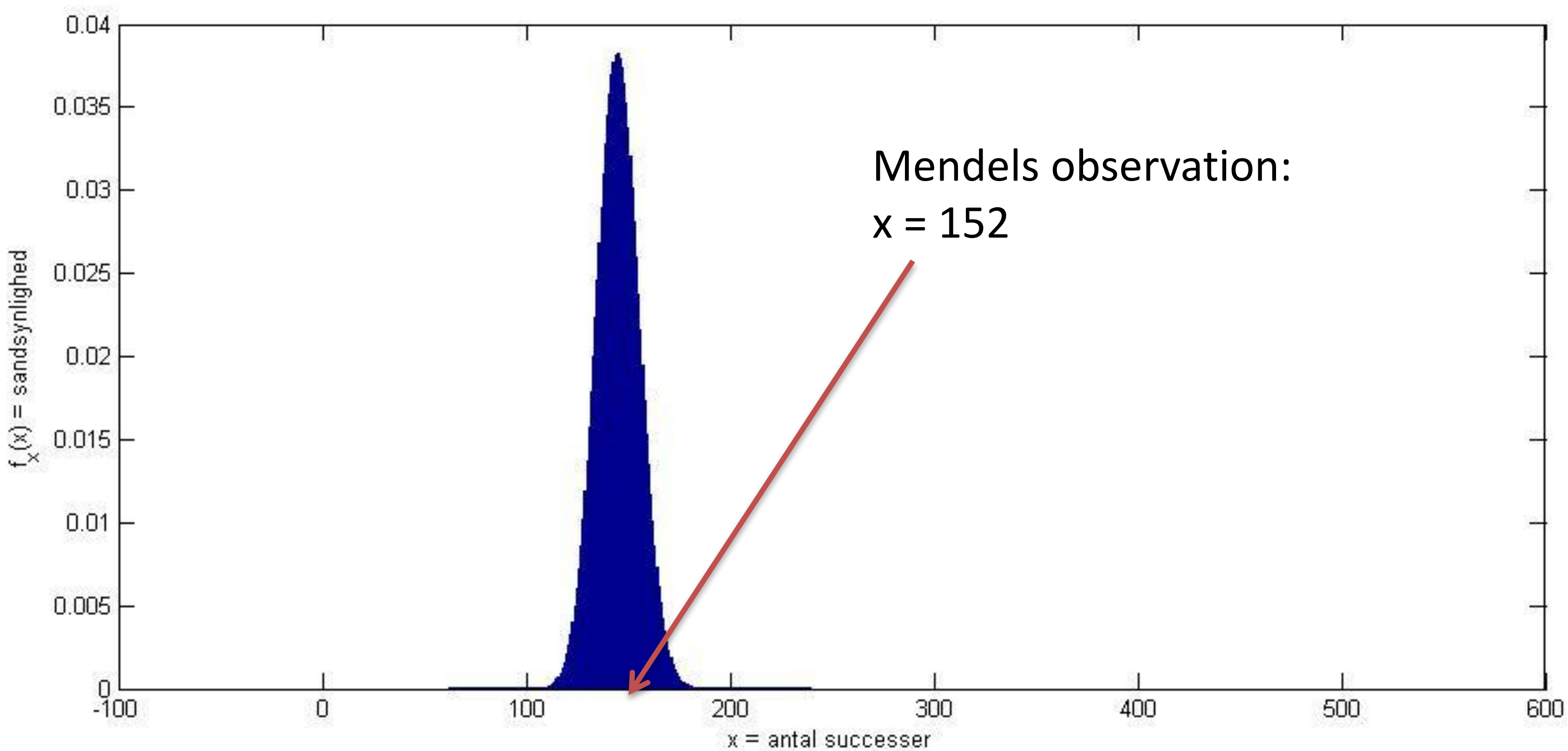
Konklusion:

På grund af tilfældig variation, vil vi ofte observere værdier, som ligger længere væk end observationen (x=152)

Teststørrelsen og dens fordeling

- Teststørrelsen antages binomialfordelt:

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



P-værdien

- Vi vurderer holdbarheden af en hypotese ved at sammenligne de observerede data med, hvad man typisk vil se, hvis hypotesen er sand.
- Dette gøres ved at beregne p-værdien.
- Stor p-værdi:
  - Data strider ikke mod hypotesen
- Lille p-værdi:
  - Data strider hypotesen

Beregning af p-værdi

Sandsynligheden for at observere en teststørrelse, som er mere ekstrem end x = 152:

$$pval = \Pr(X \leq np - |np - x| \,\cup X > np + |np - x|)$$
$$= \Pr(X \leq 145 - |145 - 152|) + \Pr(X > 145 + |145 - 152|)$$
$$= \Pr(X \leq 145 - 7) + \Pr(X > 145 + 7)$$
$$= \Pr(X \leq 138) + \Pr(X > 152)$$
$$= F_{binomial}(138) + \left(1 - F_{binomial}(152)\right)$$
$$= 0.50$$

```
>> binocdf(152,580,1/4)
ans =
     0.7652
```

Signifikansniveau

- Når man laver et hypotesetest (og det hedder altså <u>et</u> test inden for statistisk...), må man vælge et passende niveau for p-værdien.
- Dette kaldes signifikansniveauet og betegnes α.
- Hvis p-værdien er større end α
  - Data strider ikke mod hypotesen
- Hvis p-værdien er mindre end α
  - Data strider mod hypotesen
- Typiske værdier for α er 0.05 eller 0.01.

Beregning af p-værdi

Sandsynligheden for at observere en teststørrelse, som er mere ekstrem end x = 152:

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

Da pval > 0.05, strider data ikke mod hypotesen.

<u>Konklusion på Mendels eksperiment:</u>
Data strider ikke imod antagelsen om lige udspaltning af de fire genotyper.

Normalfordelte data

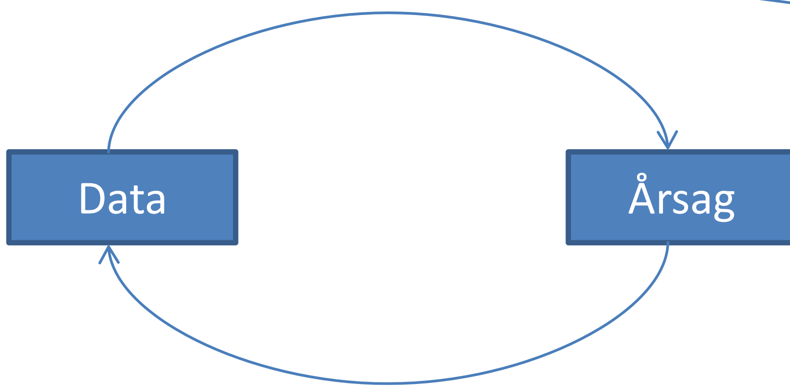
Population:
Alle studerende
på AU

Sample på 40 studerende

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

Notation:
H: $\mu = 50$

Statistik



Sandsynlighedsteori

Teststørrelse:

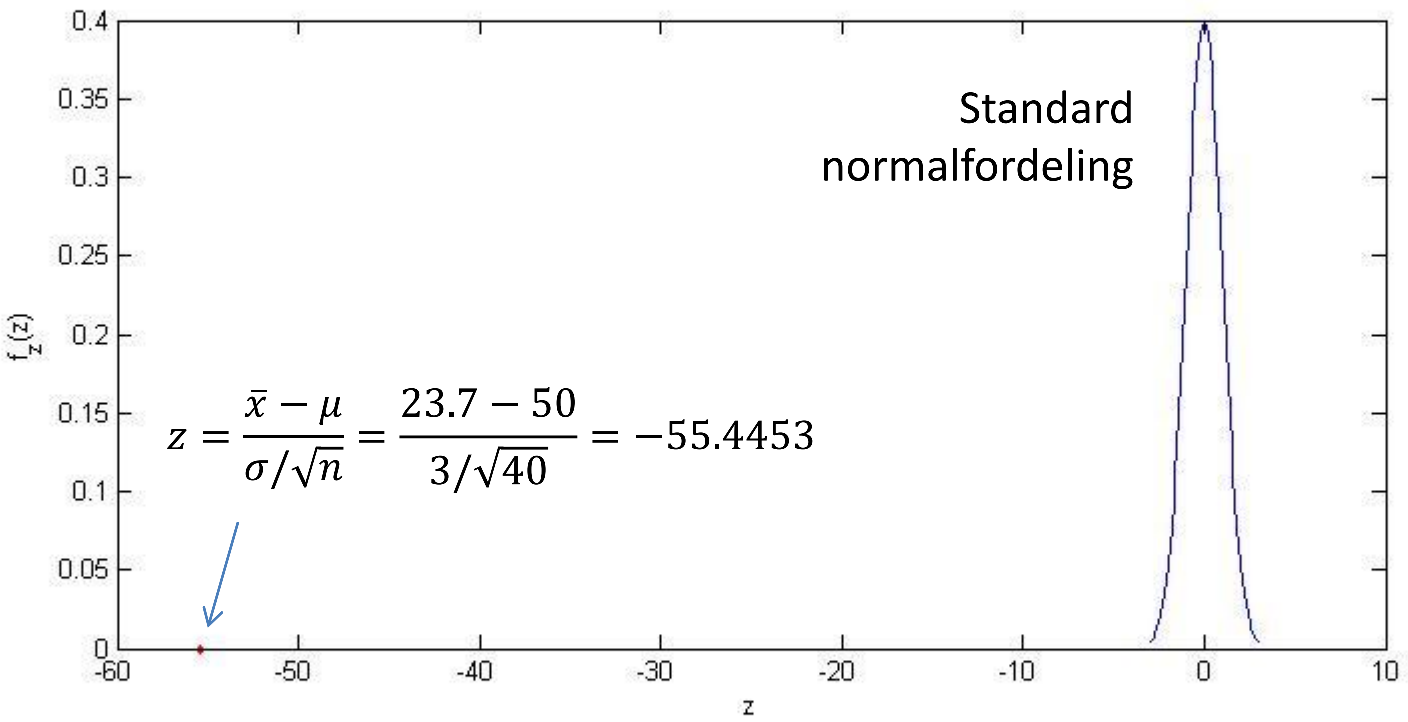
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Statistisk model:

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

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

Teststørrelsen og dens fordeling



Beregning af p-værdi

Sandsynligheden for at observere en teststørrelse, som er mere ekstrem end z = -55.4453

P-værdi:

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

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

Da pval < 0.05, strider data mod hypotesen!

Konklusion:

Data strider imod antagelsen om, at middelværdien af alderen er 50 år.

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

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

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

Standardisering af binomialfordelte data

- Hvis

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

og $n \cdot p > 5$ og $n \cdot (1 - p) > 5$, så er *X* cirka normalfordelt.

- Standardiseret teststørrelse (*z*)
  - Træk middelværdien fra den observerede værdi (*x* = antal successer)
  - Og del med standardafvigelsen

$$z = \frac{x - np}{\sqrt{np(1-p)}} \sim N(0,1)$$

- Så er

$$\Pr(X \leq x) = F_{binomial}(x) \approx \Phi(z)$$

Approximativ p-værdi

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

$$
z = \frac{x - np}{\sqrt{np(1 - p)}}
$$

Da pval > 0.05, strider data ikke mod hypotesen.

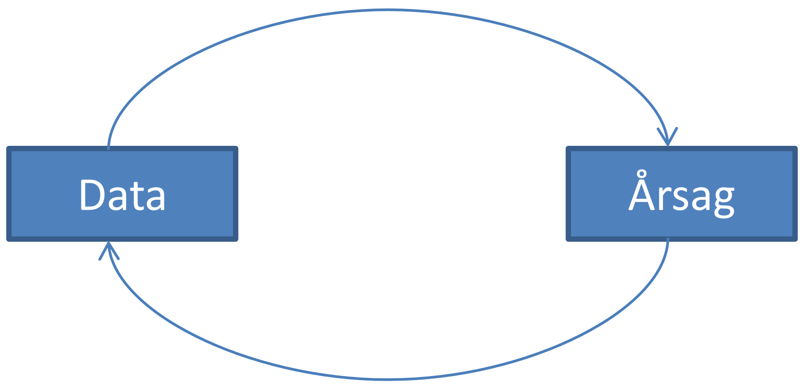
Begreber

Population:
Gule og grønne ærtebælge

Sample på 580 ærtebælge

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

Statistik



Sandsynlighedsteori

Teststørrelse:

x = antal successer = 152

Statistisk model:

$x \sim binomial(580, p)$

Begreber

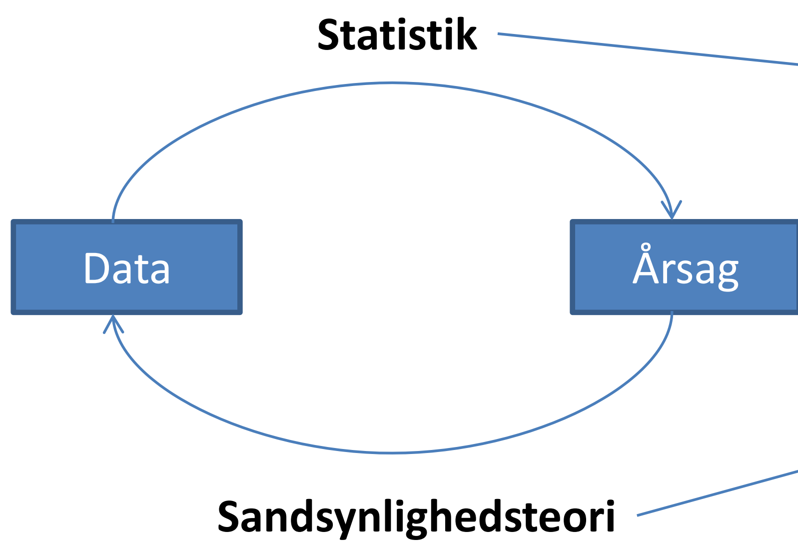
- Teststørrelse
  - Teststørrelsen er en funktion af data, $W = W(x)$.
  - Teststørrelsen er en stokastisk variabel!!!
  - Teststørrelsen bruges til at bestemme graden af overensstemmelse mellem data og hypotese.
- Statistisk model
  - Beskriver teststørrelsens sandsynlighedsfordeling.
  - Fordelingen afhænger af en eller flere *parametre*.

Begreber

Population:
Gule og grønne ærtebælge

Sample på 580 ærtebælge

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



Teststørrelse:

$x = \text{antal successer} = 152$

Statistisk model:

$x \sim binomial(580, p)$

Parameter

Begreber

- Parameter
  - Vi antager, at de samplede data kommer fra en population, hvor den sande værdi af parameteren i den statistiske model er ukendt.
  - Hvis vi gentager eksperimentet under identiske betingelser, vil den sande parameter være uændret, selvom vi får andre data.
- Parameter-skøn eller estimat
  - Et skøn af parameteren beregnes på baggrund af de observerede data.

Notation

- Generelt betegner vi den sande parameter θ.
  - Bemærk, θ kan være en vektor.
- Parameter estimatet betegnes $\hat{\theta} = \hat{\theta}(x)$
  - og er altså en funktion af data (x).
- Estimatet er en stokastisk variabel!!!

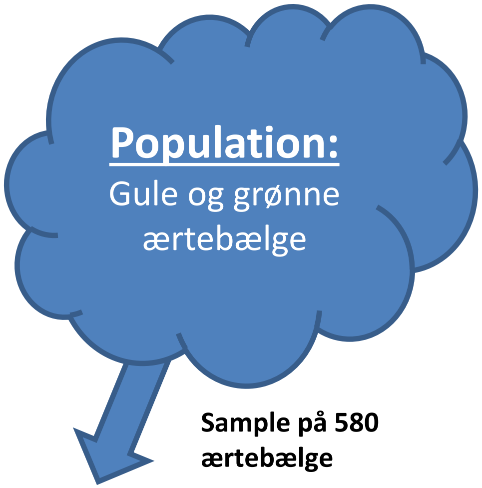
Det gode estimat

- Unbiased
  - Forventningsværdien af estimatet skal være den sande værdi af parameteren.
  - $E[\hat{\theta}] = \theta$

- Usikkerhed på estimatet
  - $Var(\hat{\theta})$ så lille som mulig.

- Optimal
  - $\hat{\theta}$ er ”optimal”, hvis den maksimerer sandsynligheden for data, givet $\hat{\theta}$.
  - Man siger, at $\hat{\theta}$ er maximum likelihood estimatet af θ.

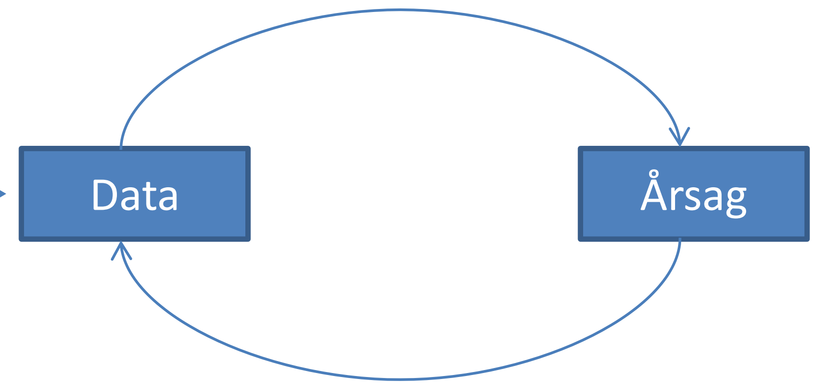
Estimation



Sample på 580 ærtebælge

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

Statistik



Sandsynlighedsteori

Teststørrelse:

$x = \text{antal successer} = 152$

Statistisk model:

$x \sim binomial(580, p)$

Givet data, hvad kan vi sige om parameteren p

- Estimat: $\hat{p} = ?$
- Usikkerhed: Konfidensinterval

Binomialfordelingen

- Data er $b_1, b_2, \dots, b_n$, men typisk observerer vi kun teststørrelsen,

$$x = \sum_{i=1}^{n} b_i \ = antal\ successer \ \sim \ binomial(n, p)$$

- Estimat af parameteren, *p*

$$\hat{p} = \hat{p}(x) = \frac{x}{n}$$

- <u>Unbiased:</u>

$$E[\hat{p}] = E\left[\frac{x}{n}\right] = \frac{1}{n} E[x] = \frac{1}{n}(n \cdot p) = p$$

- <u>Varians:</u>

$$\mathrm{Var}(\hat{p}) = Var\left(\frac{x}{n}\right) = \frac{1}{n^2} Var(x) = \frac{1}{n^2}\big(n \cdot p \cdot (1-p)\big) = \frac{1}{n} p \cdot (1-p)$$

Maximum likelihood

- Er vores estimat ($\hat{p}$) optimalt?
- Tæthedsfunktionen for de observerede data, givet $\hat{p}$, er

$$f(x|\hat{p}) = \binom{n}{x} \hat{p}^x (1-\hat{p})^{n-x}$$

- For at finde det optimale parameterskøn, maksimerer vi ovenstående.
  - Vi kan se bort fra binomialkoefficienten, da den ikke afhænger af $\hat{p}$.
  - Vi må også tage logaritmen til udtrykket, da logaritmen er monoton.
- Løsningen skal maksimere dette udtryk:

$$x \cdot \log(\hat{p}) + (n-x) \cdot \log(1-\hat{p})$$

Maximum likelihood - løsning

- Løsningen er

$$\arg\max_{\hat{p}} \; (x \cdot \log(\hat{p}) + (n - x) \cdot \log(1 - \hat{p}))$$

- Differentier og sæt lig med nul:

$$\frac{x}{\hat{p}} - \frac{n - x}{1 - \hat{p}} = 0$$

- Vi isolerer $\hat{p}$ og får:

$$\hat{p} = \frac{x}{n}$$

Hurra!!!

Opsummering - binomialfordelingen

- Det optimale estimat af parameteren, *p*, er

$$\hat{p} = \hat{p}(x) = \frac{x}{n}$$

- Unbiased:

$$E[\hat{p}] = p$$

- Varians:

$$\mathrm{Var}(\hat{p}) = \frac{1}{n} p \cdot (1 - p)$$

Opsummering - binomialfordelingen

- Matlab
  - Tæthedsfunktion: $\Pr(X = x) =$ `binopdf(x,n,p)`
  - Fordelingsfunktion: $\Pr(X \leq x) =$ `binocdf(x,n,p)`

- Bruges når man har et eksperiment med en sekvens af ja/nej hændelser.
- Ofte kender man ikke sekvensen, men får man blot oplyst en brøk, som angiver succes-raten

$$\hat{p} = \frac{x}{n}$$

Usikkerhed på estimatet

- Da parameterskønnet $\hat{\theta}(x)$ varierer fra gentagelse til gentagelse på grund af tilfældige variationer, er den skønnede værdi i sig selv ikke særlig informativ uden samtidig at angive noget om denne tilfældige variation.
- Man vælger ofte at gøre det, at i stedet for blot at angive et enkelt punkt $\hat{\theta}(x)$ i parameterrummet, så angiver man et helt interval af værdier omkring $\hat{\theta}(x)$.
- Ideen er, at enhver værdi i dette interval er, med de givne data, også et rimeligt gæt på værdien af parameteren.

95% konfidensinterval

- Definition:
  - Sandsynligheden for, at 95% konfidensintervallet indeholder den sande parameterværdi, skal være 0,95:

$$\Pr(\theta \ er \ indeholdt \ i \ intervallet \ [\theta_-; \theta_+]) = 0{,}95$$

- Bogens overordnede strategi
  - Vi ser kun på fordelinger, som har parameter θ, og som kan approksimeres med en normalfordeling.
  - Beregn den standardiserede teststørrelse: $z = \frac{x-\mu}{\sigma} \sim N(0{,}1)$
  - Så gælder der, at $\Pr(-1{,}96 \leq z \leq 1{,}96) = 0.95$
  - Brug dette til at beregne 95% konfidensintervallet, $[\theta_-; \theta_+]$.

95% konfidensinterval for binomialfordelingen

- Hvad skal der gælde om intervalgrænserne?

$$\Pr(p_-(x) \leq p \leq p_+(x)) = 0.95$$

- Hvad ved vi?
  - Antag, vi kan bruge normal approksimationen for binomialfordelte data

$$z = \frac{x - np}{\sqrt{n \cdot p \cdot (1-p)}} \sim N(0,1)$$

  - Så er

<u>Tjek:</u>

```
>> normcdf(1.96)
ans =
    0.9750
```

$$\Pr(-1{,}96 \leq z \leq 1{,}96) =$$
$$\Phi(1{,}96) - \big(1 - \Phi(1{,}96)\big) = 1 - 2\big(1 - \Phi(1{,}96)\big) = 0.95$$

95% konfidensinterval for binomialfordelingen

- Vi sætter ind

$$\Pr(-1{,}96 \leq z \leq 1{,}96) = \Pr\left(-1{,}96 \leq \frac{x - np}{\sqrt{n \cdot p \cdot (1-p)}} \leq 1{,}96\right) = 0.95$$

- Regner man lidt på dette, får man

$$\Pr\left(\frac{1}{n + 1{,}96^2}\left[x + \frac{1{,}96^2}{2} - 1.96\sqrt{\frac{x(n-x)}{n} + \frac{1{,}96^2}{4}}\right] \leq p \leq \frac{1}{n + 1{,}96^2}\left[x + \frac{1{,}96^2}{2} + 1.96\sqrt{\frac{x(n-x)}{n} + \frac{1{,}96^2}{4}}\right]\right)$$

$$= \Pr(p_-(x) \leq p \leq p_+(x)) = 0{,}95$$

95% konfidensinterval for binomialfordelingen

- Mendels eksperiment (binomialfordeling)
  - Teststørrelse $x = antal\ successer\ \sim\ binomial(n,p)$
  - Observation $x = 152$
  - Parameterskøn $\hat{p}(x) = \frac{x}{n} = \frac{152}{580} = 0{,}2621$
- 95% konfidensinterval

$$p_-(x) = \frac{1}{n + 1{,}96^2}\left[x + \frac{1{,}96^2}{2} - 1.96\sqrt{\frac{x(n-x)}{n} + \frac{1{,}96^2}{4}}\right] = 0{,}2312$$

$$p_+(x) = \frac{1}{n + 1{,}96^2}\left[x + \frac{1{,}96^2}{2} + 1.96\sqrt{\frac{x(n-x)}{n} + \frac{1{,}96^2}{4}}\right] = 0{,}3026$$

Sammenhæng mellem konfidensinterval og p-værdi

- Betragt situationen fra tidligere med en statistisk model indeholdende en parameter θ.
- Lad $pval(x; \theta_0)$ være *p*-værdien for et test af hypotesen

$$H\colon\ \theta = \theta_0$$

  baseret på observationen *x*.
- Hvis $pval(x; \theta_0) > \alpha$, strider data som bekendt ikke mod hypotesen.
- Der vil typisk være mange valg af θ_0, som opfylder denne betingelse.
- Vi definerer derfor mængden af alle sådanne parametre:

$$\{\theta | pval(x; \theta\ ) > \alpha\}$$

Sammenhæng mellem konfidensinterval og p-værdi

- Vi får altså et interval af parameterværdier, som stemmer overens med data (*x*):

$$\{\theta | pval(x; \theta\ ) > \alpha\}$$

- Dette er $(1 - \alpha) \cdot 100\%$ konfidensintervallet.
- Fx, hvis $\alpha = 0{,}05$, får vi 95% konfidensintervallet.

$(1-\alpha)\cdot 100\%$ konfidensintervallet for binomialfordelingen

- Nedre grænse

$$p_{-}(x) = \frac{1}{n+u^2}\left[x + \frac{u^2}{2} - u\sqrt{\frac{x(n-x)}{n} + \frac{u^2}{4}}\right]$$

- Øvre grænse

$$p_{+}(x) = \frac{1}{n+u^2}\left[x + \frac{u^2}{2} + u\sqrt{\frac{x(n-x)}{n} + \frac{u^2}{4}}\right]$$

- Hvor

$$u = \Phi^{-1}(1 - \frac{\alpha}{2})$$

- Matlab

$u = \Phi^{-1}\left(1 - \frac{0.05}{2}\right) =$ `norminv(1-0.05/2) = 1.96`

Testkatalog for binomialfordelingen

- Statistisk model
  - $X \sim binomial(n, p)$
  - Parameterskøn: $\hat{p} = \frac{x}{n}$
  - Hvor observationen er $x = antal\ succeser$

- Hypotesetest
  - $H: p = p_0$
  - Teststørrelse: $\mathrm{z} = \frac{x - np_0}{\sqrt{n \cdot p_0 \cdot (1 - p_0)}}$
  - Approksimativ p-værdi: $pval = 2 \cdot |1 - \Phi(|z|)|$

- Approksimativt 95% konfidensinterval
  - $[p_-(x); p_+(x)] = \left[\frac{1}{n+u^2}\left[x + \frac{u^2}{2} - u\sqrt{\frac{x(n-x)}{n} + \frac{u^2}{4}}\right]; \frac{1}{n+u^2}\left[x + \frac{u^2}{2} + u\sqrt{\frac{x(n-x)}{n} + \frac{u^2}{4}}\right]\right]$
  - Hvor $u = 1{,}96$

- Forudsætninger for approksimationen: $n \cdot p_0 > 5$ og $n \cdot (1 - p_0) > 5$.

Mendels eksperiment i Matlab

```
%% Eksempel 1 - Mendels eksperiment
x = 152;
n = 580;
p0 = 1/4;
u = 1.96;

% Hypotesetest (approksimativ p-værdi)
z = (x-n*p0)/sqrt(n*p0*(1-p0))
pval = 2*(1-normcdf(abs(z)))

% Parameterskøn
p_est = x/n

% 95% konfidensinterval
p_nedre = 1/(n+u^2) * (x + u^2/2 - u*sqrt( x*(n-x)/n + u^2/4 ) )
p_oevre = 1/(n+u^2) * (x + u^2/2 + u*sqrt( x*(n-x)/n + u^2/4 ) )
```

Eksempel

- Drenge- og pigefødsler
  - I 2005 blev der i Holme-Højbjerg-Skåde området i Aarhus født 231 personer, hvoraf 108 var piger og 123 var drenge.
  - Vi ønsker at undersøge, om pige- og drengefødsler er lige hyppige.
- Statistisk model
  - $x = antal\, pigefødsler = 108$
  - $X \sim binomial(n, p)$, hvor $n = 231$

Eksempel

- Hypotese
  - $H: p = p_0 = \frac{1}{2}$
- Teststørrelse
  - $z = \frac{x - np_0}{\sqrt{n \cdot p_0 \cdot (1 - p_0)}} = \frac{108 - 231 \cdot \frac{1}{2}}{\sqrt{231 \cdot \frac{1}{2} \cdot (1 - \frac{1}{2})}} = -0{,}9869$
- Approksimativ p-værdi
  - $pval = 2 \cdot |1 - \Phi(|z|)| = 2 \cdot (1 - 0.8382) = 0{,}3237$
- pval > 0.05: Vi kan ikke afvise hypotesen om, at pige- og drengefødsler er lige hyppige.

Eksempel

- Parameterskøn
  - $\hat{p}(x) = \frac{x}{n} = \frac{108}{231} = 0{,}4675$

- 95% konfidensinterval

$$[p_-(x); p_+(x)] = \left[\frac{1}{n+u^2}\left[x + \frac{u^2}{2} - u\sqrt{\frac{x(n-x)}{n} + \frac{u^2}{4}}\right]; \frac{1}{n+u^2}\left[x + \frac{u^2}{2} + u\sqrt{\frac{x(n-x)}{n} + \frac{u^2}{4}}\right]\right]$$

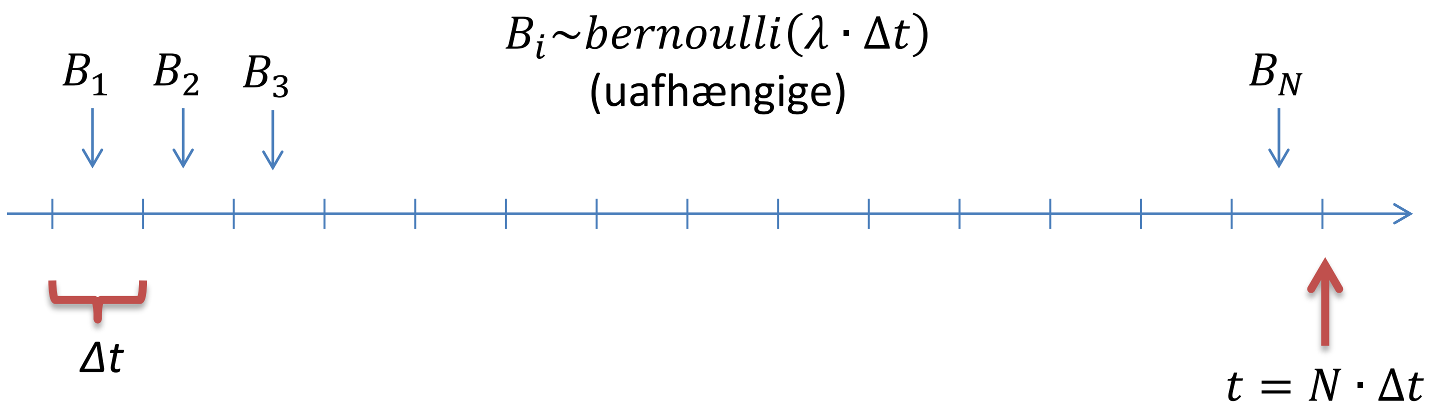
  - Hvor $u = 1{,}96$
  - $p_-(x) = 0{,}4124$
  - $p_+(x) = 0{,}5401$

Eksempel i Matlab

```
%% Eksempel 2 - Drenge- og pigefødsler
x = 108;
n = 231;
p0 = 1/2;
u = 1.96;

% Hypotesetest (approksimativ p-værdi)
z = (x-n*p0)/sqrt(n*p0*(1-p0))
pval = 2*(1-normcdf(abs(z)))

% Parameterskøn
p_est = x/n

% 95% konfidensinterval
p_nedre = 1/(n+u^2) * (x + u^2/2 - u*sqrt( x*(n-x)/n + u^2/4 ) )
p_oevre = 1/(n+u^2) * (x + u^2/2 + u*sqrt( x*(n-x)/n + u^2/4 ) )
```

Poissonfordelingen

- Bruges til en beskrive en proces med forskellige ankomsttider.
  - Atomare henfald
  - Trafiksimulering (fx tilfældig ankomst af biler ved et lyskryds)
- Model
  - Opdel tidsaksen i *N* intervaller af længde *Δt*.
  - I hvert interval er der $B_i = 1$ eller $B_i = 0$ ankomster, hvor

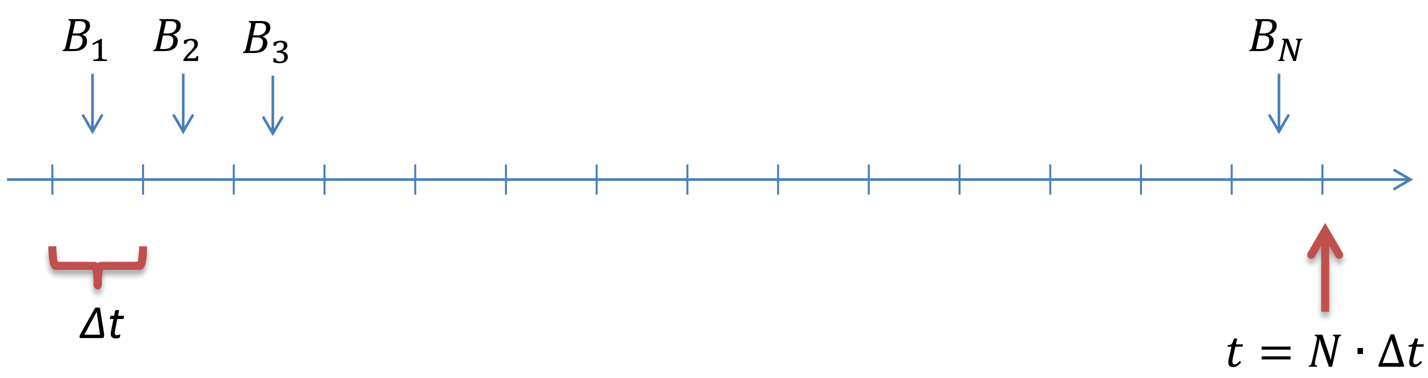


Poissonfordelingen

- Hvad er sandsynligheden for at observere $X = x$ ankomster i tidsintervallet [1; N]?

$$X = \sum_{i=1}^{n} B_i \sim binomial(N, \lambda \cdot \Delta t)$$

$$\Pr(X = x) = \binom{N}{x} (\lambda \cdot \Delta t)^x (1 - \lambda \cdot \Delta t)^{N-x}$$



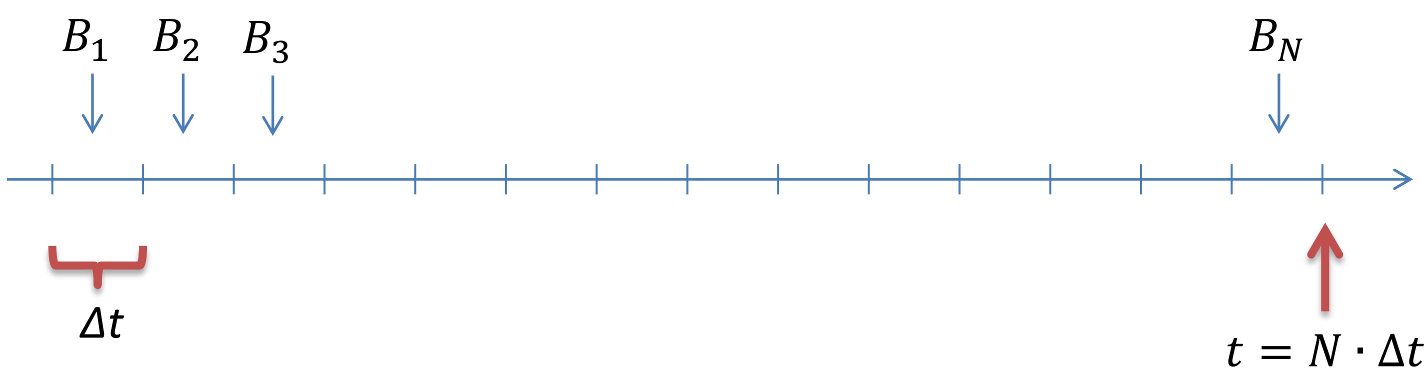
Poissonfordelingen

- Observation

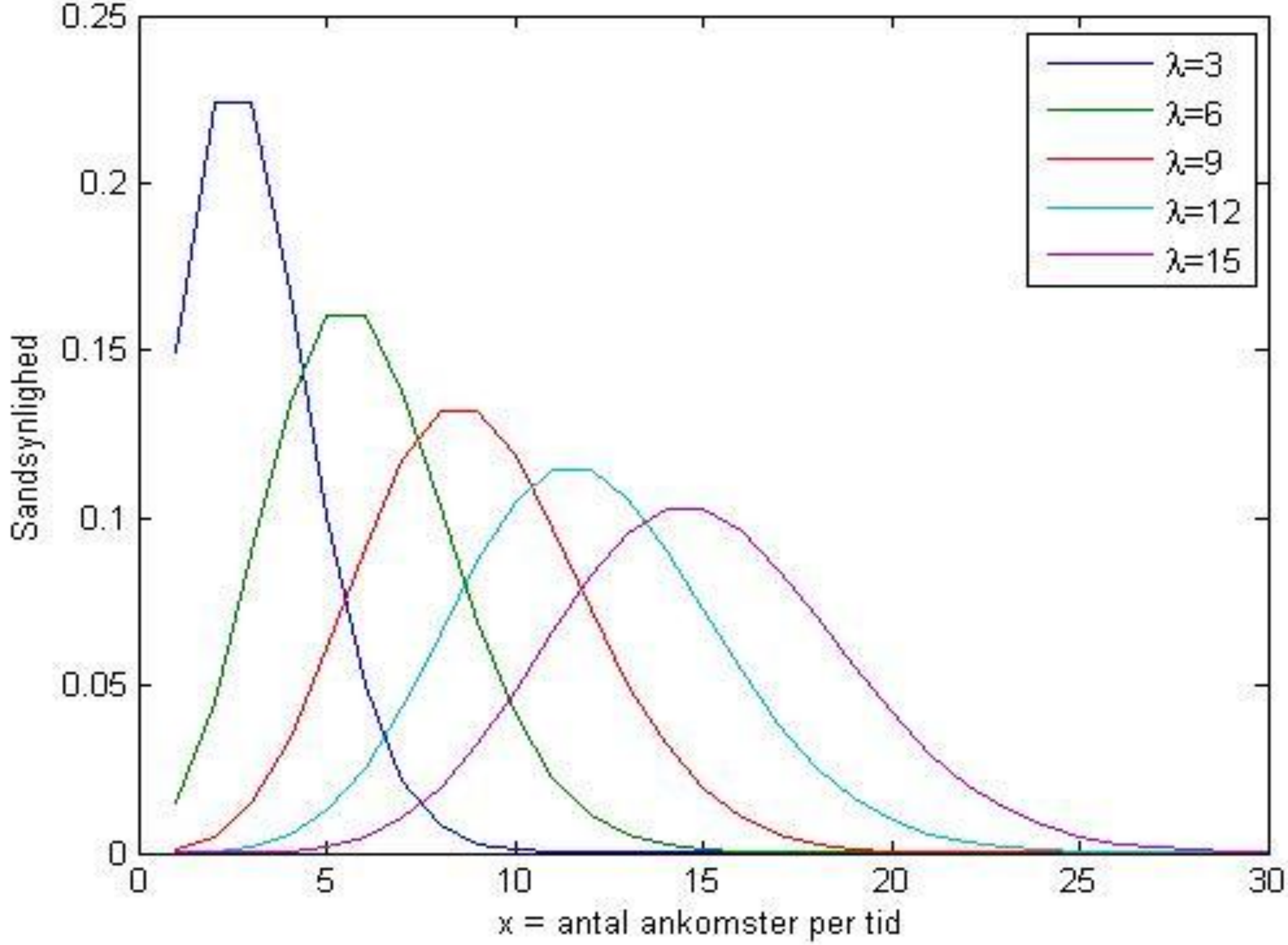
$$N \cdot (\lambda \cdot \Delta t) = konstant = \frac{t}{\Delta t} \cdot (\lambda \cdot \Delta t) = t \cdot \lambda = \gamma$$

- I grænsen $\Delta t \to 0$, kan man vise, at

$$\Pr(X = x) = \frac{(\lambda \cdot t)^x}{x!} e^{-\lambda \cdot t} = \frac{\gamma^x}{x!} e^{-\gamma}$$



Effekt af γ



Poissonfordelingen

- Notation

$$X \sim poisson(t \cdot \lambda)$$
$$X \sim poisson(\gamma)$$

- Middelværdi

$$E[X] = \gamma = t \cdot \lambda$$

- Varians

$$Var(X) = \gamma = t \cdot \lambda$$

- Der gælder også, at hvis $X_1 \sim poisson(\gamma_1)$ og $X_2 \sim poisson(\gamma_2)$ og uafhængige, så er $X_1 + X_2 \sim poisson(\gamma_1 + \gamma_2)$

Uddybning af poissonfordelingen

- Tæthedsfunktion

$$f(x) = \Pr(X = x) = \frac{(\lambda \cdot t)^x}{x!} e^{-\lambda \cdot t} = \frac{\gamma^x}{x!} e^{-\gamma}$$

- Data

$$x = antal\ ankomster$$

- Parameterskøn

$$\hat{\lambda} = \frac{x}{t}$$

- Unbiased

$$E[\hat{\lambda}] = \lambda$$

Uddybning af poissonfordelingen

- Tæthedsfunktion

$$f(x) = \Pr(X = x) = \frac{(\lambda \cdot t)^x}{x!} e^{-\lambda \cdot t} = \frac{\gamma^x}{x!} e^{-\gamma}$$

- Matlab
  - Tæthedsfunktion: $\Pr(X = x) =$ `poisspdf(x,gamma)`
  - Fordelingsfunktion: $\Pr(X \leq x) =$ `poisscdf(x,gamma)`

Eksempel

- En butik har 300 besøgende på 2 timer.
- Hvad er sandsynligheden for, at der kommer mere end 170 besøgende den næste time?
- Data: $x = 300$
- Parameterskøn: $\hat{\lambda} = \frac{300}{2} = 150$ besøgende/time
- Beregn $\Pr(X > 170)$ for t = 1 time

$$\Pr(X > 170) = 1 - \Pr(X \leq 170) = 1 - \sum_{k=0}^{170} \frac{\left(\hat{\lambda} \cdot t\right)^k}{k!} e^{-\hat{\lambda} \cdot t}$$

- Matlab

`1-poisscdf(170,150)` $\approx$ 5%

Bemærk: Her er t=1, hvilket medfører, at $\lambda = \gamma$.

Standardisering af poissonfordelte data

- Hvis

$$X \sim poisson(\gamma = \lambda \cdot t)$$

og $\gamma = \lambda \cdot t > 5$, så er *X* cirka normalfordelt.

- Standardiseret teststørrelse (*z*)
  - Træk middelværdien fra den observerede værdi (*x* = antal ankomster)
  - Og del med standardafvigelsen

$$\mathrm{z} = \frac{x - t\lambda}{\sqrt{t \cdot \lambda}} = \frac{x - \gamma}{\sqrt{\gamma}} \sim N(0,1)$$

- Så er

$$\Pr(X \leq x) = F_{poisson}(x) \approx \Phi(z)$$

Testkatalog for poissonfordelingen

- Statistisk model
  - $X \sim poisson(\lambda \cdot t)$
  - Parameterskøn: $\hat{\lambda} = \frac{x}{t}$
  - Hvor observationen er $x = antal\ ankomster\ i\ tidsintervallet\ t$

- Hypotesetest
  - $H: \lambda = \lambda_0$
  - Teststørrelse: $\mathrm{z} = \frac{x - t\lambda_0}{\sqrt{t \cdot \lambda_0}}$
  - Approksimativ p-værdi: $pval = 2 \cdot |1 - \Phi(|z|)|$

- Approksimativt 95% konfidensinterval
  - $[\lambda_-(x); \lambda_+(x)] = \left[\frac{1}{t}\left[x + \frac{u^2}{2} - u\sqrt{x + \frac{u^2}{4}}\right]; \frac{1}{t}\left[x + \frac{u^2}{2} + u\sqrt{x + \frac{u^2}{4}}\right]\right]$
  - Hvor $u = 1{,}96$

- Forudsætninger for approksimationen: $\mathrm{t} \cdot \lambda_0 > 5$.

Eksempel

- Rutherford & Geiger
  - 2608 tællinger af radioaktive henfald
  - I tidsintervaller af 72 sekunders varighed

Antal henfald	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Antal tids-intervaller	57	203	383	525	532	408	273	139	45	27	10	4	0	1	1

- Statistisk model

$$X_i \sim poisson(72 \cdot \lambda), for\ i = 0,1,\ldots,2608$$

$$X = \sum_{i=1}^{2608} X_i \sim poisson(2608 \cdot 72 \cdot \lambda)$$

Eksempel

- Parameterskøn
  - $\hat{\lambda}(x) = \frac{x}{t} = \frac{\text{det } totale\ antal\ henfald}{samlet\ tid} = \frac{11571}{187776} = 0.0616$

- 95% konfidensinterval

$$[\lambda_-(x); \lambda_+(x)] = \left[\frac{1}{t}\left[x + \frac{u^2}{2} - u\sqrt{x + \frac{u^2}{4}}\right]; \frac{1}{t}\left[x + \frac{u^2}{2} + u\sqrt{x + \frac{u^2}{4}}\right]\right]$$

  - Hvor $u = 1{,}96$
  - $\lambda_-(x) = 0.0605$
  - $\lambda_+(x) = 0.0628$