

Normalfordelte data

Læsning:

Jens Ledet Jensen kap. 4+5.1+6

Hypotesetest

Population:

Gule og grønne
ærtebælge

Sample på 580
ærtebælge

Bælg

1

2

3

.

.

.

580

Alder

grøn

gul

grøn

.

.

.

grøn

H: $p = 1/4$

Statistik

Teststørrelse:

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

Data

Årsag

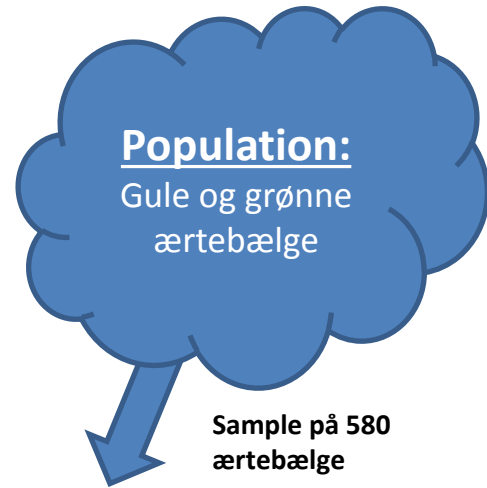
Statistisk model:

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

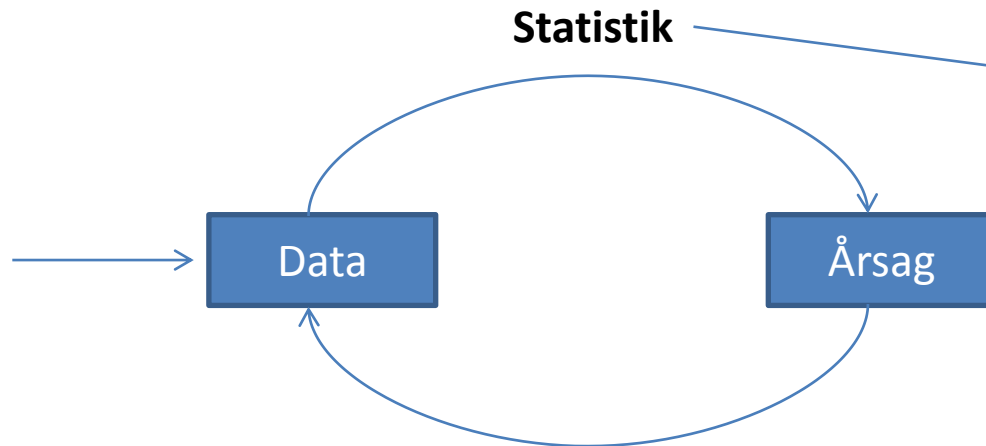
Sandsynlighedsteori

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

Estimation



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



Teststørrelse:

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

Statistisk model:

$x \sim \text{binomial}(580, p)$

95% konfidensinterval:
Hvilke parameterverdier stemmer overens med data?

Testkatalog for binomialfordelingen

- Statistisk model
 - $X \sim \text{binomial}(n, p)$
 - Parameterskøn: $\hat{p} = \frac{x}{n}$
 - Hvor observationen er $x = \text{antal successer}$
- Hypotesetest
 - $H: p = p_0$
 - Teststørrelse: $z = \frac{x - np_0}{\sqrt{n \cdot p_0 \cdot (1 - p_0)}}$
 - 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 $n \cdot (1 - p_0) > 5$.

Testkatalog for poissonfordelingen

- Statistisk model
 - $X \sim \text{poisson}(\lambda \cdot t)$
 - Parameterskøn: $\hat{\lambda} = \frac{x}{t}$
 - Hvor observationen er $x = \text{antal ankomster i tidsintervallet } t$
- Hypotesetest
 - $H: \lambda = \lambda_0$
 - Teststørrelse: $z = \frac{x - t\lambda_0}{\sqrt{t \cdot \lambda_0}}$
 - 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: $t \cdot \lambda_0 > 5$.

Normalfordelingen

- Generel normalfordeling (Gauss)
 - $X \sim N(\mu, \sigma^2)$
 - $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(x-\mu)^2/2\sigma^2}$
 - $E[X] = \bar{X} = \mu$
 - $\text{Var}(X) = \sigma^2$
- Standard normalfordeling
 - $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$
 - $E[Z] = 0$
 - $\text{Var}(Z) = 1$

Udregning af sandsynligheder

- Fordelingsfunktion

$$\Pr(X \leq x) = F_X(x) = \Pr\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

- Interval

$$\begin{aligned}\Pr(a \leq X \leq b) &= F_X(b) - F_X(a) \\ &= \Pr\left(\frac{a - \mu}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{b - \mu}{\sigma}\right) \\ &= \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)\end{aligned}$$

Det gode estimat

- Unbiased
 - Forventningsværdien af estimatet skal være den sande værdi af parameteren.
 - $E[\hat{\theta}] = \theta$
- Usikkerhed på estimatet
 - $Var(\hat{\theta})$ skal blive mindre, når antal datapunkter $n \rightarrow \infty$.
- Optimal
 - $\hat{\theta}$ er "optimal", hvis den maksimerer likelihood funktionen, $f(x|\hat{\theta})$.
 - Man siger, at $\hat{\theta}$ er maximum likelihood estimatet af θ .

Estimat af μ

- Givet målingerne x_1, x_2, \dots, x_n , hvor

$$x_i \sim N(\mu, \sigma^2)$$

- så virker det naturligt at vælge gennemsnittet

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

som estimat af middelværdien.

- Husk notationen!

$$\hat{\mu} = \bar{x}$$

Estimat af μ

- Unbiased

$$E[\hat{\mu}] = E\left[\frac{1}{n} \sum_{i=1}^n x_i\right] = \frac{1}{n} \cdot E\left[\sum_{i=1}^n x_i\right] = \frac{1}{n} \cdot \sum_{i=1}^n E[x_i] = \frac{n \cdot \mu}{n} = \mu$$

- Varians af estimatet

$$\begin{aligned} Var[\hat{\mu}] &= Var\left[\frac{1}{n} \sum_{i=1}^n x_i\right] = \frac{1}{n^2} \cdot Var\left[\sum_{i=1}^n x_i\right] = \frac{1}{n^2} \cdot \sum_{i=1}^n Var[x_i] = \frac{1}{n^2} \cdot n \cdot \sigma^2 \\ &= \sigma^2/n \end{aligned}$$

Estimat af μ

- Maximum likelihood
- Er vores estimat $\hat{\mu} = \bar{x}$ optimalt?
- Tæthedsfunktionen for de observerede data, givet $\hat{\mu}$, er

$$\begin{aligned} f(x|\hat{\mu}, \sigma^2) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(x_i - \hat{\mu})^2 / 2\sigma^2} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \hat{\mu})^2} \end{aligned}$$

- For at finde det optimale parameterskøn, maksimerer vi ovenstående mht. $\hat{\mu}$
 - Vi kan se bort fra konstante koefficienter, som ikke afhænger af $\hat{\mu}$.
 - Vi må også tage logaritmen til udtrykket, da logaritmen er monoton.

Estimat af μ

- Den optimale løsning er

$$\arg \max_{\hat{\mu}} \left(- \sum_{i=1}^n (x_i - \hat{\mu})^2 \right) = \arg \min_{\hat{\mu}} \sum_{i=1}^n (x_i - \hat{\mu})^2$$

- Hvis vi differentierer og sætter lig med nul, får vi

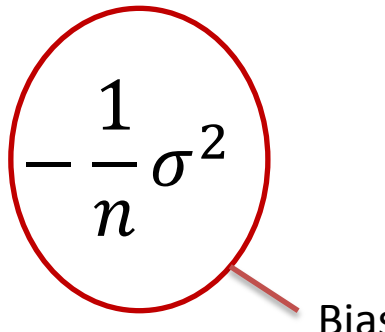
$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

Estimat af σ^2

- Maximum-likelihood estimatet af σ^2 er

$$\hat{\sigma}_{ML}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

- Men regner man efter, indser man, at dette estimat er biased:

$$E[\hat{\sigma}_{ML}^2] = \sigma^2 \left(-\frac{1}{n} \sigma^2 \right)$$


Bias

Estimat af σ^2

- Varians-estimatet kaldes den *empiriske varians*.
- Den empiriske varians s^2 er givet ved

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- Dette estimat er unbiased:

$$E[s^2] = \sigma^2$$

Spørgsmål

- Hvilken fordeling har $\hat{\mu}$ og s^2 ?

Sum af to normalfordelte variable

- Hvis $X_1 \sim N(\mu_1, \sigma_1^2)$ og $X_2 \sim N(\mu_2, \sigma_2^2)$ og uafhængige, så er

$$Z = X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Den centrale grænseværdisætning (bogens udgave – Resultat 4.5)

- Hvis X_1, X_2, \dots, X_n er uafhængige med
 - samme fordeling
 - middelværdi $E[X_i] = \mu$
 - og varians $\text{Var}(X_i) = \sigma^2$
- så i grænsen $n \rightarrow \infty$ har vi, at

$$\Pr\left(\frac{\sum_{i=1}^n X_i - n \cdot \mu}{\sqrt{n \cdot \sigma^2}} \leq z\right) = \Phi(z) \sim N(0,1)$$

Tjek selv...

- Middelværdi

$$\begin{aligned} E \left[\frac{\sum_{i=1}^n X_i - n \cdot \mu}{\sqrt{n \cdot \sigma^2}} \right] &= \frac{1}{\sqrt{n \cdot \sigma^2}} \cdot E \left[\sum_{i=1}^n X_i \right] - \frac{n \cdot \mu}{\sqrt{n \cdot \sigma^2}} \\ &= \frac{1}{\sqrt{n \cdot \sigma^2}} \cdot \sum_{i=1}^n E[X_i] - \frac{n \cdot \mu}{\sqrt{n \cdot \sigma^2}} = \frac{n \cdot \mu}{\sqrt{n \cdot \sigma^2}} - \frac{n \cdot \mu}{\sqrt{n \cdot \sigma^2}} = 0 \end{aligned}$$

- Varians

$$\begin{aligned} Var \left[\frac{\sum_{i=1}^n X_i - n \cdot \mu}{\sqrt{n \cdot \sigma^2}} \right] &= \left(\frac{1}{\sqrt{n \cdot \sigma^2}} \right)^2 \cdot Var \left[\sum_{i=1}^n X_i \right] \\ &= \frac{1}{n \cdot \sigma^2} \cdot \sum_{i=1}^n Var[X_i] = \frac{n \cdot \sigma^2}{n \cdot \sigma^2} = 1 \end{aligned}$$

Den centrale grænseværdisætning for gennemsnittet \bar{X}

- Hvis X_1, X_2, \dots, X_n er uafhængige med
 - samme fordeling
 - middelværdi $E[X_i] = \mu$
 - og varians $\text{Var}(X_i) = \sigma^2$
- så i grænsen $n \rightarrow \infty$ har vi, at

$$\Pr\left(\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \leq z\right) = \Pr\left(\frac{\frac{1}{n} \sum_{i=1}^n X_i - \mu}{\sqrt{\sigma^2/n}} \leq z\right) = \Phi(z) \sim N(0,1)$$

Tjek selv...

- Middelværdi

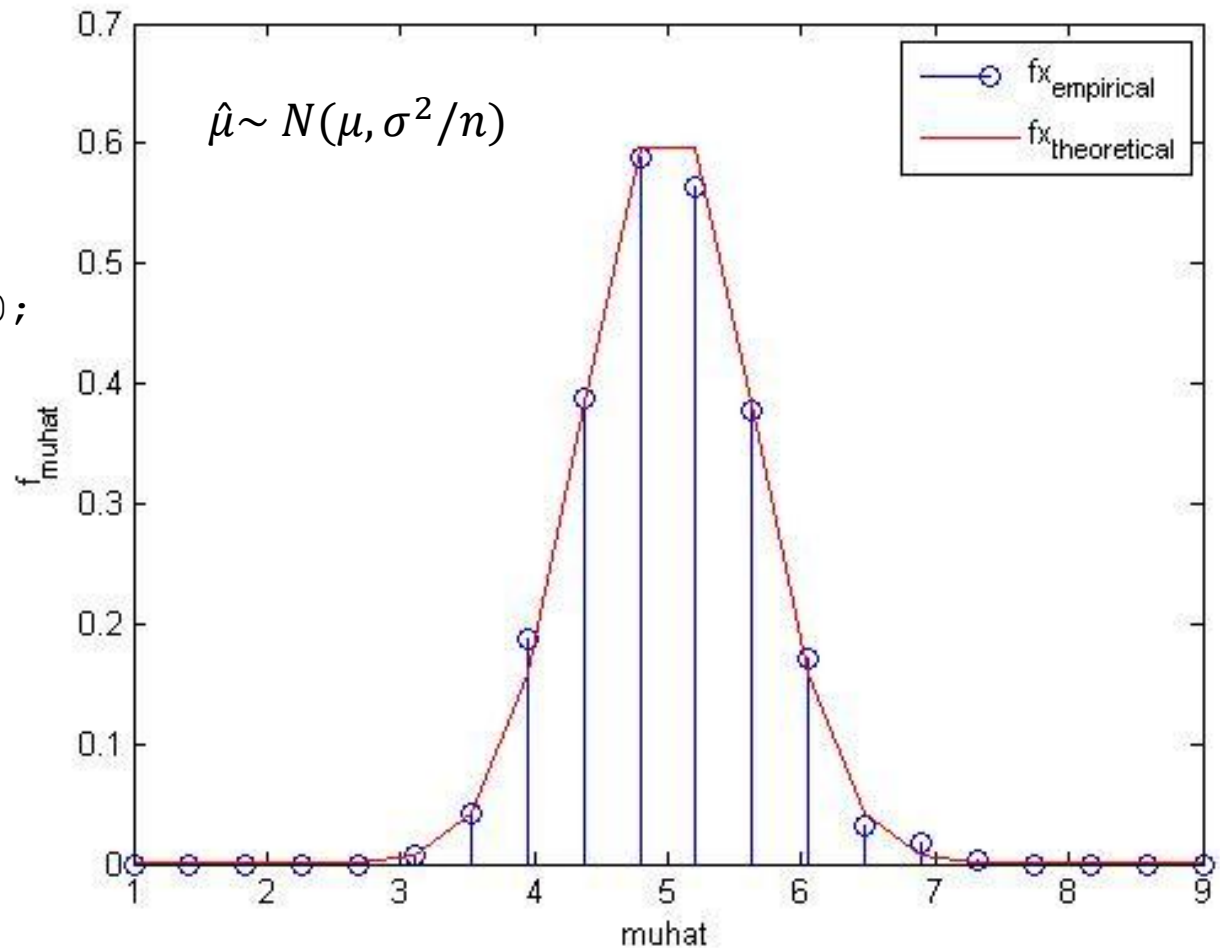
$$\begin{aligned} E \left[\frac{\frac{1}{n} \sum_{i=1}^n X_i - \mu}{\sqrt{\sigma^2/n}} \right] &= \frac{1}{\sqrt{\sigma^2/n}} \cdot E \left[\frac{1}{n} \sum_{i=1}^n X_i \right] - \frac{\mu}{\sqrt{\sigma^2/n}} \\ &= \frac{1}{\sqrt{\sigma^2/n}} \cdot \frac{1}{n} \sum_{i=1}^n E[X_i] - \frac{\mu}{\sqrt{\sigma^2/n}} = \frac{\mu}{\sqrt{\sigma^2/n}} - \frac{\mu}{\sqrt{\sigma^2/n}} = 0 \end{aligned}$$

- Varians

$$\begin{aligned} Var \left[\frac{\frac{1}{n} \sum_{i=1}^n X_i - \mu}{\sqrt{\sigma^2/n}} \right] &= \left(\frac{1}{\sqrt{\sigma^2/n}} \right)^2 \cdot Var \left[\frac{1}{n} \sum_{i=1}^n X_i \right] = \frac{1}{\sigma^2/n} \cdot \frac{1}{n^2} \sum_{i=1}^n Var[X_i] \\ &= \frac{n}{\sigma^2} \cdot \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{n}{\sigma^2} \cdot \frac{\sigma^2}{n} = 1 \end{aligned}$$

Eksempel – fordelingen af middelværdi- estimatet i normalfordelte data

```
% Population  
mu      = 5;  
sigma   = 2;  
n       = 10;  
  
num_experiments = 1000;
```



95% konfidensinterval for estimatet af middelværdien

- Fra sidst ved vi, at

$$\Pr(-1,96 \leq z \leq 1,96) = 0,95 \quad \text{hvis} \quad z \sim N(0,1)$$

- Dermed får vi

$$\Pr\left(-1,96 \leq \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}} \leq 1,96\right) = 0,95$$

$$\Leftrightarrow$$

$$\Pr\left(\bar{x} - 1,96 \cdot \sqrt{\sigma^2/n} \leq \mu \leq \bar{x} + 1,96 \cdot \sqrt{\sigma^2/n}\right) = 0,95$$

**Forudsætning: Vi skal kende den sande varians, σ^2 .
Ellers brug en t-fordeling (senere...).**

Testkatalog for middelværdien i normalfordelte data (kendt varians)

- Statistisk model
 - $X \sim N(\mu, \sigma^2)$
 - Parameterskøn: $\hat{\mu} = \bar{x} \sim N(\mu, \sigma^2/n)$
 - Hvor observationen \bar{x} er gennemsnittet
- Hypotesetest
 - $H: \mu = \mu_0$
 - Teststørrelse: $z = \frac{\bar{x} - \mu_0}{\sqrt{\sigma^2/n}} \sim N(0,1)$
 - P-værdi: $pval = 2 \cdot |1 - \Phi(|z|)|$
- 95% konfidensinterval
 - $[\mu_-(x); \mu_+(x)] = [\bar{x} - u \cdot \sqrt{\sigma^2/n}; \bar{x} + u \cdot \sqrt{\sigma^2/n}]$
 - Hvor $u = 1,96$
- **Forudsætning: Vi skal kende den sande varians, σ^2 . Ellers brug en t-fordeling (senere...).**

Cup-filling eksempel fra Wikipedia

- Antal samples:
- Standard afvigelse:
- Måling:
- Hypotese:
- Teststørrelse:
- p-værdi:
- 95% konfidensinterval:

$$n = 25$$

$$\sigma = 2,5 \text{ g} \longrightarrow \text{Kendt varians!}$$

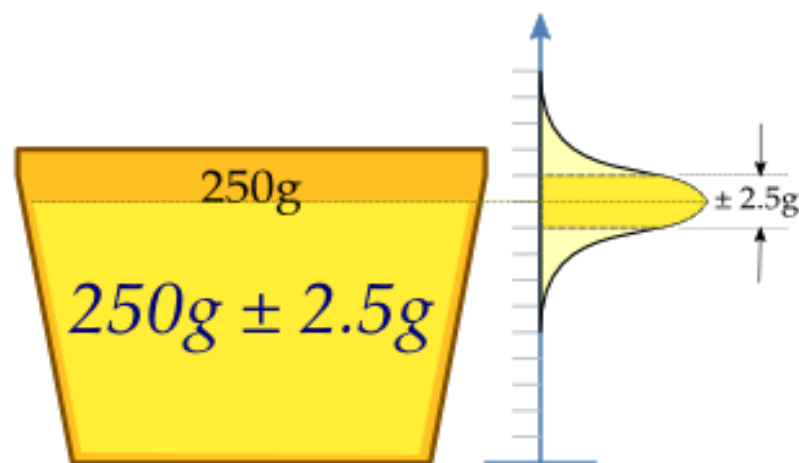
$$\bar{x} = 250,2 \text{ g}$$

$$H: \mu = \mu_0 = 250 \text{ g}$$

$$z = \frac{\bar{x} - \mu_0}{\sqrt{\sigma^2/n}} = \frac{250,2 - 250}{\sqrt{2,5^2/25}} = \frac{0,2}{2,5/5} = \frac{0,2}{0,5} = 0,4$$

$$2 \cdot (1 - \Phi(0,4)) = 2 \cdot (1 - 0,6554) > 0.05$$

$$\begin{aligned} \bar{x} \pm 1,96 \cdot \sqrt{\sigma^2/n} &= \\ 250,2 \pm 1,96 \cdot 0,5 &= \\ 250,2 \pm 0,98 &= \\ [249,22; 251,18] \end{aligned}$$



Sample size determination

- 95% konfidensinterval for en middelværdi

$$\bar{x} \pm 1,96 \cdot \sqrt{\sigma^2/n}$$

- Generel form

$$\bar{x} \pm B$$

- Hvis du som ingeniør/forsker ønsker at lave et estimat af en middelværdi (ud fra gennemsnittet), kan du selv vælge B .
- Så skal du bruge mindst

$$n \geq \left(\frac{1,96 \cdot \sigma}{B} \right)^2$$

samples.

Eksempel – sample size

Example 9.6

A large manufacturing firm is interested in estimating the average distance traveled to work by its employees. Past studies of this type indicate that the standard deviation of these distances should be in the neighborhood of 2 miles. How many employees should be sampled if the estimate is to be within 0.1 mile of the true average, with 95% confidence?

Solution The resulting interval is to be of the form $\bar{X} \pm 0.1$ with $1 - \alpha = 0.95$. Thus, $B = 0.1$ and $z_{0.025} = 1.96$. It follows that

$$n \geq \left[\frac{z_{\alpha/2}\sigma}{B} \right]^2 = \left[\frac{1.96(2)}{0.1} \right]^2 = 1,536.64, \text{ i.e., } n \geq 1,537$$

Thus, at least 1,537 employees should be sampled to achieve the desired results.

Spørgsmål

- Hvilken fordeling har $\hat{\mu}$ og s^2 ?
- Ifølge den centrale grænseværdisætning:

$$\hat{\mu} = \bar{x} \sim N(\mu, \sigma^2/n)$$

χ^2 fordelingen

- Udtales ”Ki-i-anden” fordelingen
- Lad U_1, U_2, \dots, U_f være uafhængige stokastiske variable, som alle er standard normalfordelte

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

- Så er

$$V = \sum_{i=1}^f U_i^2 \sim \chi^2(f)$$

med f frihedsgrader.

χ^2 fordelingen

- Fordelingsfunktion
- Vi har

$$V = \sum_{i=1}^f U_i^2 \sim \chi^2(f)$$

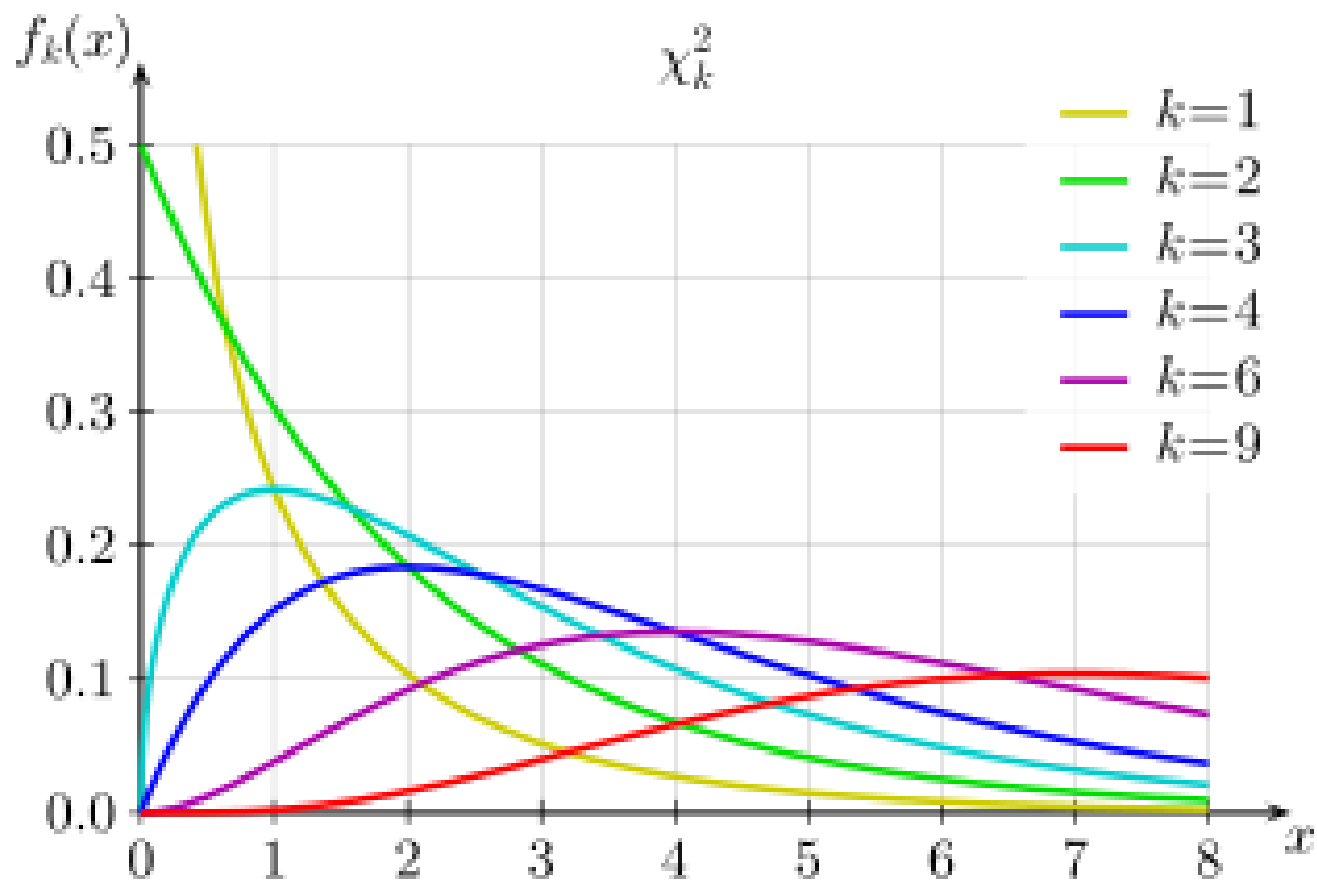
- Notation

$$\Pr(V \leq v) = \chi_{cdf}^2(v, f)$$

- Matlab

- Fordelingsfunktion $\Pr(V \leq v) = \text{chi2cdf}(v, f)$
- Invers $v = \text{chi2inv}(\alpha, f)$

χ^2 fordelingen



Spørgsmål

- Hvilken fordeling har $\hat{\mu}$ og s^2 ?
- Ifølge den centrale grænseværdisætning:

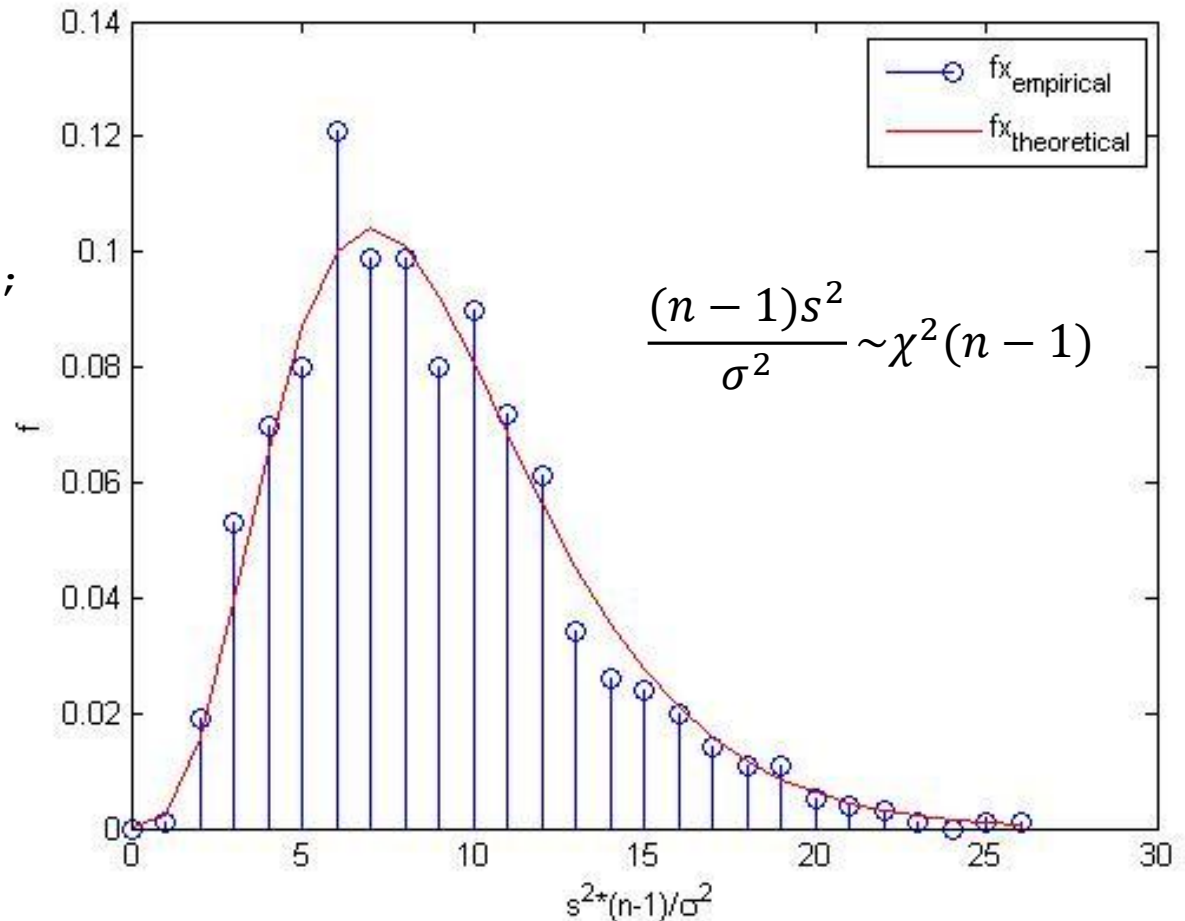
$$\hat{\mu} = \bar{x} \sim N(\mu, \sigma^2/n)$$

- Og man kan vise, at

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \sim \frac{\sigma^2}{n-1} \chi^2(n-1)$$

Eksempel – fordelingen af varians- estimatet i normalfordelte data

```
% Population  
mu      = 5;  
sigma   = 2;  
n       = 10;  
  
num_experiments = 1000;
```



Testkatalog for varians i normalfordelte data

- Statistisk model
 - $X \sim N(\mu, \sigma^2)$
 - Parameterskøn: $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \sim \frac{\sigma^2}{n-1} \chi^2(n-1)$
 - Hvor \bar{x} er gennemsnittet.
- Hypotesetest
 - $H: \sigma^2 = \sigma_0^2$
 - Teststørrelse: $v = s^2 / \sigma^2$
 - P-værdi: $pval = \begin{cases} 2 \cdot \chi_{cdf}^2((n-1) \cdot v, n-1) & v \leq 1 \\ 2 \cdot (1 - 2 \cdot \chi_{cdf}^2((n-1) \cdot v, n-1)) & v > 1 \end{cases}$
- 95% konfidensinterval
 - $[s_-(x); s_+(x)] = \left[\frac{(n-1) \cdot s^2}{\chi_{inv}^2(0,975, n-1)}; \frac{(n-1) \cdot s^2}{\chi_{inv}^2(0,025, n-1)} \right]$

Bemærkning

- Der kommer ingen eksamensopgaver, hvor man skal lave test eller beregne konfidensinterval for variansen.

Testkatalog for middelværdien i normalfordelte data (kendt varians)

- Statistisk model
 - $X \sim N(\mu, \sigma^2)$
 - Parameterskøn: $\hat{\mu} = \bar{x} \sim N(\mu, \sigma^2/n)$
 - Hvor observationen \bar{x} er gennemsnittet

- Hypotesetest

- $H: \mu = \mu_0$

- Teststørrelse: $z = \frac{\bar{x} - \mu_0}{\sqrt{\sigma^2/n}}$

- P-værdi: $pval = 2 \cdot |1 - \Phi(|z|)|$

Varsians
ukendt!

Så lad os se på
t-fordelingen

- 95% konfidensinterval

- $[\mu_-(x); \mu_+(x)] = [\bar{x} - u \cdot \sqrt{\sigma^2/n}; \bar{x} + u \cdot \sqrt{\sigma^2/n}]$

- Hvor $u = 1,96$

- **Forudsætning: Vi skal kende den sande varians, σ^2 . Ellers brug en t-fordeling (senere...).**

Ny teststørrelse

- I stedet for

$$z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

- Indfører vi teststørrelsen

$$t = \frac{\bar{x} - \mu}{\sqrt{s^2/n}} \sim t(n - 1)$$

t-fordelingen

- Fordelingsfunktion
- Vi har

$$t = \frac{\bar{x} - \mu}{\sqrt{s^2/n}} \sim t(n - 1)$$

- Notation

$$\Pr(T \leq t) = t_{cdf}(t, n - 1)$$

- Matlab

- Fordelingsfunktion $\Pr(T \leq t) = \text{tcdf}(t, f)$
- Invers $t = \text{tinv}(\alpha, f)$

Testkatalog for middelværdien i normalfordelte data (ukendt varians)

- Statistisk model
 - $X \sim N(\mu, \sigma^2)$
 - Parameterskøn: $\hat{\mu} = \bar{x}$, $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
 - Hvor observationen \bar{x} er gennemsnittet
- Hypotesetest
 - $H: \mu = \mu_0$
 - Teststørrelse: $t = \frac{\bar{x} - \mu_0}{\sqrt{s^2/n}} \sim t(n-1)$
 - P-værdi: $pval = 2 \cdot \left(1 - t_{cdf}(|t|, n-1)\right)$
- 95% konfidensinterval
 - $[\mu_-(x); \mu_+(x)] = \left[\bar{x} - t_0 \cdot \sqrt{s^2/n}; \bar{x} + t_0 \cdot \sqrt{s^2/n}\right]$
 - Hvor $t_0 = t_{inv}(0.975, n-1)$

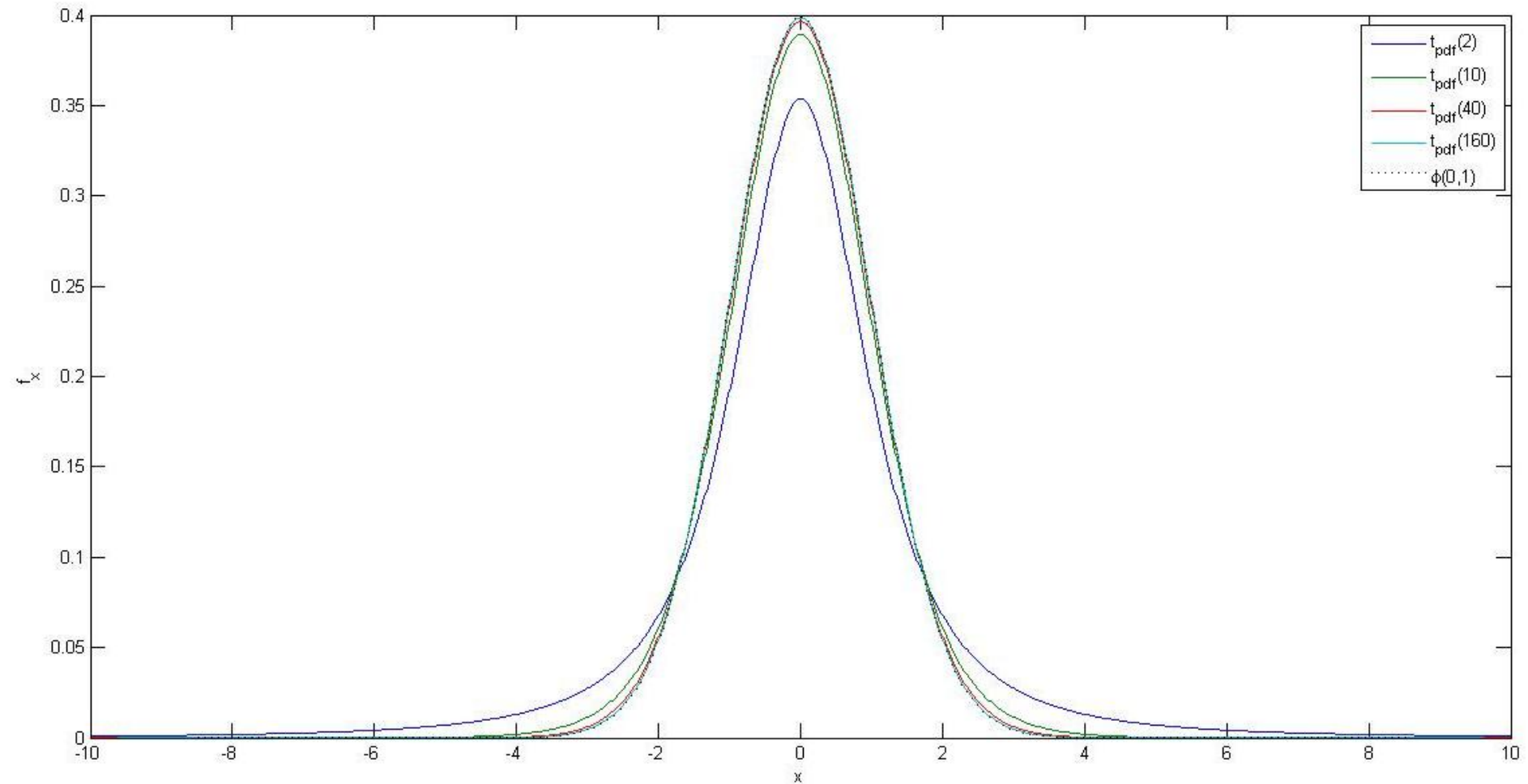
t -fordelingens konvergens mod standard normalfordelingen

- Efterhånden som antallet af datapunkter (n) bliver større, bliver estimeret (s^2) af variansen bedre.
- For små n er 95% konfidensintervallet, beregnet med `tinv`, bredere end det tilsvarende 95% konfidensinterval beregnet med `norminv`.
- For store n er 95% konfidensintervallerne cirka ens.

$f = n - 1$

f	10	20	40	80	160	∞
$t_{\text{inv}}(0.975, f)$	2.23	2.09	2.02	1.99	1.97	1.96

t -fordelingens konvergens mod standard normalfordelingen



Eksempel – jordens massefylde

```
% Eksempel 4 - t-test: jordens massefylde
% Tabel 6.1: Henry Cavendishs målinger af jordens massetæthed i 1797.
x = [ 5.36 5.29 5.58 5.65 5.57 5.53 5.62 5.29 ...
      5.44 5.34 5.79 5.10 5.27 5.39 5.42 5.47 ...
      5.63 5.34 5.46 5.30 5.75 5.68 5.85 ];
n = length(x);
mu0 = 5.517;
s2 = var(x);

% Teststørrelse (H:  $\mu = \mu_0 = 5.517$ )
x_hat = mean(x)
t = (x_hat - mu0) / sqrt(s2/n)
pval = 2*(1-tcdf(abs(t), n-1))

% 95% konfidensinterval
t0 = tinv(0.975, n-1)
mu_nedre = x_hat - t0*sqrt(s2/n)
mu_oevre = x_hat + t0*sqrt(s2/n)
```

Fraktilsammenligning

- Eksempel på statistisk model
 - Givet n uafhængige målinger x_1, x_2, \dots, x_n
 - Det kunne fx være indholdet [mL] af flasker.
 - Så kunne vi antage, at

$$X_i \sim N(\mu, \sigma^2)$$

- Men hvordan tjekker man egentlig, om dette er en rimelig antagelse?

Fraktilsammenligning

- Den empiriske fordelingsfunktion $F_n(x)$
 - Antallet af observationer med værdi mindre end eller lig med x , divideret med det samlede antal observationer n .
- Grundlæggende ide
 - Indtegn punkterne $\left(\Phi \left(\frac{x-\mu}{\sigma} \right), F_n(x) \right)$ for en række x -værdier.
 - Tjek om punkterne ligger på identitetslinjen.

Fraktilsammenligning

- Problem med den grundlæggende ide
 - Indtegn punkterne $\left(\Phi \left(\frac{x-\mu}{\sigma} \right), F_n(x) \right)$ for en række x -værdier.
 - Tjek om punkterne ligger på identitetslinjen.
 - **Vi kender ikke μ og σ .**

Fraktilsammenligning

- Så i stedet for at indtegne punkterne

$$\left(\Phi \left(\frac{x - \mu}{\sigma} \right), F_n(x) \right)$$

- indtegner man punkterne

$$\left(x_{[i]}, \Phi^{-1} \left(\frac{i - 1/2}{n} \right) \right)$$

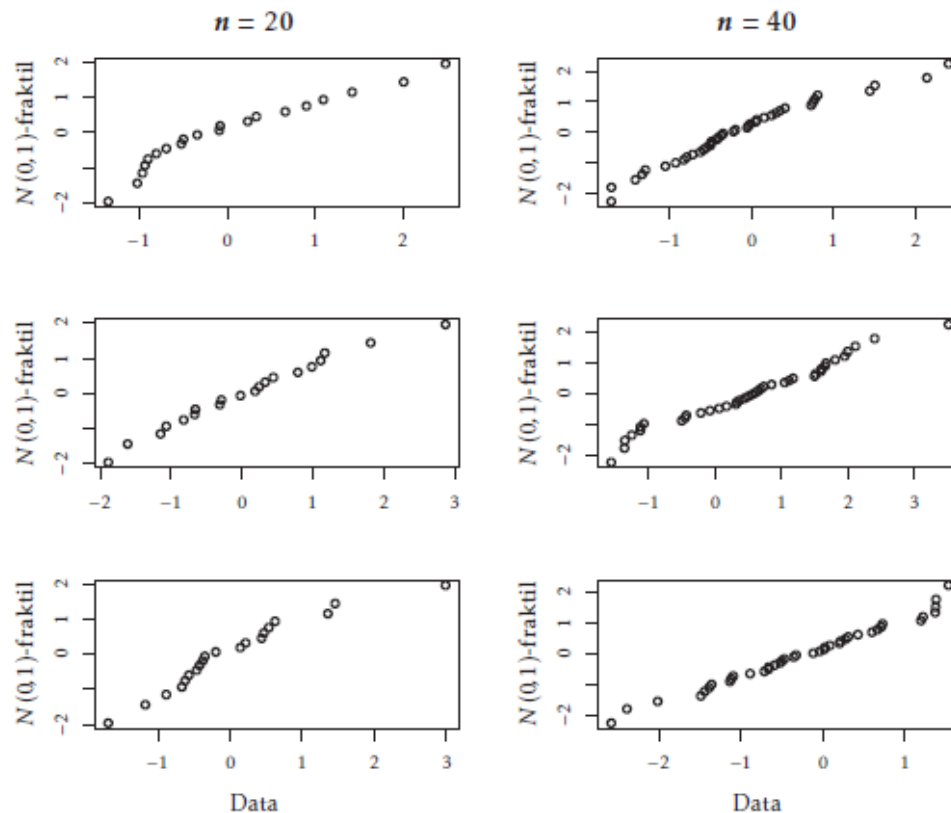
- hvor $x_{[1]} \leq x_{[2]} \leq \dots \leq x_{[n]}$ er en ascenderende sortering af de oprindelige data (x_1, x_2, \dots, x_n) .

Fraktilsammenligning

- Data kan med rimelighed antages normalfordelte, hvis punkterne i plottet snor sig omkring en ret linje.
- Omvendt, hvis der er store systematiske afvigelser, tror vi ikke på, data er normalfordelte.
- I praksis kan det være svært at afgøre, hvis antallet af datapunkter n er lille.

Eksempler på naturlige afvigelser fra en ret linje.

- Matlab: `qqplot`



Figur 6.2: Eksempler på fraktilsammenligninger med data simuleret fra en standard normalfordeling. I venstre søjle er der $n = 20$ observationer i hvert af de tre datasæt, og i højre søjle er der $n = 40$ observationer i hvert datasæt.

Eksempel – jordens massefylde

