02402: Introduktion til Statistik

Forelæsning 4: Konfidensinterval for middelværdi (og varians)

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Oversigt

1. Introduktion og eksempel
2. Fordelingen for gennemsnittet
   - t-fordelingen
3. Konfidensintervallet for μ
   - Eksempel: Højder
4. Den statistiske sprogbrug og formelle ramme
5. Ikke-normale data, *den centrale grænseværdisætning (CLT)*
6. Formel fortolkning af konfidensintervallet
7. Konfidensinterval for varians og spredning

Overview

1. Introduktion og eksempel
2. Fordelingen for gennemsnittet
   - t-fordelingen
3. Konfidensintervallet for μ
   - Eksempel: Højder
4. Den statistiske sprogbrug og formelle ramme
5. Ikke-normale data, *den centrale grænseværdisætning (CLT)*
6. Formel fortolkning af konfidensintervallet
7. Konfidensinterval for varians og spredning

Eksempel - Højde af 10 studerende:

Stikprøve, $n = 10$:

168 161 167 179 184 166 198 187 191 179

Stikprøvegennemsnit og -standardafvigelse:

$$\bar{x} = 178$$
$$s = 12.21$$

Estimater for populationens middelværdi og -standardafvigelse:

$$\hat{\mu} = 178$$
$$\hat{\sigma} = 12.21$$

NYT:**Konfidensinterval for μ**:

$$178 \pm 2.26 \cdot \frac{12.21}{\sqrt{10}} \Leftrightarrow [169.3;\ 186.7]$$

NYT:**Konfidensinterval for σ**:

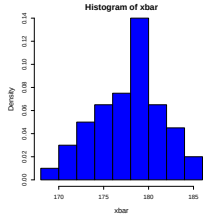
$$[8.4;\ 22.3]$$

Overview

1. Introduktion og eksempel
2. Fordelingen for gennemsnittet
   - t-fordelingen
3. Konfidensintervallet for μ
   - Eksempel: Højder
4. Den statistiske sprogbrug og formelle ramme
5. Ikke-normale data, *den centrale grænseværdisætning (CLT)*
6. Formel fortolkning af konfidensintervallet
7. Konfidensinterval for varians og spredning

(Empirisk) fordeling af stikprøvegennemsnittet

```r
# 'Sand' middelværdi og standardafvigelse
mu <- 178
sigma <- 12

# Sample size
n <- 10

# Simulér normalfordelte X_i for n = 10
x <- rnorm(n = n, mean = mu, sd = sigma)
x

# Empirisk tæthed
hist(x, prob = TRUE, col = 'blue')
# Stikprøvegennemsnit
mean(x)

# Gentag eksperimentet (100 gange)
mat <- replicate(100, rnorm(n = n, mean = mu, sd = sigma))

# Udregn gennemsnit for hver stikprøve
xbar <- apply(mat, 2, mean)
xbar

# Fordelingen af stikprøvegennemsnittene (vist til højre)
hist(xbar, prob = TRUE, col = 'blue')
# Gns. og varians af stikprøvegennemsnittene
mean(xbar)
var(xbar)
```



Sætning 3.3: Fordeling for gennemsnittet af normalfordelinger

(Stikprøve-)fordelingen for $\bar{X}$

Antag at $X_1, \ldots, X_n$ er uafhængige, ensfordelte (i.i.d) normalfordelte stokastiske variable, $X_i \sim N(\mu, \sigma^2), i = 1, \ldots, n$, så:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Middelværdi og varians følger af regneregler

Middelværdien af $\bar{X}$ (Theorem 2.56):

$$E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu$$

Variansen for $\bar{X}$ (Theorem 2.56):

$$\mathrm{Var}(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Normaliteten af $\bar{X}$ (Theorem 2.40):

Fra denne sætning følger at $\bar{X}$ er normalfordelt med middelværdi μ varians σ^2/n.

Fordelingen af den fejl, vi begår $\bar{X} - \mu$

Spredningen af $\bar{X}$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Spredningen af $(\bar{X} - \mu)$

$$\sigma_{(\bar{X}-\mu)} = \frac{\sigma}{\sqrt{n}}$$

Standardiseret version af de samme ting, Theorem 3.4:

Fordelingen for den *standardiserede* fejl, vi begår:

Antag at $X_1, \ldots, X_n$ er uafhængige, ensfordelte (i.i.d.) normalfordelte stokastiske variable $X_i \sim N\left(\mu, \sigma^2\right)$ hvor $i = 1, \ldots, n$, så:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N\left(0, 1^2\right)$$

Dvs. at det standardiserede stikprøvegennemsnittet Z følger en standard-normalfordeling.

Praktisk problem i alt dette!:

Hvordan skal resultaterne fra de foregående slides omsættes til et konkret interval for μ?

Når nu populationsspredningen σ indgår i alle formlerne?

Oplagt løsning:

Anvend estimtatet s i stedet for σ i formlerne!

MEN MEN:

Så bryder den givne teori faktisk sammen!!

HELDIGVIS:

Der findes heldigvis en udvidet teori, der kan klare det!!

Theorem 3.5: Mere anvendeligt resultat: (kopi af Theorem 2.49)

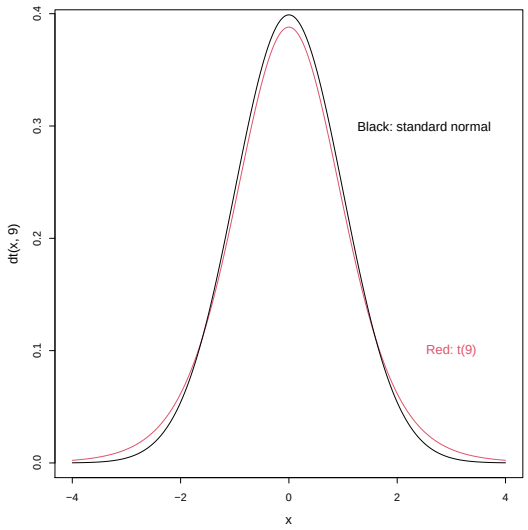
t-fordelingen tager højde for usikkerheden i at bruge s:

Antag at $X_1, \ldots, X_n$ er uafhængige og ensfordelte (i.i.d.) normalfordelte stokastiske variable, hvor $X_i \sim N\left(\mu, \sigma^2\right)$ og $i = 1, \ldots, n$, så er:

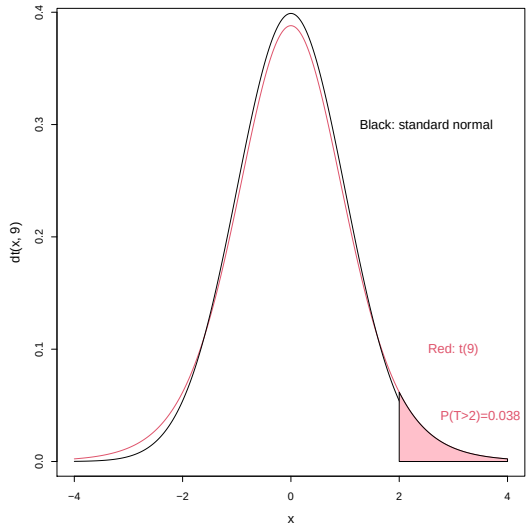
$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t$$

hvor t er t-fordelingen med $n - 1$ frihedsgrader (degrees of freedom, `df`).

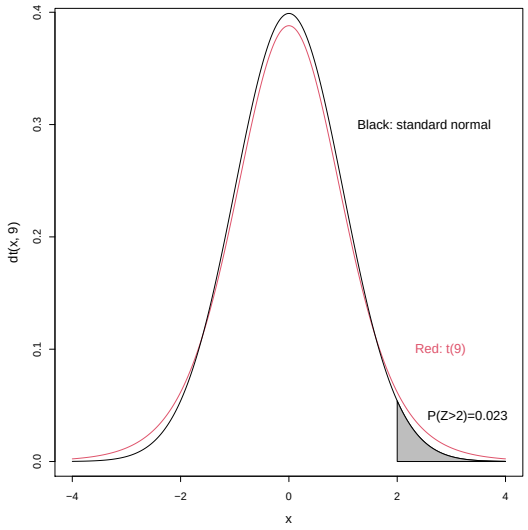
t-fordelingen med 9 frihedsgrader ($n = 10$):



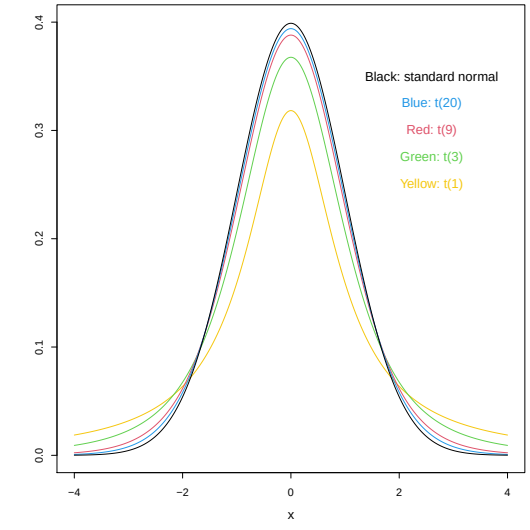
t-fordelingen med 9 frihedsgrader og standardnormalfordelingen:



t-fordelingen med 9 frihedsgrader og standardnormalfordelingen:



Forskellige t-fordelinger:



Overview

1. Introduktion og eksempel
2. Fordelingen for gennemsnittet
   - *t*-fordelingen
3. Konfidensintervallet for μ
   - Eksempel: Højder
4. Den statistiske sprogbrug og formelle ramme
5. Ikke-normale data, *den centrale grænseværdisætning (CLT)*
6. Formel fortolkning af konfidensintervallet
7. Konfidensinterval for varians og spredning

Metodeboks 3.8: One-sample konfidensinterval for μ

Brug den rigtige t-fordeling til at lave konfidensintervallet:

For en stikprøve $x_1, \ldots, x_n$ er $100(1-\alpha)\%$ konfidensintervallet givet ved:

$$\bar{x} \pm t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

hvor $t_{1-\alpha/2}$ er $100(1-\alpha)\%$ fraktilen i t-fordelingen med $n-1$ frihedsgrader.

Mest almindeligt med $\alpha = 0.05$:

Oftest bruger man 95%-konfidensintervallet:

$$\bar{x} \pm t_{0.975} \cdot \frac{s}{\sqrt{n}}$$

Højde-eksempel

```
## The t-quantiles for n=10:
qt(0.975,9)
```

[1] 2.262
Giver os at $t_{0.975} = 2.26$.
Vi kan genkende det allerede angivne resultat:

$$178 \pm 2.26 \cdot \frac{12.21}{\sqrt{10}}$$

som er:

$$178 \pm 8.74 = [169.3;\ 186.7]$$

Højde-eksempel, 99% Konfidensinterval (CI)

```
qt(0.995,9)
```

[1] 3.25
Giver resultatet $t_{0.975} = 2.26$.

I dette tilfælde fås

$$178 \pm 3.25 \cdot \frac{12.21}{\sqrt{10}}$$

som giver

$$178 \pm 12.55 = [165.4;\ 190.6]$$

Der findes en R-funktion, der kan gøre det hele (og mere til):

```
x <- c(168,161,167,179,184,166,198,187,191,179)
t.test(x,conf.level=0.99)

##
##  One Sample t-test
##
## data:  x
## t = 46, df = 9, p-value = 5e-12
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  165.5 190.5
## sample estimates:
## mean of x
##       178
```


Overview

1. Introduktion og eksempel
2. Fordelingen for gennemsnittet
   - t-fordelingen
3. Konfidensintervallet for μ
   - Eksempel: Højder
4. Den statistiske sprogbrug og formelle ramme
5. Ikke-normale data, *den centrale grænseværdisætning (CLT)*
6. Formel fortolkning af konfidensintervallet
7. Konfidensinterval for varians og spredning

Den formelle ramme for *statistisk inferens*

Fra bogen, kapitel 1:

- An *observational unit* is the single entity/level about which information is sought (e.g. a person) (**Observationsenhed**)
- The *statistical population* consists of all possible "measurements" on each *observational unit* (**Population**)
- The *sample* from a statistical population is the actual set of data collected. (**Stikprøve**)

Sprogbrug og koncepter:

- μ og σ er *parametre*, som beskriver populationen
- $\bar{x}$ er *estimatet* for μ (konkret udfald)
- $\bar{X}$ er *estimatoren* for μ (nu set som stokastisk variabel)
- Begrebet *'statistic(s)'* er en fællesbetegnelse for begge

Den formelle ramme for *statistisk inferens* - Eksempel

Fra bogen, kapitel 1, højdeeksempel

Vi måler højden for 10 tilfældige personer i Danmark

Stikprøven:

De 10 konkrete talværdier: $x_1, \ldots, x_{10}$

Populationen:

Højderne for alle mennesker i Danmark.

Observationsenheden:

Én person

Statistisk inferens = Learning from data

Learning from data:

Is learning about parameters of distributions that describe populations.

Vigtigt i den forbindelse:

Stikprøven skal på meningsfuld vis være *repræsentativ* for en eller anden veldefineret population

Hvordan sikrer man det:

F.eks. ved at sikre at stikprøven er fuldstændig *tilfældigt udtaget*

Tilfældig stikprøveudtagning (random sampling)

Definition 3.12 :

- En tilfældig stikprøve fra en (uendelig) population: Observationerne $X_1, X_2, ..., X_n$ udgør en tilfældig stikprøve af størrelse n fra den uendelige population $f(x)$ hvis:
  1. Hvert X_i er en stokastisk variabel med fordeling $f(x)$
  2. De n stokastiske variable er uafhængige

Hvad betyder det?

1. Alle observationer skal komme fra den samme population
2. De må IKKE dele information med hinanden (f.eks. hvis man havde udtaget hele familier i stedet for enkeltindivider)

Overview

1. Introduktion og eksempel
2. Fordelingen for gennemsnittet
   - t-fordelingen
3. Konfidensintervallet for μ
   - Eksempel: Højder
4. Den statistiske sprogbrug og formelle ramme
5. Ikke-normale data, *den centrale grænseværdisætning (CLT)*
6. Formel fortolkning af konfidensintervallet
7. Konfidensinterval for varians og spredning

Theorem 3.14: Den centrale Grænseværdisætning (CLT)

Gennemsnittet af en tilfældig stikprøve følger altid en normalfordeling hvis n er stor nok:

Lad $\bar{X}$ være gennemsnittet for en tilfældigt udtrukket stikprøve af størrelse n taget fra en population med middelværdi μ og varians σ^2, så gælder at

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

opfylder at fordelingen nærmer sig til standard-normalfordelingen $N(0, 1^2)$, når $n \to \infty$.

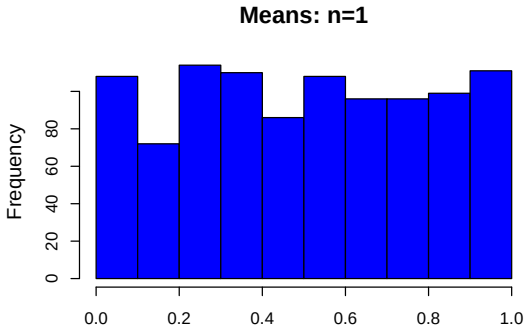
Dvs., hvis n er stor nok, kan vi (tilnærmelsesvist) antage:

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

Engelsk: *Central Limit Theorem* (CLT)

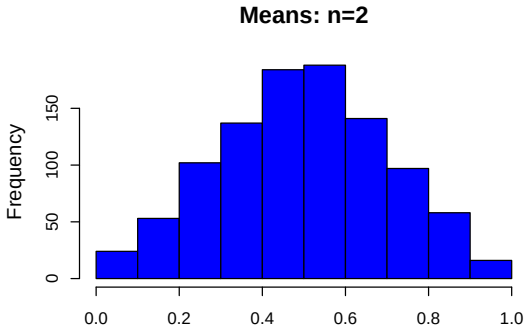
CLT in action - gennemsnit af uniformt fordelte observationer

```
n=1
k=1000
u=matrix(runif(k*n),ncol=n)
hist(apply(u,1,mean), col="blue", main="Means: n=1", xlab="")
```



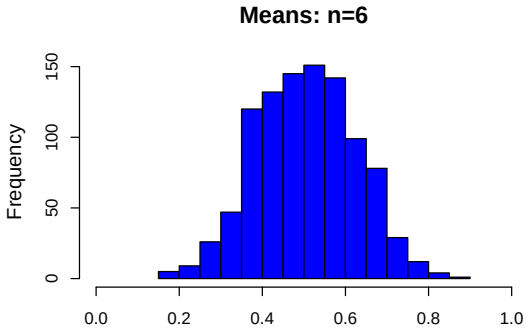
CLT in action - gennemsnit af uniformt fordelte observationer

```
n=2
k=1000
u=matrix(runif(k*n),ncol=n)
hist(apply(u,1,mean),col="blue",main="Means: n=2",xlab="")
```



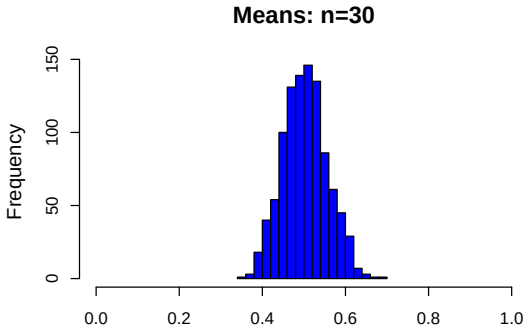
CLT in action - gennemsnit af uniformt fordelte observationer

```
n=6
k=1000
u=matrix(runif(k*n),ncol=n)
hist(apply(u,1,mean),col="blue",main="Means: n=6",xlab="", xlim=c(0,1))
```



CLT in action - gennemsnit af uniformt fordelte observationer

```
n=30
k=1000
u=matrix(runif(k*n),ncol=n)
hist(apply(u,1,mean),col="blue",main="Means: n=30",xlab="", nclass=15, xlim=c(0,1))
```



Konsekvens af CLT:

Vores CI-metode virker også for ikke-normale data:

Vi kan bruge konfidens-interval baseret på t-fordelingen i stort set alle situationer, blot n er "stor nok"

Hvornår er n "stor nok"?

Faktisk svært at svare præcist på, MEN:

- Tommelfingerregel:$n \geq 30$
- Selv for mindre n kan formlen være (næsten)gyldig for ikke-normale data.

Overview

1. Introduktion og eksempel
2. Fordelingen for gennemsnittet
   - t-fordelingen
3. Konfidensintervallet for μ
   - Eksempel: Højder
4. Den statistiske sprogbrug og formelle ramme
5. Ikke-normale data, *den centrale grænseværdisætning (CLT)*
6. Formel fortolkning af konfidensintervallet
7. Konfidensinterval for varians og spredning

'Repeated sampling' fortolkning

I det lange løb fanger vi den sande værdi i 95% af tilfældene:

Konfidensintervallet vil variere i både bredde (s) og position ($\bar{x}$) hvis man gentager sit studie.

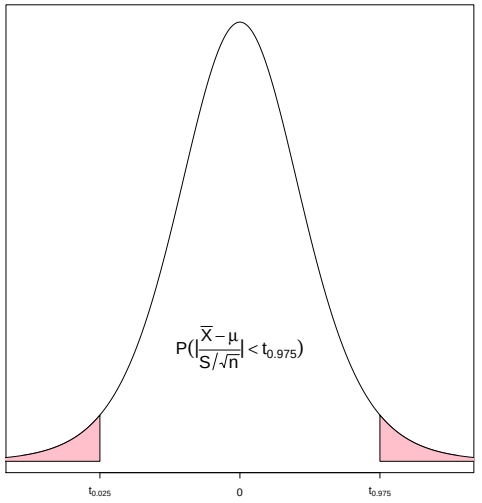
Mere formelt udtrykt (Theorem 3.4 og 2.49):

$$P\left(\frac{|\bar{X}-\mu|}{S/\sqrt{n}} < t_{0.975}\right) = 0.95$$

Som er ækvivalent med:

$$P\left(\bar{X} - t_{0.975}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{0.975}\frac{S}{\sqrt{n}}\right) = 0.95$$

'Repeated sampling' fortolkning



Overview

1. Introduktion og eksempel
2. Fordelingen for gennemsnittet
   - t-fordelingen
3. Konfidensintervallet for μ
   - Eksempel: Højder
4. Den statistiske sprogbrug og formelle ramme
5. Ikke-normale data, *den centrale grænseværdisætning (CLT)*
6. Formel fortolkning af konfidensintervallet
7. Konfidensinterval for varians og spredning

Motiverende eksempel

Produktion af tabletter.

I produktion af tabletter blandes et aktivt stof med et pulver, hvorefter blandingen formes til tabletter. Vi producerer pulverblanding og piller deraf. Det er vigtigt at blandingen er så homogen (ensartet) som muligt, så at tabletternes styrke er ens.

Vi betragter en blanding af det aktive stof og fyldpulver, hvoraf vil vil producere en stor mængde tabletter.

Vi ønsker at producere blandingerne, så at koncentrationen af det aktive stof i tabletterne skal være 1 mg/g med den mindst mulige spredning En tilfældig stikprøve udtages, hvor vi måler mængden af aktivt stof (i mg/g). Vi antager at vores målinger følger en normalfordeling.

Stikprøvefordelingen for varians-estimatet, Theorem 2.81

Antag i.i.d. normalfordelte variable, $X_i \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, n$.

Variansestimatet opfører sig som en χ^2-fordeling:

Lad

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

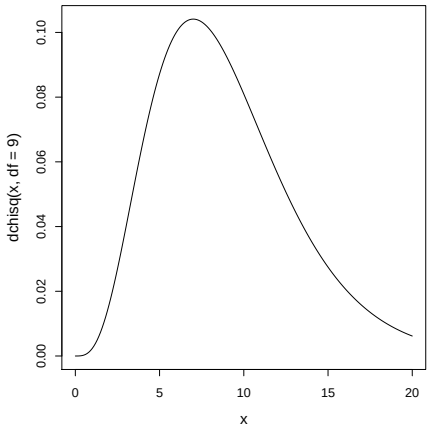
Så gælder:

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

er en stokastisk variabel, som er χ^2-fordelt med $\nu = n - 1$ frihedsgrader.

χ^2-fordelingen med $\nu = 9$ frihedsgrader (degrees of freedom)

```
x <- seq(0, 20, by = 0.1)
plot(x, dchisq(x, df = 9), type = "l")
```



Metode 3.19: Konfidensinterval for stikprøvevarians og -spredning

Antag i.i.d. normalfordelte variable, $X_i \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, n$.

Variansen:

Et $100(1-\alpha)\%$ konfidensinterval for stikprøvevariansen: $\hat{\sigma}^2$ er:

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}; \frac{(n-1)s^2}{\chi^2_{\alpha/2}} \right]$$

hvor fraktilerne kommer fra en χ^2-fordeling med $\nu = n-1$ frihedsgrader.

Standardafvigelsen:

Et $100(1-\alpha)\%$ konfidensinterval for stikprøvestandardafvigelsen $\hat{\sigma}$ er:

$$\left[\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}}; \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}} \right]$$

Eksempel

Data:

En tilfældig stikprøve med $n = 20$ tabletter er udtaget og fra denne får man:

$$\hat{\mu} = \bar{x} = 1.01, \ \hat{\sigma}^2 = s^2 = 0.07^2$$

95%-konfidensinterval for variansen - vi skal bruge χ^2-fraktilerne (19 frihedsgrader):

$$\chi^2_{0.025} = 8.9065, \ \chi^2_{0.975} = 32.8523$$

```
qchisq(c(0.025, 0.975), df = 19)
```

```
[1] 8.907 32.852
```


Eksempel

Så konfidensintervallet for variansen σ^2 bliver:

$$\left[\frac{19 \cdot 0.7^2}{32.85}; \frac{19 \cdot 0.7^2}{8.907} \right] = [0.002834; \ 0.01045]$$

Og konfidensintervallet for spredningen σ bliver:

$$\left[\sqrt{0.002834}; \sqrt{0.01045} \right] = [0.053; \ 0.102]$$

Højdeeksempel

Vi skal bruge χ^2-fraktilerne med $\nu = 9$ frihedsgrader:

$$\chi^2_{0.025} = 2.700389, \ \chi^2_{0.975} = 19.022768$$

```
qchisq(c(0.025, 0.975), df = 9)
```

```
[1] 2.70 19.02
```

Så konfidensintervallet for højdens standardafvigelse σ bliver:

$$\left[\sqrt{\frac{9 \cdot 12.21^2}{19.022768}}; \sqrt{\frac{9 \cdot 12.21^2}{2.700389}} \right] = [8.4; \ 22.3]$$

Eksempel - Højde af 10 studerende - recap:

Stikprøve, $n = 10$:

168 161 167 179 184 166 198 187 191 179

Gennemsnit og standardafvigelse for stikprøven:

$$\bar{x} = 178$$
$$s = 12.21$$

Estimater for populationsgennemsnit og standardafvigelse:

$$\hat{\mu} = 178$$
$$\hat{\sigma} = 12.21$$

NYT:**Konfidensinterval, μ**:

$$178 \pm 2.26 \cdot \frac{12.21}{\sqrt{10}} \Leftrightarrow [169.3;\ 186.7]$$

NYT:**Konfidensinterval, σ**:

$$[8.4;\ 22.3]$$

Oversigt

1. Introduktion og eksempel
2. Fordelingen for gennemsnittet
   - t-fordelingen
3. Konfidensintervallet for μ
   - Eksempel: Højder
4. Den statistiske sprogbrug og formelle ramme
5. Ikke-normale data, *den centrale grænseværdisætning (CLT)*
6. Formel fortolkning af konfidensintervallet
7. Konfidensinterval for varians og spredning