

# Introduktion til Statistik

## Forelæsning 4: Konfidensinterval for middelværdi (og spredning)

Peder Bacher

DTU Compute, Dynamiske Systemer  
Bygning 303B, Rum 009  
Danmarks Tekniske Universitet  
2800 Lyngby – Danmark  
e-mail: pbac@dtu.dk

Forår 2017

# Chapter 3: Konfidensintervaller for én gruppe/stikprøve

## Grundlæggende koncepter

- Population og tilfældig stikprøve
- Estimation (*f.eks.  $\hat{\mu}$  er estimat af  $\mu$* )
- Signifikansniveau  $\alpha$
- Konfidensintervaller (*fanger rigtige prm.  $1 - \alpha$  af gangene*)
- Stikprøvefordelinger (*stikprøvegennemsnit ( $t$ ) og empirisk varians ( $\chi^2$ )*)
- Centrale grænseværdisætning

## Specifikke metoder, én gruppe/stikprøve

- Konfidensinterval for middelværdi ( $t$ -fordeling)
- Konfidensinterval for varians ( $\chi^2$ -fordeling)

# Chapter 3: One sample confidence intervals

## General concepts

- Population and a random sample
- Estimation (*e.g.  $\hat{\mu}$  is estimate of  $\mu$* )
- Significance level  $\alpha$
- Confidence intervals (*Catches true value  $1 - \alpha$  times*)
- Sampling distributions (*sample mean ( $t$ ) and sample variance ( $\chi^2$ )*)
- Central Limit Theorem

## Specific methods, one sample

- Confidence interval for the mean ( $t$ -distribution)
- Confidence interval for the variance ( $\chi^2$ -distribution)

# Overview

- 1 Fordelingen for gennemsnittet
  - $t$ -fordelingen
- 2 Konfidensintervallet for  $\mu$ 
  - Eksempel
- 3 Den statistiske sprogbrug og formelle ramme
- 4 Ikke-normale data, Central Grænseværdisætning (CLT)
- 5 Konfidensinterval for varians og spredning

## Theorem 3.2: Fordeling for gennemsnit af normalfordelinger

(Stikprøve-) fordelingen/ The (sampling) distribution for  $\bar{X}$

Assume that  $X_1, \dots, X_n$  are independent and identically normally distributed random variables,  $X_i \sim N(\mu, \sigma^2)$  and  $i = 1, \dots, n$ , then:

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

# Middelværdi og varians følger af regneregler

Theorem 2.40: Lineær funktion af normal distribuerede variable er også normalfordelt

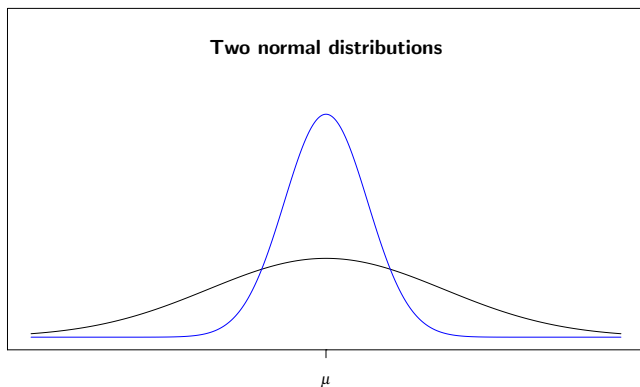
Theorem 2.53: Middelværdien af  $\bar{X}$

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

Theorem 2.53: Variansen for  $\bar{X}$

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

# Spørgsmål om stikprøvegennemsnittet (socrative.com, room: PBAC)



Den ene pdf hører til  $X_i$  og den anden til  $\bar{X}$ . Hvad kan konkluderes (for  $n > 1$ )?

- A: Den sorte hører til  $X_i$  og den blå til  $\bar{X}$

Svar A:

# Simuler i R: Middelværdi og spredning af stikprøvegennemsnit

```
## Simuler stikprøvegennemsnit af normalfordelt stokastisk variabel

## Middelværdien
mu <- -5
## Standard afvigelsen
sigma <- 2
## Stikprøvestørrelsen
n <- 50

## Simuler normalfordelte  $X_i$ 
x <- rnorm(n=n, mean=mu, sd=sigma)
## Se realiseringerne
x
## Empirisk tæthed
hist(x, prob=TRUE, col='blue')

## Beregn gennemsnittet (stikprøve middelværdien, i.e. sample mean)
mean(x)
## Beregn stikprøvevariansen (sample variance)
var(x)

## Gentag den simulerede stikprøvetagning mange gange
mat <- replicate(100, rnorm(n=n, mean=mu, sd=sigma))
## Beregn gennemsnittet for hver af dem
xbar <- apply(mat, 2, mean)
## Nu har vi mange realiseringer af stikprøvegennemsnittet
xbar
## Se deres fordeling
```



## Standardiseret fejl vi begår, Corollary 3.3:

Når vi bruger  $\bar{X}$  som estimat for  $\mu$ :

Så begår vi fejlen  $\bar{X} - \mu$

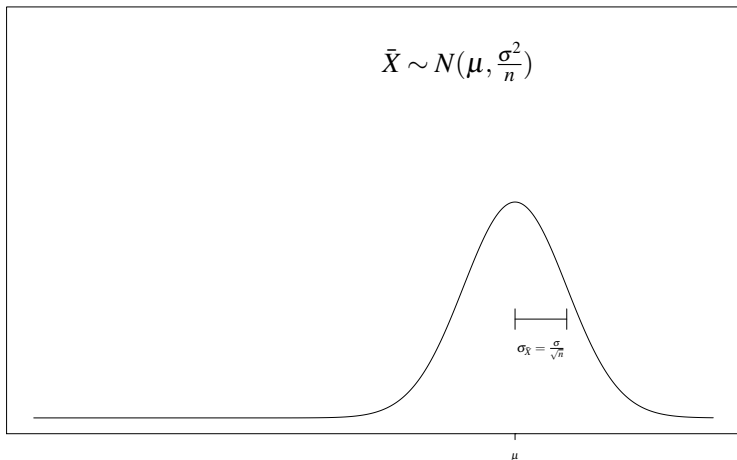
Fordelingen for den standardiserede fejl vi begår:

Assume that  $X_1, \dots, X_n$  are independent and identically normally distributed random variables,  $X_i \sim N(\mu, \sigma^2)$  where  $i = 1, \dots, n$ , then:

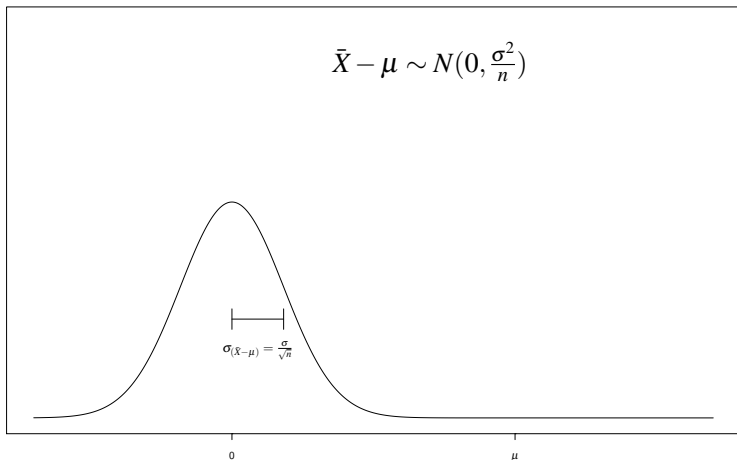
$$Z = \frac{\bar{X} - \mu}{\sigma_{(\bar{X} - \mu)}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1^2)$$

That is, the standardized sample mean  $Z$  follows a *standard normal distribution*.

## Transformation til standard normalfordeling:

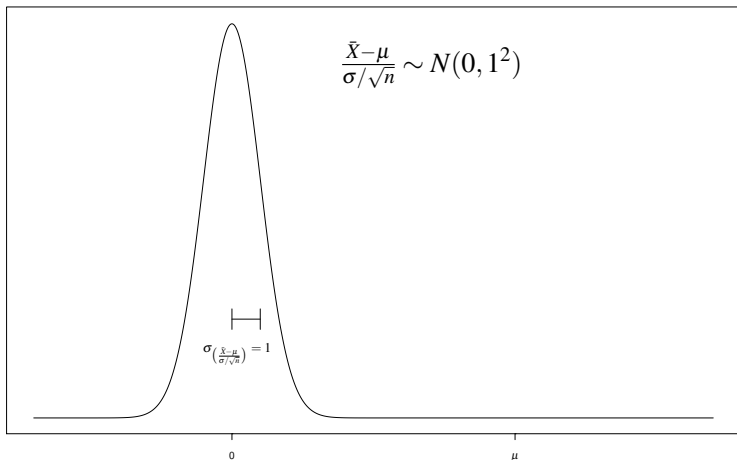
Pdf for gennemsnittet  $\bar{X}$  når  $X_i \sim N(\mu, \sigma^2)$ 

## Transformation til standard normalfordeling:

Pdf for *fejlen vi begår*  $\bar{X} - \mu$  når  $X_i \sim N(\mu, \sigma^2)$ 

Transformation til standard normalfordeling:

Pdf for *den standardiserede fejl*  $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$  når  $X_i \sim N(\mu, \sigma^2)$



Standardiseret til *standard normalfordeling* (noteres  $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1^2)$ )

# Nu kan et 95% konfidensinterval udledes

95% konfidensinterval for  $\mu$ :

$$P(z_{0.025} < Z < z_{0.975}) = 0.95 \quad \Leftrightarrow$$

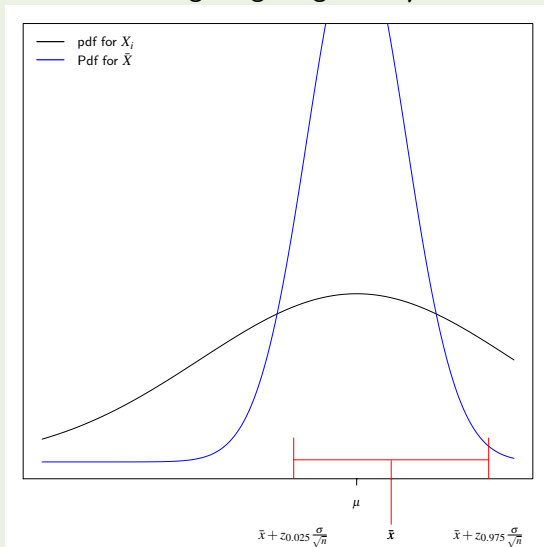
$$P\left(z_{0.025} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{0.975}\right) = 0.95 \quad \Leftrightarrow$$

$$P\left(z_{0.025} \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{0.975} \frac{\sigma}{\sqrt{n}}\right) = 0.95 \quad \Leftrightarrow$$

$$P\left(\bar{X} + z_{0.025} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{0.975} \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

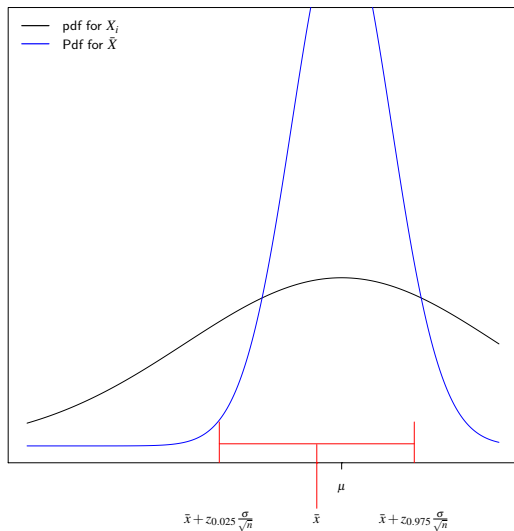
# 1. simulering: Beregning af 95% konfidensinterval

Konfidensintervallet er omkring  $\bar{x}$  og fanger her  $\mu$



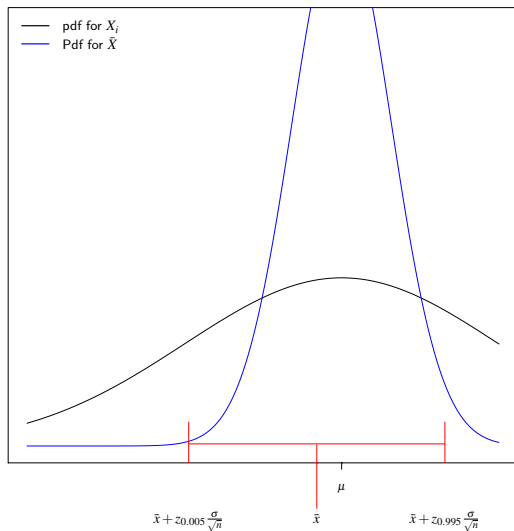
## 2. simulering: Beregning af 95% konfidensinterval

Konfidensintervallet er omkring  $\bar{x}$  og fanger her  $\mu$



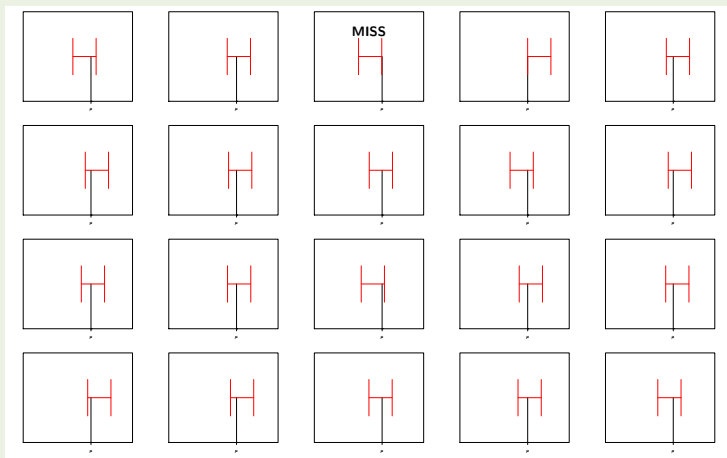
## 2. simulering: Beregning af 99% konfidensinterval

99% konfidensintervallet er breddere end 95% konfidensintervallet (det skal fange  $\mu$  of

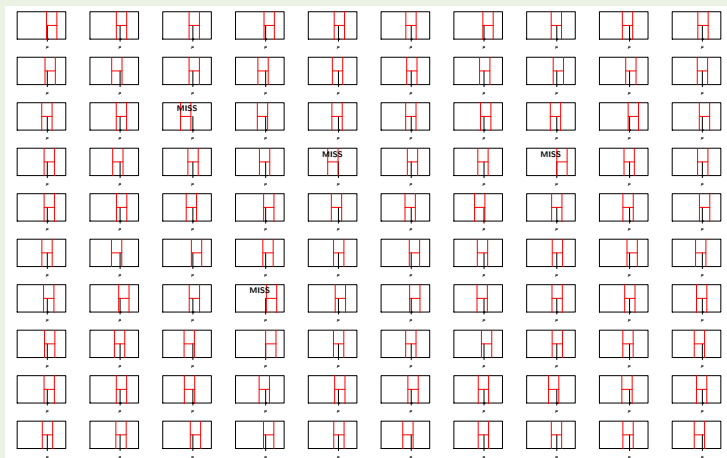




## 20 simuleringer: Beregning af 95% konfidensinterval



# 100 simuleringer: Beregning af 95% konfidensinterval



## Spørgsmål om konfidensinterval ([socrative.com](https://socrative.com), room: PBAC)

Hvis vi planlægger at beregne et 98% konfidensinterval for middelværdien, hvad er da sandsynligheden for at middelværdien *ikke* ligger inde i intervallet?

- A: 1%
- B: 2%
- C: 4%
- D: Den kender vi ikke
- E: Ved ikke

Svar B: Der er 2% for at vi ikke 'fanger' den rigtige middelværdi i 98% konfidensintervallet

# Spørgsmål om konfidensinterval ([socrative.com](https://socrative.com), room: PBAC)

Når vi så har udført eksperimentet og har stikprøven, ved vi da om middelværdien er indeholdt i det konfidensinterval vi har beregnet?

- A: Ja
- B: Nej
- C: Ved ikke

Svar B: Nej, vi ved ikke om vi har fanget den rigtige middelværdi, vi kender kun sandsynligheden for at fange den

# Praktisk problem!!

*Populationsspredningen  $\sigma$  indgår i formelen og den kender vi ikke!!*

Oplagt løsning:

Anvend estimatet  $S$  af  $\sigma$  i stedet for!

MEN MEN:

Så bryder den givne teori faktisk sammen!!

HELDIGVIS:

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

## Theorem 3.4: More applicable extension of the same stuff: (kopi af Theorem 2.49)

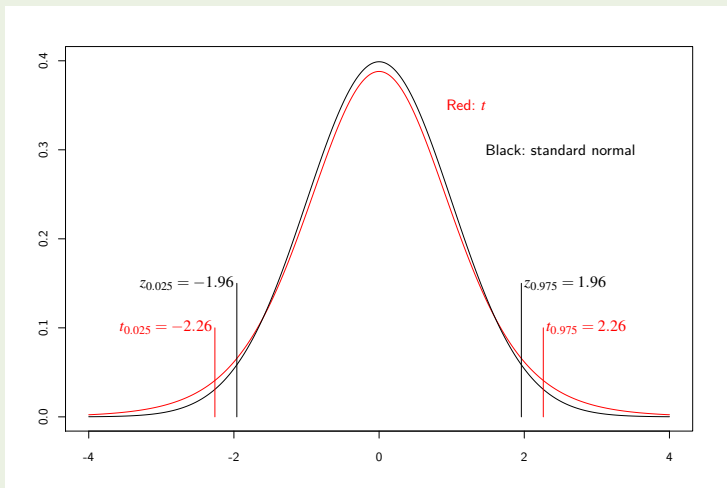
*t*-fordelingen tager højde for usikkerheden i at bruge *s*:

Assume that  $X_1, \dots, X_n$  are independent and identically normally distributed random variables, where  $X_i \sim N(\mu, \sigma^2)$  and  $i = 1, \dots, n$ , then

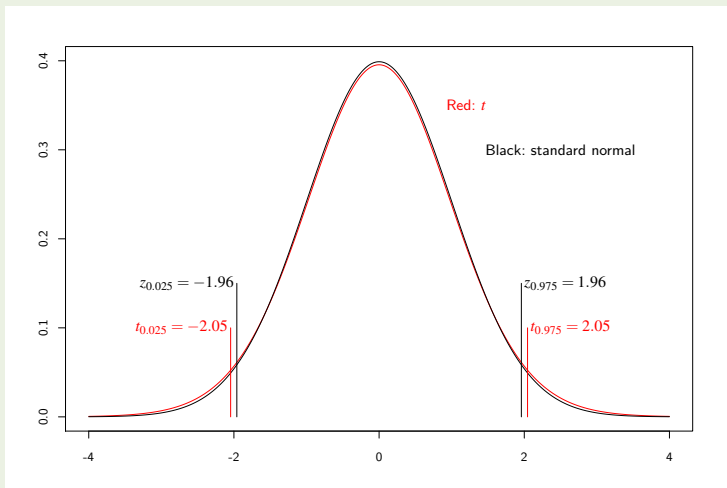
$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t$$

where  $t$  is the  $t$ -distribution with  $n - 1$  degrees of freedom.

# t-fordelingen med 9 frihedsgrader ( $n = 10$ ) og standardnormalfordelingen



# t-fordelingen med 29 frihedsgrader ( $n = 30$ ) og standardnormalfordelingen





## Metodeboks 3.8: One-sample konfidensinterval for $\mu$

Brug den rigtige  $t$ -fordeling til at lave konfidensintervallet:

For a sample  $x_1, \dots, x_n$  the  $100(1 - \alpha)\%$  confidence interval is given by:

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

where  $t_{1-\alpha/2}$  is the  $100(1 - \alpha)\%$  quantile from the  $t$ -distribution with  $n - 1$  degrees of freedom.

Mest almindeligt med  $\alpha = 0.05$ :

The most commonly used is the 95%-confidence interval:

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

## Eksempel - Højde af 10 studerende

Stikprøve,  $n = 10$ :

168 161 167 179 184 166 198 187 191 179

Sample mean og standard deviation:

$$\bar{x} = 178$$

$$s = 12.21$$

Estimer population mean og standard deviation:

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

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

## Højde-eksempel, 95% konfidensinterval (CI)

```
## 97.5% fraktilen af t-fordelingen for n=10:  
qt(p=0.975, df=9)
```

```
## [1] 2.26
```

Indsat i formlen

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

giver det

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

## Højde-eksempel, 99% Konfidensinterval (CI)

```
## 99.5% fraktilen af t-fordelingen for n=10:  
qt(p=0.995, df=9)
```

```
## [1] 3.25
```

Indsat i formlen

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

giver det

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

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

```
## Angiv data
x <- c(168,161,167,179,184,166,198,187,191,179)
## Beregn 99% konfidensinterval
t.test(x, conf.level=0.99)

##
## One Sample t-test
##
## data: x
## t = 50, df = 9, p-value = 5e-12
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
## 165 191
## sample estimates:
## mean of x
## 178
```

# Indeklimaundersøgelse

Specialestuderende er igang med en indeklimaundersøgelse, så tag jeres computer frem og gå til:

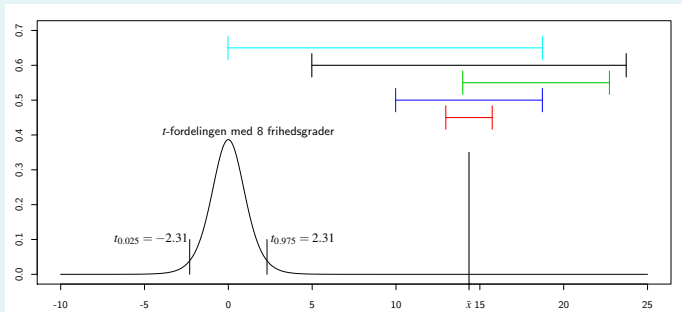
`byg-cweb2.win.dtu.dk/dtu_auditorier/`

og udfyld spørgeskemaet. TAK!

PAUSE

# Svar via socrative.com eller Socrative app. Room: PBAC

- Gennemsnit  $\bar{x} = 14.4$ , stikprøvespredningen  $s = 6$ , antal obs. er  $n = 9$
- Formlen for konfidensintervallet er  $\bar{x} \pm t_{0.975} \frac{s}{\sqrt{n}}$



Hvilket af intervallerne er det rigtige 95% konfidensinterval?

A: Turkise    B: Sorte    C: Grønne    D: Blå    E: Røde

Svar D: Blå. Fordi  $t_{0.975} \frac{s}{\sqrt{n}} = 2.31 \frac{6}{\sqrt{9}} \approx 4.6$  så nedre grænse omkring 10.



# Den formelle ramme for *statistisk inferens*

## Fra eNote, Chapter 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:

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

# Den formelle ramme for *statistisk inferens* - Eksempel

Fra eNote, Chapter 1, højdeeksempel

Vi måler højden for 10 tilfældige personer i Danmark

Stikprøven/The sample:

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

Populationen:

Højderne for alle mennesker i Danmark.

Observationsenheden:

En 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

Ved at sikre at stikprøven er fuldstændig tilfældig udtaget

# Tilfældig stikprøveudtagning

## Definition 3.11:

- A random sample from an (infinite) population: A set of observations  $X_1, X_2, \dots, X_n$  constitutes a random sample of size  $n$  from the infinite population  $f(x)$  if:
  - ① Each  $X_i$  is a random variable whose distribution is given by  $f(x)$
  - ② These  $n$  random variables are independent

## 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)

## Theorem 3.13: The Central Limit Theorem

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

Let  $\bar{X}$  be the mean of a random sample of size  $n$  taken from a population with mean  $\mu$  and variance  $\sigma^2$ , then

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

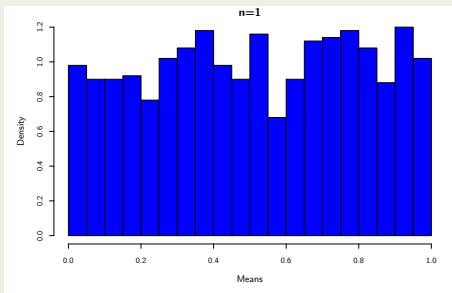
is a random variable whose distribution function approaches that of the standard normal distribution,  $N(0, 1^2)$ , as  $n \rightarrow \infty$

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

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1^2) \text{ og } \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t \text{ ved } t\text{-fordelingen med } n - 1 \text{ frihedsgrader}$$

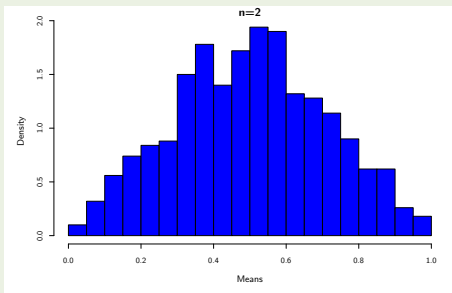
# CLT in action - gennemsnit af Uniform fordelte observationer

```
## Stikprøvestørrelse
n=1
## Antal gentagelser
k=1000
## Simuler
u=matrix(runif(k*n),ncol=n)
## Se empirisk tæthed
hist(apply(u,1,mean), col='blue', main='n=1', xlab='Means', nclass=15, prob=TRUE, x
```



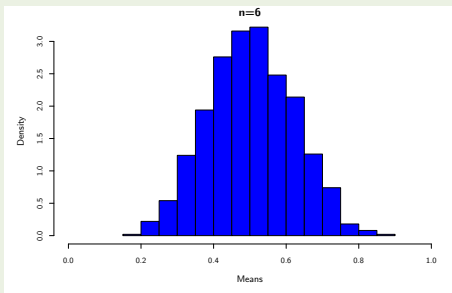
# CLT in action - gennemsnit af Uniform fordelte observationer

```
## Stikprøvestørrelse  
n=2  
## Antal gentagelser  
k=1000  
## Simuler  
u=matrix(runif(k*n),ncol=n)  
## Se empirisk tæthed  
hist(apply(u,1,mean), col='blue', main='n=2', xlab='Means', nclass=15, prob=TRUE, x
```



# CLT in action - gennemsnit af Uniform fordelte observationer

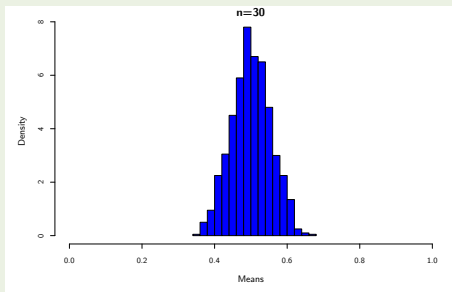
```
## Stikprøvestørrelse
n=6
## Antal gentagelser
k=1000
## Simuler
u=matrix(runif(k*n),ncol=n)
## Se empirisk tæthed
hist(apply(u,1,mean), col='blue', main='n=6', xlab='Means', nclass=15, prob=TRUE, x
```





# CLT in action - gennemsnit af Uniform fordelte observationer

```
## Stikprøvestørrelse
n=30
## Antal gentagelser
k=1000
## Simuler
u=matrix(runif(k*n),ncol=n)
## Se empirisk tæthed
hist(apply(u,1,mean), col='blue', main='n=30', xlab='Means', nclass=15, prob=TRUE, ...)
```



# 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"

Hvad er "stor nok"?

Faktisk svært at svare præcist på, MEN:

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

# Svar via socrative.com eller Socrative app. Room: PBAC

Er lydniveauet behageligt?

- A: Fino
- B: Nope, skru' op
- C: Nope, skru' ned
- D: Nope, der er dårlig og ubehagelig lyd herinde med det lydanlægget

## Svar via socrative.com eller Socrative app. Room: PBAC

Bør Peder klæde sig mere nydeligt?

- A: Ja, for den da! Det er grimt det tøj
- B: Nej, han ser faktisk rigtig checket ud
- C: Nej, det kan være lige meget med tøjet, han skal barbere sig og rede sit hår først
- D: Ved ikke, jeg har simpelthen været for optaget af statistikken til at lægge mærke til hans påklædning

# Stikprøvefordelingen for varians-estimatet (Theorem 2.56)

Variansestimater opfører sig som en  $\chi^2$ -fordeling:

Let

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

then:

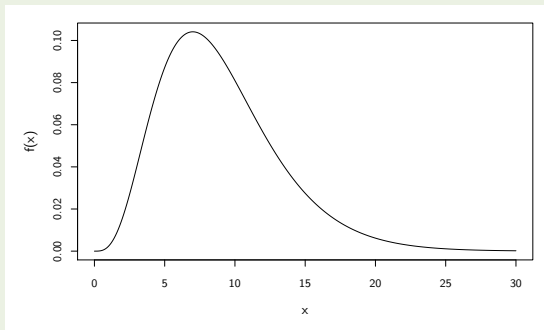
$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

is a random variable following the  $\chi^2$ -distribution with  $\nu = n - 1$  degrees of freedom.

# $\chi^2$ -fordelingen med $\nu = 9$ frihedsgrader

```
## Plot chi^2 tæthedsfunktion med 9 frihedsgrader

## En sekvens af x værdier
x <- seq(0, 30, by = 0.1)
## Plot chi^2 tæthedsfunktion
plot(x, dchisq(x, df = 9), type = 'l', ylab="f(x)")
```



## Metode 3.18: Konfidensinterval for stikprøvevariens og stikprøvespredning

### Variansen:

A  $100(1 - \alpha)\%$  confidence interval for the variance  $\sigma^2$  is:

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

where the quantiles come from a  $\chi^2$ -distribution with  $\nu = n - 1$  degrees of freedom.

### Spredningen:

A  $100(1 - \alpha)\%$  confidence interval for the sample standard deviation  $\hat{\sigma}$  is:

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

## Eksempel

### Produktion af tabletter

Vi producerer pulverblanding og tabletter deraf, så 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.

### 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  $\chi^2$ -fraktilerne:

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



## Eksempel

Så konfidensintervallet for variansen  $\sigma^2$  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  $\sigma$  bliver:

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

## Højdeeksempel

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

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

```
## 2.5% og 97.5% fraktilerne i chi^2 fordelingen for n=10
qchisq(c(0.025, 0.975), df = 9)
```

```
## [1] 2.7 19.0
```

Så konfidensintervallet for højdespredningen  $\sigma$  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

Sample mean og standard deviation:

$$\bar{x} = 178$$

$$s = 12.21$$

Estimer population mean og standard deviation:

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

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

NYT: **Konfidensinterval,  $\mu$ :**

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

NYT: **Konfidensinterval,  $\sigma$ :**

$$[8.4; 22.3]$$

# Svar via socrative.com eller Socrative app. Room: PBAC

Hvilket af følgende udsagn er korrekt?

- A: Statistik er virkelig skod, jeg tror ikke det kan bruges til noget
- B: Statistik er altså øv, man skal bare sidde og sætte en masse tal ind i nogle dumme formler
- C: Jeg burde ligge under min dyne og blive frisk til at feste igennem i aften
- D: Statistik er virkelig fedt, det er fascinerende at man ikke bare kan regne et estimat ud, men man kan også regne ud hvor præcist det estimat er

Svar D

# Overview

- 1 Fordelingen for gennemsnittet
  - $t$ -fordelingen
- 2 Konfidensintervallet for  $\mu$ 
  - Eksempel
- 3 Den statistiske sprogbrug og formelle ramme
- 4 Ikke-normale data, Central Grænseværdisætning (CLT)
- 5 Konfidensinterval for varians og spredning