

Kursus 02403

Introduktion til Statistik

## Forelæsning 5: Kapitel 7: Inferens for gennemsnit (One-sample setup)

Jan Kloppenborg Møller

DTU Informatik

Bygning 305 - rum 210

Danmarks Tekniske Universitet

2800 Lyngby – Danmark

e-mail: [jkm@imm.dtu.dk](mailto:jkm@imm.dtu.dk)

# Oversigt, Inferens for gennemsnit (One-sample setup)

- 1 Intro og generelle begreber
- 2 Maksimal fejl på et estimat
- 3 Bestemmelse af stikprøvestørrelse
- 4 Konfidensinterval
- 5 Udvidelse til "ukendt" varians-setup
- 6 Kendt eller ukendt varians
- 7 R (R note 6)

# Oversigt

- 1 Intro og generelle begreber
  - Populationer og stikprøver
  - Inferens og stikrøvefordeling
  - Central Grænseværdisætning
- 2 Maksimal fejl på et estimat
- 3 Bestemmelse af stikprøvestørrelse
- 4 Konfidensinterval
- 5 Udvidelse til "ukendt" varians-setup
- 6 Kendt eller ukendt varians

## Motiverende eksempel

En bilproducent er interesseret i at fastlægge benzinforbruget ved kørsel på motorvej for et nyt bilmærke. Fra et pilotstudie vides, at variansen af forbruget,  $\sigma^2$ , er  $0.2^2$  (l/100 km) $^2$ . Der udføres nu 25 forsøgsture (samme fører), hvor hver bil kører 100 km, og fra disse målinger estimeres middelværdien.

*Hvad bliver  $\text{Var}[\bar{X}]$ ?*

*Beregn maksimal fejl med 95% sandsynlighed, dvs.  $E_{0.95}$*

*Beregn maksimal fejl med 99% sandsynlighed, dvs.  $E_{0.99}$*

# Kapitel 7: Inferens for gennemsnit

- Estimation (7.1-7.2)
  - Maksimal fejl på estimat af middelværdi
  - Intervalestimation
- Kapitel 6: Stikprøvefordelinger
  - Definition af population og tilfældig stikprøve
  - Stikprøvefordeling for middelværdien fra en normalfordeling
  - $t$ -fordelingen

# Populationer og stikprøver

- Ved hjælp af en stikprøve forsøger man at generalisere om en population
- Det er derfor vigtigt, at stikprøven er repræsentativ for populationen

# Definition af population og tilfældig stikprøve (Kap. 6.1)

- Tilfældig stikprøve fra en **endelig** population:

Observationerne  $X_1, X_2, \dots, X_n$  er en tilfældig stikprøve af størrelse  $n$  fra en endelig population af størrelse  $N$ , såfremt værdierne er valgt således, at enhver delmængde af størrelse  $n$  af de  $N$  elementer fra populationen har den samme sandsynlighed for at blive valgt.

# Definition af tilfældig stikprøve (Kap. 6.1)

- Tilfældig stikprøve fra en **uendelig** population:

Et sæt observationer  $X_1, X_2, \dots, X_n$  er en tilfældig stikprøve af størrelse  $n$  fra en uendelig population  $f(x)$  såfremt:

- 1 Hvert  $X_i$  er en stokastisk variabel med tæthedsfunktion  $f(x)$
- 2 De  $n$  stokastiske variable er uafhængige

# Inferens for middelværdier

- Parameter: middelværdi af population  $\mu$
- Data: en tilfældig stikprøve  $x_1, x_2, \dots, x_n$
- Estimator:  $\bar{X}$
- Estimat af varians på middelværdi: (fra Kap. 6.2)

$$\text{Var}[\bar{X}] = \frac{\sigma^2}{n}$$

# Stikrøvefordeling for middelværdien når variansen er kendt (Theorem 6.1, side 180)

- Uendelig population:

Lad  $\bar{X}$  være middelværdien af en stikprøve af størrelse  $n$  fra en fordeling med middelværdi  $\mu$  og varians  $\sigma^2$

Da er  $\bar{X}$  en stokastisk variabel og følger en fordeling med middelværdi  $\mu$  og varians  $\frac{\sigma^2}{n}$

# Stikprøvefordeling for middelværdien når variansen er kendt (Theorem 6.1, side 180)

- Endelig population:

Lad  $\bar{X}$  være middelværdien af en stikprøve af størrelse  $n$  fra en fordeling med middelværdi  $\mu$  og varians  $\sigma^2$

Da er  $\bar{X}$  en stokastisk variabel og følger en fordeling med middelværdi  $\mu$  og varians  $\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$

# Begreber

## Central Estimator:

En estimator  $\hat{\theta}$  er central (eller ikke-biased), hvis og kun hvis, middelværdien af stikprøvefordelingen for estimatoren er lig  $\theta$

Efficient Estimator En estimator  $\hat{\theta}_1$  er en mere efficient estimator af  $\theta$  end estimatoren  $\hat{\theta}_2$  hvis:

- 1  $\hat{\theta}_1$  og  $\hat{\theta}_2$  begge er centrale estimatorer af  $\theta$
- 2 Variansen af stikprøvefordelingen for  $\hat{\theta}_1$  er mindre end for  $\hat{\theta}_2$

# Den centrale grænseværdisætning (Theorem 6.3, side 184)

Lad  $\bar{X}$  være middelværdien af en stikprøve fra en fordeling med middelværdi  $\mu$  og varians  $\sigma^2$ . Da vil

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

følge en  $N(0, 1^2)$  fordeling for  $n \rightarrow \infty$

HVIS data i sig selv følger en normalfordeling, så er stikprøvefordelingen også normal:

$$X_i \sim N(\mu, \sigma^2) \Rightarrow Z \sim N(0, 1).$$

# Oversigt

- 1 Intro og generelle begreber
- 2 Maksimal fejl på et estimat
  - Eksempel 1
- 3 Bestemmelse af stikprøvestørrelse
- 4 Konfidensinterval
- 5 Udvidelse til "ukendt" varians-setup
- 6 Kendt eller ukendt varians
- 7 R (R note 6)

# Maksimal fejl på et estimat

For store  $n$  gælder (den centrale grænseværdi sætning):

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

Den maksimale fejl,  $E$ , på et estimat med sandsynlighed  $(1 - \alpha)$  bliver:

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

hvor  $z_{\alpha/2}$  findes i standard normalfordelingen (tabel 3)

## Eksempel 1

En bilproducent er interesseret i at fastlægge benzinforbruget ved kørsel på motorvej for et nyt bilmærke. Fra et pilotstudie vides, at variansen af forbruget,  $\sigma^2$ , er  $0.2^2 \text{ (l/100 km)}^2$ . Der udføres nu 25 forsøgsture (samme fører), hvor hver bil kører 100 km, og fra disse målinger estimeres middelværdien.

*Hvad bliver  $\text{Var}[\bar{X}]$ ?*

*Beregn maksimal fejl med 95% sandsynlighed, dvs.  $E_{0.95}$*

*Beregn maksimal fejl med 99% sandsynlighed, dvs.  $E_{0.99}$*

# Oversigt

- 1 Intro og generelle begreber
- 2 Maksimal fejl på et estimat
- 3 Bestemmelse af stikprøvestørrelse
  - Eksempel 2
- 4 Konfidensinterval
- 5 Udvidelse til "ukendt" varians-setup
- 6 Kendt eller ukendt varians
- 7 R (R note 6)

# Bestemmelse af stikprøvestørrelse

- Den maksimale fejl med sandsynlighed (eller konfidens)  $1 - \alpha$  er:

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

- Når  $\sigma$  er kendt og vi ønsker at bestemme stikprøvestørrelse  $n$  for en maksimal fejl med sandsynlighed  $1 - \alpha$  fås:

$$n = \left( \frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2$$

## Eksempel 2

Bilproducenten beslutter sig for at afprøve nogle justeringer, der antages at have indflydelse på benzinforbruget. Baseret på pilotstudiet antages indledningsvist, at variansen af forbruget,  $\sigma^2$ , er  $0.2^2$  (l/100 km)<sup>2</sup>. Der køres igen ture á 100 km.

*Bestem den nødvendige stikprøvestørrelse (antal ture), såfremt man ønsker, at den maksimale fejl med 95% sandsynlighed højst er  $E_{95} = 0.10$  (l/100 km)?*

# Oversigt

- 1 Intro og generelle begreber
- 2 Maksimal fejl på et estimat
- 3 Bestemmelse af stikprøvestørrelse
- 4 Konfidensinterval**
- 5 Udvidelse til "ukendt" varians-setup
- 6 Kendt eller ukendt varians
- 7 R (R note 6)

# Intervalestimation

Det er ofte mere informativt at angive intervalestimer end blot en enkelt værdi, f.eks. kan man være interesseret i et interval, der dækker  $(1 - \alpha)\%$ .

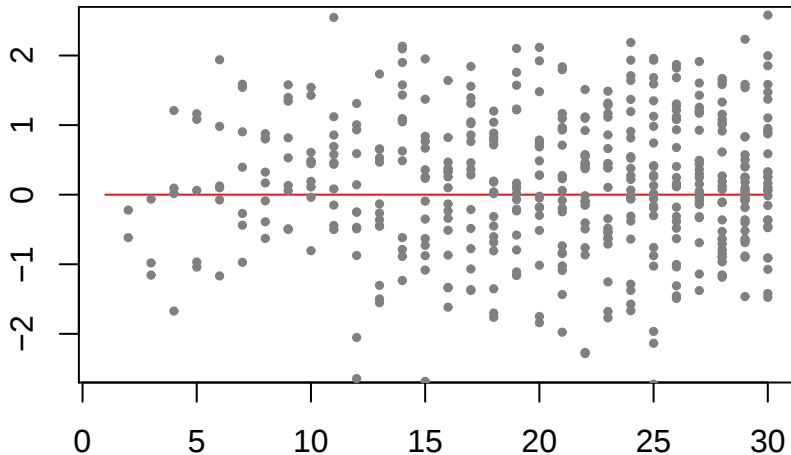
Idet vi kan skrive for kendt  $\sigma$ :

$$-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}$$

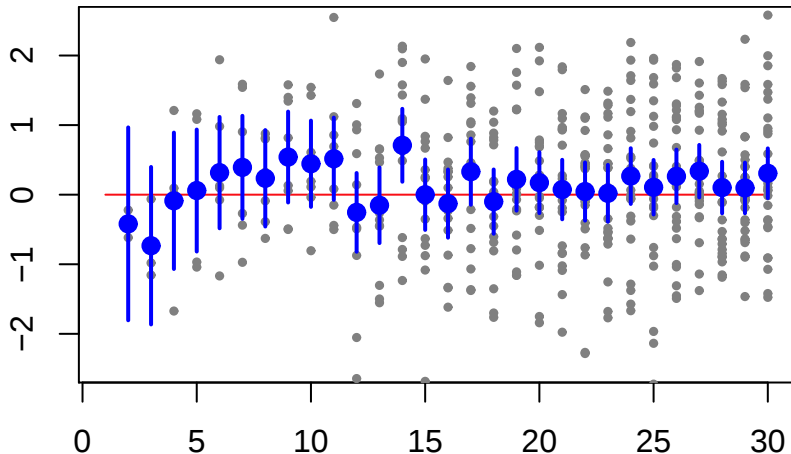
Ved omskrivning fås  $(1 - \alpha)$  konfidensintervallet:

$$\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

# Konfidensinterval 1



## Konfidensinterval 2



# Oversigt

- 1 Intro og generelle begreber
- 2 Maksimal fejl på et estimat
- 3 Bestemmelse af stikprøvestørrelse
- 4 Konfidensinterval
- 5 Udvidelse til "ukendt" varians-setup
  - $t$ -fordeling
  - Maximal fejl
  - Eksempel 3
  - Konfidensinterval
  - Eksempel 4

# Stikprøvefordeling for middelværdien når variansen ikke er kendt

Lad  $\bar{X}$  være middelværdien af en stikprøve af størrelse  $n$  fra en fordeling med middelværdi  $\mu$  og varians  $\sigma^2$  og hvor stikprøvens varians er estimeret:

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

# Stikprøvefordeling for middelværdien når variansen ikke er kendt

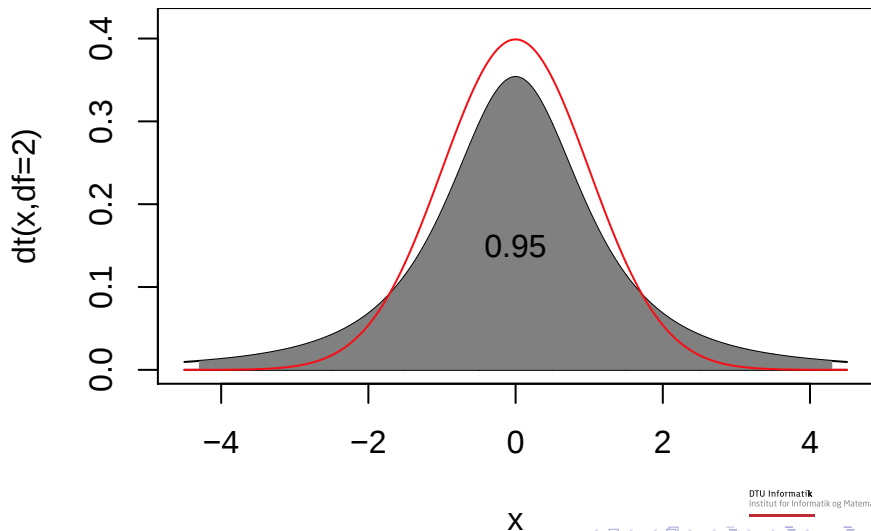
$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Da er

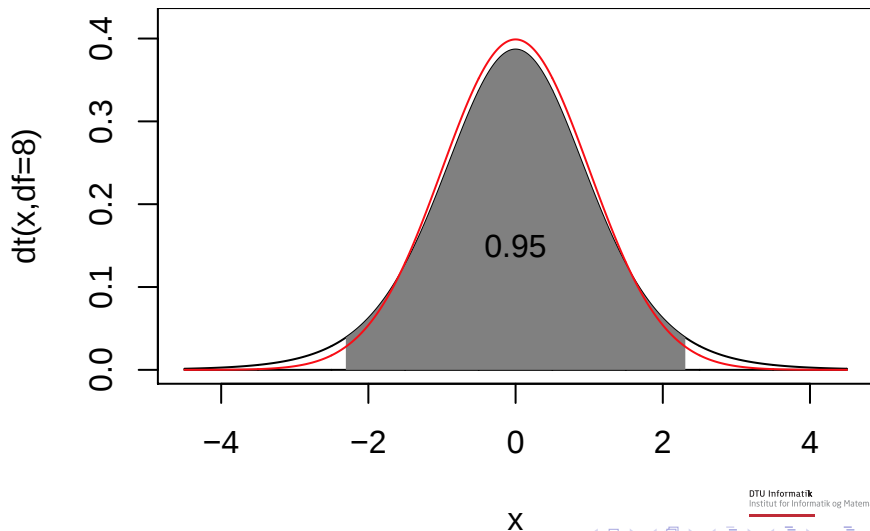
$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

en stokastisk variabel og følger en  $t$ -fordeling med parameter  $v = n - 1$

# t-fordelingen



# t-fordelingen



# Tabelopslag i $t$ -fordelingen

- Tabelopslag i  $t$ -fordelingen gøres vha tabel 4
- Ved  $t_\alpha(n-1)$  forstås den værdi, således at

$$P(t \geq t_\alpha) = \alpha$$

- R : `pt(x, df)`, `qt( $\alpha$ , df)`

# Maksimal fejl på et estimat ( $\sigma$ ukendt)

Når  $\sigma$  er ukendt, anvendes estimatet  $s$  for  $\sigma$ . I stedet for  $z_{\alpha/2}$  anvendes  $t_{\alpha/2}$  og den maksimale fejl,  $E$  med  $(1 - \alpha)$  konfidens, bliver

$$E = t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

hvor  $t_{\alpha/2} = t(n - 1)_{\alpha/2}$   
findes i t-fordelingen (tabel 4)

## Eksempel 3

Med de nye justering udføres 16 forsøg. Fra disse estimeres variansen  $s^2 = 0.25^2$ .

*Bestem den maksimale fejl,  $E_{0.95}$ , dvs. med 95% sandsynlighed.*

*Forklar hvorfor estimatet af  $E_{0.95}$  i dette tilfælde er større end 0.10 (l/100 km).*

# Konfidensinterval

Såfremt man ikke kender  $\sigma$  og har en lille stikprøve ( $n < 30$ ), erstattes  $\sigma$  med estimatet  $s$ , og  $z_{\alpha/2}$  erstattes med  $t_{\alpha/2}$ :

$$\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

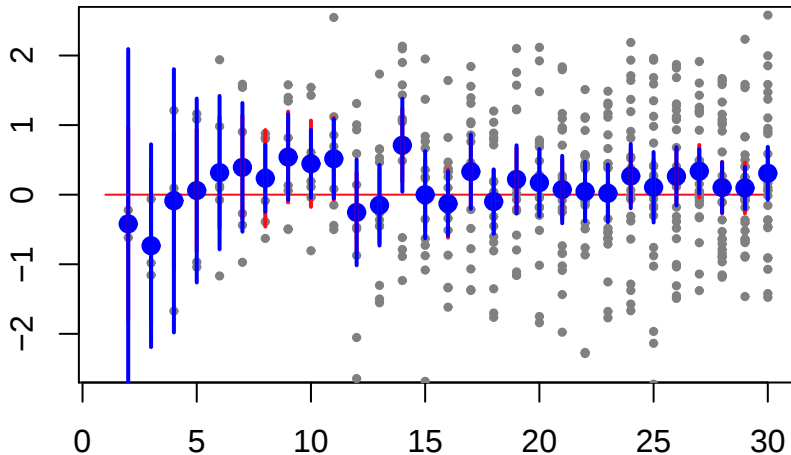
(Under antagelse af, at data er normalfordelt!)

# Konfidensinterval

Såfremt man ikke kender  $\sigma$ , men har en stor stikprøve ( $n \geq 30$ ) anvendes den samme formel, blot erstattes  $\sigma$  med estimatet  $s$ :

$$\bar{x} - z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

# Konfidensinterval $\sigma$ ukendt



## Eksempel 4

I et amerikansk studie ønskede man at sammenligne indhold af arsenik i drikkevandet ved 8 forskellige lokaliteter:

| lokalitet | vandprøve (ppm)                   |
|-----------|-----------------------------------|
| 1         | 2.2                               |
| 2         | 4.1                               |
| 3         | 2.1                               |
| 4         | 0.8                               |
| 5         | 0.1                               |
| 6         | 3.2                               |
| 7         | 2.9                               |
| 8         | 2.2                               |
|           | $\bar{x} = 2.2$ og $s_x^2 = 1.64$ |

## Eksempel 4

Det antages nu, at vandprøvemålingerne følger en normalfordeling.

*Angiv et 95% konfidensinterval for middelinholdet  $\mu$  af arsenik i drikkevandet.*

# Oversigt

- 1 Intro og generelle begreber
- 2 Maksimal fejl på et estimat
- 3 Bestemmelse af stikprøvestørrelse
- 4 Konfidensinterval
- 5 Udvidelse til "ukendt" varians-setup
- 6 Kendt eller ukendt varians**
- 7 R (R note 6)

# Generelt

I de anførte formler anvendes:

- 1 Når variansen er kendt, anvendes  $\sigma$  og  $z_{\alpha/2}$
- 2 Når variansen ikke er kendt, men stikprøven er stor, anvendes  $s$  og  $z_{\alpha/2}$
- 3 Når variansen ikke er kendt og stikprøven er lille, anvendes  $s$  og  $t_{\alpha/2}$

NB: "Ukendt" betyder at man bruger foreliggende data til at bestemme variansen  
(Og erkender begrænsningen/usikkerheden i dette)

# Oversigt

- 1 Intro og generelle begreber
- 2 Maksimal fejl på et estimat
- 3 Bestemmelse af stikprøvestørrelse
- 4 Konfidensinterval
- 5 Udvidelse til "ukendt" varians-setup
- 6 Kendt eller ukendt varians
- 7 R (R note 6)**

## R (R note 6)

| R     | Betegnelse            |
|-------|-----------------------|
| t     | t-fordelingen         |
| chisq | $\chi^2$ -fordelingen |
| f     | F-fordelingen         |

- d Tæthedsfunktion  $f(x)$  (probability distribution).
- p Fordelingsfunktion  $F(x)$  (cumulative distribution function).
- q Fraktil (quantile) i fordeling.

Eksempel:  $P(t \leq 2)$ ,  $pt(2)$

# Oversigt

- 1 Intro og generelle begreber
- 2 Maksimal fejl på et estimat
- 3 Bestemmelse af stikprøvestørrelse
- 4 Konfidensinterval
- 5 Udvidelse til "ukendt" varians-setup
- 6 Kendt eller ukendt varians
- 7 R (R note 6)

# Ugen der gik

- Nøgletal
- Grafisk fremstilling
- Stokastiske variable og deres fordelinger
- Inferens