FORELESNING 13

MET 3431

EIVIND ERIKSEN

MAR 08 2012

STATISTIKK

PLAN:

1. Estimering av populasjonsandel (7.2)
2. Estimering av gjennomsnitt (7.4)

Husk: Arbeidskrav 5 (kap 6-7)

Repetisjon: Estimering av populasjonsandel

- punkt-estimat: $\hat{p} = \frac{x}{n}$ — x: antall med pos. svar, n: utvalgsstørrelsen
- konfidensintervall: $\hat{p} \pm E$ (E feilledd)

$(\hat{p} - E, \hat{p} + E)$

$$E = \underbrace{z_{\alpha/2}} \cdot \underbrace{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

$95\% \rightarrow \alpha = 0.05$

$z_{0.025} = 1.96$

Konfidensintervall for populasjonsandelen p

Example

- La p være andelen kvinner som holder barnet med venstre arm.
- 25 av 32 kvinner på fødselsavd holdt med venstre arm. $\hat{p} = \frac{25}{32} \simeq 0.781$ $\simeq \underline{78.1\%}$
- $$E = 1.96 \cdot \sqrt{\frac{\frac{25}{32}(1 - \frac{25}{32})}{32}} = \underline{0.143}$$
- 95% konfidensintervall for andelen p:
  $\hat{p} \pm E$
  $$p = \frac{25}{32} \pm 0.143$$
- Kan også skrives $0.638 < p < 0.942$
- Vi er 95% sikre på at andelen av mødre med barnet i venstre arm er et sted mellom 63.8% og 94.2%

99% konfidensintervall. Andelen iPhone på BI

Example

Fila *klassens data alle ny*: 419 av 1937 studenter har en iPhone. Lag et 99% konfidensintervall for andelen studenter p som har iPhone på BI.

- TabellA2 : $z_{\alpha/2} = 2.576$
- $$E = 2.576 \cdot \sqrt{\frac{\frac{419}{1937}(1 - \frac{419}{1937})}{1937}} = 0.0241$$
- 99% konfidensintervall for andelen p:

$$p = \frac{419}{1937} \pm 0.0241$$

$\simeq 0.216$

$\hat{p} - E$ $\hat{p} + E$

- Kan også skrives $< 0.192, 0.240 >$
- Vi er 99% sikre på at andelen av iPhone brukere på BI er mellom 19.2% og 24.0%

90 % intervall: Andelen iPhone på BI

Example

Lag et 90% konfidensintervall for andelen studenter p som har iPhone på BI.

- TabellA2 : $z_{\alpha/2} = 1.645$
- $$E = 1.645 \cdot \sqrt{\frac{\frac{419}{1937}(1 - \frac{419}{1937})}{1937}} = 0.0154$$
- 90% konfidensintervall for andelen p:
$$p = \frac{419}{1937} \pm 0.0154$$
- Kan også skrives $< 0.201, 0.232 >$
- Vi er 90% sikre på at andelen av iPhone brukere på BI er mellom 20.1% og 23.2%

95 % intervall: Andelen iPhone på BI

Example

95% konfidensintervall for andelen studenter p som har iPhone på BI.

- TabellA2 : $z_{\alpha/2} = 1.96$ gir
  $E = 1.96 \cdot \sqrt{419/1937(1 - 419/1937)/1937} = 0.0183$
- 95% konfidensintervall er da $0.198 < p < 0.235$

Konfidensintervall for andeler i JMP

- JMP bruker en litt annen formel enn den vi bruker[a]
- *klassens data alle ny*, velg analyze>distribution og Mobiltelefon
- Rød diamant: *confidence intervals 95%* gir $0.199 < p < 0.235$ når vi runder av til 3 desimaler

[a] Agresti-Coull konfidensgrense

Bredden på intervallet

Intervallbredde versus konfidensnivå

- Jo sikrere du trenger å være på at intervallet inneholder parameteren, jo bredere blir intervallet
- Man må avveie dette slik at intervallet ikke blir for bredt
- Vanlig kompromiss er å bruke 95% konfidensnivå

Hvor stor stikkprøve trenger vi?

Størrelsen på stikkprøven

- Du har bestemt ønsket feilmargin E
- Du har bestemt konfidensnivået
- Da kan du anslå hvor stor stikkprøve du trenger
- Formelen er

$$n = \frac{z_{\alpha/2}^2 \cdot 0.25}{E^2}$$

Example

Vi ønsker et 95% konfidensintervall med feilmargin $E = 0.05$, dvs. $\pm 5\%$.

$$n = \frac{1.96^2 \cdot 0.25}{0.05^2} = 385$$

For å få en feilmargin på ca 5% bør du ha 385 objekter i stikkprøven.

Seksjon 7-4: Estimere gjennomsnittet μ

Seksjon 7-4

- Estimere gjennomsnittet μ i populasjonen
- Punktestimator er selvfølgelig $\overline{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$
- Vi vil lage konfidensintervall for μ
- Da trenger vi først *Student t-fordelingen*

Forutsetninger for å bruke t-fordeling

- Stikkprøven må være tilfeldig utvalgt
- Originaldataene x er normalfordelt, eller $n > 30$

Frihetsgrader og Student t-fordelingen

Frihetsgrader (eng: degrees of freedom)

- Stikkprøven har n objekter
- Da sier vi at den har $n-1$ *frihetsgrader*

t-fordelingen

Anta at x er normalfordelt. Med en stikkprøve av størrelse n kan vi beregne $\overline{x}$ og s. Da vil

$$\boxed{t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}}$$

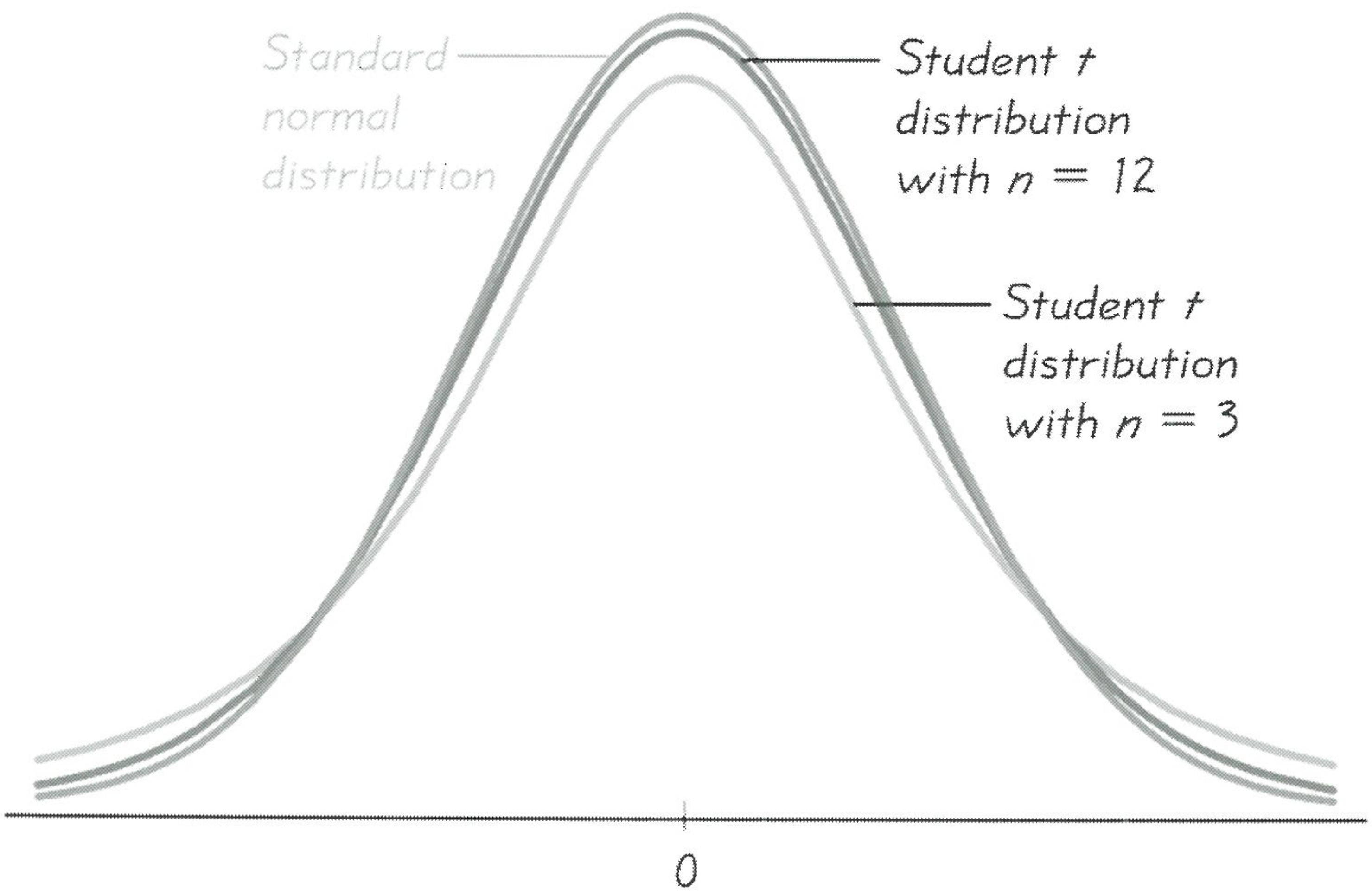
være *Student t-fordelt med* $n-1$ *frihetsgrader*

Utvalg; størrelse n:

$$\overline{x} = \frac{x_1 + \dots + x_n}{n}$$

$$s = \sqrt{\frac{(x_1-\overline{x})^2 + \dots + (x_n-\overline{x})^2}{n-1}}$$

Student t-fordelingen



Figur: t-fordelingen ift. standard normalfordeling. df=2 og 11.

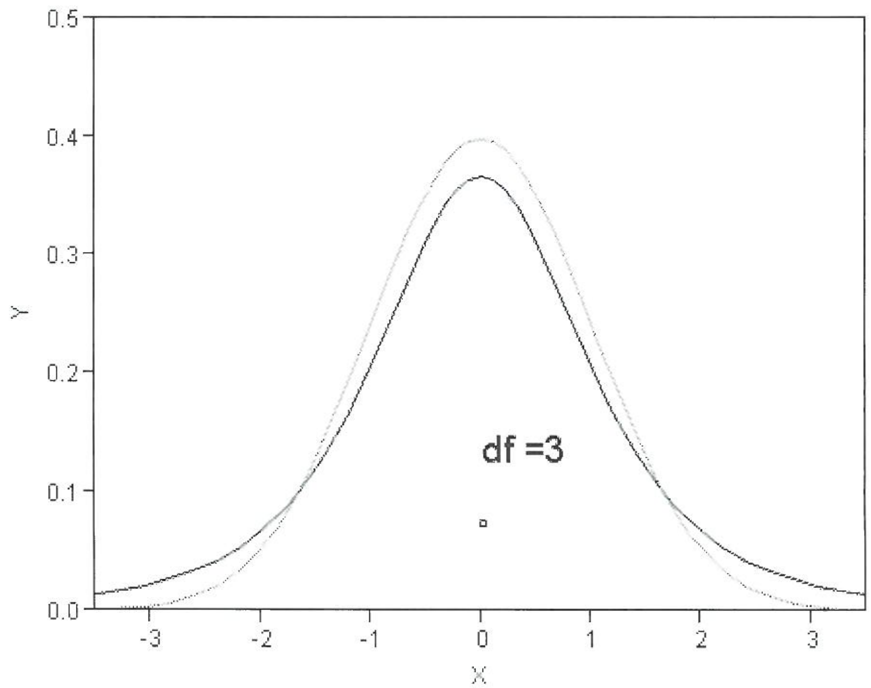
Student t-fordelingen

For hver frihetsgrad er det assosiert en t-fordeling.

- t-fordelingen er symmetrisk og ligner normalfordelingen, men har høyere standardavvik
- Når frihetsgraden vokser nærmer t-fordelingen seg en standard normalfordeling

Et JMP script

- Åpne **Normal vs t.JSL** i folderen *Sample scripts*
- *Edit > Run Script*
- Tetthetskurven til t-fordelingen for forskjellige frihetsgrader (df)
- Standard normalfordeling i rødt



Konfidensintervall for μ

$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} \longrightarrow \bar{x} \pm E$ dvs $(\bar{x} - E, \bar{x} + E)$

Feilmarginen

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

s = std avvik for utvalget

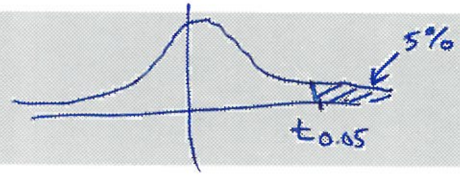
Der $t_{\alpha/2}$ har $n-1$ frihetsgrader. Finnes i tabell A-3.

Konfidensintervall for μ

$$\bar{x} - E < \mu < \bar{x} + E$$

Eksempel

$\alpha = 0.10$
$\alpha/2 = 0.05$



Example

$n =$ $\quad \bar{x} = \frac{x_1 + x_2 + \dots + x_{34}}{34}$ $\quad s = \sqrt{\frac{(x_1-\bar{x})^2 + \dots}{n-1}}$

- Du måler vekten på 34 lakrispastiller
- Gjennomsnittsvekt er $\overline{x} = 0.932$ og standardavviket er $s = 0.1$
- Lag et 90% konfidensinterval for parameteren μ

Vi har $n > 30$ så kravet er ok.

$E = t_{0.05} \cdot s/\sqrt{n}$, $t_{0.05} = 1.692$, $s/\sqrt{n} = 0.1/\sqrt{34}$

- $34 - 1 = 33$ frihetsgrader. ~~Vi runder *ned* til 32 i tabell A3~~
- Kritisk verdi $t_{\alpha/2} = \cancel{1.694}$ 1.692 for $\cancel{32}$ 33 df
- Feilmargin $E = \cancel{1.694}$ 1.692 $\cdot 0.1/\sqrt{34} = 0.029$
- Konfidensintervall:

$\bar{x} - E \qquad \bar{x} + E$

$$0.903 < \overline{x} < 0.961$$

- Vi er 90% sikre på at gjennomsnittsvekta er mellom 0.903 og 0.961 gram

Konfidensintervall for gjennomsnittet μ

Prosedyre

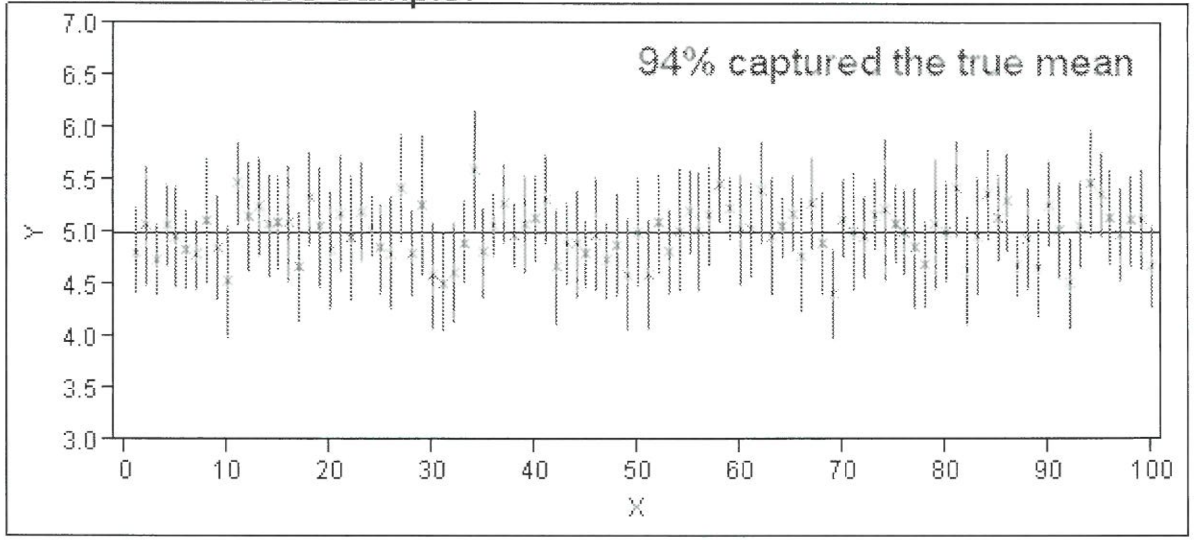
1. Sjekk at dataene er normalfordelte, eller at $n > 30$.
2. Med $n - 1$ frihetsgrader, finn kritisk verdi $t_{\alpha/2}$ i tabell A3
3. Regn ut feilmargin $E = t_{\alpha/2} \cdot s/\sqrt{n}$
4. Regn ut nedre grense $\overline{x} - E$ og øvre grense $\overline{x} + E$
5. Rund av til tre desimaler
6. Oppgi konfidensintervallet

$$\overline{x} - E < \mu < \overline{x} + E$$

Hva er et konfidensintervall?

Åpne scriptet *confidence* i folder *Sample scripts*

Press Ctrl-D to re-sample.



Click on numbers below to change their values.

Population Mean =5

Population Std Dev =1

Sample Size =20

Confidence Interval =0.95

Use Population SD? (0 for no) =0

Figur: Simulering av 100 95% konfidensintervall. 94 av dem inneholdt den sanne $\mu = 5$.

Lengde på telefonsamtaler 1

Example

- Lengden på interne telefonsamtaler i en bedrift er normalfordelt
- Stikkprøve på 5 interne samtaler: $23, 25, 12, 30, 20$ minutter.
- Lag et 95% konfidensintervall for gjennomsnittlig lengde på telefonsamtale i bedriften.

$\frac{23+25+12+30+20}{5}$ $\quad$ $\sqrt{\frac{(23-22)^2+(25-22)^2+\ldots}{4}}$

- Løsning
  1. Regn ut $\overline{x} = 22.00$ og $s = 6.671$
  2. Finn $t_{\alpha/2} = 2.776$ for 4 frihetsgrader. Tabell A3 $\quad$ $E = 2.776 \cdot \frac{6.671}{\sqrt{5}}$
  3. Feilmarginen:
  $$E = 2.776 \cdot \frac{6.671}{\sqrt{5}} = 8.282$$
  4. 95% konfidensintervall for μ går ifra $22.00 - 8.282$ til $22.00 + 8.282$:
  $$13.72 < \mu < 30.28$$
  5. Vi er 95% sikre på at gjennomsnittlig samtaletid er mellom 13.72 og 30.28 minutter