FORELESNING 12

MET 3431

STATISTIKK

EIVIND ERIKSEN

MAR 06 2012

PLAN:

(1) Normalfordeling : fordelingen til en observator (6.4–6.5)

(2) Kap. 7

Fordelingen til en observator

Observatorfordeling (eng: sampling distribution)

I seksjon 6-4 handler det om å forstå at en observator er en tilfeldig variabel. Den har derfor en sannsynlighetsfordeling. Denne kalles observatorfordelingen .

Estimator

Ofte er en observator ment å skulle *estimere* en populasjonsparameter. Da kalles observatoren for en *estimator*.

Example

Tenk deg at du tar mange stikkprøver, hver på størrelse n.

- $\overline{x}$ er en estimator for μ
- Denne estimatoren varierer fra stikkprøve til stikkprøve (eng: sampling variability)

Forventningsrette estimatorer

Table 6-7 Sampling Distributions of Statistics (for Samples of Size 2 Drawn with Replacement from the Population 1, 2, 5)

Sample	Mean $\bar{x}$	Median	Range	Variance s^2	Standard Deviation s	Proportion of Odd Numbers	Probability
1, 1	1.0	1.0	0	0.0	0.000	1	1/9
1, 2	1.5	1.5	1	0.5	0.707	0.5	1/9
1, 5	3.0	3.0	4	8.0	2.828	1	1/9
2, 1	1.5	1.5	1	0.5	0.707	0.5	1/9
2, 2	2.0	2.0	0	0.0	0.000	0	1/9
2, 5	3.5	3.5	3	4.5	2.121	0.5	1/9
5, 1	3.0	3.0	4	8.0	2.828	1	1/9
5, 2	3.5	3.5	3	4.5	2.121	0.5	1/9
5, 5	5.0	5.0	0	0.0	0.000	1	1/9
Mean of Statistic Values	8/3 =	8/3	16/9	26/9 =	1.3	2/3	
Population Parameter	8/3	2	4	26/9	1.7	2/3	
Does the sample statistic target the population parameter?	Yes	No	No	Yes	No	Yes	

Alle stikkprøver av størrelse $n = 2$

Noen estimatorer er forventningsrette

- $\bar{x}$
- s^2
- $\hat{p}$

Noen er ikke:

- Median
- Variasjonsbredde
- s

Seksjon 6-5

Prosedyrene i denne seksjone er grunnlaget for å estimere populasjonsparameter og for hypotesetesting.

Notasjon

- $\mu_{\overline{x}}$ er gjennomsnittet til $\overline{x}$
- $\sigma_{\overline{x}}$ er standardavviket til $\overline{x}$

Sentralgrenseteoremet

- Anta at vi har en stor stikkprøve av x-er fra en populasjon med gjennomsnitt μ og standardavvik σ
- Da vil gjennomsnittet $\overline{x}$ være normalfordelt med

$$\mu_{\overline{x}} = \mu, \qquad \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

Sentralgrenseteoremet

Vi antar at:

1. Den tilfeldige variabelen x har gjennomsnitt μ og standardavvik σ.
2. Man tar mange tilfeldige utvalg av størrelse n fra populasjonen

Da følger det at:

Sentralgrenseteoremet

1. Fordelingen til $\overline{x}$ blir mer og mer lik en normalfordeling når n er vokser
2. Gjennomsnittet til $\overline{x}$ er μ
3. Standardavviket til $\overline{x}$ er $\sigma/\sqrt{n}$

Sentralgrenseteorem demo i JMP

JMP Demo

1. Åpne fila *C:\Program Files \SAS \JMP \8 \Support Files English \Sample Data*
2. Velg *Rows>Add rows* og legg til 200 rader

Variabelen $N = 50$ inneholder en formel for gjennomsnittet av $n = 50$ tilfeldige variable. Når vi adderer 200 rader, så simulerer vi 200 stikkprøver på størrelse $n = 50$.

Praktiske regler + Notasjon

Når kan vi stole på sentralgrenseteoremet?

1. Når stikkprøven er større enn $n = 30$
2. Hvis x'ene selv er normalfordelte, så gjelder sentralgrenseteoremet for alle stikkprøvestørrelser.

Vekt igjen

Example

Anta at vekten på menn er normalfordelt

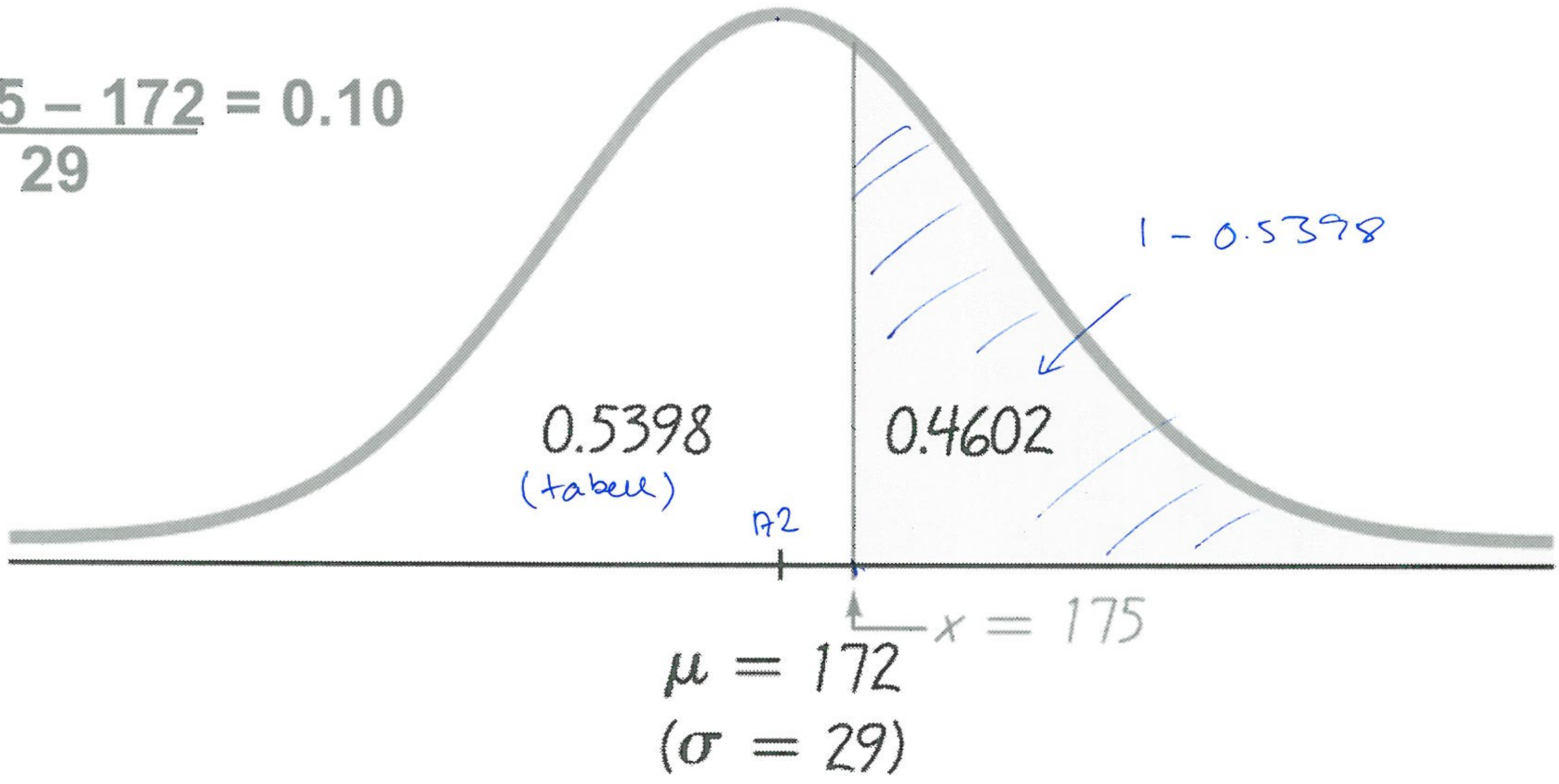
- Gjennomsnitt $\mu = 172$ pounds
- Standardavvik på $\sigma = 29$ pounds.

Finn:

1. Sannsynligheten for at tilfeldig mann veier mer enn 175 pounds?
2. 20 menn er tilfeldig utvalgt. Hva er sannsynligheten for at deres gjennomsnittsvekt er mer enn 175 pounds?

Vekt igjen, spørsmål 1

$$z = \frac{175 - 172}{29} = 0.10$$



Figur: Sannsynligheten er 0.4602 for at en mann veier mer enn 175

Vekt igjen, spørsmål 2

$$\overline{X} = \frac{x_1 + x_2 + \dots + x_{20}}{2}$$

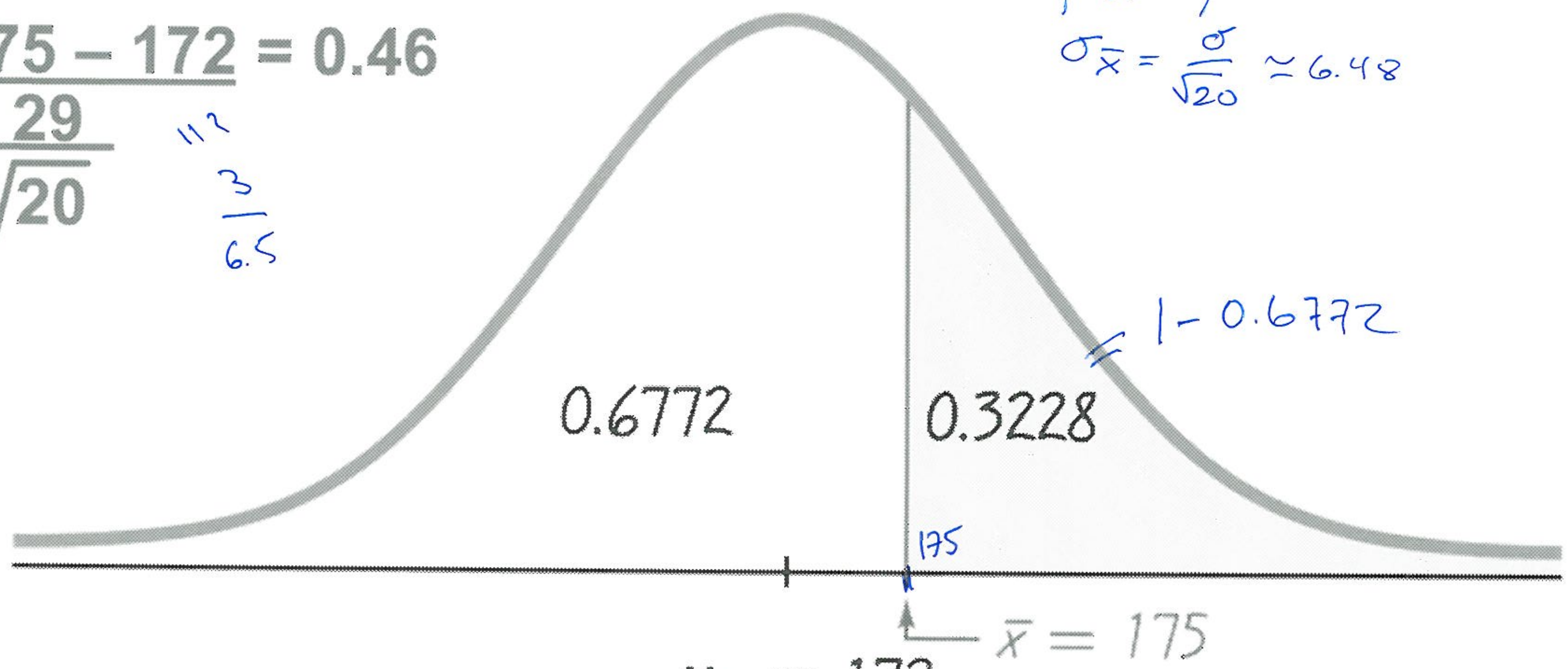
normalfordelt:

$$\mu_{\overline{X}} = \mu = 172$$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{20}} \simeq 6.48$$

$$z = \frac{175 - 172}{\frac{29}{\sqrt{20}}} = 0.46$$

$$\approx \frac{3}{6.5}$$



$$\mu_{\bar{x}} = 172$$

$$\left(\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{29}{\sqrt{20}} = 6.4845971\right)$$

Figur: Sannsynligheten er 0.3228 for at 20 mann veier mer enn 175 i snitt

Vekt igjen, Oppsummering

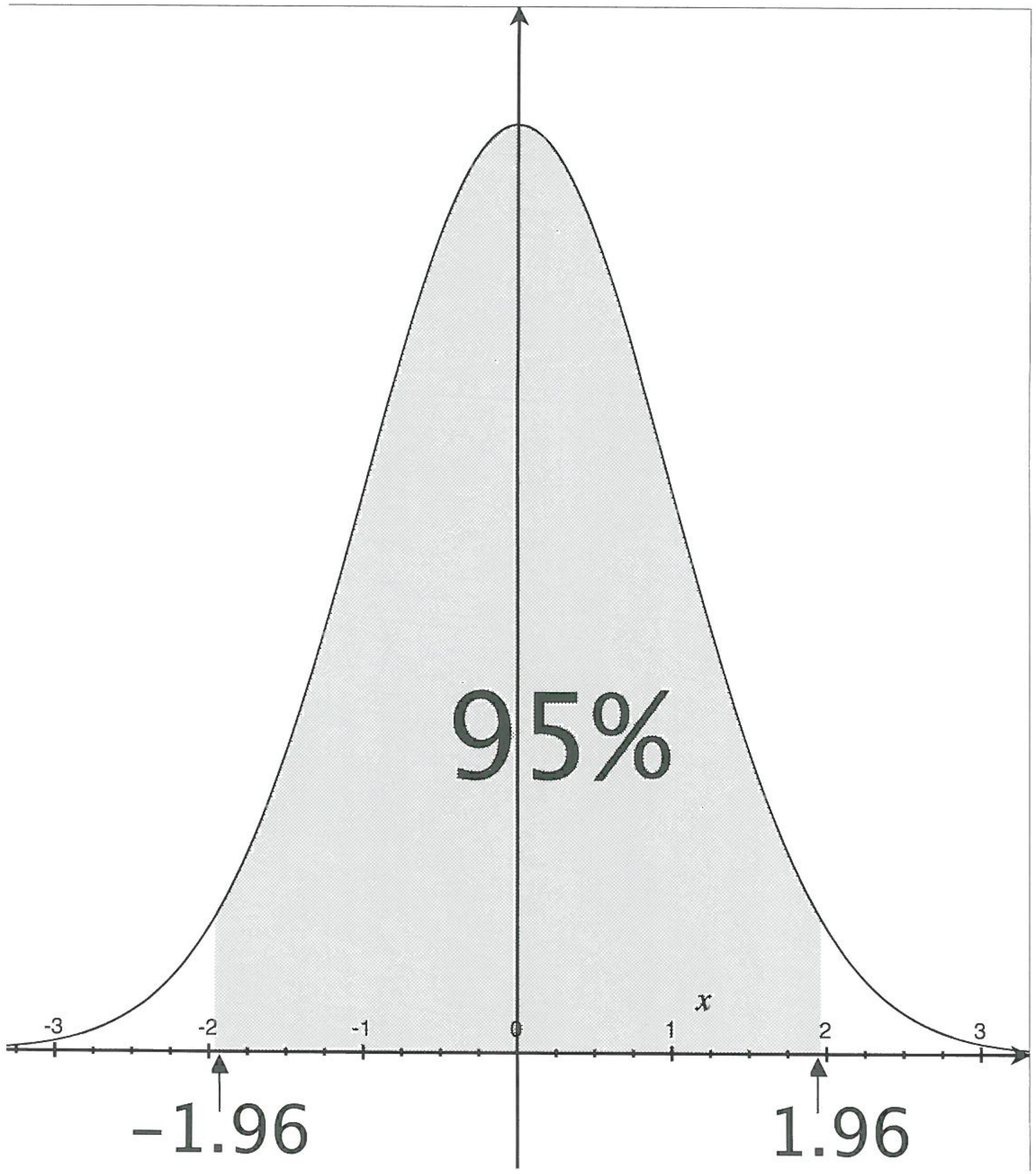
- Sannsynligheten for at 1 mann veier mer enn 175 er

$$P(x > 175) = 0.4602$$

- Sannsynligheten for at 20 mann i snitt veier mer enn 175 er

$$P(\overline{x} > 175) = 0.3228$$

- Det er lettere for en enkelt mann å veie mer enn 175 enn det er for 20 menn å veier mer enn 175 i snitt!

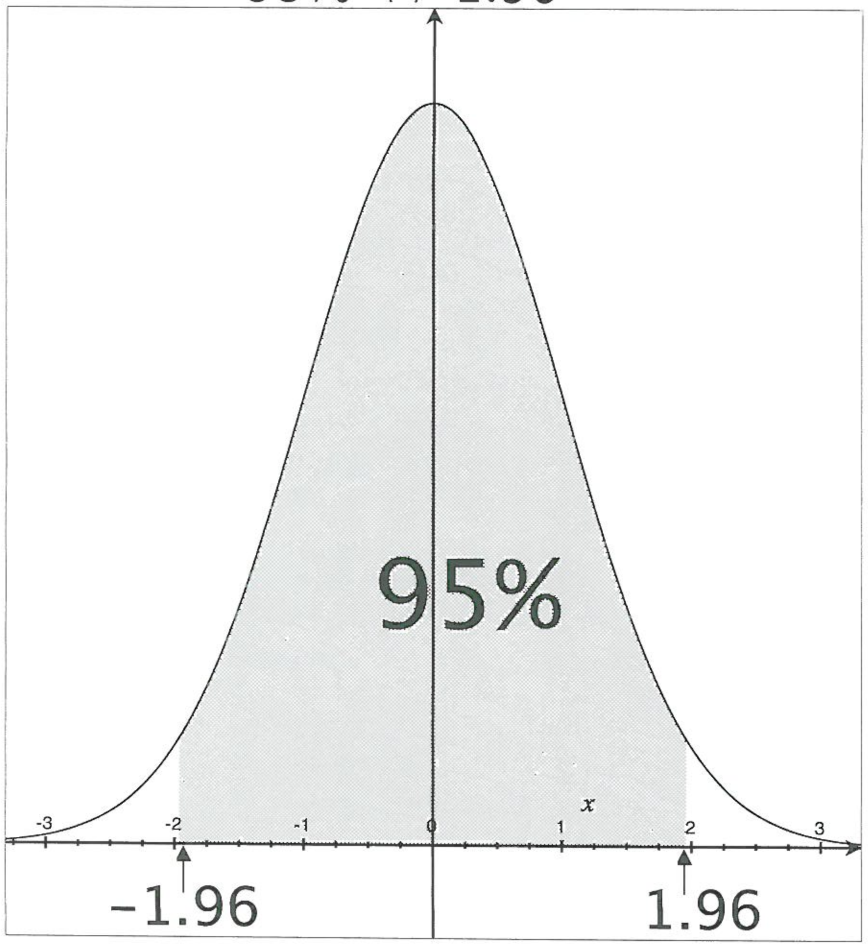


Kritiske verdier til 2.5%

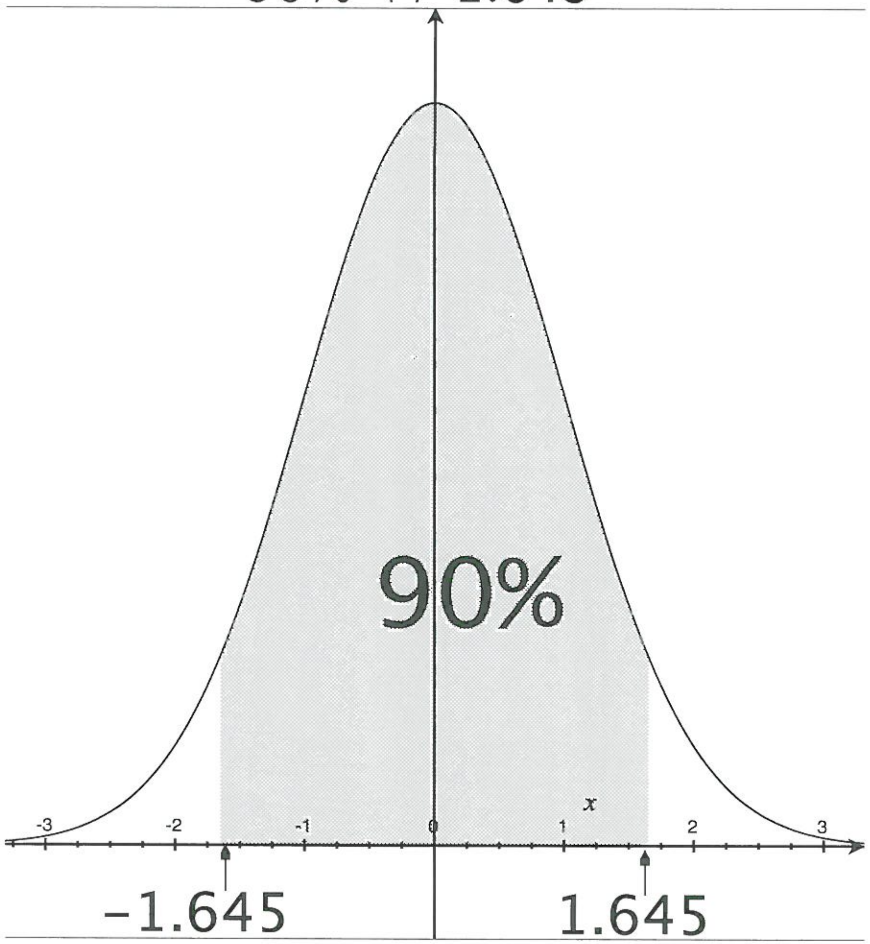
95% $\leftrightarrow$ 1.96

To viktige kritiske verdier

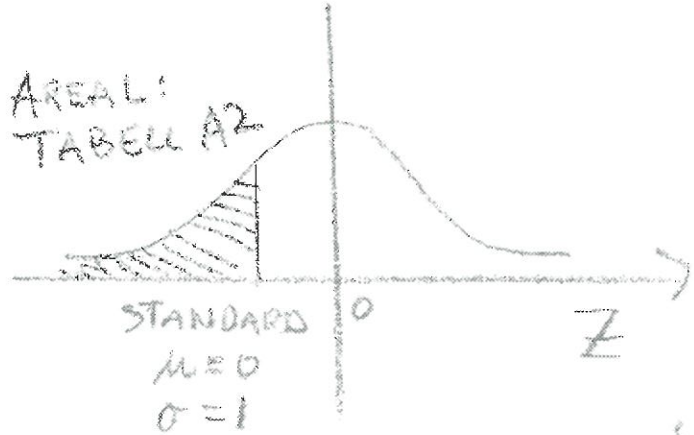
$95\% \leftrightarrow 1.96$



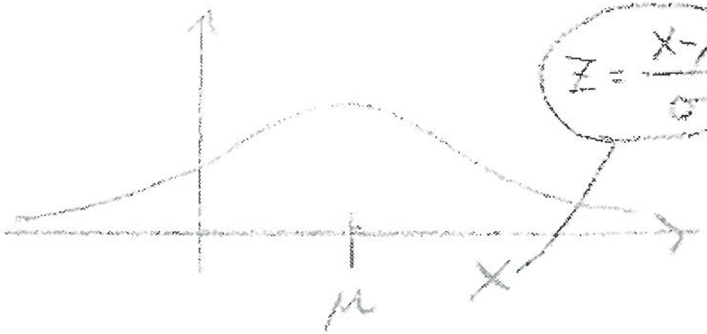
$90\% \leftrightarrow 1.645$



Oversikt Kapittel 6



GENERELL μ, σ



$$Z = \frac{X-\mu}{\sigma}$$

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

$\bar{X}$

SENTRALGRENSE TEOREM:

$E(X) = \mu$

$Var(X) = \sigma^2$

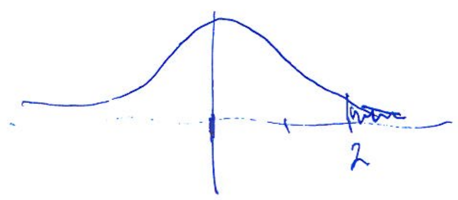
Dersom $n > 30$
eller X normalfordelt:

$E(\bar{X}) = \mu$, $Var(\bar{X}) = \sigma^2/\sqrt{n}$

Høyden til kvinnene i en populasjon er normalfordelt med gjennomsnitt 170cm og standardavvik 4cm. Hva er sannsynligheten for at en tilfeldig kvinne er over 178cm?

a) 0.02   *b*) 0.05   *c*) 0.08   *d*) 0.10   *e*) Vetkje

z-verdi: $\frac{178-170}{4} = \frac{8}{4} = 2$



1. Section 7-2: Estimere populasjonsandelen

2. Section 7-4: Estimere μ når σ er ukjent

Kapittel 7

Nå begynner vi med statistisk inferens!

Bruke stikkprøven til å

1. Estimere verdien til en parameter i populasjonen. (Kapittel 7)
2. Teste en påstand/hypotese om en parameter i populasjonen (Kapittel 8)

Hva skal vi estimere ?

- Populasjonsandelen p
- Populasjonsgjennomsnittet $\overline{x}$

Estimere populasjonsandelen

Konfidensintervall

- Vi ønsker å estimere andelen i populasjonen p
- Vi starter med andelen $\hat{p}$ i stikkprøven og lager et *konfidensintervall*.

Forutsetninger for et riktig konfidensintervall

- Stikkprøven er et tilfeldig utvalg
- Betingelsen for en binomisk forsøksrekke holder (se seksjon 5-3)
- Minst 5 suksesser og 5 fiaskoer

Notasjon

Notasjon

- Andelen i populasjonen: p
  (eng: proportion. Ikke forveksle med p i binomialfordelingen)
- Andelen i stikkprøven $\hat{p} = \frac{x}{n}$

  $\left.\begin{array}{l} n = 100 \text{ studenter} \\ x = 21 \text{ har iPhone} \end{array}\right\} \hat{p} = \frac{21}{100} = \underline{0.21}$
- x er antall suksesser i en stikkprøve med n objekter
- $\hat{q} = 1 - \hat{p}$ andelen fiaskoer

Forutsetninger for et riktig konfidensintervall for p

- Stikkprøven er et tilfeldig utvalg
- Betingelsen for en binomisk forsøksrekke holder (se seksjon 5-3)
- I stikkprøven er minst 5 med i andelen, og minst 5 er ikke med

Punktestimat

Punktestimator

En punktestimator er en enkel verdi som anslår verdien til en parameter

Punktestimator for andelen p

$\hat{p} = \frac{x}{n}$ er den beste punktestimatoren for p

Å estimere en parameter

- Du kan enten bruke et punktestimat
- Eller estimere parameteren med et intervall

Konfidensintervall

Konfidensintervall og konfidensnivå

Et *konfidensintervall* er et intervall som brukes til å estimere den sanne verdien til en populasjonsparameter.
Konfidensnivået angir hvor ofte intervallet faktisk vil inneholde den sanne populasjonsparameteren. Til hvert konfidensnivå tilhører det en α

Konfidensnivå	α
90%	$\alpha = 0.10$
95 %	$\alpha = 0.05$
99 %	$\alpha = 0.01$

Example

Et 95% konfidensintervall vil 95% av gangene inneholde parameteren

Jordbæris

Example

"Av 851 besøkende i et kjøpesenter har 51% jordbær som favorittsmak på is. "

- Sjekk at forutsetningene på side 5
- Punktestimatet for andelen er $\hat{p} = 0.51$
- 95% konfidensintervall for andelen som foretrekker jordbæris blir da

$$\overbrace{< 0.476, 0.544 >}$$

fra 51% − 3.4%
til 51% + 3.4%

- Vi er 95% sikre på at intervallet fra 0.476 til 0.544 inneholder den sanne andelen av folk som foretrekker is med jordbærsmak.
- Dette betyr at dersom vi spurte mange grupper av 829 personer, og lagde et konfidensintervall hver gang, så vil 95% av intervallene inneholde den sanne andelen p

Kritiske verdier $z_{\alpha/2}$

95% konfidensnivå

$\Downarrow$

$\alpha = 0.05$

$\Downarrow$

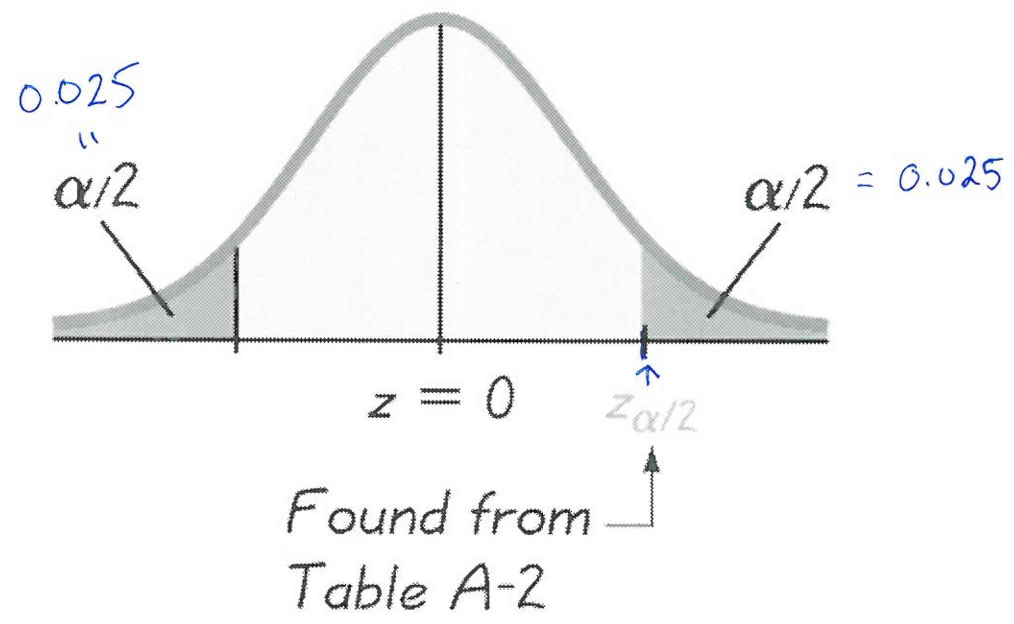
$z_{\alpha/2} = z_{0.025} = \underline{1.96}$

Prosedyren

Men hvordan fant man ut at intervallet går ifra 0.476 til 0.544?

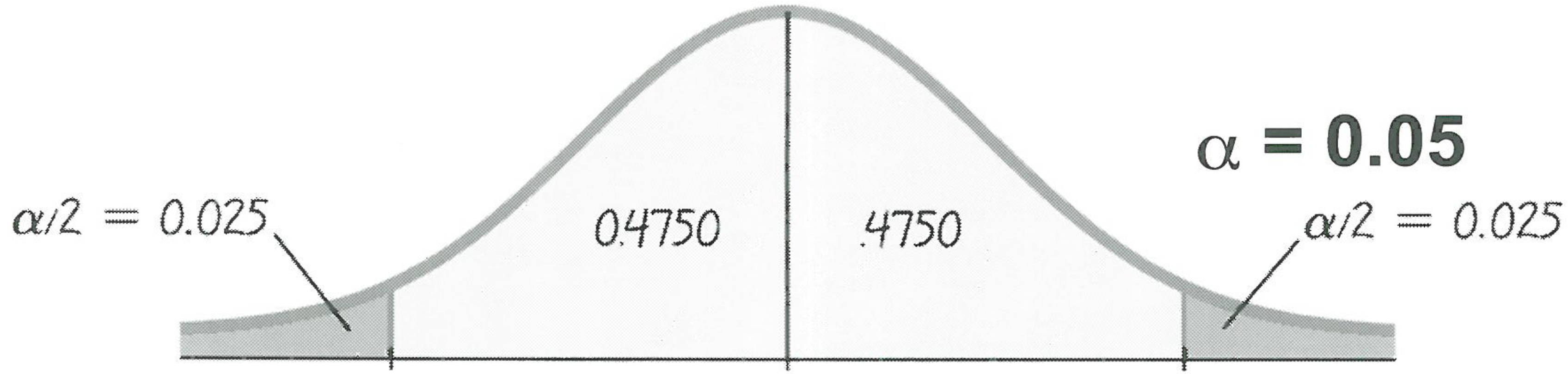
Kritiske verdier

- Sjekk krav side 5 $\rightarrow$ $\hat{p}$ normalfordelt.
- α blir da arealet i to haler
- Det gir kritiske verdier $z_{\alpha/2}$



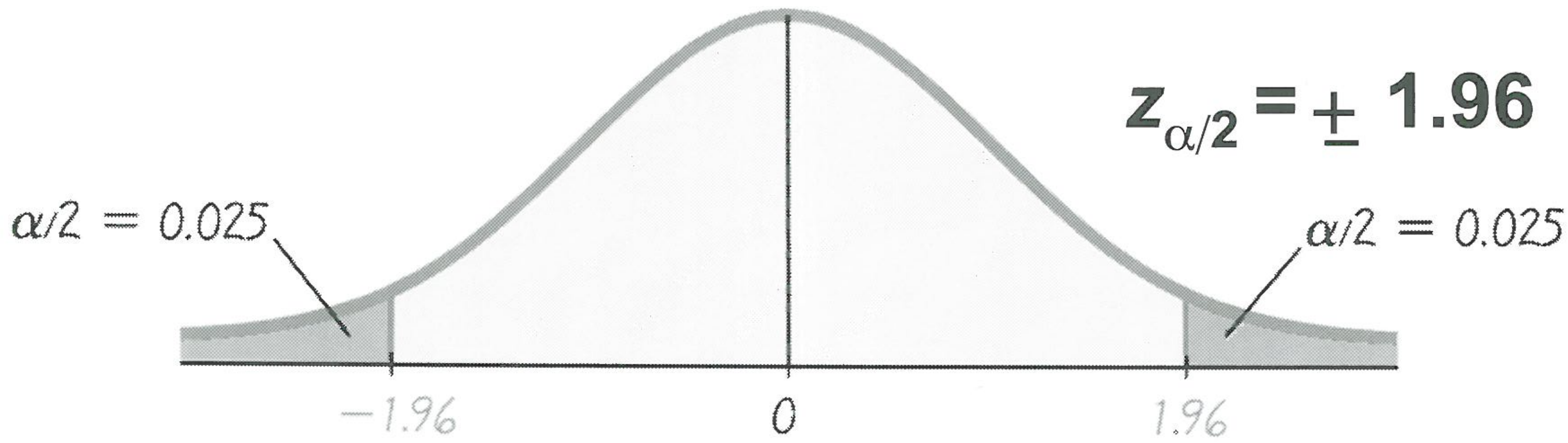
$z_{\alpha/2} = \pm 1.96$ for 95% konfidensintervall

α = 0.05



Bruk Table A-2 til å finne *z* verdien 1.96

$z_{\alpha/2} = \pm\ 1.96$



Bruke $z_{\alpha/2} = \pm 1.96$ til å finne feilmarginen

Estimatoren $\hat{p}$ er normalfordelt

Når kravene på side 5 er tilfredstilt, så er $\hat{p}$ er normalfordelt med forventning p og standardavvik $\sigma_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/n}$

Feilmarginen for konfidensintervallet for en andel

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

Example

Vi hadde $n = 821$ kunder med $\hat{p} = 0.51$, så da blir feilmarginen

$$E = 1.96 \cdot \sqrt{\frac{0.51(1-0.51)}{821}} = 0.034$$

Andel jordbæris konfidensintervall

Example

1. Vi ville ha et 95% konfidensintervall
2. Vi fant kritisk verdi $z_{\alpha/2} = 1.96$
3. Andelen i stikkprøven var $\hat{p} = 0.51$
4. Regne ut feilmarginen $E = 0.034$
5. Konfidensintervallet blir da

$$\hat{p} \pm E \leftrightarrow 0.51 \pm 0.034$$

6. Fra $0.51 - 0.034$ til $0.51 + 0.034$
7. 95% konfidensintervall: Fra 0.476 til 0.544

Forskjellige måter å oppgi konfidensintervall på

Kjært barn har mange navn

$$0.476 < p < 0.544$$

$$0.51 \pm 0.034$$

$$< 0.476, 0.544 >$$

Konfidensintervall for andelen

Prosedyre

1. Sjekk at krav på side 5 OK
2. Finn kritisk verdi $z_{\alpha/2}$ i tabell A2
3. Regn ut feilmargin $E = z_{\alpha/2} \cdot \sqrt{\hat{p}(1-\hat{p})/n}$
4. Regn ut nedre grense $\hat{p} - E$ og øvre grense $\hat{p} + E$
5. Rund av til tre desimaler
6. Oppgi konfidensintervallet

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