

# Prediksjonsintervall

TMA4240/TMA4245 Statistikk

Håkon Tjelmeland

Institutt for matematiske fag

Norges teknisk-naturvitenskapelige universitet

# Prediksjonsintervall

Observator: En observerbar funksjon av en eller flere variabler som utgjør et utvalg

★ Situasjon:

- har en populasjon, parameter  $\theta$
- observerer verdier for et utvalg,  $X_1, X_2, \dots, X_n$
- ønsker intervall som inneholder verdi av ny observasjon  $X^*$  fra samme populasjon

## Definisjon (Prediksjonsintervall)

Anta at vi har stokastiske variabler  $X_1, X_2, \dots, X_n$ , og at sannsynlighetsfordelingen til disse inneholder en parameter  $\theta$  og at verdien til  $\theta$  er ukjent. La  $x_1, x_2, \dots, x_n$  være observerte verdier for de stokastiske variabler  $X_1, X_2, \dots, X_n$ . Anta videre at  $X^*$  representerer en fremtidig observasjon og at sannsynlighetsfordelingen til denne også avhenger av verdien av parameteren  $\theta$ . Anta så at vi for to observatorer  $\hat{X}_L^*(X_1, X_2, \dots, X_n)$  og  $\hat{X}_U^*(X_1, X_2, \dots, X_n)$  har at

$$P\left(\hat{X}_L^*(X_1, X_2, \dots, X_n) \leq X^* \leq \hat{X}_U^*(X_1, X_2, \dots, X_n)\right) = 1 - \alpha,$$

for  $\alpha \in (0, 1)$ . Det numeriske intervallet

$$\left[\hat{X}_L^*(x_1, x_2, \dots, x_n), \hat{X}_U^*(x_1, x_2, \dots, x_n)\right]$$

kalles da et  $(1 - \alpha) \cdot 100\%$ -prediksjonsintervall for  $X^*$

- ★ Hvordan finne slike observatorer  $\hat{X}_L^*(X_1, X_2, \dots, X_n)$  og  $\hat{X}_U^*(X_1, X_2, \dots, X_n)$ ?
- ★ Hva er tolkningen av intervallet  $\left[\hat{X}_L^*(x_1, x_2, \dots, x_n), \hat{X}_U^*(x_1, x_2, \dots, x_n)\right]$ ?

# Eksempel: Utledding av prediksjonsintervall

$$\text{Husk: } \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

★ Situasjon:

- la  $X_1, X_2, \dots, X_n$  og  $X^*$  være et tilfeldig utvalg fra  $N(\mu, \sigma^2)$ -populasjonen
- observerte verdier for  $X_1, X_2, \dots, X_n$  er  $x_1, x_2, \dots, x_n$
- verdien til  $\mu$  er ukjent, antar verdien til  $\sigma^2$  kjent!
- ønsker et prediksjonsintervall for  $X^*$

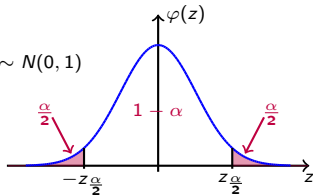
★ Estimator for  $\mu$ :  $\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

★ Tar utgangspunkt i  $X^* - \bar{X}$ :

$$X^* - \bar{X} \sim N\left(0, \sigma^2 + \frac{\sigma^2}{n}\right) \Rightarrow \frac{\bar{X} - X^*}{\sqrt{\sigma^2 \left(1 + \frac{1}{n}\right)}} \sim N(0, 1)$$

★ Dermed

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{X^* - \bar{X}}{\sqrt{\sigma^2 \left(1 + \frac{1}{n}\right)}} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$



★ Løser hver ulikhet hver for seg med hensyn på  $X^*$ :

$$\bar{X} - z_{\frac{\alpha}{2}} \sqrt{\sigma^2 \left(1 + \frac{1}{n}\right)} \leq X^* \quad \text{og} \quad X^* \leq \bar{X} + z_{\frac{\alpha}{2}} \sqrt{\sigma^2 \left(1 + \frac{1}{n}\right)}$$

★ Får at

$$P\left(\bar{X} - z_{\frac{\alpha}{2}} \sqrt{\sigma^2 \left(1 + \frac{1}{n}\right)} \leq X^* \leq \bar{X} + z_{\frac{\alpha}{2}} \sqrt{\sigma^2 \left(1 + \frac{1}{n}\right)}\right) = 1 - \alpha$$

★  $(1 - \alpha) \cdot 100\%$ -prediksjonsintervall for  $X^*$ :  $\left[ \bar{x} - z_{\frac{\alpha}{2}} \sqrt{\sigma^2 \left(1 + \frac{1}{n}\right)}, \bar{x} + z_{\frac{\alpha}{2}} \sqrt{\sigma^2 \left(1 + \frac{1}{n}\right)} \right]$

# Utleiding av prediksjonsintervall

★ Prosedyre for å utlede et prediksjonsintervall for en ny observasjon  $X^*$  basert på  $x_1, x_2, \dots, x_n$

1. Bestemme en stokastisk variabel,  $Z = h(X^*, X_1, X_2, \dots, X_n)$ 
  - inneholder ingen ukjente parametre
  - har en kjent fordeling

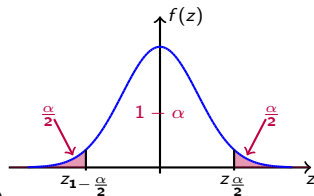
$$Z = \frac{X^* - \bar{X}}{\sqrt{\sigma^2 \left(1 + \frac{1}{n}\right)}} \sim N(0, 1)$$

2. Finne kvantiler  $z_{1-\frac{\alpha}{2}}$  og  $z_{\frac{\alpha}{2}}$  slik at

$$P\left(z_{1-\frac{\alpha}{2}} \leq h(X^*, X_1, X_2, \dots, X_n) \leq z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

3. Løse hver ulikhet med hensyn på  $X^*$
4. Sette ulikhetene sammen igjen med  $X^*$  i midten

$$P\left(\hat{X}_L^*(X_1, X_2, \dots, X_n) \leq X^* \leq \hat{X}_U^*(X_1, X_2, \dots, X_n)\right) = 1-\alpha$$



5. Konkludere ved å skrive opp  $(1 - \alpha) \cdot 100\%$ -prediksjonsintervallet for  $X^*$

$$\left[\hat{X}_L^*(x_1, x_2, \dots, x_n), \hat{X}_U^*(x_1, x_2, \dots, x_n)\right]$$

# Tolkning av prediksjonsintervall

$P(A)$ : Andel av gangene hendelsen  $A$  skjer hvis man gjentar det stokastiske forsøket uendelig mange ganger

- ★ Kan ikke tolke selve prediksjonsintervallet

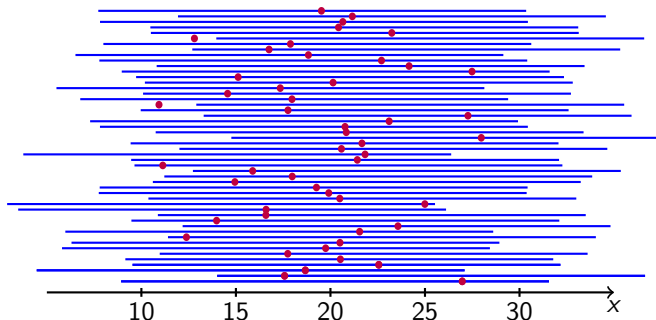
$$\left[ \hat{X}_L^*(x_1, x_2, \dots, x_n), \hat{X}_U^*(x_1, x_2, \dots, x_n) \right]$$

- ★ Må tolke sannsynlighetsuttrykket

$$P\left(\hat{X}_L^*(X_1, X_2, \dots, X_n) \leq X^* \leq \hat{X}_U^*(X_1, X_2, \dots, X_n)\right) = 1 - \alpha$$

- ★ Simuleringseksempel:

- la  $X_1, X_2, \dots, X_n$  og  $X^*$  være tilfeldig utvalg fra  $N(\mu, \sigma^2)$ -populasjonen
- verdien til  $\mu$  er ukjent, ønsker prediksjonsintervall for  $X^*$ , antar verdien til  $\sigma^2$  kjent
- $(1 - \alpha) \cdot 100\%$ -prediksjonsintervall for  $X^*$ :  $\left[ \bar{x} - z_{\frac{\alpha}{2}} \sqrt{\sigma^2 \left(1 + \frac{1}{n}\right)}, \bar{x} + z_{\frac{\alpha}{2}} \sqrt{\sigma^2 \left(1 + \frac{1}{n}\right)} \right]$
- Stokastisk simulering med  $n = 3$ ,  $\mu = 20$ ,  $\sigma^2 = 5^2$  og  $\alpha = 0.05$



# Oppsummering

- ★ Har definert begrepet
  - prediksjonsintervall
- ★ Har diskutert
  - prosedyre for å utlede et prediksjonsintervall
  - tolkningen av prediksjonsintervall
    - tenker oss at vi gjentar målingene (også fremtidig måling!) uendelig mange ganger