

STK1110 Høsten 2021

Mer om konfidensintervaller

Tilsvare Avsnitt 8.3, 8.4 og 8.5

Ingrid Hobæk Haff
Matematisk institutt
Universitetet i Oslo

Lite utvalg uif data med ukjent varians

- Vi antar at vi har observasjoner x_1, \dots, x_n av de stokastiske variablene X_1, \dots, X_n , som er uif med forventning μ og varians σ^2 , som begge er ukjent.
- Videre ser vi her på det tilfellet at n er for lav til at sentralgrenseteoremet slår inn.
- Da trenger vi i tillegg å anta at $X_i \sim N(\mu, \sigma^2)$, $i = 1, \dots, n$.
- Vi ønsker å konstruere et konfidensintervall for μ , og da trenger vi estimatorer for både μ og σ^2 , og vi kan bruke \bar{X} og S^2 .

Konfidensintervall for μ

- Vi har observatoren

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

- Når n er stor, er T tilnærmet standard normalfordelt. Det gjelder imidlertid ikke for mindre n .
- Vi så i Kap. 6.4 at

$$\bar{X} \text{ og } S^2 \text{ er uavhengige, } \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \text{ og } \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2,$$

slik at $T \sim t_{n-1}$.

Konfidensintervall for μ (forts.)

- La $t_{\alpha,\nu}$ være øvre α -kvantil i t_ν -fordelingen.
- Vi har

$$P\left(-t_{\alpha/2,n-1} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{\alpha/2,n-1}\right) = 1 - \alpha.$$

- Det kan omformes til

$$P\left(\bar{X} - t_{\alpha/2,n-1} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2,n-1} \frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

Konfidensintervall for μ (forts.)

- Dermed inneholder det stokastiske intervallet

$$\left(\bar{X} - t_{\alpha/2, n-1} \frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2, n-1} \frac{S}{\sqrt{n}} \right),$$

μ med $100 \cdot (1 - \alpha)\%$ sannsynlighet.

- Ved å sette inn observerte verdier får vi et $100 \cdot (1 - \alpha)\%$ konfidensintervall for μ :

$$\left(\bar{x} - t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \right).$$

Eksempler

Eksempel

En kjemiker er interessert i å bestemme konsentrasjonen μ av et stoff i en løsning. Hun måler konsentrasjonen i mg/l 5 ganger. Vi antar at målingene er uif med forventning μ og standardavvik σ , som er ukjent, og vi ønsker å lage et konfidensintervall for μ .

Eksempel

Tenk nå at kjemikeren vil måle konsentrasjonen én gang til. Kan vi lage et intervall som med stor sikkerhet vil inneholde den nye observasjonen? Dette kalles et prediksjonsintervall.

Prediksjonsintervall for en ny observasjon

- Generelt har vi denne situasjonen: $X_1, \dots, X_n \stackrel{\text{uif}}{\sim} N(\mu, \sigma^2)$ og vi vil predikere verdien til en ny observasjon X_{n+1} .
- En prediktor for X_{n+1} er \bar{X} .
- Prediksjonsfeilen er da $\bar{X} - X_{n+1}$ og oppfyller

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

- Videre er $\bar{X} - X_{n+1}$ og S^2 uavhengige, slik at

$$T = \frac{\bar{X} - X_{n+1}}{S \sqrt{1 + \frac{1}{n}}} \sim t_{n-1}.$$

Prediksjonsintervall for en ny observasjon (forts.)

- Ved å følge samme framgangsmåte som for konfidensintervallet for μ , får vi et $100 \cdot (1 - \alpha)\%$ prediksjonsintervall for X_{n+1} gitt ved

$$\left(\bar{x} - t_{\alpha/2, n-1} s \left(1 + \frac{1}{\sqrt{n}} \right), \bar{x} + t_{\alpha/2, n-1} s \left(1 + \frac{1}{\sqrt{n}} \right) \right).$$

- Prediksjonsintervallet for en ny observasjon er mye videre enn konfidensintervallet for forventningen.
- Konfidensintervallet vil kollapse mot den enkelte verdien μ når $n \rightarrow \infty$ da all usikkerhet om parameterne forsvinner.
- Prediksjonsintervallet, derimot, går mot $\mu \pm z_{\alpha/2} \sigma$ da det alltid er en viss usikkerhet i den nye verdien en vil predikere, selv når parameterne er kjent.

Konfidensintervall for σ^2 og σ

- Vi antar at vi har observasjoner x_1, \dots, x_n av de stokastiske variablene $X_1, \dots, X_n \stackrel{\text{uif}}{\sim} N(\mu, \sigma^2)$, der både μ og σ^2 er ukjent.
- Vi ønsker nå å lage et $100 \cdot (1 - \alpha)\%$ konfidensintervall for σ^2 og σ .
- Fra Kap 6.4 vet vi at $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$.
- La $\chi_{\alpha, \nu}^2$ være øvre α -kvantil i χ_{ν}^2 -fordelingen.
- Da har vi:

$$P\left(\chi_{1-\alpha/2, n-1}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{\alpha/2, n-1}^2\right) = 1 - \alpha.$$

Konfidensintervall for σ^2 og σ

- Det kan omformes til

$$P \left(\frac{(n-1)S^2}{\chi_{\alpha/2, n-1}^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{1-\alpha/2, n-1}^2} \right)$$
$$P \left(S \sqrt{\frac{(n-1)}{\chi_{\alpha/2, n-1}^2}} \leq \sigma \leq S \sqrt{\frac{(n-1)}{\chi_{1-\alpha/2, n-1}^2}} \right) = 1 - \alpha.$$

- Det gir $100 \cdot (1 - \alpha)\%$ konfidensintervall for:

$$\sigma^2 : \left(\frac{(n-1)s^2}{\chi_{\alpha/2, n-1}^2}, \frac{(n-1)s^2}{\chi_{1-\alpha/2, n-1}^2} \right)$$
$$\sigma : \left(s \sqrt{\frac{(n-1)}{\chi_{\alpha/2, n-1}^2}}, s \sqrt{\frac{(n-1)}{\chi_{1-\alpha/2, n-1}^2}} \right).$$

Konfidensintervaller basert på bootstrap

- Vi antar at vi har observasjoner x_1, \dots, x_n av de stokastiske variablene X_1, \dots, X_n , som er uif med en fordeling som avhenger av en ukjent parameter θ .
- Vi ønsker å estimere θ , samt lage et $100 \cdot (1 - \alpha)\%$ konfidensintervall for θ .
- Anta videre at vi har estimatoren $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$, som gir estimatet $\hat{\theta} = \hat{\theta}(x_1, \dots, x_n)$.
- Vi skal nå bruke ikke-parametrisk bootstrap til å lage konfidensintervall.

Konfidensintervaller basert på bootstrap (forts.)

- Vi begynner med å bootstrappe $\hat{\theta}$:
 - ① For $b = 1, \dots, B$
 - ② Trekk $x_{b1}^*, \dots, x_{bn}^*$ fra x_1, \dots, x_n med tilbakelegging.
 - ③ La $\hat{\theta}_b^* = \hat{\theta}(x_{b1}^*, \dots, x_{bn}^*)$.
- Basert på $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$ kan en lage konfidensintervall på flere måter.
- Dersom disse er tilnærmet normalfordelt med forventning θ og standardavvik $\sigma_{\hat{\theta}}$, kan en estimere $\sigma_{\hat{\theta}}$ med
$$s_{boot} = \sqrt{\frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}_b^* - \bar{\hat{\theta}}^*)^2}, \text{ med } \bar{\hat{\theta}}^* = \frac{1}{B} \sum_{b=1}^B \hat{\theta}_b^*.$$
- En bruker da at $\frac{\hat{\theta} - \theta}{s_{boot}} \stackrel{\text{tiln.}}{\sim} N(0, 1)$.
- Da er $\hat{\theta} \pm z_{\alpha/2} s_{boot}$ et tilnærmet $100 \cdot (1 - \alpha)\%$ konfidensintervall for θ .

Konfidensintervaller basert på bootstrap (forts.)

- Når normalfordelingsantakelsen ikke holder, kan vi i stedet bruke **persentilintervallet**.
- Det får en som følger:
 - 1 Sortér bootstrap-estimatene, slik at $\hat{\theta}_{(1)}^* \leq \dots \leq \hat{\theta}_{(B)}^*$.
 - 2 Persentilintervallet for θ er $\left(\hat{\theta}_{(B\alpha/2)}^*, \hat{\theta}_{(B(1-\alpha/2))}^* \right)$.

Konfidensintervaller basert på bootstrap (forts.)

- I visse tilfeller er det behov for å gjøre noen korreksjoner av persentilintervallet.
- F.eks. kan det være skjevhet i bootstrapfordelingen sammenlignet med det opprinnelige estimatet $\hat{\theta}$.
- Denne skjevheten kan en vurdere ved å beregne $\hat{\theta} - \bar{\theta}^*$.
- Et annet problem kan være at $\sigma_{\hat{\theta}}$ avhenger av θ .
- I slike situasjoner kan en i stedet bruke BCa-intervallet (står for "bias corrected and accelerated"), som prøver å korrigere problemene over ved å bruke andre persentiler enn $\alpha/2$ og $1 - \alpha/2$.