STK-IN1050 - Statistikk for informatikere

Repetisjon

Matematisk institutt

2026-05-20

Pensum

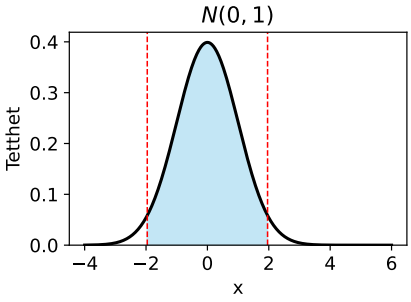
- Eksplorativ dataanalyse
- Grunnleggende sannsynlighetsteori
- Stokastiske variabler og statistiske fordelinger
- Monte Carlo simulering
- Dynamiske prosesser
- Punktestimering og estimeringsusikkerhet
- Bootstrapping
- Konfidensintervaller og hypotesetester
- Lineær regresjon
- Klassifisering

Dagens tema

- Sannsynlighetsteori: grunnprinsipper
- Konfidensintervaller
- *t*-test
- Statistiske fordelinger
- Lineær regresjon
- Logistisk regresjon

	Store talls lov	Sentralgrenseteoremet
Fokus	Verdien til utvalgsgjennomsnittet.	Formen på fordelingen til utvalgsgjennomsnitt.
Hva den sier	Når utvalgsstørrelsen øker, vil utvalgsgjennomsnittet konvergere mot den sanne forventningen i populasjonen.	Når utvalgsstørrelsen øker, vil fordelingen til utvalgsgjennomsnittene nærme seg en normalfordeling.
Hvorfor det er viktig	Det forteller oss at ved å samle inn mer data får vi et mer presist estimat på «sannheten».	Det gjør det mulig å utføre statistiske tester (beregne *p*-verdier og konfidensintervall), og bruke normalfordeling i modeller, selv om de opprinnelige dataene ikke er normalfordelte.
Når kan den brukes	Fungerer for nesten alle sannsynlighetsfordelinger.	Krever vanligvis en utvalgsstørrelse på minst $n \geq 30$ for å gi en tydelig klokkeformet kurve.

Standard normalfordeling



Hvis X er en stokastisk variabel med

$$X \sim N(0, 1)$$

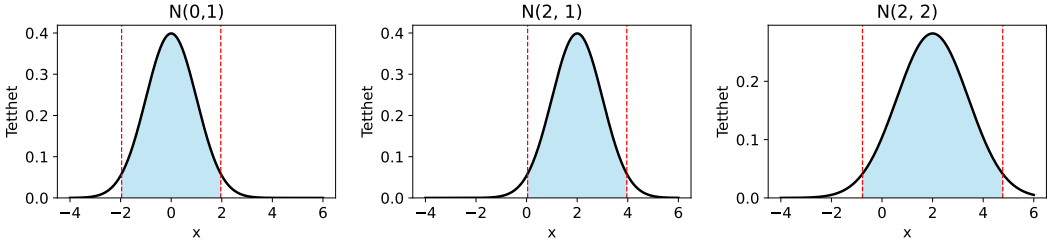
da er 95% sannsynlighet for at et utfall x ligger på intervallet

$$[-1,96, 1,96],$$

dersom, i standard normalfordelingen,

- -1,96 er 0,025-kvantilen (= 2,5-persentilen)
- 1,96 er 0,975-kvantilen (= 97,5-persentilen)

Generell normalfordeling



Hvis $X \sim N(2, 1)$, da er 95% sannsynlighet for at et utfall x ligger på intervallet

$$[2 - 1,96, 2 + 1,96] = [0,04, 3,96].$$

Hvis $X \sim N(2, 2)$, da er 95% sannsynlighet for at et utfall x ligger på intervallet

$$[2 - 1,96 \times \sqrt{2}, 2 + 1,96 \times \sqrt{2}] \approx [-0,77, 4,77].$$

Generell normalfordeling

Den generelle regelen er: Hvis $X \sim N(\mu, \sigma^2)$, da er

$$\frac{X - \mu}{\sigma} \sim N(0, 1)$$

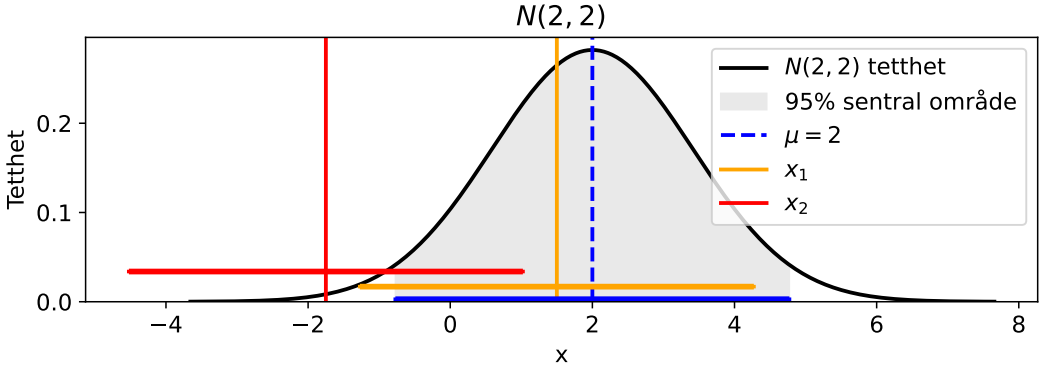
$\Rightarrow$ For et utfall $y = (x - \mu)/\sigma$ gjelder med 95% sannsynlighet, at

$$-1,96 \leq y \leq 1,96.$$

Dette er ekvivalent med

$$-1,96 \leq \frac{x - \mu}{\sigma} \leq 1,96 \quad \Longleftrightarrow \quad \mu - 1,96 \times \sigma \leq x \leq \mu + 1,96 \times \sigma.$$

(Vi kan også erstatte «95% sannsynlighet» med hvilken som helst sannsynlighet hvor vi da må erstatte -1,96 og 1,96 med de tilsvarende kvantilene.)



Vurdere intervaller sentrert på tilfeldige utvalg $x_1, x_2, \ldots$ og med bredde lik bredden til det 95% sentrale område. Forventningen μ vil ligge innenfor intervallet i 95% av tilfellene: akkurat når x_i ligger innenfor det 95% sentrale området rundt μ.

Gjennomsnitt $\longrightarrow$ forventning

Hvis $X_i \sim N(\mu, \sigma^2)$ uavhengig, da er $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$.

Hvis $X_i \nsim N(\mu, \sigma^2)$ uavhengig, da er $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \approx N(\mu, \frac{\sigma^2}{n})$ hvis n er stor nok.

Gjennomsnitt $\longrightarrow$ forventning

Hvis $X_i \sim N(\mu, \sigma^2)$ uavhengig, da er $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$.

Hvis $X_i \nsim N(\mu, \sigma^2)$ uavhengig, da er $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \approx N(\mu, \frac{\sigma^2}{n})$ hvis n er stor nok.

$\Rightarrow \frac{\overline{X}-\mu}{(\sigma/\sqrt{n})} \sim N(0,1)$ eller $\frac{\overline{X}-\mu}{(\sigma/\sqrt{n})} \approx N(0,1)$ hvis n stor nok

Hvis vi nå observerer $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, da vil μ ligge innenfor intervallet

$$\left[\overline{x} - 1,96 \times \frac{\sigma}{\sqrt{n}}, \quad \overline{x} + 1,96 \times \frac{\sigma}{\sqrt{n}}\right]$$

i 95% av de tilfellene vi gjennomfører denne prosessen.

Gjennomsnitt $\longrightarrow$ forventning

Hvis $X_i \sim N(\mu, \sigma^2)$ uavhengig, da er $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$.

Hvis $X_i \nsim N(\mu, \sigma^2)$ uavhengig, da er $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \approx N(\mu, \frac{\sigma^2}{n})$ hvis n er stor nok.

$\Rightarrow \frac{\overline{X}-\mu}{(\sigma/\sqrt{n})} \sim N(0,1)$ eller $\frac{\overline{X}-\mu}{(\sigma/\sqrt{n})} \approx N(0,1)$ hvis n stor nok

Hvis vi nå observerer $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, da vil μ ligge innenfor intervallet

$$\left[\overline{x} - 1,96 \times \frac{\sigma}{\sqrt{n}}, \quad \overline{x} + 1,96 \times \frac{\sigma}{\sqrt{n}}\right]$$

i 95% av de tilfellene vi gjennomfører denne prosessen.

$\Rightarrow$ Intervallet er et 95% konfidensintervall og gir informasjon om usikkerheten i estimatet $\overline{x} \approx \mu$.

Gjennomsnitt $\longrightarrow$ forventning

Hvis σ er ukjent, bruker vi en estimator S for σ.

Da har vi mer usikkerhet.

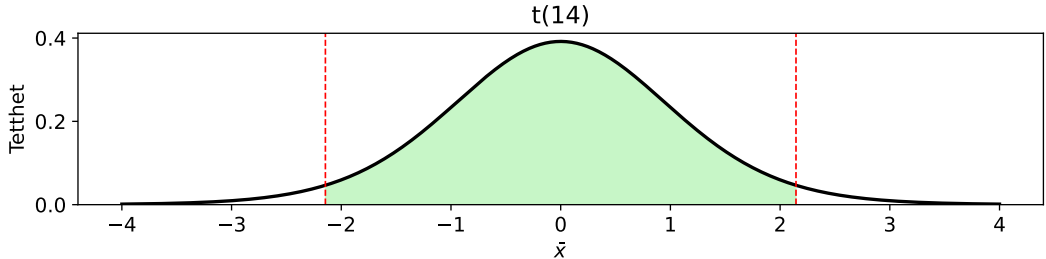
$\Rightarrow \frac{\bar{X}-\mu}{(S/\sqrt{n})} \sim t(n-1)$ eller $\frac{\bar{X}-\mu}{(S/\sqrt{n})} \approx t(n-1)$ hvis n stor nok

$\Rightarrow$ vi bruker 0,975-kvantilen t^* fra $t(n-1)$-fordelingen til å lage et 95% konfidensintervall:

$$\left[\bar{x} - t^* \times \frac{s}{\sqrt{n}}, \quad \bar{x} + t^* \times \frac{s}{\sqrt{n}}\right].$$

Vi har at $t^* > 1,96$ og $t^* \approx 1,96$ hvis n stor.

Én-utvalgs *t*-test



Én-utvalgs *t*-test med $H_0 : \mu = 0$ mot $H_a : \mu \neq 0$ med signifikansnivå $\alpha = 0.05$ bygger på akkurat samme prinsipp. Anta vi har $n = 15$. Hvis

$$\frac{\bar{x} - 0}{s/\sqrt{n}} < -2,145 \quad \text{eller} \quad \frac{\bar{x} - 0}{s/\sqrt{n}} > 2,145$$

blir *p*-verdien lavere enn signifikansnivået og vi forkaster nullhypotesen.

To-utvalgs *t*-test

En *t*-test brukes ofte til å sammenligne to grupper:

$$H_0 : \mu_1 = \mu_2, \quad H_a : \mu_1 \neq \mu_2, \quad \alpha = 0,05,$$

eller, ekvivalent,

$$H_0 : \mu_1 - \mu_2 = 0, \quad H_a : \mu_1 - \mu_2 \neq 0, \quad \alpha = 0,05.$$

To-utvalgs *t*-test

En *t*-test brukes ofte til å sammenligne to grupper:

$$H_0 : \mu_1 = \mu_2, \quad H_a : \mu_1 \neq \mu_2, \quad \alpha = 0,05,$$

eller, ekvivalent,

$$H_0 : \mu_1 - \mu_2 = 0, \quad H_a : \mu_1 - \mu_2 \neq 0, \quad \alpha = 0,05.$$

Under nullhypotesen gjelder

$$\bar{X}_1 - \bar{X}_2 \approx N(0, \sigma^2_{\text{diff}}) \quad \text{eller} \quad \frac{\bar{X}_1 - \bar{X}_2}{\text{SE}(\bar{X}_1 - \bar{X}_2)} \approx t(\text{df}),$$

hvor $\text{SE}(\bar{X}_1 - \bar{X}_2)$ og df avhenger av om vi antar lik eller forskjellig varians i de to gruppene (se forelesning 16).

Betingelser for å bruke en *t*-test

- Data uavhengig og tilfeldig utvalg
- Tilnærmet normalfordelte data eller stor utvalgsstørrelse
  - $n \leq 15$: Trygt hvis data er nær normale
  - $n \geq 15$: Kan brukes hvis ingen outliers eller sterk skjevhet
  - $n \geq 40$: Kan brukes selv med sterk skjevhet

Diskrete og kontinuerlige fordelinger

Fordeling	Nøkkelparametre	Verdimengde	Typiske egenskaper	Vanlige bruksområder
Bernoulli	p = sannsynlighet for suksess	$x \in \{0, 1\}$	Én enkelt ja/nei-hendelse.	Modellere enkeltutfall: myntkast (krone/mynt), om en kunde kjøper eller ikke, om en test er positiv.
Binomisk	n = antall forsøk, p = sannsynlighet for suksess	$x \in \{0, 1, \ldots, n\}$	Antall suksesser i n uavhengige Bernoulli-forsøk.	Antall "treff" på annonser, antall defekte varer i et parti, antall vellykkede behandlinger.
Poisson	λ = rate	$x \in \{0, 1, 2, \ldots\}$	Teller hendelser over tid/rom.	Antall ankomster per time (kunder, telefoner), antall feil på en kabel eller en maskin per tidsenhet.
Uniform	a, b	$x \in [a, b]$	Alle verdier i intervallet like sannsynlige.	Simulering, tilfeldige starttidspunkter, baseline-modeller.
Eksponensiell	λ = rate	$x \in [0, \infty)$	Tid mellom hendelser i en Poisson-prosess. «Minneløs» fordeling.	Modellerer ventetid: tid til neste kunde, tid til feil, levetid for komponenter uten aldringseffekt.
Normal	μ = forventning, σ = sd	$x \in (-\infty, \infty)$	Klokkeformet, symmetrisk.	Høyde, testscore, målefeil; brukes i statistisk inferens.
t	ν = frihetsgrader	$x \in (-\infty, \infty)$	Symmetrisk som normalfordeling, men med «tykkere haler», spesielt ved små ν.	Hypotesetesting og konfidensintervall for middelverdi når σ er ukjent og utvalget er lite.

Lineær regresjon

$$Y \sim N(\mu_x, \sigma^2)$$

hvor

$$\mu_x = \beta_0 + \beta_1 x$$

Et $(1-\alpha)100\%$ konfidensintervall for β_1 er gitt ved

$$b_1 \pm t^* \, \text{SE}(b_1)$$

hvor t^* er den kvantilen fra $t(n-2)$-fordelingen hvor arealet mellom $-t^*$ og t^* er $(1-\alpha)$.

Logistisk regresjon

$$Y \sim Bin(1, p(x))$$

hvor

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$$

Et $(1-\alpha)100\%$ konfidensintervall for β_1 er gitt ved

$$b_1 \pm z^* \, \text{SE}(b_1)$$

hvor z^* er den kvantilen fra $N(0,1)$-fordelingen hvor arealet mellom $-z^*$ og z^* er $(1-\alpha)$.

Egenskap	Lineær regresjon	Logistisk regresjon
Primært formål	Modellere og predikere en kontinuerlig respons.	Modellere og predikere en binær respons (f.eks. 0/1, ja/nei).
Responsvariabel	$Y \in (-\infty, \infty)$ er numerisk	$Y \in \{0, 1\}$
Modellform	$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon.$	$\text{logit}(p(x)) = \log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$
Antatt fordeling	Feilledd ε antas normalfordelt med konstant varians, $\varepsilon \sim N(0, \sigma^2)$.	Responsen Y følger en Bernoulli/binomisk modell med sannsynlighet $p(x)$, $Y \sim Bin(1, p(x))$.
Tolkning av koeffisient	β_j: endring i gjennomsnittlig y når x_j øker med 1 enhet, alt annet fast.	β_j: endring i log-odds; $\exp(\beta_j)$ er oddsratio når x_j øker med 1 enhet, alt annet fast.
Predikert verdi	Predikert $\hat{y}$ kan være hvilken som helst reell verdi.	Predikert sannsynlighet $\hat{p} \in (0, 1)$ for at $Y = 1$.
Typiske bruksområder	Predikere nivåer: pris, temperatur, testscore osv.	Klassifisering: sannsynlighet for kjøp, sykdom, frafall, klikk på annonse osv.
Sensitiv mot outliers	Ganske sensitiv, siden prediksjoner er lineære og ubegrenset.	Mindre sensitiv i responsrommet (0–1), men outliers i kovariater kan fortsatt påvirke.
Mål på tilpasning	Ofte R^2, residualplott, RMSE.	Ofte ACC, ROC-kurve og AUC, Brier score.
Når brukes den	Når responsen er kontinuerlig og antakelser om lineæritet og normalitet er rimelige.	Når responsen er binær (eller andeler) og vi ønsker sannsynligheter og odds for ulike utfall.