STK-IN1050 - Statistikk for informatikere

14. Konfidensintervaller

Matematisk institutt

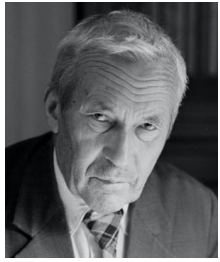
2026-03-23

Hvordan kan tilfeldige fenomener beskrives matematisk?

Kolmogorovs aksiomer (Andrey Kolmogorov, 1933)

Aksiom 1: For enhver begivenhet A er $P(A) \geq 0$.
Aksiom 2: For utfallsrommet Ω er $P(\Omega) = 1$.
Aksiom 3: Hvis $A_1, A_2, A_3, \ldots$ er en uendelig følge av disjunkte begivenheter, så er

$$P(A_1 \cup A_2 \cup \cdots) = P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i).$$



Andrey Kolmogorov (1903-1987)

Kilde: https://mathigon.org/

Aksiomene sier ikke noe om hvilke verdier sannsynlighetene skal ha i en gitt situasjon. For å angi konkrete verdier for sannsynlighetene, må vi bruke kunnskaper om den situasjonen vi studerer.

Karakteristikker vs. stokastiske variabler

Deskriptiv statistikk	Sannsynlighetsteori
karakteristikker	stokastiske variabler
verdier	verdier (fra en verdimengde)
absolutte og relative frekvenser	sannsynligheter for (mengder av) muliger verdier

- Merk: mulige verdier av en stokastisk variabel må være tall.
- Skille: diskrete vs. kontinuerlige stokastiske variabler.

Estimat og estimator

Vi observerer data $x_1, x_2, \ldots, x_n$. Et **estimat** av en ukjent parameter θ er

$$\hat{\theta} = h(x_1, x_2, \ldots, x_n),$$

en funksjon som bare avhenger av dataene.

Hvis $X_1, X_2, \ldots, X_n$ er stokastiske variabler, da er den stokastiske variabelen

$$\hat{\theta} = h(X_1, X_2, \ldots, X_n)$$

en **estimator** for θ og $h(x_1, x_2, \ldots, x_n)$ er et utfall.

Estimatoren $\hat{\theta}$ har forventning, varians og **utvalgsfordeling**.

Vi skal nå bruke alt det vi har lært så langt til

- **statistisk inferens**: trekke konklusjoner om en populasjon basert på data fra et utvalg
- **prediksjon**: bruke data til å si noe om framtidige eller ukjente verdier

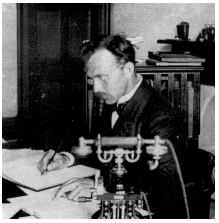
Statistisk inferens:

- **Konfidensintervaller** og **bootstrapping**: prinsipielle måter å kunne si noe om både estimater og den tilhørende usikkerheten i estimatene
- **Hypotesetesting**: systematisk metode for å vurdere hypoteser om en eller flere populasjoner basert på data

Prediksjon:

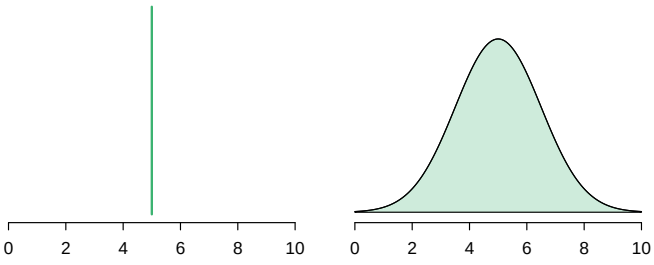
- **Regresjon**: metoder for å modellere og predikere en kontinuerlig responsvariabel ut fra én eller flere forklaringsvariabler
- **Klassifisering**: metoder for å tilordne observasjoner til én av flere kategorier basert på forklaringsvariabler

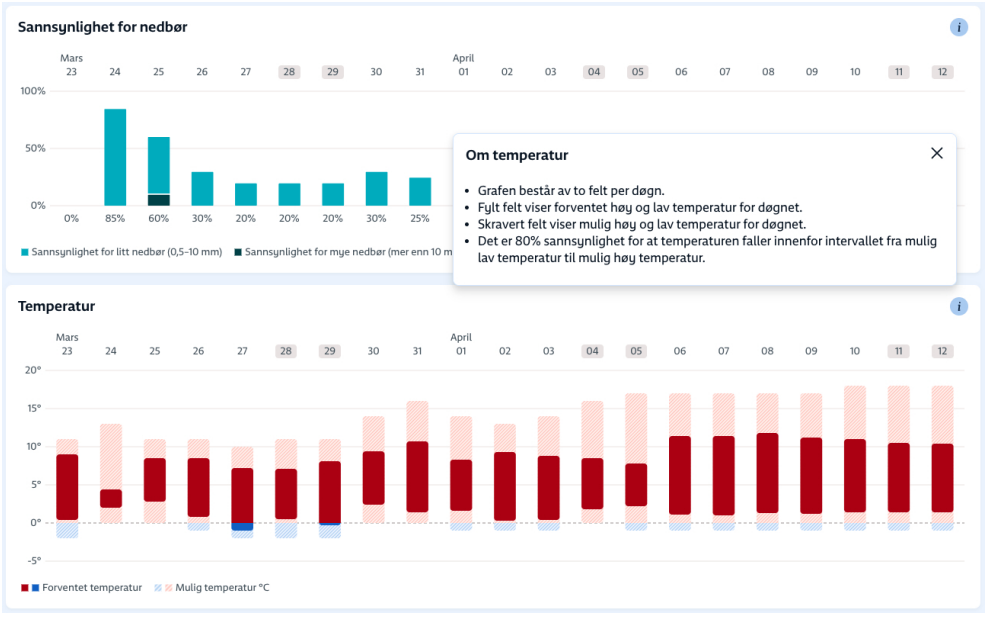
«All those whose duty it is to issue regular daily forecasts know that there are times when they feel very confident and other times when they are doubtful as to coming weather. It seems to me that the condition of confidence or otherwise forms a very important part of the prediction.» (Montly Weather Review, 1906)



W. Ernest Cooke (1863-1947)

Kilde: Wikipedia





Kilde: https://www.yr.no/nb/21-dagersvarsel/11-6332/Norge/Oslo/Oslo/Blindern

Store talls lov

Anta at $X_1, X_2, \ldots, X_n$ er uavhengige og identisk fordelte stokastiske variabler med $\mathbb{E}(X_i) = \mu$ for alle $i = 1, \ldots, n$. Da gjelder at

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \to \mu \quad \text{når } n \to \infty.$$

Sentralgrenseteoremet

Når n er stor, er fordelingen til $\overline{X}_n$ tilnærmet normal: $\overline{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$.

Standardisering av $\overline{X}_n$ gir: $Z = \dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \approx N(0,1)$ for store n.

Konfidensintervaller – hvorfor og hva?

Vi har lært å finne estimater (for eksempel et gjennomsnitt), men:

- Hvor nær sannheten er de?
- Hvor usikker er de?

Estimater i praksis

Tenk at vi samler inn data og beregner:

1. Gjennomsnittlig responstid for en webserver
2. Andel brukere som klikker på en knapp på en nettside

For hvert tilfelle har vi en populasjon:

- alle forespørsler til webserveren
- alle brukere som kunne besøkt nettsiden;

og en parameter som beskriver populasjonen:

- sann gjennomsnittlig responstid;
- sann klikkrate.

Punktestimat skjuler usikkerhet

Å rapportere bare

«Gjennomsnittlig responstid er 1.2 sekunder.»

skjuler to viktige ting:

- Estimatet er tilfeldig (hadde blitt annerledes med et annet utvalg)
- Vi sier ingenting om usikkerheten rundt dette tallet

Konsekvenser:

- Vanskelig å vurdere hvor mye vi kan stole på resultatet
- Vanskelig å sammenligne grupper eller metoder på en god måte

Tankeeksperiment: mange utvalg

Tenk at:

1. Den sanne parameteren θ (f.eks. sann gjennomsnittlig responstid) er fast, men ukjent
2. Vi gjør følgende mange ganger:
   - Trekker et tilfeldig utvalg
   - Beregner et intervall etter en fast oppskrift
   - Tegner intervallet på tallinjen

Da vil noen intervaller inneholde θ og noen vil bomme.

Et **95% konfidensintervall** er konstruert slik at:

I det lange løp vil ca. 95% av intervallene dekke θ.

Viktig presisering

Tilfeldigheten ligger i intervallet, ikke i θ:

- θ (f.eks. sann gjennomsnittstid) er fast, men ukjent
- Nye utvalg $\Rightarrow$ nye intervaller
- Metoden er laget slik at den treffer riktig en gitt andel ganger (f.eks. 95%)

Når vi sier:

«95% konfidensintervall for θ er $[0.8, 1.6]$»

mener vi:

«Vi har brukt en metode som i gjentatte forsøk vil gi intervaller som inneholder θ i 95% av tilfellene. For dette datasettet fikk vi intervallet $[0.8, 1.6]$.»

Konstruksjon av konfidensintervaller

Vi demonstrerer metoden på en enkel, men urealistisk situasjon.

Anta $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ hvor σ^2 er kjent og vi ønsker å estimere μ.

En estimator for forventningen er gitt ved $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ og vi har

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

Vi kan standardisere $\bar{X}$:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Z har altså en kjent fordeling.

Kvantiler

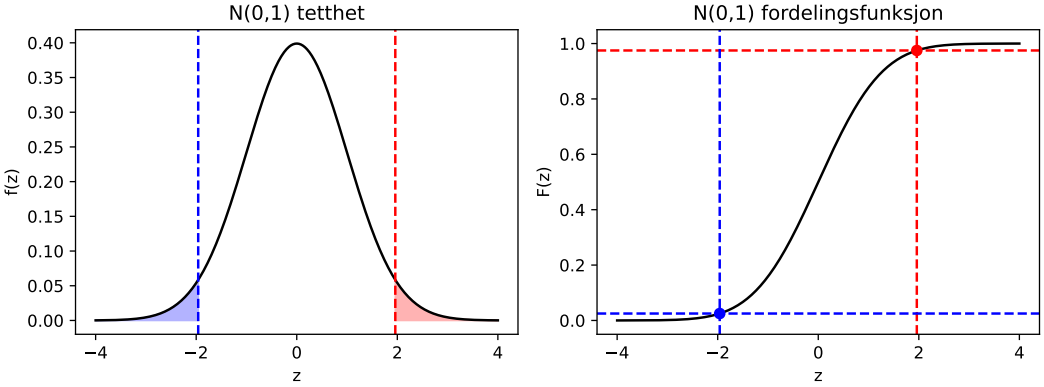
La $Z \sim N(0,1)$. Vi skal bruke $(1-p)$ kvantilen z_p gitt ved

$$P(Z \geq z_p) = p.$$

Eksempel: $z_{0.025} \approx 1.96$ fordi

$$P(Z \geq 1.96) \approx 0.025.$$

Kvantiler



Eksempel: Kvantiler $z_{0.025} = 1.96$ (rød) og $z_{0.975} = -1.96$ (blå).

z_p er altså det tallet slik at arealet til høyre for z_p er p.

Kvantiler

La $F(z) = P(Z \leq z)$ være fordelingsfunksjonen til Z.

Da har vi:

$$p = P(Z \geq z_p) = 1 - F(z_p),$$

dvs.

$$z_p = F^{-1}(1 - p).$$

På grunn av symmetri i normalfordelingen gjelder:

$$P(Z \leq -z_p) = P(Z \geq z_p) = p,$$

og dermed

$$z_{1-p} = -z_p.$$

Tosidig konfidensintervall

For et $(1-\alpha)100\%$ konfidensintervall ønsker vi:

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1-\alpha.$$

Dvs. vi deler halene i to:

- Venstre hale: sannsynlighet $\alpha/2$
- Høyre hale: sannsynlighet $\alpha/2$

Eksempler:

- 90% KI: $\alpha = 0.10 \Rightarrow \alpha/2 = 0.05 \Rightarrow z_{0.05} \approx 1.64$
- 95% KI: $\alpha = 0.05 \Rightarrow \alpha/2 = 0.025 \Rightarrow z_{0.025} \approx 1.96$
- 99% KI: $\alpha = 0.01 \Rightarrow \alpha/2 = 0.005 \Rightarrow z_{0.005} \approx 2.575$

Tosidig konfidensintervall

Vi setter inn definisjonen av Z:

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

Hvis vi isolerer μ får vi dermed:

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

Dette er definisjonen av et $(1-\alpha)100\%$-konfidensintervall for μ.

Konfidensintervall for normal forventning μ når σ^2 er kjent

Et **konfidensintervall** for μ med **konfidensnivå** $(1-\alpha)100\%$ er gitt ved

$$\left[\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \ \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

En kompakt notasjon er

$$\overline{x}\ \pm\ z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}},$$

hvor $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ kalles **feilmargin**.

Effekt av utvalgsstørrelse

Feilmarginen er

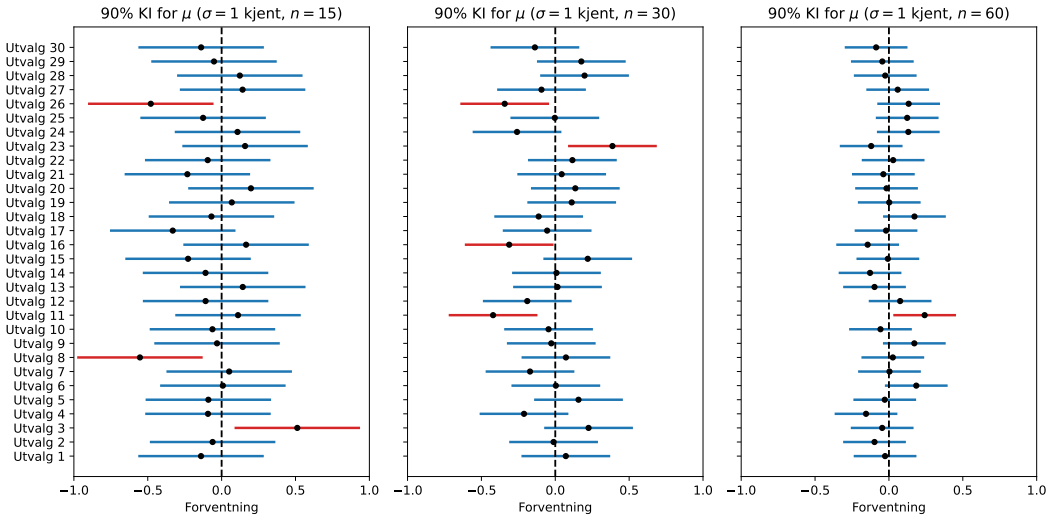
$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

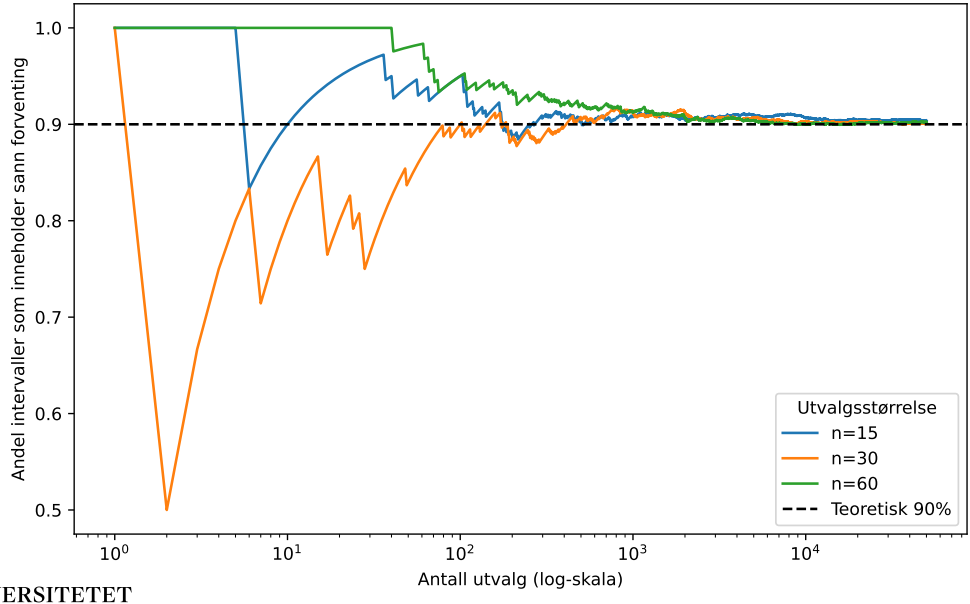
Vi har at

- $z_{\alpha/2}$ avhenger bare av konfidensnivået
- σ beskriver variabiliteten i populasjonen (antatt kjent)
- for $\sqrt{n}$ gjelder:
  - dobler vi $n \Rightarrow$ feilmarginen reduseres med faktor $\sqrt{2}$
  - firedobler vi $n \Rightarrow$ feilmarginen halveres

Intuisjon:

- mer data $\Rightarrow$ mer informasjon $\Rightarrow$ smalere intervall
- gevinsten er avtagende: å gjøre n veldig stor gir mindre og mindre ekstra innsnerving


90% KI for μ ($\sigma = 1$ kjent, $n = 15$)
90% KI for μ ($\sigma = 1$ kjent, $n = 30$)
90% KI for μ ($\sigma = 1$ kjent, $n = 60$)
Utvalg 30
Utvalg 29
Utvalg 28
Utvalg 27
Utvalg 26
Utvalg 25
Utvalg 24
Utvalg 23
Utvalg 22
Utvalg 21
Utvalg 20
Utvalg 19
Utvalg 18
Utvalg 17
Utvalg 16
Utvalg 15
Utvalg 14
Utvalg 13
Utvalg 12
Utvalg 11
Utvalg 10
Utvalg 9
Utvalg 8
Utvalg 7
Utvalg 6
Utvalg 5
Utvalg 4
Utvalg 3
Utvalg 2
Utvalg 1
−1.0
−0.5
0.0
0.5
1.0
Forventning


1.0
0.9
0.8
0.7
0.6
0.5
Andel intervaller som inneholder sann forventing
10^0
10^1
10^2
10^3
10^4
Antall utvalg (log-skala)
Utvalgsstørrelse
n=15
n=30
n=60
Teoretisk 90%

Fra kjent til ukjent varians

I praksis er σ^2 ukjent og må estimeres fra data.

Gitt data $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ estimerer vi nå både forventingen og variansen (standardavviket):

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad S = \sqrt{S^2}.$$

Vi definerer en standardisert stokastisk variabel

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

Nå er både telleren og nevneren tilfeldige; T **ikke** standard normalfordelt.

Likevel har vi et viktig resultat: Under normalantakelsen er T ***t*-fordelt**.

t-fordeling

Anta at $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ er uavhengige. Da gjelder:

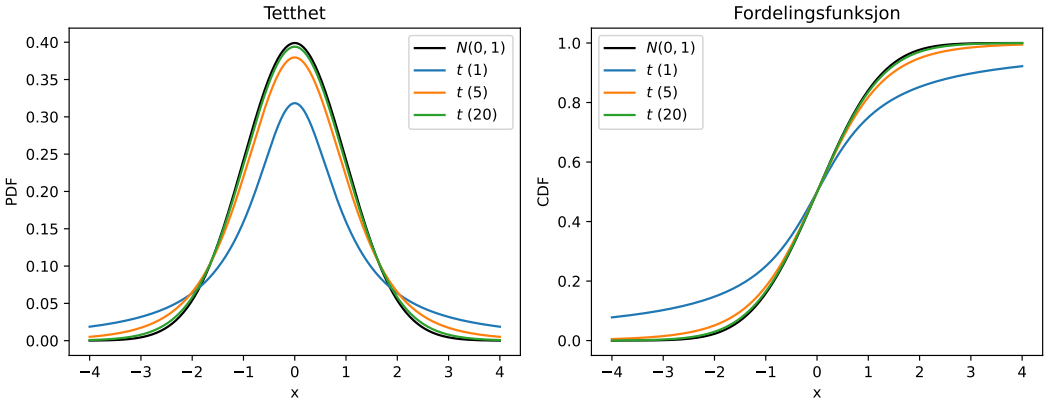
$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1),$$

dvs. T har en ***t*-fordeling med** $n-1$ **frihetsgrader**.

Denne fordelingen er symmetrisk rundt 0 og ligner normalfordelingen, men har tungere haler.

Når n blir stor nærmer t-fordelingen seg standard normalfordeling.

t vs. normal



Kritiske verdier $t_{n-1,\alpha/2}$ er større enn $z_{\alpha/2}$ for små n, men nærmer seg $z_{\alpha/2}$ for store n.

Konfidensintervall for μ basert på T

Vi velger $(1-\alpha/2)$ kvantil $t_{n-1,\alpha/2}$ slik at:

$$P\big(-t_{n-1,\alpha/2} \leq T \leq t_{n-1,\alpha/2}\big) = 1-\alpha,$$

hvor $T \sim t(n-1)$. Dvs.

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

Konfidensintervall for μ basert på T

Gitt data $x_1, \ldots, x_n$ finner vi gjennomsnittet og standardavviket til utvalget:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

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

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

som vi også kan skrive som

$$\bar{x} \ \pm \ t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}} \quad \text{eller} \quad \bar{x} \ \pm \ t_{n-1,\alpha/2}\,\widehat{\text{se}}(\bar{x}),$$

hvor $\widehat{\text{se}}(\bar{x}) = \frac{s}{\sqrt{n}}$ er **standardfeilen** til $\bar{x}$.

Sammenlikning: kjent vs. ukjent σ

- Kjent σ:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

- Ukjent σ:

$$\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}.$$

Forskjeller:

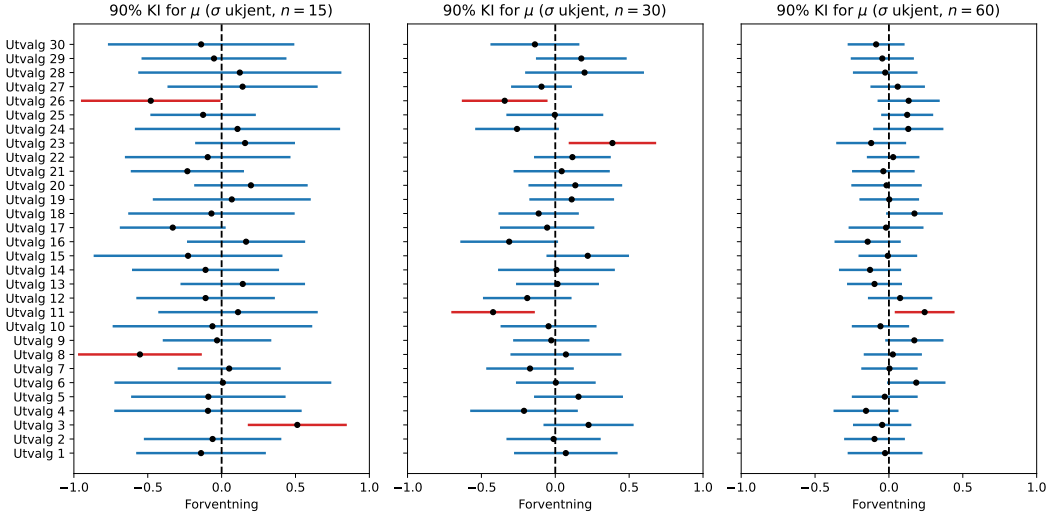
- Bruker s i stedet for σ (ulike intervaller har dermed ulik bredde)
- Bruker t-kvantil i stedet for normal-kvantil

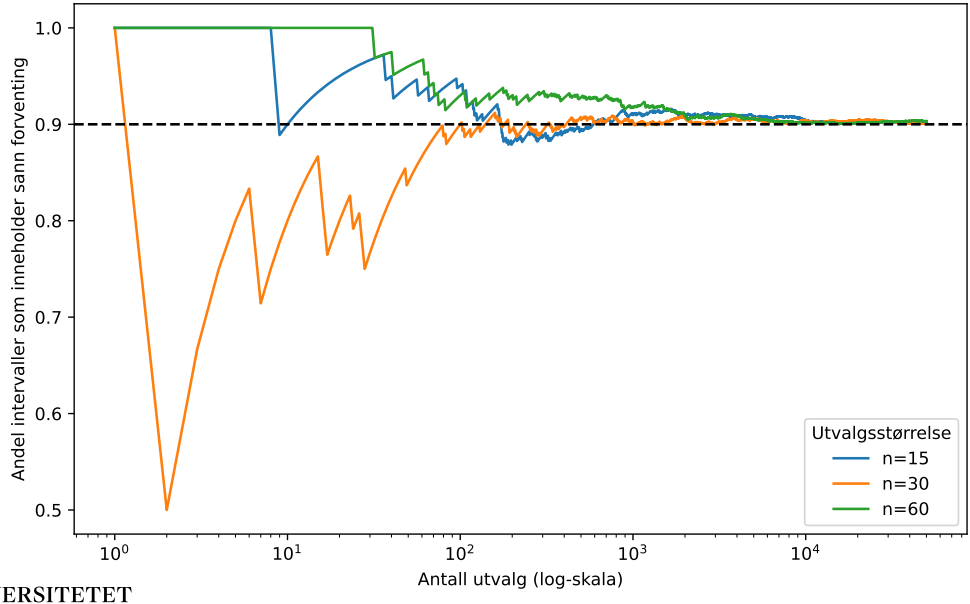
Sammenlikning: kjent vs. ukjent σ

t-kvantil $t_{n-1,\alpha/2}$ er større enn $z_{\alpha/2}$ for små n og nærmer seg $z_{\alpha/2}$ når n blir stor ($z_{0.05} \approx 1.64$):

```
from scipy.stats import t
alpha = 0.10
for n in [15, 60, 100, 1000]:
    df = n - 1
    t_crit = t.ppf(1 - alpha/2, df)
    print(f"n = {n}, df = {df}, 90% t-kritisk verdi: {t_crit:.4f}")
```

```
n = 15, df = 14, 90% t-kritisk verdi: 1.7613
n = 60, df = 59, 90% t-kritisk verdi: 1.6711
n = 100, df = 99, 90% t-kritisk verdi: 1.6604
n = 1000, df = 999, 90% t-kritisk verdi: 1.6464
```


90% KI for μ (σ ukjent, $n = 15$)
90% KI for μ (σ ukjent, $n = 30$)
90% KI for μ (σ ukjent, $n = 60$)
Utvalg 30
Utvalg 29
Utvalg 28
Utvalg 27
Utvalg 26
Utvalg 25
Utvalg 24
Utvalg 23
Utvalg 22
Utvalg 21
Utvalg 20
Utvalg 19
Utvalg 18
Utvalg 17
Utvalg 16
Utvalg 15
Utvalg 14
Utvalg 13
Utvalg 12
Utvalg 11
Utvalg 10
Utvalg 9
Utvalg 8
Utvalg 7
Utvalg 6
Utvalg 5
Utvalg 4
Utvalg 3
Utvalg 2
Utvalg 1
−1.0
−0.5
0.0
0.5
1.0
Forventning


1.0
0.9
0.8
0.7
0.6
0.5
Andel intervaller som inneholder sann forventing
10^0
10^1
10^2
10^3
10^4
Antall utvalg (log-skala)
Utvalgsstørrelse
n=15
n=30
n=60

1

Hva uttrykker et 95% konfidensintervall for en parameter?

(a) At det er 95% sannsynlighet for at parameteren ligger i intervallet
(b) At 95% av slike intervaller vil inneholde parameteren i gjentatte utvalg
(c) At parameteren alltid ligger i intervallet
(d) At utvalgsverdien alltid ligger i intervallet

2

Hva skjer med lengden på et konfidensintervall når konfidensnivået øker fra 95% til 99%?

(a) Intervallet blir kortere
(b) Intervallet blir lengre
(c) Intervallet forblir det samme
(d) Det avhenger ikke av konfidensnivået

3

Du får et 95% konfidensintervall for middelverdien: $[2, 6]$. Hva kan du si om den sanne middelverdien?

(a) Den sanne middelverdien kan ligge mellom 2 og 6
(b) Den sanne middelverdien er alltid 4
(c) Utvalgsverdien er alltid 2 eller 6
(d) Vi kan ikke si noe om middelverdien

Konfidensintervall for andeler

Vi observerer antall suksesser X i n forsøk hvor $X \sim \text{Bin}(n, p)$, der p er ukjent sannsynlighet for suksess.

En naturlig estimator for p er

$$\hat{p} = \frac{X}{n}.$$

Estimatoren har forventning og varians

$$\mathbb{E}(\hat{p}) = p, \quad \text{Var}(\hat{p}) = \frac{p(1-p)}{n}.$$

Vi ønsker et konfidensintervall for p basert på $\hat{p}$.

Normaltilnærming for $\hat{p}$

Når n er stor nok, og p ikke er for nær 0 eller 1, kan vi bruke sentralgrenseteoremet og tilnærme:

$$\hat{p} \approx N\left(p, \frac{p(1-p)}{n}\right).$$

Standardisering gir

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \approx N(0,1).$$

Utfordring: p er ukjent, også i standardavviket

Løsning: vi erstatter p med $\hat{p}$ i standardavviket

Standardfeil og z-statistikk

Vi estimerer standardfeilen:

$$\widehat{\text{se}}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Det gir en tilnærmet z-statistikk:

$$Z \approx \frac{\hat{p} - p}{\widehat{\text{se}}(\hat{p})} \approx N(0,1).$$

For $(1-\alpha)100\%$ konfidensintervall må vi da løse:

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) \approx P\Big(-z_{\alpha/2} \leq \frac{\hat{p} - p}{\widehat{\text{se}}(\hat{p})} \leq z_{\alpha/2} \Big) \approx 1 - \alpha.$$

Konfidensintervall for p (med normaltilnærming)

Det gir:

$$(1-\alpha)100\%\text{ KI for } p: \quad \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Dette er et **Wald-intervall** for p basert på normaltilnærming.

Normaltilnærmingen er vanligvis grei hvis

$$n\hat{p} \geq 10 \quad \text{og} \quad n(1-\hat{p}) \geq 10.$$

Dvs. vi trenger nok observasjoner av både suksess og feil

Eksempel: Konfidensintervall for sannsynligheten p

Anta at vi gjør $n = 100$ forsøk og observerer $X = 37$ suksesser

Estimat: $\hat{p} = \frac{37}{100} = 0.37$.

Sjekk tommelfingerregel: $n\hat{p} = 37 \geq 10, \quad n(1-\hat{p}) = 63 \geq 10$.

Vi ønsker et 95% konfidensintervall ($\alpha = 0.05$): $z_{0.025} \approx 1.96$

Standardfeil: $\widehat{\text{se}}(\hat{p}) = \sqrt{\frac{0.37 \cdot 0.63}{100}} \approx 0.0483$.

Feilmargin: $1.96 \cdot 0.0483 \approx 0.095$.

Konfidensintervall: $[0.37 - 0.095,\ 0.37 + 0.095] = [0.275,\ 0.465]$.

Rapportering:

Vi estimerer p til 0.37 *med et tilnærmet 95% KI* $[0.275, 0.465]$.

Wald-konfidensintervaller

Vi har sett to eksempler:

- μ med kjent σ:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- p i binomisk modell (normaltilnærming):

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Felles struktur:

- Estimator $\hat{\theta}$
- Estimert standardfeil $\widehat{\text{se}}(\hat{\theta})$
- Tilnærmet normalfordeling for $\hat{\theta}$

Wald-konfidensintervall

Anta:

- Vi ønsker å estimere en parameter θ
- Vi har en estimator $\hat{\theta}$ med

$$\hat{\theta} \approx N\big(\theta,\ [\widehat{\text{se}}(\hat{\theta})]^2\big),$$

  (ofte begrunnet via sentralgrenseteoremet)
- Vi kjenner/kan estimere standardfeilen $\widehat{\text{se}}(\hat{\theta})$

Da er **Wald-konfidensintervall med nivå** $(1-\alpha)100\%$ gitt ved

$$\hat{\theta} \ \pm \ z_{\alpha/2}\, \widehat{\text{se}}(\hat{\theta}).$$

Wald-metoden – fordeler og ulemper

Fordeler:

- Svært enkel formel
- Brukes mye i praksis
- Lett å generalisere til mange typer estimater

Ulemper:

- Kan gi dårlig dekning når:
  - utvalget er lite
  - fordelingen er sterkt skjev
  - parameteren er nær en grense (f.eks. p nær 0 eller 1)
- Kan til og med gi umulige verdier (f.eks. negativ sannsynlighet)

Hvorfor planlegge utvalgsstørrelse?

Konfidensintervaller har en **bredde** (feilmargin):

- Smalt intervall:
  - mer presis informasjon
  - lettere å trekke klare konklusjoner
- Bredt intervall:
  - stor usikkerhet
  - vanskelig å ta beslutninger

Vi kan stille spørsmålet:

*Hvor stort utvalg trenger vi for at feilmarginen skal være **høyest** e?*

Utvalgsstørrelse for middelverdi (kjent σ)

Vi bruker den enkleste settingen for illustrasjon: Anta $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ hvor σ^2 er kjent. $(1-\alpha)100\%$ konfidensintervall for μ er gitt ved

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Vi ønsker:

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq e.$$

Det gir:

$$\frac{\sigma}{\sqrt{n}} \leq \frac{e}{z_{\alpha/2}} \quad \Rightarrow \quad \sqrt{n} \geq \frac{z_{\alpha/2}\,\sigma}{e} \quad \Rightarrow \quad n \geq \left(\frac{z_{\alpha/2}\,\sigma}{e}\right)^2.$$

Eksempel: ønsket feilmargin for μ

- Vi ønsker et 95% konfidensintervall ($\alpha = 0.05$, $z_{0.025} \approx 1.96$)
- Vi antar at $\sigma = 10$
- Vi ønsker feilmargin $e = 2$ (maksimalt)

Beregning:

$$n \geq \left(\frac{1.96 \cdot 10}{2}\right)^2 = \left(\frac{19.6}{2}\right)^2 = 9.8^2 = 96.04.$$

Vi må altså ha $n = 97$.

Tolkning: Med $n \approx 100$ og kjent $\sigma = 10$ kan vi forvente et 95% konfidensintervall med feilmargin på omtrent 2.

Oppsummering

- **Konfidensintervall**:
  - intervall av plausible verdier for en ukjent parameter
  - konstruert slik at det dekker sannheten i en andel $(1 - \alpha)$ av gjentatte forsøk
- Normalmodell:
  - kjent σ: $\overline{x} \pm z_{\alpha/2}\, \sigma/\sqrt{n}$
  - ukjent σ: $\overline{x} \pm t_{n-1,\alpha/2}\, s/\sqrt{n}$
- Andeler (binomisk modell, normaltilnærming):
  - $\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$
- Wald-intervaller:
  - generell form: $\hat{\theta} \pm z_{\alpha/2}\, \widehat{\text{se}}(\hat{\theta})$
  - enkle og mye brukt, men kan være dårlige i små/ekstreme situasjoner
- Utvalgsstørrelse:
  - bredde på KI kan brukes til å planlegge hvor mange observasjoner vi trenger.

4

Når bruker vi *t*-fordeling i stedet for normalfordeling for å lage konfidensintervall for middelverdien?

(a) Når variansen er kjent
(b) Når variansen er ukjent og utvalget er lite
(c) Når utvalget er stort
(d) Når dataene er binomisk fordelt

5

Hva er en forutsetning for å bruke normalapproksimasjon for konfidensintervall for andelen p i en binomisk fordeling?

(a) n er stor og både np og $n(1-p)$ er større enn ca 10
(b) p er nær 0
(c) n er mindre enn 30
(d) Variansen er kjent

6

Hva er hovedforskjellen på et konfidensintervall for middelverdi med kjent og ukjent varians?

(a) Vi bruker *t*-fordeling ved ukjent varians
(b) Vi bruker alltid normalfordeling
(c) Intervallene blir alltid kortere med ukjent varians
(d) Vi trenger ikke utvalgsdata ved ukjent varians