

Oppsummering og repitisjon

TMA4245 – Parallellell 1

Geir-Arne Fuglstad, IMF

7. april 2025

Plan

- Repetisjon/oppsummering
- Mye slides
 - Få et overblikk
 - Se sammenhenger
 - Oppdag hva du må lese mer om
- Ikke prøv å skrive ned alt. Skriv ned hva du må øve mer på.
- Jeg tror jeg blir ferdig med alt i dag, hvis ikke gjør jeg ferdig på fredag.
- Fredag blir brukt til plenumsregning

Eksamen

- Onsdag 28. mai 2025 09:00—13:00
- Tillatte hjelpemidler
 - Tabeller og formuler i statistikk
 - Bestemt, enkel kalkulator
 - Hvitt A5-ark med egne notater (begge sider)
- Faglærer vil gå rundt på eksamen

Kode på eksamen

- Dere skal kunne tolke Python kode
- Dere skal kunne skrive enkel Python kode
 - Den trenger ikke være syntaktisk perfekt men den må bruke funksjoner og funksjonalitet som faktisk finnes
 - Dere trenger ikke å huske navnet perfekt på alle funksjoner
- Dere skal kunne tolke output fra kode

Se eksamenener høst 2023, vår 2024, sommer 2024 og høst 2024 for eksempler.

Noen råd til eksamen

- Regn tidligere eksamensoppgaver
 - Mai og august 2020 var bestått/ikke-bestått
 - November 2020, mai og august 2021 var hjemmeksamen
 - Noe endringer i pensum siden høst 2023
- Ta med tillatte hjelpemidler
 - IKKE kom uten formelsamling og tabeller!
 - IKKE kom uten kalkulator!
- Ofte teller deloppgaver like mye
 - Forskjellig vanskelighetsgrad og arbeidsmengde
- Les oppgaveteksten NØYE:
 - “Skriv opp” versus “Utled”
 - “Forklar” versus “bevis”
- MÅ være nok mellomregning til å forstå hva du har gjort
- Besvar alle (del)spørsmål

Grunnbegreper for mengder

- **Stokastisk forsøk:** Et forsøk der resultatet er underlagt tilfeldigheter
 - **Utfallsrom:** Mengden av alle mulige resultater, S
 - **Enkeltutfall:** Et mulig utfall, $e \in S$
 - **Hendelse:** Delmengde av utfallsrommet, $A \subset S$
 - **Venndiagram:** Illustrasjon av hendelser og operasjoner
-
- Operasjoner
 - **Snitt:** $A \cap B$
 - **Union:** $A \cup B$
 - **Komplement:** $A' = S \setminus A$
 - **Den tomme mengden:** $\emptyset$
 - **Disjunkte hendelser:** $A \cap B = \emptyset$

Sannsynlighet

- **Sannsynlighet**/sannsynlighetsmål: Reell funksjon på hendelser, P
 - $0 \leq P(A) \leq 1$
 - $P(S) = 1$
 - For parvis disjunkte hendelser $A_1, A_2, \dots$

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Eksempel (**Uniform sannsynlighetsmodell**):

$$P(A) = \frac{g}{m} = \frac{\text{gunstige}}{\text{mulige}}$$

Kombinatorikk

Se formelsamling side 32:

- $n! = 1 \cdot 2 \cdot \dots \cdot n$: antall måter å permutere n elementer
- ${}_nP_r = \frac{n!}{(n-r)!} = n \cdot (n-1) \cdot \dots \cdot (n-r+1)$: antall ordnede utvalg av r elementer blant n uten tilbakelegging
- $\binom{n}{r} = \frac{n!}{r!(n-r)!} = {}_nC_r$: antall uordnede utvalg av r elementer blant n uten tilbakelegging

Grunnleggende regneregler

- **Addisjonssetningen:** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- **Komplementærsetning:** $P(A') = 1 - P(A)$
- **Betinget sannsynlighet:** $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- **Multiplikasjonssetningen:** $P(A \cap B) = P(A)P(B|A)$

NB:

- Disjunkte: $P(A \cup B) = P(A) + P(B)$
- Uavhengige: $P(A \cap B) = P(A)P(B)$

Gjør deg kjent med hva som finnes i formelsamlingen på side 32

Total sannsynlighet

Se side 32 i formelsamling

- **Bayes' regel:** $P(A|B) = \frac{P(A)P(B|A)}{P(B)}$

- **Setning om total sannsynlighet:**

$$P(A) = \sum_{i=1}^n P(A \cap B_i) = \sum_{i=1}^n P(B_i)P(A|B_i)$$

der $B_1, B_2, \dots, B_n$ er en **partisjon** av utfallsrommet S .

Stokastiske variabler

- **Stokastisk variabel:** Reell funksjon på utfallsrommet

Disse kan være:

- **Diskret:** Endelig eller tellbart uendelig mange mulige verdier
- **Kontinuerlig:** Et kontinuum av mulige verdier. Ikke-tellbart uendelig mange mulige verdier.

Diskrete stokastiske variabler

- **Punktsannsynlighet (PMF):** $f(x) = P(X = x)$
- **Kumulativ fordelingsfunksjon (CDF):** $F(x) = P(X \leq x)$

Kontinuerlige stokastiske variabler

- **Sannsynlighetstetthet (PDF):** $f(x)$ (slik at $P(a < X < b) = \int_a^b f(x)dx$)
- **Kumulativ fordelingsfunksjon (CDF):** $F(x) = P(X \leq x)$

Simultanfordeling

- **Simultan punktsannsynlighet:**

$$f(x, y) = P(X = x, Y = y)$$

- **Simultan sannsynlighetstetthet:**

$$f(x, y) \text{ slik at } P((X, Y) \in A) = \iint_A f(x, y) dx dy$$

- **Marginalfordeling:** Summer eller integrer bort én variabel fra simultanfordeling
- **Betinget fordeling:**

$$f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

- X og Y er **uavhengige** hvis og bare hvis

$$f_{XY}(x, y) = f_X(x)f_Y(y) \text{ for alle } x \text{ og } y$$

Stokastisk simulering

- **Stokastisk simulering:** Å lage pseudo-tilfeldige tall som oppfører seg som om de kom fra en angitt fordeling

- **“Invers-CDF-metode”:** Dere skal kjenne til

$$U \sim \text{Unif}[0,1] \text{ betyr } X = F_X^{-1}(U) \sim f_X(x)$$

- **Monte Carlo (simulering):** Vi kan estimere parametere med uavhengige simuleringer,

$$E[\widehat{g(X)}] = \frac{1}{n} \sum_{i=1}^n g(X_i),$$

der $X_1, \dots, X_n \sim_{\text{iid}} f_X(x)$. Merk at dette inkluderer

- Sannsynligheter: $P(X \in A) = E[\mathbb{I}(X \in A)]$
- Varianser: $\text{Var}[g(X)] = E[(g(X) - E[g(X)])^2]$
- Sannsynligetsfordelinger: Histogrammer

Forventningsverdi og varians

For $(X, Y) \sim f(x, y)$ har vi

- **Forventningsverdi:** “gjennomsnitt av uendelig mange simuleringer”

$$\mu = E[X] = \begin{cases} \sum_x xf(x), & \text{diskret,} \\ \int_{-\infty}^{\infty} xf(x)dx, & \text{kontinuerlig.} \end{cases}$$

- **Varians:** “Mål på (kvadratisk) variasjon”

$$\text{Var}[X] = E[X^2] - E[X]^2$$

- **Kovarians:** (formelsamling side 33)

$$\text{Cov}[X, Y] = E[XY] - E[X]E[Y]$$

- **Korrelasjon:** (formelsamling side 33)

Regneregler

Se side 34 i formelsamling.

- Forventning er lineær

$$E[a + bX + cY] = a + bE[X] + cE[Y]$$

- Varians krever uavhengighet eller kjent kovarians

$$\text{Var}[a + bX + cY] = b^2\text{Var}[X] + c^2\text{Var}[Y] + 2bc\text{Cov}[X, Y]$$

Viktige fordelinger

Formelsamling har viktig informasjon om alle fordelinger i kurset.

Noen viktige diskrete fordelinger:

- **Binomisk fordeling**
- **Negativ-binomisk fordeling (og geometrisk fordeling)**
- **Poissonfordeling**

Noen viktige kontinuerte fordelinger

- **Normalfordeling**
- **Gammafordeling (og eksponentialfordeling og kjikvadratfordeling)**
- **Student t-fordeling**

Viktige sammenhenger

- For normalfordeling bruker vi hele tiden

$$X \sim N(\mu, \sigma^2) \text{ betyr } Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

- Hvis $Z \sim N(0,1)$ og $V \sim \chi_n^2$ er uavhengige så gjelder

$$\frac{Z}{\sqrt{V/n}} \sim t_n$$

- Binomisk fordeling kan tilnærmes med poissonfordeling hvis p er liten og n er stor.
- Binomisk fordeling kan tilnærmes med normalfordeling hvis n er stor (tommelfingerregel: $np \geq 5$ og $n(1 - p) \geq 5$)
- Eksponentialfordeling er tid til første hendelse i poissonprosess
- Poissonfordeling er antall hendelser i gitt tidsintervall i poissonprosess

Viktige tabeller

- Tabell for CDF
 - **Standardnormalfordeling**
 - **Binomisk fordeling**
 - **Poissonfordeling**
- Tabell for kritiske verdier
 - **Standardnormalfordeling**
 - **t-fordeling**
 - **Kjikvadratfordeling**

Sentralgrenseteoremet

Hvis $X_1, X_2, \dots, X_n \sim_{\text{iid}} f(x)$, der $\mu = E[X_i]$ og $\sigma^2 = \text{Var}[X_i]$ for $i = 1, \dots, n$, da gjelder

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \rightarrow N(0,1)$$

når $n \rightarrow \infty$.

Den praktiske konsekvensen er at vi er villig til å anta

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

Tommelfingerregel: $n \geq 30$

Se side 38 i formelsamling.

Transformasjoner av stokastiske variabler

Se side 34 i formelsamling.

Anta at $X \sim f_X(x)$ og at g er en strengt monoton reell funksjon.

Da gjelder:

- X diskret, betyr $Y = g(X)$ diskret og

$$f_Y(y) = f_X(g^{-1}(y))$$

- X kontinuertlig og g deriverbar, betyr $Y = g(X)$ kontinuertlig og

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}}{dy}(y) \right|$$

NB: Den siste reglen ser litt rar ut, men er analog med skifte av integrasjonsvariabel i Matematikk 1

Feilforplantningsformelen

Antagelser:

- X, Y og Z uavhengige stokastiske variabler
- $g: \mathbb{R}^3 \rightarrow \mathbb{R}$ er en funksjon (med nok deriverte)

Da approksimerer vi $V = g(X, Y, Z)$ med Taylors formel og får:

- $E[V] \approx g(\mu_X, \mu_Y, \mu_Z)$
- $$\begin{aligned} \text{Var}[V] \approx & \left(\frac{\partial g}{\partial x}(\mu_X, \mu_Y, \mu_Z) \right)^2 \text{Var}[X] + \left(\frac{\partial g}{\partial y}(\mu_X, \mu_Y, \mu_Z) \right)^2 \text{Var}[Y] \\ & + \left(\frac{\partial g}{\partial z}(\mu_X, \mu_Y, \mu_Z) \right)^2 \text{Var}[Z] \end{aligned}$$

Dette kalles gjerne **Gauss feilforplantningslov**. Indirekte beskrevet på side 35 i formelsamling.

Momentgenererende funksjoner

Side 33 i formelsamling: $M_X(t) = E[\exp(tX)]$

Viktig regel: For uavhengige $X_1, X_2, \dots, X_n$, gjelder

$$M_{\sum_{i=1}^n a_i X_i}(t) = \prod_{i=1}^n M_{X_i}(a_i t)$$

Brukes som regel til:

1. Finn et moment $E[X^r] = M_X^{(r)}(0)$ for $r = 0, 1, 2, \dots$
2. Finn fordelingen til $Y = \sum_{i=1}^n a_i X_i$ basert på viktig regel over gjennom
 $X \sim Y$ hvis og bare hvis $M_X(t) = M_Y(t)$ i et intervall rundt 0

Viktige resultater som kan vises fra momentgenererende funksjoner

- Lineær kombinasjon av uavhengige normalfordelte variabler er normalfordelt
- Lineær kombinasjon av uavhengige poissonfordelte variabler er poissonfordelte
- Sum av uavhengige kjikvadratfordelinger er kjikvadratfordelte

Se side 35 i formelsamlingen

Populasjoner og utvalg

- En $f(x; \theta)$ -**populasjon** har ukjente deterministiske **parametere** θ (som vi har lyst til å lære)
- Et **tilfeldig utvalg** av størrelse n fra $f(x; \theta)$ -populasjonen er
$$X_1, X_2, \dots, X_n \sim_{\text{iid}} f(x; \theta)$$
- Basert på det tilfeldige utvalget kan vi regne ut **observatorer** som ikke inkluderer ukjente parametere
- Hvis observatoren $\hat{\Theta}(X_1, \dots, X_n)$ brukes til å anslå verdien av parameteren θ , kalles $\hat{\Theta}$ en **estimator**. F.eks.:
 - $\hat{\Theta} = \bar{X}$ er en estimator for $\theta = \mu$
 - $\hat{\Theta} = S^2$ er en estimator for $\theta = \sigma^2$

NB: Estimatoren $\hat{\Theta}$ er en stokastisk variabel, og det observerte estimatet $\hat{\theta}$ er et deterministisk tall

Egenskaper til estimatorer

For en estimator $\hat{\theta}$ for θ er vi interessert i

- Forventningsverdi: $E[\hat{\theta}]$
- Varians: $\text{Var}[\hat{\theta}]$

Med **beste estimator** mener vi i dette kurset:

1. **Estimatoren er forventningsrett**, $E[\hat{\theta}] = \theta$
2. **Estimatoren er mest effisient**. Variansen, $\text{Var}[\hat{\theta}]$, er lavest mulig

Dvs. hvis vi har flere estimatorer sjekker vi

1. Hvilke estimatorer er forventningsrette?
2. For de forventningsrette estimatorene, hvilken har lavest varians?

Sannsynlighetsmaksimeringsestimatoren (SME)

Prinsipp: Finn parameterene θ som maksimerer rimeligheten til det du har observert.

Observasjoner: Trekkes fra populasjonen, $(X_1, \dots, X_n) \sim f(x_1, \dots, x_n; \theta)$

Rimelighetsfunksjon: Simultanfordelingen til observasjonene som en funksjon av parameterene θ
$$L(\theta; x_1, \dots, x_n) = f(x_1, \dots, x_n; \theta)$$

Log-rimelighetsfunksjon: $l(\theta; x_1, \dots, x_n) = \log(L(\theta; x_1, \dots, x_n))$

Matematisk: $\hat{\theta} = \operatorname{argmax}_{\theta} \{l(\theta; x_1, \dots, x_n)\}$ og $\hat{\Theta} = \operatorname{argmax}_{\theta} \{l(\theta; X_1, \dots, X_n)\}$

Tilfeldig utvalg: Som regel har vi et tilfeldig utvalg slik at

$$l(\theta; x_1, \dots, x_n) = \log\left(\prod_{i=1}^n f(x_i; \theta)\right) = \sum_{i=1}^n \log(f(x_i; \theta))$$

Konfidensintervall

Usikkerheten i et estimat må baseres på fordelingen til estimatoren under gjentatte forsøk

Definisjon:

Anta $X_1, X_2, \dots, X_n \sim_{\text{iid}} f(x; \theta)$ og $0 < \alpha < 1$. Da er

$$[\hat{\theta}_L(x_1, \dots, x_n), \hat{\theta}_U(x_1, \dots, x_n)]$$

et $100(1 - \alpha)\%$ -konfidensintervall for θ hvis

$$P\left(\hat{\theta}_L(X_1, \dots, X_n) \leq \theta \leq \hat{\theta}_U(X_1, \dots, X_n)\right) = 1 - \alpha.$$

Vi skriver at $\hat{\theta}_L(x_1, \dots, x_n) \leq \theta \leq \hat{\theta}_U(x_1, \dots, x_n)$ er et $100(1 - \alpha)\%$ -konfidensintervall for θ .

Tolkning:

Det var $100(1 - \alpha)\%$ sannsynlighet for at intervallet skulle inneholde θ før vi gjorde det tilfeldige utvalget.

Med andre ord vil $100(1 - \alpha)\%$ av intervaller vi lager på denne måten inneholde θ .

Utleddning av $100(1 - \alpha)\%$ konfidensintervall

1. Har tilfeldig utvalg: $X_1, \dots, X_n \sim_{\text{iid}} f(x; \theta)$
2. Velger **pivotal** $Z = g(X_1, \dots, X_n, \theta)$
 - Kun θ ukjent
 - Z typisk basert på en estimator for θ
3. Starter med et sannsynlighetsuttrykk
$$P(a \leq g(X_1, \dots, X_n, \theta) \leq b) = 1 - \alpha$$
der a og b typisk er kritiske verdier vi må finne i tabell.
4. Løser $a \leq g(X_1, \dots, X_n, \theta) \leq b$ til en ulikhet
$$\hat{\Theta}_L(X_1, \dots, X_n) \leq \theta \leq \hat{\Theta}_U(X_1, \dots, X_n)$$
5. Konkluder at $100(1 - \alpha)\%$ konfidensintervall for θ er
$$\hat{\theta}_L(x_1, \dots, x_n) \leq \theta \leq \hat{\theta}_U(x_1, \dots, x_n)$$

Standard eksempler

- Ett-utvalg normalfordeling
 - μ og σ^2 **ukjente**
 - Intervall for μ : $T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$
 - Intervall for σ^2 : $V = \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$
 - μ **ukjent** og σ^2 **kjent**
 - Intervall for μ : $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$
 - μ **kjent** og σ^2 **ukjent**
 - Intervall for σ^2 : $V = \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi_n^2$
- To-utvalg normalfordeling
 - $\mu_1 - \mu_2$ **ukjent**, og σ_1^2 og σ_2^2 **kjente**
 - Intervall for $\mu_1 - \mu_2$: $Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
- Ett-utvalg, binomisk fordeling
 - p **ukjent**, typisk $X \sim \text{Binom}(n, p)$ eller tilnærming $X \sim N(np, np(1 - p))$
- Regresjonsmodeller
 - $\alpha, \beta, \alpha + \beta x_0, \sigma^2$ på lignende måte som over, men OBS på antall frihetsgrader når σ^2 **ukjent**

Prediksjonsintervall

Antagelser:

- $X_1, X_2, \dots, X_n$ trekkes fra simultanfordelingen $f(x_1, \dots, x_n; \theta)$
- X^* trekkes fra $g(x^*; \theta)$
- $0 < \alpha < 1$

Prediksjonsintervall:

Intervall

$$[\hat{X}_L(x_1, \dots, x_n), \hat{X}_U(x_1, \dots, x_n)]$$

er et $100(1 - \alpha)\%$ -prediksjonsintervall for X^* hvis

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

Tolkning:

Det var $100(1 - \alpha)\%$ sannsynlighet for at intervallet skulle inneholde x^* før vi trakk $x_1, \dots, x_n$ **og** x^* . Med andre ord vil $100(1 - \alpha)\%$ av intervaller vi lager på denne måten inneholde x^* .

Typisk framgangsmåte prediksjonsintervall

For $Y^* \sim N(\mu, \sigma^2)$ har vi lyst til å bruke $Z = \frac{Y^* - \mu}{\sigma} \sim N(0,1)$

Hvis vi ikke kjenner μ må vi bytte til estimator $\hat{\mu}$, $Z = \frac{Y^* - \hat{\mu}}{\sqrt{\sigma^2 + \text{Var}[\hat{\mu}]}}$. F.eks:

- $Z = \frac{Y^* - \bar{Y}}{\sigma \sqrt{1 + 1/n}} \sim N(0,1)$
- $Z = \frac{Y^* - \hat{A} - \hat{B}x^*}{\sigma \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}} \sim N(0,1)$

Hvis vi i tillegg ikke kjenner σ^2 må vi bytte til S^2 . F.eks.:

- $T = \frac{Y^* - \bar{Y}}{S \sqrt{1 + 1/n}} \sim t_{n-1}$, mistet en frihetsgrad ved å estimere μ
- $T = \frac{Y^* - \hat{A} - \hat{B}x^*}{S \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$, mistet to frihetsgrader ved å estimere α og β

Hypotesetester – bakgrunn

- To muligheter
 - H_0 , **nullhypotesen**, er dagens tilstand som vi skal vurdere om det er grunnlag for å forkaste
 - H_1 , **alternativ hypotese**, er det vi skal vurdere om det er grunnlag for å påstå
- Utgangspunktet vårt er “uskyldig til bevist skyldig”. Altså foretrekker H_0 med mindre “bevisene” for H_1 er sterke nok.
- To typer feil
 - **Type I feil**: Forkaster H_0 selv om den er sann (dømmer uskyldig person)
 - **Type II feil**: Forkaster ikke H_0 selv om H_1 er sann (dømmer ikke skyldig person)
- **Signifikansnivået** til en hypotesetest er $\alpha = P(\text{Type I feil})$.

Hypotesetester – utledning

1. Formuler nullhypotese H_0 og alternativ hypotese H_1
2. Velg en testobservator T som har kjent fordeling under H_0
3. Bestem en beslutningsregel. For eksempel:
 - a) “Forkast H_0 hvis $T \geq k$ ”,
 - b) “Forkast H_0 hvis $T \leq k$ ”,
 - c) eller “Forkast H_0 hvis $T \leq k_L$ eller $T \geq k_R$ ”
4. Velg signifikansnivå $0 \leq \alpha \leq 1$
5. Regn ut kritisk verdi k (eller k_L og k_R) for å oppfylle signifikansnivå, og skriv opp beslutningsregelen
6. Regn ut observert verdi t og konkluder

P-verdier

Definisjon:

P-verdien er sannsynligheten, når H_0 er sann, for å observere en verdi av testobservatoren som er lik den observerte verdien eller en verdi som er mer ekstrem i retning av den alternative hypotesen.

Tolkning:

Det høyeste verdien av signifikansnivået α der konklusjonen ville være å forkaste H_0 .

Praktisk:

En svakhet ved hypotesetester er at konklusjonen “ikke forkast” er lite informativ om hvor sterke “beviser” vi hadde mot H_0 . Med p-verdien får vi mer informasjon enn “ikke forkast” ville gitt.

NB:

En-sidig: $p = P(T \geq t)$ eller $p = P(T \leq t)$

To-sidig: $p = 2P(T \geq k)$ eller $p = 2P(T \leq k)$

Standard testobservatorer

Bruk lignende pivotaler som for konfidensintervall, men sett inn kjent verdi $\theta = \theta_0$ under $H_0: \theta = \theta_0$

- Eksempler for ett-utvalg normalfordeling

- μ og σ^2 **ukjente**

- Hypotesetest for μ : $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$
 - Hypotesetest for σ^2 : $V = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi_{n-1}^2$

- μ **ukjent** og σ^2 **kjent**

- Hypotesetest for μ : $Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim N(0,1)$

- μ **kjent** og σ^2 **ukjent**

- Hypotesetest for σ^2 : $V = \frac{n\hat{\sigma}^2}{\sigma_0^2} \sim \chi_n^2$

Teststyrke

Typisk prosedyre:

1. Utled testen og kom frem til forkastningskriterium (f.eks. $T \geq k$)
2. Anta at alternativ verdi $\theta = \theta_1$ er sann (altså ikke lenger H_0)
3. Regn ut sannsynligheten for å forkaste (f.eks. $P(T \geq k | \theta = \theta_1)$).

Styrken er altså sannsynligheten for å forkaste nullhypotesen når $\theta = \theta_1$,
$$\text{styrke}(\theta_1) = P(\text{«forkast } H_0\text{»} | \theta = \theta_1)$$

Typisk oppgaveformulering:

Vi har en test med signifikansnivå α . Bestem utvalgsstørrelse n slik at styrken til testen er minst 95% når $\theta = \theta_1$

Andre ikke-parametriske hypotesetester

- **Fortegnstesten:** en test på medianen til fordeling. Bruker tilfeldig utvalg, men krever ikke antagelse om en bestemt fordeling.
- **Tosidig permutasjonstest:** sammenligner to grupper med tall basert på hvor “ekstrem” en forskjell i egenskaper er mellom de to gruppene sammenlignet med tilfeldig permutasjon under nullhypotesen om at gruppene er “like”.

Enkel lineær regresjon

Utgangspunkt:

Vi ser på Y som stokastisk og antar en lineær avhengighet på x i forventningsverdi

$$E[Y] = \alpha + x\beta, \quad \text{for } x \text{ i et rimelig område.}$$

Observasjoner:

Vi samler inn n målinger $(x_1, y_1), \dots, (x_n, y_n)$ ved **kjente** $x_1, \dots, x_n$ og **stokastiske** $Y_1, \dots, Y_n$, og antar

1. homogen varians, $\text{Var}[Y_i] = \sigma^2$, $i = 1, \dots, n$.
2. ukorrelerte målinger, $\text{Cov}[Y_i, Y_j] = 0$, $i \neq j$.
3. Normalfordelte målefeil

Oppsummert:

$$Y_i = \alpha + x_i\beta + \epsilon_i, \quad i = 1, \dots, n,$$

med antagelser om residualene

1. forventning 0
2. varians σ^2
3. ukorrelerte
4. normalfordelte.

Dere må vite

Estimatorene (side 36 i formelsamlingen):

$$\hat{B} = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{A} = \bar{Y} - \hat{B} \bar{x} \quad \text{og} \quad S^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{A} - \hat{B} \bar{x})^2$$

Fordelingene ($\hat{A}$ og $\hat{B}$ står ikke i formelsamlingen):

$$\hat{A} \sim N\left(\alpha, \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \sigma^2\right), \quad \hat{B} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \quad \text{og} \quad \frac{(n-2)S^2}{\sigma^2} \sim \chi_{n-2}^2$$

NB: Dere trenger normalfordeling på residualer for å få disse fordelingene. Dere trenger ikke utlede fordelingen til S^2 .

Egenskaper (disse skal typisk ikke utledes):

- S^2 er uavhengig av $\hat{A}$ og $\hat{B}$ (**NB:** side 36 i formelsamling)
- $\bar{Y}$ og $\hat{B}$ er ukorrelererte/uavhengige (**NB:** Står ikke i formelsamlingen og er veldig tungvint å vise med teknikkene dere kjenner)

Utleddning av estimatorer

Sannsynlighetsmaksimeringsestimatoren:

- krever normalfordelte residualer
- Standard fremgangsmåte, men krever at man løser for alle partiellderiverte er null samtidig
- Gir estimatorer for α , β og σ^2

Minste-kvadraters-metode:

- baserer seg bare på å minimere

$$\text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{a} - \hat{b}x_i)^2$$

- gir ikke en estimator for varians, men at vi velger

$$s^2 = \frac{\text{SSE}}{n - 2}$$

fordi den er forventningsrett

NB

- Denne gjennomgangen har **ikke** dekket alt
- Hvis det var noen konsepter dere ikke kunne, **øv på dem!**

Hvilket kurs burde jeg ta nå?

Disse fagene bruker typisk R (og ikke Python)

- TMA4265 – Stokastisk modellering
 - Mer sannsynlighet
 - Prosesser som utvikler seg stokastisk over tid
- TMA4267 – Lineære statistiske modeller
 - Multippel lineær regresjon med matriser og vektorer
 - Kan vise en del ting mye enklere enn vi har gjort
- TMA4268 – Statistisk læring
 - Regresjon, klassifikasjon, regularisering, og ikke-lineære sammenhenger
 - Mange metoder, og mye fokus på bruk av ulike metoder