Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

# TMA4240 Statistikk
# Høst 2018

**Anbefalt øving 11**
**Løsningsskisse**

**Oppgave 1**

**a**) En rimelig estimator for forventningsverdien $\mu$ er gjennomsnittet

$$\bar{X} = \frac{1}{5}\sum_{i=1}^{5} X_i,$$

som vil være normalfordelt med forventningsverdi $E(\bar{X}) = \mu$ og varians $\text{Var}(\bar{X}) = \sigma^2/5$. En rimelig estimator for variansen er

$$S^2 = \frac{1}{4}\sum_{i=1}^{5}(X_i - \bar{X})^2,$$

som har forventningsverdi $E(S^2) = \sigma^2$. Observasjonene $x_1, \ldots, x_5$ i tabellen gir estimatene

$$\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = \underline{\underline{4.9540}} \quad \text{og} \quad s^2 = \frac{1}{4}\sum_{i=1}^{5}(x_i - \bar{x}) = \underline{\underline{0.3440}}.$$

**b**) For å utlede et konfidensintervall for $\mu$ tar vi utgangspunkt i den tilfeldige variabelen

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/5}} \sim N(0,1)$$

som er standard normalfordelt, siden $\bar{X} \sim N(\mu, \sigma^2/5)$. Når den ukjente variansen $\sigma^2$ byttes ut med estimatoren $S^2$, får vi observatoren

$$T = \frac{\bar{X} - \mu}{\sqrt{S^2/5}} \sim t_4$$

som er $t$-fordelt med $5 - 1 = 4$ frihetsgrader. Vi ønsker et 95% konfidensintervall for $\mu$, og trenger derfor 0.975-kvantilen i $t$-fordelingen med 4 frihetsgrader, som er $t_{0.025,4} = 2.7764$. Konfidensintervallet kan nå konstrueres som følger,

$$P(-t_{0.025,4} \le T \le t_{0.025,4}) = 1 - 2 \cdot 0.025 = 0.95$$

$$P\left(-t_{0.025,4} \le \frac{\bar{X} - \mu}{\sqrt{S^2/5}} \le t_{0.025,4}\right) = 0.95$$

$$P\left(\bar{X} - t_{0.025,4} \cdot \sqrt{\frac{S^2}{5}} \le \mu \le \bar{X} + t_{0.025,4} \cdot \sqrt{\frac{S^2}{5}}\right) = 0.95.$$

Når tallverdier settes inn får vi intervallet

$$\bar{x} \pm t_{0.025,4} \cdot \sqrt{\frac{s^2}{5}} = 4.9540 \pm 2.7764 \cdot \sqrt{\frac{0.3440}{5}} = \underline{\underline{[4.2258, 5.6822]}}.$$

Anta at vi skal teste nullhypotesen $H_0 : \mu = \mu_0$ mot den alternative hypotesen $H_1 : \mu \neq \mu_0$. Vi bruker testobservatoren

$$T_0 = \frac{\bar{X} - \mu_0}{\sqrt{S^2/5}},$$

som i likhet med $T$ er $t$-fordelt med 4 frihetsgrader (for $\mu_0 = \mu$ har vi $T_0 = T$). På signifikansnivå 5% vil vi beholde $H_0$ hvis vi observerer $-t_{0.025,4} \leq T_0 \leq t_{0.025,4}$. Ellers forkastes $H_0$. Akseptansekriteriet for $H_0$ er dermed

$$-t_{0.025,4} \leq \frac{\bar{X} - \mu_0}{\sqrt{S^2/5}} \leq t_{0.025,4}$$

eller, om vi isolerer $\mu_0$ i midten,

$$\bar{X} - t_{0.025,4} \cdot \sqrt{\frac{S^2}{5}} \leq \mu_0 \leq \bar{X} + t_{0.025,4} \cdot \sqrt{\frac{S^2}{5}},$$

som er identisk med 95% konfidensintervallet over. For en gitt verdi av $\mu_0$ kan altså konfidensintervallet brukes til å teste $H_0$ mot $H_1$ på signifikansnivå 5%, ved å beholde nullhypotesen kun dersom $\mu_0$ er inneholdt i intervallet.

**Oppgave 2**

**a**) Sannsynligheten for å få 5 kron er

$$P(5 \text{ kron}) = \frac{1}{2^5} = 1/32 = \underline{\underline{0.031}}.$$

Sannsynligheten for å få 3 kron er lik punktsannsynligheten $P(X = 3)$ der $X$ er binomisk fordelt med parametre $n = 5$ og $p = 0.5$, altså

$$P(X = 3) = \binom{5}{3} 0.5^3 \cdot (1 - 0.5)^{5-3} = 10 \cdot 0.5^3 \cdot 0.5^2 = \underline{\underline{0.3125}}.$$

Fire kron på rad kan inntreffe på 3 forskjellige måter: Kron på alle 5 kastene, kron på de første 4 kastene, og mynt på siste, eller mynt på første kast og kron på de 4 siste. Antall mulige utfall av de fem kastene er $2^5 = 32$, og alle er like sannsynlige, så sannsynligheten for å få fire kron på rad er

$$P(4 \text{ kron på rad}) = \frac{3}{32} = \underline{\underline{0.0938}}.$$

**b**) Sannsynligheten for at lengste sekvens har lengde 5 eller 6 kan anslås ved å regne ut andelen utfall hvor lengste sekvens var på 5 eller 6 kast, av de 10000 simulasjonene. Fra figuren leser vi av at lengste sekvens hadde lengde 5 i omtrent 2700 tilfeller, og lengde 6 i omtrent 1700 tilfeller, og vi får estimatet

$$\widehat{P(5 \text{ eller } 6)} = \frac{2700 + 1700}{10000} = \underline{\underline{0.44}}.$$

I Miriams myntkastsekvens har den lengste uavbrutte sekvensen av kron lengde 2. For en tilfeldig generert myntkastsekvens av lengde 30, vil lengden av lengste uavbrutte sekvens av kron ha en sannsynlighetsfordeling som er svært lik den i figuren. At denne lengden er så lav som 2 er ganske usannsynlig, og Miriams myntkastsekvens er dermed mistenkelig.

Vi vil teste nullhypotesen

$$H_0: \text{ Sekvensen er tilfeldig generert}$$

mot den alternative hypotesen

$$H_1: \text{ Sekvensen er ikke tilfeldig generert.}$$

Vi antar at under nullhypotesen er lengden av lengste sammenhengende sekvens av kron fordelt som i figuren. For å avgjøre om nullhypotesen skal forkastes eller ikke, regner vi ut $p$-verdien, altså sannsynligheten for å observere et like ekstremt eller mer ekstremt utfall. Her er dette lik sannsynligheten for at lengste uavbrutte sekvens av kron er 0, 1 eller 2. Utfra figuren ser det ut som om antall utfall i søylene for 0, 1 og 2 er henholdsvis 0, 0 og 25. Vi får dermed følgende estimat for $p$-verdien:

$$\widehat{P(0, 1 \text{ eller } 2)} = \frac{25}{10000} = 0.0025.$$

Dette er en lav $p$-verdi som tilsier at nullhypotesen forkastes f.eks. på signifikansnivå 0.05. Det er altså grunn til å hevde at Miriam har funnet på tallene.

## Oppgave 3

**a**) For at $X$ skal være tilnærmet binomisk fordelt må en ha at $N >> n$ slik at det ikke betyr noe at samme person ikke vil bli spurt mer enn en gang.

$$P(X = 9) = \binom{n}{9} \theta^9 (1-\theta)^{n-9} = \binom{20}{9} 0.5^9 (1-0.5)^{20-9} = \underline{\underline{0.1602}}$$

$$P(X > 9) = 1 - P(X \leq 9) \overset{\text{tabell}}{=} 1 - 0.412 = \underline{\underline{0.588}}$$

$$\begin{aligned} P(X > 9 | X \leq 12) &= \frac{P(X > 9 \cap X \leq 12)}{P(X \leq 12)} = \frac{P(10 \leq X \leq 12)}{P(X \leq 12)} \\ &= \frac{P(X \leq 12) - P(X \leq 9)}{P(X \leq 12)} \overset{\text{tabell}}{=} \frac{0.868 - 0.412}{0.868} = \underline{\underline{0.525}} \end{aligned}$$

**b**) $X \sim b(x; n, \theta)$

$$L(\theta) = P(X = x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

$$l(\theta) = ln \binom{n}{x} + xln\theta + (n-x)ln(1-\theta)$$

$$l'(\theta) = x\frac{1}{\theta} + (n-x)\frac{1}{1-\theta}(-1) = \frac{x}{\theta} - \frac{n-x}{1-\theta}$$

$$l'(\theta) = 0$$
$$\Downarrow$$
$$\frac{x}{\theta} = \frac{n-x}{1-\theta}$$
$$x(1-\theta) = \theta(n-x)$$
$$x - x\theta = \theta n - \theta x$$
$$\theta = \frac{x}{n}$$

dvs. SME er $\underline{\underline{\hat{\theta} = \frac{X}{n}}}$

$$\mathrm{E}(\hat{\theta}) = \mathrm{E}\left(\frac{X}{n}\right) = \frac{1}{n}\mathrm{E}(X) = \frac{1}{n}n\theta = \underline{\theta}$$

$$\mathrm{Var}(\hat{\theta}) = \mathrm{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2}\mathrm{Var}(X) = \frac{1}{n^2}n\theta(1-\theta) = \underline{\underline{\frac{\theta(1-\theta)}{n}}}$$

**c**) $H_0 : \theta = \frac{1}{2}$ mot $H_1 : \theta \neq \frac{1}{2}$

Benytter testobservator

$$Z = \frac{X - n\frac{1}{2}}{\sqrt{n\frac{1}{2}(1-\frac{1}{2})}} \approx N(z; 0, 1)$$

når $H_0$ er riktig.

Dvs. forkaster $H_0$ dersom

$$Z < -z_{\frac{\alpha}{2}} \quad \text{eller} \quad Z > z_{\frac{\alpha}{2}}$$
$$\Updownarrow$$
$$\frac{X - \frac{n}{2}}{\frac{\sqrt{n}}{2}} < -z_{\frac{\alpha}{2}} \quad \text{eller} \quad \frac{X - \frac{n}{2}}{\frac{\sqrt{n}}{2}} > z_{\frac{\alpha}{2}}$$

Innsatt tall:

Observert verdi av testobservator:

$$Z = \frac{2562 - \frac{5000}{2}}{\frac{\sqrt{5000}}{2}} = 1.75$$

$\alpha = 0.10$ gir $z_{\frac{\alpha}{2}} = 1.645$

Dvs. forkaster $H_0$ og erklærer G som vinner av valget.

**Oppgave 4**

**a**) Forventingsverdien til $\hat{\alpha}$ er

$$E(\hat{\alpha}) = E\left(\frac{1}{8}\sum_{i=1}^{8} D_i\right) = \frac{1}{8}\sum_{i=1}^{8} E(D_i) = \frac{8\alpha}{8} = \alpha,$$

hvor vi bruker at $E(D_i) = E(\alpha + \epsilon_i) = E(\alpha) + E(\epsilon_i) = \alpha$, fordi $E(\epsilon_i) = 0$. Variansen til $\hat{\alpha}$ er

$$\mathrm{Var}(\hat{\alpha}) = \mathrm{Var}\left(\frac{1}{8}\sum_{i=1}^{8} D_i\right) = \frac{1}{8^2}\sum_{i=1}^{8} \mathrm{Var}(D_i) = \frac{8\sigma_1^2}{8^2} = \frac{\sigma_1^2}{8},$$

hvor vi bruker at $\mathrm{Var}(D_i) = \mathrm{Var}(\alpha+\epsilon_i) = \mathrm{Var}(\alpha)+\mathrm{Var}(\epsilon_i) = \sigma_1^2$, siden $\alpha$ er en konstant, og $\mathrm{Var}(\epsilon_i) = \sigma_1^2$.

Dybdemålingene $D_1, D_2, \ldots, D_8$ er normalfordelte. Estimatoren $\hat{\alpha}$ er en lineærkombinasjon av disse. Derfor er $\hat{\alpha}$ normalfordelt, $\hat{\alpha} \sim N(\alpha, \sigma_1^2/8)$. Merk at selv om vi vet at estimatoren er normalfordelt, så kjenner vi ikke parametrene i fordelingen, derfor kan vi ikke ta utgangspunkt i normalfordelingen hvis vi skal gjøre inferens om $\alpha$. Det at variansen $\sigma_1^2$ er ukjent, gjør at vi i stedet må basere oss på $t$-fordelingen.

**b**) Vi ønsker et 95% konfidensintervall for dypet $\alpha$. Vi vet fra **a)** at estimatoren $\hat{\alpha}$ er normalfordelt med forventningsverdi $\alpha$ og varians $\sigma_1^2/8$, slik at den normaliserte variabelen

$$\frac{\alpha - E(\hat{\alpha})}{\sqrt{\mathrm{Var}(\hat{\alpha})}} = \frac{\hat{\alpha} - \alpha}{\sqrt{\sigma_1^2/8}}$$

er standard normalfordelt. Siden variansen $\sigma_1^2$ er ukjent bytter vi den ut med estimatoren

$$S^2 = \frac{1}{7}\sum_{i=1}^{8}(D_i - \hat{\alpha})^2 = \frac{1}{7}\left[\sum_{i=1}^{8} D_i^2 - \frac{1}{8}\left(\sum_{i=1}^{8} D_i\right)^2\right],$$

og får $T$-observatoren

$$T = \frac{\hat{\alpha} - \alpha}{\sqrt{S^2/8}}$$

som er $t$-fordelt med $8 - 1 = 7$ frihetsgrader. Et 95% konfidensintervall for $T$ gis da av

$$P(-t_{0.025,7} \le T \le t_{0.025,7}) = 1 - 2 \cdot 0.025 = 0.95$$

eller, uttrykt ved $\alpha$,

$$P(-t_{0.025,7} \le \frac{\alpha - \hat{\alpha}}{\sqrt{S^2/8}} \le t_{0.025,7}) = 0.95,$$

hvor $t_{0.025,7} = 2.3646$ er 0.975-kvantilen i $t$-fordelingen med 7 frihetsgrader. Punktestimatet for $\alpha$ og estimatet for variansen blir

$$\hat{\alpha} = \frac{1}{8}\sum_{i=1}^{8} d_i = \frac{22.44}{8} = 2.8050$$

og

$$s^2 = \frac{1}{7}\left[\sum_{i=1}^{8} d_i^2 - \frac{1}{8}\left(\sum_{i=1}^{8} d_i\right)^2\right] = \frac{1}{7}\left[63.0162 - \frac{22.44^2}{8}\right] = 0.0103$$

hvor $\sum_{i=1}^{8} d_i$ og $\sum_{i=1}^{8} d_i^2$ er hentet fra oppgaveteksten. Med disse tallverdiene får vi følgende 95% konfidensintervall for $\alpha$,

$$\hat{\alpha} \pm t_{0.025,7} \cdot \sqrt{\frac{s^2}{8}} = 2.8050 \pm 2.3646 \cdot \sqrt{\frac{0.0103}{8}} = [2.7202, 2.8898].$$

**c**) Vi forutsetter at bruddet er i posisjon $l_b = 10.0$, og kan derfor sette $l_i < l_b$ for de fem første observasjonene $D_1, \ldots D_5$, og $l_i \geq l_b$ for de tre siste observasjonene $D_6, \ldots, D_8$. Den antatte modellen kan dermed forenkles til

$$D_i = \begin{cases} \alpha + \epsilon_i & \text{for } i = 1, \ldots, 5 \\ \alpha - \beta + \epsilon_i & \text{for } i = 6, \ldots, 8 \end{cases}$$

hvor $\epsilon_1, \epsilon_2, \ldots, \epsilon_8 \overset{\text{u.i.f.}}{\sim} N(0, \sigma_2^2)$.

Rimelighetsfunksjonen for $\alpha$ og $\beta$ er

$$L(\alpha, \beta) = \prod_{i=1}^{8} f(d_i; \alpha, \beta) = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2\sigma_2^2}(d_i - \alpha)^2} \cdot \prod_{i=6}^{8} \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2\sigma_2^2}(d_i - (\alpha-\beta))^2},$$

og log-rimelighetsfunksjonen er

$$\ell(\alpha, \beta) = \log L(\alpha, \beta) = -\frac{5}{2}\log(2\pi\sigma_2^2) - \frac{1}{2\sigma_2^2}\sum_{i=1}^{5}(d_i - \alpha)^2 - \frac{3}{2}\log(2\pi\sigma_2^2) - \frac{1}{2\sigma_2^2}\sum_{i=6}^{8}(d_i - \alpha + \beta)^2.$$

De partiellderiverte med hensyn til $\alpha$ og $\beta$ er

$$\frac{\partial \ell}{\partial \alpha} = \frac{1}{2\sigma_2^2} \cdot 2\sum_{i=1}^{5}(d_i - \alpha) + \frac{1}{2\sigma_2^2} \cdot 2\sum_{i=6}^{8}(d_i - \alpha + \beta)$$

og

$$\frac{\partial \ell}{\partial \beta} = -\frac{1}{2\sigma_2^2} \cdot 2\sum_{i=6}^{8}(d_i - \alpha + \beta).$$

Setter vi begge de partiellderiverte lik null, får vi likningene

$$\sum_{i=1}^{8} d_i - 8\alpha + 3\beta = 0 \quad \text{og} \quad \sum_{i=6}^{8} d_i - 3\alpha + 3\beta = 0.$$

Likningen til høyre gir

$$3\beta = 3\alpha - \sum_{i=6}^{8} d_i.$$

og setter vi dette inn i likningen til venstre, får vi

$$\begin{aligned}\sum_{i=1}^{8} d_i - 8\alpha + 3\alpha - \sum_{i=6}^{8} d_i &= 0\\ \sum_{i=1}^{5} d_i &= 5\alpha\\ \alpha &= \frac{1}{5}\sum_{i=1}^{5} d_i.\end{aligned}$$

Om vi så går tilbake til den høyre likningen over, og løser for $\beta$, får vi

$$\begin{aligned}3\beta &= 3\alpha - \sum_{i=6}^{8} d_i\\ \beta &= \alpha - \frac{1}{3}\sum_{i=6}^{8} d_i\\ &= \frac{1}{5}\sum_{i=1}^{5} d_i - \frac{1}{3}\sum_{i=6}^{8} d_i.\end{aligned}$$

Disse løsningene tilsier at sannsynlighetsmaksimeringsestimatorene for $\alpha$ og $\beta$ er

$$\hat{\alpha} = \frac{1}{5}\sum_{i=1}^{5} D_i \quad \text{og} \quad \hat{\beta} = \frac{1}{5}\sum_{i=1}^{5} D_i - \frac{1}{3}\sum_{i=6}^{8} D_i,$$

hvilket stemmer overens med uttrykkene i oppgaveteksten. Forventningsverdiene til $\hat{\alpha}$ og $\hat{\beta}$ er

$$E(\hat{\alpha}) = E\left(\frac{1}{5}\sum_{i=1}^{5} D_i\right) = \frac{1}{5}\sum_{i=1}^{5} E(D_i) = \frac{5\alpha}{5} = \alpha,$$

og

$$\begin{aligned}E(\hat{\beta}) &= E\left(\frac{1}{5}\sum_{i=1}^{5} D_i - \frac{1}{3}\sum_{i=6}^{8} D_i\right) = \frac{1}{5}\sum_{i=1}^{5} E(D_i) - \frac{1}{3}\sum_{i=6}^{8} E(D_i)\\ &= \frac{3\alpha}{3} - \frac{3(\alpha-\beta)}{3} = \beta,\end{aligned}$$

så begge SMEene er forventningsrette. Basert på de oppgitte dybdemålingene har vi estimatene

$$\hat{\alpha} = \frac{2.88 + 2.92 + 2.82 + 2.73 + 2.91}{5} = \underline{\underline{2.8520}}$$

og

$$\hat{\beta} = \frac{2.88 + 2.92 + 2.82 + 2.73 + 2.91}{5} - \frac{2.76 + 2.62 + 2.80}{3} = \underline{\underline{0.1253}}.$$

**d**) For å undersøke om de innsamlede dataene gir grunnlag for å forkaste påstanden om at $\beta = 0$, vil vi sette opp en hypotesetest for den ukjente parameteren $\beta$, med utgangspunkt i estimatoren $\hat{\beta}$, utledet i deloppgave **c)**. Vi vet at $\hat{\beta}$ har forventningsverdi $E(\hat{\beta}) = \beta$. For å sette opp hypotesetesten har vi også bruk for variansen, som er

$$\begin{aligned}\mathrm{Var}(\hat{\beta}) &= \mathrm{Var}\left(\frac{1}{5}\sum_{i=1}^{5} D_i - \frac{1}{3}\sum_{i=6}^{8} D_i\right) = \frac{1}{5^2}\sum_{i=1}^{5}\mathrm{Var}(D_i) + \frac{1}{3^2}\sum_{i=6}^{8}\mathrm{Var}(D_i) \\ &= \frac{5\sigma_2^2}{5^2} + \frac{3\sigma_2^2}{3^2} = \sigma_2^2\left(\frac{1}{5} + \frac{1}{3}\right).\end{aligned}$$

Den ukjente variansen $\sigma_2^2$ må estimeres. Til det bruker vi estimatoren

$$S_2^2 = \frac{1}{6}\left(\sum_{i=1}^{5}(D_i - \hat{\alpha})^2 + \sum_{i=6}^{8}(D_i - \hat{\alpha} + \hat{\beta})^2\right)$$

som er forventningsrett for $\sigma_2^2$. Det står 6 i nevneren fordi vi har $n = 8$ uavhengige observasjoner, og $p = 2$ parametre i modellen, som gir $n - p = 6$ gjenværende frihetsgrader. Den normaliserte tilfeldige variabelen

$$\frac{\hat{\beta} - E(\hat{\beta})}{\sqrt{\mathrm{Var}(\hat{\beta})}} = \frac{\hat{\beta} - \beta}{\sqrt{\sigma_2^2(1/5 + 1/3)}}$$

er standard normalfordelt. Vi lager en testobservator ved å bytte ut $\sigma_2^2$ med $S_2^2$, og $\beta$ med $\beta_0$,

$$T_\beta = \frac{\hat{\beta} - \beta_0}{\sqrt{S_2^2(1/5 + 1/3)}} \sim t_6.$$

Denne er altså $t$-fordelt med $n - p = 6$ frihetsgrader. Under nullhypotesen har vi $\beta_0 = 0$. Estimatet for $\sigma_2^2$ blir

$$s_2^2 = \frac{1}{6}\left[\sum_{i=1}^{5}(d_i - \hat{\alpha})^2 - \sum_{i=6}^{8}(d_i - \hat{\alpha} + \hat{\beta})^2\right] = 0.0071,$$

hvor vi har brukt estimatene $\hat{\alpha}$ og $\hat{\beta}$ funnet i **c)**. Vi antar at det kun er aktuelt å undersøke om vannledningen er hevet etter bruddposisjonen, og ikke senket. Altså lar vi den alternative hypotesen være $H_1 : \ \beta > 0$, og vi setter opp en ensidig, høyresidig test. På signifikansnivå $\alpha = 0.05$ vil vi da forkaste nullhypotesen dersom vi observerer at $T > t_{0.05,6} = 1.9432$, som er 0.95-kvantilen i $t$-fordelingen med 6 frihetsgrader. Den observerte verdien av testobservatoren er

$$t_\beta^{\mathrm{obs}} = \frac{\hat{\beta}}{\sqrt{s_2^2(1/5 + 1/3)}} = \frac{0.1253}{\sqrt{0.0071 \cdot (1/5 + 1/3)}} = 2.0368,$$

som er større enn den kritiske verdien $t_{0.05,6} = 1.9432$, og vi forkaster dermed nullhypotesen på signifikansnivå $\alpha = 0.05$. Dybdemålingene gir med andre ord belegg for å motsi kommuneingeniøren.