Oppgave 1 Sykkelruter

a)

$$P(Y > 6) = 1 - P(Y > 6) = 1 - P(\frac{Y-7}{1} > \frac{6-7}{1}) = 1 - \Phi(-1) = 1 - 0.1587 = 0.8413$$

$$P(X < 7|X < 8) = \frac{P(X < 7 \cap X < 8)}{P(X < 8)} = \frac{P(X < 7))}{P(X < 8)} =$$
$$\frac{P(\frac{X-6}{1} < \frac{7-6}{1})}{P(\frac{X-6}{1} < \frac{8-6}{1})} = \frac{\Phi(1)}{\Phi(2)} = \frac{0.8413}{0.9772} = 0.86$$

$$P(\min(X, Y) < 6) = 1 - P(\min(X, Y) > 6) = 1 - P(X > 6 \cap Y > 6) =$$
$$1 - P(X > 6) \cdot P(Y > 6) = 1 - 0.5 \cdot 0.8416$$

b) Hypoteser:
$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 < \mu_2$

Som tilsvarer
$H_0 : \mu_1 - \mu_2 = 0$
$H_1 : \mu_1 - \mu_2 < 0$

Ser på $\bar{X} - \bar{Y}$, da det er en estimator for $\mu_1 - \mu_2$. Har at $\bar{X} \sim N(\mu_1, \sigma^2/7)$ og $\bar{Y} \sim N(\mu_2, \sigma^2/8)$. Får da; $\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \sigma^2(1/7 + 1/8))$ (lineær kombinasjon av uavhengige normalfordelte stokastiske variable).

Antar at H_0 er sann, dvs $\mu_1 - \mu_2 = 0$ og får da;

$$Z = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{1/7 + 1/8}} \sim N(0, 1)$$

Forkaster H_0 dersom observert $z_{obs} < -z_\alpha$, dvs $z_{obs} < -z_{0.05} = -1.645$. Observer;

$$z_{obs} = \frac{\bar{x} - \bar{y}}{\sigma\sqrt{1/7 + 1/8}} = \frac{6.31 - 6.81}{1\sqrt{1/7 + 1/8}} = -0.96$$

Da z_{obs} ikke er i forkastningsområdet beholder vi H_0, og kan ikke konkludere med at Solan sin rute er raskest.

For å finne styrken må vi finne fordelingen til Z når $X \sim N(6, 1^2)$ og $Y \sim N(7, 1^2)$. Vi kaller denne stok.var. Z_{H_1}. Nå er $\bar{X}-\bar{Y} \sim N(-1, (1/7+1/8))$. Og vi får at $Z_{H_1} = \frac{\bar{X}-\bar{Y}}{\sqrt{1/7+1/8}}$ er normalfordelt med;

$$E(Z_{H_1}) = E(\frac{\bar{X}-\bar{Y}}{\sqrt{1/7+1/8}}) = \frac{1}{\sqrt{1/7+1/8}} E(\bar{X}-\bar{Y}) = \frac{-1}{\sqrt{1/7+1/8}} = -1.92$$

$$Var(Z_{H_1}) = Var(\frac{\bar{X}-\bar{Y}}{\sqrt{1/7+1/8}}) = \frac{1}{1/7+1/8} Var(\bar{X}-\bar{Y}) = 1$$

Altså $Z_{H_1} \sim N(-1.92, 1)$. Styrken er sannsynlihgheten for at H_0 blir forkastet, dvs

$$P(Z_{H_1} < -1.645) = P(\frac{Z_{H_1}-(-1.92)}{1} < \frac{-1.645-(-1.92)}{1}) = \Phi(0.27) = 0.6064$$

Dersom $\mu_1 = 6$ er $\mu_2 = 7$ er styrken til testen på 0.61.

c) Vi har to utvalg med felles varians, og kan da bruke estimatoren

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2 + \sum_{j=1}^{m}(Y_j-\bar{Y})^2}{n+m-2}$$

der $S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$, $S_2^2 = \frac{1}{m-1}\sum_{j=1}^{n}(Y_j-\bar{Y})^2$, $n = 7$ og $m = 8$.

Estimert varians;

$$s_p^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 + \sum_{j=1}^{m}(y_j-\bar{y})^2}{n+m-2} = \frac{6.81+5.44}{7+8-2} = 0.94$$

For å finne et konfidensintervall for σ^2 trenger vi å finne fordelingen til en stokastisk variabel der både S_p^2 og σ^2 inngår. Da det er varians vi ser på, mistenker vi at dette må bli en χ^2- fordeling.

Vi vet at $\frac{(n-1)S_1^2}{\sigma^2} \sim \chi^2_{n-1}$ og $\frac{(m-1)S_2^2}{\sigma^2} \sim \chi^2_{m-1}$. Videre har vi at

$$\frac{(n+m-2)\cdot S_p^2}{\sigma^2} = \frac{(n-1)S_1^2}{\sigma^2} + \frac{(m-1)S_2^2}{\sigma^2}.$$

En sum av χ^2-fordelte variable er χ^2-fordelt med summer av frihetsgradene. Altså er $\frac{(n+m-2)\cdot S_p^2}{\sigma^2} \sim \chi^2_{n+m-2}$, og vi får

$$P(\chi^2_{1-\alpha/2} < \frac{(n+m-2)\cdot S_p^2}{\sigma^2} < \chi^2_{\alpha/2}) = 1-\alpha P(\frac{S_p^2(n+m-2)}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{S_p^2(n+m-2)}{\chi^2_{1-\alpha/2}}) = 1-\alpha.$$

Konfidensintervall

$$[\frac{s_p^2(n+m-2)}{\chi^2_{\alpha/2}}, \frac{s_p^2(n+m-2)}{\chi^2_{1-\alpha/2}}]$$

Innsatt for verdier; $\alpha = 0.05$, $\chi^2_{0.975,13} = 5.009$, $\chi^2_{0.025,13} = 24.736$ får vi $[0.50, 2.44]$.

En hypotesetest:
$H_0 : \sigma^2 = 1$
$H_1 : \sigma^2 \neq 1$

med signifikansnivå på 5% vil ikke bli forkastet da 1 er i konfidensintervallet.

d) Enkel lineær regresjonsmodell:

$$Y_i = \alpha + \beta t_i + \epsilon_i$$

for $i = 1, 2, \ldots, n$. Antar uavhengige normalfordelte støyledd; $\epsilon_i \sim N(0, \sigma_\epsilon)$. (Trenger bare å anta at støyen er uavhengige med $E(\epsilon_i) = 0$ og lik varians.)

Minste-kvadraters estimatorar for regresjonsparametrene;

$$\hat{\beta} = \frac{\sum_{i=1}^n (t_i - \bar{t})Y_i}{\sum_{i=1}^n (t_i - \bar{t})^2}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{t}$$

Har at $E(\hat{\alpha}) = \alpha$ og $E(\hat{\beta}) = \beta$. Antar $\epsilon_i \sim N(0, 0.5^2)$, dvs $\sigma_\epsilon^2 = 0.5^2$. Har da at $Y_i \sim N(\alpha + \beta t_i, \sigma_\epsilon^2)$.

Fabian sin hypotese (tar lengre tid dess senere på morgenen han starter):
$H_0 : \beta = 0$
$H_1 : \beta > 0$

For å teste hypotesa tar vi utgangspunkt i fordelinga til $\hat{\beta}$. $\hat{\beta}$ er en lineær kombinasjon av normalfordelte stokastiske variable (Y_i), og er dermed selv normalfordelt. Vet at $E(\hat{\beta}) = \beta$, trenger $Var(\hat{\beta})$.

$$Var(\hat{\beta}) = Var(\frac{\sum_{i=1}^n (t_i - \bar{t})Y_i}{\sum_{i=1}^n (t_i - \bar{t})^2}) = \frac{1}{(\sum_{i=1}^n (t_i - \bar{t})^2)^2} Var(\sum_{i=1}^n (t_i - \bar{t})Y_i) =$$

$$\frac{1}{(\sum_{i=1}^n (t_i-\bar{t})^2)^2}\sum_{i=1}^n (t_i-\bar{t})^2 Var(Y_i) = \sigma_\epsilon^2 \frac{1}{\sum_{i=1}^n (t_i-\bar{t})^2}$$

Altså er $\hat{\beta} \sim N(\beta, \frac{\sigma_\epsilon^2}{\sum_{i=1}^n (t_i-\bar{t})^2})$. Og vi har under H_0

$$Z = \frac{\hat{\beta}-0}{\sigma_\epsilon/\sqrt{\sum_{i=1}^n (t_i-\bar{t})^2}} \sim N(0,1)$$

Vi vil forkaste H_0 dersom vår observerte $z_{obs} > z_\alpha = z_{0.01} = 2.326$

$$z_{obs} = \frac{0.037}{0.5/\sqrt{2250}} = 3.54$$

z_{obs} er i forkastningsområdet. Vi forkaster H_0, og aksepterer H_1.

e) Vår prediktor: $\hat{Y}_0 = \hat{\alpha} + \hat{\beta} t_0$.

Estimerte regresjonsparametre $\hat{\alpha} = 5.30$ og $\hat{\beta} = 0.0373$. Dermed er predikert verdi for $t_0 = 90$; $\hat{y}_0 = \hat{\alpha} + \hat{\beta} t_0 = 8.66$.

For å finne et prediksjonsintervall ser vi på prediksjonsfeileen $\hat{Y}_0 - Y_0$, som er en lineærkombinasjon av normalfordelte stokastiske variable, og dermed selv normalfordelt med

$$E(\hat{Y}_0 - Y_0) = E(\hat{\alpha} + \hat{\beta} t_0 - (\alpha + \beta t_0 + \epsilon_0)) = \alpha + \beta t_0 - (\alpha + \beta t_0) = 0$$

og

$$Var(\hat{Y}_0 - Y_0) = Var(\hat{\alpha} + \hat{\beta} t_0 - (\alpha + \beta t_0 + \epsilon_0)) = Var(\bar{Y} + \hat{\beta}(t_0 - \bar{t}) - (\alpha + \beta t_0 + \epsilon_0))$$

(bruker at $\bar{Y}$, $\hat{\beta}$ og ϵ_0 er uavhengige)

$$= Var(\bar{Y}) + (t_0-\bar{t})^2 Var(\hat{\beta}) + Var(\epsilon_0) = \frac{\sigma_\epsilon^2}{n} + (t_0-\bar{t})^2\sigma_\epsilon^2 \frac{1}{\sum_{i=1}^n (t_i-\bar{t})^2} + \sigma_\epsilon^2 = \sigma_{\hat{Y}_0 - Y_0}^2$$

Vi får dermed at

$$P(z_{\alpha/2} < \frac{\hat{Y}_0 - Y_0}{\sigma_{\hat{Y}_0 - Y_0}} < z_{\alpha/2}) = 1 - \alpha.$$

Løser ut for Y_0, og får;

$$P(\hat{Y}_0 - z_{\alpha/2}\sigma_{\hat{Y}_0 - Y_0} < Y_0 < \hat{Y}_0 + z_{\alpha/2}\sigma_{\hat{Y}_0 - Y_0}) = 1 - \alpha$$

Prediksjonsintervall: $[\hat{y}_0 - z_{\alpha/2}\sigma_{\hat{Y}_0 - Y_0}; \hat{y}_0 + z_{\alpha/2}\sigma_{\hat{Y}_0 - Y_0}] = [6.51; 10.81]$.

Kommentar: Vi har tilpasset modellen med data fra kl $7:00$ til kl $8:00$. Deretter har vi predikert for klokka $8:30$, en halv time senere. Dette blir kalt ekstrapolasjon. Modellen passer for økende morgentrafikk, men fra dataene vet vi ingenting om at trafikken fortsatt er økende fra klokka $8:00$ til $8:30$. Vi skal derfor være forsiktig med å bruke modellen utenfor tidsspennet i datasettet vårt.

Oppgave 2 Løsning: Ras ved sprengningsarbeid

a) $S = \sum_{i=1}^{4} X_i$. i) X_i er enten suksess, 1, om ras skjer, eller ikke-suksess, 0, om ras ikke skjer. ii) Uavhengige X_i-er. iii) Konstant suksess sannsynlighet $p = 0.15$. S er binomisk fordelt. $P(S=0) = (1-p)^4 = 0.52$

$P(S=1|S \geq 1) = P(S=1)/P(S \geq 1) = p(1-p)^3 4/(1-0.52) = 0.71.$

$P(S>1|S \geq 1) = 1 - P(S=1|S \geq 1) = 1 - 0.71 = 0.29$

b) Z er kostnad, $Z = 40$ mill ved $X = 1$, dvs 'RAS'. $Z = 0$ ved $X = 0$, dvs 'IKKE RAS'. $P(Z = 40) = p = 0.15$. $P(Z = 0) = 1 - p = 0.85$

$$E(Z) = 40 \cdot 0.15 + 0 \cdot 0.85 = 40 \cdot 0.15 = 6$$

$$Var(Z) = E(Z^2) - E(Z)^2 = 40^2 * 0.15 - 6^2 = 14.3^2 = 204$$

$Std(Z) = \sqrt{Var(Z)} = 14.3$ millioner

Strategi A har forventet kostnad (6 millioner), mindre enn 7 millioner som man får ved strategi B. Ved kun å se på forventet verdi, vil vi velge A. Usikkerheten i kostnad er derimot stor, og hvis man ikke liker risiko om uforutsett utgift på 40 millioner, vil man velge B.

Enten er kostnad 7 millioner, dersom grundig undersøkelse svarer 'RAS'. Dette skjer med sannsynlighet 0.15. Eller er kostnad 0, dersom grundig undersøkelse svarer 'IKKE RAS'. Dette skjer med sannsynlighet 0.85. I tillegg kommer en kostnad til ekspertene på 5 millioner.

$$E(X) = 5 + (0.15 \cdot 7 + 0.85 \cdot 0) = 6.05$$

Forventet kostnad er $6.050.000 > 6$ mill. Dersom man bruker forventet kostnad som beslutningsgrunnlag, bør den grundige undersøkelsen ikke gjennomføres. Undersøkelsen har derimot mindre forventet kostnad enn stategi B, så hvis du har valgt strategi B over, er det kanskje igjen smart å gjennomføre undersøkelsen. Underøkelsen gir utfallsrom på kostnad: $\{5, 5+7 = 12\}$.

c) Indikatoren I_i er enten 0 (ved feil uttalelse, dvs $Y_i \neq X_i$) eller 1 (ved riktig uttalelse, dvs $Y_i = X_i$). Sannsynligheten for rett uttalelse, gitt sannheten X_i er:
$P(I_i = 1) = P(Y_i = X_i | X_i) = \gamma$. Ikke-suksess er $I_i = 0$, som skjer 5 ganger, mens suksess er $I_i = 1$ som skjer 10 ganger. Rimelighetsfunksjonen (likeilhood) er

$$L(\gamma) = \prod_{i=1}^{15} P(Y_i = y_i | X_i = x_i) = [\gamma^5(1-\gamma)^{7-5}][\gamma^5(1-\gamma)^{8-5}] = \gamma^{\sum_{i=1}^{15} I_i}(1-\gamma)^{15-\sum_{i=1}^{15} I_i} \tag{1}$$

Log-likelihood er

$$l(\gamma) = \ln L(\gamma) = \sum_{i=1}^{15} I_i \ln \gamma + (15 - \sum_{i=1}^{15} I_i)\ln(1-\gamma)$$

Vi deriverer log likelihood og får: $l'(\hat{\gamma}) = \sum_{i=1}^{15} I_i/\hat{\gamma} - (15 - \sum_{i=1}^{15} I_i)/(1-\hat{\gamma}) = 0$. Løsningen er $\hat{\gamma} = \sum_{i=1}^{15} I_i/15$. Så forslaget er SME.

Det er også mulig å løse oppgaven ved å tenke at $W = \sum_{i=1}^{15} I_i$ er binomisk fordelt, med parameter 15 og γ. Likelihood blir da

$$L(\gamma) = \binom{15}{w}\gamma^w(1-\gamma)^{15-w} \tag{2}$$

med samme løsning som over.

Innsetting: $\hat{\gamma} = \sum_{i=1}^{15} I_i/15 = 10/15 = 2/3 = 0.67$.

d)

$$E(\hat{\gamma}) = \frac{\sum_i E(I_i)}{15} = 15\gamma/15 = \gamma$$
$$Var(\hat{\gamma}) = \frac{\sum_i Var(I_i)}{15^2} = 15(1-\gamma)\gamma/15^2 = (1-\gamma)\gamma/15$$

her er $E(I_i) = \gamma \cdot 1 + (1-\gamma)\cdot 0 = \gamma$, og $Var(I_i) = E(I_i^2) - E(I_i)^2 = \gamma - \gamma^2 = (1-\gamma)\gamma$.

e) Loven om total sannsynlighet:

$$P(Y = 1) = P(Y = 1|X = 1)P(X = 1) + P(Y = 1|X = 0)P(X = 0)$$

Vi får: $P(Y = 1) = 0.66 \cdot 0.15 + 0.33 \cdot 0.85 = 0.38$ $P(Y = 0) = 1 - P(Y = 1) = 0.62$

Dette gjør det lettere å bruke Bayes formel

$$P(X = 1|Y = 0) = \frac{P(Y = 0|X = 1)P(X = 1)}{0.62} = 0.33 \cdot 0.15/0.62 = 0.08$$

$$P(X=0|Y=0) = 1 - 0.08 = 0.92$$

$$P(X=1|Y=1) = \frac{P(Y=1|X=1)P(X=1)}{0.38} = 0.66 \cdot 0.15/0.38 = 0.26$$
$$P(X=0|Y=1) = 1 - 0.26 = 0.74$$

Beslutningen blir billigste løsning. Dersom $E(Z|Y=y) < 7$ mill, velges strategi A. Dersom $E(Z|Y=y) > 7$ mill, velges strategi B.

$$E(Z|Y=0) = 40 \cdot 0.08 + 0 \cdot 0.92 = 3.2 < 7$$

$$E(Z|Y=1) = 40 \cdot 0.26 + 0 \cdot 0.74 = 10.4 > 7$$

$$C = 1 + E(Z|Y=0)P(Y=0) + 7P(Y=1) = 1 + 3.2 \cdot 0.62 + 7 \cdot 0.38 = 5.65$$

Siden forventet kostnad nå er mindre enn 6 mill, bør denne undersøkelsen gjennomføres.