ST1201 Statistiske metoder

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Løsningsforslag - Eksamen desember 2011

Oppgave 1

a)

Dette er en ANOVA-tabell for k-utvalg med $k = 4$ og $n_j = 6$ for $j = 1, 2, 3, 4$. Den fullstendige ANOVA-tabellen blir

Kilde	df	SS	MS	F
Betong	$k - 1 = 3$	47203.13	15734.38	2.90
Error	$6 \cdot 4 - 4 = 20$	108671.50	5433.58	
Total	$6 \cdot 4 - 1 = 23$	155874.63		

der

$$\text{SSTR} = \text{MSTR} \cdot 3 = 47203.14,$$

$$\text{MSE} = \frac{\text{SSE}}{20} = \frac{10861.50}{20} = 5433.58,$$

$$\text{SSTOT} = \text{SSTR} + \text{SSE} = 47203.13 + 108671.50 = 155874.63$$

og

$$F = \frac{\text{MSTR}}{\text{MSE}} = \frac{15734.38}{5433.58} = 2.90.$$

Testobservatoren F relaterer seg til hypotesene

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad \text{mot} \quad H_1 : \text{ikke slik},$$

der μ_i, for $i = 1, 2, 3, 4$, er forventet opptatt fuktighet for betong av type nummer i.
Når H_0 er riktig er F Fisher fordelt med 3 og 20 frihetsgrader. Finner kritisk verdi for $\alpha = 0.05$ fra tabell til å være $f_{0.05,3,20} = 3.10$. Beslutningsregelen blir dermed at vi skal forkaste H_0 når $F > 3.10$. Betongdataene gav $F = 2.90 < 3.10$ slik at konklusjonen blir at vi skal ikke forkaste H_0.

b)
En to-utvalg t-test baserer seg på at man har observasjoner av stokastiske variabler $X_1, \ldots, X_n$ og $Y_1, \ldots, Y_m$ der alle X_i-er og Y_i er uavhengige av hverandre,

$$X_i \sim \mathrm{N}(\mu_X, \sigma^2) \quad \text{og} \quad Y_i \sim \mathrm{N}(\mu_Y, \sigma^2).$$

Man benytter da testobervatoren

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{S_p^2 \left(\frac{1}{n} + \frac{1}{m}\right)}}$$

som er Student t-fordelt med $n + m - 2$ frihetsgrader når $H_0 : \mu_X = \mu_Y$ er riktig. Variansestimatoren S_P^2 er gitt ved formelen

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n + m - 2}.$$

Vi lar X_i-ene og Y_i-ene være henholdsvis data for betong av type 3 og 4. Ved å benytte oppgitte verdier for S_X^2 og S_Y^2 i tabellen på første side av oppgavesettet får man

$$s_P^2 = \frac{5 \cdot 3593.50 + 5 \cdot 3704.27}{10} = 3648.89 \ , \ t = \frac{\frac{3663}{6} - \frac{2924}{6}}{\sqrt{3648.89 \left(\frac{1}{6} + \frac{1}{6}\right)}} = 3.53.$$

Man må her benytte en tosidig test slik at kritisk verdi blir $t_{\frac{\alpha}{2}, n+m-2} = t_{0.025,10} = 2.228$. Beslutningsregelen blir dermed at man skal forkaste H_0 dersom $T < -2.228$ eller $T > 2.228$. Vi observerte $t = 3.53 > 2.228$, slik at konklusjonen blir at vi forkaster H_0.

Det er ikke urimelig at vi i punkt **a**) konkluderer med at det ikke er signifikant forskjell mellom forventningsverdiene til de fire utvalgene, mens vi her i punkt **b**) konkluderer med at det er signifikant forskjell mellom forventningsverdiene til utvalg 3 og 4. Vi kan spesielt legge merke til at vi her i punkt **b**) sammenligner de to av de fire utvalgene som har størst avvik i gjennomsnittsverdi. Vi kan også legge merke til at empirisk varians for utvalg nummer 1 er betydelig større enn for de andre tre utvalgene. ANOVA-analysen baserer seg som kjent på antagelsen om lik varians for alle utvalg. Den store empiriske variansen for utvalg nummer 1 vil dermed føre til at estimatert ("pooled") varians i ANOVA-analysen blir betydelig større enn tilsvarende størrelse i t-testen.

Oppgave 2

a)
Rimelighetsfunksjonen blir her

$$L(p) = f_Y(y; p) = \binom{m}{y} p^y (1-p)^{m-y}.$$

Log-rimelilighetsfunksjonen blir dermed

$$l(p) = \ln[L(p)] = \ln\left(\begin{array}{c} m \\ y \end{array} \right) + y \cdot \ln p + (m-y)\ln(1-p).$$

Deriverer og setter lik null:

$$l'(p) = 0 + \frac{y}{p} + \frac{m-y}{1-p} \cdot (-1) = \frac{y}{p} - \frac{m-y}{1-p} = 0 \Rightarrow y(1-p) = p(m-y) \Rightarrow y - yp = pm - py \Rightarrow p = \frac{y}{m}.$$

Dermed får vi at SME blir

$$\widehat{p} = \frac{Y}{m}.$$

Siden vi kun har en parameter vi skal estimaere, og kun har en observasjon, finner man momentestimatoren ved å sette forventet verdi for Y lik observert verdi for Y. Siden $\mathrm{E}[Y] = mp$ får vi

$$m\widehat{p} = Y \Rightarrow \widehat{p} = \frac{Y}{m}.$$

b)
En estimator $\widehat{\theta}$ for en parameter θ er en *beste* estimator hvis den er forventningsrett og har minst like liten varians som enhver annen forventningsrett estimator.
For å vise at en estimator er en *beste* estimator kan man sjekke at den er forventningsrett og sjekke at variansen til estimatoren er lik Cramér-Raos nedre grense for forvenintningsrette estimatorer.
For $\widehat{p}$ har vi

$$\mathrm{E}[\widehat{p}] = \mathrm{E}\left[\frac{Y}{m}\right] = \frac{\mathrm{E}[Y]}{m} = \frac{mp}{m} = p,$$

og

$$\mathrm{Var}[\widehat{p}] = \mathrm{Var}\left[\frac{Y}{m}\right] \frac{\mathrm{Var}[Y]}{m^2} = \frac{mp(1-p)}{m^2} = \frac{p(1-p)}{m}.$$

Regner ut Cramér-Raos nedre grense, Starter med

$$\ln f_Y(y;p) = \ln\left(\begin{array}{c} m \\ y \end{array} \right) + y \cdot \ln p + (m-y)\ln(1-p).$$

Deriverer to ganger med hensyn på p:

$$\frac{\partial \ln f_Y(y;p)}{\partial p} = \frac{y}{p} + \frac{m-y}{1-p} \cdot (-1) = \frac{y}{p} - \frac{m-y}{1-p},$$

$$\frac{\partial^2 \ln f_Y(y;p)}{\partial p^2} = -\frac{y}{p^2} - \frac{m-y}{(1-p)^2} \cdot (-1) = \frac{m-y}{(1-p)^2} - \frac{y}{p^2}.$$

Tar forventningsverdien,

$$\mathrm{E}\left[\frac{\partial \ln f_Y(y;p)}{\partial p}\right] = \mathrm{E}\left[\frac{m-y}{(1-p)^2} - \frac{y}{p^2}\right] = \frac{m-\mathrm{E}[Y]}{(1-p)^2} - \frac{\mathrm{E}[Y]}{p^2} = \frac{m-mp}{(1-p)^2} - \frac{mp}{p^2}$$

$$= \frac{m}{1-p} - \frac{m}{p} = \frac{m(1-p+p)}{p(1-p)} = \frac{m}{p(1-p)}.$$

Cramé-Raos nedre grense for varians av forventningsrette estimatorer blir dermed (hvor vi benytter at $\widehat{p}$ er basert på kun $n = 1$ stokastiske variabler),

$$\left\{-n\mathrm{E}\left[\frac{\partial^2 \ln f_Y(Y;\theta)}{\partial \theta^2}\right]\right\}^{-1} = \frac{p(1-p)}{m}.$$

Vi ser dermed at $\widehat{p}$ var forventningsrett og at variansen for $\widehat{p}$ er lik Cramér-Raos nedre grense. Dermed er $\widehat{p}$ en beste estimator for p.

c)
Vi ser at $\widehat{\theta}$ er forventningsskjev fordi

$$\mathrm{E}\left[\widehat{\theta}\right] = \mathrm{E}\left[Y - \frac{Y^2}{m}\right] = \mathrm{E}[Y] - \frac{\mathrm{E}[Y^2]}{m} = mp - \frac{\mathrm{Var}[Y] + \mathrm{E}[Y]^2}{m}$$

$$= mp - \frac{mp(1-p) + (mp)^2}{m} = mp - p(1-p) + mp^2 == (m-1)p(1-p) = \frac{m-1}{m}\theta \neq \theta.$$

Vi ser at $\widehat{\theta}$ er assymptotisk forventnigsrett fordi

$$\lim_{m\to\infty} \mathrm{E}\left[\widehat{\theta}\right] = \lim_{m\to\infty}\left[\frac{m-1}{m}\theta\right] = \theta.$$

Vi ser at vi får en forventningsrett estimator med å dele $\widehat{\theta}$ på $(m-1)/m$. Den forventingsrette estimatoren blir dermed

$$\widehat{\theta} = \frac{\widehat{\theta}}{\frac{m-1}{m}} = \frac{mY}{m-1}\left(1 - \frac{Y}{m}\right).$$

Oppgave 3

a)
Rimelighetsfunksjonen blir

$$L(\alpha,\beta) = \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma_0}\exp\left\{-\frac{1}{2\sigma_0^2}(y_i - \alpha - \beta(x_i - \bar{x}))^2\right\}\right].$$

Logrimelighetsfunksjonen blir

$$l(\alpha,\beta) = \sum_{i=1}^{n}\left[-\frac{1}{2}\ln(2\pi) - \ln\sigma_0 - \frac{1}{2\sigma_0^2}(y_i - \alpha - \beta(x_i-\bar{x}))^2\right]$$

$$= -\frac{n}{2}\ln(2\pi) - n\ln\sigma_0 - \frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(y_i - \alpha - \beta(x_i-\bar{x}))^2.$$

Partiellderiverer med hensyn på α og setter lik null,

$$\frac{\partial l}{\partial \alpha} = -\frac{1}{2\sigma_0^2}\sum_{i=1}^{n} 2(y_i - \alpha - \beta(x_i-\bar{x}))\cdot(-1) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i - n\alpha - \beta\sum_{i=1}^{n}(x_i-\bar{x}) = 0 \Rightarrow \alpha = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

hvor vi har benyttet at $\sum_{i=1}^{n}(x_i-\bar{x}) = 0$. Partiellderiverer så med hensyn på β og setter lik null,

$$\frac{\partial l}{\partial \beta} = -\frac{1}{2\sigma_0^2}\sum_{i=1}^{n} 2(y_i - \alpha - \beta(x_i-\bar{x}))\cdot(-(x_i-\bar{x})) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i(x_i-\bar{x}) - \alpha\sum_{i=1}^{n}(x_i-\bar{x}) - \beta\sum_{i=1}^{n}(x_i-\bar{x})^2 = 0$$

$$\Rightarrow \beta = \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2},$$

hvor vi igjen har benyttet at $\sum_{i=1}^{n}(x_i-\bar{x}) = 0$.
Sannsynlighetsmaksimeringsestimorene for α og β er dermed gitt ved

$$\widehat{\alpha} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{og} \quad \widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})Y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}.$$

Ved å benytte at $Y_1,\ldots,Y_n$ er uavhengige får vi

$$\text{Var}\left[\widehat{\beta}\right] = \text{Var}\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})Y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right] = \frac{\text{Var}\left[\sum_{i=1}^{n}(x_i-\bar{x})Y_i\right]}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2} = \frac{\sum_{i=1}^{n}\text{Var}[(x_i-\bar{x})Y_i]}{(\sum_{i=1}^{n}(x_i-\bar{x})^2)^2}$$

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2\text{Var}[Y_i]}{(\sum_{i=1}^{n}(x_i-\bar{x})^2)^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2\sigma_0^2}{(\sum_{i=1}^{n}(x_i-\bar{x})^2)^2} = \frac{\sigma_0^2\sum_{i=1}^{n}(x_i-\bar{x})^2}{(\sum_{i=1}^{n}(x_i-\bar{x})^2)^2}$$

$$= \frac{\sigma_0^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}.$$

b)
$\widehat{\beta}$ er en lineærkombinasjon av Y_i-ene som er uavhengige og normalfordelte variabler. Dermed blir også $\widehat{\beta}$ normalfordelt, dvs.

$$\widehat{\beta} \sim \mathrm{N}\left(\beta, \frac{\sigma_0^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).$$

Dermed har vi også at

$$\frac{\widehat{\beta} - \beta}{\sqrt{\frac{\sigma_0^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim \mathrm{N}(0,1)$$

slik at

$$\mathrm{P}\left(-z_{\frac{a}{2}} \leq \frac{\widehat{\beta} - \beta}{\sqrt{\frac{\sigma_0^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \leq z_{\frac{a}{2}}\right) = 1 - a$$

Løser hver ulikhet med hensyn på β. Starter med den første,

$$-z_{\frac{a}{2}} \leq \frac{\widehat{\beta} - \beta}{\sqrt{\frac{\sigma_0^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \Leftrightarrow -\widehat{\beta} - z_{\frac{a}{2}} \sqrt{\frac{\sigma_0^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \leq -\beta \Leftrightarrow \widehat{\beta} + z_{\frac{a}{2}} \sqrt{\frac{\sigma_0^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \geq \beta$$

$$\frac{\widehat{\beta} - \beta}{\sqrt{\frac{\sigma_0^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \leq z_{\frac{a}{2}} \Leftrightarrow -\beta \leq -\widehat{\beta} + z_{\frac{a}{2}} \sqrt{\frac{\sigma_0^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \Leftrightarrow \beta \geq \widehat{\beta} - z_{\frac{a}{2}} \sqrt{\frac{\sigma_0^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Må dermed også ha at

$$\mathrm{P}\left(\widehat{\beta} - z_{\frac{a}{2}} \sqrt{\frac{\sigma_0^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \leq \beta \leq \widehat{\beta} + z_{\frac{a}{2}} \sqrt{\frac{\sigma_0^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}\right) = 1 - a.$$

Et $(1-a) \cdot 100\%$ konfidensintervall for β blir dermed

$$\left[\widehat{\beta} - z_{\frac{a}{2}} \sqrt{\frac{\sigma_0^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \beta + z_{\frac{a}{2}} \sqrt{\frac{\sigma_0^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}\right].$$

c)
For å finne et prediksjonsintervall tar vi utgangspunkt i

$$\widehat{\alpha} + \widehat{\beta}(x_0 - \bar{x}) - Y_0.$$

Denne vil være normalfordelt fordi det er en lineærkombinasjon av uavhengige normalfordelte variabler, nemlig $Y_1, \ldots, Y_n$ og Y_0. Dette kan vi se ved å sette inn i uttrykket over hva vi i punkt **a**) fant for $\widehat{\alpha}$ og $\widehat{\beta}$.
Forventningsverdien til dette uttrykket blir

$$\mathrm{E}\left[\widehat{\alpha}+\widehat{\beta}(x_0-\bar{x})-Y_0\right] = \mathrm{E}\left[\widehat{\alpha}\right]+\mathrm{E}\left[\widehat{\beta}\right](x_0-\bar{x})-\mathrm{E}[Y_0] = \alpha+\beta(x_0-\bar{x})-(\alpha+\beta(x_0-\bar{x})) = 0.$$

Siden Y_0 åpenbart er uavhengig av $\widehat{\alpha}$ og $\widehat{\beta}$ får vi for tilhørende varians

$$\begin{aligned}
\mathrm{Var}\left[\widehat{\alpha}+\widehat{\beta}(x_0-\bar{x})-Y_0\right] &= \mathrm{Var}\left[\widehat{\alpha}+\widehat{\beta}(x_0-\bar{x})\right]+\mathrm{Var}[Y_0] \\
&= \mathrm{Var}\left[\widehat{\alpha}\right]+\mathrm{Var}\left[\widehat{\beta}(x_0-\bar{x})\right]+2\mathrm{Cov}\left[\widehat{\alpha},\widehat{\beta}(x_0-\bar{x})\right]+\mathrm{Var}[Y_0] \\
&= \mathrm{Var}\left[\widehat{\alpha}\right]+(x_0-\bar{x})^2\mathrm{Var}\left[\widehat{\beta}\right]+2(x_0-\bar{x})\mathrm{Cov}\left[\widehat{\alpha},\widehat{\beta}\right]+\mathrm{Var}[Y_0]
\end{aligned}$$

Fra oppgaveteksten har vi uttrykk for $\mathrm{Var}\left[\widehat{\alpha}\right]$ og $\mathrm{Var}\left[\widehat{\beta}\right]$, og $\mathrm{Var}[Y_0]=\sigma_0^2$. Trenger å regne ut $\mathrm{Cov}\left[\widehat{\alpha},\widehat{\beta}\right]$. Ved å benytte at $\mathrm{Cov}[Y_i,Y_j]=\mathrm{Var}[Y_i]=\sigma_0^2$ dersom $i=j$ og lik 0 hvis $i\neq j$ får vi at

$$\begin{aligned}
\mathrm{Cov}\left[\widehat{\alpha},\widehat{\beta}\right] &= \mathrm{Cov}\left[\frac{1}{n}\sum_{i=1}^{n}Y_i, \frac{\sum_{j=1}^{n}(x_j-\bar{x})Y_j}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right] \\
&= \frac{1}{n\sum_{i=1}^{n}(x_i-\bar{x})^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}\left[Y_i,(x_j-\bar{x})Y_j\right] \\
&= \frac{1}{n\sum_{i=1}^{n}(x_i-\bar{x})^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_j-\bar{x})\mathrm{Cov}\left[Y_i,Y_j\right] \\
= \frac{1}{n\sum_{i=1}^{n}(x_i-\bar{x})^2}\sum_{i=1}^{n}(x_i-\bar{x})\sigma_0^2 &= \frac{\sigma_0^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2}\sum_{i=1}^{n}(x_i-\bar{x}) = 0.
\end{aligned}$$

Dermed får vi at

$$\mathrm{Var}\left[\widehat{\alpha}+\widehat{\beta}(x_0-\bar{x})-Y_0\right] = \frac{\sigma_0^2}{n}+\frac{\sigma_0^2(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}+\sigma_0^2 = \sigma_0^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right),$$

og dermed også

$$\widehat{\alpha}+\widehat{\beta}(x_0-\bar{x})-Y_0 \sim \mathrm{N}\left(0,\sigma_0^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)\right)$$

og

$$\frac{\widehat{\alpha}+\widehat{\beta}(x_0-\bar{x})-Y_0}{\sqrt{\sigma_0^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)}} \sim \mathrm{N}(0,1),$$

slik at

$$\mathrm{P}\left(-z_{\frac{a}{2}} \leq \frac{\widehat{\alpha}+\widehat{\beta}(x_0-\bar{x})-Y_0}{\sqrt{\sigma_0^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^n(x_i-\bar{x})^2}\right)}} \leq z_{\frac{a}{2}}\right) = 1-a.$$

Løser så hver ulikhet hver for seg med hensyn på Y_0 og setter deretter ulikhetene sammen igjen med Y_0 alene i midten, og får

$$\mathrm{P}\left(\widehat{\alpha}+\widehat{\beta}(x_0-\bar{x})-z_{\frac{a}{2}}\sqrt{\sigma_0^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^n(x_i-\bar{x})^2}\right)} \leq Y_0\right.$$

$$\left.\leq \widehat{\alpha}+\widehat{\beta}(x_0-\bar{x})+z_{\frac{a}{2}}\sqrt{\sigma_0^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^n(x_i-\bar{x})^2}\right)}\right) = 1-a.$$

Et $(1-a)\cdot 100\%$ prediksjonsintervall for Y_0 blir dermed

$$\left[\widehat{\alpha}+\widehat{\beta}(x_0-\bar{x})-z_{\frac{a}{2}}\sqrt{\sigma_0^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^n(x_i-\bar{x})^2}\right)}, \widehat{\alpha}+\widehat{\beta}(x_0-\bar{x})+z_{\frac{a}{2}}\sqrt{\sigma_0^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^n(x_i-\bar{x})^2}\right)}\right].$$

Vi ser at prediksjonsintervallet blir kortest når $x_0-\bar{x}=0$, dvs. når $x_0=\bar{x}$.