TMA4240 Statistikk
Eksamen desember 2015

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Løsningsskisse

Oppgave 1

a) Den kumulative fordelingsfunksjonen til X, $F(x) = P(X \leq x)$:

$$F(x) = P(X \leq x) = \int_1^x f(u)du = \int_1^x \theta u^{-(\theta+1)} du$$
$$= \big[-u^{-\theta} \big]_1^x = -x^{-\theta} - (-1) = 1 - \frac{1}{x^\theta}$$

for $x > 1$, ellers er $F(x) = 0$. Setter $\theta = 1.16$, og regner ut

$$P(X \leq 2) = F(2) = 1 - \frac{1}{2^{1.16}} = 0.55$$
$$P(X > 4) = 1 - F(4) = \frac{1}{4^{1.16}} = 0.20$$
$$P(X > 4 \mid X > 2) = \frac{P(X > 4 \cap X > 2)}{P(X > 2)} = \frac{P(X > 4)}{P(X > 2)}$$
$$= \frac{0.20}{1 - 0.55} = 0.44$$

b) Sannsynlighetsmaksimeringsestimatoren for θ.

Vi starter med rimelighetsfunksjonen L, og tar logaritmen til L for å lette regningen.

$$L(\theta, x_1, x_2, \ldots, x_n) = \prod_{i=1}^n f(x_i, \theta) = \prod_{i=1}^n \theta x_i^{-(\theta+1)} = \theta^n \prod_{i=1}^n x_i^{-(\theta+1)}$$
$$l(\theta, x_1, x_2, \ldots, x_n) = \ln L(\theta, x_1, x_2, \ldots, x_n) = n \ln \theta - (\theta + 1) \sum_{i=1}^n \ln x_i$$

Vi vil finne maksimum av l som funksjon av θ og deriverer l med hensyn på θ, og deretter setter lik 0 og løser ut. Vi sjekker at vi finner maksimum og ikke minimum ved å undersøke om den andrederiverte er negativ.

$$\frac{dl}{d\theta} = \frac{n}{\theta} - \sum_{i=1}^{m} \ln x_i \text{ og } \frac{dl}{d\theta} = 0$$
$$\frac{n}{\theta} - \sum_{i=1}^{n} \ln x_i = 0$$
$$\frac{n}{\theta} = \sum_{i=1}^{n} \ln x_i$$
$$\theta = \frac{n}{\sum_{i=1}^{n} \ln x_i}$$

Sjekk av maksimum: $\frac{d^2 l}{d\theta^2} = -\frac{n}{\theta^2} < 0$ alltid negativ, dvs. maksimumspunkt

Dermed har vi maksimert rimelighetsfunksjonen med hensyn på θ og funnet følgende sannsynlighetsmaksimeringsestimator:

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} \ln X_i}$$

Oppgave 2

a) Oppgitt at:

- X: normalfordelt med $\mathrm{E}(X) = 1$ og $\mathrm{Var}(X) = 4$.
- Y: normalfordelt med $\mathrm{E}(Y) = 2$ og $\mathrm{Var}(Y) = 1$.
- X og Y er uavhengige.

Dermed:

$$P(X \leq 0) = P(\frac{X-1}{\sqrt{4}} \leq \frac{0-1}{\sqrt{4}}) = \Phi(-0.5) = 0.3085$$

Videre: $X+Y$ er normalfordelt (siden en sum av to uavhengige normalfordelte variabler er normalfordelt) med $\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y) = 1+2 = 3$ og $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) = 1+4 = 5$.

$$P(X+Y > 4) = 1 - P(X+Y \leq 4) = 1 - P\left(\frac{X+Y-3}{\sqrt{5}} \leq \frac{4-3}{\sqrt{5}}\right)$$
$$= 1 - \Phi(0.45) = 1 - 0.6736 = 0.3264$$

Til slutt, $X-Y$ er også normalfordelt, og $\mathrm{E}(X-Y) = \mathrm{E}(X) - \mathrm{E}(Y) = 1-2 = -1$ og $\mathrm{Var}(X-Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) = 1+4 = 5$.

$$P(X-Y \leq -2) = P\left(\frac{X-Y+1}{\sqrt{5}} \leq \frac{-2+1}{\sqrt{5}}\right)$$
$$= \Phi(-0.45) = 0.3264$$

b) Skriver estimatorene som funksjoner av $V_X = \frac{(n-1)S_X^2}{\sigma^2}$ og $V_Y = \frac{(m-1)S_Y^2}{\sigma^2}$:

$$S^2_{\text{pooled}} = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2} = \frac{\sigma^2 V_X + \sigma^2 V_Y}{n+m-2}$$
$$S^2_{\text{mean}} = \frac{1}{2}S_X^2 + \frac{1}{2}S_Y^2 = \frac{\sigma^2 V_X}{2(n-1)} + \frac{\sigma^2 V_Y}{2(m-1)}$$

Vi bruker at når V_X er kjikvadratfordelt med parameter $(n-1)$ så er $\mathrm{E}(V_X) = (n-1)$ og $\mathrm{Var}(V_X) = 2(n-1)$, og tilsvarende at når V_Y er kjikvadratfordelt med parameter $(m-1)$ så er $\mathrm{E}(V_Y) = (m-1)$ og $\mathrm{Var}(V_Y) = 2(m-1)$ (dette står blant annet i Tabeller og formeler i statistikk under Kjikvadratfordelingen).

$$\begin{aligned}
\mathrm{E}(S^2_{\text{pooled}}) &= \mathrm{E}\big(\frac{\sigma^2 V_X + \sigma^2 V_Y}{n+m-2}\big) = \frac{\sigma^2}{n+m-2}\mathrm{E}\big(V_X + V_Y\big) = \frac{\sigma^2}{n+m-2}[\mathrm{E}\big(V_X\big) + \mathrm{E}(V_Y)] \\
&= \frac{\sigma^2}{n+m-2}[(n-1)+(m-1)] = \sigma^2 \frac{n+m-2}{n+m-2} = \sigma^2
\end{aligned}$$

Dermed er S^2_{pooled} en forventningsrett estimator for σ^2.

$$\begin{aligned}
\mathrm{E}(S^2_{\text{mean}}) &= \mathrm{E}\big(\frac{\sigma^2 V_X}{2(n-1)} + \frac{\sigma^2 V_Y}{2(m-1)}\big) = \frac{\sigma^2}{2(n-1)}\mathrm{E}(V_X) + \frac{\sigma^2}{2(m-1)}\mathrm{E}(V_Y) \\
&= \frac{\sigma^2}{2(n-1)}(n-1) + \frac{\sigma^2}{2(m-1)}(m-1) = \frac{1}{2}\sigma^2 + \frac{1}{2}\sigma^2 = \sigma^2
\end{aligned}$$

Dermed er også S^2_{mean} en forventningsrett estimator for σ^2.

Variansen finnes tilsvarende, men husk at $\mathrm{Var}(V_X) = 2(n-1)$ og $\mathrm{Var}(V_Y) = 2(m-1)$, og konstanter kvadreres: $\mathrm{Var}(aV_X) = a^2\mathrm{Var}(V_X)$.

$$\begin{aligned}
\mathrm{Var}(S^2_{\text{pooled}}) &= \mathrm{Var}\big(\frac{\sigma^2 V_X + \sigma^2 V_Y}{n+m-2}\big) = \frac{\sigma^4}{(n+m-2)^2}\mathrm{Var}\big(V_X + V_Y\big) \\
&= \frac{\sigma^4}{(n+m-2)^2}[\mathrm{Var}\big(V_X\big) + \mathrm{Var}(V_Y)] = \frac{\sigma^4}{(n+m-2)^2}[2(n-1) + 2(m-1)] \\
&= \frac{\sigma^4}{(n+m-2)^2}[2(n+m-2)] = \frac{2\sigma^4}{n+m-2}
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(S^2_{\text{mean}}) &= \mathrm{Var}\big(\frac{\sigma^2 V_X}{2(n-1)} + \frac{\sigma^2 V_Y}{2(m-1)}\big) = \frac{\sigma^4}{4(n-1)^2}\mathrm{Var}(V_X) + \frac{\sigma^4}{4(m-1)^2}\mathrm{Var}(V_Y) \\
&= \frac{\sigma^4}{4(n-1)^2}2(n-1) + \frac{\sigma^4}{4(m-1)^2}2(m-1) = \frac{\sigma^4}{2}\big(\frac{1}{n-1} + \frac{1}{m-1}\big)
\end{aligned}$$

Vi skal bare sammenligne variansene for S^2_{pooled} og S^2_{mean} når $n = 10$ og $m = 20$.

$$\mathrm{Var}(S^2_{\text{pooled}}) = \frac{2\sigma^4}{10+20-2} = \frac{2\sigma^4}{10+20-2} = \frac{2\sigma^4}{28} = 0.07\sigma^4$$
$$\mathrm{Var}(S^2_{\text{mean}}) = \frac{\sigma^4}{2}\big(\frac{1}{10-1} + \frac{1}{20-1}\big) = \frac{\sigma^4}{2}\big(\frac{1}{9} + \frac{1}{19}\big) = 0.08\sigma^4$$

Dermed er $\text{Var}(S^2_{\text{pooled}}) < \text{Var}(S^2_{\text{mean}})$ for disse verdiene av n og m, og vi vil generelt foretrekke så liten som mulig varians for en estimator. Dermed vil vi foretrekke S^2_{pooled}.

Ikke spurt om: men, det vil være slik at variansen til S^2_{pooled} alltid er mindre eller lik variansen til S^2_{mean}. Dette kan man se enklest ved å se at

$$\begin{aligned}
\text{Var}(S^2_{\text{pooled}}) &\leq \text{Var}(S^2_{\text{mean}}) \\
\frac{1}{n+m-2} &\leq \frac{1}{4}\left(\frac{1}{n-1} + \frac{1}{m-1}\right) \\
4(n-1)(m-1) &\leq ((n-1)+(m-1))(n+m-2) \\
4(n-1)(m-1) &\leq (n+m-2)^2 \\
n^2 - 4nm + m^2 &\geq 0 \\
(n-m)^2 &\geq 0
\end{aligned}$$

Likhet får vi når $n = m$.
Heller ikke spurt om: Videre så har S^2_{pooled} en fordel til over S^2_{mean} fordi $\frac{(n+m-2)S^2_{\text{pooled}}}{\sigma^2}$ er en sum av to uavhengige kjikvadratfordelte størrelser og er dermed kjikvadratfordelt med parameter $n+m-2$ og kan lett brukes til å lage en t-fordelt observator (som gjort i neste oppgave), mens dette ikke er tilfellet for S^2_{mean}.

c) Vi ønsker å undersøke om det er grunn til å tro at μ_X er større enn μ_Y, og dermed setter vi opp:

$$H_0\colon \mu_X = \mu_Y \text{ mot } H_1\colon \mu_X > \mu_Y$$

som er det samme som å teste

$$H_0\colon \mu_X - \mu_Y = 0 \text{ mot } H_1\colon \mu_X - \mu_Y > 0$$

Testobservator er

$$T_0 = \frac{\bar{X} - \bar{Y}}{S_{\text{pooled}}\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

som er t-fordelt med $n+m-2$ frihetsgrader når nullhypotesen er sann.

Vi forkaster H_0 når T_0 er stor (fordi når H_0 er falsk er $\mu_X - \mu_Y > 0$ og da vil vi forvente at $\bar{X} - \bar{Y}$ er stor, og dermed også T_0 stor), og finner forkastningsgrensen ved å løse

$$\begin{aligned}
P(\text{forkaste } H_0 \text{ når } H_0 \text{ er sann}) &= 0.05 \\
P(T_0 > k) &= 0.05
\end{aligned}$$

dermed må $k = t_{0.05,n+m-2}$, fordi det er det tallet i t-fordelingen som det er sannsynlighet 0.05 for å være større enn.
Forkastningsregelen blir: Forkast H_0 når $T_0 > t_{0.05,n+m-2}$.

Innsatt verdier: $n = 129$, $\bar{x} = 75.2$, $s_X^2 = 174.6$, $m = 141$, $\bar{y} = 61.0$ og $s_Y^2 = 292.1$.

$$
\begin{aligned}
t_{0.05,129+141-2} &= t_{0.05,268} \approx 1.645 \\
s_{\text{pooled}} &= \sqrt{\frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}} = \sqrt{\frac{128 \cdot 174.6 + 140 \cdot 292.1}{129+141-2}} \\
&= \sqrt{\frac{63242.8}{268}} = \sqrt{236.0} = 15.4 \\
t_0 &= \frac{\bar{x} - \bar{y}}{s_{\text{pooled}}\sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{75.2 - 61.0}{15.4\sqrt{\frac{1}{129} + \frac{1}{141}}} = \frac{14.2}{1.87} = 7.6
\end{aligned}
$$

Dermed har vi at $t_0 = 7.6$ er mye større en forkastningsgrensen 1.645 og vi forkaster klart H_0. P-verdien (ikke spurt om) for testen blir $P(T_0 > 7.6) = 2.5 \cdot 10^{-13}$ (det kan man ikke finne i tabell, men med software).

Oversatt til spørreundersøkelses-situasjonen: det er grunn til å tro at de som har svart «Enig» eller «Svært enig» på påstanden «Akkurat nå synes jeg at Statistikk er et artig kurs» har en høyere opplevd prosent av forelesningen de i gjennomsnitt mener de forstår enn de som svarte «Nøytral», «Uenig» eller «Svært uenig» på «artig»-spørsmålet.

Oppgave 3

a) $P(X = 0) = 1 - P(X = 1) = 1 - 0.2 = 0.8$ og $P(Y = 1|X = 0) = 1 - P(Y = 0|X = 0) = 0.1$.

$$
\begin{aligned}
P(Y = 1) &= P(X = 0 \cap Y = 1)) + P(X = 1 \cap Y = 1) \\
&= P(X = 0)P(Y = 1|X = 0) + P(X = 1)P(Y = 1|X = 1) \\
&= 0.8 \cdot 0.1 + 0.2 \cdot 0.9 = 0.26
\end{aligned}
$$

$$
P(X = 1|Y = 1) = \frac{P(X = 1 \cap Y = 1)}{P(Y = 1)} = \frac{P(X = 1)P(Y = 1|X = 1)}{P(Y = 1)} = \frac{0.2 \cdot 0.9}{0.26} = 0.69
$$

b) Vi har en binomisk fordeling fordi vi ser på dette som 1) 5 delforsøk, 2) i hvert delforsøk sjekker vi om signalet har kommet frem korrekt (suksess) eller ikke, 3) suksessannsynligheten er 0.9, og 4) de 5 delforsøkene er uavhengig av hverandre. Da er V=antall suksesser binomisk fordelt med parametere 5 og 0.9.

Binomisk fordeling gir

$$
P(V \geq 4) = P(V = 4) + P(V = 5) = 5 \cdot 0.9^4 \cdot 0.1^1 + 0.9^5 = 0.33 + 0.59 = 0.92
$$

Eller fra tabell: $P(V \geq 4) = 1 - P(X \leq 3) = 1 - 0.08 = 0.92$.

La X være kortnotasjon for $(X_1, X_2, X_3, X_4, X_5)$ og tilsvarende for Y. Bayes' regel:

$$
\begin{aligned}
&P(X = (0,0,0,0,0)|Y = (0,0,0,0,0)) \\
=&\frac{P(X = (0,0,0,0,0) \cap Y = (0,0,0,0,0))}{P(Y = (0,0,0,0,0))} \\
=&\frac{P(X = (0,0,0,0,0))P(Y = (0,0,0,0,0)|X = (0,0,0,0,0))}{\sum_x P(X = x)P(Y = (0,0,0,0,0)|X = x)}
\end{aligned}
$$

I nevneren må man summere over alle de 32 mulige sendte signaler. Fra binomisk fordeling kan mottatt signal $Y = (0,0,0,0,0)$ skje ved at ulike antall posisjoner endres: ingen elementer endres (som i teller), ett element endres (i en av de 5 posisjonene), to elementer endres (det er i alt 10 par av to blant de 5 posisjonene), etc. helt til alle posisjoner endres; $X = (1,1,1,1,1)$. Ved direkte bruk av binomisk fordeling og sannsynligheter for X har vi:

$$P(Y = (0,0,0,0,0)|X = (0,0,0,0,0))P(X = (0,0,0,0,0)) = 0.9^5 \cdot 0.35 = 0.207$$
$$P(Y = (0,0,0,0,0)|X = (1,1,1,1,1))P(X = (1,1,1,1,1)) = 0.1^5 \cdot 0.35 = 0.0000035$$

For de andre 30 andre utfallene samlet får vi følgende i nevner:

$$(5 \cdot 0.9^4 \cdot 0.1 + 10 \cdot 0.9^3 \cdot 0.1^2 + 10 \cdot 0.9^2 \cdot 0.1^3 + 5 \cdot 0.9 \cdot 0.1^4) \cdot 0.01 = 0.0041$$

Da blir

$$P(X = (0,0,0,0,0)|Y = (0,0,0,0,0)) = 0.207/(0.207 + 0.0041 + 0.00000035) = 0.98$$

Oppgave 4

a) Fra tabell for Poisson fordelingen med $\lambda = 18$:

$$\begin{aligned} P(Y \leq 18) &= 0.56 \\ P(Y > 10|Y \leq 18) &= \frac{P(10 < Y \leq 18)}{P(Y \leq 18)} = \frac{P(Y \leq 18) - P(Y \leq 10)}{0.56} \\ &= \frac{0.5622 - 0.0304}{0.5622} = 0.946 \end{aligned}$$

Ikke spurt om - men for å motivere at Poisson kan tilnærmes med normal. For normalfordelingen:

$$\begin{aligned} P(Y \leq 18) &= P(Z \leq \frac{18-18}{\sqrt{18}}) = P(Z \leq 0) = 0.5 \\ P(Y > 10|Y \leq 18) &= \frac{P(Z \leq 0) - P(Z < -8/\sqrt{18})}{P(Z \leq 0)} = \frac{0.5 - 0.0297}{0.5} = 0.941 \end{aligned}$$

b) Det ser ut til å være en positiv trend med høyere salg for høyere føreforholdindeks. Og økningen ser ut til å være omtrent konstant mellom føreforholdindeksene (1,2,3,4). Men variabiliteten øker tydelig med føreforholdsindeksen. Den er mye større for $x = 4$ enn for $x = 1$. Antakelsen om konstant varians for støyleddene er dermed neppe gyldig.

La en ny observasjon (for «fantastiske forhold») være Y_0. Prediksjonen er $\hat{Y}_0$ og føreforholdsindeksen er $x_0 = 4$. Et 90 prosent prediksjonsintervall starter med $\hat{Y}_0 - Y_0$, $E(\hat{Y}_0 - Y_0) = 0$, og $s_0^2 = Var(\hat{Y}_0 - Y_0) = \sigma^2(\frac{1}{n} + \frac{(4-\bar{x})^2}{\sum_{i=1}^{20}(x_i-\bar{x})^2} + 1)$.

$\hat{\beta}_1 = 237.15/24.95 = 9.5$ $\hat{\beta}_0 = 25.65 - 9.5 * 2.45 = 2.4$

Prediksjonen er $\hat{Y}_0 = \hat{\beta}_0 + 4 \cdot \hat{\beta}_1 = 2.4 + 4 \cdot 9.5 = 40.4$.

Estimatet av variansen σ^2 er oppgitt til $s^2 = 5.65^2$, og vi får en t-fordeling: Da har vi

$$P(t_{18,0.05} < \frac{\hat{Y}_0 - Y_0}{s_0} < t_{18,0.95}) = 0.9, \quad t_{18,0.05} = -1.734$$

$s_0^2 = s^2(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{20}(x_i-\bar{x})^2} + 1) = 5.65^2(1/20 + (4-2.45)^2/24.95 + 1) = 6.05^2.$

$$P(\hat{Y}_0 + s_0 \cdot t_{18,0.05} < Y_0 < \hat{Y}_0 + s_0 \cdot t_{18,0.95}) = 0.9$$

Prediksjonsintervallet blir: $(29.9, 50.9)$.

c)

$$\begin{aligned} E(\min\{Y, 25\}) &= \int_{-\infty}^{n} y f(y) dy + \int_{n}^{\infty} n f(y) dy \\ &= \mu\Phi((n-\mu)/\sigma) - \sigma\phi((n-\mu)/\sigma) + n(1 - \Phi((n-\mu)/\sigma)) \end{aligned}$$

Som gir $E(\min\{Y, 25\}) = 19.58$

Funksjonen som må optimeres er

$$f(n) = 20 \cdot E(\min\{Y, n\}) - 5 \cdot n = 20(n - (n-\mu) \cdot \Phi((n-\mu)/\sigma) - \sigma\phi((n-\mu)/\sigma)) - 5n$$

Prøving og feiling gir $f(20) = 260.1$, $f(25) = 266.7$, $f(22) = 266.9$, $f(24) = 267.9$, $f(23) = 268.1$. Dermed blir optimalt antall kopper forberedt $n = 23$.