Norges teknisk–naturvitenskapelige universitet
Institutt for matematiske fag

Løsningsskisse for eksamen i TMA4240 Statistikk, 3. desember 2022

Oppgave 1 Spillet Tenzies

a) Situasjonen representerer den med 10 uavhengige forsøk med suksess sansynlighet $p = 1/6$. Da er antall suksesser Y binomisk fordelt.

$$P(Y > 0) = 1 - P(Y = 0) = 1 - \binom{10}{0} p^0 (1-p)^{10} = 1 - (5/6)^{10} = 1 - 0.16 = 0.84.$$

$$P(Y = 1 | Y > 0) = \frac{P(Y = 1, Y > 0)}{P(Y > 0)} = \frac{P(Y = 1)}{P(Y > 0)} = \frac{10 \frac{1}{6} \frac{5}{6}^9}{0.84} = 0.32/0.84 = 0.38.$$

b) Geometrisk fordeling kjennetegnes ved uavhengige forsøk med konstant sannsynlighet for suksess, og Z_i er da antall forsøk til første suksess.

$$P(Z_i = z) = (1-p)^{z-1} p = q^{z-1} p, \qquad z = 1, 2, \ldots.$$

Her er $p = 1/6$ og $q = 1 - 1/6 = 5/6$.

$$F(z) = P(Z_i \leq z) = \sum_{y=1}^{z} q^{y-1} p = p \frac{1 - q^z}{1 - q} = 1 - q^z = 1 - \left(\frac{5}{6}\right)^z.$$

Vi løser for $F(z) = 0.5$.

$$1 - \left(\frac{5}{6}\right)^z = 0.5, \quad z \ln(5/6) = \ln 0.5, \quad z = \frac{\ln 0.5}{\ln(5/6)} = 3.8.$$

Verdien 3.8 er ikke en del av utfallsrommet. Den kumulative sannsynlighetsfordelingen når verdien 0.5 for $z = 4$.

c) $W = \max\{Z_1, \ldots, Z_{10}\}$.

$$P(W \leq w) = P(Z_1 \leq w, \ldots, Z_{10} \leq w) = \prod_{i=1}^{10} P(Z_i \leq w) = \left[1 - \left(\frac{5}{6}\right)^w\right]^{10}.$$

Sannsynligheten for å bli konge er da

$$P(W \leq 10) = \left[1 - \left(\frac{5}{6}\right)^{10}\right]^{10} = 0.84^{10} = 0.17.$$

Oppgave 2 Støymålinger

a)

$$P(Y > 60) = P\left(\frac{Y - 40}{10} > \frac{60 - 40}{10}\right) = P(Z > 2) = 1 - 0.98 = 0.02$$

$$P(50 < Y < 60) = P(Y < 60) - P(Y < 50) = 0.98 - P(Z < 1) = 0.98 - 0.84 = 0.14$$

$$P(Y > c) = 0.01 = P\left(\frac{Y - 40}{10} > \frac{c - 40}{10}\right) = 0.01,$$

Det betyr at $(c - 40)/10 = 2.33$, og da er $c = 10 \cdot 2.33 + 40 = 63.3$.

b) For å finne et 95 % konfidensintervall for μ tar vi utgangspunkt i at $\frac{\bar{Y}-\mu}{S/\sqrt{n}} \sim t_{29}$. Nedre og øvre 0.025 percentiler i denne fordelingen er $\pm t_{29,0.025} = \pm 2.05$. Da er

$$P\left(-2.05 < \frac{\bar{Y} - \mu}{S/\sqrt{n}} < 2.05\right) = 0.95$$

$$P\left(-2.05 \cdot S/\sqrt{n} < \bar{Y} - \mu < 2.05 \cdot S/\sqrt{n}\right) = 0.95$$

$$P\left(\bar{Y} - 2.05 \cdot S/\sqrt{n} < \mu < \bar{Y} + 2.05 \cdot S/\sqrt{n}\right) = 0.95$$

Innsatt tall blir 95 % konfidensintervall for μ: $57.0 \pm 2.05 \cdot 11.9/\sqrt{30}$ eller $(52.5, 61.5)$.

c) Klokka $t_i = 12$ er $(t_i - 8)/4 = 1$ og fordi $\cos(\pi) = -1$ så er $x_i = 0$. Derfor vil forventet verdi klokka 12 være $E(Y_i) = \beta_0$.

Vi ønsker å undersøke om det er grunn til å konkludere at forventet støynivå klokka 12; β_0, er mindre enn 50dB. Hypotesen er

$$H_0\colon \beta_0 \geq 50 \text{ mot } H_1\colon \beta_0 < 50$$

Testobservator er

$$\frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{30}(x_i-\bar{x})^2}}} \sim t_{28}$$

Ved innsetting av tall får vi: $\hat{\beta}_1 = \frac{\sum_{i=1}^{30}(x_i-\bar{x})y_i}{\sum_{i=1}^{30}(x_i-\bar{x})^2} = 227.7/20.1 = 11.3$ og $\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 = (1709/30) - (38.5/30)\cdot 11.3 = 42.4$. Under H_0 er $\beta_0 = 50$, og dermed blir

$$\frac{42.50 - 50}{7.32\sqrt{\frac{1}{30} + \frac{(38.5/30)^2}{20.1}}} = -3.03.$$

Her er $t_{28,0.05} = -1.70$. Det er klart at $-3.03 < -1.70$ og dermed forkastes H_0.

Oppgave 3 Bilverkstedet

a) Setter inn $\beta = 1$ i tetthetsfunksjonen for X og regner ut arealet under tettheten mellom 0 og 1:

$$P(X < 1) = \int_0^1 \exp(-x)dx = -\exp(-1) + \exp(0) = 1 - 0.37 = 0.63.$$

Den deriverte av momentgenererende funksjon, evaluert i 0, definerer forventninger: $E(X) = \frac{dM_X(0)}{dt}$ og $E(X^2) = \frac{d^2M_X(0)}{dt^2}$. Her er:

$$\frac{dM_X(t)}{dt} = \beta(1-\beta t)^{-2}, \quad \frac{dM_X(0)}{dt} = E(X) = \beta.$$

$$\frac{d^2M_X(t)}{dt^2} = 2\beta^2(1-\beta t)^{-3}, \quad \frac{d^2M_X(0)}{dt^2} = E(X^2) = 2\beta^2.$$

$$Var(X) = E(X^2) - [E(X)]^2 = 2\beta^2 - \beta^2 = \beta^2.$$

b) Ved transformasjon av variable er tettheten til $Z = \frac{2Y}{\beta}$ gitt ved:

$$f_Z(z) = f_Y[y(z)]\frac{dy}{dz} = \frac{1}{\beta^2}(z\beta/2)\exp[-\frac{1}{\beta}(z\beta/2)]\frac{\beta}{2} = \frac{z}{4}\exp(-z/2)$$

der $y = z\beta/2$ og $\frac{dy}{dz} = \beta/2$ brukes.

Tettheten gjenkjennes som en χ^2_4 tetthet: $f(z) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}z^{\nu/2-1}\exp(-z/2)$, der $\nu = 4$ er frihetsgradene. Her er $\Gamma(\nu/2) = \Gamma(2) = 1$.

For en χ^2_4 fordeling har vi at $E(Z) = 4$. Da er $E(Y) = \beta/2E(Z) = 2\beta$.

For en χ^2_4 fordeling har vi at $Var(Z) = 2 \cdot 4 = 8$. Da er $Var(Y) = (\beta^2/4)Var(Z) = 2\beta^2$.

(Merk at Y sin tetthet kan gjenkjennes som en gamma-fordeling med parametre $\alpha = 2$ og β. Forventning og varians stemmer med uttrykkene for forventning og varians til denne gammafordelinga.)

c) Sannsynlighetsestimatoren finnes ved å maksimere likelihood med hensyn på parameteren β. Likelihood er definert som simultan tetthet til alle data:

$$L(\beta) = \prod_{i=1}^{m_1} f_X(x_i)\prod_{j=1}^{m_2} f_Y(y_j) = \frac{1}{\beta^{m_1}}\exp\left(-\frac{\sum_{i=1}^{m_1} x_i}{\beta}\right)\frac{\prod_{j=1}^{m_2} y_j}{\beta^{2m_2}}\exp\left(-\frac{\sum_{j=1}^{m_2} y_j}{\beta}\right)$$

Log-likelihood er gitt ved

$$l(\beta) = \ln L(\beta) = -m_1\ln\beta - \frac{\sum_{i=1}^{m_1} x_i}{\beta} + \sum_{j=1}^{m_2}\ln y_j - 2m_2\ln\beta - \frac{\sum_{j=1}^{m_2} y_j}{\beta}$$

Vi deriverer denne:

$$l'(\beta) = -\frac{m_1}{\beta} + \frac{\sum_{i=1}^{m_1} x_i}{\beta^2} - \frac{2m_2}{\beta} + \frac{\sum_{j=1}^{m_2} y_j}{\beta^2}$$

Vi setter den deriverte lik 0 for å finne sannsynlighetsmaksimeringsestimatet. Løsningen er

$$\hat{\beta} = \frac{\sum_{i=1}^{m_1} x_i + \sum_{j=1}^{m_2} y_j}{m_1 + 2m_2}$$

Man kan vise at dette er et topppunkt. Vi har andrederiverte

$$l''(\beta) = -\frac{1}{\beta^2}\left(2\frac{\sum_{i=1}^{m_1} x_i + \sum_{j=1}^{m_2} y_j}{\beta} - m_1 - 2m_2\right).$$

Ved innsetting av $\hat{\beta} > 0$ får vi

$$l''(\hat{\beta}) = -\frac{1}{\hat{\beta}^2}(2m_1 + 4m_2 - m_1 - 2m_2) = -\frac{1}{\hat{\beta}^2}(m_1 + 2m_2) < 0.$$

Forventningen til $\hat{\beta}$ er ved innsetting fra oppgave a) og b):

$$E(\hat{\beta}) = \frac{\sum_{i=1}^{m_1} E(X_i) + \sum_{j=1}^{m_2} E(Y_j)}{m_1 + 2m_2} = \frac{\sum_{i=1}^{m_1} \beta + \sum_{j=1}^{m_2} 2\beta}{m_1 + 2m_2} = \beta$$

Variansen til $\hat{\beta}$ er ved innsetting fra oppgave a) og b):

$$Var(\hat{\beta}) = \frac{\sum_{i=1}^{m_1} Var(X_i) + \sum_{j=1}^{m_2} Var(Y_j)}{(m_1 + 2m_2)^2} = \frac{\sum_{i=1}^{m_1} \beta^2 + \sum_{j=1}^{m_2} 2\beta^2}{(m_1 + 2m_2)^2} = \frac{\beta^2}{(m_1 + 2m_2)}$$

d) Utfra sentralgrenseteoremet er estimatoren

$$\hat{\beta} = \frac{\sum_{i=1}^{m_1} X_i + \sum_{j=1}^{m_2} Y_j}{m_1 + 2m_2}$$

approksimativt normalfordelt. Dette teoremet sier at summer og gjennomsnitt av uavhengige identisk fordelte variable blir normalfordelt når antall ledd i summen går mot uendelig. Her er det egentlig to summer (en av X_i-er og en av Y_j-er), men de blir begge normalfordelte når m_1 og m_2 går mot uendelig.

En testobservatoren blir da

$$\frac{\hat{\beta} - \beta}{\beta/\sqrt{m_1 + 2m_2}} \sim N(0, 1).$$

Vi forkaster $H_0 : \beta = 1$ mot $H_1 : \beta > 1$ dersom vi denne testobservatoren er signifikant stor under H_0. Her er $\hat{\beta} = \frac{1.43 \cdot 100 + 2.80 \cdot 100}{100 + 200} = 1.41$.

$$\frac{1.41 - 1}{1/\sqrt{300}} = 7.1$$

P-verdien finnes ved å faktisk se hvor sannsynlig en så stor verdi er. I dette tilfellet er $P(Z > 7.1) < 0.0000000001$. Tolkningen av en så liten p-verdi er at hypotese H_0 bør forkastes.

(Dersom man setter inn $\hat{\beta} = 1.41$ i standardavviket i nevner (som også gir en approksimativ normalfordeling), så får man $\frac{1.41-1}{1.41/\sqrt{300}} = 5.04$. p-verdien er $P(Z > 5.0) < 0.000001$ som også er svært liten.)