TMA4245 Statistikk
Eksamen 20. desember 2012

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Løsningsskisse

Oppgave 1

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

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

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:

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

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 2

a)

$$P(X > 1000) = P(\frac{X - 800}{100} > 2) = P(Z > 2) = 0.023$$

$$P(500 < X < 1000) = P(X < 1000) - P(X < 500) = (1 - 0.023) - P(Z < -3) = 0.976$$

b) From the probability of 0 in the Poisson:

$$P(\text{Ingen fisk}) = e^{-3} = 0.05$$

By conditional probability and the Poisson distribution:

$$P(X > 3|X > 0) = \frac{1 - P(X \leq 3)}{1 - 0.05} = \frac{1 - (0.05 + 0.15 + 0.22 + 0.22)}{0.95} = \frac{1 - 0.647}{0.95} = 0.37$$

c)

$$P(X > 0) = 1 - P(X = 0) = 1 - (\theta + (1 - \theta) \cdot e^{-\mu}) = 1 - (0.5 + 0.5e^{-4}) = 0.49$$

$$E(X) = \sum_{x=0}^{\infty} xP(X = x) = \sum_{x=1}^{\infty} xP(X = x) = (1 - \theta) \sum_{x=1}^{\infty} x \frac{\mu^x}{x!} e^{-\mu} = (1 - \theta)\mu$$

where the last sum is the expected value in the usual Poisson distribution (Here μ). This gives $E(X) = 0.5 \cdot 4 = 2$.

d) The likelihood function is the product of the independent variables, regarded as a function of the unknown parameters:

$$L(\theta, \mu) = \prod_{i=1}^{n} P(X = x_i) = (\theta + (1 - \theta)e^{-\mu})^r \prod_{x_i > 0} (1 - \theta) \frac{\mu^{x_i}}{x_i!} e^{-\mu}$$

Log likelihood becomes:

$$l(\theta,\mu) = \ln L(\theta,\mu) = r\ln(\theta+(1-\theta)e^{-\mu})+(n-r)\ln(1-\theta)+\ln\mu\sum_{x_i>0}x_i-(n-r)\mu-\sum_{x_i>0}\ln x_i!$$

The maximum likelihood estimator for θ is found by differentiation with respect to θ:

$$\frac{dl}{d\theta} = r\frac{1-e^{-\mu}}{\theta+(1-\theta)e^{-\mu}} - \frac{(n-r)}{1-\theta}$$

Solving for $\frac{dl}{d\theta} = 0$ gives:

$$r(1-\theta)(1-e^{-\mu}) = (n-r)(\theta+(1-\theta)e^{-\mu})$$

And separating for θ gives the desired solution:

$$\hat{\theta} = \frac{r-ne^{-\mu}}{n(1-e^{-\mu})}$$

The plot peaks at about $\hat{\mu} = 3$, which is the maximum likelihood estimate for μ.

Inserting this we get

$$\hat{\theta} = \frac{8-20e^{-3}}{20(1-e^{-3})} = 0.37$$

Oppgave 3

a) The method of least squares fits the line with smallest square distances from the measurements to the line. It minimizes $\text{SSE} = \sum_i (y_i - \alpha - \beta x_i)^2$.

By differentiation we get

$$\frac{d\text{SSE}}{d\alpha} = -2\sum_i (y_i-\alpha-\beta x_i), \quad \frac{d\text{SSE}}{d\beta} = -2\sum_i x_i(y_i-\alpha-\beta x_i)$$

Setting the derivatives equal to 0, we have

$$\sum_{i=1}^n (y_i-\alpha-\beta x_i) = 0 \quad \text{and} \quad \sum_{i=1}^n x_i(y_i-\alpha-\beta x_i) = 0.$$

Dividing by n this gives

$$\bar{y}-\alpha-\beta\bar{x} = 0 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^n x_i y_i - \alpha\bar{x} - \beta\cdot\frac{1}{n}\sum_{i=1}^n x_i^2 = 0.$$

Solving the first equation with respect to α gives

$$\alpha = \bar{y} - \beta\bar{x},$$

and inserting this expression into the latter we obtain

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i - (\bar{y} - \beta\bar{x})\bar{x} - \beta \cdot \frac{1}{n}\sum_{i=1}^{n} x_i^2 = 0,$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{y}\bar{x} + \beta\left(\bar{x}^2 - \frac{1}{n}\sum_{i=1}^{n} x_i^2\right) = 0 \Rightarrow \beta = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}y}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2}.$$

Multiplying with n in the numerator and denominator in this last expression, we get

$$\beta = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}y}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}.$$

To obtain the least squares estimators we replace the y_i's with the corresponding random variables, i.e.

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

b) The starting point is

$$P\left(-t_{n-2,0.025} < \frac{(\hat{\beta} - \beta)}{s/\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} < t_{n-2,0.025}\right) = 0.95$$

Solving each of the inequalities with respect to β we get

$$-t_{n-2,0.025} < \frac{(\hat{\beta} - \beta)}{s/\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \Rightarrow -\frac{-st_{n-2,0.025}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} < \hat{\beta} - \beta$$

$$\Rightarrow \hat{\beta} + \frac{st_{n-2,0.025}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} > \beta$$

and

$$\frac{(\hat{\beta} - \beta)}{s/\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} < t_{n-2,0.025} \Rightarrow \hat{\beta} - \beta < \frac{st_{n-2,0.025}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$\Rightarrow \beta > \hat{\beta} - \frac{st_{n-2,0.025}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Thereby we have

$$P\left(\hat{\beta} - \frac{st_{n-2,0.025}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} < \beta < \hat{\beta} + \frac{st_{n-2,0.025}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right) = 0.95$$

Thus, the confidence interval is

$$\left(\hat{\beta} - \frac{st_{n-2,0.025}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \hat{\beta} + \frac{st_{n-2,0.025}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right).$$

We have $n = 28$ and the table gives us $t_{n-2,0.025} = 2.056$. Inserting numbers we get: $\hat{\beta} = -5942/36517 = -0.16$. The winning time is expected to improve by $4 \cdot 0.16 = 0.64$ seconds in successive Olympic Games.

The 95 percent confidence interval: $(-0.199, -0.126)$.

c) We set $x_0 = 2016$. We have $\hat{\alpha} = 109.26 + 0.1627 \cdot 1954.5 = 427.3$. The predicted time is $\hat{Y}_0 = \hat{\alpha} + 2016\hat{\beta} = 427.3 - 2016 \cdot 0.1627 = 99.3$, that is about 1 min, 39 sec.

The 95% prediction interval for the winning time in 2016 is $\hat{Y}_0 \pm t_{n-2,0.025} s\sqrt{1 + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2} + 1/n}$. Inserting numbers the prediction interval becomes $(91.8, 106.7)$.

d) We have $\hat{Y}_0 = \hat{\alpha} + x_0\hat{\beta} = 90$. This means:

$$x_0 = \frac{90 - \hat{\alpha}}{\hat{\beta}} = \frac{90 - 427.3}{-0.1627} = 2073$$

The next Olympic Games is in 2076.

It appears as if the winning times are nonlinear as a function of time. The model says that they spent 427 sec in year 0, which is not likely. Also, the model predicts that we will run at negative times in the future, which is not possible.

Modeling assumptions can be checked using residual plots. Do the residuals $e_i = Y_i - \hat{Y}_i$ seem to have a trend? According to the model they should be close to independent and identically Gaussian distributed. With the current plot, it seems that the early winning times are above the line, while the winning times (1950-1980) are under the line, and the latest winning times are above.

Based on residual plots it may be worth looking at another model, possibly for the log winning times, or something else.