TMA4240 Statistics
Exam November 2018

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Løsningsskisse

Oppgave 1

a) To find the cumulative distribution function $F_Y(y)$:

$$
\begin{aligned}
F_Y(y) &= P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq +\sqrt{y}) \\
&= \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1+x}{2} dx = \sqrt{y}
\end{aligned}
$$

for $y \in (0,1)$.

Then

$$
F_Y(y) = \begin{cases} 0 & y \leq 0 \\ \sqrt{y} & y \in (0,1) \\ 1 & y \geq 1 \end{cases}
$$

We then find the pdf by deriving $F_Y(y)$ wrt y:

$$
f_Y(y) = \frac{dF_Y(y)}{dy} = \begin{cases} 0 & y \leq 0 \\ \frac{1}{2} y^{-1/2} & y \in (0,1) \\ 0 & y > 1 \end{cases}
$$

Finally we get the expected value of $2Y - Y^2$ by:

$$
\begin{aligned}
E(2Y - Y^2) &= \int_{\mathcal{R}} (2y - y^2) f_Y(y) dy \\
&= \int_0^1 (2y - y^2) \frac{1}{2} y^{-1/2} dy \\
&= \frac{2}{3} - \frac{1}{5} \\
&= \frac{7}{15}
\end{aligned}
$$

Oppgave 2

a) A Poisson process must satisfy the following properties

- The number of events occurring in disjoint time intervals are independent
- The probability that a single outcome will occur in a very short time interval is proportional to the length of the time interval

$$P(X = 1 \text{ in } (0,t)) = \lambda t + o(t)$$

- The probability that more than one outcome will occur in such a short time interval is negligible

$$P(X \geq 2 \text{ in } (0,t)) = o(t)$$

The parameter λ is the expected number of cars passing by the specific point every minute.

b) We have that $P(X(t) = x) = \frac{(\lambda t)^x}{x!} \exp\{-(\lambda t)\}$ with $\lambda = 1.5$

$$P(X(1) = 2) = \frac{(\lambda 1)^2}{2!} \exp\{-(\lambda 1)\} = \frac{1.5^2}{2!} \exp\{-1.5\} = 0.25$$

$$P(X(2) \geq 2) = 1 - P(X(2) \leq 1) = 1 - \sum_{x=0}^{1} \frac{(2\ \lambda)^x}{x!} \exp\{-2\ \lambda\} = 1 - 0.1991 = 0.8$$

To solve the last question we first compute the probability that during 1-minute period there are more than 5 cars passing:

$$P(X(1) > 5) = 1 - P(X(1) \leq 5) = 1 - \sum_{x=0}^{5} \frac{(\lambda)^x}{x!} \exp\{-(\lambda)\} = 1 - 0.9955 = 0.0045$$

Let $Z = \{\text{more than 5 cars are passing in at least period}\}$, thus

$$\begin{aligned} P(Z) &= 1 - P(\text{at most 5 cars are passing in every period}) \\ &= 1 - (P(\text{at most 5 cars are passing in one period}))^{10} \\ &= 1 - (1 - 0.0045)^{10} \\ &= 0.044 \end{aligned}$$

c) The hypotheses test to perform is:

$$\begin{cases} H_0: & \lambda = 1.5 \\ H_1: & \lambda > 1.5 \end{cases}$$

We have that $X_i \sim \text{Poisson}(\lambda t_i)$ for $i = 1, \ldots, 10$ and the X_i's are independent. Therefore the stochastic variable $\sum_{i=1}^{10} X_i$ is distributed as a Poisson with mean $\mu = \lambda \sum_{i=1}^{10} t_i$ and variance $\mu = \lambda \sum_{i=1}^{10} t_i$.

Under H_0 we have that $\lambda = 1.5$, moreover, from the data is $\sum_{i=1}^{10} t_i = 100$.

Thus, $\sum_{i=1}^{10} X_i$ is approximately normally distributed with mean $\mu = \lambda \sum_{i=1}^{10} t_i$ and variance $\mu = \lambda \sum_{i=1}^{10} t_i$ since $\mu = 150$ under the null hypothesis. Alternatively, we have

$$E(\widehat{\lambda}) = E\left(\frac{\sum_{i=1}^{10} X_i}{\sum_{i=1}^{10} t_i}\right) = \frac{\lambda \sum_{i=1}^{10} t_i}{\sum_{i=1}^{10} t_i} = \lambda$$

and

$$\text{Var}(\widehat{\lambda}) = \text{Var}\left(\frac{\sum_{i=1}^{10} X_i}{\sum_{i=1}^{10} t_i}\right) = \frac{\lambda \sum_{i=1}^{10} t_i}{\left(\sum_{i=1}^{10} t_i\right)^2} = \frac{\lambda}{\sum_{i=1}^{10} t_i}$$

A test statistics is then

$$Z = \frac{\widehat{\lambda} - \lambda}{\sqrt{\frac{\lambda}{\sum_{i=1}^{10} t_i}}} = \frac{\sum_{i=1}^{10} X_i - \lambda \sum_{i=1}^{10} t_i}{\sqrt{\lambda \sum_{i=1}^{10} t_i}}$$

which is $N(0,1)$ under H_0.

With a significance of 1% we reject if $Z > z_{0.01} = 2.326$. In our case we have

$$Z = \frac{192 - 1.5 \cdot 100}{\sqrt{1.5 \cdot 100}} = 3.429$$

Therefore we reject H_0 and build a toll house.

d) We have that Z is a binomial random variable with parameters $n = 10$ and

$$p = P(X(t) > \lambda_0 t)$$

Under $H_0 : \lambda = \lambda_0 = 1.5$ we have that $X(t) \sim \text{Poisson}(1.5t)$, so

$$p = 1 - P(X(t) \leq 15) = 1 - \sum_{x=0}^{15} \frac{15^x}{x!} \exp\{-15\} = 1 - 0.5681 = 0.4319$$

So under H_0 we have $Z \sim \text{Binom}(n = 10, p = 0.4319)$

We need to find the smallest value of k such that when $\lambda = 1.5$ we have $P(Z \geq k) \leq 0.01$.

For different values of k we have:

k	$P(Z \geq k$ when H_0 is correct$)$
10	0.00022
9	0.0032
8	0.0207

So we have that $k = 9$.

In our dataset we have that $z = 8$ so we do not reject H_0.

Oppgave 3

a) Since $Y \sim n(y; 15, 4)$ we have

$$\begin{aligned}
P(Y > 20) &= 1 - P(Y \leq 20) \\
&= 1 - P\left(\frac{Y - 15}{4} \leq \frac{20 - 15}{4}\right) \\
&= 1 - \Phi(1.25) \\
&= 1 - 0.8944 \\
&= 0.1056
\end{aligned}$$

Since Y is normally distributed with expectation 15 and the normal distribution is symmetric around the mean, we get that $P(Y > 20) = P(Y < 10) = 0.1056$. Since the events $Y > 20$ and $Y < 10$ are disjoint we get

$$P(Y < 10 \cup Y > 20) = 2P(Y > 20) = 2 \cdot 0.1056 = 0.2112.$$

Finally, from the definition of conditional probability we obtain

$$\begin{aligned} P(Y > 20 | Y > 10) &= \frac{P(Y > 10 \cap Y > 20)}{P(Y > 10)} \\ &= \frac{P(Y > 20)}{P(Y > 10)} \\ &= \frac{0.1056}{1 - 0.1056} \\ &= 0.1181 \end{aligned} .$$

b) The likelihood function is given as

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} n(y_i; \beta x_i, 4) \cdot \prod_{i=1}^{n} n(z_i; c_0 + \beta x_i, 4) \\ &= \left(\frac{1}{\sqrt{2\pi \cdot 4^2}}\right)^n \exp\left\{-\frac{1}{2 \cdot 4^2}\sum_{i=1}^{n}(y_i - \beta x_i)^2\right\} \\ &\quad \times \left(\frac{1}{\sqrt{2\pi \cdot 4^2}}\right)^n \exp\left\{-\frac{1}{2 \cdot 4^2}\sum_{i=1}^{n}(z_i - c_0 - \beta x_i)^2\right\} \\ &= \left(2\pi \cdot 4^2\right)^{-n} \exp\left\{-\frac{1}{2 \cdot 4^2}\sum_{i=1}^{n}(y_i - \beta x_i)^2\right\} \exp\left\{-\frac{1}{2 \cdot 4^2}\sum_{i=1}^{n}(z_i - c_0 - \beta x_i)^2\right\} \end{aligned}$$

The log-likelihood is given as

$$l(\beta) = -n \log\left(2\pi \cdot 4^2\right) - \frac{1}{2 \cdot 4^2}\sum_{i=1}^{n}(y_i - \beta x_i)^2 - \frac{1}{2 \cdot 4^2}\sum_{i=1}^{n}(z_i - c_0 - \beta x_i)^2$$

Differentiating with respect to β

$$\begin{aligned} l'(\beta) &= -\frac{1}{2 \cdot 4^2}\sum_{i=1}^{n} 2(y_i - \beta x_i)(-x_i) - \frac{1}{2 \cdot 4^2}\sum_{i=1}^{n} 2(z_i - c_0 - \beta x_i)(-x_i) \\ &= \frac{1}{4^2}\sum_{i=1}^{n} x_i(y_i + z_i - c_0 - 2\beta x_i) \end{aligned}$$

Set $l'(\beta) = 0$ and solve for β

$$\begin{aligned} \sum_{i=1}^{n} x_i(y_i + z_i - c_0 - 2\beta x_i) &= 0 \\ \beta \cdot 2\sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i(y_i + z_i) - c_0 \sum_{i=1}^{n} x_i \\ \beta &= \frac{\sum_{i=1}^{n} x_i(y_i + z_i) - c_0 \sum_{i=1}^{n} x_i}{2\sum_{i=1}^{n} x_i^2} \end{aligned}$$

Thus, the MLE for β is

$$\widehat{\beta} = \frac{\sum_{i=1}^n x_i(Y_i + Z_i) - c_0 \sum_{i=1}^n x_i}{2\sum_{i=1}^n x_i^2}$$

Its expectation and variance is given as

$$\begin{aligned}
\mathrm{E}(\widehat{\beta}) &= \mathrm{E}\left(\frac{\sum_{i=1}^n x_i(Y_i + Z_i) - c_0 \sum_{i=1}^n x_i}{2\sum_{i=1}^n x_i^2}\right) \\
&= \frac{\sum_{i=1}^n x_i \cdot \mathrm{E}(Y_i + Z_i) - c_0 \sum_{i=1}^n x_i}{2\sum_{i=1}^n x_i^2} \\
&= \frac{\sum_{i=1}^n x_i \cdot (\beta x_i + c_0 + \beta x_i) - c_0 \sum_{i=1}^n x_i}{2\sum_{i=1}^n x_i^2} \\
&= \frac{2\beta \sum_{i=1}^n x_i^2 + c_0 \sum_{i=1}^n x_i - c_0 \sum_{i=1}^n x_i}{2\sum_{i=1} x_i^2} \\
&= \beta
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(\widehat{\beta}) &= \mathrm{Var}\left(\frac{\sum_{i=1}^n x_i(Y_i + Z_i) - c_0 \sum_{i=1}^n x_i}{2\sum_{i=1}^n x_i^2}\right) \\
&= \frac{\sum_{i=1}^n x_i^2 \cdot \mathrm{Var}(Y_i + Z_i)}{\left(2\sum_{i=1}^n x_i^2\right)^2} \\
&= \frac{2\sigma^2 \sum_{i=1}^n x_i^2}{\left(2\sum_{i=1}^n x_i^2\right)^2} \\
&= \frac{\sigma^2}{2\sum_{i=1}^n x_i^2} \\
&= \frac{4^2}{2\sum_{i=1}^n x_i^2}
\end{aligned}$$

since Y_i and Z_i are independent.

c) Since $\widehat{\beta}$ is a linear combination of independent and normally distributed random variables it is also normally distributed:

$$\widehat{\beta} \sim n\left(z; \beta, \sqrt{\frac{4^2}{2\sum_{i=1}^n x_i^2}}\right).$$

We therefore construct a $(1-\alpha)\cdot 100$ % confidence interval for β based on

$$Z = \frac{\widehat{\beta} - \beta}{\sqrt{\frac{4^2}{2\sum_{i=1}^n x_i^2}}} \sim n(z; 0, 1).$$

$$P\left(-z_{\alpha/2} \leq \frac{\widehat{\beta} - \beta}{\sqrt{\frac{4^2}{2\sum_{i=1}^n x_i^2}}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

Solving for β we get

$$P\left(\widehat{\beta} - z_{\alpha/2}\sqrt{\frac{4^2}{2\sum_{i=1}^n x_i^2}} \leq \beta \leq \widehat{\beta} + z_{\alpha/2}\sqrt{\frac{4^2}{2\sum_{i=1}^n x_i^2}}\right) = 1-\alpha$$

That is, the $(1-\alpha)\cdot 100$ % confidence interval for β is

$$\left[\widehat{\beta} - z_{\alpha/2}\sqrt{\frac{4^2}{2\sum_{i=1}^n x_i^2}}, \widehat{\beta} + z_{\alpha/2}\sqrt{\frac{4^2}{2\sum_{i=1}^n x_i^2}}\right]$$

We have

$$\widehat{\beta} = \frac{68586 + 72398 - 5\cdot 982}{2\cdot 97324} = 0.699$$

and

$$\sqrt{\frac{4^2}{2\cdot\sum_{i=1}^n x_i^2}} = \sqrt{\frac{4^2}{2\cdot 97324}} = 0.009$$

thus we get

$$[0.699 - 1.645\cdot 0.009, 0.699 + 1.645\cdot 009] = [0.684, 0.713].$$

Since $\mu_0 = 0.5$ is not inside the 90 % interval we reject the null hypothesis.

Oppgave 4

a) In the case with a random sample $X_1, X_2, \ldots, X_n \sim n(x;\mu,\sigma)$ where both μ and σ is unknown it is known that a $(1-\alpha)\cdot 100$ % prediction interval for a new observation X_0 independent of $X_1, X_2, \ldots, X_n$ is given as

$$\left[\bar{X} - t_{\alpha/2,n-1}\sqrt{S^2\left(1+\frac{1}{n}\right)}, \bar{X} + t_{\alpha/2,n-1}\sqrt{S^2\left(1+\frac{1}{n}\right)}\right]$$

where $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ and $t_{\alpha/2,n-1}$ is the $\alpha/2$ critical value in the student t-distribution with $n-1$ degrees of freedom.

In our case a 95 % prediction interval is

$$\left[\frac{53.37}{10} - 2.262\cdot\sqrt{0.73^2\left(1+\frac{1}{10}\right)}, \frac{53.37}{10} + 2.262\cdot\sqrt{0.73^2\left(1+\frac{1}{10}\right)}\right] = [3.605, 7.069]$$

b) Under the null hypothesis we have

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim n(z;0,1).$$

We reject the null hypothesis on a significance level α if $Z \geq z_\alpha$. For a given alternative hypothesis $\mu = \mu_0 + \delta$ the power of the test is

$$1 - \beta \leq P(\text{reject H}_0 \text{ when } \mu = \mu_0 + \delta)$$

that is

$$
\begin{aligned}
\beta &\geq P(\text{do not reject } \mathrm{H}_0 \text{ when } \mu = \mu_0 + \delta) \\
&= P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq z_\alpha \text{ when } \mu = \mu_0 + \delta\right) \\
&= P\left(\bar{X} \leq \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} \text{ when } \mu = \mu_0 + \delta\right) \\
&= P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{\mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} - \mu}{\sigma/\sqrt{n}} \text{ when } \mu = \mu_0 + \delta\right). \\
&= P\left(Z \leq \frac{z_\alpha \frac{\sigma}{\sqrt{n}} - \delta}{\sigma/\sqrt{n}}\right) \\
&= P\left(Z \leq z_\alpha - \frac{\delta}{\sigma/\sqrt{n}}\right)
\end{aligned}
$$

We conclude that

$$
\begin{aligned}
-z_\beta &\geq z_\alpha - \frac{\delta\sqrt{n}}{\sigma} \\
\frac{\delta\sqrt{n}}{\sigma} &\geq z_\alpha + z_\beta \qquad . \\
n &\geq \left(\frac{(z_\alpha + z_\beta)\sigma}{\delta}\right)^2
\end{aligned}
$$

In our case we have $\sigma = 1, \alpha = 0.05, z_\alpha = 1.645, \beta = 0.05, z_\beta = 1.645$ and $\delta = 0.5$, and get

$$n \geq 43.3.$$

Eva needs to weight at least 44 salmons.