Norwegian University of Science and Technology
Department of Mathematical Sciences

EKSAMEN I EMNE TMA4245 STATISTIKK

3. juni 2011

Problem 1

a) Properties of a Poisson process:

(i) The number of eruptions in a time interval is independent of the number of eruptions in other, disjoint time intervals.

(ii) The probability that an eruption will occur during a time interval is proportional to the length of the time interval.

(iii) The probability that more than one eruption will occur during a very short time interval is negligible.

Let X be the number of eruptions occurring during $t = 5$ years, or $t = 5 \cdot 12 = 60$ months.

$$P(\geq 1) = 1 - P(X = 0) = 1 - e^{-\lambda t} = 1 - e^{-0.026 \cdot 60} = 0.789$$

The question *What is the probability that the next eruption will occur more than three years after the starting date?* can be interpreted in two ways. Either that there are no eruptions for the three first years of the tenancy;

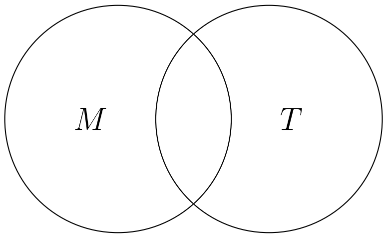
$P(\text{More than 3 years to next eruption}) = P(\text{No eruption in 3 years}) = e^{-0.024 \cdot 36} = 0.392$

Or that it is more then 4.5 years = 42 months to the next eruption;

$P(\text{More than 42 months to next eruption}) = P(\text{No eruption in 42 months}) = e^{-0.024 \cdot 42} = 0.336.$

Problem 2 Boligmarkedet i Trondheim

a) A venn diagram of the events



We have

$$P(M) = \frac{94}{381} = 0.2467$$
$$P(T) = \frac{190}{381} = 0.4987$$
$$P(M \cap T) = \frac{50}{381} = 0.1312$$

If the events M and T are disjoint, then $P(M \cap T) = 0$. Here, $P(M \cap T) > 0$ and the events are thus *not* disjoint.
If the events are independent, then $P(M \cap T) = P(M) \cdot P(T)$. Here

$$P(M) \cdot P(T) = 0.2467 \cdot 0.4987 = 0.1230 \neq P(M \cap T)$$

and the events are thus *not* independent. But they are close to independent.

b) We notice that the expected value and variance of Y is

$$\begin{aligned} \mathrm{E}(Y) &= \mathrm{E}(\beta x + \epsilon(x)) = \mathrm{E}(\beta x) + \mathrm{E}(\epsilon(x)) = \beta x + 0 = \beta x \\ \mathrm{Var}(Y) &= \mathrm{Var}(\beta x + \epsilon(x)) = \mathrm{Var}(\beta x) + \mathrm{Var}(\epsilon(x)) = 0 + \tau^2 x^2 = \tau^2 x^2 \end{aligned}$$

If $\beta > 1$ we thus expect the final price per m^2, Y, to be greater then the suggested price x, i.e. expect that the apartment will be sold at a higher price than suggested by the estate company.

We now define W as the final price for an $60m^2$ apartment,

$$W = 60Y$$

As W is a linear combination of a Gaussian variable Y, then W must be Gaussian as well. With suggested price per m^2 $x = 1.8/60 = 0.03$ and assuming $\beta = 1,1$, $\tau^2 = 0.1^2$

its expected value and variance is

$$\begin{aligned}\mathrm{E}(V) &= 60\mathrm{E}(Y) = 60\beta x = 60 \cdot 1.1 \cdot 0.03 = 1.98 \\ \mathrm{Var}(V) &= 60^2 \mathrm{Var}(Y) = 60^2\tau^2 x^2 = 60^2 \cdot 0.1^2 \cdot 0.03^2 = 0.18^2\end{aligned}$$

Thus $V \sim N(1.98, 0.18^2)$

(i) The probability of paying more than 2 mill.kr for the apartment is

$$\begin{aligned}P(W > 2) = 1 - P(W \le 2) = 1 - P\left(\frac{W - 1.98}{0.18} \le \frac{2 - 1.98}{0.18}\right) \\ = 1 - \Phi(Z \le 0.11) = 1 - 0.5438 = 0.4562\end{aligned}$$

c) The maximum likelihood estimator of β is

$$\hat{\beta} = \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i}{x_i}$$

We first notice that the estimator is a linear combination of Gaussian variables $Y's$, and must thus be Gaussian itself. The expectation and variance of $\hat{\beta}$ is

$$\begin{aligned}\mathrm{E}(\hat{\beta}) &= \mathrm{E}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{Y_i}{x_i}\right) = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{E}(Y_i)}{x_i} = \frac{1}{N}\sum_{i=1}^{N}\frac{\beta x_i}{x_i} = \frac{1}{N}\sum_{i=1}^{N}\beta = \beta \\ \mathrm{Var}(\hat{\beta}) &= \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{Y_i}{x_i}\right) = \frac{1}{N^2}\sum_{i=1}^{N}\frac{\mathrm{Var}(Y_i)}{x_i^2} = \frac{1}{N^2}\sum_{i=1}^{N}\frac{\tau^2 x_i^2}{x_i^2} = \frac{\tau^2}{N}\end{aligned}$$

Thus $\hat{\beta} \sim N(\beta, \frac{\tau^2}{N})$.
A confidence interval can be found by

$$P\left(-z_{\alpha/2} \le \frac{\hat{\beta} - \beta}{\sqrt{\frac{\tau^2}{N}}} \le z_{\alpha/2}\right) = 1 - \alpha$$

$$P\left(-z_{\alpha/2} \cdot \frac{\tau}{\sqrt{N}} \le \hat{\beta} - \beta \le z_{\alpha/2} \cdot \frac{\tau}{\sqrt{N}}\right) = 1 - \alpha$$

$$P\left(\hat{\beta} - z_{\alpha/2} \cdot \frac{\tau}{\sqrt{N}} \le \beta \le \hat{\beta} + z_{\alpha/2} \cdot \frac{\tau}{\sqrt{N}}\right) = 1 - \alpha$$

With $\alpha = 0.05$ we have $z_{0.025} = 1.960$, and the 95%-confidence interval is

$$\begin{aligned}\left[\hat{\beta} - z_{\alpha/2} \cdot \frac{\tau}{\sqrt{N}}, \hat{\beta} + z_{\alpha/2} \cdot \frac{\tau}{\sqrt{N}}\right] &= \left[\frac{1}{30} \cdot 32.98 - 1.960 \cdot \frac{0.1}{\sqrt{30}}\ ,\ \frac{1}{30} \cdot 32.98 + 1.960 \cdot \frac{0.1}{\sqrt{30}}\right] \\ &= [1.0635\ ,\ 1.135]\end{aligned}$$

d) To test if the two areas has the same proportion between suggested price and expected price we test for whether the slope parameter β is equal or not. Formulated as a hypothesis test

$$\mathrm{H}_0 \ : \beta_1 = \beta_2 \quad , \quad \mathrm{H}_1 \ : \beta_1 \neq \beta_2$$

Here we have denoted the parameter from Midtbyen as β_1. The regression model for Tyholt is equal to the model for Midtbyen, and we find the maximum likelihood estimate of β_2 similar to what we did for β_1 in c)

$$\hat{\beta}_2 = \frac{1}{N} \sum_{i=1}^{M} \frac{W_i}{x_i} \sim N\left(\beta_2, \frac{\tau^2}{M}\right)$$

The variable $\hat{\beta}_1 - \hat{\beta}_2$ is thus a linear combination of Gaussian variables, and is itself Gaussian with mean and variance by

$$\begin{aligned} \mathrm{E}(\hat{\beta}_1 - \hat{\beta}_2) &= \beta_1 - \beta_2 \\ \mathrm{Var}(\hat{\beta}_1 - \hat{\beta}_2) &= \mathrm{Var}(\hat{\beta}_1) + \mathrm{Var}(\hat{\beta}_2) = \frac{\tau^2}{N} + \frac{\tau^2}{M} \end{aligned}$$

A confidence interval for $\beta_1 - \beta_2$ is

$$P\left(-z_{\alpha/2} \leq \frac{(\hat{\beta}_1 - \hat{\beta}_2) - (\beta_1 - \beta_2)}{\sqrt{\frac{\tau^2}{N_1} + \frac{\tau^2}{N_2}}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

$$P\left(-z_{\alpha/2} \cdot \sqrt{\frac{\tau^2}{N_1} + \frac{\tau^2}{N_2}} \leq (\hat{\beta}_1 - \hat{\beta}_2) - (\beta_1 - \beta_2) \leq z_{\alpha/2} \cdot \sqrt{\frac{\tau^2}{N_1} + \frac{\tau^2}{N_2}}\right) = 1 - \alpha$$

$$P\left((\hat{\beta}_1 - \hat{\beta}_2) - z_{\alpha/2} \cdot \sqrt{\frac{\tau^2}{N_1} + \frac{\tau^2}{N_2}} \leq (\beta_1 - \beta_2) \leq (\hat{\beta}_1 - \hat{\beta}_2) + z_{\alpha/2} \cdot \sqrt{\frac{\tau^2}{N_1} + \frac{\tau^2}{N_2}}\right) = 1 - \alpha$$

With inserted values we find the 95% -confidence interval

$$\begin{aligned} &\left[(\hat{\beta}_1 - \hat{\beta}_2) - z_{\alpha/2} \cdot \sqrt{\tfrac{\tau^2}{N_1} + \tfrac{\tau^2}{N_2}} \ , \ (\hat{\beta}_1 - \hat{\beta}_2) + z_{\alpha/2} \cdot \sqrt{\tfrac{\tau^2}{N_1} + \tfrac{\tau^2}{N_2}}\right] \\ = &\left[\left(\tfrac{32.98}{30} - \tfrac{56.66}{50}\right) - 1.960 \cdot \sqrt{\tfrac{0.1^2}{30} + \tfrac{0.1^2}{50}} \ , \ \left(\tfrac{32.98}{30} - \tfrac{56.66}{50}\right) + 1.960 \cdot \sqrt{\tfrac{0.1^2}{30} + \tfrac{0.1^2}{50}}\right] \\ &= [-0.0791 \ , \ 0.0114] \end{aligned}$$

As the interval does include the value 0 we do not reject the null hypothesis of equal parameter β at a 5% significance level.

e) In the regression model we have assumed that the mean is linear with respect to the suggested price and that the error terms, ϵ, are independent Gaussian distributed with mean 0 and variance $x_i^2\tau^2$.

From the Figure it seems at there might not be a linear relationship between the mean and x as all observations for $x > 30.5$ are above the line, the last 6 all more then two standard deviations.

We can analyze this assumption by making scatter plots for the residuals $e_i = y_i - \hat{y}_i$. If these are independent, the residuals should be spread quite uniformly in the scatter plot. We can also make a histogram of the residuals and check if it resembles a Gaussian density.

Further normality can can be checked bt a qq-plot. Note that this is not strait forward as we have assumed known, but different variances.

Problem 3

a) The cumulative distribution function of an exponentially distributed variable having expected value μ is given by $F(t) = 1 - e^{-t/\mu}$, so when $\mu = 2$, $P(X < 1) = 1 - e^{-1/2} = 0.39$, where X is production time.

None of indepentent production times X_1, X_2, X_3, X_4, X_5 being less than 1 is the same as all of them being greater than 1, the probability of which is $(P(X_i > 1))^5 = (1 - (1 - e^{-1/2}))^5 = 0.082$.

b) First we note that $E\bar{X} = E(\frac{1}{n_1}\sum_{i=1}^{n_1} X_i) = \frac{1}{n_1}\sum_{i=1}^{n_1} EX_i = \frac{1}{n_1} \cdot n_1\mu = \mu$ and that $\text{Var}\bar{X} = \text{Var}(\frac{1}{n_1}\sum_{i=1}^{n_1} X_i) = \frac{1}{n_1^2}\sum_{i=1}^{n_1} \text{Var}X_i = \frac{1}{n_1^2} \cdot n_1\mu^2 = \mu^2/n_1$, and, likewise, $E\bar{Y} = \mu/c$ and $\text{Var}\bar{Y} = \mu^2/(c^2 n_2)$.

For $c = 2$, $\alpha = \frac{1}{2}$ and $\beta = 1$, $\tilde{\mu} = \frac{1}{2}\bar{X} + \bar{Y}$, so that $E\tilde{\mu} = E(\frac{1}{2}\bar{X} + \bar{Y}) = \frac{1}{2}\mu + \mu/2 = \mu$, so $\tilde{\mu}$ is unbiased, and $\text{Var}\tilde{\mu} = \frac{1}{4}\mu^2/n_1 + \mu^2/(4n_2) = \frac{\mu^2}{4}(\frac{1}{n_1} + \frac{1}{n_2})$.

By the central limit theorem, $\bar{X}$ and $\bar{Y}$ are approximately normally distributed. Since $\bar{X}$ and $\bar{Y}$ are independent, $\tilde{\mu} = \frac{1}{2}\bar{X} + \bar{Y}$ is approximately normally distributed with expected value μ and standard deviation $\frac{\mu}{2}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$. So

$$
\begin{aligned}
0.95 &\approx P\left(-z_{0.025} < \frac{\tilde{\mu}-\mu}{\frac{\mu}{2}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} < z_{0.025}\right) = P\left(-z_{0.025} < \frac{\frac{\tilde{\mu}}{\mu}-1}{\frac{1}{2}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} < z_{0.025}\right) \\
&= P\left(1-\frac{1}{2}z_{0.025}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} < \frac{\tilde{\mu}}{\mu} < 1-\frac{1}{2}z_{0.025}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right) \\
&= P\left(\frac{\tilde{\mu}}{1+\frac{1}{2}z_{0.025}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} < \mu < \frac{\tilde{\mu}}{1-\frac{1}{2}z_{0.025}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right),
\end{aligned}
$$

the double inequality defining a 95% confidence interval for μ. Note that we in the last step of solving the double inequality have assumed that $1-\frac{1}{2}z_{0.025}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} > 0$, that is, $\frac{1}{n_1}+\frac{1}{n_2} < 4/z_{0.025} \approx 1.04$, which is satisfied if $n_2 \geq 2$ and $n_2 \geq 2$, which was obviously already assumed when the central limit theorem was invoked.

When $n_1 = 30$, $n_2 = 20$, $\bar{x} = 2.07$ and $\bar{y} = 0.59$, the 95% confidence interval $(1.27, 2.27)$ is obtained.

The calculations above can be simplified if we make another assumption: that μ in the denominator can be replaced by $\tilde{\mu}$. Then

$$
\begin{aligned}
0.95 &\approx P\left(-z_{0.025} < \frac{\tilde{\mu}-\mu}{\frac{\tilde{\mu}}{2}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} < z_{0.025}\right) = P\left(-z_{0.025} < \frac{1-\frac{\mu}{\tilde{\mu}}}{\frac{1}{2}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} < z_{0.025}\right) \\
&= P\left(\tilde{\mu}\left(1-\frac{1}{2}z_{0.025}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right) < \mu < \tilde{\mu}\left(1+\frac{1}{2}z_{0.025}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)\right).
\end{aligned}
$$

Inserting numerical values, we get the confidence interval $(1.17, 2.08)$.

c) We want $\tilde{\mu}$ unbiased, that is $\mu = E\tilde{\mu} = E(\alpha\bar{X}+\beta\bar{Y}) = \alpha\mu+\beta\mu/c$, giving $\beta/c = 1-\alpha$. We want the variance, $\mathrm{Var}(\tilde{\mu}) = \mathrm{Var}(\alpha\bar{X}+\beta\bar{Y}) = \alpha^2\mathrm{Var}\bar{X}+\beta^2\mathrm{Var}\bar{Y} = \alpha^2\mu^2/n_1+(\beta/c)^2\mu^2/n_2 = \alpha^2\mu^2/n_1+(1-\alpha)^2\mu^2/n_2 = \mu^2(\alpha^2/n_1+(1-\alpha)^2/n_2)$, to be as small as possible. It is easily checked that the second degree polynomial $\alpha^2/n_1+(1-\alpha)^2/n_2$ in α has its minimum at $\alpha = n_1/(n_1+n_2)$, yielding $\beta = c(1-\alpha) = cn_2/(n_1+n_2)$, so that $\tilde{\mu} = (n_1\bar{X}+cn_2\bar{Y})/(n_1+n_2)$, and $\mathrm{Var}\tilde{\mu} = n_1^2/(n_1+n_2)^2\cdot\mu^2/n_1+n_2^2/(n_1+n_2)^2\cdot\mu^2/n_2 = \mu^2/(n_1+n_2)$.

d) We have the likelihood function

$$
L = \prod_{i=1}^{n_1}\frac{1}{\mu}e^{-x_i/\mu}\cdot\prod_{j=1}^{n_2}\frac{c}{\mu}e^{-cy_i/\mu} = c^{-n_2}\mu^{-n_1-n_2}e^{-\frac{1}{\mu}(\sum_{i=1}^{n_1}x_i+c\sum_{j=1}^{n_2}y_j)}
$$

and log-likelihood

$$\ln L = -n_2 \ln c - (n_1 + n_2) \ln \mu - \frac{1}{\mu} \left(\sum_{i=1}^{n_1} x_i + c \sum_{j=1}^{n_2} y_j \right).$$

Setting the partial derivatives

$$\frac{\partial \ln L}{\partial \mu} = -\frac{n_1 + n_2}{\mu} + \frac{1}{\mu^2} \left(\sum_{i=1}^{n_1} x_i + c \sum_{j=1}^{n_2} y_j \right) \quad \text{and} \quad \frac{\partial \ln L}{\partial c} = \frac{n_2}{c} - \frac{1}{\mu} \sum_{j=1}^{n_2} y_j$$

equal to zero we get

$$(n_1 + n_2)\mu = \sum_{i=1}^{n_1} x_i + c \sum_{j=1}^{n_2} y_j \quad \text{and} \quad n_2 \mu = c \sum_{j=1}^{n_2} y_j, \tag{1}$$

respectively. The first equation yields the maximum likelihood estimator

$$\mu^* = \frac{\sum_{i=1}^{n_1} X_i + c \sum_{j=1}^{n_2} Y_j}{n_1 + n_2} = \frac{n_1 \bar{X} + c n_2 \bar{Y}}{n_1 + n_2},$$

which we note is the same estimator as $\tilde{\mu}$ from (c), in the case that c is known. Subtracting the second equation of (1) from the first, we get $n_1 \mu = \sum_{i=1}^{n_1} x_i$ so that $\hat{\mu} = \bar{X}$. Substituting into the second equation, we get $\hat{c} = \bar{X}/\bar{Y}$.