TMA4245 Statistics
Exam May 2016

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Oppgave 1

Gustav and Margrethe have recently finished their studies at NTNU and are now planning to buy flats in Trondheim. Both are looking for a flat in a specific neighbourhood.

We assume that the price per square meter for this neighbourhood is normally distributed. In parts **a**) and **b**) we assume that the mean is $\mu = 30$ kkr (30.000 kr) and the standard deviation is $\sigma = 2.5$ kkr.

a) What is the probability that the price per square meter for a randomly chosen flat is:

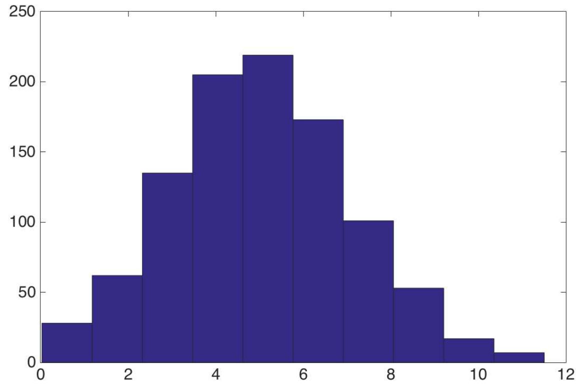
- lower than 30 kkr?
- higher than 25 kkr?
- higher than 25 kkr given that the price per square meter is lower than 30 kkr.

b) Gustav is considering a 40 square meter flat, and Margrethe is considering a 50 square meter flat. Let X_G denote the price per square meter for the flat Gustav is considering and let X_M denote the price per square meter for the flat Margrethe is considering, and use this notation to find an expression for the price (the price for the flat, not the price per square meter) for each of the flats. Further, find also an expression for the price difference between the two flats when we assume that the prices are independent. What is the probability that the flat Margrethe is considering costs less than the flat Gustav is considering?

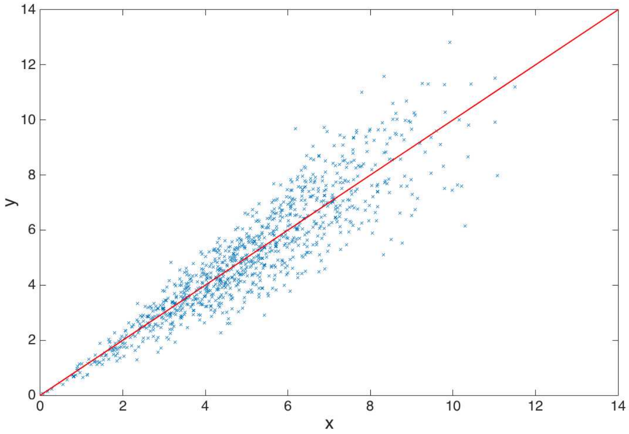
c) Gustav and Margrethe have gathered data $(x_1, x_2, \ldots, x_n)$ for the price per square meter (in kkr) for the last $n = 15$ sales in the neighbourhood, and based on these they want to find a 95% confidence interval for the mean of the price per square meter. Derive an expression for the confidence interval (you may start from a known statistic). Compute the confidence interval numerically when the observed average price per square meter is $\bar{x} = 32$ kkr and $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 74.1$.

Oppgave 2

The company GetData is testing a new sensor which will give cheap observations of the flow of water in pipes. They test the sensor in a realistic situation. In addition to the measurements made by the new sensor, $(y_1, y_2, \ldots, y_n)$, they also measure corresponding true flows of water, $(x_1, x_2, \ldots, x_n)$, for $n = 1000$ independent time periods. Figure 1 shows a histogram of the true flows, and in Figure 2 the true flows (x) are plotted against the flows measured by the new sensor (y).



Figur 1: Histogram of the measured true flows (x)



Figur 2: The observations of true flows (x) against the flow measurements made by the new sensor (y). The solid line shown is $y = x$.

a) Based on Figures 1 and 2, answer the following questions and argue briefly for each answer:

- What is the mean and standard deviation of the true flow (X)? (Both the mean and the standard deviation are integers)
- What is the mean and the standard deviation for the measurements made by the new sensor (Y) given that the true flow is $X = 6$.(Both the mean and the standard deviation are again integers)
- Is the correlation (and covariance) between the true flow (X) and the flow measured by the new sensor (Y) positive, negative or close to zero?

As known, a simple linear regression model is often defined as $Y_i = a + bx_i + \epsilon_i$, for $i = 1, 2, \ldots, n$, where Y_i is the response of interest, a and b are regression parameters, x_i is an explanatory variable which is assumed known, and the noises ϵ_i are assumed to be independent and normally distributed with mean 0 and variance σ_ϵ^2.

b) Answer the following questions and argue briefly for each answer:

- If one fits a simple linear regression model to the data in Figure 2, give approximately the resulting estimates for a and b?
- Based on these estimates for a and b, what is the predicted flow measured by the new sensor (y_0) when the true flow is $x_0 = 4$.
- Discuss whether the assumptions made for a simple linear regression model fits the data in Figure 2.

Oppgave 3

For the company SeeMe the number of visits to their web page is important. Let X_i be the number of visits during t_i hours, and let $X_1, X_2, \ldots, X_n$ be the number of visits in n disjoint time intervals. We assume that visits to the web page are described by a Poisson process with visit intensity λ. Hence $X_1, X_2, \ldots, X_n$ are independent and Poisson distributed random variables with probability distribution

$$f(x_i) = \frac{(\lambda t_i)^{x_i}}{x_i!} e^{-\lambda t_i} \quad \text{for } x_i = 0, 1, 2, \ldots.$$

a) Assume (only for this part) that $\lambda = 10$ and $t_1 = 1$. Find the probabilities

$$P(X_1 = 8) \quad , \quad P(X_1 \geq 8) \quad \text{and} \quad P(8 \leq X_1 \leq 12).$$

We now assume that the visit intensity λ is unknown. SeeMe wants to estimate the intensity λ from observed data for the number of visits in n disjoint time intervals. The following estimators have been proposed,

$$\widetilde{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad , \quad \widehat{\lambda} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} t_i} \quad \text{and} \quad \widehat{\widehat{\lambda}} = \frac{1}{n}\sum_{i=1}^{n} \frac{X_i}{t_i},$$

and it is given that $\mathrm{E}[\widehat{\lambda}] = \lambda$ and $\mathrm{Var}[\widehat{\lambda}] = \lambda / \sum_{i=1}^{n} t_i$.

b) Which of the three estimators do you prefer when $n = 5$ and $t_1 = 1$, $t_2 = 2$, $t_3 = 5$, $t_4 = 1$, $t_5 = 5$? Argue for your answer.

c) Derive the maximum likelihood estimator (MLE) for λ based on $X_1, X_2, \ldots, X_n$.

In parts **d**) and **e**) in this problem you should, regardless of your answers in parts **b**) and **c**), base your answers on the estimator $\widehat{\lambda}$ defined above. Moreover, you may consider it as known that $\lambda \sum_{i=1}^{n} t_i$ is large and use that $\widehat{\lambda}$ is then approximately normally distributed with mean and variance as given above.

SeeMe learns that the web page visit intensity for their largest competitor is $\lambda_0 = 10$ visits per hour, and they want to use the observed number of visits to their own web page, $X_1, X_2, \ldots, X_n$, to decide whether there is reason to believe that the visit intensity to their own web page is higher than this.

d) Formulate the hypotheses H_0 and H_1 for the situation described above.

Specify what test statistic you will use and what (approximate) probability distribution this test statistic has when H_0 is true.

Find the p-value for this hypothesis test when n and $t_1, t_2, \ldots, t_n$ are as given in **b**), and the observed number of visits are $x_1 = 8$, $x_2 = 20$, $x_3 = 48$, $x_4 = 10$ and $x_5 = 62$. Based on the value you find for the p-value, discuss briefly whether there is reason to believe that the visit intensity to the web page of SeeMe is higher than for web page of the competitor.

e) If one uses significance level $\alpha = 0.05$ in the hypothesis test in **d**), how large must the visit intensity for SeeMe be to give at least probability 0.9 for concluding that the intensity of SeeMe is higher than the intensity of the competitor? When answering this question, use the same values for n and $t_1, t_2, \ldots, t_n$ as in part **b**).

Fasit

1. **a**) $0.5, 0.9772, 0.9544$ **b**) $40X_G$, $50X_M$, $40X_G - 50X_M$, 0.0307 **c**) $(30.7, 33.3)$

2. **a**) $5, 2$ **b**) $a = 0, b = 1$, $\widehat{y} = 4$

3. **a**) 0.1126, 0.7798, 0.5714 **b**) Prefer $\widehat{\lambda}$ **d**) $H_0 : \lambda = \lambda_0$, $H_1 : \lambda > \lambda_0$, 0.2483 **e**) $\lambda \geq 12.6068$